Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης
Εργαστηριο Ορασης Υπολογιστων, Επικοινωνιας Λογου και Επεξεργασιας Σηματων

Οπτικοακουστική Μοντελοποίηση Προσοχής και Πολυ-αισθητηριακή Κατανόηση Ακουστικής Σκηνής

Διδακτορικη Διατριβη

της

ΑΝΤΙΓΟΝΗΣ ΤΣΙΑΜΗ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού & Μηχανικού Υπολογιστών Ε.Μ.Π.

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Δεκέμβριος 2019

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης

Οπτικοακουστική Μοντελοποίηση Προσοχής και Πολυ-αισθητηριακή Κατανόηση Ακουστικής Σκηνής

Διδακτορικη Διατριβη

της

ΑΝΤΙΓΟΝΗΣ ΤΣΙΑΜΗ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού & Μηχανικού Υπολογιστών Ε.Μ.Π.

Συμβουλευτική Επιτροπή: Καθ. Πέτρος Μαραγκός (Επιβλέπων)
Αναπλ. Καθ. Κωνσταντίνος Τζαφέστας
Αναπλ. Καθ. Γεράσιμος Ποταμιάνος

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 19/12/2019.

.....	.....	.....
Πέτρος Μαραγκός	Κωνσταντίνος Τζαφέστας	Γεράσιμος Ποταμιάνος
Καθηγητής	Αναπληρωτής Καθηγητής	Αναπληρωτής Καθηγητής
Ε.Μ.Π.	Ε.Μ.Π.	Παν/μίου Θεσσαλίας
.....	.....	.....
Ευίτα-Σταυρούλα Φωτεινέα	Παναγιώτης Τσανάκας	Αργυρώ Βατάκη
Ερευνήτρια Α΄	Καθηγητής	Επίκουρη Καθηγήτρια
Ε.Κ. Αθηνά	Ε.Μ.Π.	Παντείου Παν/μίου
.....		
Kostas Daniilidis		
Καθηγητής		
Univ. of Pennsylvania		

Αθήνα, Δεκέμβριος 2019

(Υπογραφή)

.....

Αντιγόνη Τσιάμη
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Αντιγόνη Τσιάμη, 2019.
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Στη διατριβή αυτή μελετάται η δυνατότητα ανάπτυξης ενός μοντέλου οπτικοακουστικής προσοχής/εμφάνειας βασισμένου σε ευρήματα από συμπεριφορικά πειράματα, το οποίο προβλέπει τα σημεία εστίασης του ανθρώπινου ματιού σε μια δισδιάστατη οπτικοακουστική σκηνή. Η μελέτη αυτή πραγματοποιείται με δύο τρόπους: με τεχνικές επεξεργασίας σήματος και με βαθιά νευρωνικά δίκτυα. Όσον αφορά τις τεχνικές επεξεργασίας σήματος, διερευνώνται διάφορες τεχνικές σύμμειξης οπτικής και ακουστικής εμφάνειας με ήδη υπάρχοντα μοντέλα. Τα δημιουργούμενα οπτικοακουστικά μοντέλα επαληθεύονται τόσο με ευρήματα από συμπεριφορικά πειράματα, όσο και με δεδομένα οφθαλμοκινητικής παρακολούθησης αλλά και δεδομένα fMRI. Τα αποτελέσματα τόσο από τα συμπεριφορικά πειράματα όσο και από τα πειράματα με οφθαλμοκινητικά δεδομένα υποδεικνύουν ότι η οπτικοακουστική μοντελοποίηση βελτιώνει την επίδοση του συστήματος πρόβλεψης σημείων εστίασης του ματιού. Όσον αφορά τα βαθιά νευρωνικά δίκτυα, παρουσιάζεται ένα χωρο-χρονικό οπτικοακουστικό δίκτυο εμφάνειας, το οποίο περιέχει ένα οπτικό δίκτυο εμφάνειας, ένα δίκτυο ακουστικών αναπαραστάσεων, μία μονάδα εντοπισμού του ήχου στο βίντεο και μία μονάδα σύμμειξης της οπτικής και της ακουστικής εμφάνειας. Όλα αυτά ενσωματώνονται κάτω από ένα ενιαίο δίκτυο το οποίο εκπαιδεύεται end-to-end. Το δίκτυο αξιολογείται σε βάσεις οφθαλμοκινητικών δεδομένων. Τα αποτελέσματα των συγκρίσεων με άλλες μεθόδους υποδεικνύουν την αποτελεσματικότητα της μοντελοποίησης, ανοίγοντας το δρόμο για την εκτίμηση εμφάνειας 'in-the-wild'.

Παράλληλα έχει πραγματοποιηθεί έρευνα στην κατανόηση ακουστικής σκηνής και συγκεκριμένα έχει αναπτυχθεί ένα σύστημα εντοπισμού ομιλητή, ένα σύστημα αναγνώρισης φωνής στα Ελληνικά και στα Αγγλικά και ένα σύστημα διαλόγου και κατανόησης λόγου. Τα συστήματα έχουν προσαρμοστεί και εφαρμοστεί είτε σε περιβάλλοντα έξυπνου σπιτιού ή/και σε περιβάλλοντα αλληλεπίδρασης ανθρώπων/παιδιών και ρομπότ με πολλαπλούς αισθητήρες. Επίσης, γίνεται αξιολόγησή τους σε κατάλληλες βάσεις δεδομένων.

Τέλος, εκτός από την ανάπτυξη νέων αλγοριθμικών μεθόδων για τα παραπάνω προβλήματα, σημαντικό μέρος της διατριβής αποτελεί και η συλλογή νέων μεγάλων βάσεων δεδομένων: Συλλέχθηκε μία βάση με δεδομένα οφθαλμοκινητικής παρακολούθησης από 37 βίντεο και 20 συμμετέχοντες, καθώς και μία πολυκαναλική βάση με δεδομένα φωνής στα Ελληνικά από 20 ομιλητές.

λέξεις-κλειδιά: μηχανική μάθηση, επεξεργασία σημάτων, οπτικοακουστική προσοχή

Abstract

The scope of this work is the investigation and development of a 2D computational audiovisual saliency model based on behavioral findings that will be able to produce accurate human fixation predictions in a 2D audiovisual scene, i.e. in videos. The investigation is carried out in two different ways: with signal processing techniques and with deep learning techniques. Regarding the former, several fusion schemes between visual and auditory saliency models are being investigated, and the resulting models are behaviorally validated through comparisons with results from behavioral experiments and evaluated with audiovisual human eye-tracking data and fMRI data. Results from both behavioral and eye-tracking experiments indicate that audiovisual saliency modeling indeed improves saliency estimation performance. Regarding deep learning techniques, a new spatio-temporal audiovisual saliency network is developed, that includes a visual saliency network, an audio representation network, a sound localization module, and an audiovisual saliency fusion module. All modules are integrated under a single network that is trained end-to-end. The network performance is evaluated in several eye-tracking databases and results from comparisons to other methods highlight the effectiveness of the presented network, that opens the way for estimating saliency "in-the-wild".

In parallel, research has been carried out in the direction of auditory scene understanding. Specifically, a speaker localization system has been developed, as well as a baseline distant speech recognition system in Greek and English and a speech understanding/dialog system. These systems have been adapted and applied either to a smart home environment and/or to a multi-sensory human/child - robot interaction application. They are also evaluated through experiments in appropriate databases.

Finally, except for the development and evaluation of new algorithmic methods to successfully address the above problems, an important contribution of this thesis lies in the collection of two new databases: An audiovisual human eye-tracking database employing 20 subjects and 37 videos has been collected, as well as a multi-channel speech database in Greek with data from 20 speakers.

keywords: machine learning, signal processing, audiovisual attention modeling, speech recognition

Περιεχόμενα

Περίληψη **5**

Abstract **7**

Περιεχόμενα **10**

Κατάλογος Σχημάτων **14**

Κατάλογος Πινάκων **16**

Πρόλογος **17**

1 Εισαγωγή **19**
1.1 Συνεισφορά 21
1.2 Δομή της διατριβής 26

2 Υπόβαθρο και επισκόπηση ερευνητικού πεδίου **27**
2.1 Πολυκαναλική επεξεργασία φωνής σε περιβάλλοντα έξυπνου σπιτιού 27
2.2 Συστήματα αλληλεπίδρασης ανθρώπου-ρομπότ 30
2.3 Μοντελοποίηση οπτικοακουστικής προσοχής 32

3 Πολυκαναλική επεξεργασία φωνής σε έξυπνα/ρομποτικά περιβάλλοντα **41**
3.1 Εντοπισμός ομιλητή με χρήση πολλαπλών μικροφώνων 41
3.2 Οπτικοακουστικός εντοπισμός ομιλητή 50
3.3 Πολυκαναλική αποθορυβοποίηση φωνής 53
3.4 Αναγνώριση φωνής από απόσταση 57
3.5 Σύστημα διαλόγου/κατανόησης λόγου i-Walk 63

4 Βάση 'Αθηνά': Πολυκαναλικά φωνητικά δεδομένα στα Ελληνικά **69**
4.1 Χώρος ηχογράφησης και εξοπλισμός 70
4.2 Περιγραφή βάσης 71
4.3 Διαδικασία συλλογής δεδομένων και επισημείωση 74

5 Συλλογή δεδομένων οφθαλμοκινητικής παρακολούθησης 77
5.1 Διαδικασία καταγραφής δεδομένων οφθαλμοκινητικής παρακολούθησης (eye-tracking data) . . . 77

6 Υπολογιστικό μοντέλο πρόβλεψης οπτικοακουστικής προσοχής σε βίντεο 83
6.1 Εισαγωγή . . . 83
6.2 Οπτικοακουστική μοντελοποίηση . . . 84
6.3 Αξιολόγηση . . . 91
6.4 Σύνοψη . . . 109

7 Χωρο-χρονικό οπτικοακουστικό μοντέλο εμφάνειας STAViS 111
7.1 Εισαγωγή . . . 111
7.2 Χωρο-χρονικό Οπτικοακουστικό μοντέλο εμφάνειας STAViS . . . 113
7.3 Πειράματα . . . 120
7.4 Συμπεράσματα . . . 123

8 Συμπεράσματα και μελλοντικές κατευθύνσεις 127
8.1 Συμπεράσματα . . . 127
8.2 Μελλοντικές κατευθύνσεις . . . 129

Βιβλιογραφία 131

Α΄ Λίστα δημοσιεύσεων 155

Κατάλογος Σχημάτων

3.1 Παράδειγμα της προτεινόμενης μεθόδου περιορισμού αποκλίνοντων σημείων (με μπλε $*$ συμβολίζεται η πραγματική θέση του ομιλητή, με κόκκινο $+$ η εκτίμηση της θέσης του ομιλητή, με μαύρα x οι θέσεις των μικροφώνων και με διακεκομμένη γραμμή η μέγιστη απόσταση από την εκτιμώμενη θέση προς κάποιο πιθανό αποκλίνον σημείο.) 43

3.2 Κάτοψη του έξυπνου σπιτιού DIRHA όπου απεικονίζονται όλες οι πιθανές θέσεις και προσανατολισμοί των ομιλητών καθώς και τα 40 μικρόφωνα([1]). . 46

3.3 Κάτοψη της βάσης DMN 47

3.4 Κάτοψη του χώρου των πειραμάτων. 50

3.5 Οπτικοακουστικός εντοπισμός ομιλητή. 51

3.6 Παράδειγμα του οπτικοακουστικού εντοπισμού ομιλητή. Αριστερά: Ακουστικός εντοπισμός ομιλητή (όσο υψηλότερες είναι οι SRP τιμές τόσο πιο κόκκινη είναι η εικόνα). Μέση: Οπτικός εντοπισμός ανθρώπων. Δεξιά: Οπτικοακουστικός εντοπισμός ομιλητή. Φαίνονται επίσης οι 3 αισθητήρες Kinect. 53

3.7 Σύστημα πολυκαναλικής αποθορυβοποίησης ομιλίας. 54

3.8 Οι πενταγωνικές συστοιχίες μικροφώνων της οροφής στη βάση 'Αθηνά'. . . . 56

3.9 Αριστερά: Κάτοψη του χώρου και της κατανομής των αισθητήρων και των διαθέσιμων ρομπότ στο σύστημα Multi3. Δεξιά: Φωτογραφία της διάταξης. . . 58

3.10 Πολυτροπικό, πολυ-καναλικό σύστημα ρομποτικής αντίληψης και πολυ-ρομποτική αρχιτεκτονική. Τα A1, A2, A3 αναφέρονται στη συστοιχία μικροφώνων Kinect. 58

3.11 Στιγμιότυπο του παιχνιδιού *Φτιάξε μία φάρμα*. 61

3.12 Παράδειγμα των διαφόρων ροών πληροφορίας που καταγράφηκαν από 4 κάμερες Kinect στον πολυ-καναλικό πειραματικό χώρο: απεικονίζονται τα καρέ του βίντεο, ο ήχος και ο σκελετός του χρήστη. 63

3.13 Αριστερά: Ένας ασθενής που περπατά υποστηριζόμενος από την πλατφόρμα i-Walk. Δεξιά: Ένας ασθενής πραγματοποιεί ασκήσης αποκατάστασης κάτω από την επίβλεψη του συστήματος i-Walk. 65

3.14 Επισκόπηση του πολυτροπικού έξυπνου βοηθού i-Walk 65

4.1 Κάτοψη του έξυπνου χώρου 'Αθηνά'. Απεικονίζονται οι αισθητήρες, οι θέσεις ομιλητή και ακουστικών γεγονότων. 71

4.2 RGB, Depth, Skeleton, MEMS έξοδοι για την εκφώνηση μιας φράσης ενεργοποίησης μαζί με τη χειρονομία. 74

4.3 Εργαλείο επισημείωσης ELAN: Οι διαφορετικές γραμμές αντιπροσωπεύουν λόγο, θέσεις, γεγονότα, κλπ. Το κόκκινο περίγραμμα υποδεικνύει τη φράση προς εκφώνηση. 74

5.1 Ενδεικτικά καρέ από κάθε ένα από τα βίντεο της βάσης SumMe και δίπλα η υπέρθεση όλων των οφθαλμοκινητικών δεδομένων για όλα τα καρέ του κάθε βίντεο. 80

5.2 Ενδεικτικά καρέ από κάθε ένα από τα βίντεο της βάσης ETMD και δίπλα η υπέρθεση όλων των οφθαλμοκινητικών δεδομένων για όλα τα καρέ του κάθε βίντεο. 81

6.1 Επισκόπηση του μοντέλου οπτικοακουστικής εμφάνειας. 85

6.2 Οι πάνω εικόνες από το [2] απεικονίζουν δύο ενδεικτικά καρέ από τα "pip & pop" ερεθίσματα, κατά τη στιγμή που αναβοσβήνει ο στόχος (η κάθετη γραμμή στην κάτω αριστερή γωνία). Οι κάτω εικόνες αντιστοιχούν στον οπτικό (αριστερά) και τον οπτικοακουστικό (δεξιά) χάρτη εμφάνειας που παράχθηκε με απευθείας σύμμειξη. 90

6.3 Οι πάνω εικόνες από το [3] απεικονίζουν δύο ενδεικτικά καρέ από τα "sine vs square" ερεθίσματα για την τετραγωνική διαμόρφωση τη στιγμή που αναβοσβήνει ο στόχος (η κάθετη γραμμή στην κάτω δεξιά γωνία). Στην κάτω σειρά απεικονίζονται ο οπτικός (αριστερά) και ο οπτικοακουστικός (δεξιά) χάρτης εμφάνειας που προκύπτει με απευθείας σύμμειξη. 91

6.4 (α) Πρωτότυπο σχήμα από το [2] και (β, γ, δ) τα *AUC, NSS, CC* για το set size πείραμα με τα διάφορα σχήματα σύμμειξης. Το μπλε χρώμα αντιπροσωπεύει την απευθείας σύμμειξη, το πράσινο τη σύμμειξη με ετεροσυσχέτιση, το φούξια τη σύμμειξη με αμοιβαία πληροφορία και το κόκκινο την περίπτωση όπου απουσιάζει ο ήχος. 91

6.5 (α) Πρωτότυπο σχήμα από το [2] και (β, γ, δ) τα *AUC, NSS, CC* για το temporal asynchrony πείραμα. Οι τιμές με μείον αναφέρονται στις περιπτώσεις όπου ο ήχος προηγείται της εικόνας. Τα διαφορετικά χρώματα αντιπροσωπεύουν τις διαφορετικές τεχνικές σύμμειξης όπως και παραπάνω. 92

6.6 (α) Πρωτότυπο σχήμα από το [3] για την ημιτονοειδή διαμόρφωση και (β, γ, δ) τα αποτελέσματα του *NSS* για (direct, CC, MI) σύμμειξη αντίστοιχα για το set size experiment στα ερεθίσματα sine vs square. 92

6.7 (α) Πρωτότυπο σχήμα από το [3] για την τετραγωνική διαμόρφωση και (β, γ, δ) τα αποτελέσματα του *NSS* για (direct, CC, MI) σύμμειξη αντίστοιχα για το set size experiment στα ερεθίσματα sine vs square. 92

6.8 Παράδειγμα συνεχόμενων καρέ όπου έχει γίνει υπέρθεση των δεδομένων οφθαλμοκινητικής παρακολούθησης που έχουν συλλεχθεί, για τη SumMe (αριστερά) και την ETMD (δεξιά). Η δεύτερη και τρίτη γραμμή απεικονίζουν τους αντίστοιχους οπτικούς και οπτικοακουστικούς χάρτες εμφάνειας. 101

6.9 Παράδειγμα από τη βάση δεδομένων AVAD. a) Παρουσιάζονται τα πρωτότυπα καρέ μαζί με τα δεδομένα οπτικοακουστικής παρακολούθησης μαζί με b) τους χάρτες οπτικής και c) οπτικοακουστικής εμφάνειας. d) Ακουστική κυματομορφή και e) η οπτική καμπύλη εμφάνειας. f) Απεικονίζεται η οπτική μόνο (κόκκινο χρώμα) και η οπτικοακουστική (πράσινο χρώμα) καμπύλη εμφάνειας ως προς τη μετρική NSS. Στην αρχή του βίντεο υπάρχει μόνο ακουστικός θόρυβος, που δε σχετίζεται με το οπτικό περιεχόμενο. Το πραγματικό ακουστικό γεγονός εμφανίζεται περί τα μέσα του βίντεο. Ως τότε, η οπτικοακουστική και η οπτική εμφάνεια είναι περίπου ίδιες όπως φαίνεται από τις NSS καμπύλες, ενώ όταν εμφανίζεται το ακουστικό γεγονός, η οπτικοακουστική καμπύλη ξεπερνά αισθητά την οπτική. 102

6.10 Παράδειγμα από τη βάση δεδομένων ETMD. a) Παρουσιάζονται τα πρωτότυπα καρέ μαζί με τα δεδομένα οπτικοακουστικής παρακολούθησης μαζί με b) τους χάρτες οπτικής και c) οπτικοακουστικής εμφάνειας. d) Ακουστική κυματομορφή και e) η οπτική καμπύλη εμφάνειας. f) Απεικονίζεται η οπτική μόνο (κόκκινο χρώμα) και η οπτικοακουστική (πράσινο χρώμα) καμπύλη εμφάνειας ως προς τη μετρική NSS. Επίσης το μπλε ορθογώνιο σηματοδοτεί το οπτικοακουστικό γεγονός (καλπασμός αλόγου). Παρόλο που οι διαφορές οπτικοακουστικής και οπτικής εμφάνειας είναι μικρές εδώ, μπορούμε να παρατηρήσουμε ότι κατά τη διάρκεια του ακουστικού γεγονότος, η τιμή της μετρικής NSS είναι λίγο υψηλότερη για την οπτικοακουστική καμπύλη, ενώ πριν και μετά το γεγονός είναι σχεδόν ίδιες ή εναλλάσσονται ως προς το ποια είναι καλύτερη ή χειρότερη. . . 103

6.11 Επισκόπηση του μοντέλου οπτικοακουστικής χρονικής εμφάνειας: a) σύμμειξη στο επίπεδο των χαρτών b) σύμμειξη στο επίπεδο των καμπυλών. 106

6.12 Αποτελέσματα για την οπτικοακουστική εμφάνεια με τους δύο διαφορετικούς τρόπους σύμμειξης στο επίπεδο των *χαρτών εμφάνειας* (t-test). 108

6.13 Αποτελέσματα για την οπτικοακουστική εμφάνεια με τους δύο διαφορετικούς τρόπους σύμμειξης στο επίπεδο των *καμπυλών εμφάνειας* (t-test). 108

7.1 Ενδεικτικά καρέ με τα δεδομένα οφθαλμοκινητικής παρακολούθησης, για μία καμπάνα που χτυπά. Η δεύτερη σειρά απεικονίζει τους οπτικούς χάρτες εμφάνειας, ενώ η τρίτη τους οπτικοακουστικούς χάρτες που δημιουργούνται από το προτεινόμενο μοντέλο STAViS. Είναι εμφανές ότι το τελευταίο αποτυπώνει καλύτερα την ανθρώπινη προσοχή. 112

7.2 Η αρχιτεντονική του δικτύου STAViS: το χωρο-χρονικό οπτικοακουστικό δίκτυο είναι βασισμένο στην αρχιτεκτονική ResNet και αποτελείται από ένα χωρο-χρονικό οπτικό, ένα ακουστικό μονοπάτι και τη σύμμειξή τους. 114

7.3 Η μονάδα DSAM (Deeply Supervised Attention Module) βελτιώνει τις καθολικές αναπαραστάσεις του δικτύου και παρέχει τους πολυ-επίπεδους χάρτες χωρο-χρονικής οπτικής εμφάνειας. 115
7.4 Ενδεικτικά καρέ από τη βάση Coutrot1 μαζί με τα δεδομένα οφθαλμοκινητικής παρακολούθησης που τους αντιστοιχούν, και τους αντίστοιχους ground truth, χωρο-χρονικούς οπτικούς και οπτικοακουστικούς χάρτες όπως υπολογίστηκαν από το προτεινόμενο δίκτυο STAViS (οπτικό-μόνο και οπτικοακουστικό). Επιπλέον απεικονίζεται η καμπύλη NSS σε σχέση με το χρόνο για την οπτική-μόνο και για την οπτικοακουστική περίπτωση. 124
7.5 Ενδεικτικά καρέ από τη βάση Coutrot2 μαζί με τα δεδομένα οφθαλμοκινητικής παρακολούθησης που τους αντιστοιχούν, και τους αντίστοιχους ground truth, χωρο-χρονικούς οπτικούς και οπτικοακουστικούς χάρτες όπως υπολογίστηκαν από το προτεινόμενο δίκτυο STAViS (οπτικό-μόνο και οπτικοακουστικό). Επιπλέον απεικονίζεται η καμπύλη NSS σε σχέση με το χρόνο για την οπτική-μόνο και για την οπτικοακουστική περίπτωση. 125
7.6 Ενδεικτικό καρέ, ένα για κάθε βάση από τις ETMD, AVAD, DIEM και SumMe με τα αντίστοιχα δεδομένα οφθαλμοκινητικής παρακολούθησης και τους χάρτες που προκύπτουν από το ground truth, το STAViS και άλλες state-of-the-art χωρο-χρονικές μεθόδους οπτικής εμφάνειας για συγκρίσεις. 126

Κατάλογος Πινάκων

3.1 Οι διάφορες προσεγγίσεις εντοπισμού ομιλητή που ακολουθήθηκαν. 48

3.2 Αποτελέσματα στη βάση DMN. 48

3.3 Αποτελέσματα στις βάσεις DIRHA. 50

3.4 Αξιολόγηση του συστήματος οπτικοακουστικού εντοπισμού ομιλητή. 52

3.5 Αποτελέσματα αποθορυβοποίησης στη βάση προσωμειωμένων δεδομένων DIRHA στα Ελληνικά ανά δωμάτιο και σετ δεομένων. Το LR συμβολίζει το σαλόνι και K την κουζίνα. Όλα τα αποτελέσματα είναι σε dB. 55

3.6 Πολυκαναλική αποθορυβοποίηση ομιλίας στη βάση 'Αθηνά'. Τα αποτελέσματα παρουσιάζονται ως προς το SSNRE σε dB. 57

3.7 Μέσο word (WCOR) και sentence accuracy (SCOR) (%) για την αναγνώριση φωνής σε παιδιά για το παιχνίδι της παντομίμας. Παρουσιάζονται αποτελέσματα για κάθε Kinect ξεχωριστά και για τη συγχώνευση των υποθέσεων για όλους τους τρόπους adaptation. 62

3.8 Αξιολόγηση του συστήματος αναγνώρισης φωνής για το παιχνίδι 'Φτιάξε μια φάρμα'. 63

3.9 Δημογραφικές πληροφορίες και σκορ αξιολόγησης για τους 13 ασθενείς της βάσης i-Walk. 67

4.1 Τα ηχητικά δεδομένα που συλλέχθηκαν, κατηγοριοποιημένα σε ομιλία, ακουστικά γεγονότα και θορύβους. Ο αριθμός εμφανίσεών τους (#) και η συνολική τους διάρκεια σαν ποσοστό επί των 240 λεπτών της βάσης. Τα ακουστικά γεγονότα χωρίζονται σε μικρά και μεγάλα με μέση διάρκεια 1.5sec και 3sec αντίστοιχα. 72

4.2 Οι ακολουθίες λόγου ανά συνεδρία. DK είναι η φράση ενεργοποίησης, DC η εντολή, PH η φωνητικά πλούσια πρόταση ανδ CS ο διάλογος. S1 είναι ο βασικός ομιλητής και S2 ο συνομιλητής όταν υπάρχει διάλογος. 73

6.1 Τα αποτελέσματα για τα διάφορα μοντέλα και σχήματα σύμμειξης (με ακρωνύμια και σύντομη περιγραφή) για τη βάση δεδομένων DIEM. 98

6.2 Τα αποτελέσματα για τα διάφορα μοντέλα και σχήματα σύμμειξης για τη βάση AVAD (* επαναξιολόγηση με το παρόν πλαίσιο αξιολόγησης) 99

6.3 Τα αποτελέσματα για τα διάφορα μοντέλα και σχήματα σύμμειξης για τις βάσεις Coutrot 100
6.4 Τα αποτελέσματα των διαφόρων μοντέλων και σχημάτων σύμμειξης για τη βάση SumMe και τη βάση ETMD. 101
6.5 Θέσεις των κορυφών ενεργοποίησης για το *οπτικοακουστικό* μοντέλο εμφάνειας (σύμμειξη στο επίπεδο καμπυλών και μέγιστο). 109

7.1 Ablation study: Τα L_1, L_2 και L_3 αναφέρονται στη μέθοδο εντοπισμού του ήχου: cosine, inner product και bilinear αντίστοιχα. Τα S_1^{av}, S_2^{av}, S_3^{av} και S_{fus}^{av} αναφέρονται στους διφορετικούς τρόπους σύμμειξης. Ο υπερδείκτης "mul" αφορά στους πολλαπλούς χάρτες χαρακτηριστικών. 120
7.2 Αποτελέσματα αξιολόγησης εμφάνειας για τις βάσεις δεδομένων DIEM, Coutrot1 και Coutrot2. Τα αποτελέσματα για την προτεινόμενη μέθοδο (STAViS [STA]) απεικονίζονται στην πρώτη γραμμή του πίνακα, ενώ η δεύτερη γραμμή αφορά στην εκδοχή του δικτύου μας που υπολογίζει οπτική-μόνο εμφάνεια. Στις περισσότερες περιπτώσεις, το προτεινόμενο δίκτυο ξεπερνά τις άλλες υπάρχουσες state-of-the-art μεθόδους για εκτίμηση εμφάνειας, σύμφωνα με τις 5 μετρικές αξιολόγησης. Το [STA] αναφέρεται σε χωρο-χρονικό οπτικοακουστικό, το [ST] σε χωρο-χρονικό οπτικό, ενώ το [S] υποδηλώνει χωρικό μόνο μοντέλο το οποίο εφαρμόζεται σε κάθε καρέ ανεξάρτητα. 120
7.3 Αποτελέσματα αξιολόγησης εμφάνειας για τις βάσεις δεδομένων AVAD, SumMe και ETMD. Τα αποτελέσματα για την προτεινόμενη μέθοδο (STAViS [STA]) απεικονίζονται στην πρώτη γραμμή του πίνακα, ενώ η δεύτερη γραμμή αφορά στην εκδοχή του δικτύου μας που υπολογίζει οπτική-μόνο εμφάνεια. Στις περισσότερες περιπτώσεις, το προτεινόμενο δίκτυο ξεπερνά τις άλλες υπάρχουσες state-of-the-art μεθόδους για εκτίμηση εμφάνειας, σύμφωνα με τις 5 μετρικές αξιολόγησης. Το [STA] αναφέρεται σε χωρο-χρονικό οπτικοακουστικό, το [ST] σε χωρο-χρονικό οπτικό, ενώ το [S] υποδηλώνει χωρικό μόνο μοντέλο το οποίο εφαρμόζεται σε κάθε καρέ ανεξάρτητα. 121

Πρόλογος

Το διδακτορικό μου το βίωσα σαν ένα μεγάλο ταξίδι. Όχι μόνο σαν ένα επιστημονικό ταξίδι, αλλά κυρίως σαν ένα εσωτερικό ταξίδι. Μπήκα σε αυτό με σχεδόν πλήρη άγνοια του τι θα συναντήσω, ορμώμενη από μία δημιουργική περιέργεια και όρεξη και βγαίνω από αυτό με πολύ μεγαλύτερη συνειδητότητα και καλύτερη γνώση του εαυτού μου. Όταν το ξεκίνησα, το φανταζόμουν με σαφώς ορισμένα όρια, με αρχή, μέση, και ένα όμορφο, φαντασμαγορικό τέλος. Η πραγματικότητα όμως με έκανε να εστιάσω στο ταξίδι, άλλοτε γεμάτο με ομορφιές κι άλλοτε με απογοητεύσεις, το οποίο δεν έχει σημασία από που σε παίρνει και που θα καταλήξεις, αλλά κυρίως να απολαύσεις τη διαδρομή. Και τη διαδρομή αυτή πραγματικά την απόλαυσα.

Θα ήθελα να ευχαριστήσω μέσα από την καρδιά μου τον επιβλέποντά μου Πέτρο Μαραγκό για την ευκαιρία αυτή που μου έδωσε να έρθω σε επαφή με όλον αυτό τον συναρπαστικό κόσμο, που με εμπιστεύτηκε και με άφησε με ελευθερία να περιπλανηθώ στα ερευνητικά μονοπάτια. Οι συμβουλές και η καθοδήγησή του ήταν καίριας σημασίας.

Σε αυτό το ταξίδι όμως είχα και συνταξιδιώτες. Άλλους με τους οποίους ξεκινήσαμε και τελειώνουμε μαζί αυτή τη βόλτα, άλλους που πέρασαν και έφυγαν και άλλους που θα μείνουν. Δεν έχω λόγια να εκφράσω την ευγνωμοσύνη μου και τη χαρά μου για τους ανθρώπους που γνώρισα. Όχι μόνο γιατί μοιραστήκαμε κοινές αγωνίες, χαρές, λύπες, ξενύχτια, διακοπές, ακόμη και σπίτια, αλλά κυρίως γιατί έκανα πραγματικούς φίλους, ανθρώπους που θα μείνουν στη ζωή μου. Αν αρχίσω να γράφω για τον καθένα ξεχωριστά, χρειάζομαι άλλες τόσες σελίδες όσες είναι και η διατριβή μου. Θέλω να υπερευχαριστήσω μέσα από την καρδιά μου τον Πέτρο Κούτρα. Για τη στήριξη, την εμψύχωση και τη βοήθεια τόσο πρακτικά όσο και ψυχολογικά και τα άπειρα ξενύχτια και γέλια που κάναμε όλα αυτά τα χρόνια. Ό,τι κι αν γράψω είναι πολύ λίγο. Το Νάσο για την καθοδήγηση και την όρεξη που μου μετέδωσε ειδικά στα πρώτα χρόνια. Τη Νίκη, τη Βίκυ, τη Τζο, την Αροδάμη, τον Παναγιώτη Φιλντίση, την Ξανθή, τη Γεωργία, τον Παναγιώτη Γιαννούλη, τον Ισίδωρο, το Γιώργο Ρετσινά, τον Αντώνη, το Θανάση για όλες τις στιγμές που μοιραστήκαμε όλα αυτά τα χρόνια εντός και εκτός Πολυτεχνείου. Τη Δέσποινα, το Νίκο, το Γιώργο Παυλάκο, το Χρήστο, τον Παναγιώτη Μέρμηγκα, τη Νάνσυ, το Βασίλη, το Σταύρο και τη Φωτεινή. Το λαμπ τελικά είναι οι άνθρωποι που το απαρτίζουν και πιστεύω ότι είναι αδύνατο να βρεθεί άλλο τέτοιο λαμπ σε όλη τη γη.

Θα ήθελα να ευχαριστήσω επίσης τον κύριο Μάκη Ποταμιάνο για τη συνεργασία που είχαμε και την ερευνητική καθοδήγηση που μου προσέφερε, καθώς και τον κύριο Κωνσταντίνο Τζαφέστα, που παρακολουθούσε το διδακτορικό μου σχεδόν από τα πρώτα χρόνια. Επίσης, την Ευίτα Φωτεινέα από το Ε.Κ. Αθηνά, πάντα πρόθυμη να βοηθήσει και τον κ. Παναγιώτη

Τσανάκα για τις παρατηρήσεις του από μια άλλη ερευνητική σκοπιά. Επίσης, την Αργυρώ Βατάκη, που με εισήγαγε στον κόσμο των νευροεπιστημών και αποτέλεσε το έναυσμα για να ασχοληθώ με την οπτικοακουστική προσοχή και τον κ. Κώστα Δανιηλίδη που μου προσέφερε τη μοναδική ευκαιρία να δουλέψω για λίγους μήνες με εκείνον και την ομάδα του στην Αμερική.

Θέλω επίσης ξεχωριστά να ευχαριστήσω την Έλενα και τη Μάρθα που είναι δίπλα μου όλα αυτά τα χρόνια και για τις συζητήσεις μας και τις συμβουλές που μου παρείχαν από τα διάφορα μέρη του πλανήτη στα οποία βρίσκονταν, καθώς επίσης και τον Πάνο για την στήριξη και την ηρεμία του. Την Καίτη και το Γιώργο, την Ανθή, την Πολυτίμη και την Ιλεάνα για τις πολύτιμες συζητήσεις μας.

Τέλος, τίποτα από όλα αυτά δεν θα ήταν εφικτό χωρίς την οικογένεια μου και την ανιδιοτελή, σταθερή, ανεξάντλητη αγάπη και στήριξη που μου έχουν δώσει όλα αυτά τα χρόνια. Ευχαριστώ από τα βάθη της ψυχής μου τη μαμά μου Γιώτα και τον μπαμπά μου Λευτέρη που με βοήθησαν να γίνω ο άνθρωπος που είμαι. Ξεχωριστά υπερευχαριστώ τον αδερφό μου Τάσο, μόνιμο συνοδοιπόρο στη ζωή από τότε που θυμάμαι τον εαυτό μου, ο οποίος με έχει βοηθήσει και στηρίξει ακούραστα σε όλα τα επίπεδα και τον θαυμάζω απεριόριστα.

Ο κύκλος του διδακτορικού τελειώνει, με αποσκευές γεμάτες εμπειρίες και συναισθήματα, αλλά το ταξίδι συνεχίζεται...

Κεφάλαιο 1

Εισαγωγή

Η τεχνητή νοημοσύνη έχει διεισδύσει ραγδαία στις ζωές των ανθρώπων τα τελευταία χρόνια, επαναστατικοποιώντας πάρα πολλούς τομείς της καθημερινότητας τους. Η τεχνολογία τόσο σε επίπεδο υλικού όσο και σε λογισμικού, συνεχώς εξελίσσεται, επιλύοντας προβλήματα και δημιουργώντας την ανάγκη για την επίλυση καινούριων. Ο ορισμός της τεχνητής νοημοσύνης συνεχώς αμφισβητείται: Εφόσον η νοημοσύνη των μηχανών συνεχώς αυξάνεται, εργασίες που παλιότερα θεωρούνταν ότι απαιτούν νοημοσύνη συχνά αφαιρούνται από τον ορισμό αφού έχουν επιλυθεί. Συνεπώς συχνά ως τεχνητή νοημοσύνη θεωρείται ό,τι δεν έχει πραγματοποιηθεί ακόμα. Η πιο πρόσφατη κατηγοριοποίηση μέχρι σήμερα (2017) περιλαμβάνει στις ικανότητες τεχνητής νοημοσύνης τη δυνατότητα κατανόησης της ανθρώπινης ομιλίας, την ανταγωνιστική συμμετοχή σε παιχνίδια στρατηγικής/αποφάσεων (π.χ. σκάκι), τα αυτόνομα οχήματα, τον έξυπνο διαμοιρασμό περιεχομένου σε δίκτυα, την ερμηνεία και κατανόηση πολύπλοκων δεδομένων, π.χ. εικόνες και βίντεο καθώς και την αισθητηριακή αντίληψη των μηχανών.

Σε σχέση με το τελευταίο, η αντίληψη των μηχανών είναι η ικανότητά τους να ερμηνεύουν δεδομένα με έναν παρόμοιο τρόπο όπως οι άνθρωποι, οι οποίοι χρησιμοποιούν τις αισθήσεις τους για να σχετιστούν με τον κόσμο γύρω τους. Στο παρελθόν, η αντιληπτική ικανότητα των υπολογιστών και η ικανότητα να λαμβάνουν είσοδο από το περιβάλλον τους ήταν πολύ περιορισμένη σε συσκευές όπως το πληκτρολόγιο ή το ποντίκι, αλλά η αναβάθμιση υλικού και λογισμικού, καθώς και των διάφορων αισθητήρων, επέτρεψε την είσοδο αισθητηριακής πληροφορίας με έναν τρόπο παρόμοιο με του ανθρώπου. Η αντίληψη των μηχανών επιτρέπει τη χρήση αυτής της πληροφορίας με σκοπό τη συλλογή πληροφορίας με υψηλότερη ακρίβεια και την παρουσίασή της με τέτοιο τρόπο ώστε να είναι φιλική προς το χρήστη. Μέρος της αντίληψης μηχανών αποτελούν η όραση υπολογιστών, η ικανότητα να ακούν κ.α. Ο τελικός στόχος της είναι να δημιουργηθούν μηχανές οι οποίες θα έχουν την ικανότητα να βλέπουν, να νιώθουν, να ακούν και να αντιλαμβάνονται τον κόσμο όπως οι άνθρωποι.

Παράλληλα με την αντίληψη των μηχανών αναπτύσσεται και η μηχανική μάθηση που συνίσταται στην ικανότητα των μηχανών να μαθαίνουν από τα δεδομένα, χωρίς να χρειάζεται να προγραμματιστούν απευθείας από τον άνθρωπο. Αποτελεί εξέλιξη της αναγνώρισης προτύπων και της θεωρίας υπολογιστικής μάθησης και διερευνά τη μελέτη και κατασκευή αλγορίθμων που μαθαίνουν από- και κάνουν προβλέψεις ή παίρνουν αποφάσεις από- δεδομένα. Οι αλ-

γόριθμοι αυτοί ξεπερνούν τους αυστηρούς κανόνες του προγραμματισμού, διότι είτε κάνουν δεδομενο-κεντρικές προβλέψεις/αποφάσεις, είτε χτίζουν μοντέλα από δείγματα. Η μηχανική μάθηση εφαρμόζεται σε μια πληθώρα εφαρμογών που ολοένα και αυξάνεται με ταχύτατους ρυθμούς, όπου ένας απευθείας σχεδιασμός και προγραμματισμός αλγορίθμου με καλή επίδοση είναι από δύσκολος εως αδύνατος. Τέτοια παραδείγματα εφαρμογών είναι η αναγνώριση φωνής και χειρονομιών, η αυτόματη μετάφραση, κ.α.

Οι εφαρμογές αλληλεπίδρασης ανθρώπου-μηχανής, στις οποίες είναι απαραίτητες τόσο η αντίληψη των μηχανών όσο και η μηχανική μάθηση προκειμένου να επικοινωνήσει ο άνθρωπος με τη μηχανή, έχουν αναπτυχθεί πολύ τα τελευταία χρόνια. Συχνά συνδυάζονται με την ύπαρξη αυτοματισμών ή ρομποτικών συστημάτων με αποτέλεσμα να δημιουργηθούν εφαρμογές όπως η υποβοηθούμενη διαβίωση, τα έξυπνα σπίτια με ασύρματη φωνητική διεπαφή, οι ρομποτικοί βοηθοί, τα κοινωνικά-εκπαιδευτικά ρομπότ, κ.α. Παρόλο που όλα τα παραπάνω έχουν εφαρμοστεί με μια σχετική επιτυχία, σε ρεαλιστικά σενάρια δεν αρκεί μόνο η σωστή αντίληψη και μάθηση αλλά οι μηχανές θα πρέπει να έχουν και κάποιο είδος νοημοσύνης/γνωσιακής λειτουργίας.

Για παράδειγμα, τα καινούρια ρομπότ είναι εφοδιασμένα με μια ποικιλία από αισθητήρες (3D και HD κάμερες, μικρόφωνα, αισθητήρες λέιζερ) που τους επιτρέπει να επεξεργάζονται πολυτροπική πληροφορία από το περιβάλλον τους. Πολλές φορές προορίζονται για χρήση σε διάφορες εργασίες που απαιτούν περίπλοκη επικοινωνία ανθρώπου-ρομπότ, όπως π.χ. σε εφαρμογές βοηθού στο σπίτι ή βοηθού ηλικιωμένων. Παρ΄ όλο που έχει επιτευχθεί μεγάλη βελτίωση στις τεχνολογίες επικοινωνίας ανθρώπου-μηχανής, η συμπεριφορά των συστημάτων σε τέτοια ρεαλιστικά σενάρια απέχει πολύ από το να θεωρηθεί ανθρώπινη. Αυτό οφείλεται στο ότι τα ρομπότ δεν είναι ικανά να φιλτράρουν και να θέσουν στη σωστή προτεραιότητα πολυάριθμα οπτικοακουστικά και άλλα ερεθίσματα από ένα πολύπλοκο περιβάλλον με τον τρόπο που το κάνουν οι άνθρωποι βασισμένοι σε πολυπλοκούς μηχανισμούς προσοχής. Η ενσωμάτωση τέτοιων μηχανισμών θα επιτρέψει στα ρομπότ να ενσωματωθούν ουσιαστικά μέσα στο κοινωνικό περιβάλλον και να αλληλεπιδράσουν με τους ανθρώπους με έναν πιο φυσικό τρόπο.

Μια τέτοια διεπιστημονική προσέγγιση παρουσιάζει πάρα πολλές προκλήσεις και απαιτεί τη συνέργεια πολλών διαφορετικών και ανομοιογενών πεδίων έρευνας: επεξεργασία σήματος, μηχανική μάθηση, υπολογιστική μοντελοποίηση, στατιστική, αλληλεπίδραση ανθρώπου-μηχανής/ρομπότ, ρομποτική, νευροεπιστήμες, γνωσιακές επιστήμες, είναι μόνο μερικά από αυτά. Ο στόχος αυτής της εργασίας είναι να μοντελοποιήσει την ανθρώπινη οπτικοακουστική προσοχή με σκοπό την εφαρμογή της σε περιβάλλοντα αλληλεπίδρασης ανθρώπου-μηχανής/ρομπότ, ταυτόχρονα με τη συνύπαρξη μηχανισμών αντίληψης και εκμάθησης. Συγκεκριμένα, η διατριβή επικεντρώνεται στο ρόλο του ήχου και της φωνής ως μέσο επικοινωνίας και αλληλεπίδρασης με μία μηχανή/ρομπότ, καθώς και στην επίδρασή του στην ανθρώπινη προσοχή, ειδικά όταν συνυπάρχει με οπτικά ερεθίσματα. Μελετάται και αναλύεται η οπτικοακουστική σκηνή με έμφαση στον ήχο και επιχειρείται να γεφυρωθεί το χάσμα που υπάρχει ανάμεσα στην εκτενή μελέτη της οπτικοακουστικής προσοχής από τη σκοπιά των νευροεπιστημών και την υπολογιστική μοντελοποίησή της, που έχει μελετηθεί ελάχιστα. Παράλληλα,

αναπτύσσονται τεχνικές και αλγόριθμοι που σχετίζονται με την επεξεργασία ηχητικών σημάτων και την αντίληψή τους από ένα έξυπνο σύστημα ή ρομπότ.

1.1 Συνεισφορά

Όπως αναφέρθηκε, η διδακτορική διατριβή αυτή επικεντρώνεται στην οπτικοακουστική μοντελοποίηση της ανθρώπινης προσοχής και την πολυ-αισθητηριακή αντίληψη/κατανόησης ακουστικής σκηνής. Πιο συγκεκριμένα, η εργασία αυτή μπορεί να χωριστεί σε τρεις κατηγορίες:

1. Πολυκαναλική επεξεργασία φωνής σε έξυπνα περιβάλλοντα.
2. Φωνητική διεπαφή και επεξεργασία φωνής σε συστήματα αλληλεπίδρασης ανθρώπου-ρομπότ.
3. Υπολογιστική μοντελοποίηση οπτικοακουστικής προσοχής μέσω οπτικοακουστικής εμφάνειας με τεχνικές επεξεργασίας σήματος και με βαθιά νευρωνικά δίκτυα.

Από αυτές ο κύριος ερευνητικός κορμός και το πιο καινοτόμο μέρος της διατριβής αφορά κυρίως στην τελευταία περιοχή. Θα αναφερθεί με συντομία η συνεισφορά της παρούσας διατριβής σε καθεμία από τις τρεις αυτές περιοχές:

1.1.1 Πολυκαναλική επεξεργασία φωνής σε έξυπνα περιβάλλοντα

Μία πολύ σημαντική εφαρμογή της σύγχρονης τεχνολογίας επεξεργασίας και αναγνώρισης φωνής είναι οι φωνητικές διεπαφές για αλληλεπίδραση ανθρώπου-μηχανής σε έξυπνα σπίτια. Σε τέτοια περιβάλλοντα είναι απαραίτητο οι χρήστες να μπορούν να επικοινωνούν με το σύστημα άμεσα, με ασφάλεια αλλά και με άνεση. Η ύπαρξη μιας διεπαφής που 'ακούει' συνέχεια μέχρι ο χρήστης να την χρειαστεί και να την ενεργοποιήσει και που δεν προϋποθέτει την χρήση κάποιου μέσου πέρα από τη φωνή για να επιτευχθεί η επικοινωνία είναι το αντικείμενο αρκετών ερευνητικών ομάδων. Σχετικά πρόσφατα έχουν παρουσιαστεί και εμπορικές ολοκληρωμένες λύσεις από μεγάλες εταιρίες που πρσφέρουν ακριβώς τέτοιου είδους επικοινωνία: Σύστημα που είναι συνεχώς ενεργό, επομένως ασύρματη επικοινωνία, ενεργοποίηση του συστήματος με λέξη-κλειδί, εντοπισμός του χρήστη από το σύστημα και ένα εύρος εντολών που μπορούν να εκτελεστούν από αυτό. Σε τέτοιου είδους εφαρμογές φωνητικής αλληλεπίδρασης μακρινού πεδίου (από απόσταση) χρησιμοποιούνται σχεδόν εξ' ολοκλήρου συστοιχίες μικροφώνων, πολλές φορές κατανεμημένες στο χώρο, επομένως η επεξεργασία των δεδομένων είναι πολυκαναλική.

Στην παρούσα εργασία, έχει μελετηθεί και υλοποιηθεί ένα σύστημα αναγνώρισης φωνής από απόσταση για περιβάλλον έξυπνου σπιτιού, με χρήση πολλών συστοιχιών μικροφώνων, κατανεμημένων στο χώρο. Το σύστημα αυτό αποτελείται από τα εξής υποσυστήματα:

- Ακουστικός εντοπισμός ομιλητή με χρήση πολλών μικροφώνων.
- Πολυκαναλική αποθορυβοποίηση ομιλίας με χρήση δύο διαφορετικών beamformers, τον MVDR (Minimum Variance Distortionless Response) και τον Delay-and-Sum.

- Αναγνώριση φωνητικών εντολών με γραμματική.

Οι συνεισφορές της εργασίας σε αυτό το πεδίο είναι τόσο ερευνητικές, με εστίαση στον εντοπισμό ομιλητή με χρήση πολλών μικροφώνων, όπου έχει αναπτυχθεί μια 2D μέθοδος εντοπισμού ομιλητή, εύρωστη σε περιβάλλοντα με υψηλά επίπεδα θορύβου και αντήχησης, όσο και τεχνικές, όπου αναπτύχθηκε ένα συνολικό σύστημα αναγνώρισης φωνής πραγματικού χρόνου το οποίο είναι σε θέση να αναγνωρίζει φωνητικές εντολές στα Ελληνικά και τα Αγγλικά και στο οποίο ενσωματώθηκαν τα διάφορα υποσυστήματα. Στο σύστημα αυτό, το οποίο ονομάστηκε 'Σπιτάκι μου άκουσε', ενσωματώθηκε επίσης και μια διεπαφή αυτοματισμών για έξυπνα σπίτια, για σκοπούς επίδειξης. Στα πλαίσια της διερεύνησης και ανάπτυξης όλων των παραπάνω, συλλέχθηκε και μία πλούσια βάση πολυκαναλικών δεδομένων ομιλίας στα Ελληνικά, η βάση 'Αθηνά', η οποία αποτελεί μία ακόμη συνεισφορά αυτής της διατριβής. Η βάση αυτή είναι δημόσια διαθέσιμη.

1.1.2 Φωνητική διεπαφή και επεξεργασία φωνής σε συστήματα αλληλεπίδρασης ανθρώπου-ρομπότ

Η ραγδαία ανάπτυξη και βελτίωση των ρομποτικών συστημάτων έχει επιφέρει την αναγκαιότητα ενσωμάτωσης μηχανισμών αντίληψης και αλληλεπίδρασης σε αυτά, ώστε να διευκολύνεται και να γίνεται πιο ποιοτική και φυσική η επικοινωνία τους με τους ανθρώπους. Συνήθως σε ρομποτικές εφαρμογές υπάρχουν πολλοί και διαφορετικοί μεταξύ τους αισθητήρες, διότι η ανάγκη για επικοινωνία μέσω διαφορετικών τροπικοτήτων είναι μεγαλύτερη. Επίσης, τα ανθρωπόμορφα ρομπότ και ιδιαίτερα τα κοινωνικά ρομπότ (δηλαδή αυτά που προορίζονται για σκοπούς διασκέδασης, εκπαίδευσης, κτλ.) επειδή παραπέμπουν και θυμίζουν πολύ περισσότερο τον άνθρωπο σε σχέση με ένα 'αόρατο' σύστημα φωνής σε έξυπνο σπίτι, αυξάνουν και τις απαιτήσεις από την πλευρά του ανθρώπου για πιο διαισθητική επικοινωνία. Επομένως, είναι απαραίτητο και οι δυνατότητές τους να είναι πολυποίκιλες. Συνεπώς, ένα σύστημα αλληλεπίδρασης ανθρώπου-ρομπότ πρέπει να περιλαμβάνει την ύπαρξη συστήματος αναγνώρισης φωνής, διαλόγου, αναγνώρισης χειρονομιών/δράσεων, καθώς και άλλες γνωσιακές λειτουργίες, όπως π.χ. σύστημα προσοχής, ώστε να στρέφει το κεφάλι του προς τον ομιλητή ή προς το τρέχων σημείο ενδιαφέροντος.

Αλληλεπίδραση ρομπότ-παιδιών

Σε αυτή τη διατριβή, το σύστημα που αναφέρθηκε στην προηγούμενη υποενότητα το οποίο εφαρμόστηκε σε έξυπνα σπίτια έχει τροποποιηθεί και προσαρμοστεί κατάλληλα ώστε να καλύπτει απαιτήσεις αναγνώρισης φωνής και αντίληψης σε εφαρμογές ρομποτικής αλληλεπίδρασης, όπου κυρίως παιδιά καλούνται να αλληλεπιδράσουν με ρομπότ μέσω κατάλληλα σχεδιασμένων παιχνιδιών. Το σύστημα αυτό είναι ενσωματωμένο κάτω από ένα γενικότερο σύστημα, το οποίο διαθέτει επίσης αναγνώριση χειρονομιών και δράσεων, διεπαφή διαλόγου και τη δυνατότητα να δίνει εντολές στο ρομπότ καθώς και να λαμβάνει την απαραίτητη πληροφορία από αυτό. Επίσης, μπορεί να πραγματοποιεί οπτικοακουστικό εντοπισμό ομιλητή και

να δίνει εντολή στο ρομπότ να στρέψει το κεφάλι του ώστε να κοιτάει/προσέχει τον ενεργό ομιλητή.

Το σύστημα αλληλεπίδρασης με φωνή που αναπτύχθηκε περιλαμβάνει τα εξής:

- 3D οπτικοακουστικό εντοπισμό ομιλητή ώστε να χρησιμοποιείται η θέση του ομιλητή σαν στοιχείο προσοχής από το ρομπότ.
- Πολυκαναλική αποθορυβοποίηση ομιλίας με χρήση Delay-and-Sum beamforming.
- Αναγνώριση φωνητικών φράσεων με γραμματική.

Οι συνεισφορές σε αυτό το πεδίο είναι οι εξής: Αναπτύχθηκε ένα σύστημα οπτικοακουστικού εντοπισμού ομιλητή στον τρισδιάστατο χώρο με το συνδυασμό ακουστικού εντοπισμού ομιλητή και οπτικού εντοπισμού προσώπου, προσαρμόστηκε κατάλληλα το σύστημα αναγνώρισης φωνής ώστε να αναγνωρίζει επιτυχώς τις φωνές παιδιών αφού προηγήθηκε σχετική διερεύνηση. Σε επίπεδο υλοποίησης, σημαντικό κομμάτι υπήρξε η ενσωμάτωση του συστήματος αυτού στο συνολικό. Επιπλέον, πραγματοποιήθηκε μια μικρή συλλογή φωνητικών δεδομένων από παιδιά.

Αλληλεπίδραση ανθρώπου-ρομποτικού βοηθού

Η αύξηση του μέσου όρου ηλικίας σε συνδυασμό με τη μείωση του προσωπικού των νοσοκομείων, αυξάνει την ανάγκη ύπαρξης ρομποτικών βοηθών που μπορούν να επιβλέπουν τους ασθενείς και να τους βοηθούν στη βάδιση αλλά και στην επικοινωνία. Ένας ρομποτικός περιπατητήρας (robotic rollator) μπορεί να φανεί πολύ χρήσιμος σε ανθρώπους με κινητικά ή/και νοητικά προβλήματα καθώς διαδραματίζει το ρόλο του προσωπικού φροντιστή, επιβλέποντας, αλληλεπιδρώντας και βοηθώντας τον ασθενή. Σε τέτοιους ρομποτικούς βοηθούς απατείται η συνύπαρξη και ταυτόχρονη λειτουργία πολλών αισθητήρων και συστημάτων, οι οποίοι πρέπει να λειτουργούν σε πραγματικό χρόνο, τόσο αυτόνομα όσο και σε συνέργεια. Στην παρούσα διδακτορική διατριβή αναπτύχθηκε ένα σύστημα κατανόησης του λόγου/διαλογικό σύστημα (speech understanding/dialog) το οποίο μέσω αναγνώρισης φωνής και κατανόησης του λόγου είναι σε θέση να αναγνωρίζει τις προθέσεις (intents) του ασθενούς και να δίνει ανάλογη ανατροφοδότηση ή να κάνει την αντίστοιχη ενέργεια.

Το σύστημα αλληλεπίδρασης με φωνή που αναπτύχθηκε περιλαμβάνει τα εξής:

- Αναγνώριση φωνής μέσω ενός state-of-the-art συστήματος.
- Ένα σύστημα κατανόησης του λόγου το οποίο μετατρέπει το αποτέλεσμα της αναγνώρισης φωνής σε μία πρόθεση.
- Ένα διαλογικό σύστημα το οποίο ανάλογα με την πρόθεση που έχει αναγνωριστεί δίνει την κατάλληλη ανατροφοδότηση ή πραγματοποιεί την κατάλληλη ενέργεια.

Στο ρομποτικό βοηθό συνυπάρχουν εκτός από το σύστημα κατανόησης λόγου, ένα σύστημα αναγνώρισης δράσεων και χειρονομιών και ένα σύστημα ανάλυσης βάδισης. Οι συνεισφορές

σε αυτή την ενότητα είναι οι εξής: Αναπτύχθηκε ένα σύστημα κατανόησης του λόγου και ένα σύστημα διαλόγου το οποίο προσαρμόστηκε ώστε να καλύπτει τις ανάγκες της παρούσας εφαρμογής. Πραγματοποιήθηκε μια αρκετά μεγάλη συλλογή δεδομένων από πραγματικούς ασθενείς αλλά και απο υγιείς ανθρώπους, στην οποία βασίστηκε η ανάπτυξη και η αξιολόγηση των συστημάτων που αναπτύχθηκαν.

1.1.3 Υπολογιστική μοντελοποίηση οπτικοακουστικής προσοχής μέσω οπτικοακουστικής εμφάνειας

Όπως αναφέρθηκε και προηγουμένως, η αντίληψη των μηχανών και η μηχανική μάθηση δεν αρκούν ώστε να επιτευχθεί μια ρεαλιστική, ποιοτική επικοινωνία μεταξύ των ανθρώπων και των διάφορων, ρομποτικών και μη, συστημάτων. Οι μηχανές δεν είναι ακόμα ικανές να φιλτράρουν και να θέσουν στη σωστή προτεραιότητα πολυάριθμα οπτικοακουστικά και άλλα ερεθίσματα από ένα πολύπλοκο περιβάλλον με τον τρόπο που το κάνουν οι άνθρωποι βασισμένοι σε πολυπλοκούς μηχανισμούς προσοχής.

Στην ενότητα αυτή εστιάζουμε στην μελέτη και δημιουργία ενός ανοδικού (bottom-up) οπτικοακουστικού μοντέλου προσοχής, αρχικά για μία 2D οπτικοακουστική σκηνή, π.χ. τα βίντεο. Το μοντέλο αυτό βασίζεται στη μοντελοποίηση της προσοχής μέσω της έννοιας της εμφάνειας που έχει μελετηθεί αρκετά, τόσο για την οπτική όσο και για την ακουστική τροπικότητα, αλλά όχι τόσο πολύ για το συνδυασμό τους. Ο στόχος μας είναι η δημιουργία ενός μοντέλου οπτικοακουστικής εμφάνειας το οποίο προβλέπει που στρέφει την προσοχή του/που εστιάζει το βλέμμα του ο άνθρωπος όταν παρακολουθεί μια δισδιάστατη οπτικοακουστική σκηνή. Οι συνεισφορές σε αυτό το πεδίο είναι κυρίως ερευνητικές διότι αποτελεί το πιο καινοτόμο κομμάτι της δουλειάς αυτής. Η ερευνητική δραστηριότητα κινήθηκε σε δύο άξονες, της ανάπτυξης ενός μοντέλου με τεχνικές επεξεργασίας σημάτων και με βαθιά νευρωνικά δίκτυα.

Με τεχνικές επεξεργασίας σημάτων

Όσον αφορά τον άξονα των τεχνικών επεξεργασίας σημάτων, μελετήθηκαν εκτενώς συμπεριφορικά πειράματα και ευρήματα από τις νευροεπιστήμες που αφορούν την οπτικοακουστική προσοχή και το φαινόμενο της οπτικοακουστικής ενσωμάτωσης, ώστε να εξαχθούν παρατηρήσεις και παράμετροι που μπορούν να χρησιμοποιηθούν στη μοντελοποίηση. Στη συνέχεια διερευνήθηκαν διάφοροι τρόποι σύμμειξης της οπτικής με την ακουστική εμφάνεια με βάση τα προηγούμενα ευρήματα. Η αξιολόγηση των προτεινόμενων μοντέλων γίνεται σε τρία επίπεδα: Πρώτα συγκρίνονται τα αποτελέσματα του μοντέλου για ερεθίσματα που προέρχονται από συμπεριφορικά πειράματα με τα ήδη δημοσιευμένα αποτελέσματα από ανθρώπινα συμπεριφορικά πειράματα σε ένα ποιοτικό επίπεδο, ώστε να ελεγχθεί αν η μοντελοποίηση της οπτικοακουστικής εμφάνειας συμφωνεί με τις παρατηρήσεις. Δεύτερον, τα μοντέλα εφαρμόζονται σε βάσεις δεδομένων οφθαλμοκινητικής παρακολούθησης (eye-tracking) με οπτικοακουστικά ερεθίσματα, ώστε να επαληθευτεί με έναν ακόμη τρόπο αν οι έξοδοι που παράγουν βελτιώνουν τα αποτελέσματα εμφάνειας σε σχέση με τη χρήση οπτικών μόνο μοντέλων. Για την επαλήθευση αυτή διερευνώνται και συγκρίνονται διάφορα μοντέλα με διάφορους τρόπους σύμμειξης. Επίσης

για την αξιολόγηση με δεδομένα οφθαλμοκινητικής παρακολούθησης πραγματοποιήθηκε συλλογή δεδομένων για δύο νέες βάσεις από 20 άτομα, το οποίο αποτελεί ακόμα μία συνεισφορά αυτής της εργασίας. Η βάση αυτή είναι δημόσια διαθέσιμη. Τρίτον, πραγματοποιήθηκε επαλήθευση με δεδομένα από fMRI ώστε να διαπιστωθεί με έναν ακόμη τρόπο η δυνατότητα του μοντέλου να μοντελοποιεί την ανθρώπινη οπτικοακουστική προσοχή.

Με βαθιά νευρωνικά δίκτυα

Στη συνέχεια εξετάστηκε η δυνατότητα περαιτέρω βελτίωσης του προηγούμενου μοντέλου το οποίο αναπτύχθηκε με τεχνικές επεξεργασίας σήματος, με τη χρήση βαθιών νευρωνικών δικτύων, τα οποία έχουν επιτύχει πολύ καλά αποτελέσματα σε διάφορα προβλήματα της όρασης υπολογιστών. Συγκεκριμένα, αναπτύχθηκε το νεό χωρο-χρονικό οπτικοακουστικό δίκτυο STAViS (Spatio-Temporal AudioVisual Saliency), το οποίο συνδυάζει τη χωρο-χρονική οπτική με την ακουστική πληροφορία προκειμένου να αντιμετωπίσει το πρόβλημα της εκτίμησης της εμφάνειας σε βίντεο. Από όσο μπορούμε να γνωρίζουμε, το προτεινόμενο σύστημα είναι το πρώτο βαθύ νευρωνικό μοντέλο οπτικοακουστικής εμφάνειας. Η δική μας προσέγγιση χρησιμοποιεί ένα μόνο πολυτροπικό δίκτυο που συνδυάζει οπτική και ακουστική πληροφορία σε διάφορα στάδια και μαθαίνει να πραγματοποιεί κατάλληλη σύμμειξη των δύο τροπικοτήτων ώστε να αποκτήσουμε έναν τελικό χάρτη εμφάνειας. Προτείνονται τρόποι εντοπισμού των ήχων στο βίντεο και κατόπιν η σύμμειξη της χωρο-χρονικής ακουστικής με την χωρο-χρονική οπτική εμφάνεια ώστε να αποκτηθεί ο τελικός χάρτης εμφάνειας.

Η προσέγγιση που υιοθετήθηκε για την οπτική εμφάνεια είναι βασισμένη, αλλά ταυτόχρονα επεκτείνει ένα state-of-the-art χωρο-χρονικό δίκτυο οπτικής εμφάνειας που είναι μέρος ενός μεγαλύτερου δικτύου πολλαπλών εργασιών. Όσον αφορά την ακουστική εμφάνεια, αποκτούμε ακουστικά χαρακτηριστικά από ένα state-of-the-art CNN δίκτυο που έχει δώσει καλές αναπαραστάσεις για το πρόβλημα της ταξινόμησης ακουστικών γεγονότων. Για τον εντοπισμό ακουστικών πηγών στο βίντεο, στόχος μας είναι να βρούμε τις πιθανές συσχετίσεις μεταξύ της ακουστικής και οπτικής ροής πληροφορίας, με την έννοια των κοινών σημασιολογικών εννοιών. Ερευνούμε τρεις διαφορετικούς τρόπους για να εντοπίζουμε τους ήχους στο βίντεο και να αποκτούμε έναν ακουστικό χάρτη εμφάνειας. Κατόπιν διερευνούμε τρόπους σύμμειξης προκειμένου να ενσωματωθούν οι δύο εμφάνειες και να αποκτηθεί ένας τελικός χάρτης εμφάνειας. Και εδώ εξερευνούμε τρεις διαφορετικούς τρόπους.

Το δίκτυο σχεδιάστηκε, εκαπιδεύτηκε end-to-end και αξιολογήθηκε σε 6 διαφορετικές βάσεις με οπτικοακουστικά δεδομένα οφθαλμοκινητικής παρακολούθησης. Οι βάσεις αυτές περιέχουν μεγάλη ποικιλία από διαφορετικά βίντεο, από απλά μικρά βίντεο ως ταινίες του Χόλιγουντ και από βίντεο φτιαγμένα από χρήστες μέχρι ντοκιμαντέρ. Η μέθοδος μας συγκρίνεται με άλλες 8 διαφορετικές state-of-the-art μεθόδους οπτικής εμφάνειας και στις 6 βάσεις. Τα αποτελέσματα για όλες τις βάσεις υποδεικνύουν ότι το προτεινόμενο μοντέλο STAViS ξεπερνά σε απόδοση τόσο την οπτική μόνο παραλλαγή μας αλλά και τις άλλες μεθόδους στην πλειοψηφία των περιπτώσεων. Επίσης, η συνεπής καλή επίδοση που εμφανίζει για όλες τις βάσεις υποδεικνύει την καταλληλότητα της μοντελοποίησης αυτής για την εκτίμηση εμφάνειας σε

βίντεο “in-the-wild”.

1.2 Δομή της διατριβής

Η διατριβή αυτή είναι οργανωμένη ως εξής:

- Στο κεφάλαιο 2 παρουσιάζεται η σχετική βιβλιογραφία που αντιστοιχεί στις τρεις βασικές περιοχές που αναφέρθηκαν στην προηγούμενη ενότητα, χωρισμένη ανά περιοχή.
- Στο κεφάλαιο 3 περιγράφεται το σύστημα επεξεργασίας και αναγνώρισης φωνής με τα διάφορα υπο-συστήματα. Λόγω επικάλυψης των συστημάτων αναγνώρισης φωνής σε έξυπνα σπίτια και σε ρομποτικές εφαρμογές, εφόσον χρησιμοποιούν τους ίδιους βασικούς αλγορίθμους και διαφέρουν κυρίως σε επίπεδο υλοποίησης, παρουσιάζονται στο ίδιο κεφάλαιο. Οι διαφορετικές εφαρμογές τους αναφέρονται ξεχωριστά σε υποκεφάλαια, όπου γίνονται και οι ανάλογες διευκρινίσεις όπου χρειάζεται.
- Στο κεφάλαιο 4 περιγράφεται η βάση πολυκαναλικών δεδομένων φωνής ’Αθηνά’. Αναφέρονται τα περιεχόμενά της καθώς και ο χώρος, οι συνθήκες και η διαδικασία συλλογής της.
- Στο κεφάλαιο 5 περιγράφεται η βάση δεδομένων οφθαλμοκινητικής παρακολούθησης που συλλέχθηκε. Αναφέρεται η διαδικασία καταγραφής και τα περιεχόμενά της.
- Το κεφάλαιο 6 περιέχει μία ιδιαίτερα σημαντική συνεισφορά αυτής της διδακτορικής διατριβής, που είναι η ανάπτυξη του μοντέλου οπτικοακουστικής εμφάνειας με τεχνικές επεξεργασίας σήματος. Περιγράφεται αναλυτικά η διερεύνηση, η μεθοδολογία και οι τρόποι αξιολόγησης των μοντέλων.
- Το κεφάλαιο 7 περιέχει επίσης μία καινοτόμα συνεισφορά αυτής της διδακτορικής διατριβής, που είναι η ανάπτυξη του μοντέλου οπτικοακουστικής εμφάνειας STAViS με βαθιά νευρωνικά δίκτυα. Περιγράφετα επίσης αναλυτικά η διερεύνηση, η μεθοδολογία και οι τρόποι αξιολόγησης των μοντέλων.
- Τέλος, στο κεφάλαιο 8 εξάγονται κάποια συμπεράσματα και δίνονται μελλοντικές κατευθύνσεις έρευνας.

Κεφάλαιο 2

Υπόβαθρο και επισκόπηση ερευνητικού πεδίου

2.1 Πολυκαναλική επεξεργασία φωνής σε περιβάλλοντα έξυπνου σπιτιού

Κατά τη διάρκεια των τελευταίων χρόνων, έχει δαπανηθεί πολλή ερευνητική προσπάθεια για το σχεδιασμό διεπαφών φωνής με σκοπό τη φυσική, ασύρματη επικοινωνία ανθρώπου-υπολογιστή. Τέτοιες διεπαφές συνήθως χρησιμοποιούνται σε συστήματα φωνητικής αλληλεπίδρασης σε τηλεφωνικά κέντρα και πιο πρόσφατα, σε εφαρμογές εξατομικευμένης βοήθειας σε υπολογιστές ή κινητά τηλέφωνα [4]. Οι πρόσφατες βελτιώσεις στην ακουστική μοντελοποίηση για αναγνώριση φωνής [5] έχουν συμβάλει ιδιαίτερα ώστε οι διεπαφές φωνής να γίνουν εύχρηστες σε πολλά καθημερινά περιβάλλοντα. Παρ΄ όλα αυτά, η ασύρματη, συνεχώς ενεργή (always listening) και ταυτόχρονα έυρωστη σε θόρυβο επικοινωνία μακρινού πεδίου, παραμένει ένα δύσκολο πρόβλημα που περιορίζει την καθολική εφαρμογή τους.

Ειδικά σε εφαρμογές υποβοηθούμενης διαβίωσης σε έξυπνα περιβάλλοντα, οι προκλήσεις αυτές είναι ακόμα μεγαλύτερες, διότι οι διεπαφές φωνής είναι καίριας σημασίας για την ασφάλεια και υγεία του ενοίκου [6, 7, 8], ενώ ταυτόχρονα υπάρχουν παρεμβολές από γειτονικά δωμάτια, αντήχηση και διάφορα άλλα ακουστικά γεγονότα και θόρυβοι που επικαλύπτουν την ομιλία η οποία μας ενδιαφέρει. Αντίστοιχες συνθήκες παρουσιάζονται και σε έξυπνα γραφεία. Κάτω από τέτοιες συνθήκες, η επίδοση των συστημάτων αναγνώρισης φωνής από απόσταση χειροτερεύει δραματικά σε σχέση με σενάρια στα οποία έχουμε κοντινά μικρόφωνα ή απουσία θορύβου [9].

Βελτίωση του ποσοστού αναγνώρισης σε εσωτερικούς χώρους μπορεί να επιτευχθεί αν χρησιμοποιηθεί πληροφορία από πολλαπλά κανάλια ήχου από συστοιχίες μικροφώνων [10], κατανεμημένες μέσα στους έξυπνους χώρους. Σκοπός πολλών εργασιών που ασχολούνται με παρόμοιες διεπαφές φωνής είναι η δημιουργία μιας διεπαφής για αυτοματισμούς σπιτιού που υποστηρίζει αλληλεπίδραση φωνής από απόσταση και διεκολύνει ανθρώπους με κινητικά προβλήματα. Η βασική χρήση τέτοιων συστημάτων περιλαμβάνει τον έλεγχο αυτοματισμών όπως

φώτα, θερμοκρασία, πόρτα, κλπ, με φωνητικές εντολές. Συνήθως τα συστήματα αυτά είναι συνεχώς ενεργά και χρησιμοποιούν αναγνώριση φράσης ενεργοποίησης ώστε να διασφαλίζεται μια ευέλικτη και απρόσκοπτη λειτουργία.

Αρχικά, στη βιβλιογραφία μελετήθηκαν χώροι με ένα μόνο δωμάτιο, π.χ. χώροι όπου πραγματοποιούνται συναντήσεις/διαλέξεις, οι οποίοι αν και είναι περισσότερο ελεγχόμενα περιβάλλοντα από σπίτια πολλών δωματίων, οι περισσότερες οπτικοακουστικές σκηνές τους περιέχουν πολλούς ομιλητές, επομένως υπάρχει επικαλυπτόμενη ομιλία σε μεγάλο βαθμό. Ο στόχος σε τέτοιου είδους εφαρμογές είναι η αναγνώριση συνεχούς λόγου μεγάλου λεξιλογίου (large-vocabulary continuous speech recognition) που βασίζεται στη χρήση λεξιλογίων με περίπου 50000 λέξεις. Ένα ενδεικτικό αποτέλεσμα αναγνώρισης από τέτοιες εφαρμογές είναι το 33.2% λάθος ανανγώρισης (Word Error Rate, WER) [11] σε μη επικαλυπτόμενη ομιλία, που αντιστοιχεί στο πλήρες σύστημα μετατροπής λόγου σε κείμενο, το οποίο περιέχει beamforming, προσαρμογή ομιλητή κ.α.. Στα [12] και [13] όπου το beamforming έχει αντικατασταθεί από πολυκαναλική επεξεργασία βασισμένη σε συνελικτικά νευρωνικά δίκτυα για την εκπαίδευση των ακουστικών μοντέλων, έχουν αναφερθεί βελτιώσεις στα αποτελέσματα.

Όσον αφορά χώρους με πολλά δωμάτια και συνεπώς περισσότερο περίπλοκες ακουστικές συνθήκες, ένα ιεραρχικό σύστημα ανάλυσης ήχου παρουσιάζεται στο [14], το οποίο έχει στόχο την αναγνώριση φωνητικών εντολών και τον εντοπισμό επείγουσων καταστάσεων σε χώρους όπου διαμένουν ηλικιωμένοι ή άνθρωποι που χρειάζονται βοήθεια. Σχετικά με το πρόβλημα της αναγνώρισης φωνητικών εντολών, οι συγγραφείς έχουν εστιάσει κυρίως στην διερεύνηση τεχνικών decoding για αυτόματη αναγνώριση φωνής ώστε να συδυάσουν αποτελέσματα από διαφορετικά δωμάτια, χρησιμοποιώντας ένα μικρόφωνο ανά δωμάτιο για εξοικονόμηση. Στη πιο πρόσφατη δουλειά τους [15] χρησιμοποιούν εύρωστη εκπαίδευση και προσαρμογή στο περιβάλλον και επιτυγχάνουν ένα ποσοστό λάθος αναγνώρισης εντολών της τάξης του 13%. Το σκορ αυτό αποκτήθηκε αφότου οι λέξεις που αναγνωρίστηκαν λάθος διορθώθηκαν στο επίπεδο των συλλαβών ώστε να ταιριάζουν με τις λέξεις των προκαθορισμένων εντολών.

Ένα παρόμοιο σύστημα αναπτύχθηκε στο [16] για αναγνώριση φωνητικών εντολών και καταστάσεων έκτακτης ανάγκης. Το ποσοστό του 63.85% σε ακρίβεια αναγνώρισης πρότασης (sentence accuracy) αντιστοιχεί σε αναγνώριση με γραμματική 15 εντολών σε ηχογραφήσεις από ένα μικρόφωνο που βρίσκεται σε ελεγχόμενο δωμάτιο με την ύστερη προσθήκη ραδιοφωνικού θορύβου. Στο [17] το ίδιο πρόβλημα αντιμετωπίζεται με τη χρήση beamforming και με έμφαση στην αναγνώριση φωνής αντί για την αναγνώριση του ενεργού δωματίου. Ο στόχος ήταν ο εντοπισμός και η αναγνώριση έξι εντολών με θόρυβο σε προσομειωμένες ηχογραφήσεις του ενός λεπτού σε περιβάλλον 4 δωματίων. Παρ΄ ότι οι ακουστικές σκηνές ήταν σχετικά απλές, χωρίς επικαλυπτόμενη ομιλία, το επιτευχθέν σκορ του 39% WER είναι ενδεικτικό της δυσκολίας που προσθέτουν παράγοντες όπως θόρυβος, αντήχηση, κ.α.

Προς την κατεύθυνση της εύρωστης τεχνολογίας φωνής και προκειμένου να χτιστεί ένα κοινό πλαίσιο αξιολόγησης των τεχνολογιών που περιλαμβάνει βάσεις δεδομένων, εργασίες και μετρικές αξιολόγησης, εμφανίστηκαν οι διαγωνισμοί REVERB [18] και CHiME [19]. Οι προτεινόμενες μέθοδοι για ευρωστία σε θορυβώδεις συνθήκες μπορούν να χωριστούν σε δύο κατηγορίες: α) εύρωστη μοντελοποίηση και β) πολυκαναλική επεξεργασία [20]. Η εύρωστη

μοντελοποίηση αναφέρεται σε μεθόδους παραμόρφωσης των δεδομένων (data contamination) και σε περιβαλλοντική προσαρμογή. Πιο συγκεκριμένα, όταν δεν υπάρχουν αρκετά δεδομένα εκπαίδευσης, η απόκλιση μεταξύ του περίπλοκου ακουστικού περιβάλλοντος και των γενικών φωνητικών μοντέλων μπορεί να μειωθεί με την τεχνητή παραμόρφωση των δεδομένων εκπαίδευσης [21, 22] ή/και με την προσαρμογή στο περιβάλλον [21, 23]. Η άλλη κατηγορία της πολυκαναλικής επεξεργασίας περιλαμβάνει μεθόδους όπως το beamforming, η επιλογή καναλιού, κ.α.

Η έρευνα, ανάπτυξη και αξιολόγηση των πολυκαναλικών υπο-μονάδων επεξεργασίας φωνής που είναι μέρος μιας αλυσίδας αναγνώρισης ομιλίας εξαρτώνται από την ύπαρξη βάσεων δεδομένων, είτε προσομειωμένων, είτε πραγματικών. Εξαιτίας της πολυπλοκότητας των ρεαλιστικών ακουστικών σκηνών, η συλλογή δεδομένων είναι αρκετά απαιτητική, τόσο από άποψη σχεδιασμού, όσο και εξοπλισμού αλλά και επισημείωσης. Υπάρχει η δυνατότητα δημιουργίας προσομειωμένων δεδομένων μέσω της συνέλιξης προ-ηχογραφημένων καθαρών σημάτων με τις κρουστικές αποκρίσεις των δωματίων και την προσθήκη θορύβου [24]. Παρά την ευκολία δημιουργίας τέτοιων δεδομένων, ο πειραματισμός με πραγματικά δεδομένα είναι απαραίτητος ώστε να αξιολογηθεί το σύστημα σε πραγματικές συνθήκες. Όσον αφορά βάσεις πραγματικών δεδομένων, οι περισσότερες έχουν αποκτηθεί σε ένα δωμάτιο με πολλαπλά μικρόφωνα με σκοπό είτε την ανάλυση συναντήσεων [25, 26, 27], είτε την αναγνώριση ακουστικών γεγονότων [28]. Ένας μικρός αριθμός βάσεων δεδομένων με επίκεντρο τους αυτοματισμούς σπιτιών έχουν γίνει δημόσια διαθέσιμες: Στα [29, 30] έχει ηχογραφηθεί φωνή στα Γαλλικά από πάσχοντες και από μη-πάσχοντες συμμετέχοντες που εκτελούσαν καθημερινές εργασίες ενώ αλληλεπιδρούσαν με μια διεπαφή φωνής μέσω εντολών σε ένα έξυπνο σπίτι πέντε δωματίων.

Συγκεκριμένα για τα Ελληνικά, υπάρχουν πολύ λίγες εργασίες που μελετούν διεπαφές φωνής στα Ελληνικά για περιβάλλοντα έξυπνων σπιτιών. Για παράδειγμα, στο [31] αναφέρονται αποτελέσματα σε αναγνώριση εντολών από απόσταση για τον έλεγχο οικιακών συσκευών χρησιμοποιώντας μονάδες όπως εντοπισμός ομιλητή, beamforming, κ.α. Οι συγγραφείς εστιάζουν περισσότερο σε θέματα υλοποίησης του ενοπισμού ομιλητή και του beamforming ώστε να εντοπίσουν και να βελτιώσουν τη φωνή του χρήστη σε ένα δωμάτιο με αντήχηση όπου ο χρήστης περπατάει και εκφωνεί εντολές κατά τη διάρκεια συνομιλίας του με άλλους ομιλητές στο ίδιο δωμάτιο. Επιτυγχάνεται ένα ποσοστό αναγνώρισης της τάξης του 80% αλλά τα πειράματα περιορίζονται σε τρεις περιπτώσεις συνομιλίας, είναι σχεδιασμένα για 20 εντολές και η γραμμική συστοιχία μικροφώνων που χρησιμοποιείται στρέφεται προς την περιοχή όπου λαμβάνει χώρα η συνομιλία, ενώ στο υπόλοιπο δωμάτιο επικρατεί ησυχία.

Γενικά η αναγνώριση φωνής στα Ελληνικά είναι αρκετά δύσκολη, εξαιτίας της πλούσιας μορφολογίας της γλώσσας και των περιορισμένων διαθέσιμων πόρων για ακουστική και γλωσσική μοντελοποίηση [32]. Ελάχιστες δουλειές έχουν ασχοληθεί με το πρόβλημα της αναγνώρισης συνεχούς λόγου μεγάλου λεξιλογίου στα Ελληνικά. Ενδεικτικά αποτελέσματα αναφέρονται στα [33, 34] για άρθρα εφημερίδων εκφωνημένα από ανθρώπους, με ποσοστό λάθος αναγνώρισης να κυμαίνεται μεταξύ 11.5–21%. Προσεγγίσεις πολυγλωσσικών ακουστικών μοντέλων, όπου συμπεριλαμβάνονται και τα Ελληνικά μεταξύ άλλων γλωσσών, περιγράφονται στο [35].

2.2 Συστήματα αλληλεπίδρασης ανθρώπου-ρομπότ

Η αλληλεπίδραση ανθρώπου με ρομπότ (Human-Robot Interaction, HRI) έχει προκαλέσει μεγάλο ερευνητκό ενδιαφέρον τα τελευταία χρόνια για διάφορους λόγους, μεταξύ των οποίων είναι και η ολοένα αυξανόμενη ένταξη των κοινωνικών ρομπότ στην καθημερινή ζωή των ανθρώπων [36]. Παρ΄ ότι η έρευνα και η τεχνολογία σε αυτές τις κατευθύνσεις έχει εξελιχθεί πολύ και έχει σημειωθεί μεγάλη πρόοδος σε τεχνολογίες όπως η υποβοηθούμενη διαβίωση [37, 38], η διαρκής ανάγκη για πιο διαισθητική και φυσική επικοινωνία επιφέρει συνεχώς νέες προκλήσεις. Οι πιο κοινές μορφές της ανθρώπινης επικοινωνίας περιλαμβάνουν την ανταλλαγή οπτικοακουστικής πληροφορίας, επομένως συστήματα αναγνώρισης φωνής και χειρονομίας είναι αναγκαία συστατικά στοιχεία σε ένα σύστημα ρομποτικής αλληλεπίδρασης [39]. Η αλληλεπίδραση ρομπότ και ανθρώπου θα πρέπει να μιμείται την τυπική μορφή επικοινωνίας μεταξύ ανθρώπων.

Την ίδια στιγμή τα κοινωνικά ρομπότ εισέρχονται όλο και πιο ραγδαία στη ζωή μας για σκοπούς διασκέδασης ή και εκπαίδευσης [40, 41] και ακόμα περισσότερο στη ζωή των παιδιών [42, 43], τα οποία παρουσιάζουν κάποιες ιδιαιτερότητες και προκλήσεις σε σχέση με τους ενήλικες. Τα παιδιά είναι πολύ προσαρμοστικά, μαθαίνουν γρήγορα και εξοικειώνονται εύκολα με τις νέες τεχνολογίες. Έχουν μοναδικούς τρόπους επικοινωνίας, καθώς μπορούν πολύ εύκολα να μεταδίδουν και να λαμβάνουν περίπλοκη πληροφορία, χωρίς να χρησιμοποιούν πολλή ομιλία. Αποτελούν ένα ιδανικό σύνολο για να μελετηθούν οι συνέπειες και τα αποτελέσματα της αλληλεπίδρασης ανθρώπου-ρομπότ στην ανάπτυξη και βελτίωση των επικοινωνιακών δεξιοτήτων [44].

Συνεπώς η αλληλεπίδραση παιδιού-ρομπότ (Child-Robot Interaction, CRI) είναι ένα ανερχόμενο ερευνητικό πεδίο που ταυτόχρονα εγείρει νέες προκλήσεις, όπως για παράδειγμα η αναγνώριση δράσεων και ομιλίας σε παιδιά [45], διότι η πλειοψηφία τέτοιων συστημάτων είναι σχεδιασμένη για ενήλικες. Τα παιδιά συνήθως συμπεριφέρονται διαφορετικά από ότι οι ενήλικες, επομένως μοντέλα που έχουν εκπαιδευτεί σε ενήλικες δε λειτουργούν ικανοποιητικά σε παιδιά. Ένα σύστημα αλληλεπίδρασης ανρθώπου-ρομπότ θα πρέπει να είναι εύρωστο ως προς τη χρήση του από διαφορετικά σύνολα, π.χ. ηλικιακά σύνολα [46]. Από την άλλη πλευρά, για τη βελτίωση τέτοιων συστημάτων που στοχεύουν σε παιδιά, η ύπαρξη και καταγραφή δεδομένων από παιδιά είναι απαραίτητη αλλά ταυτόχρονα και πολύ χρονοβόρα.

Η ανάπτυξη τεχνολογιών όπως η αναγνώριση δράσης, χειρονομιών και φωνής, παράλληλα με τη βελτίωση και κατασκευή νέων αισθητήρων, όπως π.χ. οι κάμερες Kinect, έχει δημιουργήσει μία νέα γενιά πολυτροπικών και πολυ-αισθητηριακών συστημάτων, τα οποία μπορούν να αυξήσουν την ευρωστία και τη συνολική απόδοση [47]. Ολοένα και περισσότερα μονο-τροπικά συστήματα ενσωματώνουν επιπλέον τροπικότητες σε μια προσπάθεια να βελτιωθεί η επίδοση μέσω της σύμμειξης [48, 49].

Σε σχέση με την αντίληψη συστημάτων αλληλεπίδρασης ρομπότ με ανθρώπους, αναφέρθηκαν ήδη δύο βασικά συστατικά, η ικανότητα αναγνώρισης φωνής και χειρονομιών. Άλλο ένα εξίσου βασικό χαρακτηριστικό είναι ο εντοπισμός ομιλητή, που είναι απαραίτητος για την εύρεση του ενεργού ομιλητή σε σενάρια που περιλαμβάνουν πολλούς χρήστες, καθώς και το

στρέψιμο της ρομποτικής προσοχής προς τον άνθρωπο που μιλάει, ώστε η επικοινωνία να γίνεται πιο φυσική [50, 51]. Όλα αυτά τα στοιχεία πρέπει να μπορούν να υποστηρίζουν φυσικά σενάρια αλληλεπίδρασης με πολλαπλούς χρήστες χωρίς να περιορίζουν την κίνησή τους ή να τους υποχρεώνουν να είναι συνδεδεμένοι με το ρομπότ. Επίσης, είναι απαραίτητο η απόδοση του συστήματος να είναι εύρωστη σε σχέση με το θόρυβο από το περιβάλλον και την πολυπλοκότητα του σεναρίου, η οποία μπορεί να υπονοεί απόσταση από τα μικρόφωνα, ακουστική αντήχηση, οπτική επικάλυψη ή αποκλίσεις στην εκτέλεση των χειρονομιών.

Συνήθως οι ρομποτικές πλατφόρμες διαθέτουν ενσωματωμένους οπτικοακουστικούς αισθητήρες, των οποίων όμως οι αντιληπτικές ικανότητες είναι εξαρτημένες από την πλατφόρμα και αρκετά περιορισμένες, παρόλο που σε ορισμένες περιπτώσεις έχουν αντιμετωπίσει επιτυχώς τέτοια προβλήματα όπως τα προαναφερθέντα [52, 53, 54]. Μία εναλλακτική, που υιοθετείται και σε αυτή την εργασία, είναι η χρήση εξωτερικών αισθητήρων που δεν είναι ενσωματωμένοι στο ρομπότ και οι οποίοι είναι πολλαπλοί οπτικοακουστικοί αισθητήρες μακρινού πεδίου, η διάταξη και τοποθέτηση των οποίων δημιουργεί ένα περιβάλλον 'έξυπνου' χώρου. Μία τέτοια επιλογή επιτρέπει τη σύμμειξη διαφόρων ροών πληροφορίας, τόσο εντός της ίδιας τροπικότητας όσο και μεταξύ διαφόρων τροπικοτήτων.

Ρομποτικοί προσωπικοί βοηθοί

Ο σχεδιασμός και η ανάπτυξη ρομποτικών προσωπικών βοηθών για τους ηλικιωμένους απόσχολεί πολύ την ερευνητική κοινότητα τα τελευταία χρόνια [55, 56].

Ο ρόλος τέτοιων συστημάτων είναι πολλαπλός, διότι πρέπει να προσφέρουν τόσο φυσική όσο και αισθητηριακή και νοητική υποβοήθηση [57], αλλά και επίβλεψη της υγείας και της συμπεριφοράς και να δρουν σαν συντροφιά στον ηλικιωμένο [58]. Οι περισσότερες έξυπνες πλατφόρμες υποβοήθησης συνήθως στοχεύουν στο να επιλύσουν συγκεκριμένα προβλήματα. Για παράδειγμα το σύστημα GUIDO [59] και το iWalker [60] προσφέρουν βοήθεια στην πλοήγηση. Μεγάλο μέρος της έρευνας επικεντρώνεται στην ανάλυση ανθρωπομετρικών δεδομένων από διάφορους αισθητήρες για να εκτιμήσουν την ανθρώπινη κατάσταση [61] και έτσι να ελέγξουν μια ρομποτική πλατφόρμα, όπως το CAIROW[62]. Το σύστημα ISR-AIWALKER χρησιμοποιεί RGB-D δεδομένα για την επίβλεψη των χρηστών [63]. Το σύστημα MOBOT [64] διαθέτει πολλαπλούς αισθητήρες και επιχειρεί να μοντλοποιήσει τις ανθρώπινες δραστηριότητες μέσω πολυτροπικών δεδομένων [65, 66] και πραγματοποιεί ανάλυση βάδισης και ευστάθειας [67, 68].

Η ρομποτική όραση σε συστήματα υποβοήθησης είναι ένα σημαντικό θέμα στην αλληλεπίδραση ανθρώπου-ρομπότ [69], ενώ πρόσφατα έχουν αναπτυχθεί πολυ-αισθητηριακά συστήματα τα οποία ενσωματώνουν τόσο οπτική αναγνώριση δράσεων, όσο και αναγνώρισης φωνής για την επικοινωνία με το ρομπότ [70, 71]. Η αναγνώριση ανθρώπινων δράσεων είναι ένα βασικό ερευνητικό πρόβλημα εδώ και αρκετά χρόνια [72], για την επίλυση του οποίου έχουν χρησιμοποιηθεί εκτενώς αναπαραστάσεις του ανθρώπινου σκελετού από εικόνες [73, 74, 75, 76].

Η ταχεία ανάπτυξη μεθόδων και τεχνικών βαθιάς εκμάθησης έχει οδηγήσει στην ανάπτυξη αποτελεσματικών μεθόδων εκτίμησης πόζας [77], ενώ έχει ωφεληθεί και η αναγνώριση δράσεων, χρησιμοποιώντας βελτιωμένες εκτιμήσεις του σκελετού και χρησιμοποιώντας recur-

rent μεθόδους όπως είναι τα δίκτυα LSTM (Long Short Term Memory) τα οποία έχουν την ικανότητα να μοντελοποιούν τη χρονική πληροφορία από ακολουθίες [78, 79, 80, 81, 82, 83].

Παρ΄ όλα αυτά, οι μέθοδοι αυτές έχουν δυσκολίες ενσωμάτωσης σε ρομποτικές πλατφόρμες, διότι οι περισσότερες από αυτές βασίζονται σε σύνολα δεδομένων από περιορισμένα και εξειδικευμένα περιβάλλοντα, με στατικές κάμερες, τα οποία είναι ευαίσθητα στην κίνηση της κάμερας. Επιπλέον, οι περισσότερες υλοποιήσεις που ενσωματώνουν εκτίμηση πόζας και αναγνώριση δράσεων σε ρομποτικές πλατφόρμες [84, 85, 86, 87, 88] συνήθως επικεντρώνονται σε συγκεκριμένες εργασίες αντί να προσεγγίζουν ολιστικά το πρόβλημα, κάτι που απαιτεί πολυτροπική αντίληψη και ανατροφοδότηση πραγματικού χρόνου στο χρήστη από το σύστημα.

2.3 Μοντελοποίηση οπτικοακουστικής προσοχής

2.3.1 Συμπεριφορικά πειράματα

Η προσοχή μπορεί να οριστεί ως η συμπεριφορική και γνωσιακή διαδικασία κατά την οποία ο άνθρωπος μπορεί να συγκεντρώνεται επιλεκτικά σε ένα συγκεκριμένο κομμάτι πληροφορίας, αγνοώντας όλα τα υπόλοιπα ερεθίσματα που αντιλαμβάνεται. Ο ρόλος της προσοχής είναι ζωτικής σημασίας στους ανθρώπους και επομένως αποτελεί ένα πολύ σημαντικό ερευνητικό πεδίο. Παρ΄ ότι συχνά διάφορα οπτικά ή ακουστικά ερεθίσματα τραβούν την ανθρώπινη προσοχή ανεξάρτητα το ένα από το άλλο, τις περισσότερες φορές τα ερεθίσματα είναι πολυτροπικά και πολυ-αισθητηριακά. Αυτό οδηγεί στην πολυτροπική ανθρώπινη προσοχή, π.χ. στην οπτικοακουστική προσοχή. Εκτός από τα καθημερινά παραδείγματα, συστηματικά συμπεριφορικά πειράματα καθώς και πειράματα νευροαπεικόνισης που υπάρχουν στη βιβλιογραφία παρέχουν αρκετές ενδείξεις/αποδείξεις ότι η ανθρώπινη συμπεριφορά εξαρτάται σημαντικά από τους πολυτροπικούς συνδυασμούς των ερεθισμάτων που αντιλαμβάνεται [89, 90, 91, 92]. Η πολυαισθητηριακή αλληλεπίδραση μεταξύ παρατηρήσεων που είναι αταίριαστες μεταξύ τους μπορεί να οδηγήσει σε ψευδαισθήσεις, όπως συμβαίνει π.χ. στο ventriloquist ή το McGurk [93] φαινόμενο, ενώ αντίθετα, ευθυγραμμισμένα πολυ-αισθητηριακά ερεθίσματα που αντιστοιχούν στο ίδιο συμβάν μπορούν να βελτιώσουν πολύ την αντίληψη και την επίδοση σε εργασίες όπως ο προσανατολισμός ή η οπτική αναζήτηση.

Στη βιβλιογραφία υπάρχει μία πληθώρα συμπεριφορικών πειραμάτων που έχουν πραγματοποιηθεί ώστε να διερευνηθούν οι ισχυρές αλληλεπιδράσεις μεταξύ τροπικοτήτων. Μεταξύ αυτών, ένας σημαντικό αριθμός έχει ασχοληθεί με το φαινόμενο της 'ενσωμάτωσης' διαφορετικών τροπικοτήτων (cross-modal integration) ή την αντίληψη του συγχρονισμού και των ορίων του. Ενδιαφερόμαστε ιδιαίτερα για τα ευρήματα αυτά που αφορούν τις οπτικοακουστικές αλληλεπιδράσεις και το πως τέτοια αποτελέσματα συνδέονται με την εμφάνεια (saliency) των παρατηρούμενων γεγονότων [94], δηλαδή το πώς διάφορα γεγονότα τραβούν την ανθρώπινη προσοχή με ανοδικό (bottom-up) τρόπο. Για παράδειγμα, σε μια εργασία οπτικής αναζήτησης όπου οι άνθρωποι πρέπει να εντοπίσουν ένα στόχο μέσα σε μια αλληλουχία από εικόνες που περιέχουν πολλά άλλα ερεθίσματα που μπορούν να αποσπάσουν την προσοχή (distractors) και που αλλάζουν δυναμικά, έχει παρατηρηθεί ότι οι χρόνοι απόκρισης για τον εντοπισμό του

στόχου μπορούν να επιταχυνθούν όταν οι αλλαγές του στόχου είναι συγχρονισμένες με έναν απλό ηχητικό τόνο (auditory pip), διότι ο τελευταίος βοηθάει τον οπτικό στόχο να 'ξεχωρίσει' (pop out), να γίνει δηλαδή περισσότερο προεξέχων. Αυτό το φαινόμενο ονομάζεται "pip and pop" και έχει μελετηθεί και αναλυθεί συστηματικά στο [2]. Παρόμοια φαινόμενα οπτικοακουστικής αλληλεπίδρασης και ενσωμάτωσης (integration) βρίσκονται στο επίκεντρο της γνωσιακής έρευνας σχεδόν εδώ και δύο δεκαετίες, σε μια προσπάθεια να γίνουν κατανοητοί οι μηχανισμοί πίσω από αυτά.

Το φαινόμενο "pip and pop" καθώς και άλλα παρόμοια αποτελέσματα από πειράματα οπτικής αναζήτησης [95, 96, 97] αποτελούν εκδήλωση του φαινομένου που ονομάζεται temporal ventriloquism, το οποίο αφορά την επίδραση του ακουστικού ερεθίσματος στο οπτικό ή τη μετακίνηση της αντίληψης για το χρόνο που ξεκίνησε ένα οπτικό ερέθισμα προς ένα ελαφρά ασύγχρονο ακουστικό ερέθισμα. Αντίστοιχα, στο χωρικό πεδίο το ανάλογο φαινόμενο ονομάζεται spatial ventriloquism. Μία σύνοψη των δύο αυτών φαινομένων και παραδειγμάτων τους παρουσιάστηκε στο [98], όπου μελετήθηκαν τα αποτελέσματα τους και τα χωρο-χρονικά κριτήρια για να εκδηλωθούν. Βρέθηκε ότι το φαινόμενο του temporal ventriloquism επηρεάζεται από τα χρονικά παράθυρα (η ασυγχρονία μεταξύ οπτικής και ακουστικής πληροφορίας δε μπορεί να ξεπερνάει τα $200ms$) αλλά δεν επηρεάζεται καθόλου από χωρική ασυμφωνία. Οι συγγραφείς ανέφεραν επίσης κάποια θεωρητικά μοντέλα που περιγράφουν τα δύο αυτά φαινόμενα.

Μία άλλη ενδιαφέρουσα προσέγγιση είναι ότι η 'ενσωμάτωση' των ερεθισμάτων ακολουθεί νόμους του Bayes [99]. Η βασική ιδέα είναι ότι τα οπτικά και ακουστικά χαρακτηριστικά συνδυάζονται με βέλτιστο τρόπο, με κατάλληλα βάρη. Κάποιες μελέτες επίσης έχουν μοντελοποιήσει τις οπτικοακουστικές αλληλεπιδράσεις μέσω Bayesian inference όπου η πιθανοφάνεια αντιπροσωπεύει τον αισθητηριακό θόρυβο στο περιβάλλον ή τον εγκέφαλο, ενώ η prior εκφράζει στατιστικά των γεγονότων του περιβάλλοντος [100].

Στο [101] μελετήθηκε το βέλτιστο χρονικό παράθυρο οπτικοακουστικής ενσωμάτωσης σε σχέση με το χρόνο απόκρισης. Οι συγγραφείς βασίζονται στη bayesian μοντελοποίηση της ενσωμάτωσης και επεκτείνουν την προηγούμενη δουλειά που αναφέρθηκε [99]. Αποδεικνύουν ότι το χρονικό παράθυρο λειτουργεί σα φίλτρο το οποίο αποφασίζει αν η πληροφορία που παράγεται από τους διαφορετικούς αισθητήρες είναι αρκετά κοντά στο χρόνο ώστε να προκαλέσει πολυ-αισθητηριακή ενσωμάτωση. Στο [102] πραγματοποιήθηκαν συμπεριφορικά πειράματα ώστε να μετρηθεί το χρονικό παράθυρο της ενσωμάτωσης στην αντίληψη του οπτικοακουστικού λόγου. Τα πειράματα αυτά υποδεικνύουν παράθυρο ενσωμάτωσης των $200ms$.

Οι συγγραφείς του [103] εξετάζουν την ανθρώπινη κρίση σχετικά με τον οπτικοακουστικό συγχρονισμό: Αναφέρεται ότι όταν οι άνθρωποι εκτίθενται σε μια σταθερή χρονική υστέρηση/ασυγχρονία μεταξύ οπτικού και ακουστικού ερεθίσματος, πειράματα δείχνουν ότι μετακινούν την υποκειμενική τους αντίληψη περί συγχρονισμού προς την χρονική αυτή υστέρηση. Η βασική ιδέα είναι ότι ο εγκέφαλος προσπαθεί να προσαρμόσει την υποκειμενική αίσθηση του συγχρονισμού μεταξύ διαφορετικών τροπικοτήτων, εντοπίζοντας και μειώνοντας τις χρονικές υστερήσεις από ερεθίσματα που μάλλον προέρχονται από το ίδιο φυσικό γεγονός. Ένα ενδιαφέρον εύρημα που σχετίζεται με τα παραπάνω είναι η ικανότητα του εγκεφάλου να

ανα-βαθμονομείται πολύ γρήγορα στην οπτικοακουστική ασυγχρονία, ακόμα και όταν εκτίθεται σε μία μοναδική σύντομη ασυγχρονία [104, 105]. Οι συγγραφείς ισχυρίζονται ότι πρόκειται για ένα αισθητηριακό φαινόμενο που δρα πολύ γρήγορα, και όχι για μια γνωσιακή διαδικασία υψηλού επιπέδου. Επίσης αποδεικνύουν ότι αυτή η χρονική ανα-βαθμονόμηση είναι ασύμμετρη ως προς τη σειρά εμφάνισης των ερεθισμάτων, διότι εμφανίζεται σε πολύ μεγαλύτερο χρονικό εύρος όταν το οπτικό ερέθισμα προηγείται σε σχέση με όταν το ακουστικό ερέθισμα προηγείται. Οι υποθέσεις αυτές επιβεβαιώνονται από μια σειρά πειραμάτων και παρουσιάζεται και ένα μοντέλο που εξηγεί τα αποτελέσματα.

Στο [106] οι συγγραφείς ισχυρίζονται ότι η χρονική διακριτική ικανότητα της αντίληψης συγχρονισμού είναι πολύ υψηλότερη (10Hz) για ένα ακουστικο-απτικό (audio-tactile) ερέθισμα σε σχέση με ένα οπτικοακουστικό ή οπτικο-απτικό (visuo-tactile) (4Hz). Άλλες σχετικές έρευνες [107] προσπαθούν να διακρίνουν το πώς επηρεάζεται χρονικά η αντίληψη ενός οπτικοακουστικού ερεθίσματος όταν υπάρχει ένα πρόσθετο ταυτόχρονο οπτικό ή ακουστικό ερέθισμα.

.

2.3.2 Οπτικά μοντέλα εμφάνειας (saliency)

Παράλληλα με τα συμπεριφορικά πειράματα, έχει πραγματοποιηθεί εκτεταμένη έρευνα που εστιάζει στους μηχανισμούς οπτικής και ακουστικής εμφάνειας ξεχωριστά. Πολλά ευρήματα σε αυτές τις περιοχές έχουν ήδη ενσωματωθεί σε υπολογιστικά μοντέλα που βοηθούν ακόμα περισσότερο στην κατανόηση τέτοιων εννοιών και την εφαρμογή τους σε πραγματικές εφαρμογές. Οι σημαντικές εργασίες [108, 109] έθεσαν τις βάσεις για τη δημιουργία και ανάπτυξη ενός μοντέλου οπτικής εμφάνειας που προβλέπει τα σημεία εστίασης (fixations) του ματιού κατά τη διάρκεια ελεύθερης θέασης εικόνων, βασισμένο σε χαρακτηριστικά των εικόνων. Αυτά τα μοντέλα που επιβεβαιώθηκαν, αλλά και προέκυψαν, από συμπεριφορικές παρατηρήσεις σε ψυχοφυσικά πειράματα, ενέπνευσαν περαιτέρω παραλλαγές και βελτιώσεις την τελευταία δεκαετία, οι οποίες εφαρμόστηκαν σε αναγνώριση αντικειμένων σε εικόνες καθώς και σε εντοπισμό σημαντικού περιεχομένου στην ομιλία.

Στο [110] παρουσιάζεται μια ανασκόπηση του state-of-the-art της οπτικής μοντελοποίησης προσοχής, όπου αναφέρονται περίπου 65 διαφορετικά μοντέλα και σχετικές συγκρίσεις μεταξύ τους. Αρχικά τα υπάρχοντα μοντέλα ταξινομούνται με βάση 13 παράγοντες: ανοδικά/καθοδικά (bottom-up/top-down), χωρικά/χωρο-χρονικά, overt/covert προσοχή, space-based/object-based, τα χηρσιμοποιούμενα χαρακτηριστικά, τα ερεθίσματα και το είδος της εργασίας, τις μετρικές και τις βάσεις που χρησιμοποιούνται για αξιολόγηση. Σε αυτή την ενότητα θα εστιάσουμε στα χωρο-χρονικά μοντέλα, αναφέροντας περιληπτικά τους διάφορους τύπους: τα εμπνευσμένα από τη βιολογία, αυτά που βασίζονται σε θεωρία πληροφορίας και τα επιλεκτικά ως προς συχνότητα/φάση. Αρχικά αναφέρονται με συντομία τα χωρικά και ύστερα με περισσότερες λεπτομέρειες τα χωρο-χρονικά.

Δύο πολύ σημαντικές εργασίες [111], [108] αποτέλεσαν τη βάση πολλών μοντέλων προσοχής εμπνευσμένων από τη βιολογία [112, 113, 114, 115], ενώ ο Itti et al. [109] υλοποίησε

ένα ανοδικό (bottom-up) υπολογιστικό μοντέλο χωρικής οπτικής εμφάνειας χρησιμοποιώντας τρία κανάλια χαρακτηριστικών: ένταση, χρώμα, προσανατολισμό. Το μοντέλο αυτό στη συνέχεια επεκτάθηκε σε χωρο-χρονικό μοντέλο γαι πρόβλεψη οπτικής εμφάνειας σε βίντεο με την προσθήκη δύο επιπλέον χρακτηριστικών: κίνηση και τρεμόπαιγμα (flicker) [116]. Υπάρχουν πολλά άλλα βιολογικά εμπνευσμένα μοντέλα που βασίζονται στο μοντέλο Itti et al., είτε χωρικά [117, 118, 119, 120], είτε χωρο-χρονικά [121, 122, 123, 124].

Όσον αφορά τα μοντέλα που χρησιμοποιούν θεωρία πληροφορίας, τα περισσότερα βασίζονται σε ένα Bayesian πλαίσιο [125, 126, 127, 128]. Οι Zhang et al. [129] πρότειναν ένα γενικό πλαίσιο για μοντελοποίηση εμφάνειας χρησιμοποιώντας ανάλυσης στατιστικής σε φυσικά δεδομένα (Saliency Using Natural (SUN)), ενώ αργότερα επέκτειναν και αυτοί το μοντέλο συμπεριλαμβάνοντας χωρο-χρονικά χαρακτηριστικά [130]. Άλλες εργασίες εκμεταλλεύονται μετρικές από τη θεωρία πληροφορίας, όπως η εντροπία (entropy), η αμοιβαία πληροφορία (mutual information) για χωρικά [131, 132, 133, 134, 135, 136, 137] ή χωρο-χρονικά [138, 133, 139, 136] μοντέλα.

Μία άλλη κατηγορία προσεγγίσεων εκτιμά την εμφάνεια στο πεδίο της συχνότητας με ρύθμιση του χάρτη εμφάνειας ώστε να είναι επιλεκτικός ως προς τη συχνότητα ή τη φάση [140, 141, 142]. Κάποια μοντέλα βασίζονται στο μετασχηματισμό Fourier ή Διακριτού Συνημιτόνου [140, 143], ενώ έχει χρησιμοποιηθεί και ο quaternion μετασχηματισμός Fourier για το συνδυασμό χαρακτηριστικών χρώματος, έντασης και κίνησης [144, 145, 142].

Όσον αφορά τα χωρο-χρονικά μοντέλα, συγκριτικά με τα χωρικά έχει γίνει λιγότερη έρευνα, και μάλιστα στις περισσότερες περιπτώσεις τα χωρο-χρονικά μοντέλα είναι απλά μία επέκταση των χωρικών ενσωματώνοντας δυναμικά χαρακτηριστικά. Στα [116, 126, 120], οι διαφορές μεταξύ των χαρτών χωρικού προσανατολισμού χρησιμοποιούνται ως χρονικά χαρακτηριστικά για ανίχνευση εμφάνειας σε βίντεο. Στο [136] οι συγγραφείς επεκτείνουν τη μέθοδο τους που βασίζεται στην αυτο-ομοιότητα, χρησιμοποιώντας $3D$ τοπικούς πυρήνες για ανίχνευση δράσεων και εμφάνειας σε βίντεο. Στο [146] προτείνεται ένα χωρο-χρονικό φιλτράρισμα χρησιμοποιώντας χρονικά αθροίσματα με βάρη για την ανίχνευση περίεργων κινήσεων σε πολύ γεμάτες σκηνές, ενώ στο [147] συνδυάζεται η πληροφορία κίνησης της κάμερας με στατικά χαρακτηριστικά ώστε να μελετηθούν οι διαφορές στατικής και δυναμικής εμφάνειας σε βίντεο. Στο [148] παρουσιάζεται ένα χωρο-χρονικό υπολογιστικό μοντέλο οπτικής εμφάνειας που βασίζεται στην αντίληψη, που παράγει τόσο χωρο-χρονική όσο και στατική ενέργεια χρησιμοποιώντας την ίδια πολυ-κλιμακωτή συστοιχία φίλτρων που βασίζεται σε τετραγωνικά Gabor φίλτρα τριών διαστάσεων. Στο [149] έχει προταθεί ένα ανοδικό μοντέλο εμφάνειας που βασίζεται στη δομή του ανθρώπινου συστήματος όρασης.

Τα τελευταία χρόνια, ένα μεγάλο ποσοστό εργασιών προσεγγίζει την οπτική εμφάνεια με βαθιά νευρωνικά δίκτυα. Στο [150] χαρακτηριστικά από διαφορετικά επίπεδα (layers) του δικτύου χρησιμοποιούνται για να εκπαιδευτούν Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines (SVMs)) για τις περιοχές που υπήρχαν σημεία εστίασης του ματιού (fixations) και για αυτές που δεν υπήρχαν. Άλλες προσεγγίσεις χρησιμοποιούν προσαρμογή προ-επαιδευμένων συνελικτικών νευρωνικών δικτύων (CNN) από εργασίες οπτικής αναγνώρισης [151], ενώ στο [152] εκπαιδεύονται τόσο ρηχά όσο και βαθιά συνελικτικά νευρωνικά

δίκτυα για πρόβλεψη εμφάνειας. Στο [153], πολυ-κλιμακωτά συνελικτικά νευρωνικά δίκτυα εκπαιδεύονται βελτιστοποιώντας κάποιες καθιερωμένες μετρικές εμφάνειας, ενώ στο [154] οι συγγραφείς εξάγουν περιοχές της εικόνας με σημεία εστίασης του ματιού και μη, για να εκπαιδευτεί ένα δυαδικό συνελικτικό δίκτυο πολλαπλής ανάλυσης (binary multiresolution CNN). Η εργασία [155] ανέδειξε ότι τα κριτήρια σφάλματος που βασίζονται σε μετρικές πιθανοτικής απόστασης είναι καταλληλότερες για προβλήματα που σχετίζονται με εμφάνεια σε σχέση με τις πρότυπες συναρτήσεις σφάλματος παλινδρόμησης. Στο [156] χρησιμοποιούνται γεννησιακά ανταγωνιστικά δίκτυα (generative adversarial networks (GAN)) ώστε να εκπαιδευτούν καλύτερα δίκτυα για πρόβλεψη σημείων εστίασης. Στο [157] οι συγγραφείς προτείνουν ένα συνελικτικό νευρωνικό δίκτυων δύο ροών που βασίζεται σε έγχρωμες εικόνες και χάρτες οπτικής ροής για πρόβλεψη δυναμικής εμφάνειας. Οι συγγραφείς στο [158] χρησιμοποίησαν πολυ-επίπεδη πληροφορία εμφάνειας από διαφορετικά layers μέσω skip conncections. Επίσης, έχουν χρησιμοποιηθεί και δίκτυα Long Short-Time Memory (LSTM) για την εκτίμηση της εμφάνειας τόσο σε στατικές εικόνες [159] όσο και σε βίντεο [160, 161, 162]. Προκειμένου να βελτιωθεί η εκτίμηση της εμφάνειας σε βίντεο, πολλές προσεγγίσεις χρησιμοποιούν δίκτυα πολλαπλών ροών (multi-stream) όπως τα RGB/Optical Flow (OF) [163], RGB/OF/Depth [164], ή πολλαπλά υποδίκτυα όπως το objectness/motion [165] ή το saliency/gaze [166]. Μια πιο πρόσφατη προσέγγιση [167] χρησιμοποιεί ενα τρισδιάστατο fully-convolutional δίκτυο με χρονικό aggregation, βασισμένο στην υπόθεση ότι ο χάρτης εμφάνειας για κάθε καρέ μπορεί να προβλεφθεί λαμβάνοντας υπόψη έναν περιορισμένο αριθμό παρελθοντικών καρέ.

2.3.3 Ακουστικά μοντέλα εμφάνειας

Το πρώτο βιολογικά εμπνευσμένο ακουστικό μοντέλο εμφάνειας προτάθηκε από τον Kayser et al. [168] και είναι δομικά παρόμοιο με το οπτικό μοντέλο Itti et al. [109, 116]. Το ακουστικό ερέθισμα μετατρέπεται σε μία χρονο-συχνοτική αναπαράσταση η οποία μπορεί να απεικονιστεί σαν 'εικόνα έντασης', η οποία ουσιαστικά είναι το ηχητικό σπεκτρόγραμμα, και η οποία είναι η είσοδος στο μοντέλο. Η έξοδος είναι ένας χάρτης εμφάνειας, ο οποίος απεικονίζει την εξέλιξη της ακουστικής εμφάνειας ως προς το χρόνο και τη συχνότητα. Είναι δηλαδή δομικά παρόμοιο με το οπτικό μοντέλο αλλά έχει μια διαφορετική ερμηνεία εφόσον ενσωματώνει την έννοια του χρόνου. Τα εξαγόμενα χαρακτηριστικά είναι η ένταση, η χρονική αντίθεση και η συχνοτική αντίθεση σε διάφορες κλίμακες (εμπνευσμένο από τη λειτουργία των ακουστικών νευρώνων). Κάθε χαρακτηριστικό εξάγεται με φίλτρα που μοντελοποιούν ευρήματα από την ακουστική φυσιολογία: το φίλτρο έντασης αντιστοιχεί στους υποδοχείς που διαθέτουν μια φάση διέγερσης, τα φίλτρα συχνοτικής αντίθεσης στους υποδοχείς με φάση διέγερσης και ταυτόχρονα μια παράπλευρη ζώνη εξασθένησης (side band inhibition) ενώ τα χρονικής αντίθεσης σε υποδοχείς με μία φάση διέγερσης και κατόπιν μία φάση εξασθένησης. Τα φίλτρα αυτά αντιστοιχούν σε φίλτρα Gabor με κατάλληλους προσανατολισμούς. Η έξοδος του μοντέλου είναι ένας δισδιάστατος χάρτης που παράγεται αθροίζοντας τους επιμέρους χάρτες χαρακτηριστικών.

Στο [169], περιγράφεται ένα υπολογιστικό μοντέλο ακουστικής εμφάνειας που διερευνά το

χώρο της ακουστικής εμφάνειας με βάση την πρωτεύουσα συχνότητα (pitch), την ένταση και το ηχόχρωμα. Βασίζεται στην υπόθεση ότι η αντίληψη ακολουθεί την εξέλιξη των ακουστικών γεγονότων σε ένα πολυδιάστατο χώρο χαρακτηριστικών και σηματοδοτεί κάθε απόκλιση από τα στατιστικά ως προεξέχον (salient) γεγονός. Επιπλέον, εξετάζεται ο ρόλος της προβλεπτικής κωδικοποίησης (predictive coding) ως μηχανισμός εντοπισμού εμφάνειας. Η προβλεπτική κωδικοποίηση αντιστοιχεί στην ελαχιστοποίηση του λάθους μεταξύ των ανοδικών αισθητήριων και των καθοδικών (top-down) προβλέψεων. Οι προκύπτουσες αναντιστοιχίες σηματοδοτούν τον εντοπισμό αποκλίσεων, δηλαδή προεξεχόντων γεγονότων. Η έξοδος του μοντέλου είναι ένα μονοδιάστατο σήμα που υποδεικνύει την πιθανότητα προεξέχοντων γεγονότων στο χρόνο. Το μοντέλο έχει επαληθευτεί μέσω συμπεριφορικών πειραμάτων.

Στο [170] οι συγγραφείς προτείνουν ένα βιολογικά εμπνευσμένο μοντέλο ακουστικής εμφάνειας βασισμένο στο [116, 168]. Εκτός από τα χαρακτηριστικά που υπολογίζονται στο [168], υπολογίζουν επίσης χαρακτηριστικά προσανατολισμού και πρωτεύουσας συχνότητας στον δισδιάστατο χάρτη εμφάνειας. Στη συνέχεια ενσωματώνουν τα διάφορα χαρακτηριστικά σε έναν ενιαίο χάρτη εμφάνειας χρησιμοποιώντας έναν μη γραμμικό αλγόριθμο τοπικής κανονικοποίησης, εμπνευσμένο από τη βιολογία και προσαρμοσμένο από το οπτικό μοντέλο [171]. Για να μπορέσουν να χρησιμοποιήσουν τον ακουστικό χάρτη, αθροίζουν τις τιμές τους στα συχνοτικά κανάλια για κάθε χρονικό σημείο, υποθέτοντας ότι η εμφάνεια προκύπτει αθροιστικά από τα κανάλια αυτά. Επαληθεύουν το μοντέλο μέσω μιας εργασίας ανίχνευσης σημαντικών συλλαβών στην ομιλία.

Στο [172] εφαρμόζεται Bayesian έκπληξη ώστε να εντοπιστούν τα προεξέχοντα ακουστικά γεγονότα. Χρησιμοποιείται η Kullback-Leibler απόκλιση μεταξύ των ύστερων (posterior) και πρότερων (prior) κατανομών ως μετρική του πόσο 'αναπάντεχο' είναι ένα καινούριο ακουστικό δείγμα. Με αυτό τον τρόπο ανιχνεύονται αποδοτικά τα αναπάντεχα και καινούρια ακουστικά γεγονότα.

2.3.4 Οπτικοακουστικά μοντέλα εμφάνειας

Στη βιβλιογραφία έχουν παρουσιαστεί διάφορες απόπειρες να μοντελοποιηθεί η οπτικοακουστική προσοχή. Οι περισσότερες από αυτές, όμως, είτε εφαρμόζονται μόνο σε εξειδικευμένες συνθήκες, είτε προσπαθούν να συγχωνεύσουν την οπτική πληροφορία με τη χωρική ακουστική πληροφορία, π.χ. σε ρομποτικές εφαρμογές.

Η πρώτη φορά που εμφανίζεται ένα υπολογιστικό οπτικοακουστικό μοντέλο εμφάνειας, το οποίο προσπαθεί να προβλέψει τι τραβάει την προσοχή σε μια οπτικοακουστική σκηνή είναι στο [173]. Το μοντέλο αυτό αναπτύχθηκε προκειμένου να καθοδηγεί ένα ανθρωποειδές ρομπότ. Η εκτίμηση της οπτικής εμφάνειας βασίζεται στο μοντέλο Itti et al. [109], ενώ όσον αφορά την ακουστική εμφάνεια, μόνο οι χωρικές ιδιότητες της πηγής λαμβάνονται υπόψη, δηλαδή το που βρίσκεται η πηγή, όπως την υπολογίζει ένας αλγόριθμος εντοπισμού θέσης. Ο χάρτης εμφάνειας της εξόδου (υπολογίζεται με μία πράξη μεγίστου) στην ουσία καθοδηγείται από τον οπτικό χάρτη εκτός αν εμφανιστεί στη σκηνή κάποιο ακουστικό γεγονός.

Σε μια παρόμοια εφαρμογή το μοντέλο που έχει προταθεί στο [174] βασίζεται στην Bayesian

έκπληξη και σε αλγόριθμο εντοπισμού θέσης ακουστικού γεγονότος για τη δημιουργία του χάρτη ακουστικής εμφάνειας ενώ για τον οπτικό χάρτη εφαρμόζεται μια προσέγγιση βασισμένη στη φάση. Τα προεξέχοντα ακουστικά και οπτικά γεγονότα ομαδοποιούνται ετερο-τροπικά (cross-modally) και η τελική εμφάνεια αυτών των οπτικοακουστικών συστάδων υπολογίζεται συνδυάζοντας γραμμικά τις ξεχωριστές τιμές εμφάνειας.

Στο [175] ο ακουστικός χάρτης υπολογίζεται επίσης με εντοπισμό θέσης πηγής και στη συνέχεια συγχωνεύεται με τον οπτικό μέσω πολλαπλασιασμού. Στο [176] παρουσιάζεται ένα τρισδιάστατο οπτικοακουστικό μοντέλο προσοχής το οποίο συνδυάζει οπτικά δεδομένα, δεδομένα βάθους και πολυκαναλικά ακουστικά δεδομένα μέσω δύο ανεξάρτητων στατικών και δυναμικών δρόμων ανάλυσης, ενώ στο [177] περιγράφεται ένα οπτικοακουστικό μοντέλο που προσπαθεί να προβλέψει το ανθρώπινο βλέμμα (gaze) σε εφαρμογές τηλεδιάσκεψης. Το τελευταίο βασίζεται στη συγχώνευση χωρικών, χρονικών και ακουστικών χαρτών προσοχής με τους τελευταίους να προκύπτουν από οπτικοακουστικό εντοπισμό ομιλητή πραγματικού χρόνου.

Από μία διαφορετική οπτική γωνία, το οπτικοακουστικό μοντέλο που εισάγεται στο [178, 179] για περίληψη ταινιών και που βελτιώθηκε περαιτέρω στο [180] στοχεύει στην πρόβλεψη του πότε και όχι του που, ελκύεται η προσοχή σε μια δυναμική σκηνή.

Όλα τα παραπάνω μοντέλα έχουν δημιουργηθεί ώστε να εξυπηρετήσουν συγκεκριμένες εφαρμογές και παρόλο που μπορεί να έχουν εμπνευστεί από γνωσιακές επιστήμες δεν έχει γίνει κάποια προσπάθεια ώστε να επαληθευτούν μέσω συμπεριφορικών ευρημάτων. Οι Coutrot και Guyader [181, 182, 183] καθώς και η Song [184] έχουν επιχειρήσει πιο άμεσα να επαληθεύσουν τα μοντέλα τους με ανθρώπους. Τα ευρήματα που χρησιμοποιούν υποδεικνύουν ότι το ανθρώπινο βλέμμα σε ταινίες ελκύεται από τα ομιλούντα πρόσωπα και τους μουσικούς οργανοπαίχτες. Για να μπορέσουν να μοντελοποιήσουν αυτά τα ευρήματα, αφού υπολογίσουν το χάρτη οπτικής εμφάνειας, σταθμίζουν κατάλληλα τις περιοχές που περιέχουν πρόσωπα ώστε να δημιουργηθεί ένας οπτικοακουστικός χάρτης που να εξηγεί τα δεδομένα οφθαλμοκινητικής παρακολούθησης (eye-tracking) που έχουν καταγραφεί κατά τη διάρκεια θέασης ταινιών.

Τέλος οι Min et al. [185, 186] ανέπτυξαν ένα οπτικοακουστικό μοντέλο προσοχής για την πρόβλεψη των σημείων εστίασης του ανθρώπινου ματιού σε σκηνές που περιέχουν κινούμενα αντικείμενα που παράγουν ήχο. Για την ακουστική προσοχή, χρησιμοποιούν μία μέθοδο εντοπισμού ήχου ώστε να εντοπίζουν χωρικά το κινούμενο αντικείμενο πάνω στην εικόνα και μετα πραγματοποιούν σύμμειξη με την οπτική εμφάνεια.

2.3.5 Άλλες οπτικοακουστικές προσεγγίσεις με βαθιά νευρωνικά δίκτυα

Κατά τη διάρκεια των τελευταίων τριών χρόνων, η ενσωμάτωση της ακουστικής πληροφορίας σε μοντέλα και προβλήματα που ήταν παραδοσιακά οπτικά μόνο, όπως ο εντοπισμός αντικειμένου ή η κατάτμηση, προσελκύει ολοένα και περισσότερο την προσοχή. Στο [187, 188], η οπτικοακουστική αντιστοιχία σε βίντεο μαθαίνεται χωρίς επίβλεψη και έχει ως αποτέλεσμα καλές οπτικές και ακουστικές αναπαραστάσεις, καθώς και υψηλό ποσοστό σωστής ταξινόμησης

και στις δύο τροπικότητες. Στο [189, 190] προτείνονται κάποιες οπτικοακουστικές αναπαραστάσεις που μαθαίνονται με έναν αυτο-επιβλεπόμενο (self-supervised) τρόπο και χρησιμοποιούνται για την ανάλυση οπτικοακουστικής σκηνής, ενώ στο [191], οι συγγραφείς προσπαθούν να αντιμετωπίσουν το πρόβλημα του εντοπισμού οπτικοακουστικού γεγονότος σε βίντεο χωρίς περιορισμούς. Επίσης, παρουσιάζεται στο [192, 193] ένα σύστημα το οποίο μαθαίνει να εντοπίζει περιοχές στην εικόνα οι οποίες παράγουν ήχους και κατόπιν δημιουργεί μια αναπαράσταση του ήχου που προέρχεται από κάθε εικονοστοιχείο της εικόνας (pixel). Το μοντέλο SoundNet [194] εκμεταλλεύεται τη φυσική συγχρονία που υπάρχει μεταξύ της οπτικής και της ακουστικής πληροφορίας στα βίντεο, ώστε να μάθει μια ακουστική αναπαράσταση χρησιμοποιώντας 2000000 μη επισημειωμένα βίντεο. Η αξιολόγηση του μοντέλου επιδεικνύει βελτιωμένη επίδοση σε προβλήματα ταξινόμησης ακουστικής σκηνής και αντικειμένων. Τέλος, στο [195] οι συγγραφείς αξιοποιούν την ακουστική και οπτική πληροφορία προκειμένου να χτίσουν αποτελεσματικά μοντέλα τόσο για οπτική όσο και για ακουστική ανάλυση, εκμεταλλευόμενα τη χρονική τους συγχρονία με έναν αυτο-επιβλεπόμενο τρόπο.

Κεφάλαιο 3

Πολυκαναλική επεξεργασία φωνής σε έξυπνα/ρομποτικά περιβάλλοντα

Όπως αναφέρθηκε και εισαγωγικά, η συγκεκριμένη ενότητα εστιάζει στην εφαρμογή αλγορίθμων και μεθόδων φωνητικής αλληλεπίδρασης σε έξυπνους χώρους ή σε εφαρμογές αλληλεπίδρασης ανθρώπου-ρομπότ, όπου πολλαπλοί αισθητήρες επιτρέπουν την αυτόματη αναγνώριση της κατάστασης και των προθέσεων του χρήστη καθώς και την υποστήριξη διάφορων αυτοματισμών [196, 197, 198]. Στη συνέχεια αναφέρονται αναλυτικά οι συνεισφορές αυτής της εργασίας στο πεδίο της πολυκαναλικής επεξεργασίας φωνής, δηλαδή η ανάπτυξη συστήματος εντοπισμού ομιλητή για έξυπνα περιβάλλοντα, ο οπτικοακουστικός εντοπισμός ομιλητή με εφαρμογή σε αλληλεπίδραση ανθρώπου-ρομπότ, το σύστημα πολυκαναλικής αποθορυβοποίησης ομιλίας, ένα σύστημα αναγνώρισης φωνής και ένα σύστημα διαλόγου/κατανόησης λόγου με εφαρμογή σε ρομποτικούς βοηθούς.

3.1 Εντοπισμός ομιλητή με χρήση πολλαπλών μικροφώνων

Στο περιβάλλον των έξυπνων σπιτιών, η θέση του ομιλητή είναι πολύ σημαντική τόσο γιατί μπορεί να χρησιμοποιηθεί σαν είσοδος σε συστήματα αποθορυβοποίησης ενισχύοντας τη φωνή και μειώνοντας το θόρυβο (και κατ΄ επέκταση βελτιώνοντας το αποτέλεσμα της αναγνώρισης φωνής), όπως επίσης για να αναγνωριστεί το δωμάτιο στο οποίο βρίσκεται ο ομιλητής προκειμένου το σύστημα να ανταποκριθεί στην εντολή του χρήστη με τη σωστή ενέργεια.

Για τον εντοπισμό του ομιλητή σε ένα χώρο και ιδιαίτερα σε ένα έξυπνο σπίτι με πολλά δωμάτια, είναι απαραίτητη η ύπαρξη πολλαπλών μικροφώνων, σε συστοιχίες κατανεμημένες κατάλληλα ώστε να καλύπτουν το χώρο. Παρόλο που έχει γίνει αρκετά εκτενής έρευνα προς αυτή την κατεύθυνση, όταν οι συνθήκες του περιβάλλοντος είναι δυσμενείς, δηλαδή υπάρχει πολύς θόρυβος και υψηλοί χρόνοι αντήχησης ($0.6s < T_{60}$ [1]$< 2s$) το πρόβλημα του εντοπισμού

[1] Ως T_{60} ορίζεται ο χρόνος που χρειάζεται ώστε η ενέργεια της αντήχησης να μειωθεί κατά 60 dB [199]

ομιλητή γίνεται αρκετά δυσκολότερο.

Οι υπάρχοντες αλγόριθμοι εντοπισμού ομιλητή με πολλαπλά μικρόφωνα μπορούν να χωριστούν σε τρεις μεγάλες κατηγορίες: μεθόδους που βασίζονται σε εκτίμηση: α) Steered Response Power (SRP), β) High Resolution Spectral Estimation (HRSE) και γ) Time Difference of Arrival (TDOA). Μια αναλυτική παρουσίαση όλως των μεθόδων μπορεί να βρεθεί στο [200]. Στο [201] διερευνάται η απόδοση της μεθόδου εκτίμησης TDOA σε σχέση με την αντήχηση των δωματίων και παρατηρείται ότι η αντήχηση δημιουργεί προβλήματα ακόμα και όταν βρίσκεται σε χαμηλά επίπεδα. Η εργασία [202] μοντελοποιεί την αντήχηση σε σχέση με τον εντοπισμό ομιλητή αλλά η συγκεκριμένη ανάλυση απαιτεί τα μικρόφωνα να βρίσκονται μακριά από τους τοίχους, κάτι που είναι δύσκολο σε περιβάλλοντα έξυπνου σπιτιού. Οι περισσότερες δημοσιευμένες εργασίες αναφέρουν πειράματα και αποτελέσματα σε πολύ μικρές βάσεις δεδομένων με περιορισμένο αριθμό πιθανών θέσεων και μέτριες συνθήκες.

Η συγκεκριμένη εργασία μελετά το συγκεκριμένο πρόβλημα σε περιβάλλοντα με υψηλή αντήχηση (μέχρι $T_{60} = 2s$) και στοχεύει να αυξήσει την ευρωστία του συστήματος χωρίς να έχει γνώση των χαρακτηριστικών και της γεωμετρίας του περιβάλλοντα χώρου. Προτείνεται ένα στάδιο απο-αντήχησης (dereverberation) προκειμένου να βελτιωθεί η ποιότητα των δεδομένων και μια επαναληπτική μέθοδος περιορισμού αποκλίνοντων σημείων (outliers) ώστε να βελτιωθεί η ακρίβεια εντοπισμού του ομιλητή. Γίνεται πειραματισμός τόσο με προσομειωμένα όσο και με πραγματικά δεδομένα που περιλαμβάνουν ένα μεγάλο αριθμό πηγών ήχου και σχετικά μικρό αριθμό μικροφώνων, καταλήγοντας σε ενδιαφέρουσες παρατηρήσεις. Στις επόμενες υπο-ενότητες παρουσιάζονται αναλυτικά η μέθοδος εντοπισμού ομιλητή, τα χαρακτηριστικά των βάσεων, τα πειράματα και τα αποτελέσματα.

3.1.1 Προτεινόμενη μέθοδος

Η παρούσα διατριβή[2] εστιάζει στην κατηγορία των μεθόδων εκτίμησης TDOA και συγκεκριμένα βασίζεται στην ευρέως χρησιμοποιούμενη μέθοδο γενικευμένης ετεροσυσχέτισης με μετασχηματισμό φάσης (Generalized Cross Correlation - PHAse Transform (GCC-PHAT)) [203], διότι υπάρχει η ανάγκη το σύστημα να είναι υπολογιστικά αποδοτικό και να τρέχει σε πραγματικό χρόνο. Επίσης, έχει παρατηρηθεί [204] ότι ο μετασχηματισμός φάσης είναι βέλτιστος σε σχέση με άλλες μεθόδους εκτίμησης TDOA όταν υπάρχει υψηλή αντήχηση.

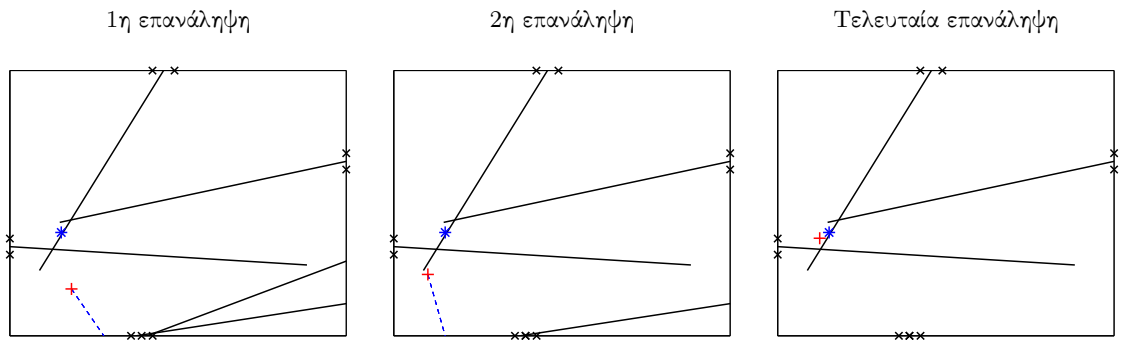
Εκτίμηση TDOA

Δοθέντος ενός μικροφώνου i, η έξοδός του μπορεί να εκφραστεί ως:

$$x_i(t) \quad = a_i s(t - \tau_i) + u_i(t) \tag{3.1}$$

όπου x_i είναι η έξοδος, s το πηγαίο σήμα, τ_i είναι το time-of-flight (TOF) δηλαδή ο χρόνος που κάνει να φτάσει το σήμα από την πηγή στο μικρόφωνο, u_i ο θόρυβος και a_i ο παράγοντας

[2]Η παρούσα εργασία υλοποιήθηκε στα πλαίσια του FP7 Ευρωπαϊκού Προγράμματος DIRHA (Distant-Speech Interaction for Robust Home Applications) https://dirha.fbk.eu



Σχήμα 3.1: Παράδειγμα της προτεινόμενης μεθόδου περιορισμού αποκλίνοντων σημείων (με μπλε $*$ συμβολίζεται η πραγματική θέση του ομιλητή, με κόκκινο $+$ η εκτίμηση της θέσης του ομιλητή, με μαύρα x οι θέσεις των μικροφώνων και με διακεκομμένη γραμμή η μέγιστη απόσταση από την εκτιμώμενη θέση προς κάποιο πιθανό αποκλίνον σημείο.)

απόσβεσης εξαιτίας της καθυστέρησης διάδοσης του σήματος από την πηγή στο μικρόφωνο.

Το πρόβλημα της εκτίμησης TDOA εστιάζεται στην εκτίμηση του $\Delta\tau_{ij} = \tau_i - \tau_j$ για ένα ζεύγος μικροφώνων (i, j). Πολλές μέθοδοι έχουν προταθεί για αυτό το σκοπό, μερικές από τις οποίες παρουσιάζονται στο [205]. Το παρόν σύστημα χρησιμοποιεί το Crosspower Spectrum Phase - Coherence Measure (CSP-CM) [206], που βασίζεται στο GCC-PHAT. Είναι κατάλληλο για εφαρμογές πραγματικού χρόνου γιατί είναι υπολογιστικά αποδοτικό.

Με $X_i(f, t)$ συμβολίζεται ο Μετασχηματισμός Fourier βραχέος χρόνου (Short-Time Fourier Transform (STFT)) του σήματος $x_i(t)$. Η μέθοδος CSP-CM υπολογίζει το

$$C_{ij}(\tau, t) = \int_{-\infty}^{\infty} \frac{X_i(f,t)X_j^*(f,t)}{|X_i(f,t)X_j(f,t)|} \exp^{j2\pi f\tau} df \tag{3.2}$$

και κάνει εκτίμηση του TDOA ως $\Delta\tau_{ij} = \arg\max_\tau C_{ij}(\tau, t)$ επειδή το $C_{ij}(\tau, t)$ αναμένεται να έχει τοπικό μέγιστο στο $\tau = \Delta\tau_{ij}$.

Εκτίμηση κατεύθυνσης άφιξης (Direction of Arrival (DOA))

Αφού υπολογιστεί το TDOA για κάθε ζεύγος μικροφώνων, μπορεί να εξαχθούν οι κατευθύνσεις άφιξης του πηγαίου σήματος (DOA) σε σχέση με τα διάφορα ζεύγη. Όταν τα μικρόφωνα και οι πηγές των σημάτων βρίσκονται στο ίδιο επίπεδο, ο γεωμετρικός τόπος των σημείων που αναπαριστούν πιθανές θέσεις ομιλητή είναι ένα μισό υπερβολοειδές στον $2D$ χώρο. Υποθέτοντας μοντέλο διάδοσης μακρινού πεδίου (far-field), η DOA μπορεί να αναπαρασταθεί με μια γραμμή που ενώνει την πηγή και το μέσο του ευθύγραμμου τμήματος που ενώνει τα δύο μικρόφωνα του ζεύγους. Τότε έχουμε [207]:

$$\frac{d\cos\theta}{c} = \Delta\tau_{ij} \Rightarrow \theta = \cos^{-1}\left(\frac{c\Delta\tau_{ij}}{d}\right) \tag{3.3}$$

όπου θ είναι η γωνία της DOA, d η απόσταση των δύο μικροφώνων και c η ταχύτητα του ήχου.

Εκτίμηση θέσης ομιλητή με ελάχιστα τετράγωνα

Εάν οι εκτιμήσεις TDOA ήταν ιδανικές και με απόλυτη ακρίβεια σωστές, όλες οι DOA γραμμές θα τέμνονταν σε ένα κοινό σημείο. Σε πραγματικές συνθήκες αυτό δε μπορεί να συμβεί οπότε αναζητείται ένας τρόπος κατάλληλου συνδυασμού των σημείων ώστε να γίνει η τελική εκτίμηση της θέσης του ομιλητή. Η προσέγγιση που ακολουθήθηκε βασίζεται στην εύρεση του σημείου εκείνου το οποίο ελαχιστοποιεί το άθροισμα των τετραγωνικών αποστάσεων από τις DOA γραμμές.

Με $\boldsymbol{\rho}_k$ συμβολίζεται το μοναδιαίο διάνυσμα που είναι παράλληλο προς την k-DOA γραμμή και με $\boldsymbol{\pi}_{0k}$ το σημείο στο μέσο του ευθύγραμμου τμήματος που ενώνει τα δύο μικρόφωνα του ζεύγους k. Για ένα τυχαίο σημείο $\boldsymbol{\alpha}$ και με $\boldsymbol{\alpha}_{proj_k}$ την προβολή του $\boldsymbol{\alpha}$ στην k-DOA γραμμή, υπολογίζεται η απόσταση $D_k^2(\boldsymbol{\alpha}) = ||\boldsymbol{\alpha} - \boldsymbol{\alpha}_{proj_k}||^2$, για όλα τα k. Θέτοντας $\mathbf{A}_k = \mathbf{I} - \boldsymbol{\rho}_k\boldsymbol{\rho}_k^T$ όπου $\mathbf{I}$ ο μοναδιαίος πίνακας:

$$D_k^2(\boldsymbol{\alpha}) = (\boldsymbol{\alpha} - \boldsymbol{\pi}_{0k})^T \mathbf{A}_k (\boldsymbol{\alpha} - \boldsymbol{\pi}_{0k}) \tag{3.4}$$

Θέλουμε να ελάχιστοποιήσουμε:

$$E(\boldsymbol{\alpha}) = \sum_{k=1}^{M} D_k^2(\boldsymbol{\alpha}) \tag{3.5}$$

όπου M είναι ο αριθμός των DOA γραμμών. Πρακτικά καταλήγουμε σε μια λύση κλειστής μορφής χρησιμοποιώντας ελάχιστα τετράγωνα (Least Squares (LS)) [208].

Βελτιώνοντας την ευρωστία του συστήματος

Όπως αναφέρθηκε και στην αρχή της ενότητας, η ακρίβεια της εκτίμησης TDOA εξαρτάται σημαντικά από τους χρόνους αντήχησης, τα επίπεδα θορύβου καθώς και από τον προσανατολισμό του ομιλητή. Οποιαδήποτε από αυτές τις παραμέτρους μπορεί να επηρεάσει αρνητικά την απόδοση του συστήματος. Προκειμένου να βελτιωθεί η ευρωστία του προτείνονται μια απλή μέθοδος απο-αντήχησης dereverberation για να βελτιωθεί η ποιότητα των δεδομένων, καθώς και μία μέθοδος περιορισμού αποκλίνοντων σημείων διότι οι εκτιμήσεις που προέρχονται από κάποια ζεύγη μικροφώνων οι οποίες δεν είναι αξιόπιστες αλλοιώνουν σημαντικά την τελική εκτίμηση.

1. Προ-φιλτράρισμα

Η μέθοδος απο-αντήχησης εφαρμόζεται πριν την εκτίμηση TDOA. Το φαινόμενο της αντήχησης μπορεί να εκφραστεί ως:

$$x_i(t) = s(t) * h_i(t) + u_i(t) \tag{3.6}$$

όπου $h_i(t)$ είναι η κρουστική απόκριση (impulse response) μεταξύ του μικροφώνου i και της θέσης της πηγής του σήματος. Η μέθοδος απο-αντήχησης βασίζεται σε cepstral φιλτράρισμα [209]. Υποθέτοντας ακίνητες πηγές και γραμμικά κανάλια που μεταβάλλονται αργά στο χρόνο, μπορούμε να μεταβούμε στο πεδίο του cepstrum, μετατρέποντας το συνελικτικό στοιχείο σε άθροιση, την οποία μπορούμε να αντιμετωπίσουμε με γραμμικό φιλτράρισμα. Το

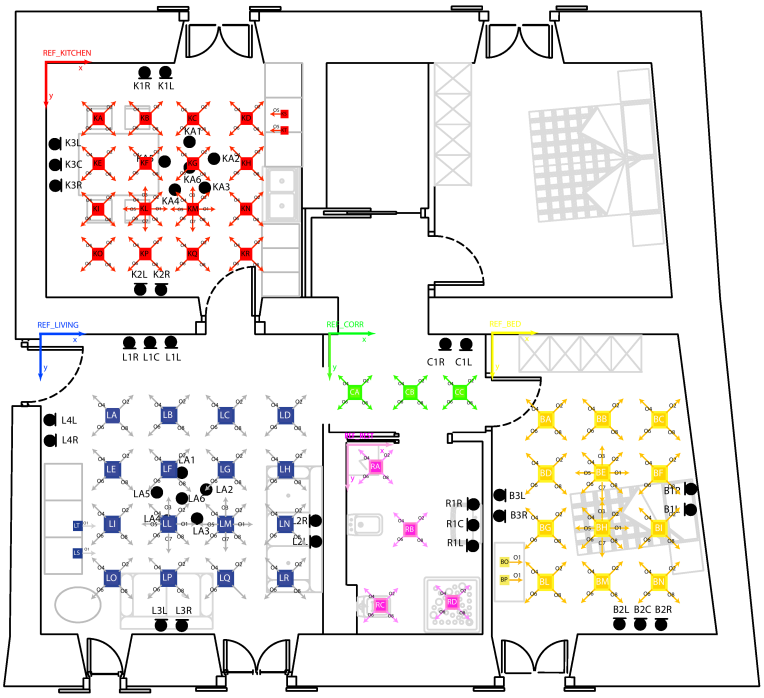
μιγαδικό cepstrum προτιμάται έναντι του πραγματικού καθώς διατηρεί πληροφορίες σχετικά με τη φάση που είναι απαραίτητες για την ανακατασκευή του σήματος. Συμβολίζοντας με $\hat{h}_i[k]$ το cepstrum της κρουστικής απόκρισης, αποδεικνύεται ότι $\hat{h}_i[k] = \hat{h}_{i,ap}[k] + \hat{h}_{i,min}[k]$, όπου $\hat{h}_{i,ap}[k]$ είναι ένα all-pass στοιχείο και $\hat{h}_{i,min}[k]$ ένα στοιχείο ελάχιστης φάσης (Minimum phase component (MPC)). Αποδεικνύεται ότι η απο-αντήχηση μπορεί να επιτευχθεί εάν αφαιρεθεί το MPC από το cepstrum του καναλιού, υποθέτοντας αμελητέο προσθετικό θόρυβο σε σχέση με το συνελικτικό, δηλαδή την αντήχηση και ότι το MPC του πηγαίου σήματος έχει μηδενική μέση τιμή.

2. Περιορισμός των outliers

Όπως αναφέρθηκε προηγουμένως η αντήχηση μπορεί να οδηγήσει σε λανθασμένες εκτιμήσεις TDOA για κάποια ζεύγη μικροφώνων. Με την προηγούμενη μέθοδο περιορίζεται η αντήχηση αλλά όχι εντελώς. Ανάμεσα στα διαθέσιμα ζεύγη μικροφώνων υπάρχουν κάποια τα οποία είναι αξιόπιστα και μπορούν με ακρίβεια να εκτιμήσουν σωστά την πραγματική θέση της πηγής. Η μέθοδος που θα παρουσιαστεί στη συνέχεια επικεντρώνεται στην εύρεση και απομάκρυνση των ζευγών αυτών που αλλοιώνουν το αποτέλεσμα, τα οποία αναφέρονται ως αποκλίνοντα σημεία (outliers). Προκειμένου να βρεθεί κάποιος αντικειμενικός τρόπος ανίχνευσης των αναξιόπιστων ζευγών μικροφώνων, δοκιμάστηκαν τρεις μέθοδοι: α) επιλογή ζευγών βασισμένη στο σηματοθορυβικό λόγο (Signal to Noise Ratio (SNR)) η οποία υποθέτει ότι τα πιο αξιόπιστα ζεύγη έχουν υψηλότερο SNR (βασισμένη στο [207]), β) επιλογή ζευγών με βάση την τιμή της κορυφής της ετεροσυσχέτισης μεταξύ των δύο μικροφώνων θεωρώντας ότι μεγαλύτερη κορυφή εγγυάται καλύτερη εκτίμηση και γ) επιλογή ζευγών βασισμένη στη διασπορά του TDOA ώστε να απομακρύνονται τα ζεύγη που δίνουν σημαντικά διαφορετικές εκτιμήσεις για συναπτά παράθυρα ηχητικού σήματος. Κάποιες από τις υποθέσεις επιβεβαιώθηκαν από ένα μέρος των δεδομένων αλλά καμία μέθοδος δε μπόρεσε να εφαρμοστεί και να αποδώσει για το σύνολο των δεδομένων λόγω της δυσκολίας και πολυπλοκότητας των συνθηκών. Στο [210] οι συγγραφείς προτείνουν μια επαναληπτική μέθοδο επίλυσης του συστήματος των DOA εξισώσεων που επιτρέπει την απομάκρυνση των αποκλίσεων, τη μέθοδο προβολής. Εκφράζοντας το σύστημα των γραμμών DOA ως $\mathbf{A}\boldsymbol{\xi} = \boldsymbol{\beta}$ όπου $\boldsymbol{\xi}$ η άγνωστη θέση της πηγής, ο προτεινόμενος αλγόριθμός επαναλαμβάνεται για όλες τις DOA εξισώσεις προβάλλοντας τη λύση στα ημιεπίπεδα που αντιπροσωπεύει η κάθε εξίσωση. Στο βήμα $i+1$, η προβαλλόμενη λύση είναι:

$$\boldsymbol{\xi}^{i+1} = \boldsymbol{\xi}^{i} + \frac{e_i}{|\boldsymbol{\alpha}_p|^2}\boldsymbol{\alpha}_p^T \tag{3.7}$$

όπου $\boldsymbol{\alpha}_p$ είναι η p^{th} γραμμή του πίνακα $\mathbf{A}$. Στην i^{th} επανάληψη, χρησιμοποιείται η p^{th} γραμμή και το $e_i = b_p - \boldsymbol{\alpha}_p\boldsymbol{\xi}^i$ συμβολίζει το λάθος, όπου b_p είναι το p^{th} στοιχείο του διανύσματος $\boldsymbol{\beta}$. Η εξίσωση με τη μεγαλύτερη απόσταση από το σημείο προβολής είναι πιθανό αποκλίνον σημείο και απομακρύνεται αν το λάθος ξεπερνάει ένα συγκεκριμένο κατώφλι. Ο αλγόριθμος περιορίζει ένα σημείο σε κάθε επανάληψη και τερματίζει όταν όλα τα λάθη βρίσκονται κάτω από το ορισμένο κατώφλι, ή αν έχει ολοκληρωθεί ο επιθυμητός αριθμός επαναλήψεων. Η μέθοδος αυτή είναι αρκετά αποδοτική όταν ο αριθμός των εξισώσεων είναι μεγάλος γιατί συγκλίνει ασυμπτωτικά στη λύση των ελαχίστων τετραγώνων. Στην περίπτωση που εξετάζουμε, εξαιτίας του μικρού



Σχήμα 3.2: Κάτοψη του έξυπνου σπιτιού DIRHA όπου απεικονίζονται όλες οι πιθανές θέσεις και προσανατολισμοί των ομιλητών καθώς και τα 40 μικρόφωνα([1]).

αριθμού των μικροφώνων, η μέθοδος αυτή δε δίνει πολύ καλά αποτελέσματα. Προτείνεται μια εναλλακτική επαναληπτική μέθοδος:

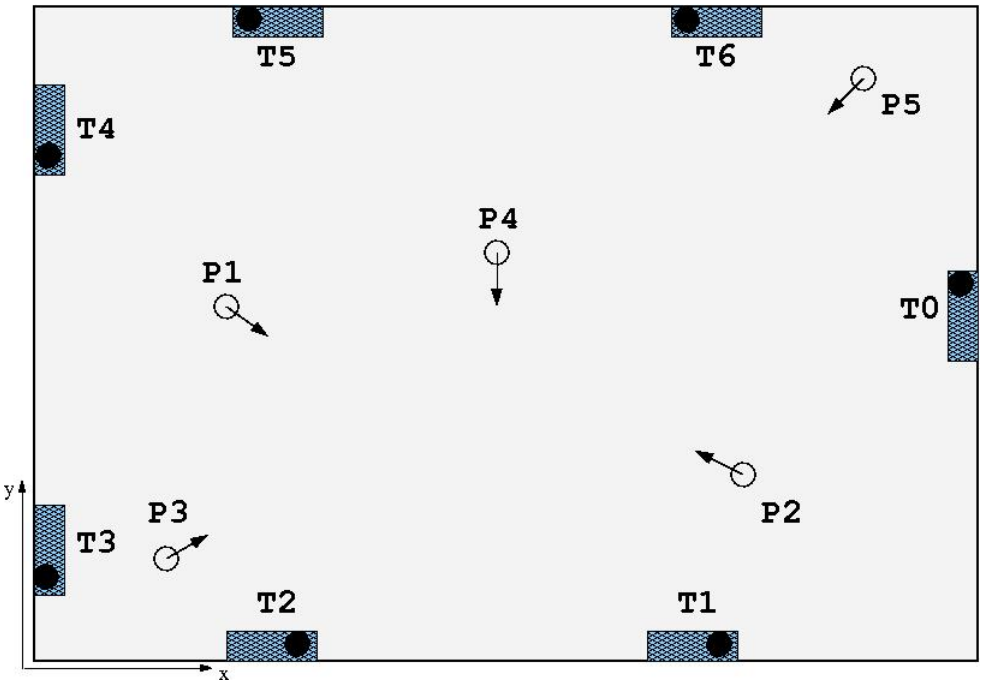
Αρχικά γίνεται μια εκτίμηση της θέσης της πηγής χρησιμοποιώντας όλες τις διαθέσιμες γραμμές DOA, όπως περιγράφηκε προηγουμένως. Κατόπιν υπολογίζονται οι αποστάσεις μεταξύ κάθε γραμμής DOA από την εκτιμώμενη θέση και επιλέγεται η μέγιστη απόσταση η οποία συγκρίνεται με ένα προκαθορισμένο κατώφλι. Αν το ξεπερνά, η γραμμή DOA που αντιστοιχεί σε αυτή τη γραμμή απομακρύνεται και υπολογίζεται μία νέα εκτίμηση της θέσης της πηγής με τις γραμμές DOA που έχουν απομείνει. Αυτή η διαδικασία επαναλαμβάνεται μέχρι είτε να μην υπάρχει καμία γραμμή που να ξεπερνά το κατώφλι είτε όταν απομείνουν μόνο δύο γραμμές. Ένα παράδειγμα απεικονίζεται στο Σχήμα 3.1.

3.1.2 Βάσεις δεδομένων

1. Προσομειωμένα και πραγματικά δεδομένα DIRHA

Όσον αφορά το πειραματικό μέρος, χρησιμοποιήθηκαν δύο σετ δεδομένων [24], προσομειωμένα και πραγματικά, που δημιουργήθηκαν στα πλαίσια του έργου DIRHA [211] σε ένα έξυπνο σπίτι στην πόλη Τρέντο της Ιταλίας. Το σπίτι αυτό διαθέτει 40 μικρόφωνα κατανεμημένα σε 12 συστοιχίες των δύο ή τριών μικροφώνων στους τοίχους καθώς και δύο συστοιχίες των 6 μικροφώνων που βρίσκονται εγκατεστημένες στο ταβάνι δύο δωματίων, του σαλονιού και της κουζίνας (Σχήμα 3.2). Το διαμέρισμα έχει πολύ υψηλή αντήχηση ($0.60s \leq T_{60} \leq 2s$).

Στην περίπτωση των προσομειωμένων δεδομένων, αρχικά ηχογραφήθηκε καθαρός λόγος σε ήσυχο περιβάλλον σε τέσσερις γλώσσες (Αυστριακά, Ελληνικά, Ιταλικά και Πορτογαλικά).



Σχήμα 3.3: Κάτοψη της βάσης DMN

Τα δεδομένα αυτά στη συνέχεια συνελίχθηκαν με τις κρουστικές αποκρίσεις του έξυπνου σπιτιού DIRHA για ένα μεγάλο εύρος στατικών πηγών, ενώ ταυτόχρονα έγινε υπέρθεση άλλων ακουστικών γεγονότων παρασκηνίου και θορύβου σε υψηλά επίπεδα, ώστε να δημιουργηθούν ρεαλιστικά πολυκαναλικά δεδομένα φωνής από απόσταση. Συνολικά διατίθενται 40 συνεδρίες του ενός λεπτού, 10 για κάθε γλώσσα, οι οποίες περιέχουν πάνω από ένα φωνητικό απόσμασμα σε διαφορετικές θέσεις/προσανατολισμούς. Για τον εντοπισμό θέσης ομιλητή πρέπει να αναφερθεί ότι υπάρχουν λίγα διαθέσιμα ζεύγη μικροφώνων που μπορούν να χρησιμοποιηθούν γιατί είναι αραιά κατανεμημένα στα δωμάτια. Τα πραγματικά δεδομένα περιέχουν 10 συνεδρίες αλληλεπίδρασης wizard-of-Oz μεταξύ χρηστών και του συστήματος ομιλίας του σπιτιού στα Ιταλικά. Σε όλες τις συνεδρίες ο χρήστης βρίσκεται είτε στο σαλόνι είτε στην κουζίνα και συνήθως κινείται. Αυτά τα δεδομένα δεν περιέχουν άλλα ακουστικά γεγονότα [1].

2. Βάση DMN

Αρχικά πραγματοποιήθηκε ένα πρώτο πείραμα επιβεβαίωσης στη βάση Distributed Microphone Network database (DMN), που μας παραχωρήθηκε από το Fondazione Bruno Kessler (FBK) [212]. Η βάση αυτή, σε αντίθεση με τις βάσεις DIRHA, περιέχει έναν αρκετά μεγάλο αριθμό διαθέσιμων μικροφώνων σε ένα μόνο δωμάτιο και είναι αρκετά μικρή. Συγκεκριμένα διατίθενται 21 μικρόφωνα διατεταγμένα σε 7 συστοιχίες των τριών μικροφώνων (Σχήμα 3.3) και περιλαμβάνονται 5 ηχογραφήσεις σε 5 διαφορετικές θέσεις. Επίσης έχει μέτριο χρόνο αντήχησης ($T_{60} = 0.7s$) και θορύβου. Πειράματα εντοπισμού ομιλητή στη συγκεκριμένη βάση έχουν γίνει και στην εργασία [213].

Oracle	Upper bound without pre-filtering
Oracle-D	Upper bound with pre-filtering
CSP	CSP-CM
CSP-O	CSP-CM with outlier elimination
CSP-D	CSP-CM with pre-filtering
CSP-D-O	CSP-CM with pre-filtering and outlier elimination
SRP	SRP-PHAT
SRP-D	SRP-PHAT with pre-filtering

Πίνακας 3.1: Οι διάφορες προσεγγίσεις εντοπισμού ομιλητή που ακολουθήθηκαν.

	Pcor	RMSE (in cm)
CSP	100%	34
SRP	100%	9

Πίνακας 3.2: Αποτελέσματα στη βάση DMN.

3.1.3 Πειραματική αξιολόγηση

Στο πρόβλημα του εντοπισμού της θέσης του ομιλητή, υπάρχει διαχωρισμός μεταξύ των μικρών λαθών (fine errors) και των μεγάλων (gross errors) με τα πρώτα να αντιστοιχούν σε εκτιμήσεις με απόκλιση μικρότερη από 50cm ενώ τα δεύτερα σε εκτιμήσεις με απόκλιση μεγαλύτερη από 50cm. Το ποσοστό των fine errors (τα οποία πρακτικά θεωρούνται σωστές εκτιμήσεις) στο σύνολο των εκτιμήσεων αναφέρεται ως μετρική Pcor. Επίσης υπολογίζεται το Μέσο τετραγωνικό σφάλμα (Root Mean Square Error), τόσο για τα fine errors (RMSEf) όσο και το συνολικό (RMSE). Στον Πίνακα 3.1 αναφέρονται συνοπτικά οι μέθοδοι που υλοποιήθηκαν.

Τα αποτελέσματα που αφορούν τη βάση DMN παρουσιάζονται στον Πίνακα 3.2. Εδώ εξετάστηκαν μόνο δύο μέθοδοι αναφοράς (baseline systems), το CSP-CM όπως περιγράφηκε παραπάνω αλλά χωρίς τις βελτιώσεις που προτείνονται, και το σύστημα SRP-PHAT [214, 215], το οποίο ορίζεται ως:

$$P(q,t) = \sum_{i=1}^{M}\sum_{j=1}^{M}\int_{-\infty}^{\infty}\frac{X_i(f,t)X_j^*(f,t)}{|X_i(f,t)X_j(f,t)|}\exp^{j2\pi f\tau}\,df \tag{3.8}$$

για ένα συγκεκριμένο σημείο q του τρισδιάστατου χώρου, όπου M ο αριθμός των διαθέσιμων μικροφώνων. Στην ουσία, η μέθοδος SRP-PHAT βασίζεται στο στρέψιμο ενός beamformer σε κάθε σημείο του χώρου και τον υπολογισμό της ενέργειας, σύμφωνα με την εξίσωση 3.8. Η ζητούμενη πηγή του ομιλητή βρίσκεται παίρνοντας το σημείο εκείνο το οποίο δίνει τη μέγιστη ενέργεια.

Και τα δύο συστήματα επιτυγχάνουν 100% σωστό εντοπισμό ομιλητή, ενώ το δεύτερο δίνει πιο αξιόπιστες εκτιμήσεις με RMSE μόλις 9cm.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα που αφορούν στις βάσεις DIRHA. Για να αξιολογήσουμε καλύτερα τις μεθόδους, αρχικά υπολογίζεται ένα άνω όριο των καλύτερων εκτιμήσεων που θα μπορούσαμε να έχουμε από τον προτεινόμενο αλγόριθμο, εάν μπορούσαν

να βρεθούν τα πιο αξιόπιστα ζεύγη μικροφώνων. Ως κίνητρο λειτούργησε η παρατήρηση ότι ενώ η μέθοδος CSP-CM αποτυγχάνει στις περισσότερες περιπτώσεις να δώσει ικανοποιητικές εκτιμήσεις, ανάμεσα στα διαθέσιμα ζεύγη μικροφώνων υπήρχαν κάποια τα οποία θα έδιναν σωστές εκτιμήσεις εάν ήταν τα μόνα διαθέσιμα. Γνωρίζοντας τις πραγματικές θέσεις των πηγών από τα δεδομένα μας, πειραματιστήκαμε με όλους τους συνδυασμούς ζευγών μικροφώνων και αποκτήσαμε ένα oracle αποτέλεσμα που λειτουργεί ως άνω όριο της απόδοσης του αλγορίθμου. Επίσης για λόγους σύγκρισης έγινε πειραματισμός και με τον αλγόριθμο SRP-PHAT με και χωρίς το στάδιο απο-αντήχησης.

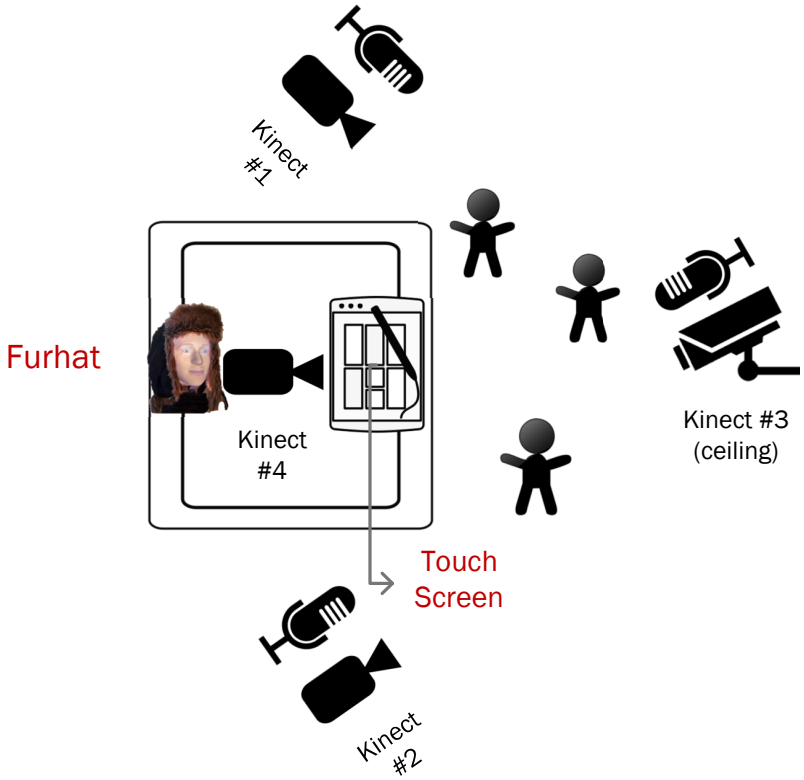
Στον Πίνακα 3.3 συνοψίζονται τα αποτελέσματα για όλες τις υλοποιημένες μεθόδους στις βασεις DIRHA. Όπως γίνεται αντιληπτό, το πρόβλημα είναι αρκετά δύσκολο και όλες οι βασικές μέθοδοι αναφοράς (baseline methods) αποτυγχάνουν να παράξουν σωστές εκτιμήσεις για το σύνολο των δεδομένων. Ωστόσο, το αποτέλεσμα "Oracle-D" υποδεικνύει ότι αν γνωρίζαμε τα πιο αξιόπιστα μικρόφωνα θα μπορούσαμε να πάρουμε αρκετά αξιόπιστες εκτιμήσεις. Οι δύο προτεινόμενες βελτιώσεις, δηλαδή το στάδιο απο-αντήχησης και η μέθοδος περιορισμού αποκλίνοντων σημείων αυξάνουν την ευρωστία του βασικού συστήματος αλλά δεν καταφέρνουν να επιτύχουν ένα υψηλό ποσοστό "Pcor". Στην περίπτωση των προσομειωμένων δεδομένων, η καλύτερη επίδοση επιτυγχάνεται από το προτεινόμενο σύστημα "CSP-D-O" με 51% fine errors, ενώ το "SRP-D" επιτυγχάνει 48% fine errors. Όσον αφορά τα πραγματικά δεδομένα όπου ο ομιλητής κινείται αργά στο χρόνο, τα καλύτερα αποτελέσματα προκύπτουν από τα "CSP","CSP-D", "CSP-D-O". Εδώ φαίνεται ότι οι προτεινόμενες βελτιώσεις δεν προσφέρουν πολλά στην ευρωστία του συστήματος. Αυτό μπορεί να εξηγηθεί με δύο τρόπους: α) η προσέγγιση απο-αντήχησης που ακολουθήθηκε υποθέτει στατικό ομιλητή, το οποίο σημαίνει ότι η υπόθεση του αργά μεταβαλλόμενου καναλιού δεν είναι ακριβής και β) αντίθετα από τα προσομειωμένα δεδομένα όπου η πηγή είναι πολύ πιο κατευθυντική, στα πραγματικά δεδομένα ο περιορισμός των αποκλίσεων δε φαίνεται να προσθέτει πολλά. Επίσης η χαμηλή επίδοση του συστήματος SRP-PHAT στις βάσεις DIRHA σε αντίθεση με τη βάση DMN μπορεί να αποδοθεί στο μικρό αριθμό των διαθέσιμων ζευγών μικροφώνων. Ενώ το σύστημα αυτό δίνει ακριβείς εκτιμήσεις, χρειάζεται έναν αρκετά μεγάλο αριθμό ζευγών για να λειτουργεί εύρωστα. Τέλος, πρέπει να αναφερθεί ο χρόνος εκτέλεσης των αλγορίθμων: Οι μέθοδοι CSP δίνουν αποτελέσματα σε πραγματικό χρόνο ενώ το σύστημα SRP-PHAT όπως υλοποιήθηκε με βάση την εργασία [215] είναι αρκετά αργό και επομένως όχι κατάλληλο για τέτοιες εφαρμογές.

3.1.4 Συμπεράσματα

Τα αποτελέσματα που παρουσιάστηκαν αναδεικνύουν τη δυσκολία εντοπισμού του ομιλητή όταν οι συνθήκες περιλαμβάνουν υψηλή αντήχηση και θόρυβο και ο διαθέσιμος αριθμός μικροφώνων είναι μικρός. Παρουσιάστηκαν δύο βελτιώσεις που αυξάνουν την ευρωστία του συστήματος, μία μέθοδος απο-αντήχησης και μία περιορισμού των αποκλίσεων. Τα αποτελέσματα δείχνουν ότι μπορεί να επιτευχθεί ικανοποιητική ακρίβεια ακόμα και κάτω από δύσκολες συνθήκες εάν μπορούν να ανιχνευθούν σωστά οι αποκλίσεις. Χρειάζεται περαιτέρω έρευνα γύρω από τη μοντελοποίηση της αντήχησης στα έξυπνα περιβάλλοντα καθώς και την εύρεση τρόπων

	Simulated Data			Real Data		
	Pcor	RMSEf (in cm)	RMSE (in cm)	Pcor	RMSEf (in cm)	RMSE (in cm)
Oracle	77.3%	16	37	89.8%	20	27
Oracle-D	84.3%	15	28	82.3%	17	28
CSP	30.0%	27	109	49.1%	32	63
CSP-O	40.3%	19	210	46.8%	29	69
CSP-D	44.7%	27	75	49.1%	27	98
CSP-D-O	51.0%	18	78	49.1%	27	75
SRP	16.9%	19	178	11.0%	23	166
SRP-D	48.0%	21	105	18.7%	21	159

Πίνακας 3.3: Αποτελέσματα στις βάσεις DIRHA.

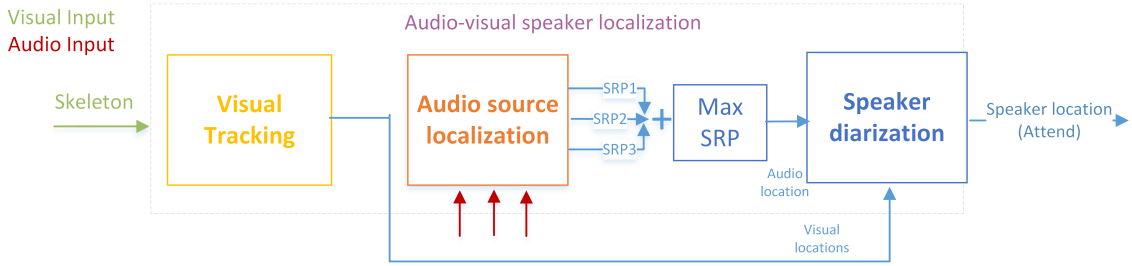


Σχήμα 3.4: Κάτοψη του χώρου των πειραμάτων.

αποτελεσματικού περιορισμού των αποκλίσεων.

3.2 Οπτικοακουστικός εντοπισμός ομιλητή

Η δουλειά που παρουσιάστηκε στην προηγούμενη ενότητα αφορούσε εντοπισμό ομιλητή στο 2D επίπεδο μόνο με συστοιχίες μικροφώνων χωρίς τη χρήση οπτικής πληροφορίας. Ο αλγόριθμος που παρουσιάστηκε είναι χρήσιμος σε εφαρμογές όπου ισχύει η υπόθεση ότι ο ομιλητής βρίσκεται στο ίδιο επίπεδο με τα μικρόφωνα και όπου δεν είναι απαραίτητο να γνωρίζουμε



Σχήμα 3.5: Οπτικοακουστικός εντοπισμός ομιλητή.

τη διάσταση z του ομιλητή, καθώς επίσης και η υπόθεση ότι έχουμε μόνο έναν ομιλητή ο οποίος είναι ακίνητος. Παρ΄ όλα αυτά υπάρχουν αρκετές εφαρμογές όπως οι εφαρμογές αλληλεπίδρασης ανθρώπου με ρομπότ και ακόμα περισσότερο παιδιού με ρομπότ όπου είναι δεν ισχύουν αυτές οι υποθέσεις. Συνήθως σε μια ακουστική ή οπτικοακουστική σκηνή υπάρχουν πολλοί άνθρωποι οι οποίοι είτε είναι ενεργοί ομιλητές είτε όχι και μπορεί να κινούνται ή να είναι στατικοί. Ταυτόχρονα μας ενδιαφέρει να γνωρίζουμε και το ύψος του ομιλητή γιατί είναι επιθυμητό το ρομπότ να στρέφει το κεφάλι και την προσοχή του προς τον ομιλητή. Για όλους αυτούς τους λόγους η προηγούμενη μεθοδολογία δεν επαρκεί.

Από την άλλη πλευρά, σε τέτοιες εφαρμογές, όπως σε αυτή που παρουσιάζεται σε αυτή την ενότητα, υπάρχουν τόσο συστοιχίες μικροφώνων όσο και κάμερες που καταγράφουν οπτική πληροφορία η οποία είναι απαραίτητη για άλλες εφαρμογές της αλληλεπίδρασης ανθρώπου-ρομπότ. Μπορούμε να εκμεταλλευτούμε αυτή την οπτική πληροφορία που συχνά είναι πιο αξιόπιστη από την ακουστική, και συνδυάζοντάς τη με την ακουστική να βελτιώσουμε τα αποτελέσματα του εντοπισμού θέσης ομιλητή. Η οπτική πληροφορία από μόνη της δεν επαρκεί γιατί μπορεί μεν να δίνει αξιόπιστα αποτελέσματα όσον αφορά την ύπαρξη ανθρώπων στο χώρο αλλά δε μπορεί να ανιχνεύσει το άτομο που μιλάει ανάμεσά τους. Οπότε χρειάζεται ένας συνδυασμός των δύο τροπικοτήτων ώστε να γίνει αποτελεσματικά ο εντοπισμός και η ανίχνευση ομιλητή.

Στη βιβλιογραφία υπάρχουν αρκετές τεχνικές για ακουστικό εντοπισμό ομιλητή [216], με κάποιες από αυτές να έχουν αναπτυχθεί ειδικά για εφαρμογές αλληλεπίδρασης ανθρώπου ρομπότ [54, 52] όπου τα μικρόφωνα βρίσκονται πάνω στα ρομπότ. Επίσης όσον αφορά τεχνικές οπτικοακουστικού εντοπισμού ομιλητή έχουν αναπτυχθεί διάφορες μέθοδοι [217, 218, 219], οι περισσότερες από τις οποίες βασίζονται σε Bayesian φιλτράρισμα ή σε σύμμειξη μεταξύ οπτικών και ακουστικών χαρακτηριστικών. Αυτές οι τεχνικές έχουν αναπτυχθεί κυρίως για συμβατικές κάμερες ή κάμερες ενσωματωμένες σε ρομπότ.

Η εφαρμογή η οποία μελετάται στο συγκεκριμένο κεφάλαιο χρησιμοποιεί ταυτόχρονα σαν κάμερες και σαν συστοιχίες μικροφώνων αισθητήρες Kinect V2 κατανεμημένους στο χώρο σε συγκεκριμένες σταθερές θέσεις. Κάθε αισθητήρας Kinect περιλαμβάνει μια κάμερα υψηλής ανάλυσης (με χρονική ανάλυση των 30 πλαισίων/δευτερόλεπτο), έναν αισθητήρα βάθους που χρησιμοποιεί τη μέθοδο Time-of-Flight, και τέσσερα μικρόφωνα. Όπως φαίνεται και στο Σχήμα 3.4, δύο αισθητήρες Kinect είναι τοποθετημένοι συμμετρικά δεξιά και αριστερά από το ρομπότ, ένας στη μέση και ένας στο ταβάνι με την κάμερα να κοιτάζει προς το πάτωμα.

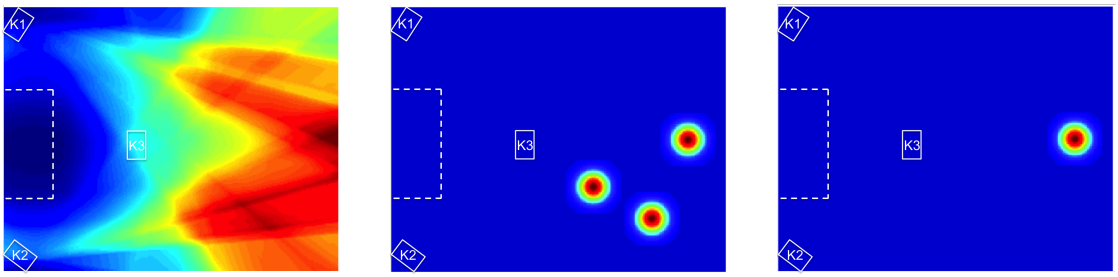
Audio Source Localization			Audio-visual speaker diarization
Pcor	RMSE	RMSEf	Pcor
45.51	0.60m	0.35m	85.58

Πίνακας 3.4: Αξιολόγηση του συστήματος οπτικοακουστικού εντοπισμού ομιλητή.

Για τον εντοπισμό ομιλητή με ακουστική πληροφορία χρησιμοποιείται ο αλγόριθμος SRP-PHAT (Steered Response Power - Phase Transform) [215, 213], ο οποίος αναφέρθηκε και στο προηγούμενο κεφάλαιο αλλά προσαρμόστηκε ώστε να λειτουργεί στον τρισδιάστατο χώρο και σε πραγματικό χρόνο και δεδομένου του αριθμού των διαθέσιμων μικροφώνων αναμένουμε να είναι αρκετά εύρωστος σε θόρυβο. Στο συγκεκριμένο πειραματικό πλαίσιο, θεωρώντας ένα καθολικό σύστημα συντεταγμένων, υπολογίζεται ανά Kinect συστοιχία η 3D SRP ενέργεια και στη συνέχεια αθροίζονται αυτές οι ενέργειες. Το σημείο που έχει τη μέγιστη SRP ενέργεια αποτελεί την εκτίμηση της θέσης του ομιλητή. Η οπτική παρακολούθηση/εντοπισμός ανθρώπου στο συγκεκριμένο πλαίσιο γίνεται μέσω της παρακολούθησης σκελετού που έχει αναπτυχθεί ειδικά για τον αισθητήρα Kinect. Η ένωση της οπτικής με την ακουστική πληροφορία γίνεται ως εξής: Αρχικά εντοπίζονται όλοι οι άνθρωποι που βρίσκονται στην οπτικοακουστική σκηνή μέσω των σκελετών τους. Παράλληλα, μέσω των πολυκαναλικών δεδομένων φωνής γίνεται η εκτίμηση της θέσης του ομιλητή. Ο συνδυασμός των δύο πληροφοριών γίνεται διαλέγοντας τη θέση του σκελετού που βρίσκεται πιο κοντά στην εκτίμηση θέσης με ακουστικά δεδομένα. Σχηματικά η διαδικασία παρουσιάζεται στο Σχήμα 3.5

Ένα παράδειγμα του αλγορίθμου απεικονίζεται στο Σχήμα 3.6, ενώ αριθμητικά αποτελέσματα από πειράματα παρουσιάζονται στον Πίνακα 3.4. Τα δεδομένα τα οποία χρησιμοποιήθηκαν για την αξιολόγηση του συστήματος περιλαμβάνουν οπτικοακουστική πληροφορία από 20 ενήλικες και 28 παιδιά. Στα δεδομένα αυτά, κάθε άτομο εκφώνησε περίπου 130 προτάσεις/φράσεις, ενώ στη σκηνή βρίσκονταν περισσότεροι άνθρωποι οι οποίοι όμως δε μιλούσαν.

Στον Πίνακα 3.4 επιχειρείται μια σύγκριση μεταξύ του ακουστικού εντοπισμού και του οπτικοακουστικού. Όσον αφορά τον πρώτο, οι μετρικές είναι οι ίδιες με την προηγούμενη ενότητα, δηλαδή τα Pcor (Percentage correct), RMSE (Root Mean Square Error), RMSEf. Όσον αφορά το δεύτερο, εφόσον οι θέσεις των ατόμων υπολογίζονται απευθείας από τον αισθητήρα Kinect ουσιαστικά διαλέγουμε τη θέση του ομιλητή ανάμεσα από κάποιες πιθανές θέσεις οπότε οι δύο τελευταίες μετρικές δεν μπορούν να εφαρμοστούν στο πρόβλημα αυτό. Επομένως μετράμε το σωστό εντοπισμό ομιλητή ανάμεσα στο σύνολο των θέσεων ατόμων που έχουν εντοπιστεί. Ο ακουστικός εντοπισμός δε λειτουργεί ικανοποιητικά, δίνοντας Pcor μόλις 45%, αλλά με RMSE 60cm, που σημαίνει ότι το μέσο λάθος εντοπισμού είναι λίγο μεγαλύτερο από αυτό το κατώφλι που διαχωρίζει τα μεγάλα από τα μικρά λάθη εντοπισμού ($50cm$). Ο οπτικοακουστικός εντοπισμός αυξάνει πολύ την απόδοση του συστήματος, επιτυγχάνοντας ποσοστό 86% σωστών εκτιμήσεων.



Σχήμα 3.6: Παράδειγμα του οπτικοακουστικού εντοπισμού ομιλητή. Αριστερά: Ακουστικός εντοπισμός ομιλητή (όσο υψηλότερες είναι οι SRP τιμές τόσο πιο κόκκινη είναι η εικόνα). Μέση: Οπτικός εντοπισμός ανθρώπων. Δεξιά: Οπτικοακουστικός εντοπισμός ομιλητή. Φαίνονται επίσης οι 3 αισθητήρες Kinect.

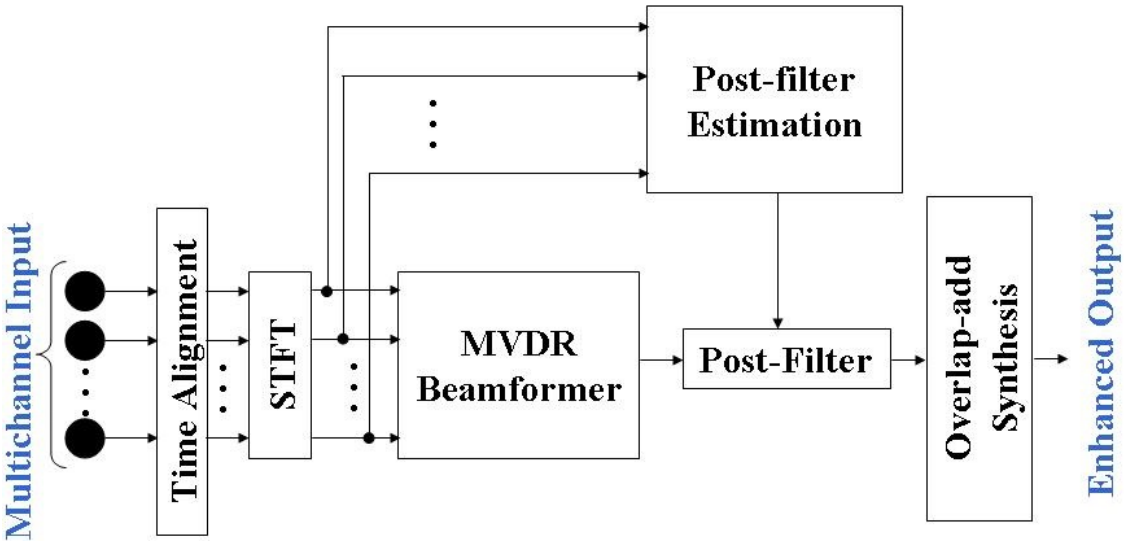
3.3 Πολυκαναλική αποθορυβοποίηση φωνής

Στην ενότητα αυτή παρουσιάζεται το σύστημα αποθορυβοποίησης ομιλίας που αποτελεί μέρος του συνολικού συστήματος πολυκαναλικής επεξεργασίας φωνής τόσο για έξυπνα περιβάλλοντα όσο και για περιβάλλοντα αλληλεπίδρασης ανθρώπου-ρομπότ. Στη συγκεκριμένη ενότητα η συνεισφορά έγκειται στην εφαρμογή ήδη υπάρχοντων αλγορίθμων σε διάφορες βάσεις δεδομένων προκειμένου να διερευνηθεί η καταλληλότητά τους και να συγκριθούν μεταξύ τους ως προς την απόδοση.

3.3.1 Σύστημα πολυκαναλικής αποθορυβοποίησης

Τα πολυκαναλικά δεδομένα φωνής επιτρέπουν την αξιοποίηση της χωρικής πληροφορίας ώστε να γίνεται αποθορυβοποίηση μέσω τεχνικών beamforming [220, 10, 221, 205]. Συνήθως για τέτοιες τεχνικές είναι απαραίτητη η γνώση/εκτίμηση της θέσης του ομιλητή, η οποία γίνεται με διάφορες τεχνικές, κάποιες από τις οποίες αναφέρθηκαν στην προηγούμεη ενότητα. Για την αποθορυβοποίηση πολυκαναλικής ομιλίας έχει αναπτυχθεί ένα σύστημα [222] το οποίο χρησιμοποιεί MVDR (Minimum Variance Distortionless Response) beamforming και μετα-φιλτράρισμα (post-filter) με βέλτιστους εκτιμητές κυματομορφής [221, 223, 224] MMSE, STSA [225], και log-STSA [226]. Στη συνέχεια παρουσιάζεται εν συντομία το σύστημα και η πειραματική αξιολόγησή του σε κάποιες βάσεις δεδομένων. Οι λεπτομέρειες του συστήματος και πιο αναλυτική παρουσίασή του μπορούν να βρεθούν στο [222].

Το σύστημα αποτελείται από ένα υποσύστημα εντοπισμού θέσης ομιλητή, ένα υποσύστημα ευθυγράμμισης χρόνου, τον MVDR beamformer και ένα μετα-φίλτρο (post-filter), όπως φαίνεται και στο Σχήμα 3.7. Το πρώτο υποσύστημα περιγράφηκε στις προηγούμενες ενότητες. Το δεύτερο υποσύστημα ευθυγραμμίζει χρονικά τα πολλαπλά κανάλια με βάση την εκτίμηση της θέσης του ομιλητή. Στη συνέχεια γίνεται ανάλυση βραχέος χρόνου, η οποία ακολουθείται από beamforming και post-filtering. Η έξοδος ανασυντίθεται χρησιμοποιώντας τη μέθοδο overlap-add.



Σχήμα 3.7: Σύστημα πολυκαναλικής αποθορυβοποίησης ομιλίας.

3.3.2 Πειραματική αξιολόγηση

Για την πειραματική αξιολόγηση του συστήματος χρησιμοποιήθηκε η βάση προσομειωμένων δεδομένων DIRHA που περιγράφηκε στην προηγούμενη ενότητα. Συγκεκριμένα ο αλγόριθμος αποθορυβοποίησης εφαρμόστηκε στη συστοιχία μικροφώνων που βρίσκεται στο ταβάνι (σαλόνι και κουζίνα). Ο στόχος είναι η αποθορυβοποίηση των συνεδριών που περιέχουν φωνητικά δεδομένα από το σαλόνι και την κουζίνα και η σύγκριση του αποτελέσματος αποθορυβοποίησης με έναν Delay-and-Sum (DSB) beamformer που είναι σχετικά απλός αλλά έχει αποδειχθεί αρκετά αποτελεσματικός, ο οποίος εφαρμόζεται στα ίδια μικρόφωνα. Εξετάστηκαν δύο περιπτώσεις σε σχέση με τη θέση του ομιλητή, η οποία αποτελεί είσοδο στο σύστημα: 1) Στην πρώτη περίπτωση χρησιμοποιήθηκαν οι πραγματικές θέσεις πηγής, όπως αυτές δίνονται από τη βάση δεδομένων. 2) Στη δεύτερη εφαρμόστηκε ο αλγόριθμος εντοπισμού ομιλητή που αναφέρθηκε στην Ενότητα 3.1 για να υπολογιστούν οι θέσεις του ομιλητή.

Ως μετρική αξιολόγησης χρησιμοποιήθηκε μια μορφή του σηματοθορυβικού λόγου (SNR) η οποία ονομάζεται Segmental Signal-to-Noise Ratio (SSNR) που είναι πιο κατάλληλη για φωνητικά δεδομένα. Πιο συγκεκριμένα τα αποτελέσματα παρουσιάζονται και ως προς το μέσο SSNR αλλά και ως προς το μέσο SSNRE το οποίο υπολογίζεται ως η διαφορά της εξόδου προς την είσοδο σε dB. Θεωρείται ως μικρόφωνο αναφοράς το μικρόφωνο που βρίσκεται στο κέντρο της εξαγωνικής συστοιχίας. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 3.5. Στις στήλες 4, 5, 6 παρουσιάζεται ο MVDR beamformer με τα MMSE, STSA, log-STSA post-filters αντίστοιχα, χρησιμοποιώντας τις πραγματικές θέσεις ομιλητή ενώ στις τρεις επόμενες στήλες έχουν χρησιμοποιηθεί οι εκτιμήσεις των θέσεων του ομιλητή. Τέλος, στην τελευταία στήλη παρουσιάζονται τα αποτελέσματα του DSB beamforming με τις πραγματικές θέσεις.

Συνολικά ο MVDR beamformer ακολουθούμενος από οποιοδήποτε post-filter επιτυγχάνει να αυξήσει σημαντικά το SSNR της εξόδου σε σχέση με τον DSB. Σε απόλυτο αριθμό βέβαια τα μέσα SSNR είναι αρνητικά σε όλες τις περιπτώσεις, γεγονός που εξηγείται από την

Test set	Room	Metric: mean	MVDR+postfilter						DSB
			Ground truth locations			Estimated locations			Ground truth locations
			MMSE	STSA	log-STSA	MMSE	STSA	log-STSA	
dev1	LR	SSNR	−3.50	−3.54	−3.51	−4.06	−4.18	−4.11	−5.23
		SSNRE	1.69	1.65	1.67	1.16	1.03	1.10	−0.03
	K	SSNR	−3.54	−3.57	−3.55	−4.23	−4.26	−4.24	−5.49
		SSNRE	1.88	1.85	1.87	1.20	1.17	1.19	−0.06
test1	LR	SSNR	−3.38	−3.42	−3.40	−3.97	−4.09	−4.02	−5.33
		SSNRE	2.08	2.04	2.06	1.48	1.36	1.43	0.12
	K	SSNR	−3.42	−3.46	−3.43	−4.53	−4.57	−4.54	−5.65
		SSNRE	2.22	2.18	2.21	1.11	1.07	1.10	−0.01

Πίνακας 3.5: Αποτελέσματα αποθορυβοποίησης στη βάση προσωμειωμένων δεδομένων DIRHA στα Ελληνικά ανά δωμάτιο και σετ δεομένων. Το LR συμβολίζει το σαλόνι και K την κουζίνα. Όλα τα αποτελέσματα είναι σε dB.

ύπαρξη πολύ υψηλού θορύβου, τόσο διάχυτου όσο και εντοπισμένου, άλλων ακουστικών γεγονότων που διαδραματίζονται ταυτόχρονα, καθώς και υψηλούς χρόνους αντήχησης. Επίσης, λόγω της απόστασης των μικροφώνων από την πηγή ομιλίας, το σήμα όταν καταγράφεται από τα μικρόφωνα έχει υποστεί αρκετή εξασθένηση. Παρά τα αρνητικά SSNR, το σύστημα επιτυγχάνει να μειώσει αρκετά το θόρυβο, σε αντίθεση με τον DSB που δεν καταφέρνει να αποθορυβοποιήσει ικανοποιητικά τα σήματα ομιλίας. Ακόμα και στην περίπτωση των εκτιμώμενων θέσεων ομιλητή όπου αναμένουμε να έχουμε χειρότερα αποτελέσματα αποθορυβοποίησης, το προτεινόμενο σύστημα ξεπερνά σε απόδοση τον DSB με πραγματικές θέσεις ομιλητή. Όσον αφορά την απόδοση των διαφορετικών post-filters όλα έχουν περίπου την ίδια επίδοση, με λίγο καλύτερο το MMSE post-filter. Παρ΄ ότι τα συμπεράσματα αυτά μπορούν να επιβεβαιωθούν και από το ηχητικό αποτέλεσμα, με εμφανή την αποθορυβοποίηση και την ενίσχυση του επιθυμητού σήματος, εμφανίζονται σε αυτό κάποιες μικρές παραμορφώσεις ως αποτέλεσμα της επεξεργασίας των σημάτων από τον MVDR beamformer και τα post-filters.

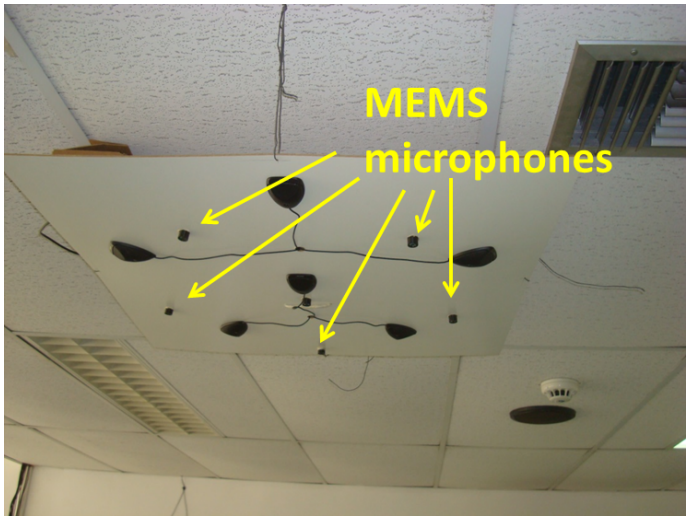
3.3.3 Σύγκριση μικροφώνων MEMS με πυκνωτικά μικρόφωνα (ECMs) στο πρόβλημα της αποθορυβοποίησης ομιλίας

Οι συστοιχίες μικροφώνων γίνονται ολοένα και πιο δημοφιλείς και αρχίζουν να χρησιμοποιούνται σε πολλές και διαφορετικές εφαρμογές, κυρίως για εντοπισμό ομιλητή, αποθορυβοποίηση και αναγνώριση φωνής. Παρ΄ όλα αυτά το κόστος και η πολυπλοκότητα που παρουσιάζουν οι συστοιχίες που είναι διαθέσιμες στο εμπόριο είναι υπερβολικά υψηλά. Η χρήση πολλαπλών μικροφώνων εκτός από τα πλεονεκτήματά της παρουσιάζει και κάποιες δυσκολίες: όσο αυξάνει ο αριθμός των σημάτων, τόσο αυξάνει και η πολυπλοκότητα των ηλεκτρονικών κυκλωμάτων που επεξεργάζονται τα δεδομένα.

Η καινούρια τεχνολογία των μικροφώνων MEMS [227] διευκολύνει σε μεγάλο βαθμό την ενσωμάτωση ηχητικών αισθητήρων σε έξυπνα περιβάλλοντα διάχυτης ευφυίας και έχουν κάποια σημαντικά πλεονεκτήματα σε σχέση με τα πυκνωτικά μικρόφωνα: έχουν λιγότερη ευαισθησία ως προς τη θερμοκρασία και επίσης είναι πολύ φθηνότερα [228]. Εφόσον το κόστος τους είναι τριών τάξεων μικρότερο από τα πυκνωτικά, αποτελούν μια ελκυστική επιλογή.

Τα MEMS μικρόφωνα που έχουν χρησιμοποιηθεί σε αυτή την έρευνα έχουν κατασκευαστεί από την STMicroelectronics και πιο συγκεκριμένα είναι τα MP34DT01 [229]: πολύ συμπαγή στο μέγεθος, χαμηλής ενέργειας, λαμβάνουν σήματα από όλες τις κατευθύνσεις (omnidirectional) και είναι ψηφιακά με ενσωματωμένη κάρτα ήχου. Η αρχιτεκτονική τους αναδεικνύει την έμφαση που έχει δοθεί στο χαμηλό κόστος και στην ευκολία χρήσης. Η πλακέτα διαθέτει μέχρι 8 μικρόφωνα που είναι συνδεδεμένα με έναν μικροελεγκτή και η ροή δεδομένων μπορεί να γίνει αν συνδεθεί η πλακέτα με υπολογιστή.

Λόγω όλων των παραπάνων πλεονεκτημάτων που παρουσιάζουν τα μικρόφωνα MEMS σε σχέση με τα πυκνωτικά (ECMs) στην παρούσα ενότητα εξετάζεται εάν αυτά μπορούν να χρησιμοποιηθούν έναντι των πυκνωτικών σε εφαρμογές έξυπνου περιβάλλοντος. Συγκεκριμένα, συγκρίνεται η επίδοσή τους στο πρόβλημα της πολυκαναλικής αποθορυβοποίησης ομιλίας. Για να επιτευχθεί μια δίκαιη σύγκριση, πραγματοποιήθηκε συλλογή δεδομένων φωνής στα πλαίσια της βάσης 'Αθηνά' [230], η οποία παρουσιάζεται αναλυτικά στο Κεφάλαιο 4. Περιληπτικά η βάση αυτή περιέχει 4 ώρες ομιλίας στα Ελληνικά από 20 ομιλητές (10 άντρες, 10 γυναίκες) οι οποίες ηχογραφήθηκαν σε ένα πραγματικό έξυπνο περιβάλλον. Χρησιμοποιήθηκαν 20 πυκνωτικά μικρόφωνα κατανεμημένα στους τοίχους (14 μικρόφωνα) και στο ταβάνι (6 μικρόφωνα), καθώς και 6 MEMS στο ταβάνι. Τα MEMS τοποθετήθηκαν σε πενταγωνικό σχηματισμό με ένα μικρόφωνο στο κέντρο, ακριβώς όπως και τα ECMs της οροφής, όπως απεικονίζεται στο Σχήμα 3.8. Όσον αφορά την πειραματική διαδικασία που περιγράφεται παρακάτω, η σύγκριση αφορούσε μόνο τα MEMS και τα ECMs της οροφής.



Σχήμα 3.8: Οι πενταγωνικές συστοιχίες μικροφώνων της οροφής στη βάση 'Αθηνά'.

Το σύστημα πολυκαναλικής αποθορυβοοποίησης ομιλίας που περιγράφηκε στην προηγούμενη ενότητα εφαρμόστηκε στα δεδομένα που ηχογραφήθηκαν από τις δύο πενταγωνικές συστοιχίες. Χρησιμοποιήθηκαν οι πραγματικές θεσεις πηγής και μικροφώνων. Ο θόρυβος μοντελοποιήθηκε ως διάχυτος θόρυβος, διότι κατά τη διάρκεια των ηχογραφήσεων, προκειμένου τα δεδομένα να είναι ρεαλιστικά, υπήρχε και θόρυβος αλλά και άλλα ακουστικά γεγονότα ταυτόχρονα με την ομιλία. Τα αποτελέσματα της αποθορυβοποίησης παρουσιάζονται στον

Πίνακα 3.6 ως προς τη μετρική SSNRE. Παρατηρούμε ότι η επίδοση των χαμηλού κόστους MEMS είναι άμεσα συγκρίσιμη με αυτή των ακριβών ECMs με μια μικρή μείωση της τάξης του 0.04dB του μέσου SSNRE. Επομένως τα MEMS μπορούν να αποτελέσουν μία εναλλακτική πρόταση χαμηλού κόστους για το πρόβλημα της αποθορυβοποίησης ομιλίας.

Sensor Type	SSNRE (dB)
ECM	2.09
MEMS	2.05

Πίνακας 3.6: Πολυκαναλική αποθορυβοποίηση ομιλίας στη βάση 'Αθηνά'. Τα αποτελέσματα παρουσιάζονται ως προς το SSNRE σε dB.

3.4 Αναγνώριση φωνής από απόσταση

3.4.1 Το πολυ-καναλικό, πολυ-ρομποτικό σύστημα Multi3

Σε αυτή την ενότητα παρουσιάζεται ένα πολυ-καναλικό, πολυτροπικό, πολυ-ρομποτικό σύστημα το οποίο έχει αναπτυχθεί για την επικοινωνία ανθρώπων και ιδιαίτερα παιδιών με ρομπότ, και το οποίο ονομάζεται Multi3[3]. Τα παιδιά μπορούν να επιοκοινωνούν με το σύστημα με φυσικό τρόπο χρησιμοποιώντας τη φωνή τους, καθώς και χειρονομίες και κινήσεις. Για την καταγραφή των απαιτούμενων οπτικών αλλά και ηχητικών δεδομένων χρησιμοποιούνται πολλαπλές κάμερες Kinect, κατανεμημένες στο χώρο, οι οποίες επιτρέπουν την ασύρματη και απρόσκοπτη επικοινωνία των παιδιών με τα ρομπότ καλύπτοντας όλο το διαθέσιμο χώρο. Πιο συγκεκριμένα, το υλοποιημένο σύστημα αντίληψης για την αλληλεπίδραση παιδιού και ρομπότ είναι πολυτροπικό (δράσεις, χειρονομίες, φωνή), πολυ-καναλικό (πολλαπλά Kinect) και έχει τη δυνατότητα να ενσωματώσει την επικοινωνία με μια πληθώρα διαφορετικών ρομπότ, διότι είναι υλοποιημένο σε ROS (Robot Operating System [231]).

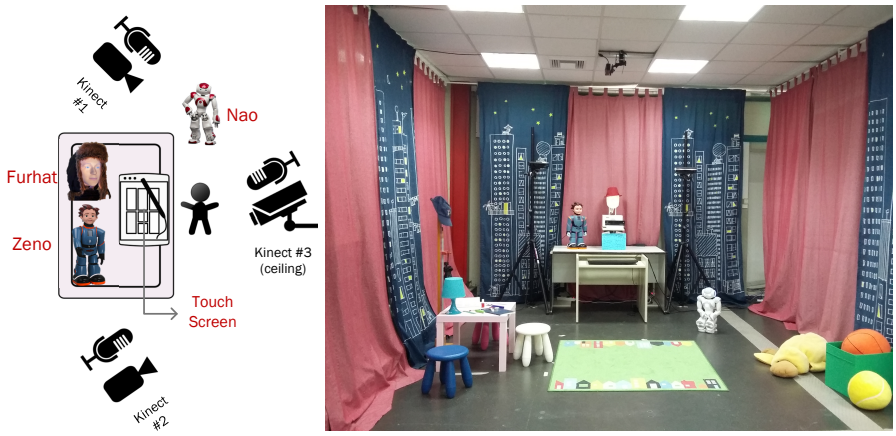
Η παρούσα εργασία εστιάζει στη συνιστώσα της φωνής, επομένως θα παρουσιαστεί η δομή του συστήματος αναγνώρισης φωνής από απόσταση και στη συνέχεια η αξιολόγηση με δεδομένα που συλλέχθηκαν από παιδιά.

Όπως αναφέρθηκε προηγουμένως, μια εφαρμογή αλληλεπίδρασης παιδιού-ρομπότ καθιστά απαραίτητη τη χρήση πολλαπλών αισθητήρων, κατανεμημένων στο χώρο, διότι για παράδειγμα, τα παιδιά μπορεί να κινούνται στο χώρο ή να υπάρχουν διάφοροι θόρυβοι που δυσκολεύουν τη διαδικασία αναγνώρισης φωνής. Η χωρική κατανομή των αισθητήρων για τη συγκεκριμένη εγκατάσταση απεικονίζεται στο Σχήμα 3.9.

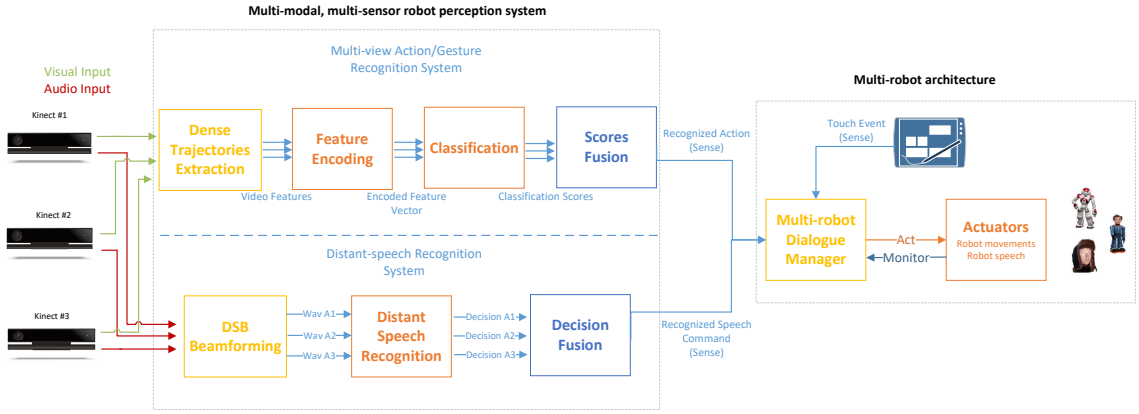
Χρησιμοποιούνται τρεις κάμερες Kinect V2 για τη λήψη τόσο οπτικών όσο και ακουστικών δεδομένων.

Το σύστημα σχεδιάστηκε και προσαρμόστηκε ειδικά για παιδιά. Στο Σχήμα 3.10 απεικονίζονται οι διάφορες συνιστώσες του συστήματος, όπως το σύστημα αναγνώρισης φωνής και

[3]Η παρούσα εργασία υλοποιήθηκε στα πλαίσια του EU Horizon 2020 προγράμματος BabyRobot https://babyrobot.eu



Σχήμα 3.9: Αριστερά: Κάτοψη του χώρου και της κατανομής των αισθητήρων και των διαθέσιμων ρομπότ στο σύστημα Multi3. Δεξιά: Φωτογραφία της διάταξης.



Σχήμα 3.10: Πολυτροπικό, πολυ-καναλικό σύστημα ρομποτικής αντίληψης και πολυ-ρομποτική αρχιτεκτονική. Τα A1, A2, A3 αναφέρονται στη συστοιχία μικροφώνων Kinect.

τα συστήματα αναγνώρισης δράσεων και χειρονομιών.

3.4.2 Πολυ-καναλική αναγνώριση φωνής από απόσταση

Η ενσωμάτωση μιας διεπαφής φωνής σε ένα σύστημα αλληλεπίδρασης ανθρώπου-ρομπότ είναι απαραίτητη ώστε να εξασφαλιστεί η φυσική επικοινωνία, αλλά ταυτόχρονα παρουσιάζει πολλές προκλήσεις εξαιτίας του θορύβου, της αντήχησης και της απόστασης μεταξύ των ομιλητών και των αισθητήρων [232]. Για να μπορέσουμε να αντιμετωπίσουμε όλα αυτά τα προβλήματα χρησιμοποιούμε πολυ-καναλική αναγνώριση φωνής από απόσταση (Distant Speech Recognition - DSR) [233, 234, 235] με τις τρεις συστοιχίες μικροφώνων από τις κάμερες Kinect.

Τα παιδιά/Οι χρήστες επικοινωνούν με το ρομπότ μέσω ενός συνόλου εντολών/φράσεων που δημιουργείται για την εκάστοτε εφαρμογή. Ένα σύστημα συνεχούς αναγνώρισης θα χρειαζόταν ένα πολύ μεγάλο όγκο δεδομένων για να εκπαιδευτεί/προσαρμοστεί σε παιδιά και ταυτόχρονα να έχει καλή επίδοση. Επομένως, αναπτύσσουμε ένα σύστημα αναγνώρισης φωνής με γραμματική.

Επεκτείνοντας μια παλαιότερη δουλειά [235], το σύστημα DSR 'ακούει' συνέχεια, δηλαδή είναι συνέχεια ενεργό και μπορεί να αναγνωρίσει κάποια φράση ανά πάσα στιγμή, ακόμα και αν υπάρχουν και άλλες πηγες ομιλίας ή θορύβου/αντήχησης. Για την παρούσα εφαρμογή, το σύνολο των εντολών που μπορούν να αναγνωριστούν περιέχουν πιθανές απαντήσεις των παιδιών σε κάποια παιχνίδια που έχουν σχεδιαστεί ώστε να παίζονται μεταξύ παιδιών και ρομπότ καθώς και κάποιες γενικές, συχνά χρησιμοποιούμενες φράσεις. Όλες οι φράσεις/προτάσεις στη συγκεκριμένη εφαρμογή είναι στα Ελληνικά. Περισσότερες λεπτομέρειες αναφέρονται στη συνέχεια.

Προκειμένου να ανατιμετωπιστούν αποτελεσματικά οι θόρυβοι και η αντήχηση και να ενισχυθεί η ευρωστία του συστήματος, γίνεται αφενός αποθορυβοποίηση των σημάτων μέσω beamforming και συγκεκριμένα DSB που είναι κατάλληλος για εφαρμογές πραγματικού χρόνου και αφετέρου προσαρμογή των μοντέλων φωνής.

Αρχικά για την αναγνώριση κάποιας προκαθορισμένης φράσης χρησιμοποιείται κυλιόμενο παράθυρο των 2.5sec με 0.6sec shift. Το αποτέλεσμα της αναγνώρισης μπορεί να είναι είτε μια προκαθορισμένη φράση είτε κενό, που σημαίνει είτε ότι δεν υπάρχει ομιλία είτε ότι η ομιλία είναι εκτός του προκαθορισμένου λεξιλογίου. Το DSB beamforming πραγματοποιείται σε κάθε συστοιχία των 4 μικροφώνων και βασίζεται στην εισαγωγή καθυστερήσεων στα διαφορετικά σήματα $a_n(t)$ ώστε να ευθυγραμμιστούν χρονικά και να ενισχύσουν το σήμα που έρχεται από μια συγκεκριμένη κατεύθυνση. Για ομοιόμορφες γραμμικές συστοιχίες με N μικρόφωνα, αν η επιθυμητή κατεύθυνση συμβολιστεί με ϕ, η χρονική καθυστέρηση που εισάγεται σε κάθε μικρόφωνο είναι

$$\tau_n = \frac{(n-1)d\cos\phi}{c}, \tag{3.9}$$

όπου c η ταχύτητα του ήχου και d η απόσταση μεταξύ δύο μικροφώνων. Το σήμα που προκύπτει μετά την εφαρμογή του beamforming είναι:

$$y(t) = \frac{1}{N}\sum_{n=1}^{N} a_n(t-\tau_n) \tag{3.10}$$

Όσον αφορά την ακουστική μοντελοποίηση, έχουν εκπαιδευτεί τριφωνικά GMM-HMM (Gaussian Mixture Models - Hidden Markov Models) cross-word μοντέλα, βασισμένα στα χαρακτηριστικά MFCC συν τις παραγώγους τους στη βάση δεδομένων Λογοτυπογραφία [33]. Συγκεκριμένα, τα δεδομένα της βάσης αυτής τροποποιήθηκαν τεχνητά συνελίσσοντας τον καθαρό λόγο με κρουστικές αποκρίσεις ενός δωματίου και προσθέτοντας γκαουσιανό θόρυβο έτσι ώστε τα μοντέλα να εκπαιδευτούν σε ομιλία από απόσταση, που είναι και το ζητούμενο. Οι κρουστικές αποκρίσεις έχουν μετρηθεί σε πραγματικό περιβάλλον χρησιμοποιώντας το σήμα exponential sine sweep [235, 236, 22]. Επιπλέον χρησιμοποιείται MLLR adaptation (Maximum Likelihood Linear Regression) για να μετασχηματιστούν οι μέσες τιμές των γκαουσιανών στις καταστάσεις των μοντέλων ώστε να μειωθεί η διαφορά ανάμεσα στα εκπαιδευμένα μοντέλα και στα δεδομένα προσαρμογής [237].

Κάθε συστοιχία παράγει ένα αποτέλεσμα αναγνώρισης. Επομένως για να μπορέσουμε να συνδυάσουμε αυτά τα διαφορετικά αποτελέσματα και ταυτόχρονα για να αυξηθεί ακόμα

περισσότερο η ευρωστία του συστήματος, συγχωνεύουμε τις υποθέσεις των τριών διαφορετικών συστοιχιών χρησιμοποιώντας 'ψηφοφορία πλειοψηφίας'. Εαν και οι τρεις υποθέσεις των διαφορετικών συστοιχιών διαφωνούν μεταξύ τους, λαμβάνονται υπόψη οι δεύτερες καλύτερες υποθέσεις.

Σενάρια χρήσης

Έχουν σχεδιαστεί δύο διαφορετικά σενάρια χρήσης-παιχνίδια ώστε να μπορέσει να αξιολογηθεί το σύστημα σε πραγματικά δεδομένα των συγκεκριμένων εφαρμογών. Το πρώτο αφορά ένα παιχνίδι παντομίμας μεταξύ ενός παιδιού και ενός ρομπότ ενώ το δέυτερο είναι ένα παιχνίδι που παίζεται μεταξύ πολλών ατόμων και ενός ρομπότ, το οποίο ονομάζεται *Φτιάξε μια φάρμα*. Παρουσιάζονται με συντομία τα δύο αυτά σενάρια.

1) *Παντομίμα*: Στο παιχνίδι αυτό, οι ρόλοι εναλλάσσονται συνεχώς μεταξύ παιδιού και ρομπότ. Εμφανίζονται δώδεκα κάρτες πάνω σε μία οθόνη αφής οι οποίες αναπαριστούν κάποια εργασία. Το ρομπότ διαλέγει τυχαία μία κάρτα και προσπαθεί να μιμηθεί την εργασία που απεικονίζεται ενώ το παιδί προσπαθεί να μαντέψει και να πει ποια εργασία είναι. Αν το παιδί δε βρει τη σωστή απάντηση, το ρομπότ το ρωτάει αν θέλει να ξαναπροσπαθήσει ή αν θέλει να του αποκαλύψει ποια ήταν η σωστή απάντηση. Μετά τον πρώτο γύρο, το ρομπότ και το παιδί αλλάζουν ρόλους. Το παιδί διαλέγει όποια κάρτα θέλει και κάνει παντομίμα, ενώ το ρομπότ προσπαθεί να μαντέψει ποια κίνηση είναι. Οι δώδεκα διαφορετικές κινήσεις είναι: βάψιμο τοίχου, καθάρισμα παραθύρων, οδήγηση λεωφορείου, κολύμπι, χορός, γυμναστική, κιθάρα, σκάψιμο μιας τρύπας, σφουγγάρισμα πατώματος, σιδέρωμα, κάρφωμα στον τοίχο και διάβασμα βιβλίου.

2) *Φτιάξε μια φάρμα*: Το παιχνίδι αυτό είναι της μορφής 'μάντεψε το αντικείμενο' και περιλαμβάνει δύο ρόλους, που μπορούν να παιχτούν ισότιμα από τους ανθρώπους και το ρομπότ. Όποιος αναλάβει τον πρώτο ρόλο πρέπει να διαλέξει ένα ζώο και να εκφωνεί διάφορα χαρακτηριστικά του. Ο δεύτερος ρόλος πρέπει να μαντέψει ποιο ζώο είναι. Όταν το ρομπότ έχει τον πρώτο ρόλο, οι άνθρωποι προσπαθούν να συνεργαστούν και να μαντέψουν ποιο ζώο μπορεί να είναι. Κάθε φορά που δε μαντεύουν σωστά, το ρομπότ αποκαλύπτει άλλο ένα χαρακτηριστικό του ζώου. Όταν βρεθεί το σωστό ζώο, το ρομπότ τους προτρέπει να τοποθετήσουν το ζώο στη σωστή θέση του στη φάρμα (π.χ. στο νερό, στους στάβλους, στα δέντρα, κλπ) σε μία οθόνη αφής που βρίσκεται μπροστά τους. Επομένως, ο σκοπός του παιχνιδιού είναι τόσο διασκέδαση όσο και εκπαίδευση γιατί το παιδί παρακινείται να μάθει χαρακτηριστικά διάφορων ζώων. Κατόπιν οι ρόλοι εναλλάσσονται: Οι άνθρωποι διαλέγουν συνεργατικά ένα ζώο και αποκαλύπτουν ένα χαρακτηριστικό τη φορά μέχρι το ρομπότ να μαντέψει τη σωστή απάντηση. Υπάρχουν 19 διαφορετικά ζώα και τα χαρακτηριστικά τους ανήκουν σε 5 διαφορετικές κατηγορίες: χρώμα, μέγεθος, είδος, αριθμός ποδιών και μία ξεχωριστή τους ιδιότητα. Ένα στιγμιότυπο της φάρμας με κάποια ζώα σωστά τοποθετημένα μπορεί να βρεθεί στο Σχήμα 3.11.

Σχήμα 3.11: Στιγμιότυπο του παιχνιδιού *Φτιάξε μία φάρμα*.

Αξιολόγηση του συστήματος

Για να επιτευχθεί ευρωστία και καλή επίδοση του συστήματος επιβάλλεται η ύπαρξη δεδομένων από παιδιά, διότι τα χαρακτηριστικά της φωνής τους διαφέρουν αρκετά από των ενηλίκων, π.χ. το pitch τους δεν έχει ακόμα διαμορφωθεί. Αν και υπάρχουν αρκετές βάσεις με δεδομένα ενηλίκων, δεν είναι κατάλληλα για την εκπαίδευση/προσαρμογή ενός συστήματος που θα απευθύνεται σε παιδιά. Την ίδια στιγμή η συλλογή δεδομένων από παιδιά είναι μια αρκετά δύσκολη και χρονοβόρα διαδικασία.

Πραγματοποιήθηκε μια συλλογή φωνητικών δεδομένων από παιδιά για κάθε ένα από τα σενάρια χρήσης που παρουσιάστηκαν. Συγκεκριμένα, ο αριθμός των παιδιών που συμμετείχαν ήταν 28 και καθένα από αυτά εκφώνησε περίπου 40 προτάσεις/φράσεις από το σύνολο και για τα δύο σενάρια των 120 φράσεων. Επίσης για λόγους αξιολόγησης και σύγκρισης, συλλέχθηκαν δεδομένα και από 20 ενήλικες, οι οποίοι εκφώνησαν περίπου 120 φράσεις.

1) *Παντομίμα*

Εδώ, η αξιολόγηση γίνεται μόνο στα παιδιά, διότι είναι ένα παιχνίδι που απευθύνεται μόνο σε παιδιά. Τα δεδομένα των παιδιών που συλλέχθηκαν χρησιμοποιούνται και για την προσαρμογή των φωνητικών μοντέλων αλλά και για την offline αξιολόγηση του συστήματος. Τα αποτελέσματα παρουσιάζονται στο Πίνακα 3.7 σε σχέση με word και sentence accuraccies, τα οποία συμβολίζονται με WCOR και SCOR αντίστοιχα. Οι τρεις πρώτες γραμμές του πίνακα περιέχουν τα αποτελέσματα για κάθε Kinect συστοιχία ξεχωριστά, ενώ η τελευταία περιέχει τα αποτελέσματα της συγχώνευσης υποθέσεων.

Χρησιμοποιήθηκαν τρία διαφορετικά σχήματα προσαρμογής: Στην περίπτωση "No-adapt" τα μοντέλα που αξιολογήθηκαν είχαν εκπαιδευτεί στη βάση Λογοτυπογραφία που περιέχει δεδομένα από ενήλικες και δεν έγινε κάποια περαιτέρω προσαρμογή τους. Τα διαθέσιμα δεδομενα των παιδιών χρησιμοποιήθηκαν για αξιολόγηση.

Στην περίπτωση "Adapt-all", τα δεδομένα από 20 παιδιά από τα 28 χρησιμοποιήθηκαν

	No-adapt		Adapt-all		Adapt-per-array	
	WCOR	SCOR	WCOR	SCOR	WCOR	SCOR
Kinect #1	79.30	70.53	98.41	95.95	98.30	95.95
Kinect #2	81.04	72.48	97.56	95.95	97.35	95.95
Kinect #3	76.85	66.83	97.45	94.60	97.56	94.60
Fusion	-	64.17	-	**98.97**	-	96.30

Πίνακας 3.7: Μέσο word (WCOR) και sentence accuracy (SCOR) (%) για την αναγνώριση φωνής σε παιδιά για το παιχνίδι της παντομίμας. Παρουσιάζονται αποτελέσματα για κάθε Kinect ξεχωριστά και για τη συγχώνευση των υποθέσεων για όλους τους τρόπους adaptation.

για την καθολική προσαρμογή των ακουστικών μοντέλων, δηλαδή τα δεδομένα και από τις 3 συστοιχίες χρησιμοποιήθηκαν για να προσαρμοστεί ένα κοινό φωνητικό μοντέλο. Τα δεδομένα από τους υπόλοιπους συμμετέχοντες χρησιμοποιοήθηκαν για αξιολόγηση. Η διαδικασία προσαρμογής/αξιολόγησης έγινε 4-fold cross-validated.

Τέλος, πειραματιστήκαμε και με την προσαρμογή ανά συστοιχία "adapt-per-array", προσαρμόζοντας ένα μοντέλο ανά συστοιχία με δεδομένα μόνο από τη συγκεκριμένη συστοιχία. Και εδώ χρησιμοποιήθηκε ο ίδιος αριθμός παιδιών για προσαρμογή και για αξιολόγηση και έγινε επίσης 4-fold cross-validation.

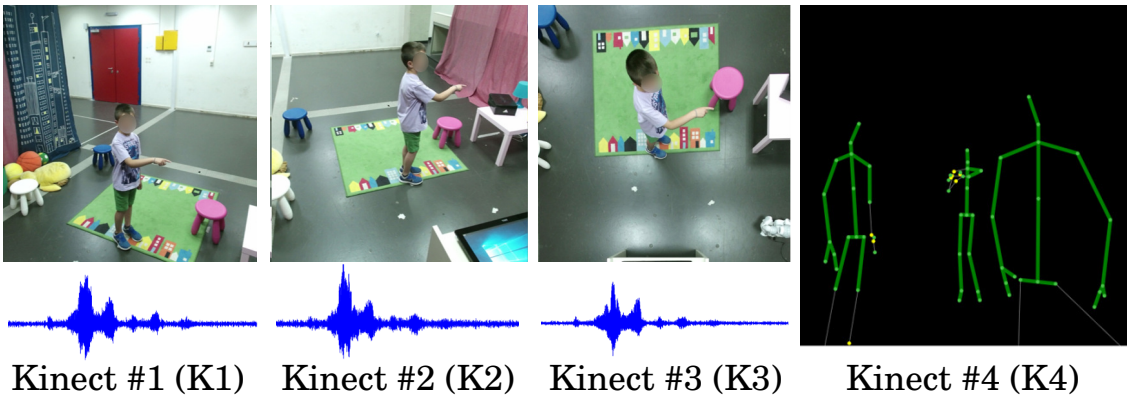
Στην πρώτη περίπτωση "no-adapt" παρατηρούμε ότι τα αποτελέσματα δεν είναι ικανοποιητικά, ενώ αντιθέτως η επίδοση του συστήματος είναι αρκετά υψηλή όταν γίνεται προσαρμογή με δεδομένα παιδιών. Το γεγονός αυτό αναδεικνύει ακόμα περισσότερο την αναγκαιότητα και σημαντικότητα της συλλογής δεδομένων από παιδιά. Τα άλλα δύο σχήματα προσαρμογής έχουν παρόμοια επίδοση, επομένως η προσαρμογή ανά συστοιχία δεν προσφέρει κάτι περισσότερο στην ευρωστία του συστήματος. Η επίδοση των διαφορετικών συστοιχιών είναι σχεδόν το ίδιο καλή και για τις τρεις, με την τελευταία να αποδίδει λίγο χειρότρα από τις άλλες. Όσον αφορά τη συγχώνευση υποθέσεων, αποδίδει χειρότερα στην περίπτωση της μη-προσαρμογής από ότι οι ξεχωριστές συστοιχίες, γιατί βασίζεται στη συγχώνευση αναξιόπιστων αποτελεσμάτων. Αντιθέτως, όταν τα μοντέλα έχουν προσαρμοστεί και οι υποθέσεις είναι καλύτερες, επιτυγχάνει σωστή αναγνώρισης της τάξης του 98.97% SCOR, που είναι η καλύτερη επίδοση που επιτυγχάνει το σύστημά μας συνολικά.

2) *Φτιάξε μια φάρμα*

Για το συγκεκριμένο παιχνίδι, χρησιμοποιώντας και τα δεδομένα των ενηλίκων, πειραματιστήκαμε με την προσαρμογή μοντέλων σε ενήλικες, παιδιά και μικτά και τα αξιολογήσαμε και με τα δύο σύνολα δεδομένων.

Τα αποτελέσματα παρουσιάζονται στο Πίνακα 3.8 μόνο για τη μετρική SCOR για τα δύο ηλικιακά γκρουπ ως σετ αξιολόγησης. Το "Average" αναφέρεται στο μέσο όρο των αποτελεσμάτων από όλες τις συστοιχίες, ενώ το "Fusion" είναι το αποτέλεσμα της συγχώνευσης υποθέσεων όπως και παραπάνω. Επίσης, όσον αφορά τα σχήματα προσαρμογής, αξιολογούνται τόσο τα μη-προσαρμοσμένα μοντέλα, όσο και τα προσαρμοσμένα σε ενήλικες, παιδιά και μικτά. Η διαδικασία της προσαρμογής/αξιολόγησης έγινε επίσης 4-fold cross-validated.

Η επίδοση του συστήματος αναγνώρισης φωνής στους ενήλικες είναι ικανοποιητική ακόμα



Σχήμα 3.12: Παράδειγμα των διαφόρων ροών πληροφορίας που καταγράφηκαν από 4 κάμερες Kinect στον πολυ-καναλικό πειραματικό χώρο: απεικονίζονται τα καρέ του βίντεο, ο ήχος και ο σκελετός του χρήστη.

		DSR-Adaptation scheme			
		No-adapt	Adults	Children	Mixed
Test		SCOR	SCOR	SCOR	SCOR
	K1	91.76	98.95	94.52	98.69
Adults	K2	90.60	98.70	90.99	97.85
	K3	91.39	98.95	94.11	98.75
	Avg	91.25	**98.87**	93.20	98.43
	Fuse	92.41	**99.82**	94.42	99.77
Children	K1	70.53	72.31	95.95	82.95
	K2	72.48	73.85	95.95	82.52
	K3	66.83	67.63	94.60	80.70
	Avg	69.95	71.20	**95.50**	82.06
	Fuse	64.17	66.02	**98.97**	95.51

Πίνακας 3.8: Αξιολόγηση του συστήματος αναγνώρισης φωνής για το παιχνίδι 'Φτιάξε μια φάρμα'.

και χωρίς προσαρμογή των μοντέλων. Παρ΄ όλα αυτά παρατηρούμε ότι η προσαρμογή των μοντέλων βελτιώνει την επίδοση σε όλες τις περιπτώσεις σε σχέση με τη μη-προσαρμογή, ακόμα κι αν γίνεται σε διαφορετικό ηλικιακό γκρουπ από την αξιολόγηση. Τα αποτελέσματα επίσης υποδεικνύουν ότι τα καλύτερα αποτελέσματα προκύπτουν όταν προσαρμόζουμε και αξιολογούμε στο ίδιο ηλικιακό γκρουπ, κάτι που είναι αναμενόμενο. Η συγχώνευση των αποφάσεων βελτιώνει ακόμα περισσότερο την επίδοση στις περισσότερες περιπτώσεις με τα καλύτερα αποτελέσματα να είναι 99.82% για τους ενήλικες και 98.97% για τα παιδιά.

3.5 Σύστημα διαλόγου/κατανόησης λόγου i-Walk

Η ραγδαία αύξηση του πληθυσμού με ειδικές ανάγκες και ειδικά των ηλικιωμένων, σε συνδυασμό με την παράλληλη μείωση του προσωπικού των φροντιστών ενισχύουν ακόμη περισσότερο την ανάγκη ύπαρξης ρομποτικών βοηθών [64]. Κατά το σχεδιασμό μιας έξυπνης πλατφόρμας

υποβοήθησης για ανθρώπους με κινητικά ή/και νοητικά προβλήματα, ειδική φροντίδα πρέπει να δίνεται ώστε το σύστημα να επιβλέπει και να βοηθά στην αποκατάσταση του ασθενούς με έναν φυσικό και μη επεμβατικό τρόπο.

Προκειμένου ένα έξυπνος ρομποτικός βοηθός να επιτύχει αυτούς τους στόχους, είναι απαραίτητη η δημιουργία και ανάπτυξη εξελιγμένων συστημάτων αλληλεπίδρασης ανθρώπου-ρομπότ (HRI) και η ενσωμάτωσή τους κάτω από ένα ενιαίο αυτόματο σύστημα. Η ρομποτική πλατφόρμα πρέπει να είναι σε θέση να αναγνωρίζει και να καταλαβαίνει τις ανθρώπινες δραστηριότητες, τις προθέσεις και τις ανάγκες, όπως επίσης και να μπορεί να αναλύει πολυαισθητηριακά σήματα βάδισης και στάσης του σώματος, ώστε να παρέχει υποστήριξη και επικοινωνία.

Το διαλογικό σύστημα που παρουσιάζεται σε αυτή την ενότητα είναι μέρος μιας ρομποτικής πλατφόρμας που ονομάζεται i-Walk[4] (Σχήμα 3.13). Η πλατφόρμα αυτή έχει σχεδιαστεί προσεκτικά ώστε να καλύπτει τις ανάγκες αυτές, συνδυάζοντας πολυ-αισθητηριακά σήματα ώστε να πραγματοποιεί ταυτόχρονα πολλές εργασίες που αφορούν την κατανόηση της ανθρώπινης συμπεριφοράς, δηλαδή αναγνώριση λεκτικής πρόθεσης, γενικευμένη αναγνώριση δράσεων και κινησιολογική ανάλυση. Το πολυτροπικό πλαίσιο αλληλεπίδρασης i-Walk έχει ως στόχο να προσφέρει φυσική επικοινωνία και ανατροφοδότηση στους χρήστες αλλά και στο έμπειρο ιατρικό προσωπικό, σχετικά με την πορεία της αποκατάστασης, με τρόπο που να παραπέμπει σε ένα προσωπικό φροντιστή.

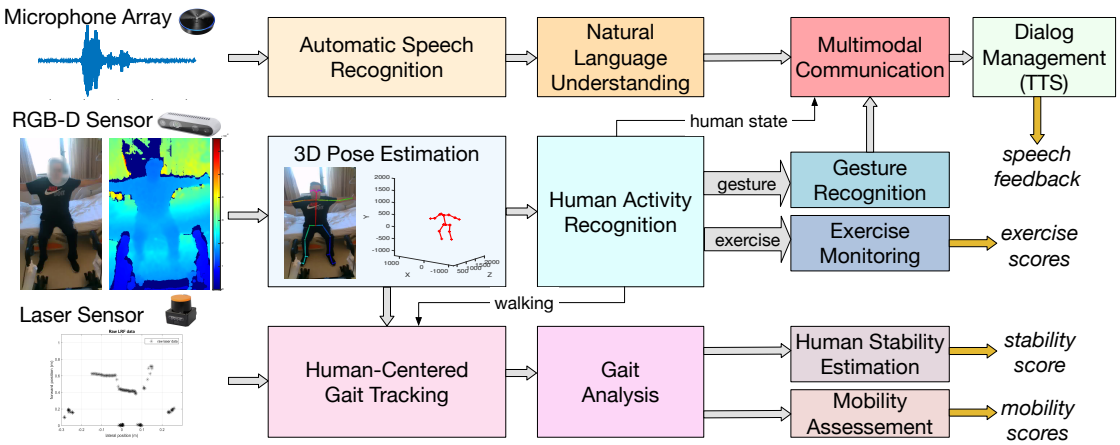
Η πλατφόρμα i-Walk είναι μια πολυ-αισθητηριακή, πολυτροπική πλατφόρμα αλληλεπίδρασης ανθρώπου-ρομπότ, η οποία έχει ως στόχο να προσδώσει ευφυία σε απλούς περιπατητήρες(rollators) και να είναι χρήσιμη σε ασθενείς με κινητικά και/ή νοητικά προβλήματα. Το σύστημα έχει τρεις βασικούς στόχους: να επιβλέπει τον ασθενή ως προς τις δράσεις του, την κινητικότητα του, να αλληλεπιδρά μαζί του επιτρέποντας στον ασθενή να επικοινωνεί τις προθέσεις του λεκτικά, να κάνει διάλογο με το σύστημα αλλά και να παίρνει ανατροφοδότηση από αυτό και τέλος να εκτιμά την κατάσταση του ασθενή στα πλαίσια μιας διαδικασίας αποκατάστασης.

Όλα τα υποσυστήματα (κατανόηση λόγου, αναγνώριση δράσεων, ανάλυση κινητικότητας) αλληλεπιδρούν μεταξύ τους αλλά δρουν και αυτόνομα και παρέχουν ανατροφοδότηση στο χρήστη. Εδώ παρουσιάζεται συγκεκριμένα το διαλογικό σύστημα/σύστημα κατανόησης του λόγου που αναπτύχθηκε για την πλατφόρμα i-Walk στα πλαίσια της παρούσας διδακτορικής διατριβής. Το προτεινόμενο σύστημα έχει αξιολογηθεί χρησιμοποιώντας πολυ-ασισθητηριακά δεδομένα τόσο από υγιείς χρήστες όσο και από πραγματικούς ασθενείς από ένα κέντρο αποκατάστασης οι οποίοι συμμετείχαν στα πειραματικά σενάρια με προβλήματα της καθημερινότητάς τους.

Η συνολική πλατφόρμα i-Walk με τα διάφορα υποσυστήματα, απεικονίζεται στο Σχήμα 3.14. Η ροή που εμφανίζεται πάνω πάνω με το πορτοκαλί χρώμα αντιπροσωπεύει το υποσύστημα της κατανόησης λόγου (*Speech Understanding*), η μεσαία ροή με το γαλάζιο χρώμα την αναγνώριση δράσεων (*Activity Recognition*) ενώ η πιο κάτω με το ανοιχτό ροζ χρώμα την

[4]Υλοποιήθηκε στα πλαίσια του προγράμματος i-Walk (Intelligent Robotic Walker for mobility and cognitive assistance of elderly and motor-impaired people) https://robotics.ntua.gr/i-walk/

Σχήμα 3.13: Αριστερά: Ένας ασθενής που περπατά υποστηριζόμενος από την πλατφόρμα i-Walk. Δεξιά: Ένας ασθενής πραγματοποιεί ασκήσης αποκατάστασης κάτω από την επίβλεψη του συστήματος i-Walk.



Σχήμα 3.14: Επισκόπηση του πολυτροπικού έξυπνου βοηθού i-Walk .

κινησιολογική ανάλυση (*Mobility Analysis*). Κάθε υποσύστημα παράγει κάποιες εξόδους για τις οποίες είναι υπεύθυνο το μπλοκ με το κόκκινο χρώμα, της πολυτροπικής επικοινωνίας (*Multimodal Communication*) η οποία ενεργοποιεί το διαλογικό σύστημα (ανοιχτό πράσινο), παρέχοντας λεκτική ανατροφοδότηση στους χρήστες.

Ένα υποσύστημα κατανόησης λόγου είναι απαραίτητο σε μία τέτοια ρομποτική πλατφόρ-

μα, διότι ο λόγος είναι η πιο φυσική και άμεση μορφή επικοινωνίας. Ο ασθενής πρέπει να μπορεί να επικοινωνεί τις προθέσεις του στο ρομπότ. Το παρόν σύστημα που αναπτύχθηκε περιέχει δύο μέρη, όπως φαίνεται και στο Σχήμα 3.14: Μία μονάδα αυτόματης αναγνώρισης φωνής (ASR) και μία μονάδα κατανόησης φυσικής γλώσσας (Natural Language Understanding (NLU)). Όσον αφορά την πρώτη μονάδα, χρησιμοποιείται ένα state-of-the-art σύστημα αναγνώρισης φωνής, το Google speech-to-text API [238], το οποίο παίρνει ως είσοδο τα σήματα που καταγράφονται απο μία συστοιχία μικροφώνων και δίνει ως έξοδο τη μεταγραφή τους σε κείμενο. Στη συνέχεια, το κείμενο αυτό αποτελεί την είσοδο στη μονάδα του NLU προκειμένου να 'μεταφραστεί' σε κάποια πρόθεση του χρήστη. Το σύστημα NLU δημιουργήθηκε και αναπτύχθηκε με το framework RASA [239, 240, 241]: Σχεδιάστηκε ένα σύνολο προκαθορισμένων προθέσεων, τόσο γενικού σκοπού όσο και ειδικού στην παρούσα εφαρμογή. Η κατηγορία γενικού σκοπού περιέχει 7 γενικές προθέσεις, δηλαδή 'χαιρετάω', 'λέω το όνομά μου', 'λέω αντίο', 'ευχαριστώ', 'ναι', 'όχι', 'ζητώ να επαναλάβεις', ενώ η ειδικού σκοπού άλλες 7 προθέσεις, δηλαδή 'θέλω να σηκωθώ', 'θέλω να καθίσω', 'θέλω να περπατήσω', 'σταμάτα', 'τέλος', 'θέλω να πάω στο μπάνιο', 'θέλω να κάνω ασκήσεις'.

Κάθε πρόθεση είναι συνδεδεμένη με μια πληθώρα παραδειγμάτων φράσεων έκφρασης της συγκεκριμένης πρόθεσης. Για παράδειγμα, ο χρήστης μπορεί να εκφράσει την πρόθεσή του να σηκωθεί λέγοντας 'Θέλω να σηκωθώ' ή 'Μπορείς να με βοηθήσεις να σηκωθώ' και οποιαδήποτε άλλη παραλλαγή.

Ένα NLU pipeline του RASA το οποίο ονομάζεται tensorflow embeddings χρησιμοποιείται ώστε να προβλέψει την παρούσα πρόθεση βασισμένη στο αποτέλεσμα της αναγνώρισης φωνής. Η πρόβλεψη της πρόθεσης στη συνέχεια αποτελεί την είσοδο στη μονάδα της πολυτροπικής επικοινωνίας.

Βάση δεδομένων και πλαίσιο αξιολόγησης

Για την εκπαίδευση και την αξιολόγηση του συστήματος κατανόησης λόγου αλλά και των υπόλοιπων τμημάτων του συστήματος, πραγματοποιήθηκε εκτενής συλλογή δεδομένων τόσο από υγιείς συμμετέχοντες, όσο και από πραγματικούς ασθενείς με διάφορα κινητικά και/ή νοητικά προβλήματα. Η πειραματική πλατφόρμα του Σχήματος 3.13 ήταν εξοπλισμένη με μία κάμερα RealSense, ένα λέιζερ Hokuyo UST-LX10 και μία συστοιχία μικροφώνων eMeet, τα οποία παρείχαν τα πολυ-αισθητηριακά δεδομένα. Η επικοινωνία των πολλαπλών αισθητήρων πραγματοποιήθηκε μέσω ROS [231]. Η πλατφόρμα ήταν επίσης εξομπλισμένη με δύο μονάδες NVIDIA Jetson TX2 που προσδίδουν δυνατότητα για βαθιά εκμάθηση πάνω στον περιπατητήρα. Η βάση που συλλέχθηκε περιγράφεται με συντομία παρακάτω:

i-Walk DB: Η βάση αυτή περιέχει δεδομένα τόσο από υγιείς συμμετέχοντες όσο και από ασθενείς. Η συλλογή δεδομένων με ασθενείς πραγματοποιήθηκε στο κέντρο αποκατάστασης Διάπλαση [242] στην Καλαμάτα. Στα πειράματα συμμετείχαν 13 ασθενείς, κατόπιν έγκρισης από το ιατρικό προσωπικό. Τα δημογραφικά τους δεδομένα (ηλικία και φύλο) μαζί με τους δείκτες Mini-Mental Mean Score (MMSE) και POMA οι οποίοι μετρούν τη νοητική ικανότητα και την κινητική κατάσταση αντίστοιχα αναγράφονται στον Πίνακα 3.9. Όλοι οι ασθενείς

Patients	Age	Gender	MMSE	POMA	Speech Intent Recognition (%)
1	80	M	29/30	18/28	80.00
2	86	M	27/30	18/28	77.77
3	25	F	29/30	18/28	100.00
4	84	F	17/30	11/28	78.57
5	85	F	19/30	14/28	62.50
6	50	M	29/30	13/28	93.33
7	78	M	27/30	15/28	54.54
8	73	M	18/30	12/28	86.67
9	72	F	19/30	11/28	46.15
10	75	M	25/30	16/28	56.25
11	83	F	23/30	11/28	84.62
12	55	F	28/30	16/26	100.00
13	75	M	28/30	11/28	46.15
Average	-	-	-	-	74.25

Πίνακας 3.9: Δημογραφικές πληροφορίες και σκορ αξιολόγησης για τους 13 ασθενείς της βάσης i-Walk.

υπέγραψαν ότι δίνουν τη συγκατάθεσή τους και προστατεύονται από το GDPR. Για την ανάπτυξη των μοντέλων δειξάχθηκε επιπλέον συλλογή δεδομένων με 20 υγιείς συμμετέχοντες με ηλικίες 23-32. Η συλλογή δεδομένων πραγματοποιήθηκε στα πλαίσια δύο σεναρίων σχεδιασμένων σε συνεργασία με τους γιατρούς, ώστε να ανταποκρίνονται στις πραγματικές ανάγκες αποκατάστασης των ασθενών και να είναι μέσα στην καθημερινότητά τους.

Σενάριο 1: Ασκήσεις αποκατάστασης. Οι συμμετέχοντες είναι καθιστοί στο κρεβάτι ή στην καρέκλα και ζητούν να κάνουν κάποιες ασκήσεις οι οποίες είναι μέρος του προγράμματος αποκατάστασης. Παραδείγματα τέτοιων ασκήσεων είναι το να σηκώσουν τα χέρια, να στρέψουν τον κορμό, να σηκωθούν σε όρθια θέση και διάφορα άλλα. Επιπλέον τους ζητήθηκε να εκφράζουν τις προθέσεις τους μέσω φωνής ελεύθερα, χωρίς να τους υποδείξουμε συγκεκριμένες φράσεις. Οι προθέσεις αναφέρθηκαν νωρίτερα στην παρούσα ενότητα.

Σενάριο 2: Μεταφορά στο μπάνιο. Το σενάριο αυτό είναι απαραίτητο στους ασθενείς με κινητικά προβλήματα γιατί είναι μία δύσκολη εργασία, η οποία μπορεί να βοηθηθεί ουσιαστικά από έναν ρομποτικό περιπατητήρα. Οι χρήστες είναι καθιστοί σε κρεβάτι/καρέκλα και εκφράζουν την πρόθεση να σηκωθούν και να πάνε στο μπάνιο. Το περπάτημα περιλαμβάνει την πλοήγηση μέσα στο δωμάτιο, είσοδο στο μπάνιο, κάθισμα, σήκωμα, επιστροφή στο κρεβάτι/καρέκλα.

Αποτελέσματα αξιολόγησης

Για την εξετάσουμε την επίδοση του συστήματος κατανόησης λόγου, αξιολογούμε το ποσοστό των σωστά αναγνωρισμένων προθέσεων από τη φωνή. Για το σκοπό αυτό χρησιμποιούμε τη βάση που περιγράφηκε προηγουμένως, η οποία περιέχει ελεύθερο λόγο, όπως αυτός εκφράστηκε από τους υγιείς και ασθενείς συμμετέχοντες. Για τους πρώτους, η βάση περιέχει 445 εκφωνήσεις και η ακρίβεια που επιτυγχάνεται είναι 94.83%, που είναι αρκετά υψηλή. Για

τους ασθενείς, τα αποτελέσματα αναφέρονται στον Πίνακα 3.9, για κάθε ασθενή ξεχωριστά αλλά και για το μέσο όρο. Η συνολική επίδοση για 173 εκφωνήσεις φτάνει το 74.25% που είναι μεν χαμηλότερο από τους υγιείς αλλά αυτό μπορεί να αποδοθεί σε δύο παράγοντες: Πρώτον, κάποιοι ασθενείς δεν είχαν μόνο κινητικά αλλά και νοητικά προβλήματα, όπως φαίνεται και στο Πίνακα 3.9, επομένως ο λόγος τους μπορεί να ήταν πολλές φορές ακατάληπτος. Δεύτερον, επειδή η ηλικία των ασθενών είναι μεγαλύτερη από των υγιών, οι εκφωνήσεις τους για τις ίδιες προθέσεις είναι διαφορετικές από των υγιών και έχουν μεγαλύτερη ποικιλία. Συνολικά η επίδοση είναι ικανοποιητική, αλλά μπορεί να βελτιωθεί συλλέγοντας περισσότερα δεδομένα από περισσότερους ασθενείς.

Κεφάλαιο 4

Βάση 'Αθηνά': Πολυκαναλικά φωνητικά δεδομένα στα Ελληνικά

Για να είναι εφικτή η κατασκευή υποδομών και αλγορίθμων φωνητικής αλληλεπίδρασης ενός χρήστη με ένα περιβάλλον έξυπνου σπιτιού, είναι απαραίτητη η ύπαρξη κατάλληλων βάσεων δεδομένων φωνής από το εσωτερικό τέτοιων έξυπνων χώρων, προκειμένου να δημιουργηθούν τα μοντέλα φωνής ή να προσαρμοστούν στο περιβάλλον, καθώς και για να αξιολογηθούν οι αλγόριθμοι. Παρόλο που υπάρχει η δυνατότητα για κατασκευή τέτοιων δεδομένων με προσομειώσεις, αυτά δε μπορούν να υποκαταστήσουν την ανάγκη για πραγματικά δεδομένα διότι παρουσιάζουν πολλούς περιορισμούς. Σε αυτή την ενότητα θα παρουσιαστεί η βάση πολυκαναλικών δεδομένων ομιλίας που συλλέχθηκε στα Ελληνικά σε έξυπνα περιβάλλοντα, η βάση 'Αθηνά'. Περιγράφονται τόσο η διαδικασία που ακολουθήθηκε πριν, κατά τη διάρκεια και μετά τη συλλογή δεδομένων καθώς και τα περιεχόμενα της βάσης.

Ενώ σε άλλες γλώσσες και κυρίως στα Αγγλικά διατίθενται κάποιες βάσεις πολυκαναλικών φωνητικών δεδομένων, στα Ελληνικά δεν είχε αναπτυχθεί μέχρι σήμερα μια τέτοια βάση. Όσον αφορά τα Αγγλικά, μια γνωστή βάση πραγματικού λόγου σε ένα έξυπνο γραφείο όπου πραγματοποιούνται συναντήσεις παρουσιάζεται στο [25]. Εκεί η συστοιχία μικροφώνων είναι τοποθετημένη στο τραπέζι και υπάρχει μεγάλη επικάλυψη ανάμεσα στα σήματα φωνής των διαφόρων ομιλητών. Μία άλλη βάση δεδομένων που περιέχει τόσο μεμονωμένα όσο και επικαλυπτόμενα τμήματα ομιλίας παρουσιάζεται στο [243].

Για έξυπνους χώρους με πολλά δωμάτια, έχει συλλεχθεί μία πολυτροπική βάση με διάφορα ακουστικά γεγονότα σε ένα έξυπνο σπίτι [244], ενώ η βάση προσομειωμένων δεδομένων DIRHA που παρουσιάστηκε σε προηγούμενη ενότητα περιέχει προσομειωμένα δεδομένα σε 4 διαφορετικές γλώσσες από 4 διαφορετικά δωμάτια [24]. Επίσης, μία βάση στα Γαλλικά [29] ηχογραφήθηκε σε ένα έξυπνο σπίτι 5 δωματίων όπου ο λόγος επικαλύπτεται τόσο με θόρυβο όσο και με άλλα ακουστικά γεγονότα αλλά περιέχει μη φυσικό λόγο αφού ηχογραφήθηκαν μόνο εντολές αυτοματισμού η μία μετά την άλλη.

Παρακάτω περιγράφεται η βάση 'Αθηνά', μία βάση πραγματικών δεδομένων ομιλίας στα Ελληνικά που ηχογραφήθηκε στο έξυπνο γραφείο Athena-RC δύο δωματίων. Φράσεις ενεργοποίησης, εντολές αυτοματισμού καθώς και φωνητικά πλούσιες προτάσεις εκφωνήθηκαν από

20 ομιλητές από διάφορες θέσεις στο χώρο. Επίσης, κάποιες συνεδρίες περιλάμβαναν αυθόρμητο διάλογο μεταξύ δύο ομιλητών. Τα περισσότερα τμήματα ομιλίας επικαλύπτονται με άλλα ακουστικά γεγονότα και θόρυβο, επομένως οι συνθήκες είναι αρκετά ρεαλιστικές αλλά και δύσκολες. Καταγράφηκαν επίσης οπτικά δεδομένα, τόσο για εντοπισμό ομιλητή όσο και για καταγραφή χειρονομιών. Η βάση αυτή είναι διαθέσιμη δημόσια[1].

4.1 Χώρος ηχογράφησης και εξοπλισμός

Το περιβάλλον έξυπνου γραφείου Athena-RC περιλαμβάνει δύο δωμάτια: Το κύριο δωμάτιο είναι ένα γραφείο και το δευτερεύον είναι μέρος του διαδρόμου. Η κάτοψη του χώρου απεικονίζεται στο Σχήμα 4.1, ενώ στο Σχήμα 4.2 απεικονίζεται μια εικόνα του χώρου κατά τη διάρκεια των ηχογραφήσεων.

4.1.1 Εξοπλισμός και συγχρονισμός

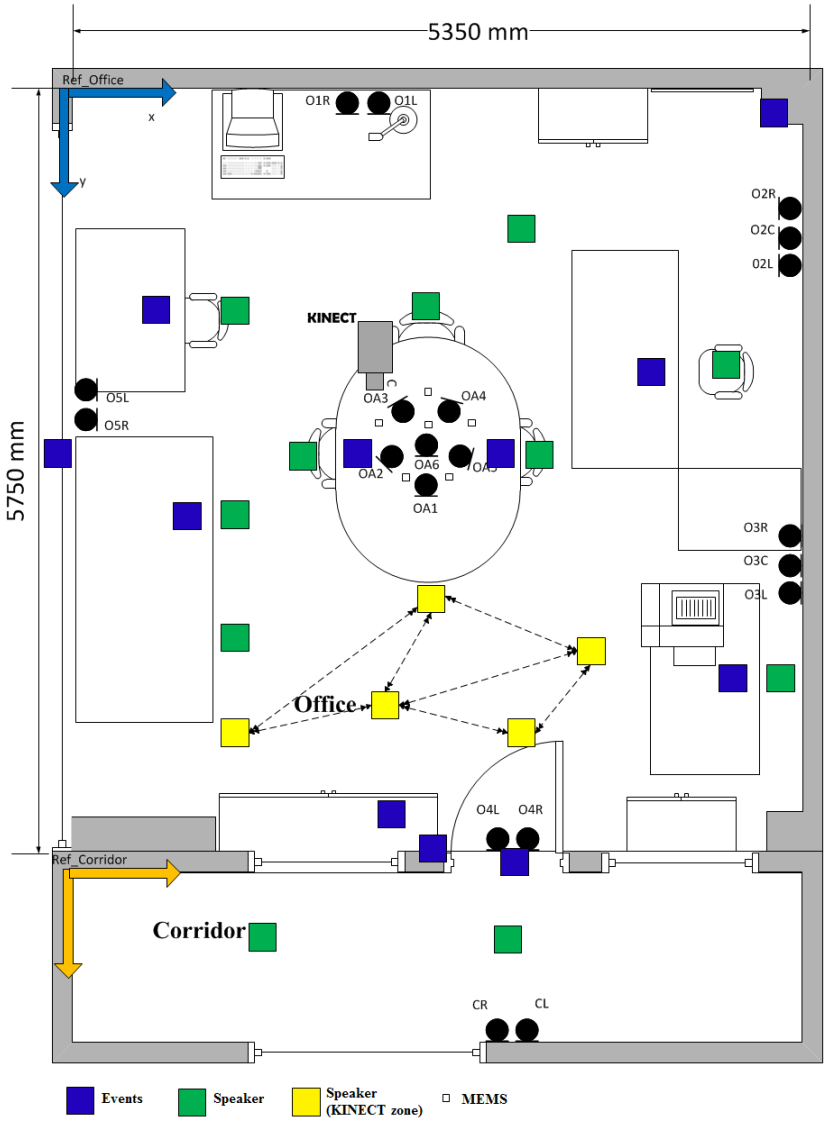
Όπως φαίνεται και στο Σχήμα 4.1, χρησιμοποιήθηκαν συνολικά 20 πυκνωτικά μικρόφωνα (Shure MX391/O, omnidirectional, 48kHz rate) τα οποία κατανεμήθηκαν στους τοίχους και την οροφή. Συγκεκριμένα 6 μικρόφωνα τοποθετήθηκαν στην οροφή σε πενταγωνική γεωμετρία με ένα μικρόφωνο στο κέντρο και τα υπόλοιπα τοποθετήθηκαν στους τοίχους ανά δυάδες ή τριάδες. Επίσης, 2 close-talk (Sennheiser ew172G3, 48kHz rate) μικρόφωνα κατέγραφαν το πραγματικό καθαρό σήμα φωνής που εκφωνούσαν οι ομιλητές. Όλα αυτά τα μικρόφωνα τοποθετήθηκαν σε 3 προενισχυτές των 8 καναλιών (Focusrite OctoPre MkII). Εκτός από τα πυκνωτικά μικρόφωνα, τοποθετήθηκαν και 6 MEMS (48kHz) μικρόφωνα στην οροφή σε παρόμοια γεωμετρία με τα πυκνωτικά. Για την καταγραφή οπτικών δεδομένων χρησιμοποιήθηκε μία κάμερα Kinect (10fps), η οποία κατέγραφε τρεις ροές δεδομένων, τα RGB, Depth, Skeleton. Τέλος, χρησιμοποιήθηκε και ένα ηχείο (Genelec 8030BPM) ώστε να μετρηθούν οι κρουστικές αποκρίσεις του χώρου.

Ο συγχρονισμός των 20 πυκνωτικών μικροφώνων με τα 2 close-talk επιτεύχθηκε μέσω της σύνδεσής τους στους συγχρονισμένους προενισχυτές ενώ η καταγραφή των MEMS και των οπτικών δεδομένων της Kinect πραγματοποιήθηκε μέσω του Robot Operating System (ROS) [231] ώστε να είναι συγχρονισμένα.

4.1.2 Δωμάτιο και θέσεις ομιλητή

Συνολικά χρησιμοποιήθηκαν 16 διαφορετικές θέσεις ομιλητή για την εκφώνηση των φράσεων, με διάφορους προσανατολισμούς. Όπως φαίνεται και από το Σχήμα 4.1, οι θέσεις αυτές χωρίζονται σε 2 διαφορετικές ζώνες, την A (κίτρινα τετράγωνα) και τη B (πράσινα τετράγωνα), που αφορά το πεδίο θέασης της κάμερας Kinect, η οποία βλέπει τη Ζώνη A αλλά όχι τη B. Για θέσεις κοντά σε γραφείο ή τραπέζι είχαμε δύο καταστάσεις: μία για όρθιο και μία για καθιστό ομιλητή. Όλες οι θέσεις ήταν σημειωμένες με ταινία στο πάτωμα ώστε να γνωρίζουμε τις πραγματικές θέσεις ομιλητή.

[1] http://cvsp.cs.ntua.gr/research/athenadb



Σχήμα 4.1: Κάτοψη του έξυπνου χώρου 'Αθηνά'. Απεικονίζονται οι αισθητήρες, οι θέσεις ομιλητή και ακουστικών γεγονότων.

Εκτός από τις θέσεις ομιλητή, συμπεριλήφθηκαν και κάποιες θέσεις (με μπλε χρώμα στο σχήμα), για πραγματοποίηση ακουστικών γεγονότων. Αυτές είναι κυρίως θέσεις πάνω σε τραπέζια, πάγκους εργασίας ή κοντά σε πόρτες και παράθυρα, ώστε να μοιάζει όσο το δυνατό περισσότερο με ένα πραγματικό έξυπνο σπίτι και τις καθημερινές δραστηριότητες που συμβαίνουν.

4.2 Περιγραφή βάσης

Η βάση αυτή περιέχει 20 ομιλητές, 10 άντρες και 10 γυναίκες. Καθένας από αυτούς συμμετείχε σε 12 συνεδρίες του ενός λεπτού. Επομένως συνολικά η βάση περιέχει ηχογραφήσεις διάρκειας 240 λεπτών. Η βάση περιέχει τόσο ακουστικά όσο και οπτικά δεδομένα, καθώς και

Audio type	#	%
Speech		
Conversation	100	20.9
Phonetically rich	140	4.0
System activation	240	2.3
Commands	240	3.6
Total	960	30.8
Long Events		
Walking steps	112	3.6
Cellphone ring	104	3.0
Keyboard	124	4.3
Glass fill	108	2.0
Coffee spoon	116	2.9
Skype call	100	3.2
Cough	96	1.9
Paper work	100	2.8
Window open/close	96	2.4
Total	956	26.1

Audio type	#	%
Short Events		
Mouse click	152	1.9
Keys	176	3.8
Knock	148	1.2
Chair moving	168	2.2
Switch on/off	156	1.2
Door open/close	144	2.1
Total	944	12.4

Background Noises		
Ambient noise	46	20.0
Fan	54	22.0
Radio music	70	29.0
Vacuum cleaner	47	20.0
Silence	21	9.0
Total	238	100

Πίνακας 4.1: Τα ηχητικά δεδομένα που συλλέχθηκαν, κατηγοριοποιημένα σε ομιλία, ακουστικά γεγονότα και θορύβους. Ο αριθμός εμφανίσεών τους (#) και η συνολική τους διάρκεια σαν ποσοστό επί των 240 λεπτών της βάσης. Τα ακουστικά γεγονότα χωρίζονται σε μικρά και μεγάλα με μέση διάρκεια 1.5sec και 3sec αντίστοιχα.

τις εκτιμήσεις των κρουστικών αποκρίσεων του χώρου. Όσον αφορά τα ηχητικά δεδομένα, τα σενάρια καταγραφής περιλάμβαναν διάφορους τύπους λόγου, ακουστικών γεγονότων και θορύβων, των οποίων τα χαρακτηριστικά παρουσιάζονται στον Πίνακα 4.1. Πρέπει να σημειωθεί ότι το 40% των συνολικών δεδομένων φωνής επικαλύπτεται με άλλα ηχητικά γεγονότα. Τα οπτικά δεδομένα περιέχουν χειρονομίες σχετικές με την ενεργοποίηση του συστήματος καθώς και τις συντεταγμένες του ομιλητή που κινείται. Η βάση έχει χωριστεί σε ένα development και ένα test σετ που περιλαμβάνουν 10 ομιλητές το καθένα.

4.2.1 Φωνητικά δεδομένα

Οι διάφοροι τύποι λόγου που χρησιμοποιήθηκαν είναι φράσεις ενεργοποίησης (λέξεις-κλειδιά), φωνητικές εντολές, φωνητικά πλούσιες προτάσεις και διάλογος μεταξύ δύο ομιλητών. Όσον αφορά τους 3 πρώτους τύπους, οι διαθέσιμες διαφορετικές φράσεις ήταν αντίστοιχα 12, 170 και 190. Για το διάλογο, χρησιμοποιήθηκαν κάποια θέματα συζήτησης ώστε να παρακινηθεί ο ομιλητής να ξεκινήσει έναν αυθόρμητο διάλογο με το συνομιλητή του, όπως κινηματογράφος, μουσική, επιστήμη, κτλ. Οι φωνητικά πλούσιες προτάσεις είναι αποσπάσματα από μια Ελληνική βάση ομιλίας μεγάλου λεξιλογίου που ονομάζεται Λογοτυπογραφία [33]. Οι φράσεις ενεργοποίησης είναι μικρές φράσεις που περιλαμβάνουν τη λέξη-κλειδί Ντίρα (DIRHA [211]) ενώ οι εντολές αφορούν αυτοματισμούς του σπιτιού. Ο Πίνακας 4.2 περιγράφει τις διαφορετικές συνεδρίες που καταγράφηκαν.

Type	Session	Sequence	Position Zone
Static Single Spk	1	S1: PH - DK - DC	B
	2	S1: PH - DK - DC	B
	3	S1: PH - DK - DC	B
	4	S1: PH - DK - DC	A
	5	S1: PH - DK - DC	A
Static 2 Spks	6	S1: CS - DK - DC & S2: CS	B
	7	S1: CS - DK - DC & S2: CS	B
	8	S1: CS - DK - DC & S2: CS	B
Moving Single Spk	9	S1: PH - DK - DC	A
	10	S1: PH - DK - DC	A
Moving 2 Spks	11	S1: CS - DK - DC & S2: CS	A
	12	S1: CS - DK - DC & S2: CS	A

Πίνακας 4.2: Οι ακολουθίες λόγου ανά συνεδρία. DK είναι η φράση ενεργοποίησης, DC η εντολή, PH η φωνητικά πλούσια πρόταση ανδ CS ο διάλογος. S1 είναι ο βασικός ομιλητής και S2 ο συνομιλητής όταν υπάρχει διάλογος.

4.2.2 Ακουστικά γεγονότα και θόρυβοι

Προκειμένου να προσομειωθούν όσο το δυνατόν πιο ρεαλιστικά οι συνθήκες και τα γεγονότα που συμβαίνουν σε ένα πραγματικό έξυπνο σπίτι, εισάγαμε στις ηχογραφήσεις κάποια ακουστικά γεγονότα και θορύβους τα οποία πραγματοποιούσαν άνθρωποι. Ο Πίνακας 4.1 παρουσιάζει τα διάφορα γεγονότα, κατηγοριοποιημένα ανά διάρκεια σε μικρά και μεγάλα. Σε κάθε συνεδρία υπήρχε ένα μικρό και ένα μεγάλο με 4 επαναλήψεις το καθένα, τυχαία κατανεμημένες μέσα στο ένα λεπτό. Οπότε υπάρχουν τόσο επικαλυπτόμενα με φωνή γεγονότα όσο και μεμονωμένα.

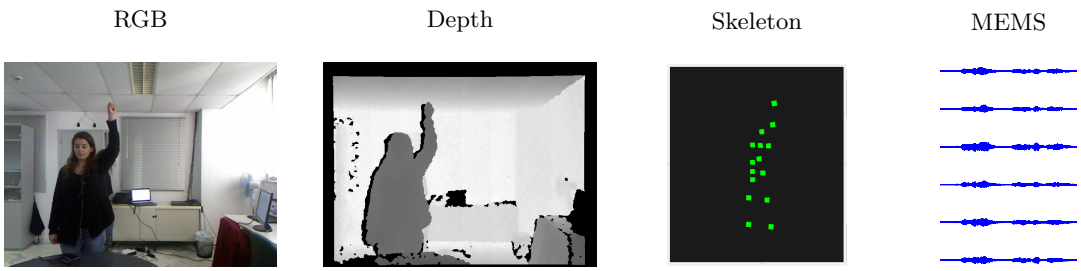
Όσον αφορά τους θορύβους, χρησιμοποιήθηκαν 5 διαφορετικά είδη: α) σιωπή, β) ραδιόφωνο που έπαιζε από υπολογιστή, γ) ανεμιστήρας, δ) διάχυτος θόρυβος από ανοιχτό παράθυρο και ε) ηλεκτρική σκούπα στο διάδρομο. Οι θόρυβοι διαρκούσαν για το σύνολο της συνεδρίας, δηλαδή για ολόκληρο το ένα λεπτό.

4.2.3 Χειρονομία

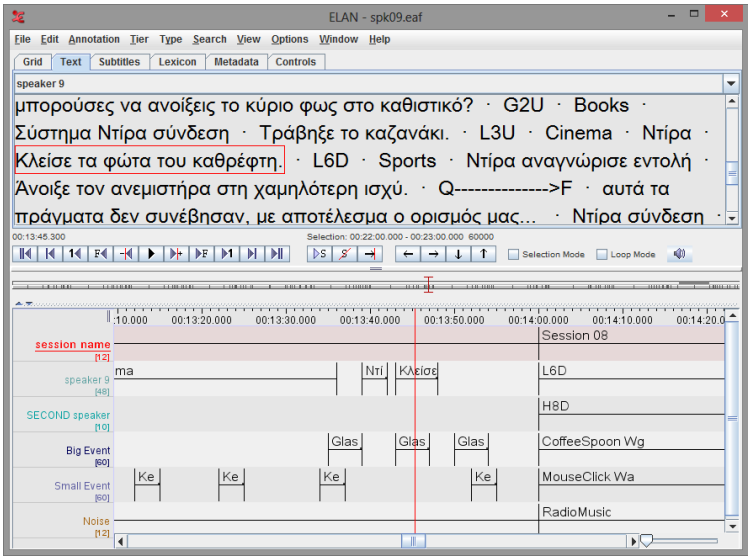
Για να μελετηθεί η πολυτροπική επεξεργασία και αλληλεπίδραση καθώς και για να βελτιωθεί η ενεργοποίηση του συστήματος, χρησιμοποιήθηκε μια χειρονομία ταυτόχρονα με την εκφώνηση μιας φράσης ενεργοποίησης από τον ομιλητή, στις συνεδρίες όπου κατέγραφε η κάμερα Kinect. Η χειρονομία είναι ένα σηκωμένο χέρι σε γροθιά, ώστε η κάμερα Kinect να μπορεί να καταγράφει τη χειρονομία ανεξάρτητα από τον προσανατολισμό του ομιλητή. Ένα στιγμιότυπο φαίνεται στο Σχήμα 4.2, όπου απεικονίζονται οι τρεις ροές πληροφορίας της κάμερας και οι έξοδοι των μικροφώνων MEMS για τη συγκεκριμένη φράση ενεργοποίησης.

4.2.4 Εκτίμηση κρουστικών αποκρίσεων

Εκτός από τη συλλογή των πραγματικών δεδομένων, μετρήθηκαν και οι κρουστικές αποκρίσεις IRs του δωματίου, από κάθε θέση ομιλίας και προσανατολισμό προς όλα τα μικρόφωνα. Οι



Σχήμα 4.2: RGB, Depth, Skeleton, MEMS έξοδοι για την εκφώνηση μιας φράσης ενεργοποίησης μαζί με τη χειρονομία.



Σχήμα 4.3: Εργαλείο επισημείωσης ELAN: Οι διαφορετικές γραμμές αντιπροσωπεύουν λόγο, θέσεις, γεγονότα, κλπ. Το κόκκινο περίγραμμα υποδεικνύει τη φράση προς εκφώνηση.

μετρήσεις βασίστηκαν σε ένα επαγγελματικό ηχείο (Genelec 8030A) ικανό να διεγείρει το περιβάλλον με μακρές ακουλουθίες του σήματος Exponential Sine Sweep (ESS) [236]. Όπως αναφέρεται στο [22], η μέθοδος με ESS εξασφαλίζει μετρήσεις αποκρίσεων με υψηλό SNR και ευρωστία σε αρμονικές παραμορφώσεις.

4.3 Διαδικασία συλλογής δεδομένων και επισημείωση

Για να επιτευχθεί μία δίκαιη κατανομή των θέσεων, φράσεων, γεγονότων και χρονικών ορίων σε όλες τις συνεδρίες, αναπτύχθηκε ένας αλγόριθμος που τα κατένειμε τυχαία, εξασφαλίζοντας ότι όλες οι φράσεις θα εμφανιστούν τουλάχιστον μία φορα. Η διαδικασία καταγραφής πραγματοποιήθηκε με τη βοήθεια του εργαλείου επισημείωσης ELAN [245] ώστε να καθοδηγούνται οι ομιλητές να ακολουθούν το σενάριο κατά τη διάρκεια των καταγραφών. Το εργαλείο υποδείκνυε τη θέση του ομιλητή, τις φράσεις που θα εκφωνηθούν καθώς και τα ακουστικά γεγονότα με τη μορφή μιας μπάρας χρόνου που κινούνταν, όπως φαίνεται στο Σχήμα 4.3.

Η καθοδήγηση του ομιλητή ήταν εφικτή μέσω της χρήσης συγχρονισμένων οθονών κατανεμημένων στους χώρους, που απεικόνιζαν την απαραίτητη πληροφορία. Πριν την αρχή της συνεδρίας, ο ομιλητής λάμβανε την οδηγία να σταθεί στη θέση που έδειχνε το εργαλείο επισημείωσης, κοιτώντας προς την κατεύθυνση που υποδείκνυε ένας αριθμός δίπλα στη θέση.

Μετά τις ηχογραφήσεις ξαναεπισημειώθηκαν τα δεδομένα ώστε να διορθωθούν τα χρονικά όρια και να επισημειωθούν και άλλα εξωτερικά γεγονότα που δε συμπεριλαμβάνονταν στο σενάριο. Τέτοια γεγονότα για παράδειγμα ήταν ομιλία από άλλα γραφεία, πόρτες που άνοιγαν ή έκλειναν, βήματα, κτλ.

Κεφάλαιο 5

Συλλογή δεδομένων οφθαλμοκινητικής παρακολούθησης

5.1 Διαδικασία καταγραφής δεδομένων οφθαλμοκινητικής παρακολούθησης (eye-tracking data)

Όπως αναφέρθηκε εισαγωγικά, το πιο καινοτόμο κομμάτι της παρούσας διδακτορικής διατριβής αφορά στη μοντελοποίηση της οπτικοακουστικής προσοχής μέσω την εμφάνειας, το οποίο θα παρουσιαστεί στα επόμενα δύο κεφάλαια. Στο πρόβλημα αυτό, το μοντέλο προσπαθεί να προβλέψει τα σημεία εστίασης του ανθρώπινου ματιού σε ένα βίντεο. Η επαλήθευση και η αξιολόγηση της εγκυρότητας και καταλληλότητας ενός μοντέλου γίνεται μέσω δεδομένων οφθαλμοκινητικής παρακολούθησης. Αυτά συλλέγονται μέσω ειδικών συσκευών που μπορούν να παρακολουθούν και να καταγράφουν τις κινήσεις του ματιού, όσο ένας άνθρωπος παρακολουθεί ένα οπτικό ερέθισμα σε μια οθόνη. Τα δεδομένα αυτά ουσιαστικά αποτελούνται από τις συντεταγμένες της κόρης των ματιών πάνω στην οθόνη, με τη συχνότητα καταγραφής της κάθε συσκευής. Στη βιβλιογραφία υπάρχουν αρκετές βάσεις με οπτικά δεδομένα οφθαλμοκινητικής παρακολούθησης, δηλαδή οι συμμετέχοντες παρακολουθούσαν τα βίντεο με οπτική πληροφορία μόνο και όχι ήχο. Στην παρούσα διδακτορική διατριβή μελετήθηκε η μοντελοποίηση της οπτικοακουστικής ανθρώπινης προσοχής, για την αξιολόγηση της οποίας χρειάζονται οπτικοακουστικά δεδομένα οφθαλμοκινητικής παρακολούθησης.

Για τους σκοπούς της πειραματικής αξιολόγησης, λόγω της ύπαρξης ελάχιστων βάσεων με οπτικοακουστικά δεδομένα οφθαλμοκινητικής παρακολούθησης, αποφασίσαμε να συλλέξουμε τέτοια δεδομένα για δύο βάσεις δεδομένων βίντεο, τη SumMe [246] και την ETMD [148]. Η πρώτη περιέχει 25 ανεπεξέργαστα βίντεο από χρήστες, ενώ η δεύτερη περιέχει 12 βίντεο-αποσπάσματα από έξι διαφορετικές ταινίες του Hollywood. Συνολικά συλλέχθηκαν δεδομένα από 37 βίντεο με διάρκεια 2 ώρες και περίπου $171,000$ καρέ. Τα διαφορετικά βίντεο κυμαίνονταν από 38 εως 388 δευτερόλεπτα σε διάρκεια και μετατράπηκαν από την πρωτότυπη μορφοποίηση

τους σε 30 frames per second (FPS) MOV format.

Για αυτό το σκοπό, το σύνολο των συμμετεχόντων χωρίστηκε σε δύο υποσύνολα που περιείχαν τον ίδιο αριθμό ατόμων και βίντεο. Επομένως, για κάθε βίντεο συλλέχθηκαν δεδομένα από 10 διαφορετικούς συμμετέχοντες, οι οποίοι ήταν εθελοντές από το Εθνικό Μετσόβιο Πολυτεχνείο, με ηλικίες από 23 εως 55 χρονών (με μέση ηλικία τα 35 χρόνια). Οι περισσότεροι από τους συμμετέχοντες δε γνώριζαν το σκοπό του πειράματος και όλοι είχαν κανονική όραση.

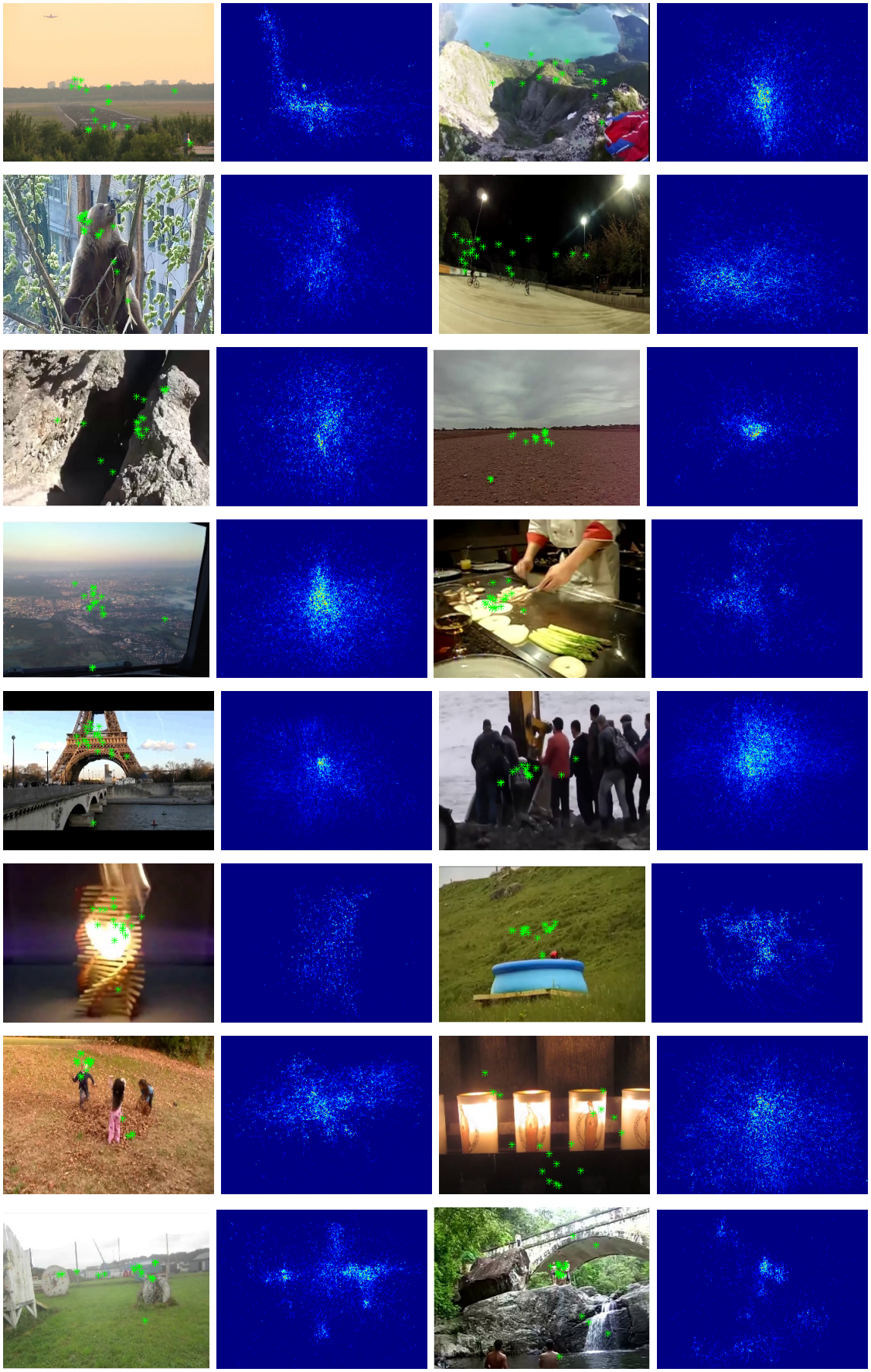
Οι κινήσεις των ματιών και από τις δύο κόρες καταγράφηκαν από έναν SR Research Eyelink 2000 desktop mounted eye tracker με 1,000 Hz συχνότητα δειγματοληψίας για κάθε μάτι. Τα βίντεο προβλήθηκαν σε μία οθόνη διάστασης 1600x900 σε απόσταση 90 εκατοστά από το θεατή. Ο ήχος ήταν στερεοφωνικός και οι χρήστες είχαν τη δυνατότητα να ακούν μέσω ακουστικών. Χρησιμοποιήθηκε επίσης μια ειδική συσκευή όπου οι συμμετέχοντες μπορούσαν να ακουμπούν το πηγούνι και το κεφάλι τους ώστε να περιορίσουν τις κινήσεις του κεφαλιού τους κατά τη διάρκεια των καταγραφών. Η παρουσίαση ελεγχόταν από τον υπέυθυνο του πειράματος μέσω του λογισμικού SR Research Experiment Builder.

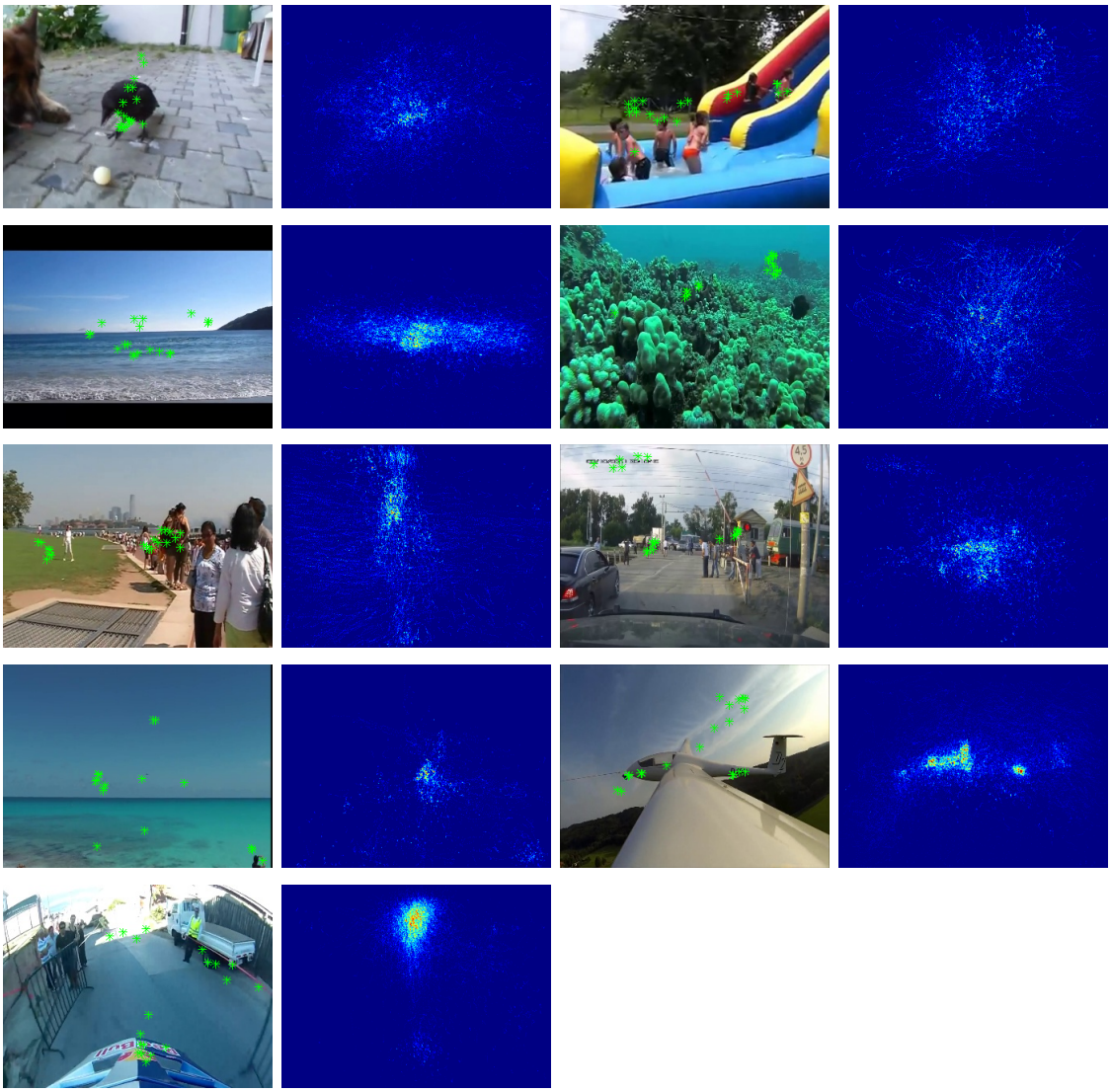
Στα άτομα που συμμετείχαν στη συλλογή δεδομένων δε δόθηκαν κάποιες ειδικές οδηγίες: Πληροφορήθηκαν ότι θα έβλεπαν κάποια βίντεο στα οποία θα μπορούσαν να κοιτούν ελεύθερα όπου ήθελαν και να είναι προσεχτικοί να μην κουνάνε πολύ το κεφάλι τους κατά τη διάρκεια ενός βίντεο ώστε να μη χαλάσει η βαθμονόμηση που προηγείται ενός βίντεο. Η σειρά των βίντεο ήταν τυχαία για κάθε χρήστη και η όλη διαδικασία διαρκούσε περίπου 90 λεπτά για κάθε χρήστη μαζί με τις οδηγίες, τη βαθμονόμηση, τυχόν διορθώσεις και διαλείμματα.

Σχετικά με τη βαθμονόμηση, χρησιμοποιήθηκε σύστημα βαθμονόμησης και των δύο ματιών, 13 σημείων. Πριν από κάθε βίντεο εάν η ακρίβεια του σημείου εστίασης ξεπερνούσε ένα κατώφλι απόκλισης των 0.5°, πραγματοποιούνταν εκ νέου μία πλήρης βαθμονόμηση. Επίσης ο κεντρικός σημαδευτής εστίασης (fixation marker) αποτελούσε ένα προαιρετικό σημείο για διάλειμμα. Για την μετέπειτα επεξεργασία και την ευκολότερη χρήση των δεδομένων που συλλέχθηκαν, πραγματοποιήθηκε υποδειγματοληψία των 1,000Hz ώστε να ταιριάζουν με τη συχνότητα δειγματοληψίας του βίντεο, δηλαδή τα καρέ/δευτερόλεπτο. Τα δεδομένα είναι δημόσια διαθέσιμα και μπορούν να βρεθούν στη σελίδα:

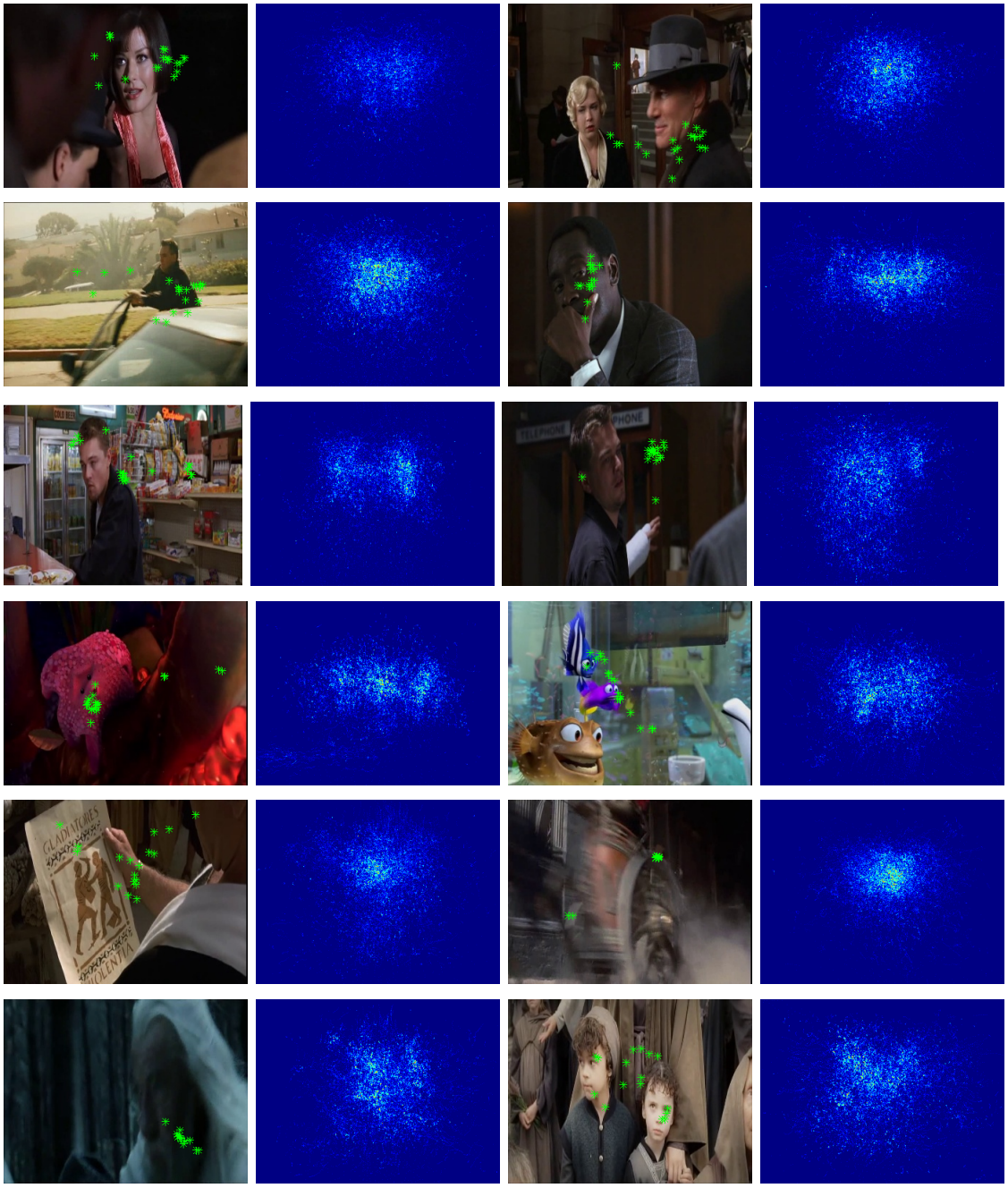
http://cvsp.cs.ntua.gr/antigoni/summe_etmd_eyetracking

Στα Σχήματα 5.1 και 5.2 παρουσιάζονται ενδεικτικά καρέ για κάθε βίντεο που υπάρχει στη βάση, καθώς και η κατανομή των οφθαλμοκινητικών δεδομένων για όλα τα καρέ του κάθε βίντεο.





Σχήμα 5.1: Ενδεικτικά καρέ από κάθε ένα από τα βίντεο της βάσης SumMe και δίπλα η υπέρθεση όλων των οφθαλμοκινητικών δεδομένων για όλα τα καρέ του κάθε βίντεο.



Σχήμα 5.2: Ενδεικτικά καρέ από κάθε ένα από τα βίντεο της βάσης ETMD και δίπλα η υπέρθεση όλων των οφθαλμοκινητικών δεδομένων για όλα τα καρέ του κάθε βίντεο.

Κεφάλαιο 6

Υπολογιστικό μοντέλο πρόβλεψης οπτικοακουστικής προσοχής σε βίντεο

6.1 Εισαγωγή

Όπως αναφέρθηκε και εισαγωγικά, η μεγάλη πλειοψηφία των μοντέλων προσοχής/εμφάνειας βασίζονται κατά κύριο λόγο στην οπτική πληροφορία αγνοώντας την ακουστική. Κατ΄ επέκταση, στη μεγάλη πλειοψηφία των δεδομένων οφθαλμοκινητικής παρακολούθησης (eye-tracking), οι συμμετέχοντες έβλεπαν τα βίντεο χωρίς ήχο. Από την άλλη μεριά, η οπτικοακουστική σύμμειξη έχει παρατηρηθεί ότι αυξάνει την επίδοση διαφόρων συστημάτων, σε πληθώρα εφαρμογών όπου τα οπτικά και ακουστικά ερεθίσματα συσχετίζονται και αποτελούν μέρος του ίδιου γεγονότος, όπως για παράδειγμα στη φωνή [49], στην περίληψη ταινιών [179], στην αλληλεπίδραση ανθρώπου-ρομπότ [247, 66, 70] και άλλα.

Σε αυτή την ενότητα, επιθυμούμε να γεφυρώσουμε το χάσμα μεταξύ της συμπεριφορικής έρευνας, όπου η οπτικοακουστική ενσωμάτωση έχει μελετηθεί αρκετά, και της υπολογιστικής μοντελοποίησης της προσοχής, η οποία μέχρι σήμερα βασιζόταν εξ ολοκλήρου στην οπτική πληροφορία. Στόχος είναι να δημιουργηθεί ένα δισδιάστατο οπτικοακουστικό μοντέλο προσοχής που θα μπορεί να εξηγήσει και να αναπαράξει τόσο ευρήματα από συμπεριφορικά πειράματα (επεκτείνοντας το [248]), όσο και από δεδομένα οφθαλμοκινητικής παρακολούθησης, μέσω της μοντελοποίησης με οπτικοακουστική εμφάνεια. Οι συνεισφορές αυτής της δουλειάς συνοψίζονται στα παρακάτω[1]:

- Οπτικοακουστική μοντελοποίηση της ανθρώπινης ανοδικής (bottom-up) προσοχής μέσω υπολογιστικής μοντελοποίησης με οπτικοακουστική εμφάνεια με τεχνικές επεξεργασίας σήματος.

- Διερεύνηση διάφορων τρόπων οπτικοακουστικής σύμμειξης μεταξύ της οπτικής και της

[1] Υλοποιήθηκε στα πλαίσια του προγράμματος COGNIMUSE (Multi-modal Signal and Event Processing in Perception and Cognition) http://cognimuse.cs.ntua.gr/the-project

μη χωρικής ακουστικής εμφάνειας η οποία δημιουργεί ένα δισδιάστατο οπτικοακουστικό χάρτη, σε αντίθεση με τεχνικές σύμμειξης στο επίπεδο της απόφασης ή των χαρακτηριστικών. Το προτεινόμενο σχήμα οπτικοακουστικής σύμμειξης είναι γενικό αφού μπορεί να εφαρμοστεί σε οποιοδήποτε ακουστικό μοντέλο και οποιοδήποτε οπτικό χωροχρονικό μοντέλο εμφάνειας.

- Η αξιολόγηση του προτεινόμενου μοντέλου γίνεται σε τρία επίπεδα:

  1. Το μοντέλο συγκρίνεται με ανθρώπινα πειραματικά ευρήματα από συμπεριφορικά πειράματα, σε ποιοτικό επίπεδο, διότι ο στόχος μας είναι η δημιουργία ενός μοντέλου που να μπορεί να εξηγεί και να αναπαράγει τέτοια ευρήματα και στοιχεία της ανθρώπινης προσοχής.

  2. Το μοντέλο επίσης συγκρίνεται με ανθρώπινα δεδομένα οφθαλμοκινητικής παρακολούθησης σε 6 διαθέσιμες βάσεις με οπτικοακουστικά δεδομένα, δηλαδή τις βάσεις DIEM, AVAD, Coutrot1, Coutrot2, SumMe, ETMD.

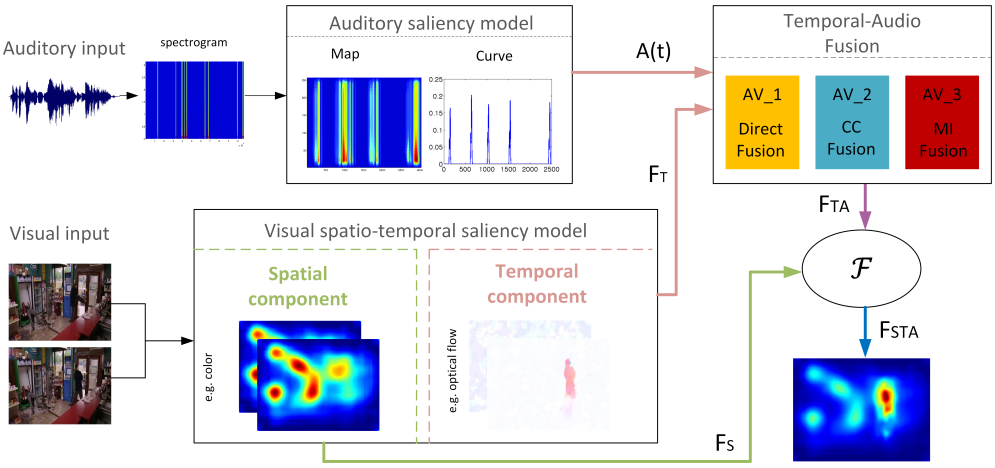
  3. Τέλος χρησιμοποιούνται δεδομένα fMRI για να διαπιστωθεί η εγκυρότητα της οπτικοακουστικής μοντελοποίησης που προτείνεται.

6.2 Οπτικοακουστική μοντελοποίηση

Όπως αναφέρθηκε και εισαγωγικά, η οπτικοακουστική ενσωμάτωση είναι ένα πεδίο έρευνας το οποίο έχει μελετηθεί αρκετά μέσω συμπεριφορικών πειραμάτων με στόχο να διερευνηθεί το πώς, που και πότε συνδυάζονται η οπτική με την ακουστική πληροφορία. Ένα παράδειγμα οπτικοακουστικής ολοκλήρωσης σε μια εργασία οπτικής αναζήτησης εμφανίζεται όταν υπάρχει ένας συγχρονισμένος ακουστικός τόνος που ακούγεται όταν αλλάζει και ο οπτικός στόχος. Σε διάφορα τέτοια πειράματα έχουν παρατηρηθεί μικρότεροι χρόνοι απόκρισης όταν υπάρχει και η ακουστική πληροφορία. Αυτή η παρατήρηση αποτέλεσε την αφετηρία ανάπτυξης του οπτικοακουστικού μας μοντέλου.

Στόχος μας είναι να συνδυάσουμε κατάλληλα μεμονωμένα μοντέλα ακουστικής εμφάνειας και οπτικής χωρο-χρονικής εμφάνειας ώστε να δημιουργήσουμε ένα οπτικοακουστικό μοντέλο το οποίο θα επαληθεύσουμε αρχικά με αποτελέσματα από συμπεριφορικά πειράματα. Στο σχεδιασμό του μοντέλου αυτού, γίνεται προσπάθεια να συνδυαστούν θεωρητικά και πειραματικά ευρήματα από τις νευροεπιστήμες με τεχνικές της επεξεργασίας σημάτων.

Σε αυτή την ενότητα αναφέρονται και εξηγούνται κάποιες επιλογές που έχουν γίνει όσον αφορά στην επεξεργασία της πληροφορίας αλλά και σε κάποιες σημαντικές παραμέτρους, καθώς και στα πιο σημαντικά στοιχεία της οπτικοακουστικής σύμμειξης, τα οποία έχουν βρεθεί ή εμπνευστεί από τη γνωσιακή έρευνα και τις νευροεπιστήμες. Μια επισκόπηση υψηλού επιπέδου του μοντέλου μας απεικονίζεται στο Σχήμα 6.1.



Σχήμα 6.1: Επισκόπηση του μοντέλου οπτικοακουστικής εμφάνειας.

6.2.1 Από τον ακουστικό χάρτη εμφάνειας στην ακουστική καμπύλη εμφάνειας

Όπως έχει αναφερθεί και προηγουμένως, τα περισσότερα μοντέλα εμφάνειας που έχουν αναπτυχθεί, είτε οπτικά είτε ακουστικά, παράγουν ως έξοδο δισδιάστατους χάρτες εμφάνειας. Στην περίπτωση των οπτικών μοντέλων, ο δισδιάστατος χάρτης έχει και φυσική υπόσταση, αφού αντιστοιχεί στην εικόνα και τις διαστάσεις της. Όσον αφορά τα ακουστικά μοντέλα, παρόλο που ο δισδιάστατος χάρτης είναι ενδιαφέρων και χρήσιμος αν θέλει κανείς να εξετάσει πως συνεισφέρουν οι διάφορες συχνότητες στην εμφάνεια, στη συγκεκριμένη εργασία μας ενδιαφέρει περισσότερο το πως αυτή μεταβάλλεται με το χρόνο και αν σε συγκεκριμένες χρονικές στιγμές υπάρχουν ή όχι προεξέχοντα γεγονότα. Επίσης, όπως στην περίπτωση των οπτικών μοντέλων ένας χάρτης δηλώνει ποιες περιοχές της εικόνας είναι προεξέχουσες και πόσο, αντίστοιχα και στα ακουστικά μοντέλα θέλουμε να ξέρουμε ποια τμήματα του ήχου είναι προεξέχοντα και πόσο.

Για τους παραπάνω λόγους, ο διασδιάστατος ακουστικός χάρτης υφίσταται μια περαιτέρω επεξεργασία ώστε να προκύψει από αυτόν μια μονοδιάστατη καμπύλη εμφάνειας. Η έννοια της καμπύλης εμφάνειας δεν είναι καινούρια [249, 170]. Στο [170] η καμπύλη προκύπτει από το χάρτη συνδυάζοντας αθροιστικά τις τιμές εμφάνειας των διαφορετικών συχνοτήτων, ενώ στο [249] για κάθε χρονική στιγμή η εμφάνεια προκύπτει παίρνοντας τη μέγιστη τιμή ανάμεσα σε όλες τις συχνότητες. Η ικανότητα να εντοπίζονται τα προεξέχοντα γεγονότα όταν πραγματοποιείται μεγιστοποίηση πάνω σε ολόκληρο το χάρτη έχει επαληθευτεί συμπεριφορικά στο [168] και εδώ ακολουθείται επίσης αυτή η προσέγγιση. Εαν ορίσουμε με $M_a(\ell, f)$ τον ακουστικό χάρτη εμφάνειας που είναι συνάρτηση του χρόνου ℓ και της συχνότητας f και με $S_a(\ell)$ την ακουστική καμπύλη εμφάνειας, τότε αυτή υπολογίζεται ως:

$$S_a(\ell) = \max_f M_a(\ell, f)$$

Η ίδια προσέγγιση ακολουθείται και στο [169], όπου οι χάρτες χαρακτηριστικών αθροίζονται και το τελικό χρονικό σκορ είναι το μέγιστο κάθε χρονικής στιγμής.

6.2.2 Οπτικοακουστικό χρονικό παράθυρο ενσωμάτωσης

Ένα σημαντικό εύρημα που προέρχεται από τις γνωσιακές επιστήμες και έχει μελετηθεί εκτενώς μέσα στην τελευταία δεκαετία σχετίζεται με το οπτικοακουστικό χρονικό παράθυρο ενσωμάτωσης. Όπως αναφέρθηκε και εισαγωγικά, πειραματικά έχει βρεθεί ότι η οπτικοακουστική ενσωμάτωση είναι πιο αποτελεσματική όταν το οπτικό και το ακουστικό ερέθισμα είναι απόλυτα συγχρονισμένα. Εξακολουθεί να πραγματοποιείται, όμως, ακόμα και αν οι δύο τροπικότητες είναι ελαφρώς ασυγχρόνιστες. Σχετικά συμπεριφορικά πειράματα υποδεικνύουν μέγιστο χρονικό παράθυρο (μέγιστο χρονικό περιθώριο χρονικής ασυγχρονίας) των 200ms [2, 102, 98].

Υπάρχουν επίσης ενδείξεις ότι σε χρονικές καταστάσεις/εργασίες η ακουστική πληροφορία κυριαρχεί έναντι της οπτικής [250, 251, 252]. Αυτές μαζί με τις ενδείξεις ότι ο ήχος επηρεάζει την όραση ακόμα και αν είναι μεταξύ τους ασύγχρονοι, υποδεικνύει ότι πρέπει να ληφθούν υπόψη τόσο παρελθοντικές όσο και μελλοντικές τιμές στη μοντελοποίηση.

Φιλτράρουμε κατάλληλα την καμπύλη ακουστικής εμφάνειας ώστε να μπορέσουμε να ενσωματώσουμε αυτό το εύρημα στο υπολογιστικό μοντέλο μας. Χρησιμοποιούμε ένα παράθυρο Hanning με 200ms μήκος και το κεντράρουμε στην τρέχουσα χρονική στιγμή πάνω στην ακουστική καμπύλη εμφάνειας. Η μορφή του παραθύρου Hanning είναι κατάλληλη για να μοντελοποιήσει το χρονικό παράθυρο της ενσωμάτωσης γιατί ενισχύει τα συγχρονισμένα ερεθίσματα και εξασθενεί τα ασυγχρόνιστα. Αν το οπτικό και ακουστικό ερέθισμα είναι συγχρονισμένα, εφόσον η τιμή της ακουστικής εμφάνειας για την τρέχουσα χρονική στιγμή παραμένει αναλλοίωτη μετά τον πολλαπλασιασμό της με το παράθυρο (στο κέντρο η τιμή είναι 1), το φαινόμενο της ολοκλήρωσης θα είναι μέγιστο, αλλιώς θα έχει υπάρξει κάποια απόσβεση μέσω του πολλαπλασιασμού που μοντελοποιεί την ασυγχρονία. Μετά την παραθύρωση, εφαρμόζουμε ένα φίλτρο κινούμενου μέσου (moving average), του οποίου το αποτέλεσμα είναι μια νέα καμπύλη εμφάνειας που έχει ενσωματώσει την επιρροή του χρόνου και της ασυγχρονίας.

Εάν ορίσουμε με $H(l)$ το παράθυρο Hanning μήκους $2N + 1$ και με $A(t)$ την τελική καμπύλη εμφάνειας, αυτή υπολογίζεται ως:

$$A(t) = \frac{1}{2N+1} \sum_{\ell=-N}^{N} S_a(t+\ell)H(\ell)$$

όπου t είναι ο χρονικός δείκτης της οπτικής πληροφορίας του βίντεο (video frame index), ενώ ℓ είναι ο χρονικός δείκτης του ήχου. Στη θέση του παραθύρου Hanning θα μπορούσαν να εφαρμοστούν και άλλα είδη παραθύρων με αντίστοιχες ιδιότητες, π.χ. το Hamming χωρίς να υπάρξουν σημαντικές διαφορές στα αποτελέσματα.

6.2.3 Σύμμειξη οπτικοακουστικής εμφάνειας

Το πιο σημαντικό μέρος της μοντελοποίησης μας συνιστάται στην απόφαση του πώς και που η οπτική και η ακουστική εμφάνεια θα συγχωνευθούν ώστε να δημιουργηθεί ένα οπτικοακου-

στικό μοντέλο εμφάνειας. Στο [253] περιγράφονται διάφορες συνδυαστικές τεχνικές για να ενσωματώνονται χάρτες χαρακτηριστικών από διαφορετικές οπτικές εισόδους. Τα χαρακτηριστικά αυτά αντιπροσωπεύουν μη-συγκρίσιμες τροπικότητες και έχουν διαφορετικό δυναμικό εύρος. Το ίδιο ακριβώς ισχύει και στην περίπτωση της σύμμειξης οπτικής και ακουστικής εμφάνειας. Τα προτεινόμενα σχήματα συγχώνευσης βασίζονται στην υπόθεση ότι αφού τα ακουστικά χαρακτηριστικά είναι δυναμικά, η παρουσία ακουστικού ερεθίσματος επηρεάζει κυρίως τα χρονικά/δυναμικά οπτικά χαρακτηριστικά. Παράλληλα, υπάρχουν ενδείξεις ότι τα δυναμικά οπτικά χαρακτηριστικά, όπως το τρεμόπαιγμα και η κίνηση, επηρεάζονται πολύ από την ύπαρξη ακουστικής πληροφορίας, εφόσον μάλιστα τα ακουστικά ερεθίσματα κυριαρχούν των οπτικών σε χρονικές εργασίες. Διάφορα παραδείγματα παρουσιάζονται στη βιβλιογραφία [254], όπως η ψευδαίσθηση της μπάλας που αναπηδά [251] που υποδεικνύει ότι ο ήχος αλληλεπιδρά τόσο με το τρεμόπαιγμα (flicker) [250, 255] όσο και με την κίνηση (motion) [256]. Με αφορμή τα ευρήματα αυτά, συγχωνεύουμε την ακουστική εμφάνεια με την εμφάνεια των χρονικών οπτικών χαρακτηριστικών.

Πολύ μεγάλης σημασίας είναι επίσης το πώς αυτές οι δύο διαφορετικές τροπικότητες θα συγχωνευτούν εφόσον είναι μη-συγκρίσιμες, έχουν άλλες διαστάσεις, κλπ. Όταν ο ήχος απουσιάζει, τότε οι οπτικοί χάρτες θα πρέπει να παραμένουν ως έχουν, ενώ όταν είναι παρών, η ακουστική εμφάνεια πρέπει να τους σταθμίζει κατάλληλα. Πειραματιστήκαμε με τρία διαφορετικά σχήματα σύμμειξης, εμπνευσμένα από γνωστές τεχνικές συνδυασμού διαφορετικών τροπικοτήτων. Συγχώνευση πραγματοποιείται μεταξύ της ακουστικής εμφάνειας και της εμφάνειας κάθε χρονικού οπτικού χαρακτηριστικού ξεχωριστά. Αυτή η συγχώνευση έχει ως αποτέλεσμα έναν κοινό χρονικο-ακουστικό χάρτη F_{TA}, όπου η επίδραση του ήχου έχει ενσωματωθεί στον δισδιάστατο χρονικό-οπτικό χάρτη. Μετά την χρονικο-ακουστική σύμμειξη, η χωρική συνιστώσα εισέρχεται και αυτή κατάλληλα, σύμφωνα με τη στρατηγική σύμμειξης κάθε μοντέλου $\mathscr{F}$ (για οπτική-μόνο εμφάνεια), καταλήγοντας έτσι στον τελικό χωρο-χρονικο-ακουστικό χάρτη (saliency spatio-temporal-audio, STA saliency map), τον οποίο συμβολίζουμε με F_{STA} και ο οποίος απεικονίζεται και στο κάτω μέρος του Σχήματος 6.1:

$$F_{STA} = \mathscr{F}(F_S, F_{TA}) \tag{6.1}$$

όπου F_S είναι η χωρική οπτική συνιστώσα. αντίστοιχα. Εστιάζουμε στην χρονικο-ακουστική σύμμειξη, διότι η τελική σύμμειξη $\mathscr{F}$ εξαρτάται από το εκάστοτε χωρο-χρονικό οπτικό μοντέλο εμφάνειας που χρησιμοποιείται. Για παράδειγμα, στο μοντέλο Itti et al. [109], το $\mathscr{F}$ είναι ένας μέσος όρος όλων των ξεχωριστών χαρτών εμφάνειας. Παρακάτω περιγράφονται αναλυτικά τα προτεινόμενα σχήματα σύμμειξης μεταξύ της ακουστικής και της χρονικής οπτικής εμφάνειας ώστε να υπολογιστεί το F_{TA}.

Απευθείας σύμμειξη εμφάνειας (Direct fusion)

Πειραματιζόμαστε με την απευθείας σύμμειξη της ακουστικής καμπύλης εμφάνειας με τους δυναμικούς οπτικούς χάρτες εμφάνειας. Οι δύο εμφάνειες συνδυάζονται πολλαπλασιαστικά,

ξεχωριστά για κάθε οπτικό χαρακτηριστικό:

$$F_{TA}(x, y, t) = F_T(x, y, t)(1 + A(t)) \tag{6.2}$$

όπου x, y είναι οι συντεταγμένες των εικονοστοιχείων της εικόνας, το F_{TA} είναι ο συνδυασμένος χάρτης, το F_T αντιπροσωπεύει το χάρτη εμφάνειας ενός μεμονωμένου χρονικού οπτικού χαρακτηριστικού και A είναι η ακουστική καμπύλη εμφάνειας. Αυτή η ιδέα πρωτοεμφανίστηκε με ένα παρόμοιο αλλά όχι ταυτόσημο τρόπο στο [175]. Οι συγγραφείς σε αυτή τη δουλειά πειραματίζονται μόνο με χωρική ακουστική εμφάνεια και ο ακουστικός χάρτης εμφάνειας που παράγουν είναι μόνο η θέση της ακουστικής πηγής στο χώρο. Επομένως συνδυάζουν τους δύο χάρτες με πολλαπλασιασμό στοιχείο προς στοιχείο. Εμείς επεκτείνουμε αυτή την ιδέα ώστε να μπορεί να εφαρμοστεί στο μοντέλο και τα δεδομένα μας.

Ετεροσυσχέτιση (Cross-correlation) μεταξύ ακουστικής και οπτικής εμφάνειας (CC fusion)

Στο [257] προτείνεται η ετεροσυσχέτιση σαν μέτρο της οπτικοακουστικής ενσωμάτωσης. Η ετεροσυσχέτιση μεταξύ πολλαπλών σημάτων από διαφορετικούς αισθητήρες είναι ένας σημαντικός δείκτης της σχέσης τους: σήματα που προέρχονται από το ίδιο γεγονός συνήθως εμφανίζουν μια πολύ στενή χρονική σχέση, εξαιτίας της εξάρτησής τους από το ίδιο αυτό γεγονός. Αντίθετα, όταν πολλαπλά σήματα έχουν παραχθεί από διαφορετικά μεταξύ τους γεγονότα, οι χρονικές δομές τους συνήθως ειναι ασυσχέτιστες. Συμπερασματικά, σήματα με παρόμοια χρονική μικροδομή και επομένως υψηλή ετεροσυσχέτιση είναι πολύ πιθανό να προέρχονται από το ίδιο γεγονός και να ενσωματώνονται πιο ισχυρά.

Στην περίπτωση της ετεροσυσχέτισης, σύμφωνα με τη μοντελοποίηση του χρονικού παραθύρου της ενσωμάτωσης που περιγράφηκε παραπάνω, πρέπει να χρησιμοποιηθεί η ετεροσυσχέτιση με περιορισμένη καθυστέρηση (lag) τ μεταξύ των δύο τροπικοτήτων, διότι αν ο ήχος και η εικόνα προκύπτουν από το ίδιο γεγονός, μπορούν να γίνουν αντιληπτά ως ένα ενιαίο γεγονός αν η ασυγχρονία τους δεν ξεπερνά τα 200ms. Επομένως, η καθυστέρηση τ της ετεροσυσχέτισης δε μπορεί να ξεπερνάει τα 200ms. Από εδώ και στο εξής συμβολίζουμε την ετεροσυσχέτιση με R_{TA} και την υπολογίζουμε σε χρονικά παράθυρα των οπτικών και ακουστικών εμφανειών. Πιο συγκεκριμένα επιλέγεται ένα παράθυρο του 1 δευτερολέπτου (όπου αντιστοιχούν $2k + 1$ καρέ) και υπολογίζεται η ετεροσυσχέτιση μεταξύ των χρονικών σειρών της ακουστικής εμφάνειας και καθενός εικονοστοιχείου ενός χρονικού οπτικού χάρτη εμφάνειας. Στη συνέχεια η μέγιστη τιμή της ετεροσυσχέτισης χρησιμοποιείται ως σταθμιστής της τρέχουσας τιμής του εικονοστοιχείου αυτού.

$$F_{TA}(x, y, t) = F_T(x, y, t)(1 + \max_{\tau} R_{TA}(x, y, t, \tau)),$$
$$t - n_{cc} \leq \tau \leq t + n_{cc} \tag{6.3}$$

με

$$R_{TA}(x,y,t,\tau) = \frac{1}{2k+1}\sum_{m=t-k}^{t+k} F_T(x,y,m)A(m-\tau),$$
$$t - n_{cc} \leq \tau \leq t + n_{cc} \tag{6.4}$$

όπου n_{cc} οι πιθανές τιμές της καθυστέρησης τ (που δε μπορούν να ξεπερνούν τα 200ms) και R_{AT} η ετεροσυσχέτιση μεταξύ της ακουστικής και της χρονικής οπτικής εμφάνειας.

Αμοιβαία πληροφορία (Mutual Information) μεταξύ ακουστικής και οπτικής εμφάνειας (MI fusion)

Σε μια προσπάθεια να εξερευνήσουμε περισσότερες εναλλακτικές που εκφράζουν κάποια σχέση οπτικοακουστικών ερεθισμάτων και με αφορμή τις εργασίες [258, 259], εξετάζουμε τη δυνατότητα της αμοιβαίας πληροφορίας (mutual information) να εκφράσει τη σχέση αυτή. Υποθέτουμε ότι οι δύο εμφάνειες προέρχονται από μια κοινή πιθανοτική διαδικασία που είναι στατική και γκαουσιανή για ένα μικρό χρονικό διάστημα [258]. Εάν συμβολίσουμε με $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ αυτή την κοινή γκαουσιανή κατανομή, τα $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$ μπορούν να υπολογιστούν από τα οπτικοακουστικά δεδομένα ανά καρέ. Για ένα συγκεκριμένο καρέ t:

$$\boldsymbol{\mu}(x,y,t) = b\begin{bmatrix} A(t) \\ F_T(x,y,t) \end{bmatrix} + (1-b)\boldsymbol{\mu}(x,y,t-1) \tag{6.5}$$

$$\boldsymbol{\Sigma}(x,y,t) = \frac{1}{1+a}\left(a\left(\begin{bmatrix} A(t) \\ F_T(x,y,t) \end{bmatrix} - \boldsymbol{\mu}(x,y,t-1)\right)\left(\begin{bmatrix} A(t) \\ F_T(x,y,t) \end{bmatrix} - \boldsymbol{\mu}(x,y,t-1)\right)^T + \boldsymbol{\Sigma}(x,y,t-1)\right) \tag{6.6}$$

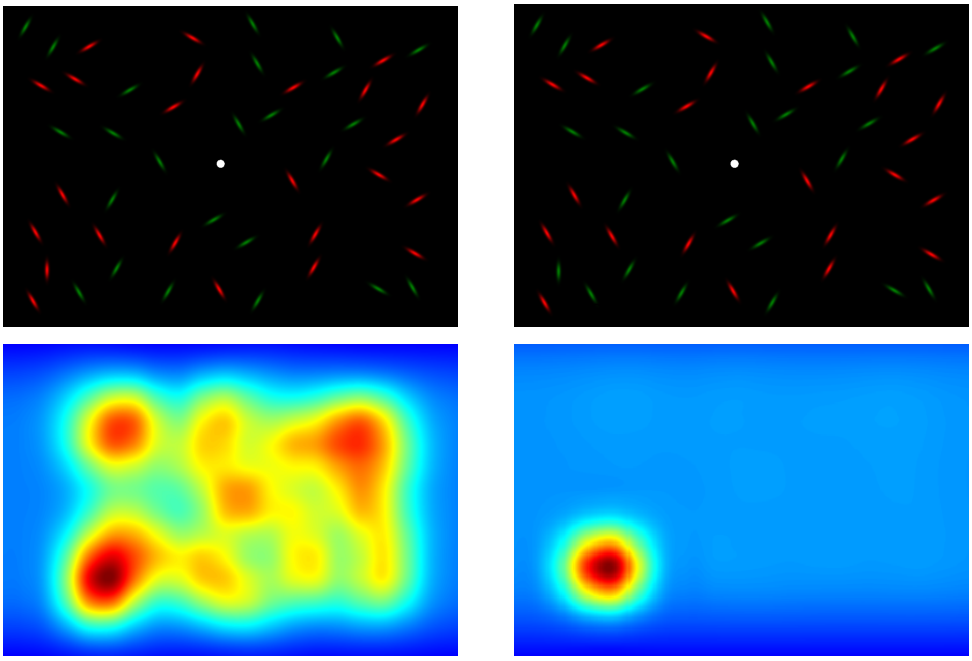
όπου a, b είναι προκαθορισμένα βάρη με τιμή μεταξύ $[0,1]$, που ελέγχουν την εξάρτηση των τρέχουσων τιμών από τις παρελθοντικές. Η αμοιβαία πληροφορία μεταξύ των ακουστικών και οπτικών δεδομένων υπολογίζεται ως εξής:

$$I(x,y,t) = -\frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_A(t)||\boldsymbol{\Sigma}_{F_T}(x,y,t)|}{|\boldsymbol{\Sigma}(x,y,t)|}\right) \tag{6.7}$$

και το $\boldsymbol{\Sigma}$ εκφράζεται ως εξής:

$$\boldsymbol{\Sigma}(x,y,t) = \begin{bmatrix} \boldsymbol{\Sigma}_A(t) & \boldsymbol{\Sigma}_{AF_T}(x,y,t) \\ \boldsymbol{\Sigma}_{AF_T}(x,y,t)^T & \boldsymbol{\Sigma}_{F_T}(x,y,t) \end{bmatrix} \tag{6.8}$$

Στην περίπτωση που έχουμε ένα ακουστικό και ένα οπτικό χαρακτηριστικό, που είναι και η περίπτωσή μας, οι Εξ.(6.7) και (6.8) απλοποιούνται ως εξής:



Σχήμα 6.2: Οι πάνω εικόνες από το [2] απεικονίζουν δύο ενδεικτικά καρέ από τα "pip & pop" ερεθίσματα, κατά τη στιγμή που αναβοσβήνει ο στόχος (η κάθετη γραμμή στην κάτω αριστερή γωνία). Οι κάτω εικόνες αντιστοιχούν στον οπτικό (αριστερά) και τον οπτικοακουστικό (δεξιά) χάρτη εμφάνειας που παράχθηκε με απευθείας σύμμειξη.

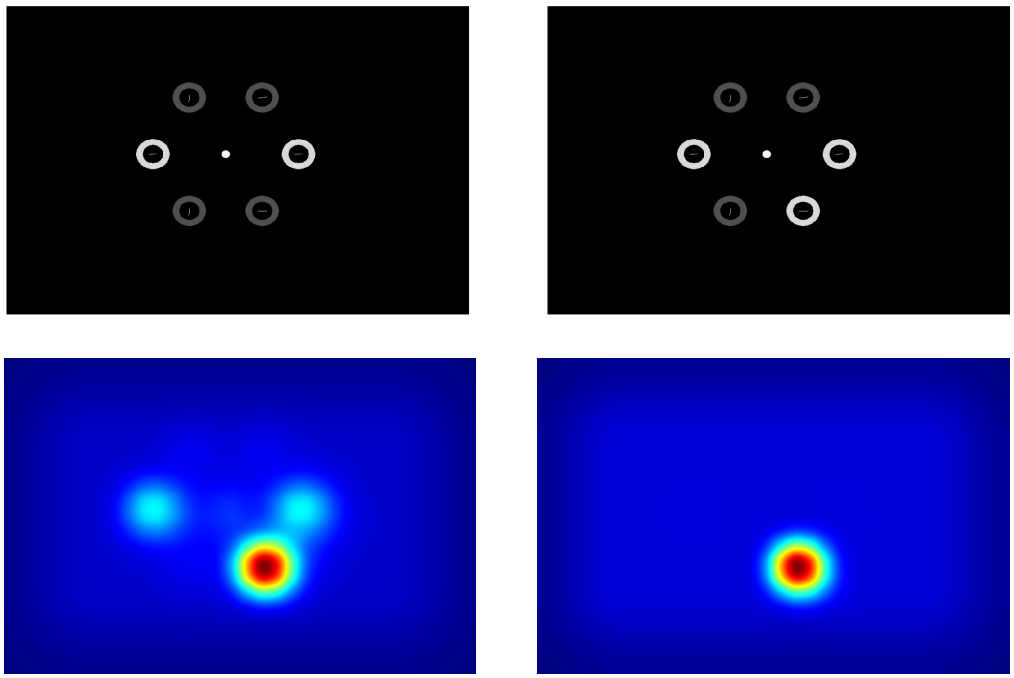
$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_A(t)^2 & \sigma_{AF_T}(x,y,t) \\ \sigma_{AF_T}(x,y,t) & \sigma^2_{F_T}(x,y,t) \end{bmatrix} \tag{6.9}$$

$$I(x,y,t) = -\frac{1}{2}\log\left(1-\rho^2(x,y,t)\right) \tag{6.10}$$

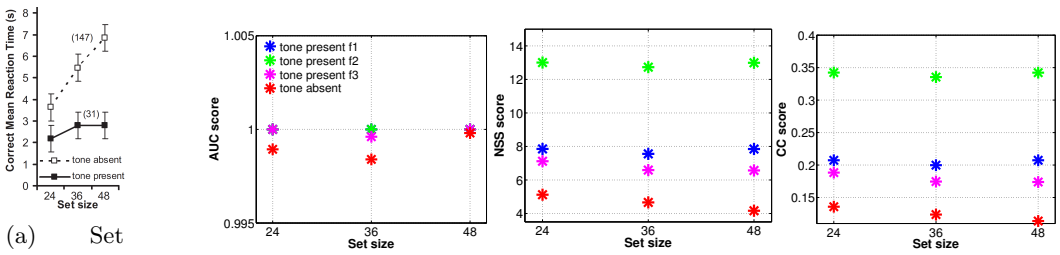
$$\rho(x,y,t) = \frac{\sigma_{AF_T}(x,y,t)}{\sqrt{\sigma_A(t)\sigma_{F_T}(x,y,t)}} \tag{6.11}$$

όπουσ_{AF_T}, σ_A, και σ_{F_T} είναι οι βαθμωτές εκτιμήσεις της συνδιακύμανσης των οπτικοακουστικών χαρακτηριστικών και οι διακυμάνσεις των ακουστικών και των οπτικών χαρακτηριστικών αντίστοιχα, ενώ ρ είναι η Pearson's Correlation Coefficient. Ο χάρτης της συγχώνευσης υπολογίζεται ως:

$$F_{TA}(x,y,t) = F_T(x,y,t)(1+I(x,y,t)) \tag{6.12}$$



Σχήμα 6.3: Οι πάνω εικόνες από το [3] απεικονίζουν δύο ενδεικτικά καρέ από τα "sine vs square" ερεθίσματα για την τετραγωνική διαμόρφωση τη στιγμή που αναβοσβήνει ο στόχος (η κάθετη γραμμή στην κάτω δεξιά γωνία). Στην κάτω σειρά απεικονίζονται ο οπτικός (αριστερά) και ο οπτικοακουστικός (δεξιά) χάρτης εμφάνειας που προκύπτει με απευθείας σύμμειξη.



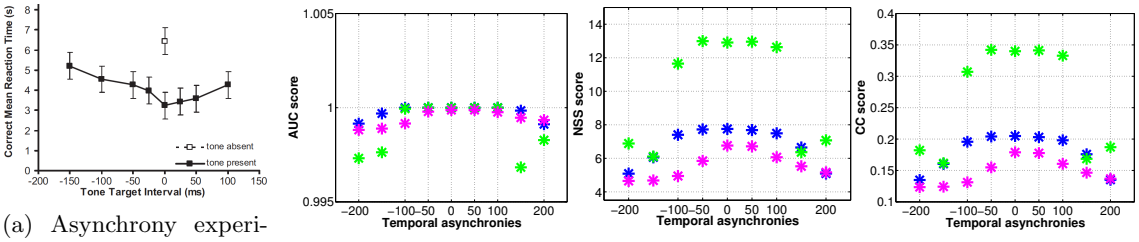
(a) Set size experiment (b) AUC (c) NSS (d) CC

Σχήμα 6.4: (α) Πρωτότυπο σχήμα από το [2] και (β, γ, δ) τα *AUC, NSS, CC* για το set size πείραμα με τα διάφορα σχήματα σύμμειξης. Το μπλε χρώμα αντιπροσωπεύει την απευθείας σύμμειξη, το πράσινο τη σύμμειξη με ετεροσυσχέτιση, το φούξια τη σύμμειξη με αμοιβαία πληροφορία και το κόκκινο την περίπτωση όπου απουσιάζει ο ήχος.

6.3 Αξιολόγηση

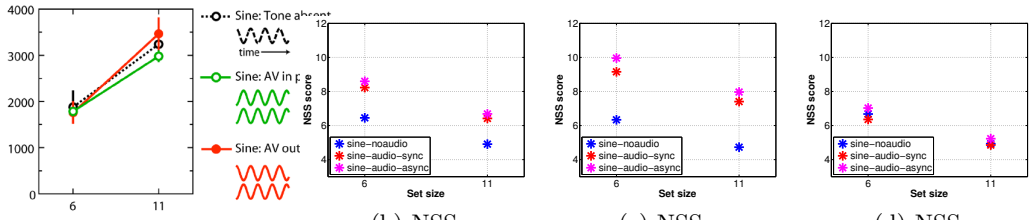
6.3.1 Μετρικές αξιολόγησης

Υιοθετούμε ευρέως χρησιμοποιούμενες μετρικές εμφάνειας για να ποσοτικοποιήσουμε τα αποτελέσματα και να επαληθεύσουμε το μοντέλο μας με συμπεριφορικά πειράματα. Εφόσον η έξοδος μας είναι ένας δισδιάστατος χάρτης εμφάνειας όπου η ακουστική εμφάνεια έχει ενσω-



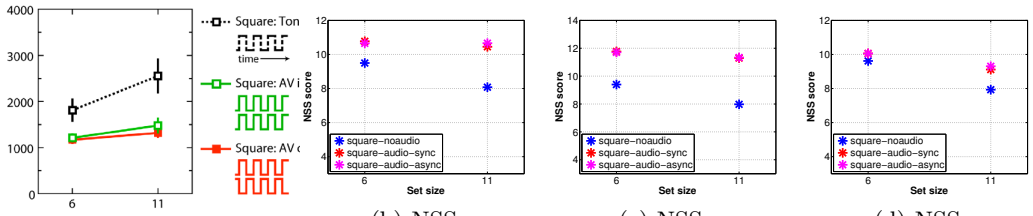
(a) Asynchrony experiment (b) AUC (c) NSS (d) CC

Σχήμα 6.5: (α) Πρωτότυπο σχήμα από το [2] και (β, γ, δ) τα *AUC, NSS, CC* για το temporal asynchrony πείραμα. Οι τιμές με μείον αναφέρονται στις περιπτώσεις όπου ο ήχος προηγείται της εικόνας. Τα διαφορετικά χρώματα αντιπροσωπεύουν τις διαφορετικές τεχνικές σύμμειξης όπως και παραπάνω.



(a) Set size - sine modulation (b) NSS (c) NSS (d) NSS

Σχήμα 6.6: (α) Πρωτότυπο σχήμα από το [3] για την ημιτονοειδή διαμόρφωση και (β, γ, δ) τα αποτελέσματα του *NSS* για (direct, CC, MI) σύμμειξη αντίστοιχα για το set size experiment στα ερεθίσματα sine vs square.



(a) Set size - square modulation (b) NSS (c) NSS (d) NSS

Σχήμα 6.7: (α) Πρωτότυπο σχήμα από το [3] για την τετραγωνική διαμόρφωση και (β, γ, δ) τα αποτελέσματα του *NSS* για (direct, CC, MI) σύμμειξη αντίστοιχα για το set size experiment στα ερεθίσματα sine vs square.

ματωθεί στο δισδιάστατο οπτικό χάρτη, υιοθετούμε μετρικές που χρησιμοποιούνται για την αξιολόγηση της οπτικής εμφάνειας και συγκεκριμένα αυτές που έχουν παρουσιαστεί αναλυτικά στο [260, 110].

Συμβολίζουμε με *ESM* (Estimated Saliency Map) την έξοδο του μοντέλου. Συνήθως, ο *GSM* (Ground-truth Saliency Map) είναι χάρτης που δημιουργείται από τα δεδομένα οφθαλμοκινητικής παρακολούθησης. Για την περίπτωση των συμπεριφορικών πειραμάτων όπου δεν διαθέτουμε τέτοια δεδομένα, αντιπροσωπεύει την πραγματική θέση του στόχου με την έννοια ότι ο στόχος είναι προεξέχων εφόσον ξεχωρίζει από το περιβάλλον του. Μία αντίστοιχη

συλλογιστική εφαρμόστηκε και στο [94]. Οι χρησιμοποιούμενες μετρικές είναι οι ακόλουθες (λεπτομέρειες στα [260, 110, 261]):

1) Linear Correlation Coefficient (CC): Μετράει την ισχύ μια γραμμικής σχέσης μεταξύ των GSM και ESM. Όταν το CC είναι κοντά στο $+1/-1$ υπάρχει μία σχεδόν τέλεια γραμμική σχέση μεταξύ των δύο μεταβλητών.

2) Normalized Scanpath Saliency (NSS): Είναι ο μέσος όρος των τιμών των αποκρίσεων στα σημεία των θέσεων του ματιού σε ένα χάρτη ESM κανονικοποιημένου με μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση. Δείχνει πόσες φορές μεγαλύτερη είναι η τιμή του ESM για κάθε σημείο εστίασης του ματιού σε σχέση με το συνολικό μέσο όρο του ESM. Ως τελική τιμή NSS θεωρούμε τη μέση τιμή στις εστιάσεις όλων των θεατών.

3) Area Under Curve shuffled (AUCs): Είναι το εμβαδό κάτω από την καμπύλη Receiver Operating Characteristics (ROC). Εδώ η τοποθεσία του στόχου θεωρείται ως το θετικό σύνολο, ενώ δειγματοληπτούνται σημεία της εικόνας από τις θέσεις όπου βρίσκονται οι distractors, ώστε αποκτούμε το *shuffled AUC* [260, 129]. Ο χάρτης ESM αντιμετωπίζεται ως ένας δυαδικός ταξινομητής που διαχωρίζει τα θετικά από τα αρνητικά δείγματα. Η καμπύλη ROC δημιουργείται κατωφλιώνοντας τον ESM και σχεδιάζοντας το ποσοστό true positive προς το false positive. Το εμβαδό κάτω από το μέσο όλων των ROC καμπυλών είναι το AUC. Όταν $AUC = 1$ η πρόβλεψη είναι τέλεια.

6.3.2 Συμπεριφορικά πειράματα

Για την επαλήθευση του μοντέλου μέσω αποτελεσμάτων από συμπεριφορικά πειράματα, αρχικά θεωρούμε απλά ερεθίσματα από πειράματα οπτικής αναζήτησης. Σε τέτοια πειράματα, δίνεται οδηγία στους συμμετέχοντες να εστιάζουν το βλέμμα τους σε κάποιο συγκεκριμένο σημείο της οθόνης και να προσπαθούν να ανιχνεύσουν ένα στόχο που περιτριγυρίζεται από distractors, χωρίς όμως να σκανάρουν σειριακά την εικόνα. Όταν ανιχνεύσουν το στόχο πρέπει να πατήσουν ένα κουμπί. Η επίδοση συνήθως μετριέται με το μέσο χρόνο απόκρισης (mean Response Time (RT)) που είναι ο χρόνος από τη στιγμή που εμφανίζεται ο στόχος μέχρι να πατηθεί το κουμπί. Ο μέσος χρόνος απόκρισης έχει συνδεθεί ξανά στο παρελθόν με την εμφάνεια, με την έννοια ότι όταν αυτός μειώνεται σημαίνει ότι η εμφάνεια του στόχου αυξάνεται, γιατί είναι πιο εύκολο για το συμμετέχοντα να τον εντοπίσει [111, 262, 263]. Τα πειράματά μας στοχεύουν να αναπαράξουν αυτή τη συμπεριφορά και τις παρατηρήσεις των συμπεριφορικών πειραμάτων σε σχέση με την εμφάνεια χρησιμοποιώντας τις μετρικές που αναφέρθηκαν προηγουμένως.

Τα ερεθίσματα που χρησιμοποιούνται εδώ προέρχονται από πειράματα οπτικής αναζήτησης [2, 3] και συγκεκριμένα από τα "pip and pop" και "sine vs square" ερεθίσματα. Τα πρώτα αποτελούνται από μικρές πλάγιες ευθείες γραμμές που συνεχώς αλλάζουν χρώμα από κόκκινο σε πράσινο και αντίστροφα, ενώ ο στόχος είναι μία κάθετη ή οριζόντια ευθεία γραμμή που αλλάζει και αυτή χρώμα, την ίδια στιγμή που συγχρονισμένα ακούγεται ένας ηχητικός τόνος. Ο συμμετέχων πρέπει να αναγνωρίσει αν ο στόχος είναι οριζόντιος ή κάθετος και να πατήσει ένα κουμπί. Οι υπόλοιπες γραμμές έχουν διάφορους άλλους προσανατολισμούς και

το χρώμα τους αλλάζει χωρίς να παίζετται ταυτόχρονα ήχος. Στο Σχήμα. 6.2 απεικονίζονται δύο καρέ όπου φαίνεται η αλλαγή χρώματος του στόχου. Για τα πειράματα, γνωρίζουμε την πραγματική θέση του στόχου καθώς και τα καρέ στα οποία αλλάζει χρώμα, δηλαδή τα καρέ στα οποία γίνεται προεξέχων. Στο ίδιο σχήμα φαίνεται και ο οπτικός και οπτικοακουστικός χάρτης που αντιστοιχεί στο δεύτερο καρέ που απεικονίζεται.

Όσον αφορά τα ερεθίσματα "sine vs square", είναι και αυτά ευθείες γραμμές που περικλείονται από έναν δακτύλιο του οποίου η φωτεινότητα αλλάζει συνεχώς με το χρόνο στην κλίμακα grayscale, ακολουθώντας ημιτονοειδή ή τετραγωνική διαμόρφωση. Οι αλλαγές φωτεινότητας του στόχου είναι είτε συγχρονισμένοι με το άκουσμα ενός μη χωρικού ακουστικού τόνου με συμφωνία φάσης ή με 180° διαφορά, είτε χωρίς το άκουσμα τόνου. Ο στόχος είναι πάλι είτε οριζόντιος είτε κάθετος ενώ οι υπόλοιπες γραμμές μπορούν να έχουν οποιοδήποτε άλλο προσανατολισμό. Το πείραμα αυτό είναι συγκριτικό: Οι συγγραφείς ισχυρίζονται ότι η οπτικοακουστική ενσωμάτωση απαιτεί την ύπαρξη απότομων γεγονότων και γι΄ αυτό συγκρίνουν το ίδιο οπτικοακουστικό ερέθισμα με δύο διαφορετικά είδη διαμόρφωσης, την ημιτονοειδή που είναι σταδιακή και την τετραγωνική που είναι απότομη. Ένα παράδειγμα απεικόνισης των ερεθισμάτων αυτών βρίσκεται στο Σχήμα. 6.3, όπου απεικονίζονται δύο καρέ μια τετραγωνικής διαμόρφωσης.

Πείραμα μεγέθους συνόλου "pip and pop"

Στο [2] τα πειράματα υποδεικνύουν ότι όταν δεν υπάρχουν ηχητικοί τόνοι, ο χρόνος απόκρισης αυξάνει ανάλογα με τον αριθμό των γραμμών, δηλαδή μειώνεται η εμφάνεια του στόχου, πιθανώς επειδή απαιτείται σειριακή οπτική αναζήτηση. Αντιθέτως όταν η αλλαγή χρώματος του στόχου συνοδεύεται από ένα σύντομο συγχρονισμένο τόνο, το μέγεθος του συνόλου των γραμμών δεν παίζει σπουδαίο ρόλο. Το πρωτότυπο σχήμα από το [2] παρουσιάζει τα ευρήματα αυτά στο Σχήμα 6.4α. Διερευνούμε εάν το μοντέλο μας με τα διάφορα σχήματα σύμμειξης μπορεί να αναπαράξει αυτή τη συμπεριφορά. Η είσοδος είναι τα ερεθίσματα "pip and pop", η έξοδος ειναι οι οπτικοακουστικοί χάρτες εμφάνειας και η αξιολόγηση πραγματοποιείται με τις μετρικές που αναφέρθηκαν στην ενότητα 6.3. Στο Σχήμα 6.4 παρουσιάζονται τα αποτελέσματα για τα *AUC, NSS*, και *CC*.

Παρατηρούμε ότι τα NSS και CC αναπαράγουν αρκετά καλά τα ευρήματα που παρουσιάζονται στο Σχήμα 6.4α. Παρ΄ ότι οι κλίσεις των ευθειών δεν είναι οι ίδιες, η τάση της καμπύλης είναι παρόμοια με αυτή των συμπεριφορικών πειραμάτων για όλες τις τεχνικές σύμμειξης. Επίσης όταν υπάρχει ηχητικός τόνος, η εμφάνεια είναι μεγαλύτερη σε σχέση με τις περιπτώσεις που δεν υπάρχει, το οποίο συμφωνεί με τα συμπεριφορικά πειράματα. Το AUC είναι πολύ υψηλό σε όλες τις περιπτώσεις, διότι όταν ο στόχος αλλάζει χρώμα, είναι η μόνη γραμμή η οποία αλλάζει χρώμα. Φαίνεται ότι αυτή η μετρική δεν μπορεί να αντικατοπτρίσει τα ζητούμενα αποτελέσματα για τα συγκεκριμένα ερεθίσματα.

Πείραμα χρονικής ασυγχρονίας "pip and pop"

Ένα δεύτερο συμπεριφορικό πείραμα από το [2] διερευνά την οπτικοακουστική ενσωμάτωση σε σχέση με την ανεκτικότητα στην ασυγχρονία, δηλαδή όταν ακουστικά και οπτικά τμήματα πληροφορίας που ανήκουν στο ίδιο γεγονός είναι ασύγχρονα μεταξύ τους. Τα ευρήματα δείχνουν ότι η ενσωμάτωση μπορεί να 'ανεχθεί' ένα συγκεκριμένο ποσοστό ασυγχρονίας. Επίσης απεικονίζουν πως η ασυγχρονία επηρεάζει τους χρόνους απόκρισης, όπου όσο μεγαλύτερη είναι η ασυγχρονία, τόσο πέφτει η επίδοση και αυξάνει ο χρόνος απόκρισης. Τα παραπάνω απεικονίζονται στο Σχήμα 6.5α. Οι συγγραφείς σχολιάζουν επίσης την ελαφριά ασσυμετρία της καμπύλης προς όφελος της περίπτωσης που το ηχητικό ερέθισμα ακολουθεί το οπτικό, υποδεικνύοντας ότι ο στόχος γίνεται περισσότερο προεξέχων όταν ο ήχος ακολουθεί παρά όταν προηγείται της αλλαγής χρώματος του. Στόχος μας είναι να αναπαράξουμε αυτό το πείραμα με παρόμοιο τρόπο όπως και το προηγούμενο. Τα αποτελέσματα εμφανίζονται στο Σχήμα 6.5.

Όσον αφορά το AUC, παρατηρούμε ότι ο στόχος εξακολουθεί να είναι το ίδιο προεξέχων σε όλες τις περιπτώσεις. Παρ΄όλα αυτά, για τα NSS και CC παρατηρούμε ότι τα αποτελέσματά μας είναι συνεπή ως προς τα συμπεριφορικά. Η εμφάνεια αυξάνεται όταν ο συγχρονισμός μεταξύ των δύο ροών πληροφορίας αυξάνεται και αντίστροφα. Η κορυφή της καμπύλης βρίσκεται στο σημείο του απόλυτου συγχρονισμού τους. Επιπρόσθετα, η εμφάνεια του στόχου είναι υψηλότερη όταν ο τόνος ακολουθεί την αλλαγή χρώματος και όχι το ανάποδο για τα σχήματα της απευθείας σύμμειξης και σύμμειξης με αμοιβαία πληροφορία για το ίδιο ποσοστό ασυγχρονίας, παρουσιάζοντας έτσι παρόμοια συμπεριφορά με τα συμπεριφορικά πειράματα. Η τεχνική σύμμειξης με ετεροσυσχέτιση δε μπορεί να απεικονίσει το ίδιο καλά τα συμπεριφορικά ευρήματα αλλά επιτυγχάνει υψηλότερη απόλυτη τιμή εμφάνειας από τις υπόλοιπες.

Πείραμα μεγέθους συνόλου "sine vs square"

Αντίστοιχα με το πρώτο "pip and pop" πείραμα, στο [3] τα αποτελέσματα υποδεικνύουν ότι όταν δεν υπάρχουν ηχητικοί τόνοι, ο χρόνος απόκρισης αυξάνεται αναλογικά με τον αριθμό του συνόλου των γραμμών, δηλαδή η εμφάνεια του στόχου μειώνεται. Το ίδιο συμβαίνει και όταν η φωτεινότητα και ο τόνος αλλάζουν με ημιτονοειδή διαμόρφωση, δηλαδή σταδιακά, ακόμα κι αν υπάρχει συγχρονισμένος τόνος. Οι συγγραφείς αποδίδουν αυτό το γεγονός στη μη πραγματοποίηση οπτικοακουστικής ενσωμάτωσης. Αντιθέτως, όταν ένας τόνος συνοδεύει την αλλαγή φωτεινότητας του στόχου αλλά και τα δύο αλλάζουν με τετραγωνική διαμόρφωση, δηλαδή απότομα, το μέγεθος του συνόλου δεν παίζει ιδιαίτερο ρόλο. Το ίδιο συμβαίνει και όταν ο τόνος έχει διαφορά φάσης 180° από τη φωτεινότητα. Τα πρωτότυπα σχήματα από το [3] βρίσκονται στα Σχήματα 6.6α και 6.7α: Εξετάζουμε για άλλη μία φορά αν το μοντέλο μας παρουσιάζει αντίστοιχη συμπεριφορά. Στο Σχήμα 6.6 παρουσιάζονται τα αποτελέσματα για τη μετρική NSS (παρόμοια συμπεριφορά με την CC) για την ημιτονοειδή διαμόρφωση ενώ στο Σχήμα 6.7 τα αντίστοιχα αποτελέσματα για την τετραγωνική διαμόρφωση.

Και στις δύο περιπτώσεις, τα αποτελέσματα μας έχουν παρόμοια συμπεριφορά με τα συμπεριφορικά πειράματα. Όταν δεν υπάρχει ακουστικός τόνος, η εμφάνεια μειώνεται με την αύξηση του μεγέθους του συνόλου των γραμμών σε όλες τις περιπτώσεις. Το ίδιο ακριβώς συμβαίνει

και με την ημιτονοειδή διαμόρφωση, είτε υπάρχει συγχρονισμένος τόνος είτε τόνος με διαφορά φάσης 180° από την αλλαγή της φωτεινότητας. Αντιθέτως, στην περίπτωση της τετραγωνικής διαμόρφωσης, παρατηρούμε ότι η εμφάνεια παραμένει σχεδόν σταθερή και ανεξάρτητη του μεγέθους του συνόλου, για όλες τις περιπτώσεις, όπως ακριβώς απεικονίζεται και στο πρωτότυπο σχήμα. Σχετικά με τις τεχνικές σύμμειξης, η απευθείας σύμμειξη φαίνεται να δίνει καλύτερα αποτελέσματα από τις άλλες δύο στην περίπτωση της τετραγωνικής διαμόρφωσης.

6.3.3 Πειράματα με δεδομένα οφθαλμοκινητικής παρακολούθησης

Για να αξιολογηθούν οι διάφορες τεχνικές σύμμειξης που αναπτύχθηκαν, πειραματιζόμαστε με το μόνο διαθέσιμο οπτικοακουστικό μοντέλο εμφάνειας Min et al. [186] αλλά και διάφορα state-of-the-art οπτικά μοντέλα εμφάνειας που είναι δημόσια διαθέσιμα και τα οποία περιέχουν τόσο χωρικές όσο και χρονικές συνιστώσες. Διαλέγουμε ένα μοντέλο από καθεμία από τις βασικές προσεγγίσεις για οπτική εμφάνεια: ένα βιολογικά εμπνευσμένο μοντέλο, το Itti et al. [109, 116], ένα από τα μοντέλα που βασίζονται στη θεωρία πληροφορίας, το SDSR [136], ένα από αυτά που εκτιμούν την εμφάνεια στο πεδίο της συχνότητας, το PQFT [142], καθώς και ένα μοντέλο βαθιάς εκμάθησης που αναπτύχθηκε βασισμένο στο [152].

Για το Itti et al. μοντέλο, η χωρική συνιστώσα περιλαμβάνει τα χαρακτηριστικά χρώματος, έντασης και προσανατολισμού, ενώ η χρονική το τρεμόπαιγμα και την κίνηση. Για το μοντέλο SDSR, ο διαθέσιμος κώδικας περιλαμβάνει ένα χωρικό και ένα χρονικό μοντέλο ξεχωριστά, οπότε για να αποκτήσουμε ένα χωρο-χρονικό οπτικό μοντέλο προσθέτουμε απλά το χωρικό και χρονικό χάρτη εμφάνειας. Όσον αφορά το μοντέλο PQFT, από την αναπαράσταση της εικόνας με quaternia χρησιμοποιούμε την κίνηση σαν τη χρονική συνιστώσα και δύο κανάλια χρώματος και ένα έντασης σαν τη στατική συνιστώσα. Έπειτα, υπολογίζουμε το χωρο-χρονικό χάρτη εμφάνειας χρησιμοποιώντας τον Quaternion Fourier μετασχηματισμό, όπως στο [142]. Για το μοντέλο βαθιάς εκμάθησης, χρησιμοποιείται μια υβριδική προσέγγιση που περιλαμβάνει ένα συνελικτικό νευρωνικό δίκτυο (CNN) για τη χωρική εμφάνεια και εκτίμηση της οπτικής ροής για τη χρονική εμφάνεια. Για τη στατική συνιστώσα χρησιμοποιήθηκε το μοντέλο βαθιάς εκμάθησης από το [152] που είναι δημόσια διαθέσιμο. Ως χρονική συνιστώσα, υπολογίστηκαν χάρτες οπτικής ροής σύμφωνα με το [264], οι οποίοι βασίζονται στον αλγόριθμο οπτικής ροής TVL1 [265]. Εφαρμόζεται ένα φίλτρο χρονικά μετακινούμενου μέσου όρου σε 10 διαδοχικά καρέ οπτικής ροής ώστε να αφαιρεθεί ο θόρυβος από τις εκτιμήσεις και στις δύο κατευθύνσεις x και y ανεξάρτητα και τελικά εφαρμόζεται και ένα φίλτρο Difference-of-Gaussians (DoG) στο πλάτος της οπτικής ροής [266]. Ο τελικός χωρο-χρονικός χάρτης εμφάνειας προκύπτει αθροίζοντας τους δύο κανονικοποιημένους χάρτες.

Σχετικά με την ακουστική εμφάνεια, η διαθεσιμότητα των δημόσια διαθέσιμων μοντέλων είναι πολύ περιορισμένη. Επιπλέον, το πρόβλημα με το οποίο ασχολούμαστε είναι κατά κύριο λόγο ένα οπτικό πρόβλημα, στο οποίο επιχειρούμε να εξετάσουμε πόσο η προσθήκη της ακουστικής πληροφορίας επηρεάζει την οπτική. Επομένως, μόνο το βιολογικα εμπνευσμένο μοντέλο του Kayser et al. [168] χρησιμοποιείται εδώ, και τα χαρακτηριστικά που υπολογίζονται είναι η ένταση, η χρονική αντίθεση και η συχνοτική αντίθεση.

Η αξιολόγηση των τριών διαφορετικών σχημάτων σύμμειξης, εφαρμοσμένα για σύμμειξη ακουστικής εμφάνειας με 4 διαφορετικά χωρο-χρονικά μοντέλα οπτικής εμφάνειας και η σύγκριση με το οπτικοακουστικό μοντέλο Min et al. [186] πραγματοποιούνται σε 6 διαφορετικές βάσεις δεδομένων: τη DIEM [267], την AVAD [186] τις Coutrot1, Coutrot2 [181, 183], τη SumMe [246, 268] και την ETMD [148, 268], οι οποίες περιέχουν αρκετά περίπλοκα και δύσκολα ερεθίσματα. Στην αξιολόγηση αυτή που πραγματοποιούμε, δε στοχεύουμε στην πλήρη βελτιστοποίηση των μοντέλων ώστε να αποκτήσουμε τα καλύτερα δυνατά αποτελέσματα σε απόλυτους αριθμούς, αλλά κυρίως στο να διερευνήσουμε αν η επίδοση των οπτικών μοντέλων αυξάνεται και βελτιώνεται με την σύμμειξη της ακουστικής πληροφορίας. Επομένως, είναι περισσότερο μια συγκριτική μελέτη μεταξύ οπτικών και οπτικοακουστικών συνδυασμών.

Βάση δεδομένων DIEM

Η βάση δεδομένων DIEM [267] περιέχει 84 βίντεο από διάφορα είδη, που έχουν αποκτηθεί από δημόσια προσβάσιμα σημεία και περιλαμβάνουν διαφημιστικά σποτ, ντοκιμαντέρ, περιλήψεις ταινιών, μουσικά βίντεο κ.α. Επομένως η πλειοψηφία των βίντεο σε αυτή τη βάση δεν περιέχει φυσικό ήχο που να αντιστοιχεί στο οπτικό ερέθισμα που προβάλλεται, διότι στα περισσότερα βίντεο ο ήχος είναι είτε μουσική, είτε κάποια ομιλία τύπου αφήγησης, οπότε η βάση αυτή παρουσιάζει δύσκολες συνθήκες. Καταγράφηκαν οι οφθαλμοκινήσεις από 42 άτομα για κάθε βίντεο, μέσω ενός Eyelink eye-tracker, με τυχαία σειρά αναπαραγωγής των βίντεο και με ήχο. Γίνεται αξιολόγηση των διάφορων μοντέλων σε αυτή τη βάση και συγκρίνονται οπτικά μοντέλα με οπτικοακουστικά. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 6.1. Στην περίπτωση των μοντέλων Itti et al. και SDSR, τα οπτικοακουστικά μοντέλα έχουν καλύτερη επίδοση από τα οπτικά για όλες τις μετρικές και τα σχήματα σύμμειξης, ενώ τα βαθιά νευρωνικά δίκτυα (deep models) δίνουν το υψηλότερο αποτέλεσμα εμφάνειας για το σχήμα της απευθείας σύμμειξης.

Βάση δεδομένων AVAD

Εφαρμόζουμε τα μοντέλα μας στην βάση δεδομένων AVAD [186], η οποία περιέχει 45 μικρά βίντεο διάρκειας 5-10sec με διάφορες οπτικοακουστικές σκηνές, π.χ. κάποιον που χορεύει, που παίζει κιθάρα, πουλιά που κελαηδούν, κλπ. Η πλειοψηφία των βίντεο περιέχουν μουσική ή άλλον ατμοσφαιρικό ήχο στο υπόβαθρο, αλλά περιέχουν και έναν κυρίαρχο διακριτό ήχο που αντιστοιχεί σε κάποιο οπτικό γεγονός. Ένα άλλο σημαντικό χαρακτηριστικό της συγκεκριμένης βάσης είναι ότι το κοινό οπτικοακουστικό γεγονός είναι παρόν και στο προσκήνιο κατά τη διάρκεια ολόκληρου του βίντεο. Περιέχονται οπτικοακουστικά δεδομένα οφθαλμοκινητικής παρακολούθησης από 16 άτομα. Για τη συγκεκριμένη βάση έχει επαναξιολογηθεί το μοντέλο Min et al.(SR) [186] με το δικό μας πλαίσιο αξιολόγησης. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 6.2.

Τα οπτικοακουστικά μοντέλα ξεπερνούν τα οπτικά για όλους σχεδόν τους συνδυασμούς και τις μετρικές, εκτός από τα βαθιά νευρωνικά δίκτυα, όπου τα οπτικά μοντέλα έχουν καλύτερη επίδοση από τα οπτικοακουστικά για τις δύο πρώτες μετρικές. Ξεπερνούν επίσης και το μοντέλο Min et al.(SR) [186], το οποίο επιτυγχάνει τη δεύτερη καλύτερη επίδοση σε σχέση

	Model / Fusion scheme	CC	NSS	AUCs
Itti_V	[116] / -	0.195	1.121	0.507
Itti_AV1	[116] / Direct	0.199	1.150	0.588
Itti_AV2	[116] / CC	0.196	1.127	0.521
Itti_AV3	[116] / MI	0.196	1.128	0.511
PQFT_V	[142] / -	0.119	0.725	0.555
PQFT_AV1	[142] / Direct	0.118	0.713	0.554
PQFT_AV2	[142] / CC	0.119	0.717	0.552
PQFT_AV3	[142] / MI	0.118	0.711	0.554
SDSR_V	[136] / -	0.088	0.524	0.542
SDSR_AV1	[136] / Direct	0.089	0.528	0.542
SDSR_AV2	[136] / CC	0.108	0.642	0.556
SDSR_AV3	[136] / MI	0.108	0.645	0.559
Deep_V	Modif.[152] / -	0.265	1.563	0.636
Deep_AV1	Modif.[152] / Direct	**0.270**	**1.585**	**0.636**
Deep_AV2	Modif.[152] / CC	0.243	1.4192	0.611
Deep_AV3	Modif.[152] / MI	0.258	1.518	0.633
Min et al. (SR)	[186, 185]	0.121	0.722	0.593

Πίνακας 6.1: Τα αποτελέσματα για τα διάφορα μοντέλα και σχήματα σύμμειξης (με ακρωνύμια και σύντομη περιγραφή) για τη βάση δεδομένων DIEM.

με τις μετρικές CC και NSS. Αυτό ίσως συμβαίνει διότι τα βαθιά νευρωνικά δίκτυα γενικά μαθαίνουν καλά τη σημασιολογική πληροφορία, όπως τα πρόσωπα, τα μουσικά όργανα, κλπ., ειδικά όταν αυτά παρουσιάζονται στο κέντρο μιας εικόνας. Τα συγκεκριμένο βίντεο είναι πολύ μικρά σε διάρκεια και το οπτικοακουστικό γεγονός παρουσιάζεται στο κέντρο της εικόνας χωρίς κάποια μετάβαση από το οπτικό στο οπτικοακουστικό γεγονός. Επομένως, πιθανώς να μην υπάρχει κάτι επιπλέον που μπορεί να τονιστεί από την ακουστική εμφάνεια που να μην έχει ήδη συμπεριληφθεί στην οπτική εμφάνεια όπως προκύπτει από τα βαθιά νευρωνικά δίκτυα. Όσον αφορά το AUC που είναι πιο εύρωστη μετρική στο center bias, υπάρχει μία ελαφριά βελτίωση.

Βάσεις δεδομένων Coutrot

Εφαρμόζουμε τα ίδια μοντέλα στις βάσεις δεδομένων Coutrot [181, 183] που είναι επίσης δημόσια διαθέσιμες: η Coutrot1 περιέχει 60 βίντεο με δυναμικές σκηνές χωρισμένες σε 4 κατηγορίες: ένα κινούμενο αντικείμενο, πολλαπλά κινούμενα αντικείμενα, τοπία και πρόσωπα. Καταγράφηκαν δεδομένα οφθαλμοκινητικής παρακολούθησης από 72 άτομα. Η Coutrot2 περιλαμβάνει 15 βίντεο τεσσάρων ατόμων που πραγματοποιούν meeting και τα αντίστοιχα δεδομένα οφθαλμοκινητικής παρακολούθησης από 40 άτομα. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 6.3.

Αντίθετα από τη βάση DIEM, η πλειοψηφία των βίντεο σε αυτή τη βάση περιέχει πραγματικά οπτικοακουστικά γεγονότα, όπου τόσο η οπτική όσο και η ακουστική πληροφορία προέρχονται από ίδια γεγονότα. Επομένως, αναμένουμε τα αποτελέσματα για τους οπτικοα-

	CC	NSS	AUCs
Itti_V	0.154	1.436	0.529
Itti_AV1	0.172	1.627	0.542
Itti_AV2	0.161	1.502	0.532
Itti_AV3	0.154	1.434	0.528
PQFT_V	0.093	0.886	0.527
PQFT_AV1	0.095	0.911	0.527
PQFT_AV2	0.095	0.909	0.525
PQFT_AV3	0.096	0.920	0.528
SDSR_V	0.098	0.922	0.521
SDSR_AV1	0.099	0.929	0.522
SDSR_AV2	0.117	1.091	0.526
SDSR_AV3	0.115	1.072	0.526
Deep_V	**0.199**	**1.893**	0.552
Deep_AV1	0.194	1.844	**0.553**
Deep_AV2	0.192	1.830	0.551
Deep_AV3	0.196	1.859	0.552
Min et al. (SR)*	0.174	1.652	0.550

Πίνακας 6.2: Τα αποτελέσματα για τα διάφορα μοντέλα και σχήματα σύμμειξης για τη βάση AVAD (* επαναξιολόγηση με το παρόν πλαίσιο αξιολόγησης)

κουστικούς συνδυασμούς να είναι καλύτερα από τα οπτικά, κάτι που πράγματι επιβεβαιώνεται για όλα τα μοντέλα και σχεδόν όλες τις μετρικές. Και στην Coutrot1 αλλά και στην Coutrot2 το καλύτερο αποτέλεσμα επιτυγχάνεται από τα βαθιά μοντέλα με απευθείας σύμμειξη.

Βάση δεδομένων SumMe

Η βάση δεδομένων SumMe [246, 268] περιέχει 25 βίντεο χρηστών καθώς και τις περιλήψεις τους που έχουν πραγματοποιηθεί από διάφορους ανθρώπους μέσα σε ένα ελεγχόμενο ψυχολογικό πείραμα. Όπως αναφέρθηκε προηγουμένως, συλλέξαμε δεδομένα οφθαλμοκινητικής παρακολούθησης, τα οποία χρησιμοποιούμε για την αξιολόγηση. Τα αποτελέσματα απεικονίζονται στον Πίνακα 6.4. Όπως και στην περίπτωση της DIEM, και σε αυτή τη βάση η πλειοψηφία των βίντεο περιέχει τεχνητό ήχο, π.χ. μουσική, και όχι πραγματικά οπτικοακουστικά γεγονότα.

Ενδεικτικά καρέ με τα αντίστοιχα δεδομένα οφθαλμοκινητικής παρακολούθησης και τους προκύπτοντες οπτικούς και οπτικοακουστικούς χάρτες εμφάνειας χρησιμοποιώντας τα βαθιά νευρωνικά μοντέλα απεικονίζονται στο αριστερό μέρος του Σχήματος 6.8. Οι συμμετέχοντες έχουν προσέξει περισσότερο το παιδί που τρέχει παρά το αυτοκίνητο, κάτι που αντικατοπτρίζεται καλύτερα στους οπτικοακουστικούς χάρτες σε σχέση με τους οπτικούς.

Παρ΄ όλα αυτά, οι οπτικοακουστικοί συνδυασμοί επιτυγχάνουν καλύτερα αποτελέσματα από τα οπτικά μοντέλα για τις περισσότερες περιπτώσεις: Για το μοντέλο Itti et al. το καλύτερο αποτέλεσμα επιτυγχάνεται με τη σύμμειξη ετεροσυσχέτισης, ενώ για το SDSR και το PQFT η απευθείας σύμμειξη και η αμοιβαία πληροφορία δίνουν τα καλύτερα αποτελέσματα

	Coutrot1			Coutrot2		
	CC	NSS	AUCs	CC	NSS	AUCs
Itti_V	0.181	1.015	0.544	0.164	1.362	0.593
Itti_AV1	0.183	1.062	0.559	0.239	2.005	0.632
Itti_AV2	0.187	1.055	0.548	0.177	1.475	0.600
Itti_AV3	0.182	1.013	0.543	0.166	1.373	0.593
PQFT_V	0.128	0.845	0.543	0.162	1.584	0.588
PQFT_AV1	0.130	0.853	0.544	0.166	1.615	0.588
PQFT_AV2	0.128	0.845	0.542	0.159	1.550	0.582
PQFT_AV3	0.131	0.862	0.544	0.169	1.640	0.588
SDSR_V	0.115	0.674	0.539	0.072	0.622	0.560
SDSR_AV1	0.116	0.677	0.540	0.073	0.627	0.561
SDSR_AV2	0.128	0.748	0.542	0.096	0.829	0.586
SDSR_AV3	0.128	0.744	0.544	0.098	0.839	0.589
Deep_V	0.266	1.585	0.586	0.248	2.094	0.624
Deep_AV1	**0.270**	**1.650**	**0.591**	**0.253**	**2.157**	**0.627**
Deep_AV2	0.264	1.598	0.588	0.249	2.113	0.623
Deep_AV3	0.266	1.595	0.589	0.251	2.132	0.626
Min et al. (SR)	0.115	0.666	0.550	0.127	1.043	0.605

Πίνακας 6.3: Τα αποτελέσματα για τα διάφορα μοντέλα και σχήματα σύμμειξης για τις βάσεις Coutrot

αντίστοιχα. Σε σχέση με τα βαθιά νευρωνικά μοντέλα, μόνο η απυθείας σύμμειξη δίνει ελαφρώς βελτιωμένα αποτελέσματα σε σύγκριση με το οπτικά-μόνο μοντέλο για το $AUCs$, κάτι το οποίο έχει πρόσφατα παρατηρηθεί και για τη σύγκριση των χωρικών-μόνο με τα χωρο-χρονικά βαθιά μοντέλα εμφάνειας [269].

Βάση δεδομένων ETMD

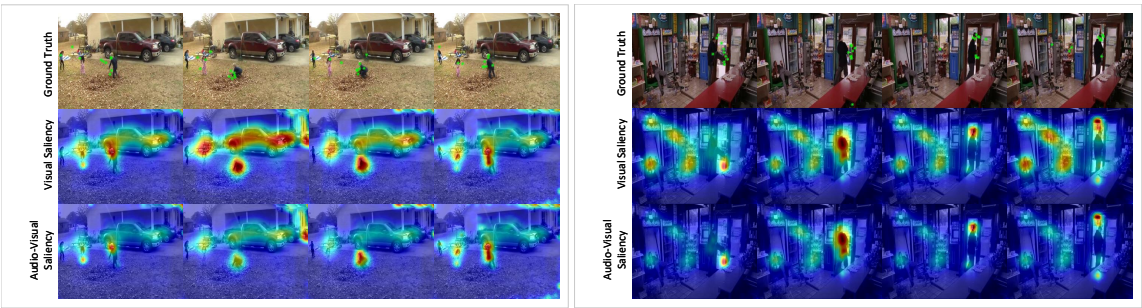
Η τελευταία βάση, η ETMD [148, 268] περιέχει 12 αποσπάσματα ταινιών από 6 ταινίες του Hollywood. Τα αποσπάσματα αυτά περιέχουν υψηλού επιπέδου πληροφορία με περίπλοκο σημασιολογικό περιεχόμενο.

Ενδεικτικά καρέ με τα αντίστοιχα δεδομένα οφθαλμοκινητικής παρακολούθησης και τους προκύπτοντες οπτικούς και οπτικοακουστικούς χάρτες εμφάνειας χρησιμοποιώντας τα βαθιά νευρωνικά μοντέλα απεικονίζονται στο δεξί μέρος του Σχήματος 6.8. Οι συμμετέχοντες προσέχουν περισσότερο την πόρτα που κλείνει, κάτι που αντικατοπτρίζεται καλύτερα στους οπτικοακουστικούς χάρτες σε σχέση με τους οπτικούς.

Τα αποτελέσματα παρουσιάζονται στον Πίνακα 6.4. Τα οπτικοακουστικά μοντέλα έχουν καλύτερη επίδοση από τα οπτικά για όλες τις περιπτώσεις και μετρικές, εκτός από τα βαθιά νευρωνικά μοντέλα, όπου τα οπτικοακουστικά μοντέλα είναι απλά συγκρίσιμα σε επίδοση με το οπτικό-μόνο. Αυτό μπορεί να συμβαίνει λόγω του ότι τα εν λόγω βίντεο περιέχουν πολύ μεγάλη σημασιολογική πληροφορία η οποία έχει ήδη εκμαθηθεί από το χωρικό-μόνο μοντέλο (κατά τη διάρκεια της εκπαίδευσής του σε μεγάλες βάσεις από εικόνες). Τα αποτελέσματα

	SumMe			ETMD		
	CC	NSS	AUCs	CC	NSS	AUCs
tti_V	0.157	1.290	0.628	0.166	1.216	0.617
Itti_AV1	0.157	1.289	0.628	0.167	1.221	0.619
Itti_AV2	0.156	1.294	0.631	0.166	1.218	0.620
Itti_AV3	0.157	1.289	0.629	0.166	1.217	0.618
PQFT_V	0.095	0.874	0.586	0.101	0.816	0.577
PQFT_AV1	0.096	0.877	0.587	0.102	0.820	0.577
PQFT_AV2	0.096	0.846	0.589	0.099	0.800	0.571
PQFT_AV3	0.072	0.667	0.564	0.102	0.823	0.576
SDSR_V	0.092	0.747	0.591	0.066	0.484	0.555
SDSR_AV1	0.093	0.751	0.593	0.067	0.490	0.557
SDSR_AV2	0.098	0.805	0.605	0.083	0.619	0.576
SDSR_AV3	0.099	0.809	0.607	0.084	0.623	0.580
Deep_V	**0.194**	**1.595**	0.653	**0.254**	**1.880**	0.703
Deep_AV1	**0.194**	1.592	**0.654**	0.253	1.868	**0.704**
Deep_AV2	0.175	1.482	0.653	0.218	1.627	0.694
Deep_AV3	0.178	1.482	0.656	0.223	1.653	0.696
Min et al. (SR)	0.080	0.650	0.605	0.117	0.857	0.634

Πίνακας 6.4: Τα αποτελέσματα των διαφόρων μοντέλων και σχημάτων σύμμειξης για τη βάση SumMe και τη βάση ETMD.

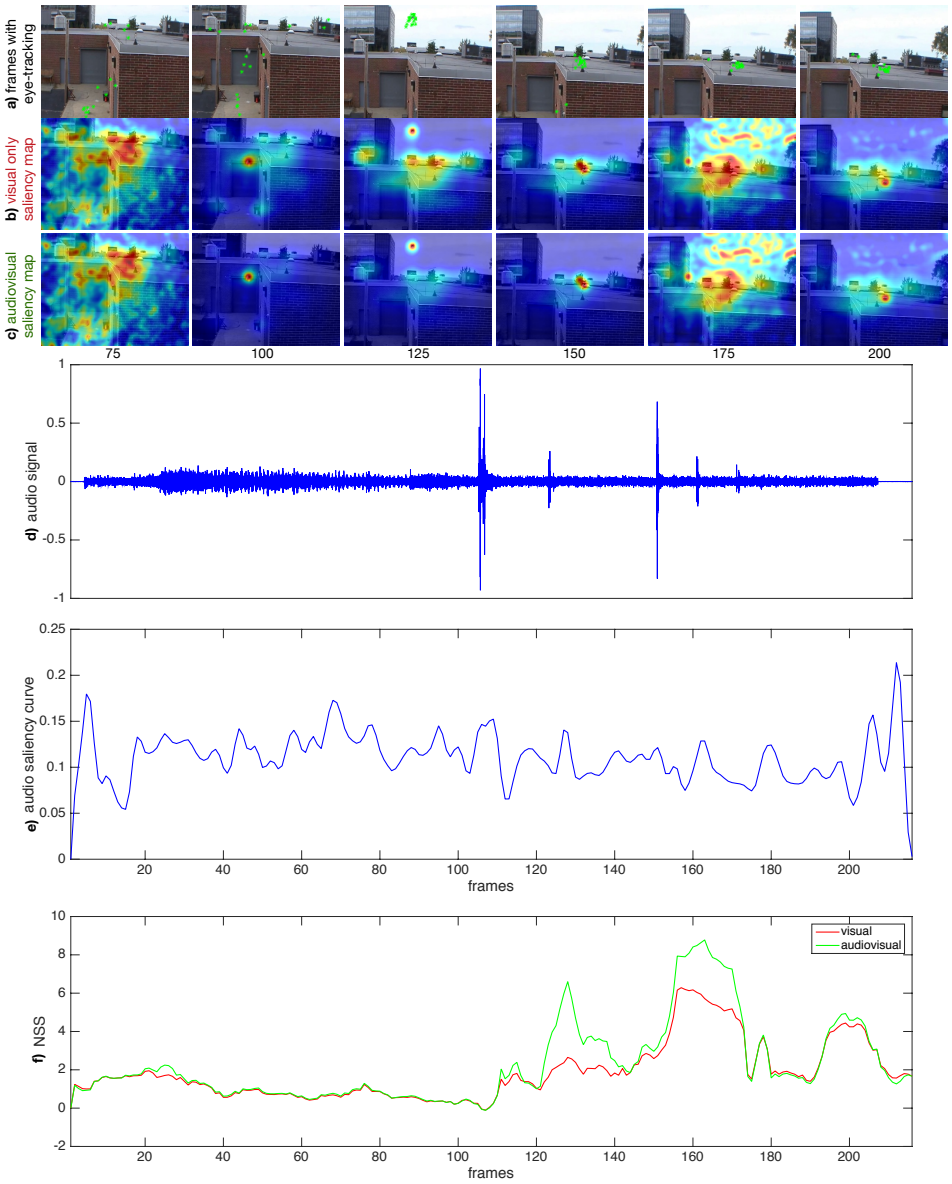


Σχήμα 6.8: Παράδειγμα συνεχόμενων καρέ όπου έχει γίνει υπέρθεση των δεδομένων οφθαλμοκινητικής παρακολούθησης που έχουν συλλεχθεί, για τη SumMe (αριστερά) και την ETMD (δεξιά). Η δεύτερη και τρίτη γραμμή απεικονίζουν τους αντίστοιχους οπτικούς και οπτικοακουστικούς χάρτες εμφάνειας.

επίσης υποδεικνύουν ότι για την οπτικοακουστική ενσωμάτωση χρειάζεται επίσης top-down πληροφορία, γεγονός που αναδεικνύει επίσης την καταλληλότητα των οφθαλμοκινητικών δεδομένων που συλλέξαμε για περαιτέρω έρευνα.

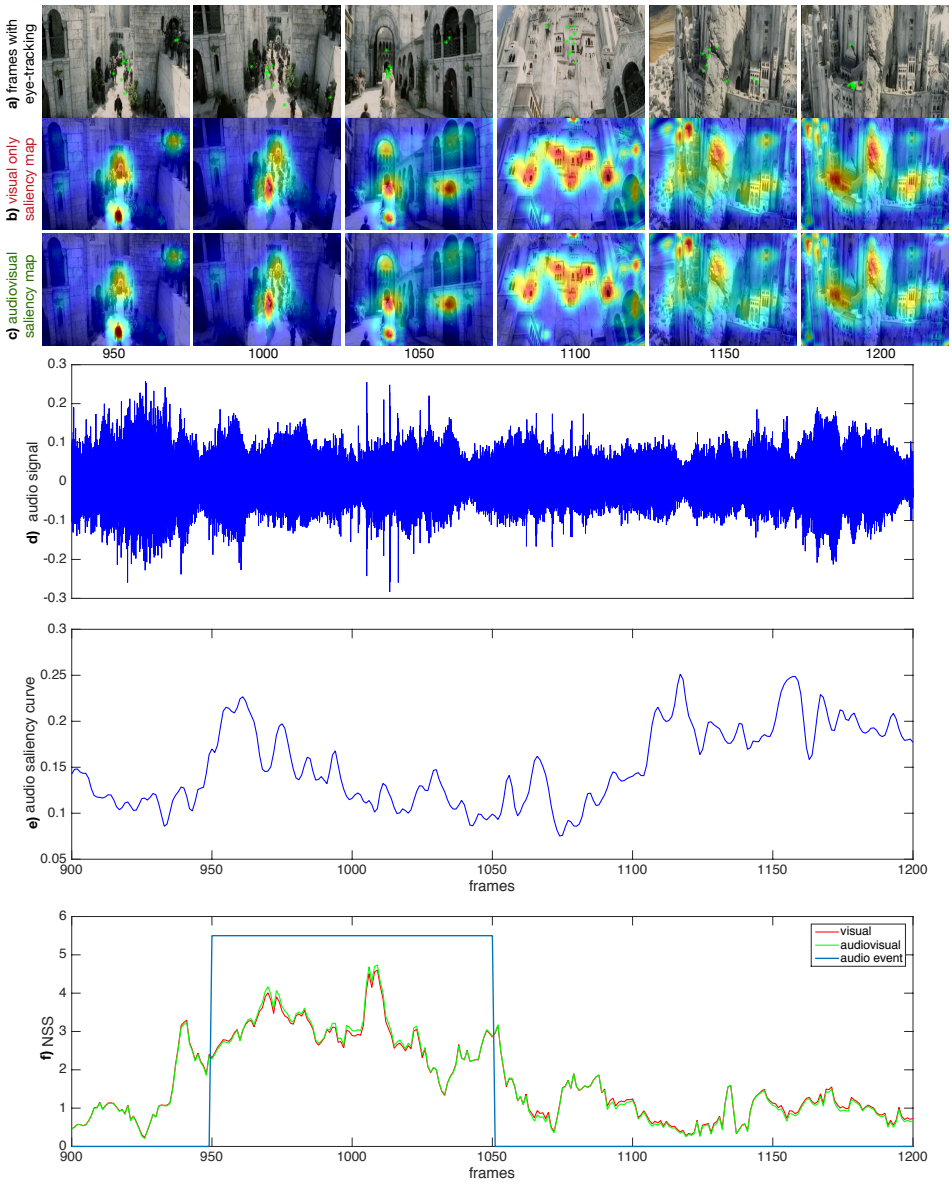
Ανάλυση και συζήτηση

Ο στόχος μας είναι να αναλύσουμε την επίδοση των διαφόρων μοντέλων σε όλα τα σύνολα δεδομένων μας και να διαπιστώσουμε σε ποιες περιπτώσεις και κάτω από ποιες συνθήκες η ενσωμάτωση της ακουστικής εμφάνειας βελτιώνει πράγματι τη μοντελοποίηση της ανθρώπι-



Σχήμα 6.9: Παράδειγμα από τη βάση δεδομένων AVAD. a) Παρουσιάζονται τα πρωτότυπα καρέ μαζί με τα δεδομένα οπτικοακουστικής παρακολούθησης μαζί με b) τους χάρτες οπτικής και c) οπτικοακουστικής εμφάνειας. d) Ακουστική κυματομορφή και e) η οπτική καμπύλη εμφάνειας. f) Απεικονίζεται η οπτική μόνο (κόκκινο χρώμα) και η οπτικοακουστική (πράσινο χρώμα) καμπύλη εμφάνειας ως προς τη μετρική NSS. Στην αρχή του βίντεο υπάρχει μόνο ακουστικός θόρυβος, που δε σχετίζεται με το οπτικό περιεχόμενο. Το πραγματικό ακουστικό γεγονός εμφανίζεται περί τα μέσα του βίντεο. Ως τότε, η οπτικοακουστική και η οπτική εμφάνεια είναι περίπου ίδιες όπως φαίνεται από τις NSS καμπύλες, ενώ όταν εμφανίζεται το ακουστικό γεγονός, η οπτικοακουστική καμπύλη ξεπερνά αισθητά την οπτική.

νης προσοχής. Συνολικά για όλες τις βάσεις δεδομένων, τα αποτελέσματα στα πιο σύνθετα ερεθίσματα υποδεικνύουν ότι τα οπτικοακουστικά μοντέλα εμφάνειας μπορούν πράγματι να βελτιώσουν τα αποτελέσματα πρόβλεψης της εστίασης της προσοχής σε σχέση με τα οπτικά μόνο μοντέλα. Σε κάποιες περιπτώσεις η βελτίωση είναι πολύ μικρή, π.χ. για το οπτικοα-



Σχήμα 6.10: Παράδειγμα από τη βάση δεδομένων ETMD. a) Παρουσιάζονται τα πρωτότυπα καρέ μαζί με τα δεδομένα οπτικοακουστικής παρακολούθησης μαζί με b) τους χάρτες οπτικής και c) οπτικοακουστικής εμφάνειας. d) Ακουστική κυματομορφή και e) η οπτική καμπύλη εμφάνειας. f) Απεικονίζεται η οπτική μόνο (κόκκινο χρώμα) και η οπτικοακουστική (πράσινο χρώμα) καμπύλη εμφάνειας ως προς τη μετρική NSS. Επίσης το μπλε ορθογώνιο σηματοδοτεί το οπτικοακουστικό γεγονός (καλπασμός αλόγου). Παρόλο που οι διαφορές οπτικοακουστικής και οπτικής εμφάνειας είναι μικρές εδώ, μπορούμε να παρατηρήσουμε ότι κατά τη διάρκεια του ακουστικού γεγονότος, η τιμή της μετρικής NSS είναι λίγο υψηλότερη για την οπτικοακουστική καμπύλη, ενώ πριν και μετά το γεγονός είναι σχεδόν ίδιες ή εναλλάσσονται ως προς το ποια είναι καλύτερη ή χειρότερη.

κουστικό PQFT στις βάσεις SumMe και ETMD, αλλά σε κάποιες άλλες περιπτώσεις είναι σημαντική. Στα Σχήματα 6.9, 6.10 παρουσιάζονται δύο παραδείγματα από δύο διαφορετικές βάσεις δεδομένων, την AVAD και την ETMD, όπου στις τρεις πρώτες σειρές απεικονίζονται

ενδεικτικά καρέ, ομοιόμορφα δειγματοληπτημένα από όλο το βίντεο με τα αντίστοιχα δεδομένα οφθαλμοκινητικής παρακολούθησης και οι αντίστοιχοι οπτικοί και οπτικοακουστικοί χάρτες εμφάνειας για τα συγκεκριμένα καρέ. Στην τέταρτη γραμμή βρίσκεται η ακουστική κυματομορφή για όλο το βίντεο ενώ στην πέμπτη η αντίστοιχη καμπύλη ακουστικής εμφάνειας που προκύπτει. Τέλος, στην έκτη γραμμή παρουσιάζονται οι NSS καμπύλες κατά τη χρονική διάρκεια του βίντεο για την οπτική περίπτωση (κόκκινο χρώμα) και για την οπτικοακουστική (πράσινο χρώμα), όπως αυτές προέκυψαν από τα μοντέλα Deep_V και Deep_AV1 models.

Τα δύο αυτά σχήματα μπορούν να δώσουν μια πιο βαθιά ματιά στο πώς η εμφάνεια εξελίσσεται κατά τη διάρκεια του χρόνου και πως η ακουστική εμφάνεια συνεισφέρει στη συνολική εμφάνεια σύμφωνα με τη μοντελοποίηση που έχει γίνει. Το πώς και πότε η ακουστική εμφάνεια επιδρά στην οπτική έχει μελετηθεί και παλιότερα στη βιβλιογραφία [270]. Τα σχήματα αυτά υποδεικνύουν ότι η ακουστική εμφάνεια μπορεί να ενισχύσει τη συνολική, όταν υπάρχουν πραγματικά οπτικοακουστικά γεγονότα. Για παράδειγμα, στο Σχήμα 6.9, όταν το ακουστικό γεγονός ξεκινά, παρατηρείται μία αύξηση στην επίδοση (η καμπύλη με το πράσινο χρώμα) και η οπτικοακουστική εμφάνεια δείχνει να μοντελοποιεί καλύτερα την ανθρώπινη προσοχή σε σχέση με την οπτική, κάτι που ανταναχλάται επίσης στα δεδομένα οφθαλμοκινητικής παρακολούθησης. Πριν ή μετά το ακουστικό γεγονός, η ακουστική εμφάνεια δεν φαίνεται να ενισχύει την οπτική, αλλά ταυτόχρονα δε χειροτερεύει και την απόδοση, που είναι η επιθυμητή συμεριφορά του μοντέλου που αναπτύχθηκε.

Σε τέτοιες περιπτώσεις μη ύπαρξης ακουστικού γεγονότος, η οπτικοακουστική επίδοση είναι σχεδόν ίδια με την οπτική. Στο Σχήμα 6.10, όπου το ερέθισμα είναι πιο περίπλοκο και το βαθύ νευρωνικό δίκτυο έχει ήδη επιτύχει μια υψηλή επίδοση όταν ξεκινάει το σχετικό ακουστικό γεγονός (καλπασμός αλόγου), υπάρχει μία μικρή βελτίωση του οπτικοακουστικού σε σχέση με το οπτικό (πράσινη και κόκκινη καμπύλη). Αυτή η μικρή βελτίωση εξαφανίζεται πριν και μετά το ακουστικό γεγονός, ενώ σε κάποιες χρονικές στιγμές, η οπτική εμφάνεια πετυχαίνει καλύτερη επίδοση από την οπτικοακουστική.

Σε σχέση με τα διάφορα μοντέλα που αξιολογήθηκαν, η ενσωμάτωση της ακουστικής πληροφορίας στο μοντέλο Itti et al. [116] επιτυγχάνει καλύτερη επίδοση συγκρινόμενη με το οπτικό μόνο μοντέλο, για όλες τις βάσεις. Πραγματοποιήθηκαν κάποια ενδεικτικά τεστ ANOVA για να επιβεβαιωθεί η στατιστική σημαντικότητα των αποτελεσμάτων. Στη βάση Coutrot1, το ANOVA τεστ μεταξύ των μοντέλων Itti_V και Itti_AV1 έχει ως αποτέλεσμα ένα F-statistic $F = 8.267$ ($p < 0.005$) για τη μετρική CC, ενώ για την Coutrot2 το αντίστοιχο αποτέλεσμα είναι $F = 3810$ ($p < 0.001$) για τη μετρική NSS.

Συνολικά για όλες τις βάσεις, τα μοντέλα που αναπτύχθηκαν με βαθιά νευρωνικά δίκτυα πετυχαίνουν τα καλύτερα αποτελέσματα σε απόλυτους αριθμούς, ενώ η οπτικοακουστική σύμμειξη στις περισσότερες περιπτώσεις ξεπερνά την οπτική μόνο εμφάνεια. Σε πολλές περιπτώσεις το Deep_V και το Deep_AV1 έχουν πολύ κοντινή επίδοση, επομένως πραγματοποιήθηκε ένα ANOVA τεστ για να διαπιστωθεί αν οι διαφορές αυτές μεταξύ τους είναι στατιστικά σημαντικές. Για τη βάση Coutrot1 το τεστ έχει ως αποτέλεσμα $F = 9.5113$ ($p < 0.005$), 52.822 ($p < 0.001$), $F = 9.890$ ($p < 0.001$) για τις μετρικές CC, NSS, AUCs ενώ για την Coutrot2 είναι $F = 32.642$ ($p < 0.001$), $F = 43.941$ ($p < 0.001$), $F = 5.518$ ($p < 0.05$) αντίστοιχα.

Στη βάση ETMD τα αντίστοιχα αποτελέσματα είναι $F = 85.284$ ($p < 0.001$), $F = 65.350$ ($p < 0.001$) και $F = 4.726$ ($p < 0.05$).

Όσον αφορά τη σύγκριση με το state-of-the-art οπτικοακουστικό μοντέλο Min et al. [186], έχει τη δεύτερη καλύτερη επίδοση στη βάση AVAD [186], αλλά η επίδοσή του για τις μετρικές CC και NSS μειώνεται όταν το μοντέλο εφαρμόζεται σε πιο περίπλοκα ερεθίσματα ή σε ερεθίσματα μεγαλύτερης διάρκειας, όπως οι ταινίες, τα οποία μπορεί να περιέχουν πολύ λίγα πραγματικά οπτικοακουστικά γεγονότα. Παρ΄ όλα αυτά, στη μετρική AUCs, το Min et al. δίνει καλύτερα αποτελέσματα από το Itti_AV1 και έχει συγκρίσιμη απόδοση με τα βαθιά νευρωνικά μοντέλα, ειδικά στις βάσεις DIEM και ETMD.

6.3.4 Αξιολόγηση με δεδομένα fMRI

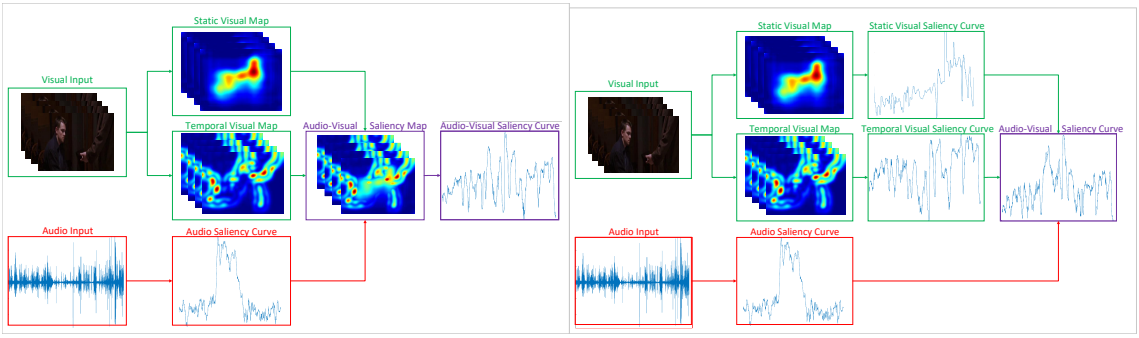
Όπως αναφέρθηκε και εισαγωγικά στην ενότητα, ο σκοπός της δημιουργίας ενός τέτοιου οπτικοακουστικού μοντέλου εμφάνειας είναι και η επαλήθευσή του με ανθρώπινα δεδομένα. Προς αυτό το σκοπό, εκτός από την επαλήθευση που πραγματοποιήθηκε τόσο με συμπεριφορικά πειράματα όσο και με δεδομένα οφθαλμοκινητικής παρακολούθησης, εδώ επιχειρείται και η επαλήθευσή του με τεχνικές απεικόνισης του εγκεφάλου όπως η λειτουργική μαγνητική απεικόνιση του εγκεφάλου, το γνωστό functional Magnetic Resonance Imaging (fMRI). Το fMRI μπορεί να λειτουργήσει ως ένα μη επεμβατικό εργαλείο επίβλεψης της νευρικής δραστηριότητας κατά τη διάρκεια εξωτερικών ερεθισμάτων και συνεπώς μπορεί να φωτίσει τη δομική και λειτουργική αρχιτεκτονική του εγκεφάλου. Πρόσφατα έχει γίνει μια μετακίνση προς πιο περίπλοκα και φυσικά ερεθίσματα, όπως πραγματικές εικόνες, αποσπάσματα βίντεο και ήχου. Η απόπειρα μελέτης τέτοιων ερεθισμάτων στοχεύει στην κατανόηση της αναπαράστασής τους στον ανθρώπινο εγκέφαλο και τελικά στη σύνδεση των χαρακτηριστικών χαμηλού επιπέδου με την υψηλού επιπέδου σημασιολογική πληροφορία που μεταδίδουν, έτσι ώστε να προταθούν και να βελτιωθούν τα υπολογιστικά μοντέλα για διάφορες εφαρμογές στην όραση υπολογιστών [271]. Η προβολή βίντεο με ταυτόχρονη καταγραφή δεδομένων fMRI προσφέρει ένα ημι-φυσικό τρόπο να συμπεράνει κανείς τους περίπλοκους μηχανισμούς που διαθέτει ο άνθρωπος για να αναπαριστά και να καταλαβαίνει τέτοια ερεθίσματα, ενώ ταυτόχρονα δημιουργεί νέες προκλήσεις για την ανάπτυξη αποτελεσματικών και συμπεριφορικά επαληθεύσιμων υπολογιστικών μοντέλων.

Εδώ επιχειρείται να γεφυρωθεί το χάσμα ανάμεσα στην τεράστια πρόοδο των υπολογιστικών μοντέλων και των νευροβιολογικών και ψυχοφυσικών ενδείξεων αναλύοντας δεδομένα fMRI. Με λίγα λόγια, ο σκοπός αυτής της υποενότητας είναι η επαλήθευση του οπτικοακουστικού μοντέλου που παρουσιάστηκε, μέσω δεδομένων fMRI. Η επαλήθευση στοχεύει στο να επιβεβαιώσει ότι το μοντέλο πράγματι μοντελοποιεί την οπτικοακουστική ανθρώπινη προσοχή όταν αυτή εκτίθεται σε οπτικοακουστικά ερεθίσματα. Ουσιαστικά ερευνάται το κατά ποσο τα δεδομένα fMRI υποδεικνύουν ενεργοποίηση σε περιοχές οι οποίες αναμένεται να ενεργοποιούνται όταν οι άνθρωποι εκτίθενται σε συγκεκριμένα ερεθίσματα.

Μία ενδιαφέρουσα ερώτηση είναι κατά πόσο οι διαφορετικοί άνθρωποι έχουν την ίδια αντίληψη για τα ερεθίσματα που τους παρουσιάζονται, καθώς επίσης και το αν οι νευρικές

αναπαραστάσεις που δημιουργούν είναι στην ουσία τους διαφορετικές ή αν μοιράζονται κάποια κοινή δομή. Εάν αποδειχτεί το τελευταίο, τότε τα δεδομένα fMRI θα μπορούσαν να χρησιμοποιηθούν ώστε να επαυξήσουν τα υπολογιστικά μοντέλα σύμφωνα με την ανθρώπινη αντίληψη. Γι αυτό το σκοπό συλλέχθηκε ένας μεγάλος αριθμός δεδομένων χρησιμοποιώντας βίντεο σαν ερεθίσματα και στα οποία συμμετείχαν αρκετοί εθελοντές.

Επειδή τα δεδομένα fMRI είναι χρονοσειρές, μετατρέπουμε το οπτικοακουστικό χωροχρονικό μοντέλο σε οπτικοακουστικό χρονικό, πηγαίνοντας από τον χάρτη εμφάνειας σε καμπύλη εμφάνειας. Στην παρούσα ενότητα επιλέγουμε σαν οπτικοακουστικό μοντέλο την παραλλαγή με τα βαθιά νευρωνικά δίκτυα που είχαν καλύτερη επίδοση και ως τρόπο σύμμειξης επιλέγουμε την απευθείας σύμμειξη.



(α′) (β′)

Σχήμα 6.11: Επισκόπηση του μοντέλου οπτικοακουστικής χρονικής εμφάνειας: a) σύμμειξη στο επίπεδο των χαρτών b) σύμμειξη στο επίπεδο των καμπυλών.

Αρχικά ο οπτικοακουστικός χάρτης εμφάνειας πρέπει να μετατραπεί σε μονοδιάστατη καμπύλη. Ο απλός χωρικός μέσος όρος για κάθε καρέ δεν είναι κατάλληλος γιατί οι χάρτες εμφάνειας μπορεί να περιέχουν αρκετές μηδενικές τιμές σε μη προεξέχουσες περιοχές, οι οποίες μπορεί να επηρεάσουν την καμπύλη εμφάνειας. Για το σκοπό αυτό πραγματοποιείται χωρικός μέσος όρος μόνο στις προεξέχουσες περιοχές του χάρτη εμφάνειας. Αρχικά ορίζουμε τον τελεστή $B : \mathbb{R}^{\mathbb{E}} \rightarrow \{0,1\}^{\mathbb{E}}$ ο οποίος μετατρέπει τον χάρτη σε δυαδική εικόνα εφαρμόζοντας το κατώφλι του Otsu [272], όπου το $\mathbb{E}$ υποδηλώνει ένα καρέ μεγέθους $m \times n$. Στη συνέχεια λαμβάνουμε την 1D καμπύλη $C(t)$, εφαρμόζοντας την αντιστοίχιση (mapping) $G : \mathbb{R}^{\mathbb{E}} \rightarrow \mathbb{R}$ στον χάρτη εμφάνειας $F(x, y, t)$:

$$C(t) = G(F(x, y, t)) = \frac{\sum_{x,y} B(F(x, y, t)) \cdot F(x, y, t)}{\sum_{x,y} B(F(x, y, t))} \tag{6.13}$$

Τέλος εφαρμόζουμε ένα median φίλτρο μήκους 151 καρέ στην καμπύλη $C(t)$ και κανονικοποιούμε τις τιμές της στο $[0, 1]$.

Όπως αναφέρθηκε και προηγουμένως, με την απευθείας σύμμειξη, ο χρονικο-ακουστικός χάρτης εμφάνειας προκύπτει ως:

$$F_{TA}(x, y, t) = F_T(x, y, t)(1 + A(t)) \tag{6.14}$$

Επομένως, η αντίστοιχη χρονικο-ακουστική καμπύλη προκύπτει ως:

$$C_{TA}(t) = (1 + C_A(t)) \cdot C_T(t), \tag{6.15}$$

όπου $C_T(t) = G(M_T(x, y, t))$.

Όσον αφορά την ενσωμάτωση της χωρικής συνιστώσας, πειραματιζόμαστε με δύο συναρτήσεις, το μέσο όρο (average) και το μέγιστο (max), οι οποίες αντιστοιχούν σε διαφορετικές προσεγγίσεις στην ενσωμάτωση χαρακτηριστικών. Με το πρώτο θέλουμε να έχουμε υψηλές τιμές εμφάνειας και στο χρονικο-ακουστικό και στο χωρικό χάρτη, ενώ με το δεύτερο αναζητούμε περιοχές που είναι προεξέχουσες τουλάχιστον σε έναν από τους δύο χάρτες. Η σύμμειξη μπορεί να πραγματοποιηθεί σε δύο διαφορετικά επίπεδα: α) το επίπεδο του χάρτη ή β) το επίπεδο της καμπύλης αφού εφαρμοστεί ο μετασχηματισμός της Εξ. (6.13).

Στην πρώτη περίπτωση το αποτέλεσμα παραμένει δισδιάστατος χάρτης εμφάνειας $F_{STA}(x, y, t)$ για τον οποίο η καμπύλη εμφάνειας υπολογίζεται ως:

$$\text{Aver.}: \quad F_{STA}(x, y, t) = (F_S(x, y, t) + F_{TA}(x, y, t))/2 \tag{6.16}$$

$$\text{Max}: \quad F_{STA}(x, y, t) = \max(F_S(x, y, t), F_{TA}(x, y, t)) \tag{6.17}$$

$$C_{STA}(t) = G(F_{STA}(x, y, t)) \tag{6.18}$$

Στη δεύτερη περίπτωση, η σύμμειξη πραγματοποιείται μεταξύ των 1D καμπυλών $C_S(t) = G(F_S(x, y, t))$, $C_{TA}(t) = G(F_{TA}(x, y, t))$:

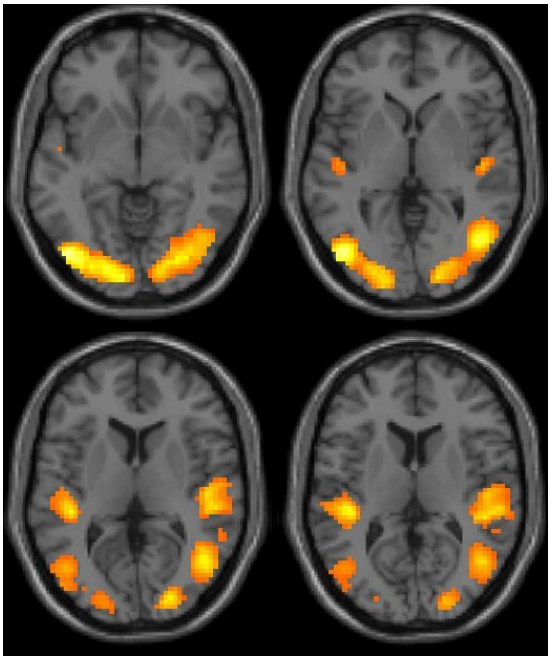
$$\text{Aver.}: \quad \tilde{C}_{STA}(t) = (C_S(t) + C_{TA}(t))/2 \tag{6.19}$$

$$\text{Max}: \quad \tilde{C}_{STA}(t) = \max(C_S(t), C_{TA}(t)) \tag{6.20}$$

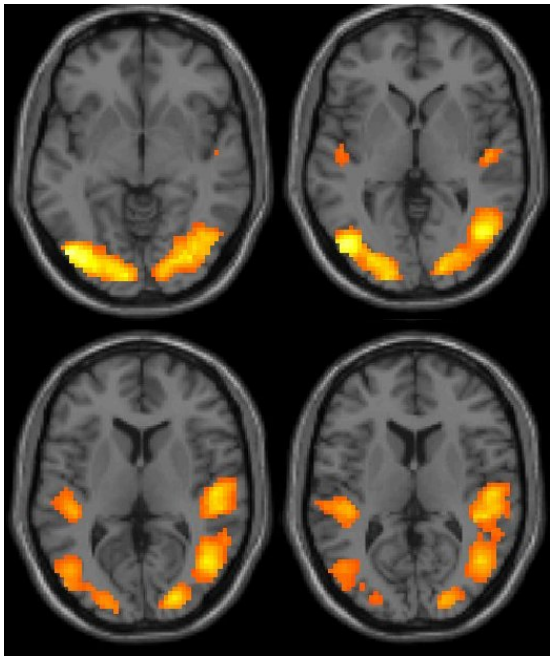
Στο Σχήμα 6.11 παρουσιάζονται οι δύο αυτές προσεγγίσεις σύμμειξης. Στο Σχήμα 6.11α′ παρατηρούμε ότι ο οπτικοακουστικός χάρτης εμφάνειας είναι πιο συσχετισμένος με το χρονικό οπτικό κομμάτι εξαιτίας της επίδρασης της ακουστικής εμφάνειας. Στην άλλη περίπτωση (Σχήμα 6.11β′) η ακουστική καμπύλη εμφάνειας διαμορφώνει την οπτική χρονική καμπύλη που στη συνέχεια συγχωνεύεται με τη χωρική.

Για την πειραματική αξιολόγηση, πραγματοποιήθηκε συλλογή δεδομένων με έξι συμμετέχοντες και δεδομένα από μία βάση με αποσπάσματα από ταινίες, κάποια από τα οποία ανήκουν στη βάση ETMD [273, 268]. Ακολούθησε η προ-επεξεργασία δεδομένων για το fMRI, για την οποία οι λεπτομέρειες μπορούν να βρεθούν στα [274, 275].

Στα Σχήματα 6.12, 6.13 παρουσιάζονται τα αποτελέσματα για το οπτικοακουστικό μοντέλο

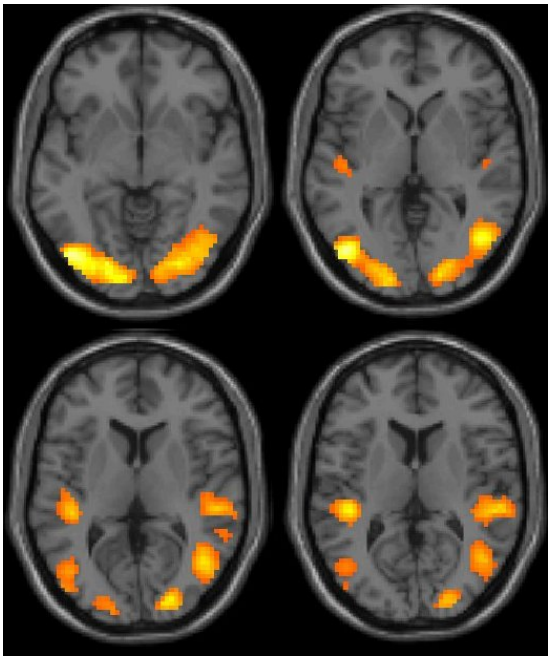


(α′) Οπτικοακουστική εμφάνεια - Μέσος όρος

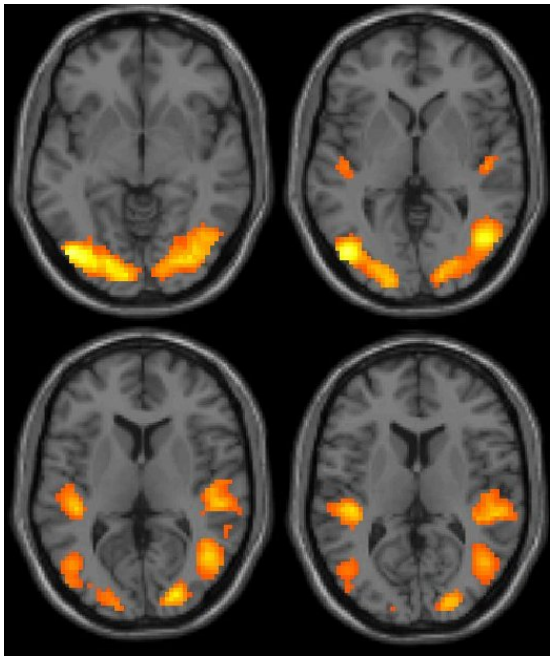


(β′) Οπτικοακουστική εμφάνεια - Μέγιστο

Σχήμα 6.12: Αποτελέσματα για την οπτικοακουστική εμφάνεια με τους δύο διαφορετικούς τρόπους σύμμειξης στο επίπεδο των *χαρτών εμφάνειας* (t-test).



(α′) Οπτικοακουστική εμφάνεια - Μέσος όρος



(β′) Οπτικοακουστική εμφάνεια - Μέγιστο

Σχήμα 6.13: Αποτελέσματα για την οπτικοακουστική εμφάνεια με τους δύο διαφορετικούς τρόπους σύμμειξης στο επίπεδο των *καμπυλών εμφάνειας* (t-test).

MNI Coordinates *X Y Z*	*Hemisphere*	*Functional or [Anatomical] area*	*T-value*
-45 -76 -2	Left	V4 ventral	16.69
-21 -82 -8	Left	V3 ventral	12.22
48 -67 1	Right	V4 ventral	12.07
36 -73 -8	Right	V5/MT	11.72
-9 -91 -5	Left	V1	10.87
-39 -31 13	Left	Area TE 1.1	10.70
-48 -31 16	Left	Area TE 1.1	10.67
-42 -28 10	Left	Area TE 1.1	10.43

Πίνακας 6.5: Θέσεις των κορυφών ενεργοποίησης για το *οπτικοακουστικό* μοντέλο εμφάνειας (σύμμειξη στο επίπεδο καμπυλών και μέγιστο).

για τους δύο διαφορετικούς τρόπους σύμμειξης. Όπως φαίνεται στα σχήματα, παρατηρούνται συστάδες ενεργοποίησης τόσο στον οπτικό όσο και στον ακουστικό φλοιό, ειδικά όταν η σύμμειξη γίνεται στο επίπεδο των καμπυλών. Σε σχέση με το επίπεδο στο οποίο πραγματοποιείται η σύμμειξη, φαίνεται ότι η σύμμειξη στο επίπεδο των καμπυλών λειτουργεί καλύτερα για το πρόβλημα της πρόβλεψης της χρονικής εμφάνειας από τη σύμμειξη στο επίπεδο των χαρτών, η οποία χρησιμοποιείται εκτενώς για την πρόβλεψη των σημείων εστίασης. Επίσης παρατηρούμε ότι το μέγιστο δίνει ελαφρώς καλύτερα αποτελέσματα από το μέσο όρο, αφού δίνει πιο συγκεντρωμένη ενεργοποίηση στον οπτικό και ακουστικό φλοιό.

Στον Πίνακα 6.5 παρουσιάζονται οι θέσεις των μέγιστων κορυφών ενεργοποίησης για το οπτικοακουστικό μοντέλο με σύμμειξη στο επίπεδο των καμπυλών και χρησιμοποιώντας το μέγιστο ως τελεστή.

Επομένως, συνοψίζοντας, τα αποτελέσματα υποδεικνύουν ότι το προτεινόμενο οπτικοακουστικό μοντέλο εμφάνειας έχει υψηλή συσχέτιση με τις ενεργοποιήσεις που παρουσιάστηκαν στον εγκέφαλο, εφόσον εντοπίζονται ενεργοποιήσεις τόσο στον οπτικό όσο και στον ακουστικό φλοιό. Άρα το προτεινόμενο μοντέλο επαληθεύτηκε και με δεδομένα fMRI.

6.4 Σύνοψη

Στην ενότητα αυτή παρουσιάστηκε ένα υπολογιστικό οπτικοακουστικό μοντέλο εμφάνειας, το οποίο υλοποιήθηκε με τεχνικές επεξεργασίας σήματος και βασίστηκε σε εμπνευσμένους από συμπεριφορικά πειράματα τρόπους σύμμειξης, εφαρμόζοντάς τους σε γνωστά μοντέλα οπτικής και ακουστικής εμφάνειας. Προτάθηκαν τρεις διαφορετικοί τρόποι σύμμειξης οπτικής και ακουστικής πληροφορίας και πραγματοποιήθηκε εκτενής αξιολόγηση τόσο με ανθρώπινα συμπεριφορικά πειράματα, όσο και με δεδομένα οφθαλμοκινητικής παρακολούθησης αλλά και δεδομένα fMRI. Η πρώτη προσπάθεια αξιολόγησης αφορά στα συμπεριφορικά πειράματα και συγκεκριμένα στα φαινόμενα “pip and pop” και “sine vs. square”, όπου το μοντέλο μας επιβεβαιώνει τα πειράματα αυτά, καθώς επιδεικνύει παρόμοια συμπεριφορά σε ποιοτικό επίπεδο. Σε σχέση με την δεύτερη στρατηγική αξιολόγησης με τα ανθρώπινα οπτικοακουστικά

δεδομένα οφθαλμοκινητικής παρακολούθησης, αξιολογήθηκε η επίδοση των διαφόρων τρόπων σύμμειξης και μοντέλων εμφάνειας σε 6 βάσεις δεδομένων ποικίλης περιπλοκότητας, τις DIEM, AVAD, Coutrot1, Coutrot2, SumMe και ETMD. Τέλος, σε σχέση με τα fMRI δεδομένα, το μοντέλο μας επαληθεύτηκε, διότι εντοπίστηκαν ενεργοποιήσεις τόσο στον οπτικό όσο και στον ακουστικό φλοιό. Τα αποτελέσματα και για τις τρεις στρατηγικές αξιολόγησης είναι υποσχόμενα και αναδεικνύουν την καταλληλότητα της οπτικοακουστικής εμφάνειας σε σχέση με την οπτική μόνο, για τη μοντελοποίηση της προσοχής, ακόμα και σε περίπλοκα ερεθίσματα.

Κεφάλαιο 7

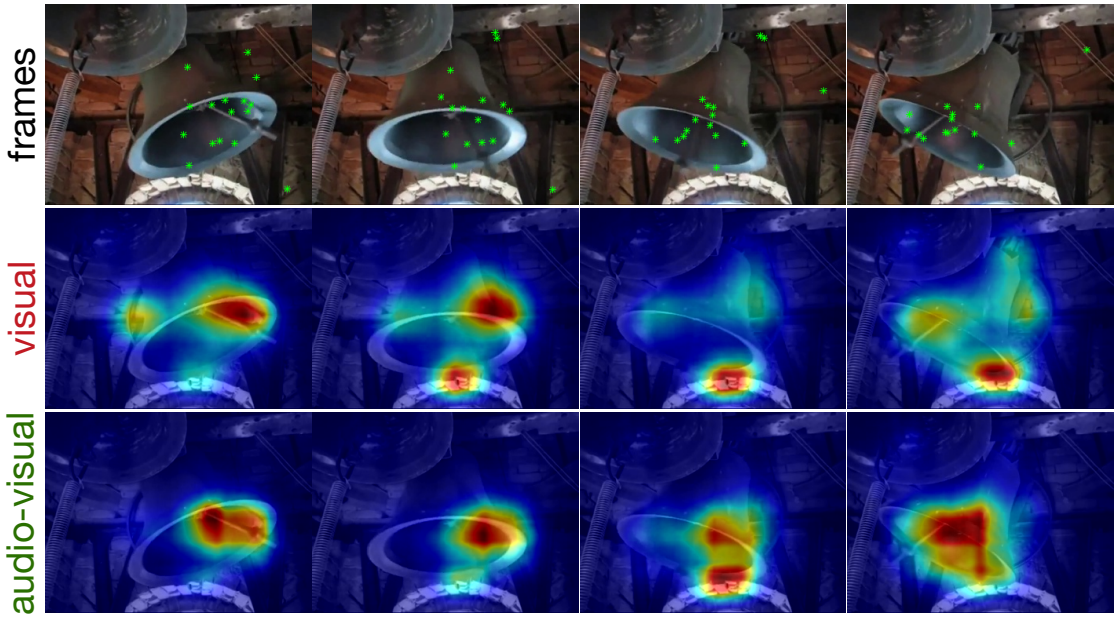
Χωρο-χρονικό οπτικοακουστικό μοντέλο εμφάνειας STAViS

7.1 Εισαγωγή

Όπως αναφέρθηκε και σε προηγούμενη ενότητα, η ακουστική πληροφορία είναι παρούσα στα περισσότερα γεγονότα που προσελκύουν την ανθρώπινη προσοχή και παρά τις ισχυρές αποδείξεις, οι περισσότερες ερευνητικές δουλειές στην όραση υπολογιστών εξακολουθούν να την αγνοούν στη μοντελοποίηση της προσοχής και να επικεντρώνονται μόνο στην οπτική. Προηγουμένως παρουσιάστηκαν τρόποι σύμμειξης οπτικής και ακουστικής εμφάνειας προκειμένου να δημιουργηθεί ένα μοντέλο οπτικοακουστικής εμφάνειας που θα προβλέπει τα σημεία εστίασης της ανθρώπινης προσοχής και να διερευνηθεί εάν αυτό μπορεί να μοντελοποιήσει καλύτερα την προσοχή σε σχέση με τα οπτικά μόνο μοντέλα.

Η διερεύνηση που πραγματοποιήθηκε, καθώς και οι τρόποι σύμμειξης, βασίστηκαν σε τεχνικές επεξεργασίας σήματος και ανακάλυψης συσχετίσεων ανάμεσα στις δύο τροπικότητες. Τα μοντέλα που είχαν βασιστεί σε βαθιά νευρωνικά δίκτυα εμφάνισαν καλύτερη επίδοση από τα υπόλοιπα σε απόλυτο αριθμό, υποδεικνύοντας έτσι την ανάγκη διερεύνησης της μοντελοποίησης της οπτικοακουστικής προσοχής αμιγώς με βαθιά νευρωνικά δίκτυα. Εδώ πάλι ο κύριος στόχος είναι η εκτίμηση της εμφάνειας σε βίντεο "in-the-wild", δηλαδή σε οποιοδήποτε βίντεο χωρίς πρότερη γνώση του περιεχομένου του. Το Σχήμα 7.1 αναδεικνύει το πρόβλημα που θέλουμε να αντιμετωπίσουμε: οι καμπάνες που χτυπούν είναι ένα οπτικοακουστικό γεγονός. Τα καρέ που απεικονίζονται με τα δεδομένα οφθαλμοκινητικής παρακολούθησης είναι ένα παράδειγμα της ανθρώπινης προσοχής σε μια οπτικοακουστική σκηνή. Η προσοχή μπορεί να μοντελοποιηθεί καλύτερα από ένα μοντέλο που λαμβάνει υπόψη και την ακουστική πληροφορία σε σχέση με ένα οπτικό μόνο μοντέλο, όπως υποδεικνύεται και από τους αντίστοιχους χάρτες εμφάνειας.

Τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks, CNNs) έχουν διεισδύσει στο πεδίο της Όρασης Υπολογιστών και έχουν βελτιώσει δραματικά την επίδοση της πλειοψηφίας των 'χωρικών' προβλημάτων, όπως η αναγνώριση αντικειμένων ή η σημασιολογική κατάτμηση [276, 277, 278]. Ο τεράστιος όγκος των δεδομένων βίντεο που γίνονται



Σχήμα 7.1: Ενδεικτικά καρέ με τα δεδομένα οφθαλμοκινητικής παρακολούθησης, για μία καμπάνα που χτυπά. Η δεύτερη σειρά απεικονίζει τους οπτικούς χάρτες εμφάνειας, ενώ η τρίτη τους οπτικοακουστικούς χάρτες που δημιουργούνται από το προτεινόμενο μοντέλο STAViS. Είναι εμφανές ότι το τελευταίο αποτυπώνει καλύτερα την ανθρώπινη προσοχή.

ολοένα και πιο διαθέσιμα μέσω διαδικτυακών πηγών, καθιστούν πιο εύκολη και ταυτόχρονα απαιτούν συνεχώς τη βελτίωση της απόδοσης σε προβλήματα όπως η αναγνώριση δράσεων, η εκτίμηση εμφάνειας σε βίντεο, η ανάλυση οπτικής σκηνής ή η περίληψη βίντεο, επιβάλλοντας την ανάγκη αξιοποίησης όχι μόνο της χωρικής πληροφορίας, αλλά και της χρονικής [279, 280, 281]. Αντίστοιχες βελτιώσεις απόδοσης έχουν επιτευχθεί και στον τομέα επεξεργασίας ακουστικής πληροφορίας, όπως η αναγνώριση ακουστικών γεγονότων [282], η αναγνώριση φωνής [283, 284], ο εντοπισμός ηχητικών πηγών [191], κλπ, με χρήση βαθιών νευρωνικών δικτύων.

Ως μια επιπρόσθετη, χωρίς επιπλέον κόστος πηγή πληροφορίας, η ακουστική πληροφορία έχει ενσωματωθεί σε πολλές εφαρμογές σχετιζόμενες με βίντεο κατά τη διάρκεια των τελευταίων τριών χρόνων [187, 192]. Ο ήχος περιέχει πλούσια πληροφορία η οποία όταν ενσωματώνεται με την χωρο-χρονική οπτική, βοηθάει στην εκμάθηση των δικτύων και βελτιώνει την επίδοση σε πολλά προβλήματα που για χρόνια θεωρούνταν καθαρά οπτικά προβλήματα. Εξίσου σημαντικό πλεονέκτημα είναι ότι ο ήχος έρχεται ουσιαστικά χωρίς επιπλέον κόστος, εφόσον είναι ήδη ενσωματωμένος στην πλειοψηφία των βίντεο και συγχρονισμένος με αυτό.

Στη συγκεκριμένη ενότητα παρουσιάζεται και προτείνεται το δίκτυο STAViS (Spatio-Temporal AudioVisual Saliency Network)[1], ένα νέο χωρο-χρονικό οπτικοακουστικό μοντέλο εμφάνειας με νευρωνικά δίκτυα, το οποίο συνδυάζει χωρο-χρονική οπτική πληροφορία

[1] A. Tsiami, P. Koutras and P. Maragos, "STAViS: Spatio-Temporal AudioVisual Saliency Network", submitted in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2020)

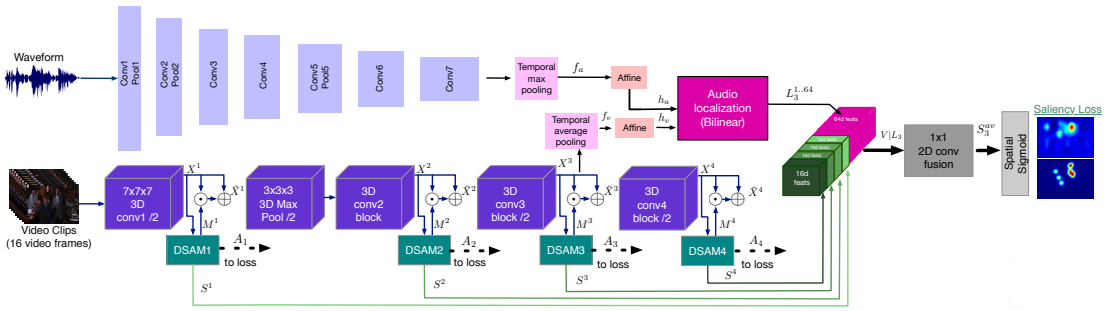
με ακουστική, προκειμένου να αντιμετωπίσει αποτελεσματικά το πρόβλημα της εκτίμησης της εμφάνειας σε βίντεο. Από όσο μπορούμε να γνωρίζουμε, το προτεινόμενο σύστημα είναι το πρώτο βαθύ νευρωνικό μοντέλο εμφάνειας, εκτός ίσως από μία ταυτόχρονη εργασία [285] όπου όμως χρησιμοποιούνται δύο ανεξάρτητα δίκτυα για την κάθε τροπικότητα και οι έξοδοί τους απλά συνενώνονται σαν ένα σχήμα late fusion. Η δική μας προσέγγιση χρησιμοποιεί ένα μόνο πολυτροπικό δίκτυο που συνδυάζει οπτική και ακουστική πληροφορία σε διάφορα στάδια και μαθαίνει να πραγματοποιεί κατάλληλη σύμμειξη των δύο τροπικοτήτων ώστε να αποκτήσουμε έναν τελικό χάρτη εμφάνειας. Στην παρούσα ενότητα προτείνονται τρόποι εντοπισμού των ήχων στο βίντεο και κατόπιν η σύμμειξη της χωρο-χρονικής ακουστικής με την χωρο-χρονική οπτική εμφάνεια ώστε να αποκτηθεί ο τελικός χάρτης εμφάνειας.

Η προσέγγισή μας για την οπτική εμφάνεια είναι βασισμένη, αλλά ταυτόχρονα επεκτείνει ένα state-of-the-art χωρο-χρονικό δίκτυο οπτικής εμφάνειας, το SUSiNet [281] που είναι μέρος ενός μεγαλύτερου δικτύου πολλαπλών εργασιών που αντιμετωπίζει από κοινού τρία διαφορετικά προβλήματα: εκτίμηση εμφάνειας, αναγνώριση δράσεων και περίληψη βίντεο. Όσον αφορά την ακουστική εμφάνεια, αποκτούμε ακουστικά χαρακτηριστικά από ένα state-of-the-art CNN δίκτυο για ταξινόμηση ακουστικών γεγονότων, που ονομάζεται SoundNet [194]. Για τον εντοπισμό ακουστικών πηγών στο βίντεο, στόχος μας είναι να βρούμε τις πιθανές συσχετίσεις μεταξύ της ακουστικής και οπτικής ροής πληροφορίας, με την έννοια των κοινών σημασιολογικών εννοιών. Ερευνούμε τρεις διαφορετικούς τρόπους για να εντοπίζουμε τους ήχους στο βίντεο και να αποκτούμε έναν ακουστικό χάρτη εμφάνειας. Κατόπιν διερευνούμε τρόπους σύμμειξης προκειμένου να ενσωματωθούν οι δύο εμφάνειες και να αποκτηθεί ένας τελικός χάρτης εμφάνειας. Και εδώ εξερευνούμε τρεις διαφορετικούς τρόπους.

Το δίκτυο σχεδιάστηκε, εκαπιδεύτηκε end-to-end και αξιολογήθηκε σε 6 διαφορετικές βάσεις με οπτικοακουστικά δεδομένα οφθαλμοκινητικής παρακολούθησης. Οι βάσεις αυτές είναι οι ίδιες που αναφέρθηκαν και σε προηγούμενη ενότητα και περιέχουν μεγάλη ποικιλία από διαφορετικά βίντεο, από απλά μικρά βίντεο ως ταινίες του Χόλιγουντ και από βίντεο φτιαγμένα από χρήστες μέχρι ντοκιμαντέρ. Η μέθοδος μας συγκρίνεται με άλλες 8 διαφορετικές state-of-the-art μεθόδους οπτικής εμφάνειας και στις 6 βάσεις. Τα αποτελέσματα για όλες τις βάσεις υποδεικνύουν ότι το προτεινόμενο μοντέλο STAViS ξεπερνά σε απόδοση τόσο την οπτική μόνο παραλλαγή μας αλλά και τις άλλες μεθόδους στην πλειοψηφία των περιπτώσεων. Επίσης, η συνεπής καλή επίδοση που εμφανίζει για όλες τις βάσεις υποδεικνύει την καταλληλότητα της μοντελοποίησης αυτής για την εκτίμηση εμφάνειας σε βίντεο "in-the-wild".

7.2 Χωρο-χρονικό Οπτικοακουστικό μοντέλο εμφάνειας STAViS

Το προτεινόμενο χωρο-χρονικό οπτικοακουστικό δίκτυο για εκτίμηση εμφάνειας που σχεδιάστηκε (Σχήμα 7.2) αποτελείται από μία χωρο-χρονική οπτική μονάδα η οποία υπολογίζει την οπτική εμφάνεια, μία μονάδα ακουστικής αναπαράστασης βασισμένη στο [194] η οποία υπολογίζει ακουστικά χαρακτηριστικά, μία μονάδα εντοπισμού ήχου η οποία υπολογίζει τη χωρο-χρονική ακουστική εμφάνεια, μία μονάδα εκτίμησης της οπτικοακουστικής εμφάνειας η



Σχήμα 7.2: Η αρχιτεντονική του δικτύου STAViS: το χωρο-χρονικό οπτικοακουστικό δίκτυο είναι βασισμένο στην αρχιτεκτονική ResNet και αποτελείται από ένα χωρο-χρονικό οπτικό, ένα ακουστικό μονοπάτι και τη σύμμειξή τους.

οποία συνδυάζει και συγχωνεύει την οπτική και την ακουστική εμφάνεια και τέλος τα κατάλληλα losses. Όλα αυτά περιγράφονται με λεπτομέρεια στις επόμενες ενότητες.

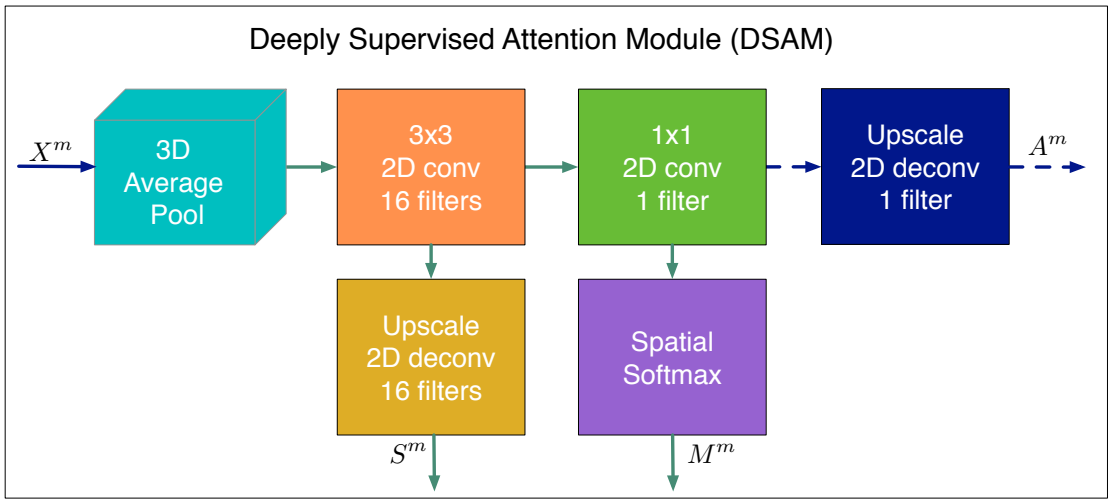
7.2.1 Χωρο-χρονικό οπτικό δίκτυο

Η αρχιτεκτονική του χωρο-χρονικού δικτύου οπτικής εμφάνειας που σχεδιάστηκε παρουσιάζεται στο Σχήμα 7.2. Χρησιμοποιείται η γενική αρχιτεκτονική ResNet [277] και πιο συγκεκριμένα η 3D επέκταση του που προτάθηκε αρχικά για το πρόβλημα της αναγνώρισης δράσεων στο [286]. Το μονοπάτι του οπτικού δικτύου με παραμέτρους $\mathbf{W}_{res}$ υποδηλώνεται με το σκούρο μωβ χρώμα και περιλαμβάνει τα 4 πρώτα συνελικτικά μπλοκ ResNet conv1, conv2, conv3, conv4 τα οποία δίνουν εξόδους $X^m, m = 1, \ldots, 4$ αντίστοιχα, σε διαφορετικές χωρικές και χρονικές κλίμακες. Παράλληλα, εφαρμόζεται και ένας μηχανισμός 'προσοχής' (attention mechanism) λεγόμενος Deeply Supervised Attention Module (DSAM), λαμβάνοντας το στοιχείο-προς-στοιχείο γινόμενο μεταξύ κάθε καναλιού του χάρτη χαρακτηριστικών X^m και του χάρτη 'προσοχής' M^m, ώστε να ενισχυθούν οι πιο προεξέχουσες περιοχές αυτών των αναπαραστάσεων:

$$\tilde{X}_m = (1 + M^m) \odot X_m, \quad m = 1, \ldots, 4. \tag{7.1}$$

Η ιδέα της βαθιάς επίβλεψης (deep supervision) που είναι ο πυρήνας της μονάδας DSAM έχει χρησιμοποιηθεί στο παρελθόν για εντοπισμό ακμών [287], κατάτμηση αντικειμένου [288] και στατικής εμφάνειας [158], αλλά σε αντίθεση με αυτές τις εργασίες, εδώ ο ρόλος της DSAM είναι διπλός: Χρησιμοποιείται τόσο για να βελτιώσει τις οπτικές αναπαραστάσεις των χαρακτηριστικών όσο και για να δώσει ως έξοδο τους πολυ-επίπεδους χάρτες εμφάνειας, όπως φαίνεται στο Σχήμα 7.2 από τις αποχρώσεις του πράσινου. Επομένως, οι παράμετροι $\mathbf{W}^m_{am}$ της DSAM εκπαιδεύονται τόσο από το βασικό μονοπάτι του οπτικού δικτύου και από τα δεδομένα οφθαλμοκινητικής παρακολούθησης μέσω των skip connections στο Σχήμα 7.2.

Το Σχήμα 7.3 παρουσιάζει την αρχιτεκτονική της μονάδας DSAM στο επίπεδο m. Περιλαμβάνει ένα average pooling στη χρονική διάσταση, το οποίο ακολουθείται από δύο χωρικά συνελικτικά φίλτρα τα οποία υπολογίζουν τα χαρακτηριστικά εμφάνειας S^m και το χάρτη ε-



Σχήμα 7.3: Η μονάδα DSAM (Deeply Supervised Attention Module) βελτιώνει τις καθολικές αναπαραστάσεις του δικτύου και παρέχει τους πολυ-επίπεδους χάρτες χωρο-χρονικής οπτικής εμφάνειας.

νεργοποίησης A^m. Και οι δύο αναπαραστάσεις υπερδειγματοληπτούνται (up-sampled) στις αρχικές διαστάσεις τις εικόνας χρησιμοποιώντας κατάλληλα layers αποσυνέλιξης και χρησιμοποιούνται για τη βαθιά επίβλεψη της μονάδας αλλά και για την εκτίμηση εμφάνειας πολλαπλών επιπέδων. Ο χάρτης 'προσοχής' $M^m(x, y)$ προκύπτει εφαρμόζοντας ένα spatial softmax στο χάρτη ενεργοποίησης $A^m(x, y)$:

$$M^m(x, y) = \frac{\exp(A^m(x, y))}{\sum_x \sum_y \exp(A^m(x, y))} \tag{7.2}$$

7.2.2 Δίκτυο Ακουστικής Αναπαράστασης

Όσον αφορά την ακουστική αναπαράσταση, επιλέγεται η ενασχόληση απευθείας με τις ηχητικές κυματομορφές και χρήση ενός 1-D fully convolutional δικτύου αντί να υπολογίζονται οι χωρο-χρονικές αναπαραστάσεις και να εφαρμόζονται 2D CNNs [195, 187, 192]. Αρχικά το σήμα του ήχου περικόβεται ώστε να ταιριάζει στη διάρκεια των οπτικών καρέ (16 καρέ). Αξίζει να σημειωθεί ότι το δίκτυο που χρησιμοποιούμε μπορεί να χειριστεί ήχο μεταβλητής διάρκειας, οπότε δεν απαιτείται καμία επιπλέον επεξεργασία όπως υποδειγματοληψία ή αλλου είδους ενέργεια. Στη συνέχεια, εφαρμόζεται ένα Hanning παράθυρο ώστε να δοθεί βάρος στις κεντρικές τιμές του ήχου που αφορούν την παρούσα χρονική στιγμή, αλλά ταυτόχρονα να συμπεριληφθούν παρελθοντικές και μελλοντικές στιγμές ως context με κάποια απόσβεση [268]. Κατόπιν, για την κωδικοποίηση της πληροφορίας υψηλού επιπέδου, χρησιμοποιείται μια αρχιτεκτονική δικτύου με παραμέτρους $\mathbf{W}_a$ βασισμένη στα 7 πρώτα layers του SoundNet [194]. Τα layers αυτά ακολουθούνται από ένα temporal max-pooling layer ώστε να λάβουμε ένα διάνυσμα $f_a \in \mathbb{R}^{D_a}$ σταθερής διάστασης για όλο το ακουστικό σήμα εισόδου.

7.2.3 Εντοπισμός ηχητικής πηγής σε βίντεο

Ο στόχος της μονάδας της προηγούμενης υπο-ενότητας είναι να δημιουργήσει καλές αναπαραστάσεις του ήχου. Ο επόμενος στόχος είναι η αξιοποίηση των αναπαραστάσεων αυτών ώστε να εντοπιστούν οι cross-modal σημασιολογικές έννοιες στο βίντεο. Ουσιαστικά, προσπαθούμε να ανακαλύψουμε τις αντιστοιχίες μεταξύ της ακουστικής και της οπτικής χωρο-χρονικής πληροφορίας και έτσι να αποκτήσουμε έναν χωρο-χρονικό ακουστικό χάρτη εμφάνειας. Από μια άλλη σκοπιά, το πρόβλημα αυτό είναι ένα πρόβλημα εντοπισμού ηχητικής πηγής. Για παράδειγμα, ένα κινούμενο αυτοκίνητο, ένα πρόσωπο που μιλάει, ή ένα φύλλο που θροϊζει είναι τέτοια cross-modal γεγονότα, όπου είναι αναγκαίο να εντοπίσουμε από πού προέρχεται ο ήχος πάνω στα καρέ του βίντεο.

Επιλέγουμε ως οπτικά χαρακτηριστικά την έξοδο από το 3D conv3 μπλοκ X^3 με διάσταση χαρακτηριστικών D_v, διότι σε αυτό το layer έχουμε και πλούσια σημασιολογική πληροφορία αλλά και σχετικά μεγάλη ανάλυση σε χωρικό επίπεδο. Εφαρμόζεται και εδώ ένα temporal average pooling προκειμένου να αφαιρεθεί η χρονική διάσταση και να αποκτήσουμε μια καθολική αναπαράσταση $f_v \in \mathbb{R}^{D_v \times N_X \times N_Y}$ για όλη τη χρονική ακολουθία. Εφόσον τα οπτικά και τα ακουστικά χαρακτηριστικά έχουν διαφορετικές διαστάσεις, επαναπροβάλλουμε και τα δύο σε μια 'κρυφή' διάσταση D_h, εφαρμόζοντας δύο διαφορετικούς αφινικούς (affine) μετασχηματισμούς:

$$\tilde{h}_a = \mathbf{U}_a \cdot f_a + \mathbf{b}_a, \quad h_v(x,y) = \mathbf{U}_v \cdot f_v(x,y) + \mathbf{b}_v, \tag{7.3}$$

όπου τα $\tilde{h}_a \in \mathbb{R}^{D_h}$, $h_v \in \mathbb{R}^{D_h \times N_X \times N_Y}$ και $\mathbf{U}_a, \mathbf{b}_a, \mathbf{U}_v, \mathbf{b}_v$ είναι οι αντίστοιχες παράμετροι εκμάθησης. Εφαρμόζουμε επίσης ένα χωρικό tiling στα ακουστικά χαρακτηριστικά ώστε να ταιριάζουν με τις χωρικές διαστάσεις των οπτικών χαρακτηριστικών $N_X \times N_Y$, οπότε προκύπτει το $h_a \in \mathbb{R}^{D_h \times N_X \times N_Y}$.

Διερευνήθηκαν τρεις διαφορετικές προσεγγίσεις προκειμένου να βρούμε ή να 'μάθουμε' τις αντιστοιχίες ακουστικών και οπτικών χαρακτηριστικών. Στην πρώτη, απλώς υπολογίζουμε το cosine similarity μεταξύ των δύο διανυσμάτων, το οποίο δεν απαιτεί καμία παράμετρο εκμάθησης και παρέχει έναν χάρτη εντοπισμού $L_1 \in \mathbb{R}^{N_X \times N_Y}$.

Σε σχέση με τη δεύτερη προσέγγιση, λαμβάνουμε το σταθμισμένο εσωτερικό γινόμενο των διανυσμάτων h_a, h_v για κάθε εικονοστοιχείο (x,y), από όπου μπορούμε να λάβουμε έναν ή περισσότερους χάρτες εντοπισμού $L_2^j \in \mathbb{R}^{N_X \times N_Y}, \quad j = 1, \ldots, N_{out}$:

$$L_2^j(x,y) = \sum_{k=1}^{D_h} s^{j,k} \cdot h_v^k(x,y) \cdot h_a^k(x,y) + \beta^j, \tag{7.4}$$

όπου $s^{j,k}, \beta^j$ είναι οι παράμετροι εκμάθησης.

Στην τρίτη περίπτωση, που αποτελεί και την προτεινόμενη μέθοδο εφαρμόζεται ένας bilinear μετασχηματισμός στα πολυτροπικά δεδομένα εισόδου, από τον οποίο επίσης μπορούμε να

λάβουμε έναν ή περισσότερους χάρτες εξόδου $L_3^j \in \mathbb{R}^{N_X \times N_Y}, \quad j = 1, \ldots, N_{out}$:

$$\begin{aligned} L_3^j(x,y) &= h_v(x,y)^T \cdot \mathbf{M}^j \cdot h_a(x,y) + \mu^j \\ &= \sum_{l=1}^{D_h} \sum_{k=1}^{D_h} M^{j,l,k} \cdot h_v^l(x,y) \cdot h_a^k(x,y) + \mu^j, \end{aligned} \tag{7.5}$$

όπου $M^{j,l,k}, \mu^j$ είναι οι παράμετροι εκμάθησης. Πρέπει να σημειωθεί ότι οι προηγούμενες προσεγγίσεις (L_1, L_2) είναι ειδικές περιπτώσεις αυτού του bilinear μετασχηματισμού (L_3), ο οποίος επιτρέπει πλουσιότερες αλληλεπιδράσεις μεταξύ των εισόδων. Πιο συγκεκριμένα, όταν οι πίνακες $\mathbf{M}^j$ είναι διαγώνιοι με διαγώνια στοιχεία $s^{j,k}$, έχουμε την περίπτωση του σταθμισμένου εσωτερικού γινομένου (L_2). Όταν ο πίνακας $\mathbf{M}$ είναι ο μοναδιαίος πίνακας, το αποτέλεσμα είναι πάρα πολύ κοντά (πλην ενός παράγοντα κανονικοποίησης) στο cosine similarity.

7.2.4 Εκτίμηση οπτικοακουστικής εμφάνειας

Έχουμε τώρα υπολογίσει έναν ακουστικό χάρτη εμφάνειας, εκφρασμένο μέσω χαρτών εντοπισμού ηχητικών πηγών. Παρ΄ όλα αυτά, υπάρχουν πολλά στοιχεία σε ένα βίντεο που μπορεί να προσελκύσουν την ανθρώπινη προσοχή, που ίσως δεν σχετίζονται με ακουστικά γεγονότα. Επομένως για να δημιουργήσουμε ένα δίκτυο για την εκτίμηση της πολυτροπικής εμφάνειας το οποίο να έχει καλή επίδοση "in-the-wild" με οποιοδήποτε είδος ερεθίσματος, πρέπει να ενσωματώσουμε και την οπτική πληροφορία, όπως μοντελοποιείται από το δίκτυο χωρο-χρονικής οπτικής εμφάνειας. Για το λόγο αυτό, μία ακόμη σημαντική συνεισφορά της διατριβής αυτής είναι η διερεύνηση τριών διαφορετικών τρόπων οπτικοακουστικής σύμμειξης.

Ο πιο απλός τρόπος σύμμειξης είναι η εκμάθηση ενός γραμμικού σταθμισμένου αθροίσματος του οπτικού χάρτη S^v και του ακουστικού S^a, που επιτυγχάνεται εφαρμόζοντας ανεξάρτητα fully convolution layers στους πολυ-επίπεδους συνενωμένους οπτικούς χάρτες χαρακτηριστικών $V^j = (S^1|\ldots|S^m|\ldots|S^M)$ και στους ακουστικούς L^j. Το αποτέλεσμα της σύμμειξης είναι $S_1^{av} = w_v \cdot \sigma(S^v) + w_a \cdot \sigma(S^a)$, όπου το $\sigma(\cdot)$ υποδηλώνει μία sigmoid activation function.

Επιπλέον, εμπνευσμένοι από παλιότερες προσεγγίσεις με τεχνικές επεξεργασίας σήματος για οπτικοακουστική εμφάνεια [268, 274], διερευνήθηκε ένας τρόπος σύμμειξης βασισμένος στην 'προσοχή', όπου η ηχητική πληροφορία 'διαμορφώνει' (modulates) την οπτική: $S_2^{av} = \sigma(S^v) \cdot (1 + \sigma(S^a))$. Στην περίπτωση πολλαπλών χαρτών εντοπισμού ήχου, μπορούμε να πολλαπλασιάσουμε έναν προς ένα τους συνενωμένους οπτικούς χάρτες χαρακτηριστικών V^j με τους χάρτες εντοπισμού ήχου L^j και στη συνέχεια να εφαρμόσουμε ένα fully convolutional layer ώστε να πάρουμε ως έξοδο έναν χάρτη εμφάνειας: $\tilde{S}_2^{av,j} = \sigma(V^j) \cdot (1 + \sigma(L^j))$.

Παρ΄ όλα αυτά, όπως απεικονίζεται και στο Σχήμα 7.2, η κύρια και πιο γενική προσέγγισή μας που αποτελεί και την τρίτη προσέγγιση και η οποία επιτρέπει πιο ελεύθερη αλληλεπίδραση μεταξύ των οπτικών και των ακουστικών χαρακτηριστικών είναι η συνένωση των χαρτών πολυτροπικών χαρακτηριστικών και η εφαρμογή ενός convolutional layer για τη σύμμειξή τους σε έναν χάρτη εμφάνειας: $S_3^{av} = \mathbf{W}_{cat} \cdot (V|L) + \beta_{av}$. Τέλος, εφαρμόζουμε και ένα late

fusion scheme μεταξύ όλων των προηγούμενων προσεγγίσεων, το οποίο μπορεί να εκφραστεί από ένα σταθμισμένο άθροισμα: $S^{av}_{fus} = \tilde{w}_v \cdot \sigma(S^v) + \tilde{w}_a \cdot \sigma(S^a) + \cdot w_{av} \cdot \sigma(S^3)$.

7.2.5 Losses

Για την εκπαίδευση των παραμέτρων $\mathbf{W}_v$ του δικτύου που σχετίζονται με την οπτική πληροφορία, κατασκευάζουμε ένα loss το οποίο συγκρίνει το χάρτη εμφάνειας S^v και τις ενεργοποιήσεις A^m με τους ground truth χάρτες Y_{sal} που αποκτώνται από τα δεδομένα οφθαλμοκινητικής παρακολούθησης:

$$\begin{aligned} \mathcal{L}_v(\mathbf{W}_v) = \mathcal{D}(\mathbf{W}_v|\sigma(S^v), Y_{sal}) + \\ \sum_{m=1}^{4} \mathcal{D}(\mathbf{W}^m_{AM}|\sigma(A^m), Y_{sal}), \end{aligned} \tag{7.6}$$

όπου το $\sigma(\cdot)$ υποδηλώνει την sigmoid μη γραμμικότητα και το $\mathcal{D}(\cdot)$ είναι το loss function μεταξύ των 2D χαρτών που υπολογίστηκαν και των αντίστοιχων ground truth. Όταν εκπαιδεύονται οι παράμετροι $\mathbf{W}_{av}$ του οπτικοακουστικού δικτύου, χρησιμοποιείται το εκπαιδευμένο οπτικό μόνο δίκτυο ως σημείο εκκίνησης και δε χρησιμοποιούνται οι skip connections των μονάδων DSAM:

$$\mathcal{L}_{av}(\mathbf{W}_{av}) = \mathcal{D}(\mathbf{W}_{av}|\sigma(S^{av}), Y_{sal}). \tag{7.7}$$

Για να αξιολογηθεί η εκτίμηση εμφάνειας, χρησιμοποιούνται διάφορες μετρικές οι οποίες συγκρίνουν τους χάρτες πρόβλεψης εμφάνειας $P \in [0,1]^{N_X \times N_Y}$ με τα δεδομένα οφθαλμοκινητικής παρακολούθησης [261]. Ως χάρτες ground truth χρησιμοποιούνται είτε οι χάρτες των σημείων εστίασης $Y_{fix} \in \{0,1\}^{N_X \times N_Y}$ στο πεδίο της εικόνας, μεγέθους $N_X \times N_Y$, είτε οι πυκνοί χάρτες εμφάνειας $Y_{den} \in [0,1]^{N_X \times N_Y}$ οι οποίοι παράγονται μέσω της συνέλιξης ενός δυαδικού χάρτη εστίασης με έναν Gaussian πυρήνα. Επομένως, ως $\mathcal{D}(\cdot)$ χρησιμοποιούμε τρία loss functions που σχετίζονται με τους διαφορετικούς τρόπους αξιολόγησης της εμφάνειας. Το πρώτο είναι το cross-entropy loss μεταξύ του χάρτη πρόβλεψης P και του πυκνού χάρτη Y_{den}:

$$\begin{aligned} \mathcal{D}_{CE}(\mathbf{W}|P, Y_{den}) = -\sum_{x,y} Y_{den}(x,y) \odot \log(P(x,y;\mathbf{W})) \\ +(1 - Y_{den}(x,y)) \odot (1 - \log(P(x,y;\mathbf{W}))). \end{aligned} \tag{7.8}$$

Το δεύτερο βασίζεται στο linear Correlation Coefficient (CC) [261] που περιγράφηκε και σε προηγούμενη ενότητα και χρησιμοποιείται ευρέως στην αξιολόγηση της εμφάνειας. Μετράει τη γραμμική σχέση μεταξύ της πρόβλεψης P και του πυκνού χάρτη Y_{den}:

$$\mathcal{D}_{CC}(\mathbf{W}|P, Y_{den}) = -\frac{\mathrm{cov}(P(x,y;\mathbf{W}), Y_{den}(x,y))}{\rho(P(x,y;\mathbf{W})) \cdot \rho(Y_{den}(x,y))}, \tag{7.9}$$

όπου το cov, ρ συμβολίζει τη συμμεταβλητότητα (covariance) και την τυπική απόκλιση (standard deviation) αντίστοιχα.

Το τελευταίο loss προκύπτει από τη μετρική Normalized Scanpath Saliency (NSS) [261], η οποία όπως έχει αναφερθεί και σε προηγούμενη ενότητα υπολογίζεται με βάση τις εκτιμώμενες τιμές των χαρτών $\tilde{P}(x, y; \mathbf{W}) = \frac{P(x,y;\mathbf{W})-\mu(P(x,y;\mathbf{W}))}{\rho(P(x,y;\mathbf{W}))}$, αφού κανονικοποιηθούν με μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση στα σημεία εστίασης της ανθρώπινης προσοχής ($Y_{fix}(x, y) = 1$):

$$\mathcal{D}_{NSS}(\mathbf{W}|\tilde{P}, Y_{fix}) = -\frac{1}{N_f} \sum_{x,y} \tilde{P}(x, y; \mathbf{W}) \odot Y_{fix}(x, y), \tag{7.10}$$

όπου το $N_f = \sum_{x,y} Y_{fix}(x, y)$ υποδηλώνει το συνολικό αριθμό των σημείων εστίασης.

Το τελικό loss του i-οστού δείγματος εισόδου δίνεται από ενα σταθμισμένο συνδυασμό των losses $\mathcal{L}^i_{CE}, \mathcal{L}^i_{CC}, \mathcal{L}^i_{NSS}$ που δίνονται είτε από την Εξ.(7.6) είτε από την Εξ.(7.7) χρησιμοποιώντας τις αντίστοιχες loss functions $\mathcal{D}^i_{CE}, \mathcal{D}^i_{CC}, \mathcal{D}^i_{NSS}$:

$$\mathcal{L}^i_{sal}(\mathbf{W}) = w_1\mathcal{L}^i_{CE} + w_2\mathcal{L}^i_{CC} + w_3\mathcal{L}^i_{NSS}, \tag{7.11}$$

όπου w_1, w_2, w_3 είναι τα βάρη για το κάθε loss.

7.2.6 Υλοποίηση

Η υλοποίηση και ο πειραματισμός μας όσον αφορά το οπτικό δίκτυο χρησιμοποιεί ως ραχοκοκαλιά την αρχιτεκτονική 3D ResNet-50 [286] η οποία έχει επιδείξει ανταγωνιστική επίδοση σε σχέση με πιο βαθιές αρχιτεκτονικές για το πρόβλημα της αναγνώρισης δράσεων, τόσο σε σχέση με την επίδοση όσο και με το υπολογιστικό κόστος. Σαν σημείο εκκίνησης για το $\mathbf{W}_{res}$ χρησιμοποιήθηκαν τα βάρη ενός προ-εκπαιδευμένου μοντέλου στη βάση δεδομένων Kinetics 400. Για την ακουστική ροή χρησιμοποιείται το SoundNet [194] (με χρήση των 7 από τα 8 layers), το οποίο βασίζεται σε 1D χρονικές συνελίξεις και το οποίο έχει εφαρμοστεί με επιτυχία σε ταξινόμηση ακουστικής σκηνής και γεγονότων. Σαν σημείο εκκίνησης για τις παραμέτρους $\mathbf{W}_a$ του δικτύου ακουστικής αναπαράστασης χρησιμοποιούμε τα βάρη από το προ-εκπαιδευμένο μοντέλο το οποίο έχει εκπαιδευτεί σε 2000000 βίντεο από το Flickr [194].

Εκπαίδευση: Για την εκπαίδευση χρησιμοποιούμε stochastic gradient descent με momentum 0.9, ενώ θέτουμε το weight decay στο $1e-5$ για κανονικοποίηση. Επίσης χρησιμοποιούμε effective batchsizes των 128 δειγμάτων, ενώ το learning rate ξεκινάει από 0.01 και διαιρείται δια 10 όταν υπάρχει κορεσμός στο loss. Αρχικά εκπαιδεύουμε το οπτικό χωρο-χρονικό δίκτυο και στη συνέχεια ολόκληρο το οπτικοακουστικό δίκτυο χρησιμοποιώντας τα βάρη του οπτικού ως σημείο εκκίνησης. Τα skip connections των μονάδων DSAM χρησιμοποιούνται μόνο κατά την εκπαίδευση του οπτικού μόνο δικτύου. Τα βάρη w_1, w_2, w_3 του loss της εμφάνειας επιλέγονται ως 0.1, 2, 1 αντίστοιχα, κατόπιν πειραματισμού.

Επαύξηση δεδομένων: Τα δεδομένα εισόδου του δικτύου αποτελούνται από RGB βίντεο κλιπς των 16 καρέ, τα οποία έχουν αλλάξει μέγεθος στη χωρική διάστασή τους, στα 112×112 εικονοστοιχεία, καθώς και από τις αντίστοιχες ακουστικές κυματομορφές με δειγματοληψία στα 22050 Hz. Επίσης εφαρμόστηκε επαύξηση δεδομένων για την τυχαία παραγωγή δειγμάτων εκπαίδευσης. Πρώτον, τα αρχικά μεγάλης διάρκειας βίντεο χωρίστηκαν σε μη επικαλυπτόμενα

Method \ Dataset	Overall				
	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑
Visual	0.5260	2.54	0.8922	0.6187	0.4088
L_1_AudioOnly	0.5132	2.48	0.8923	0.6220	0.4100
L_2_AudioOnly	0.5381	2.63	0.8969	0.6171	0.4157
L_3_AudioOnly	0.5226	2.59	0.8935	0.6228	0.4027
L_1_S_1^{av}	0.5243	2.62	0.8924	0.6224	0.4103
L_2_S_1^{av}	0.5344	2.69	0.8944	0.6215	0.4249
L_3_S_1^{av}	0.5354	2.71	0.8959	0.6236	0.4278
L_1_S_2^{av}	0.5069	2.43	0.8687	0.6223	0.2985
L_2_S_2^{av}	0.5082	2.44	0.8699	0.6225	0.2996
L_3_S_2^{av}	0.5066	2.46	0.8622	0.6219	0.2981
L_3^{mul}_$\tilde{S}_2^{av}$	0.5271	2.56	0.8943	0.6242	0.4132
$\mathbf{L_3^{mul}}$_$\mathbf{S_3^{av}}$	**0.5414**	**2.73**	**0.8983**	**0.6267**	0.4241
L_3^{mul}_S_{fus}^{av}	0.5401	2.70	0.8979	0.6261	**0.4290**

Πίνακας 7.1: Ablation study: Τα L_1, L_2 και L_3 αναφέρονται στη μέθοδο εντοπισμού του ήχου: cosine, inner product και bilinear αντίστοιχα. Τα S_1^{av}, S_2^{av}, S_3^{av} και S_{fus}^{av} αναφέρονται στους διφορετικούς τρόπους σύμμειξης. Ο υπερδείκτης "mul" αφορά στους πολλαπλούς χάρτες χαρακτηριστικών.

κομμάτια των 90 καρέ από τα οποία προήλθαν τα δείγματα των 16 καρέ, εφαρμόζοντας ένα τυχαίο flipping, όπως περιγράφεται στο [264], αλλά χωρίς τυχαίες περικοπές. Εφαρμόστηκαν στη συνέχεια οι ίδιοι χωρικοί μετασχηματισμοί και στα τμήματα των 16 καρέ αλλά και στον χάρτη εμφάνειας που προκύπτει από τα δεδομένα οφθαλμοκινητικής παρακολούθησης του μεσαίου καρέ, ο οποίος θεωρείται και ο ground truth χάρτης ολόκληρου του κλιπ.

Testing: Κατά τη διάρκεια της φάσης του testing αποκτούμε έναν χάρτη εκτίμησης της εμφάνειας ανά καρέ, χρησιμοποιώντας ένα κυλιόμενο παράθυρο των 16 καρέ με βήμα 1 χωρίς τυχαίο flipping.

7.3 Πειράματα

Method \ Dataset	DIEM					Coutrot1					Coutrot2				
	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑
STAViS [STA]	**0.5795**	**2.26**	**0.8838**	**0.6741**	**0.4824**	0.4722	2.11	**0.8686**	**0.5847**	**0.3935**	**0.7349**	**5.28**	**0.9581**	**0.7106**	**0.5111**
STAViS [ST]	0.5665	2.19	0.8792	0.6648	0.4719	0.4587	1.99	0.8617	0.5764	0.3842	0.6529	4.19	0.9405	0.6895	0.4470
DeepNet [152] [S]	0.4075	1.52	0.8321	0.6227	0.3183	0.3402	1.41	0.8248	0.5597	0.2732	0.3012	1.82	0.8966	0.6000	0.2019
DVA [158] [S]	0.4779	1.97	0.8547	0.641	0.3785	0.4306	2.07	0.8531	0.5783	0.3324	0.4634	3.45	0.9328	0.6324	0.2742
SAM [159] [S]	0.4930	2.05	0.8592	0.6446	0.4261	0.4329	2.11	0.8571	0.5768	0.3672	0.4194	3.02	0.9320	0.6152	0.3041
SalGAN [289] [S]	0.4868	1.89	0.8570	0.6609	0.3931	0.4161	1.85	0.8536	0.5799	0.3321	0.4398	2.96	0.9331	0.6183	0.2909
ACLNet [160, 161] [ST]	0.5229	2.02	0.8690	0.6221	0.4279	0.4253	1.92	0.8502	0.5429	0.3612	0.4485	3.16	0.9267	0.5943	0.3229
DeepVS [165] [ST]	0.4523	1.86	0.8406	0.6256	0.3923	0.3595	1.77	0.8306	0.5617	0.3174	0.4494	3.79	0.9255	0.6469	0.2590
TASED [167] [ST]	0.5579	2.16	0.8812	0.6579	0.4615	**0.4799**	**2.18**	0.8676	0.5808	0.3884	0.4375	3.17	0.9216	0.6118	0.3142

Πίνακας 7.2: Αποτελέσματα αξιολόγησης εμφάνειας για τις βάσεις δεδομένων DIEM, Coutrot1 και Coutrot2. Τα αποτελέσματα για την προτεινόμενη μέθοδο (STAViS [STA]) απεικονίζονται στην πρώτη γραμμή του πίνακα, ενώ η δεύτερη γραμμή αφορά στην εκδοχή του δικτύου μας που υπολογίζει οπτική-μόνο εμφάνεια. Στις περισσότερες περιπτώσεις, το προτεινόμενο δίκτυο ξεπερνά τις άλλες υπάρχουσες state-of-the-art μεθόδους για εκτίμηση εμφάνειας, σύμφωνα με τις 5 μετρικές αξιολόγησης. Το [STA] αναφέρεται σε χωρο-χρονικό οπτικοακουστικό, το [ST] σε χωρο-χρονικό οπτικό, ενώ το [S] υποδηλώνει χωρικό μόνο μοντέλο το οποίο εφαρμόζεται σε κάθε καρέ ανεξάρτητα.

Method \ Dataset	AVAD					SumMe					ETMD				
	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑
STAViS [STA]	**0.6086**	**3.18**	**0.9196**	**0.5936**	**0.4578**	0.4220	2.04	**0.8883**	**0.6562**	**0.3373**	**0.5690**	**2.94**	**0.9316**	0.7317	**0.4251**
STAViS [ST]	0.6041	3.07	0.9157	0.5900	0.4431	0.4180	1.98	0.8848	0.6477	0.3325	0.5602	2.84	0.9290	0.7278	0.4121
DeepNet [152] [S]	0.3831	1.85	0.8690	0.5616	0.2564	0.3320	1.55	0.8488	0.6451	0.2274	0.3879	1.90	0.8897	0.6992	0.2253
DVA [158] [S]	0.5247	3.00	0.8887	0.5820	0.3633	0.3983	2.14	0.8681	0.6686	0.2811	0.4965	2.72	0.9039	0.7288	0.3165
SAM [159] [S]	0.5279	2.99	0.9025	0.5777	0.4244	0.4041	**2.21**	0.8717	0.6728	0.3272	0.5068	2.78	0.9073	0.7310	0.3790
SalGAN [289] [S]	0.4912	2.55	0.8865	0.5799	0.3608	0.3978	1.97	0.8754	0.6882	0.2897	0.4765	2.46	0.9035	**0.7463**	0.3117
ACLNet [160, 161] [ST]	0.5809	3.17	0.9053	0.5600	0.4463	0.3795	1.79	0.8687	0.6092	0.2965	0.4771	2.36	0.9152	0.6752	0.3290
DeepVS [165] [ST]	0.5281	3.01	0.8968	0.5858	0.3914	0.3172	1.62	0.8422	0.6120	0.2622	0.4616	2.48	0.9041	0.6861	0.3495
TASED [167] [ST]	0.6006	3.16	0.9146	0.5898	0.4395	**0.4288**	2.10	0.8840	0.6570	0.3337	0.5093	2.63	0.9164	0.7117	0.3660

Πίνακας 7.3: Αποτελέσματα αξιολόγησης εμφάνειας για τις βάσεις δεδομένων AVAD, SumMe και ETMD. Τα αποτελέσματα για την προτεινόμενη μέθοδο (STAViS [STA]) απεικονίζονται στην πρώτη γραμμή του πίνακα, ενώ η δεύτερη γραμμή αφορά στην εκδοχή του δικτύου μας που υπολογίζει οπτική-μόνο εμφάνεια. Στις περισσότερες περιπτώσεις, το προτεινόμενο δίκτυο ξεπερνά τις άλλες υπάρχουσες state-of-the-art μεθόδους για εκτίμηση εμφάνειας, σύμφωνα με τις 5 μετρικές αξιολόγησης. Το [STA] αναφέρεται σε χωρο-χρονικό οπτικοακουστικό, το [ST] σε χωρο-χρονικό οπτικό, ενώ το [S] υποδηλώνει χωρικό μόνο μοντέλο το οποίο εφαρμόζεται σε κάθε καρέ ανεξάρτητα.

7.3.1 Πειραματικά Αποτελέσματα

Για να εκπαιδεύσουμε και να αξιολογήσουμε το προτεινόμενο οτπικοακουστικό δίκτυο εμφάνειας, χρησιμοποιούμε 6 βάσεις δεδομένων που περιέχουν οπτικοακουστικά δεδομένα οφθαλμοκινητικής παρακολούθησης: τις DIEM, AVAD, Coutrot1, Coutrot2, SumMe και ETMD. Οι βάσεις αυτές περιγράφηκαν σε προηγούμενη ενότητα.

Για την αξιολόγηση του προτεινόμενου δικτύου STAViS αρχικά μελετάμε τηην επίδοση των διαφόρων μεθόδων εντοπισμού ηχητικής πηγής και σύμμειξης που περιγράφηκαν στην ενότητα 7.2. Κατόπιν, επιλέγουμε το συνδυασμό αυτόν που επιτυγχάνει την καλύτερη επίδοση και συγκρίνουμε τη μέθοδό μας με 8 state-of-the-art μεθόδους οπτικής εμφάνειας χρησιμποιώντας τους κώδικες που έχουν γίνει δημόσια διαθέσιμοι, στις 6 βάσεις δεδομένων που αναφέρθηκαν και χρησιμοποιώντας 5 ευρέως χρησιμοποιούμενες μετρικές εμφάνειας [261]: Τις CC, NSS, AUC-Judd (AUC-J), shuffled AUC (sAUC) και SIM (similarity). Για το sAUC επιλέγουμε τα αρνητικά δείγματα από την ένωση των σημείων εστίασης όλων των συμμετεχόντων για όλα τα καρέ εκτός από αυτό για το οποίο υπολογίζουμε το AUC.

Ablation study: Όσον αφορά τον Πίνακα 7.1, μπορούμε να παρατηρήσουμε ότι ο bilinear μετασχηματισμός L_3 πετυχαίνει καλύτερη επίδοση σε σχέση με τις υπόλοιπες μεθόδους εντοπισμού του ήχου, εφόσον όπως αναφέρθηκε και προηγουμένως, οι υπόλοιπες μέθοδοι μπορούν να θεωρηθούν απλότερες εκδόσεις του bilinear μετασχηματισμού. Επιπλέον, αξίζει να σημειωθεί ότι οι μέθοδοι που βασίζονται μόνο στον χάρτη που προκύπτει από τον εντοπισμό του ήχου πετυχαίνουν αρκετά υψηλή επίδοση, ενώ βελτιώνονται περαιτέρω όταν συγχωνευτούν και με την οπτική εμφάνεια. Ειδικά όταν πολλαπλοί χάρτες συγχωνεύονται με τους οπτικούς με συνένωση S_3^{av}, επιτρέποντας έτσι τη μέγιστη αλληλεπίδραση μεταξύ των χαρακτηριστικών, τότε επιτυγχάνεται η καλύτερη επίδοση σχεδόν σε όλες τις μετρικές. Αρκετά καλή επίδοση επιτυγχάνεται και με το late fusion όπως φαίνεται στην τελευταία γραμμή του Πίνακα 7.1.

Συνολικά, η μέθοδος cosine similarity L_1 για τον εντοπισμό ηχητικής πηγής στο βίντεο, έχει τη χειρότερη επίδοση σε όλες τις περιπτώσεις, εφόσον δεν έχει και παραμέτρους για εκμάθηση, ενώ η τεχνική σύμμειξης S_2^{av} που στο παρελθόν έχει πετύχει καλά αποτελέσματα

με τεχνικές επεξεργασίας σήματος [268], δε φαίνεται κατάλληλη για προσεγγίσεις με βαθιά νευρωνικά δίκτυα. Επομένως, επιλέγουμε τη μέθοδο $L_3^{mul}S_3^{av}$ για το προτεινόμενο δίκτυο STAViS. Τα Σχήματα 7.4 και 7.5 απεικονίζουν κάποια ενδεικτικά καρέ από τις βάσεις δεδομένων Coutrot1 και Coutrot2, μαζί με τα δεδομένα οφθαλμοκινητικής παρακολούθησης που τους αντιστοιχούν και τους ground truth, οπτικούς και οπτικοακουστικούς χάρτες εμφάνειας όπως αυτοί προκύπτουν από το STAViS. Όπως μπορούμε να παρατηρήσουμε, το προτεινόμενο μοντέλο STAViS μοντελοποιεί καλύτερα την ανθρώπινη προσοχή. Τα σχήματα απεικονίζουν επίσης την εξέλιξη της καμπύλης NSS στο χρόνο για την οπτική και την οπτικοακουστική περίπτωση, για λόγους οπτικοποίησης και σύγκρισης.

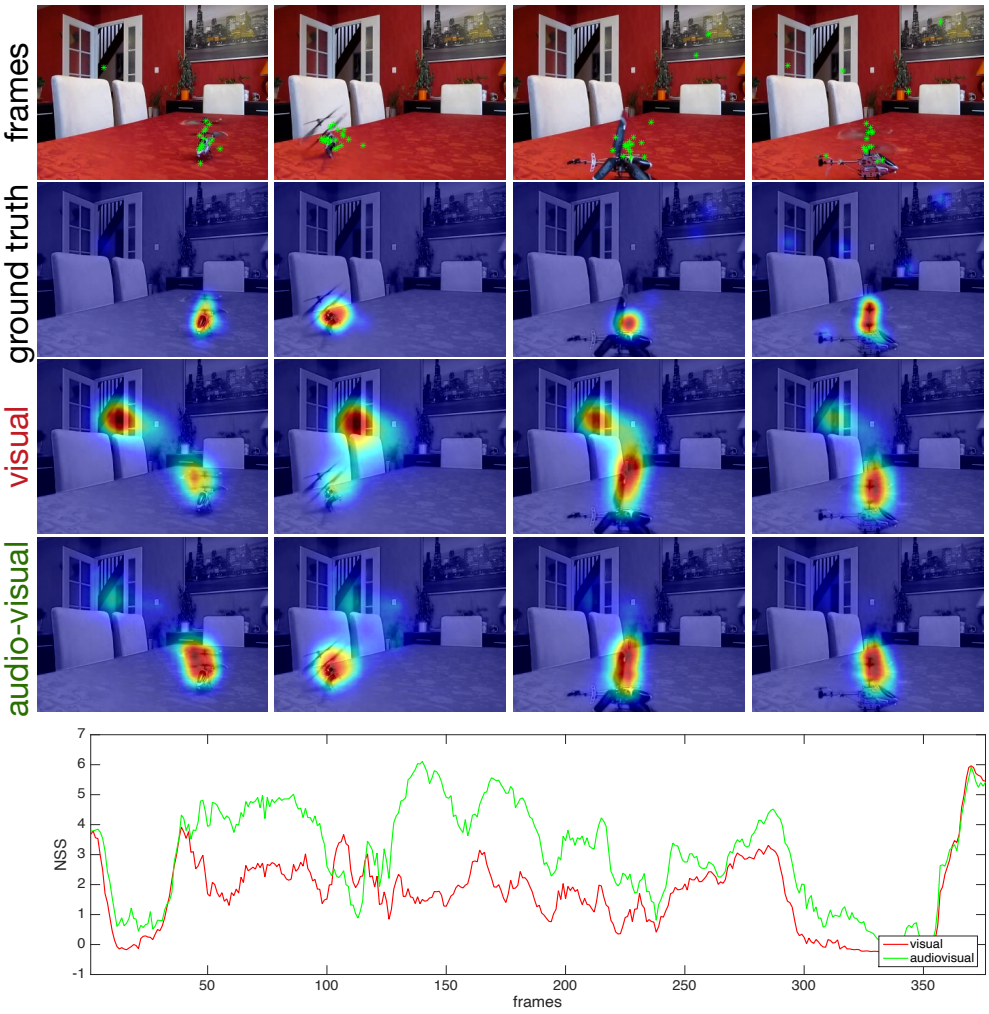
Σύγκριση με το state-of-the-art: Στη συνέχεια συγκρίνουμε την προτεινόμενη μέθοδο STAViS με 8 διαφορετικές state-of-the-art μεθόδους, τα αποτελέσματα της οποίας παρουσιάζονται ανά βάση για τις 5 μετρικές που υιοθετήθηκαν στους Πίνακες 7.2 και 7.3. Τα αποτελέσματα υπογραμμίζουν την αποτελεσματικότητα της προτεινόμενης οπτικοακουστικής μεθόδου, αφού ξεπερνά για όλες σχεδόν τις βάσεις και τις μετρικές τις υπόλοιπες state-of-the-art μεθόδους. Μπορούμε να παρατηρήσουμε ότι οι υπόλοιπες μέθοδοι εκτός από την TASED πετυχαίνουν μεν καλά αποτελέσματα σε κάποιες βάσεις μεμονωμένα, αλλά όχι με συνέπεια, κάτι που μπορεί να υποδηλώνει ότι έχουν καλή επίδοση σε μεμονωμένα είδη βίντεο αλλά όχι σε όλα τα είδη. Για παράδειγμα η μέθοδος ACLNet ενώ έχει καλή επίδοση στη βάση AVAD, δεν έχει το ίδιο καλή στη βάση SumMe η οποία περιέχει πιο αδόμητα, ερασιτεχνικά βίντεο. Η μέθοδος TASED πετυχαίνει καλή απόδοση, ξεπερνώντας την προτεινόμενη για κάποιες μετρικές στη βάση Coutrot1 και SumMe αλλά δεν επιτυγχάνει καλή επίδοση στην Coutrot2. Συνολικά, μια αρκετά καλή επίδοση επιτυγχάνεται και από την οπτική-μόνο μέθοδό μας.

Συζήτηση: Η προτεινόμενη χωρο-χρονική οπτικοακουστική μέθοδος STAViS έχει πολύ καλή με συνέπεια επίδοση σε όλες τις βάσεις δεδομένων, επιτυγχάνοντας την καλύτερη επίδοση σε σχέση με τις υπόλοιπες μεθόδους για την πλειοψηφία των βάσεων και των μετρικών. Αξίζει να σημειωθεί πάλι ότι οι βάσεις αυτές περιέχουν έναν μεγάλο αριθμό από ποικίλα και ανομοιόμορφα είδη βίντεο: Από ντοκιμαντέρ, όπου δεν υπάρχει οπτικοακουστική αντιστοιχία, μέχρι ταινίες του Χόλιγουντ και μη-επεξεργασμένα ερασιτεχνικά βίντεο. Επομένως το μοντέλο μας επιτυγχάνει καλή επίδοση "in-the-wild" , χωρίς πρότερη γνώση του περιεχομένου του βίντεο. Η επίδοση αυτή για ένα τόσο μεγάλο εύρος θα μπορούσε να υποδεικνύει ότι το μοντέλο μας είναι ικανό να μαθαίνει πότε, που και αν πρέπει να γίνεται η ενσωμάτωση της ακουστικής πληροφορίας στην οπτική. Επιπλέον, ειδικά στη βάση Coutrot2 που υπό μία έννοια είναι η πιο 'οπτικοακουστική' εφόσον όλα τα γεγονότα που περιέχει είναι οπτικοακουστικά, το μοντέλο μας ξεπερνά όλες τις άλλες μεθόδους με σημαντική διαφορά, ίσως επειδή μπορεί να μοντελοποιεί καλύτερα τις οπτικοακουστικές αντιστοιχίες. Για λόγους οπτικοποίησης, στο Σχήμα 7.6 απεικονίζεται ένα καρέ για κάθε βάση από τις ETMD, AVAD, DIEM και SumMe μαζί με τα αντίστοιχα δεδομένα οφθαλμοκινητικής παρακολούθησης στην πρώτη σειρά. Η δεύτερη σειρά περιέχει τον ground truth χάρτη εμφάνειας. Η τρίτη σειρά τους αντίστοιχους χάρτες εμφάνειας από το προτεινόμενο οπτικοακουστικό μοντέλο STAViS, ενώ οι υπόλοιπες 3 σειρές, τους ίδιους χάρτες για τις οπτικές μεθόδους ACLNet, DeepVS και TASED. Μπορούμε εύκολα να παρατηρήσουμε ότι τα αποτελέσματα μας είναι πιο κοντά στους πραγματικούς χάρτες

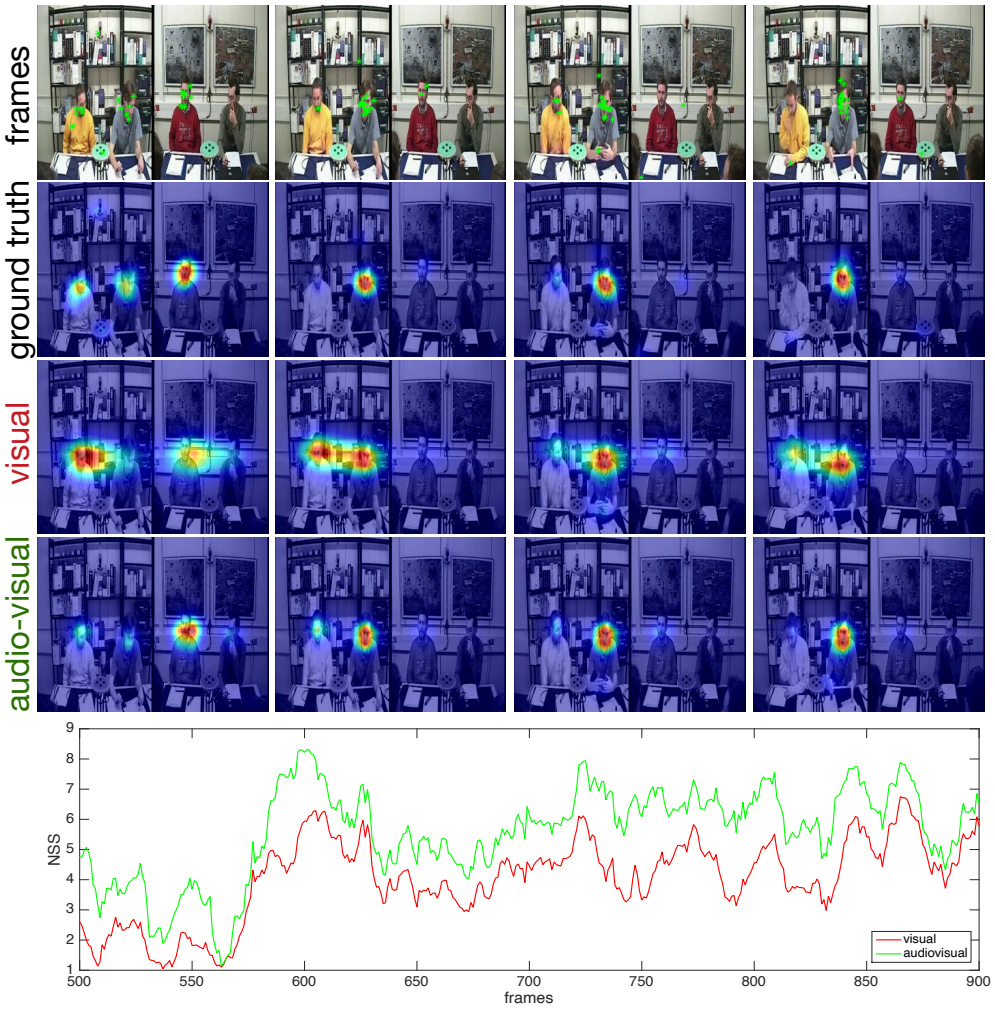
από ότι των υπόλοιπων μεθόδων.

7.4 Συμπεράσματα

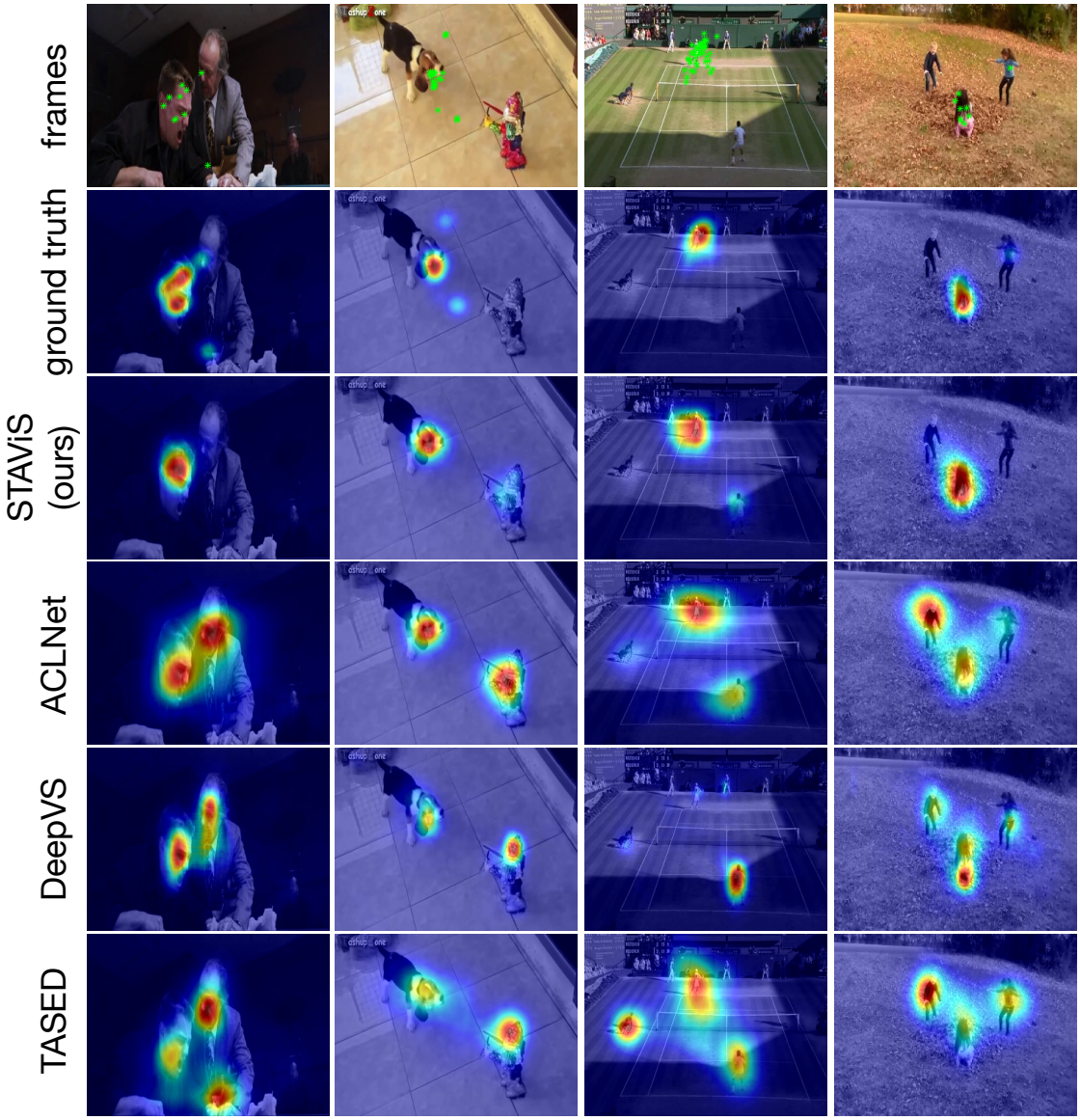
Στο κεφάλαιο αυτό παρουσιάστηκε η μέθοδος STAViS (Spatio-Temporal AudioVisual Saliency), ένα νέο χωρο-χρονικό οπτικοακουστικό δίκτυο το οποίο αντιμετωπίζει αποτελεσματικά το πρόβλημα της εκτίμησης των σημείων εστίασης της ανθρώπινης προσοχής σε βίντεο, δηλαδή της εμφάνειας. Αποτελείται από ένα state-of-the-art μοντέλο οπτικής εμφάνειας το οποίο έχει επεκταθεί στην παρούσα δουλειά, ένα state-of-the-art δίκτυο ακουστικής αναπαράστασης και δύο νέες μονάδες, μία για τον εντοπισμό ακουστικής πηγής σε βίντεο που παράγει έναν ή περισσότερους χάρτες ακουστικής εμφάνειας και μία για τη συγχώνευση των ακουστικών και οπτικών χαρτών ώστε να προκύψει ένας τελικός οπτικοακουστικός χάρτης εμφάνειας. Όλα τα συστατικά μέρη σχεδιάστηκαν, εκπαιδεύτηκαν end-to-end και αξιολογήθηκαν ενδελεχώς σε μια μεγάλη ποικιλία από βίντεο. Αποτελέσματα για 5 μετρικές σε 6 βάσεις δεδομένων και σύγκριση με 8 άλλες state-of-the-art μεθόδους υποδεικνύουν την καταλληλότητα και αποτελεσματικότητα της προτεινόμενης οπτικοακουστικής προσέγγισης για τη μοντελοποίηση της εμφάνειας.



Σχήμα 7.4: Ενδεικτικά καρέ από τη βάση Coutrot1 μαζί με τα δεδομένα οφθαλμοκινητικής παρακολούθησης που τους αντιστοιχούν, και τους αντίστοιχους ground truth, χωρο-χρονικούς οπτικούς και οπτικοακουστικούς χάρτες όπως υπολογίστηκαν από το προτεινόμενο δίκτυο STAViS (οπτικό-μόνο και οπτικοακουστικό). Επιπλέον απεικονίζεται η καμπύλη NSS σε σχέση με το χρόνο για την οπτική-μόνο και για την οπτικοακουστική περίπτωση.



Σχήμα 7.5: Ενδεικτικά καρέ από τη βάση Coutrot2 μαζί με τα δεδομένα οφθαλμοκινητικής παρακολούθησης που τους αντιστοιχούν, και τους αντίστοιχους ground truth, χωρο-χρονικούς οπτικούς και οπτικοακουστικούς χάρτες όπως υπολογίστηκαν από το προτεινόμενο δίκτυο STAViS (οπτικό-μόνο και οπτικοακουστικό). Επιπλέον απεικονίζεται η καμπύλη NSS σε σχέση με το χρόνο για την οπτική-μόνο και για την οπτικοακουστική περίπτωση.



Σχήμα 7.6: Ενδεικτικό καρέ, ένα για κάθε βάση από τις ETMD, AVAD, DIEM και SumMe με τα αντίστοιχα δεδομένα οφθαλμοκινητικής παρακολούθησης και τους χάρτες που προκύπτουν από το ground truth, το STAViS και άλλες state-of-the-art χωρο-χρονικές μεθόδους οπτικής εμφάνειας για συγκρίσεις.

Κεφάλαιο 8

Συμπεράσματα και μελλοντικές κατευθύνσεις

8.1 Συμπεράσματα

Στην εργασία αυτή παρουσιάστηκε η έρευνα που έχει γίνει σε δύο μεγάλες κατευθύνσεις: Την κατανόηση ακουστικής σκηνής σε περιβάλλοντα αλληλεπίδρασης ανθρώπου-μηχανής και την υπολογιστική μοντελοποίηση της οπτικοακουστικής προσοχής μέσω της οπτικοακουστικής εμφάνειας.

Όσον αφορά την πρώτη κατεύθυνση, παρουσιάστηκε ένα συνολικό σύστημα επεξεργασίας και αναγνώρισης φωνής, το οποίο περιλαμβάνει υπο-συστήματα εντοπισμού ομιλητή (ακουστικό ή οπτικοακουστικό), αποθορυβοποίησης ομιλίας με δύο είδη beamformer, τους MVDR, DSB ανάλογα με την εφαρμογή και αναγνώρισης φωνητικών εντολών στα Ελληνικά (με δυνατότητα αναγνώρισης και στα Αγγλικά). Η ερευνητική συνεισφορά είναι μεγαλύτερη όσον αφορά το υπο-σύστημα του εντοπισμού ομιλητή, όπου υλοποιήθηκε και παρουσιάστηκε μία μέθοδος εντοπισμού πιο εύρωστη σε θόρυβο και αντήχηση σε σχέση με άλλες μεθόδους για περιβάλλοντα έξυπνου σπιτιού. Επίσης, παρουσιάστηκε η εξέλιξη αυτού του συστήματος ώστε να ενσωματώνει και οπτική πληροφορία, όπου αυτή είναι διαθέσιμη, μέσω αισθητήρων Kinect με εφαρμογή σε αλληλεπίδραση ανθρώπου/παιδιού και ρομπότ. Ο οπτικοακουστικός εντοπισμός σημείωσε καλύτερη επίδοση από τον ακουστικό για τη συγκεκριμένη εφαρμογή. Στην κατεύθυνση της πολυκαναλικής επεξεργασίας φωνής, παρουσιάστηκαν επίσης αποτελέσματα αποθορυβοποίησης ομιλίας σε κάποιες βάσεις με δύσκολες συνθήκες, όπου συγκρίθηκαν οι δύο μέθοδοι beamforming αλλά και δύο είδη μικροφώνων μεταξύ τους, τα MEMS με τα πυκνωτικά. Επιπλεόν, παρουσιάστηκε το baseline σύστημα αναγνώρισης φωνητικών εντολών με γραμματική, του οποίου τα ακουστικά μοντέλα είναι GMM-HMM. Το σύστημα αυτό δουλεύει σε πραγματικό χρόνο και μπορεί να αναγνωρίσει εντολές που θέλει να δίνει ο χρήστης όποτε εκείνος θέλει, αφού είναι συνέχεια ενεργό. Το σύστημα αυτό έχει εφαρμοστεί τόσο σε έξυπνα περιβάλλοντα ('Σπιτάκι μου άκουσε'), όσο και σε εφαρμογές αλληλεπίδρασης ανθρώπου/παιδιών με ρομπότ. Για την τελευταία εφαρμογή, παρουσιάζονται τα αποτελέσματα της αναγνώρισης φωνής σε παιδιά για δύο παιχνίδια που σχεδιάστηκαν για αυτό το σκοπό.

Και στις δύο περιπτώσεις τα αποτελέσματα της αναγνώρισης είναι αρκετά υψηλά. Συγκρίνονται επίσης διάφοροι τρόποι προσαρμογής των μοντέλων σε παιδιά, ενήλικες, μικτά.

Τέλος, παρουσιάστηκε το σύστημα κατανόησης λόγου/διαλογικό σύστημα που ενσωματώθηκε στο ρομποτικό περιπατητήρα i-Walk. Το σύστημα αυτό αποτελείται από ένα state-of-the-art σύστημα αναγνώρισης φωνής, μία μονάδα κατανόησης λόγου, η οποία αναγνωρίζει την πρόθεση του χρήστη την οποία εκφράζει με τη φωνή του και έχει τη δυνατότητα να δίνει στο χρήστη ανατροφοδότηση ή να κάνει κάποια ενέργεια. Λειτουργεί σε πραγματικό χρόνο και είναι μέρος ενός μεγαλύτερου συστήματος, το οποίο διαθέτει επίσης αναγνώριση δράσεων και χειρονομιών και ανάλυση βάδισης. Τα συστήματα αυτά λειτουργούν και αυτόνομα αλλά και σε συνέργεια ώστε να εξασφαλίσουν την καλύτερη εμπειρία του χρήστη. Παρουσιάστηκαν επίσης τα δεδομένα που συλλέχθηκαν τόσο από υγιείς χρήστες όσο και από ασθενείς στο κέντρο αποκατάστασης Διάπλαση, τα οποία χρησιμοποιήθηκαν για να αναπτυχθούν οι σχετικοί αλγόριθμοι αλλά και για να αξιολογηθεί η επίδοση του συστήματος. Τα σχετικά πειράματα έδωσαν αρκετά ικανοποιητικά αποτελέσματα.

Ακόμη, συλλέχθηκε μία βάση πολυκαναλικών δεδομένων φωνής στα Ελληνικά σε ένα έξυπνο γραφείο δύο χώρων, η βάση 'Αθηνά', με δεδομένα από 20 ομιλητές, η οποία έγινε δημόσια διαθέσιμη. Για όλους τους παραπάνω αλγορίθμους, έχουν υλοποιηθεί κώδικες που έχουν γίνει δημόσια διαθέσιμοι.

Όσον αφορά τη δεύτερη ερευνητική κατεύθυνση, που αποτελεί και τον κύριο κορμό αυτής της διδακτορικής διατριβής και το πιο καινοτόμο κομμάτι της, παρουσιάστηκε η προσπάθεια υπολογιστικής μοντελοποίησης της προσοχής μέσω της οπτικοακουστικής εμφάνειας με σκοπό τη δημιουργία ενός δισδιάστατου μοντέλου πρόβλεψης των σημείων εστίασης του ματιού σε βίντεο. Με αφορμή την πληθώρα των συμπεριφορικών πειραμάτων και των νευροεπιστημών που έχουν μελετήσει το φαινόμενο της οπτικοακουστικής ενσωμάτωσης, όπου δηλαδή η ύπαρξη ακουστικού ερεθίσματος επηρεάζει την επεξεργασία του οπτικού από τον εγκέφαλο και αντίστροφα, επιχειρήσαμε να μοντελοποιήσουμε αυτό το φαινόμενο με υπολογιστικό τρόπο.

Αρχικά μελετήθηκε η μοντελοποίηση με τεχνικές επεξεργασίας σήματος. Συγκεκριμένα, αφού μελετήθηκαν εκτενώς διάφορα συμπεριφορικά πειράματα από τα οποία εξάχθηκαν συμπεράσματα, παρατηρήσεις και παράμετροι, διερευνήθηκαν διάφοροι τρόποι σύμμειξης της ακουστικής και της οπτικής (και συγκεκριμένα της χρονικής οπτικής) εμφάνειας, ώστε να αποδοθεί κατάλληλα η επιρροή του ακουστικού ερεθίσματος στο οπτικό και στη συνέχεια να συνδυαστούν και με τη χωρική οπτική συνιστώσα, καταλήγοντας σε ένα χωρο-χρονικο-ακουστικό μοντέλο εμφάνειας.

Οι διάφοροι τρόποι σύμμειξης είναι γενικοί και εφαρμόσιμοι σε μια πληθώρα χωρο-χρονικών οπτικών μοντέλων και ακουστικών. Επιλέχθηκαν κάποια μοντέλα, αντιπροσωπευτικά από κάθε κατηγορία για την αξιολόγηση. Η αξιολόγηση των μοντέλων αυτών έγινε σε τρία επίπεδα: Πρώτον, μέσω ποιοτικών συγκρίσεων των αποτελεσμάτων των μοντέλων με τα αποτελέσματα από συμπεριφορικά πειράματα, όπου υποδεικνύεται ότι τα μοντέλα μας έχουν την ίδια συμπεριφορά με αυτά, δεύτερον, μέσω πειραμάτων με δεδομένα οφθαλμοκινητικής παρακολούθησης σε πέντε διαφορετικές βάσεις δεδομένων. Σε όλες τις βάσεις και για όλες σχεδόν τις μετρικές, οι οπτικοακουστικοί συνδυασμοί των μοντέλων απέδωσαν καλύτερα από τους οπτικούς. Ανάμε-

σα στα διάφορα μοντέλα, την καλύτερη επίδοση παρουσίασαν τα βαθιά νευρωνικά δίκτυα για την πλειοψηφία των περιπτώσεων. Το τρίτο και τελευταίο επίπεδο αξιολόγησης διεξάχθηκε με δεδομένα fMRI ώστε να διαπιστωθεί αν πράγματι εμφανίζονται αντίστοιχες ενεργοποιήσεις στον ανθρώπινο εγκέφαλο. Τα αποτελέσματα έδειξαν ενεργοποιήσεις τόσο στον οπτικό όσο και στον ακουστικό φλοιό, κάτι το οποίο υποδεικνύει την καταλληλότητα της μοντελοποίησης που πραγματοποιήθηκε.

Στη συνέχεια, για την περαιτέρω βελτίωση της επίδοσης και την καλύτερη πρόβλεψη των σημείων εστίασης της ανθρώπινης προσοχής διερευνήθηκε η μοντελοποίηση με βαθιά νευρωνικά δίκτυα. Συγκεκριμένα, αναπτύχθηκε και παρουσιάστηκε το χωρο-χρονικό οπτικοακουστικό μοντέλο εμφάνειας STAViS, το οποίο αποτελείται από διάφορες μονάδες: ένα state-of-the-art δίκτυο χωρο-χρονικής οπτικής εμφάνειας, ένα state-of-the-art δίκτυο ακουστικών αναπαραστάσεων, από το οποίο αντλήσαμε τα ακουστικά χαρακτηριστικά, μία μονάδα η οποία ανιχνεύει από που προέρχεται ο ήχος πάνω στο βίντεο και η οποία έχει ως αποτέλεσμα έναν χωρο-χρονικό χάρτη ακουστικής εμφάνειας και μία μονάδα η οποία πραγματοποιεί τη συγχώνευση της οπτικής με την ακουστική εμφάνεια για την παραγωγή του τελικού χάρτη. Για τις δύο τελευταίες μονάδες εξετάστηκαν τρεις διαφορετικοί τρόποι παραγωγής του ακουστικού χάρτη και τρεις διαφορετικοί τρόποι σύμμειξης. Όλες οι παραπάνω μονάδες μαζί με τα αντίστοιχα losses είναι μέρος ενός συνολικού δικτύου, το οποίο εκπαιδεύτηκε end-to-end και αξιολογήθηκε σε 6 διαφορετικές βάσεις δεδομένων, για 5 διαφορετικές μετρικές και συγκρίθηκε με 8 διαφορετικές state-of-the-art μεθόδους. Τα αποτελέσματα της αξιολόγησης έδειξαν ότι το μοντέλο μας ξεπερνά σε απόδοση όλα τα υπόλοιπα σχεδόν σε όλες τις βάσεις και μετρικές. Η συνεπής επίδοση του σε όλες τις βάσεις, οι οποίες περιέχουν πολλά και διαφορετικά είδη βίντεο, υποδεικνύει πως είναι κατάλληλο και αποτελεσματικό για την πρόβλεψη των σημείων εστίασης της προσοχής του ανθρώπου 'in-the-wild'.

Τέλος, για τους σκοπούς της αξιολόγησης των δύο παραπάνω μοντέλων που αναπτύχθηκαν, συλλέχθηκε μία βάση με δεδομένα οφθαλμοκινητικής παρακολούθησης για δύο νέες βάσεις με οπτικοακουστικά ερεθίσματα, δηλαδή βίντεο. Συμμετείχαν 20 άτομα και το σύνολο των βίντεο ήταν 37 με συνολική διάρκεια 2 ώρες. Τα δεδομένα αυτά επίσης είναι δημόσια διαθέσιμα.

8.2 Μελλοντικές κατευθύνσεις

Η διατριβή αυτή επικεντρώθηκε σε δύο βασικές κατευθύνσεις: τη μοντελοποίηση της ανθρώπινης προσοχής και την πολυ-αισθητηριακή κατανόηση ακουστικής σκηνής. Μία πολύ σημαντική ερευνητική κατεύθυνση για το μέλλον θα ήταν η ενσωμάτωση των δύο αυτών κατευθύνσεων σε ρομποτικά συστήματα. Η μοντελοποίηση της προσοχής μπορεί να ανοίξει ένα νέο πεδίο για πιο έξυπνους ρομποτικούς βοηθούς οι οποίοι θα μπορούν να αντιλαμβάνονται τι συμβαίνει στο γύρω περιβάλλον και τι τραβάει την προσοχή του ανθρώπου, με σκοπό τη δημιουργία μιας αλληλεπίδρασης ανθρώπου-ρομπότ που θα πλησιάζει όσο το δυνατόν περισσότερο τη φυσική ανθρώπινη αλληλεπίδραση. Πιο συγκεκριμένα, λαμβάνοντας υπόψη τα διάφορα οπτικοακουστικά ερεθίσματα μέσω της οπτικοακουστικής εμφάνειας, καθώς και την ομιλία και τη θέση

των ανθρώπων, το ρομπότ ιδανικά θα στρέφει την 'προσοχή' του, δηλαδή το κεφάλι του ή το σώμα του, προς το σημείο ενδιαφέροντος, το οποίο σύμφωνα με το μοντέλο προσοχής είναι το πιο προεξέχον. Αυτό μπορεί να υλοποιηθεί με δύο βήματα:

- Αρχικά θα δημιουργηθεί ένα δισδιάστατο οπτικοακουστικό μοντέλο, το οποίο πέρα από την εμφάνεια, θα ενσωματώνει τη θέση του ομιλητή όπως αυτή υπολογίζεται από τον αλγόριθμο οπτικοακουστικού ή ακουστικού εντοπισμού ομιλητή, σαν βάρος που θα σταθμίζει κατάλληλα τους δισδιάστατους οπτικοακουστικούς χάρτες εμφάνειας. Η αξιολόγηση θα γίνει σε πραγματικά δεδομένα αλληλεπίδρασης ανθρώπου-ρομπότ από μία συμβατική κάμερα.

- Τελικός στόχος είναι η δημιουργία ενός τρισδιάστατου μοντέλου που θα ενσωματώνει όλα τα παραπάνω αλλά θα αξιοποιεί επιπλέον οπτικά και ακουστικά δεδομένα σε τρεις διαστάσεις με τη χρήση πολλαπλών αισθητήρων μικροφώνων και πολλαπλών καμερών βάθους. Η αξιολόγηση θα πραγματοποιηθεί σε πραγματικά δεδομένα αλληλεπίδρασης ανθρώπου-ρομπότ απο πολλαπλούς αισθητήρες.

Ο τελικός στόχος ενός τέτοιου συστήματος είναι να αποκτήσει ακόμα περισσότερη 'νοημοσύνη', επιτελώντας πολλές και διαφορετικές εργασίες και συνδυάζοντάς τες μεταξύ τους για να πάρει αποφάσεις, να κάνει κάποια ενέργεια, να δώσει κάποια ανατροφοδότηση. Όπως ακριβώς λειτουργεί και ο άνθρωπος.

Μία άλλη σημαντική κατεύθυνση θα ήταν η αξιοποίηση του οπτικοακουστικού μοντέλου προσοχής για τη δημιουργία περιλήψεων βίντεο. Υπάρχει μεγάλη βιβλιογραφία σε αυτό το πεδίο, οπότε θα μπορούσε να αξιολογηθεί εάν το μοντέλο αυτό, το οποίο έδωσε πολύ καλά αποτελέσματα στους διάφορους τρόπους με τους οποίους αξιολογήθηκε, λειτουργεί καλύτερα από παλιότερα μοντέλα για τη συγκεκριμένη εργασία. Ο όγκος των βίντεο που γίνεται διαθέσιμος καθημερινά είναι τεράστιος, επομένως ολοένα και αυξάνεται η ανάγκη για καλές και περιεκτικές περιλήψεις που θα μπορούν να βλέπουν οι χρήστες ώστε να αποφασίζουν εάν πραγματικά τους ενδιαφέρει το βίντεο και αν αξίζει να το παρακολουθήσουν ολόκληρο.

Τέλος, μία άλλη ερευνητική κατεύθυνση θα ήταν η δημιουργία ενός δικτύου πολλαπλών εργασιών, το οποίο θα μοιράζεται κοινά οπτικοακουστικά χαρακτηριστικά αλλά θα επιτελεί πολλές και διαφορετικές εργασίες, μεταξύ των οποίων η οπτικοακουστική χωρο-χρονική εμφάνεια, η περίληψη βίντεο, η αναγνώριση αντικειμένου/περιβάλλοντος, η αναγνώριση και ο εντοπισμός οπτικοακουστικού γεγονότος.

Βιβλιογραφία

[1] "Speech detection and speaker localization in domestic environments," [Online] Available at: `http://dirha.fbk.eu/hscma`.

[2] E. Van der Burg, C. N. L. Olivers, A. W. Bronkhorst, and J. Theeuwes, "Pip and pop: Nonspatial auditory signals improve spatial visual search," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 34, pp. 1053–1065, 2008.

[3] E. Van der Burg, J. Cass, C. N. L. Olivers, J. Theeuwes, and D. Alais, "Efficient visual search from synchronized auditory signals requires transient audiovisual events," *PLoS One*, vol. 5, no. 5, p. e10664, 2010.

[4] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, ""Your word is my command": Google search by voice: A case study," in *Advances in Speech Recognition*. Springer, 2010, pp. 61–90.

[5] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] W. K. Edwards and R. E. Grinter, "At home with ubiquitous computing: seven challenges," in *Ubicomp 2001: Ubiquitous Computing*, vol. LNCS-2201. Springer, 2001, pp. 256–272.

[7] M. Chan, D. Estève, C. Escriba, and E. Campo, "A review of smart homes - Present state and future challenges," *Computer Methods and Programs in Biomedicine*, vol. 91, no. 1, pp. 55–81, 2008.

[8] M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, B. Lecouteux, and P. Chahuara, "Evaluation of a context-aware voice interface for ambient assisted living: Qualitative user study vs. quantitative system evaluation," *ACM Trans. Access. Comput.*, vol. 7, no. 2, pp. 1–36, 2015.

[9] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 127–140, 2012.

[10] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications.* Springer, 2001.

[11] T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grezl, A. E. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[12] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2014, pp. 5542–5546.

[13] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," in *Proc. Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 172–176.

[14] M. Sehili, B. Lecouteux, M. Vacher, F. Portet, D. Istrate, B. Dorizzi, and J. Boudy, "Sound environment analysis in smart home," in *Ambient Intelligence.* Springer, 2012, vol. LNCS-7683, pp. 208–223.

[15] M. Vacher, B. Lecouteux, and F. Portet, "Multichannel automatic recognition of voice command in a multi-room smart home: An experiment involving seniors and users with visual impairment," in *Proc. Int. Conf. on Speech Communication and Technology*, 2014, pp. 1008–1012.

[16] E. Principi, S. Squartini, R. Bonfigli, G. Ferroni, and F. Piazza, "An integrated system for voice command recognition and emergency detection based on audio signals," *Journal of Expert Systems with Applications*, vol. 42, no. 13, pp. 5668–5683, 2015.

[17] J. A. Morales-Cordovilla, H. Pessentheiner, M. Hagmüller, and G. Kubin, "Room localization for distant speech recognition," in *Proc. Int. Conf. on Speech Communication and Technology*, 2014, pp. 2450–2453.

[18] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[19] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: An overview of challenge systems and outcomes," in *Proc. Automatic Speech Recognition and Understanding Work. (ASRU)*, 2013, pp. 162–167.

[20] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, and T. Nakatani, "Strategies for distant speech

recognition in reverberant environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, 2015.

[21] M. Matassoni, M. Omologo, D. Giuliani, and P. Svaizer, "Hidden Markov model training with contaminated speech material for distant-talking speech recognition," *Computer Speech and Language*, vol. 16, no. 2, pp. 205–223, 2002.

[22] M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo, "Impulse response estimation for robust speech recognition in a reverberant environment," in *Proc. European Signal Processing Conf.*, 2012, pp. 1668–1672.

[23] B. Lecouteux, M. Vacher, and F. Portet, "Distant speech recognition in a smart home: Comparison of several multisource ASRs in realistic conditions," in *Proc. Int. Conf. on Speech Communication and Technology*, 2011, pp. 2273–2276.

[24] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. Language Resources and Evaluation Conf.*, 2014, pp. 2629–2634.

[25] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2003, pp. 364–367.

[26] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia *et al.*, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Journal of Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 389–407, 2007.

[27] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction.* Springer, 2006, vol. LNCS-3869, pp. 28–39.

[28] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," in *Multimodal Technologies for Perception of Humans.* Springer, 2007, vol. LNCS-4112, pp. 311–322.

[29] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, "The sweet-home speech and multimodal corpus for home automation interaction," in *Proc. Language Resources and Evaluation Conf.*, 2014, pp. 4499–4506.

[30] A. Fleury, M. Vacher, F. Portet, P. Chahuara, and N. Noury, "A French corpus of audio and multimodal interactions in a health smart home," *Journal on Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 93–109, 2013.

[31] T. Giannakopoulos, N. Tatlas, T. Ganchev, and I. Potamitis, "A practical, real-time speech-driven home automation front-end," *IEEE Trans. on Consumer Electronics*, vol. 51, no. 2, pp. 514–523, 2005.

[32] M. Gavrilidou, G. Rehm, and H. Uszkoreit, *The Greek language in the digital age.* Springer, 2012.

[33] V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, and V. Diakoloukas, "Large vocabulary continuous speech recognition in greek: Corpus and an automatic dictation system," in *Proc. Int. Conf. on Speech Communication and Technology*, 2003, pp. 1565–1568.

[34] I. Rodomagoulakis, G. Potamianos, and P. Maragos, "Advances in large vocabulary continuous speech recognition in Greek: Modeling and nonlinear features," in *Proc. European Signal Processing Conf.*, 2013, pp. 1–5.

[35] D. Imseng, H. Bourlard, and P. N. Garner, "Using kl-divergence and multilingual information to improve asr for under-resourced languages," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2012, pp. 4869–4872.

[36] M. Goodrich and A. Schultz, "Human-robot interaction: a survey," *Foundations and Trends in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[37] E. Broadbent, R. Stafford, and B. MacDonald, "Acceptance of healthcare robots for the older population: review and future directions," *Int. J. Social Robotics*, vol. 1, no. 4, pp. 319–330, 2009.

[38] A. Zlatintsi, I. Rodomagoulakis, V. Pitsikalis, P. Koutras, N. Kardaris, X. Papageorgiou, C. Tzafestas, and P. Maragos, "Social human-robot interaction for the elderly: two real-life use cases," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2017, pp. 335–336.

[39] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural human-robot interaction using speech, head pose and gestures," in *IEEE Int. Conf. on Intelligent Robots and Systems*, 2004, pp. 2422–2427.

[40] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme, "Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions," in *Int. Conf. on Social Robotics*, 2015, pp. 327–336.

[41] M. Saerbeck, T. Schut, C. Bartneck, and M. Janse, "Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor," in *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, 2010, pp. 1613–1622.

[42] T. Belpaeme, P. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu *et al.*, "Multimodal child-robot

interaction: Building social bonds," *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 33–53, 2012.

[43] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard, "Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?" *Universal Access in the Information Society*, vol. 4, no. 2, pp. 105–120, 2005.

[44] G. Gordon, C. Breazeal, and S. Engel, "Can children catch curiosity from a social robot?" in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2015, pp. 91–98.

[45] S. Fernando, R. Moore, D. Cameron, E. Collins, A. Millings, A. Sharkey, and T. Prescott, "Automatic recognition of child speech for robotic applications in noisy environments," *arXiv preprint arXiv:1611.02695*, 2016.

[46] G. Skantze, "Predicting and regulating participation equality in human-robot conversations: Effects of age and gender," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2017, pp. 196–204.

[47] N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis, and P. Maragos, "A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands," in *ACM on Multimedia Conf.*, 2016, pp. 1169–1173.

[48] I. Bayer and T. Silbermann, "A multi modal approach to gesture recognition from audio and video data," in *Proc. Int. Conf. on Multimodal Interfaces*, 2013, pp. 461–466.

[49] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[50] C. Oertel, J. Gustafson, and A. Black, "Towards building an attentive artificial listener: On the perception of attentiveness in feedback utterances," in *Proc. Int. Conf. on Speech Communication and Technology*, 2016, pp. 2915–2919.

[51] J. Hemminghaus and S. Kopp, "Towards adaptive social behavior generation for assistive robots using reinforcement learning," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2017, pp. 332–340.

[52] J. Cech, R. Mittal, A. Deleforge, J. Sanchez-Riera, X. Alameda-Pineda, and R. Horaud, "Active-speaker detection and localization with microphones and cameras embedded into a robotic head," in *Proc. Humanoid Robots*, 2013, pp. 203–210.

[53] H. Löllmann, H. Barfuss, A. Deleforge, S. Meier, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," in *Proc. Speech Communication*, 2014, pp. 1–4.

[54] C. Evers, Y. Dorfan, S. Gannot, and P. Naylor, "Source tracking using moving microphone arrays for robot audition," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2017, pp. 6145–6149.

[55] S. Dubowsky, F. Genot, S. Godding, H. Kozono, A. Skwersky, H. Yu, and L. S. Yu, "Pamm-a robotic aid to the elderly for mobility assistance and monitoring: a" helping-hand" for the elderly," in *Int. Conf. on Robotics and Automation.* IEEE, 2000, pp. 570–576.

[56] A. Morris, R. Donamukkala, A. Kapuria, A. Steinfeld, J. T. Matthews, J. Dunbar-Jacob, and S. Thrun, "A robotic walker that provides guidance," in *Int. Conf. on Robotics and Automation.* IEEE, 2003, pp. 25–30.

[57] S. Jenkins and H. Draper, "Care, monitoring, and companionship: Views on care robots from older people and their carers," *International Journal of Social Robotics*, vol. 7, no. 5, pp. 673–683, 2015.

[58] A. Sharkey and N. Sharkey, "Children, the elderly, and interactive robots," *IEEE Robotics & Automation Magazine*, vol. 18, no. 1, pp. 32–38, 2011.

[59] D. Rodriguez-Losada, F. Matia, A. Jimenez, R. Galan, and G. Lacey, "Implementing map based navigation in guido, the robotic smartwalker," in *Int. Conf. on Robotics and Automation.* IEEE, 2005, pp. 3390–3395.

[60] V. Kulyukin, A. Kutiyanawala, E. LoPresti, J. Matthews, and R. Simpson, "iwalker: Toward a rollator-mounted wayfinding system for the elderly," in *IEEE Int. Conf. on RFID*, 2008, pp. 303–311.

[61] A. Frizera-Neto, R. Ceres, E. Rocon, and J. L. Pons, "Empowering and assisting natural human mobility: The simbiosis walker," *Int. Journal of Advanced Robotic Systems*, vol. 8, no. 3, 2011.

[62] M.-F. Chang, W.-H. Mou, C.-K. Liao, and L.-C. Fu, "Design and implementation of an active robotic walker for parkinson's patients," in *Proc. of SICE.* IEEE, 2012, pp. 2068–2073.

[63] J. Paulo, P. Peixoto, and U. J. Nunes, "Isr-aiwalker: Robotic walker for intuitive and safe mobility assistance and gait analysis," *IEEE Trans. Hum. Mach. Syst.*, vol. 47, no. 6, pp. 1110–1122, 2017.

[64] X. S. Papageorgiou, G. Chalvatzaki, A. C. Dometios, C. S. Tzafestas, and P. Maragos, "Intelligent assistive robotic systems for the elderly: Two real-life use cases," in *Proc. Int. Conf. PErvasive Technologies Related to Assistive Environments*, 2017, pp. 360–365.

[65] A. Guler, N. Kardaris, S. Chandra, V. Pitsikalis, C. Werner, K. Hauer, C. Tzafestas, P. Maragos, and I. Kokkinos, "Human joint angle estimation and gesture recognition for assistive robotic vision," in *Proc. European Conf. on Computer Vision.* Springer, 2016, pp. 415–431.

[66] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal human action recognition in assistive human-robot interaction," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2016, pp. 2702–2706.

[67] G. Chalvatzaki, X. S. Papageorgiou, C. S. Tzafestas, and P. Maragos, "Augmented human state estimation using interacting multiple model particle filters with probabilistic data association," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1872–1879, 2018.

[68] G. Chalvatzaki, P. Koutras, J. Hadfield, X. S. Papageorgiou, C. S. Tzafestas, and P. Maragos, "Lstm-based network for human gait stability prediction in an intelligent robotic rollator," in *Int. Conf. on Robotics and Automation.* IEEE, 2019, pp. 4225–4232.

[69] N. Efthymiou, P. Koutras, P. P. Filntisis, G. Potamianos, and P. Maragos, "Multi-view fusion for action recognition in child-robot interaction," in *Proc. IEEE Int. Conf. on Image Processing*, 2018, pp. 455–459.

[70] A. Tsiami, P. Koutras, N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos, "Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots," in *Int. Conf. on Robotics and Automation.* IEEE, 2018, pp. 1–8.

[71] A. Tsiami, P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Far-field audio-visual scene perception of multi-party human-robot interaction for children and adults," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing.* IEEE, 2018, pp. 6568–6572.

[72] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, 2015.

[73] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 588–595.

[74] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 914–927, 2013.

[75] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang, "Graph based skeleton motion representation and similarity measurement for action recognition," in *Proc. European Conf. on Computer Vision.* Springer, 2016, pp. 370–385.

[76] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Proc. IEEE Int. Conf. on Computer Vision*, 2013, pp. 2752–2759.

[77] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[78] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 1110–1118.

[79] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 1010–1019.

[80] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 1647–1656.

[81] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proc. IEEE Int. Conf. on Computer Vision*, 2015, pp. 4041–4049.

[82] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *Proc. AAAI Conf. on Artificial Intelligence*, 2016, pp. 3697–3703.

[83] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI Conf. on Artificial Intelligence*, 2017, pp. 4263–4270.

[84] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3d human pose estimation in rgbd images for robotic task learning," in *Int. Conf. on Robotics and Automation.* IEEE, 2018, pp. 1986–1992.

[85] A. Roitberg, A. Perzylo, N. Somani, M. Giuliani, M. Rickert, and A. Knoll, "Human activity recognition in the context of industrial human-robot interaction," in *Signal and Information Processing Association Ann. Summ. Conf.*, 2014, pp. 1–10.

[86] G. G. Chalvatzaki, G. Pavlakos, K. Maninis, X. S. Papageorgiou, V. Pitsikalis, C. S. Tzafestas, and P. Maragos, "Towards an intelligent robotic walker for assisted living using multimodal sensorial data," in *Int. Conf. Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies.* IEEE, 2014, pp. 156–159.

[87] F. Rezazadegan, S. Shirazi, B. Upcrofit, and M. Milford, "Action recognition: From static datasets to moving robots," in *Int. Conf. on Robotics and Automation.* IEEE, 2017, pp. 3185–3191.

[88] L. Piyathilaka and S. Kodagoda, "Human activity recognition for domestic robots," in *Field and Service Robotics.* Springer, 2015, pp. 395–408.

[89] M. A. Meredith and B. E. Stein, "Interactions among converging sensory inputs in the superior colliculus," *Science*, vol. 221, no. 4608, pp. 389–391, 1983.

[90] ——, "Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration," *Journal of Neurophysiology*, vol. 56, no. 3, pp. 640–662, 1986.

[91] A. Vatakis and C. Spence, "Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli," *Perception & Psychophysics*, vol. 69, no. 5, pp. 744–756, 2007.

[92] P. Maragos, A. Gros, A. Katsamanis, and G. Papandreou, "Cross-modal integration for performance improving in multimedia: A review," in *Multimodal Processing and Interaction: Audio, Video, Text.* Springer-Verlag, 2008, pp. 1–46.

[93] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[94] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[95] M. Keetels and J. Vroomen, "Sound affects the speed of visual processing," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 3, pp. 699–708, 2011.

[96] S. Gleiss and C. Kayser, "Eccentricity dependent auditory enhancement of visual stimulus detection but not discrimination," *Frontiers in Integrative Neuroscience*, vol. 7, no. 52, pp. 1–8, 2013.

[97] Q. Li, H. Yang, F. Sun, and J. Wu, "Spatiotemporal relationships among audiovisual stimuli modulate auditory facilitation of visual target discrimination," *Perception abstract*, vol. 44, no. 3, pp. 232–242, 2015.

[98] L. Chen and J. Vroomen, "Intersensory binding across space and time: A tutorial review," *Attention, Perception, & Psychophysics*, vol. 75, no. 5, pp. 790–811, 2013.

[99] M. O. Ernst, "A bayesian view on multimodal cue integration," *Human Body Perception From the Inside Out*, pp. 105–131, 2006.

[100] D. Alais and D. Burr, "The ventriloquist effect results from near-optimal bimodal integration," *Current Biology*, vol. 14, no. 3, pp. 257 – 262, 2004.

[101] H. Colonius and A. Diederich, "The optimal time window of visual-auditory integration: a reaction time analysis," *Frontiers in Integrative Neuroscience*, vol. 4, no. 11, pp. 1–8, 2010.

[102] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Temporal window of integration in auditory-visual speech perception," *Neuropsychologia*, vol. 45, no. 3, pp. 598–607, 2007.

[103] W. Fujisaki, S. Shimojo, M. Kashino, and S. Nishida, "Recalibration of audiovisual simultaneity," *Nature neuroscience*, vol. 7, no. 7, pp. 773–778, 2004.

[104] E. Van der Burg, D. Alais, and J. Cass, "Rapid recalibration to audiovisual asynchrony," *The Journal of Neuroscience*, vol. 33, no. 37, pp. 14 633–14 637, 2013.

[105] E. Van der Burg, E. Orchard-Mills, and D. Alais, "Rapid temporal recalibration is unique to audiovisual stimuli," *Experimental brain research*, vol. 233, no. 1, pp. 53–59, 2015.

[106] W. Fujisaki and S. Nishida, "Audio–tactile superiority over visuo–tactile and audio–visual combinations in the temporal resolution of synchrony perception," *Experimental brain research*, vol. 198, no. 2-3, pp. 245–259, 2009.

[107] W. Roseboom, S. Nishida, and D. H. Arnold, "The sliding window of audio–visual simultaneity," *Journal of Vision*, vol. 9, no. 12, pp. 1–8, 2009.

[108] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[109] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[110] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, 2013.

[111] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[112] R. Milanese, "Detecting salient regions in an image: From biological evidence to computer implementation," Ph.D. dissertation, University of Geneva, 1993.

[113] S. Baluja and D. Pomerleau, "Using a saliency map for active spatial selective attention: Implementation & initial results," in *Advances in neural information processing systems*, 1994, pp. 451–458.

[114] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, Oct. 1995.

[115] E. Niebur and C. Koch, "Control of selective visual attention: Modeling the where pathway," in *Proc. Advances in Neural Information Processing Systems*, 1995.

[116] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI*, vol. 5200. International Society for Optics and Photonics, 2003, pp. 64–79.

[117] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *J. Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[118] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, ser. Lecture Notes in Computer Science. Springer, 2006, vol. 3899.

[119] O. L. Meur, P. L. Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, 2006.

[120] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2007, pp. 545–552.

[121] O. L. Meur, P. L. Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, no. 19, pp. 2483 – 2498, 2007.

[122] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *Int'l. J. Comput. Vis.*, vol. 82, no. 3, pp. 231–243, 2009.

[123] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Spatiotemporal features for action recognition and salient event detection," *Cognitive Computation, special issue on Saliency, attention, visual search and picture scanning*, vol. 3, no. 1, pp. 167–184, 2011.

[124] A. Garcia-Diaz, X. R. Fernandez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization." *Image and Vision Computing*, vol. 30, no. 1, pp. 51–64, 2012.

[125] A. Torralba, "Modeling global scene factors in attention," *J. Opt. Soc. Amer. A*, vol. 20, pp. 1407–1418, 2003.

[126] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," in *Advances in neural information processing systems*, 2006, pp. 547–554.

[127] I. Gkioulekas, G. Evangelopoulos, and P. Maragos, "Spatial bayesian surprise for image saliency and quality assessment," in *Proc. IEEE Int. Conf. on Image Processing*, 2010, pp. 1081–1084.

[128] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proc. IEEE Int. Conf. on Image Processing*, 2003, pp. 250–253.

[129] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of vision*, vol. 8, no. 7, pp. 1–20, 2008.

[130] L. Zhang, M. H. Tong, and G. W, "Sunday: Saliency using natural statistics for dynamic analysis of scenes," in *Proc. Cognitive Science Society Conference*, 2009, pp. 2944–2949.

[131] T. Kadir and M. Brady, "Saliency, scale and image description," *Int'l. J. Comput. Vis.*, vol. 45, no. 2, pp. 83–105, 2001.

[132] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, 2006, pp. 155–162.

[133] X. Hou and L. Zhang, "Dynamic visual attention: searching for coding length increments," in *Proc. Advances in Neural Information Processing Systems*, 2009, pp. 681–688.

[134] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Advances in neural information processing systems*, 2004, pp. 481–488.

[135] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 989–1005, 2009.

[136] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, pp. 1–27, 2009.

[137] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit, "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Process.: Image Communication*, vol. 28, no. 6, pp. 642 – 658, 2013.

[138] N. Bruce and J. Tsotsos, "Spatiotemporal saliency: Towards a hierarchical representation of visual saliency," in *Int'l Workshop on Attention and Performance in Comp. Vis.*, 2008, pp. 98–111.

[139] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, 2010.

[140] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[141] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[142] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, 2010.

[143] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, 2012.

[144] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Proc. European Conf. on Computer Vision*, 2012, pp. 116–129.

[145] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[146] M. Mancas, N. Riche, J. Leroy, and B. Gosselin, "Abnormal motion selection in crowds using bottom-up saliency," in *Proc. IEEE Int. Conf. on Image Processing*, 2011, pp. 175–178.

[147] T. V. Nguyen, M. Xu, G. Gao, M. Kankanhalli, Q. Tian, and S. Yan, "Static saliency vs. dynamic saliency: a comparative study," in *Proc. ACM Int. Conf. on Multimedia*, 2013, pp. 987–996.

[148] P. Koutras and P. Maragos, "A perceptually based spatio-temporal computational framework for visual saliency estimation," *Signal Process.: Image Communication*, vol. 38, pp. 15–31, 2015.

[149] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, 2006.

[150] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 2798–2805.

[151] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet," in *Int. Conf. on Learning Representations Workshops (ICLRW)*, 2015.

[152] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 598–606.

[153] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. on Computer Vision*, 2015, pp. 262–270.

[154] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 362–370.

[155] S. Jetley, N. Murray, and E. Vig, "End-to-end saliency mapping via probability distribution prediction," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 5753–5761.

[156] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. a. Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition Workshop*, 2017.

[157] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1688–1698, 2018.

[158] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, 2018.

[159] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, 2018.

[160] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 4894–4903.

[161] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–17, 2019.

[162] P. Linardos, E. Mohedano, J. J. Nieto, K. McGuinness, X. Giro-i Nieto, and N. E. O'Connor, "Simple vs complex temporal recurrences for video saliency prediction," in *British Machine Vision Conf. (BMVC)*, 2019.

[163] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatio-temporal residual attentive networks," *IEEE Trans. Image Process.*, vol. 29, pp. 1113–1126, 2019.

[164] G. Leifman, D. Rudoy, T. Swedish, E. Bayro-Corrochano, and R. Raskar, "Learning gaze transitions from depth to improve video saliency estimation," in *Proc. IEEE Int. Conf. on Computer Vision*, 2017, pp. 1698–1707.

[165] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "Deepvs: A deep learning based video saliency prediction approach," in *Proc. European Conf. on Computer Vision*, 2018, pp. 602–617.

[166] S. Gorji and J. J. Clark, "Going from image to video saliency: Augmenting image salience with dynamic attentional push," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 7501–7511.

[167] K. Min and J. J. Corso, "Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *Proc. IEEE Int. Conf. on Computer Vision*, 2019, pp. 2394–2403.

[168] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[169] E. M. Kaya and M. Elhilali, "A temporal saliency map for modeling auditory attention," in *Proc. Information Sciences and Systems (CISS)*, 2012, pp. 1–6.

[170] O. Kalinli and S. S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech." in *Proc. Int. Conf. on Speech Communication and Technology*, 2007, pp. 1941–1944.

[171] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[172] B. Schauerte and R. Stiefelhagen, ""wow!" bayesian surprise for salient acoustic event detection," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2013, pp. 6402–6406.

[173] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub," in *Int. Conf. on Robotics and Automation*, 2008, pp. 962–967.

[174] B. Schauerte, B. Kühn, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention for object-based scene analysis," in *IEEE Int. Conf. on Intelligent Robots and Systems*, 2011, pp. 1173–1179.

[175] S. Ramenahalli, D. R. Mendat, S. Dura-Bernal, E. Culurciello, E. Nieburt, and A. Andreou, "Audio-visual saliency map: Overview, basic models and hardware implementation," in *Proc. Information Sciences and Systems (CISS)*, 2013, pp. 1–6.

[176] R. Ratajczak, D. Pellerin, Q. Labourey, and C. Garbay, "A fast audiovisual attention model for human detection and localization on a companion robot," in *Int. Conf. on Applications and Systems of Visual Paradigms*, 2016.

[177] N. Sidaty, M.-C. Larabi, and A. Saadane, "Toward an audiovisual attention model for multimodal video content," *Neurocomputing*, pp. 94–111, 2017.

[178] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis, "Video event detection and summarization using audio, visual and text saliency," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2009, pp. 3553–3556.

[179] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[180] P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, and A. Potamianos, "Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization," in *Proc. IEEE Int. Conf. on Image Processing*, 2015, pp. 4361–4365.

[181] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *Journal of Vision*, vol. 14, no. 8, pp. 1–17, 2014.

[182] ——, "An audiovisual attention model for natural conversation scenes," in *Proc. IEEE Int. Conf. on Image Processing*, 2014, pp. 1100–1104.

[183] ——, "Multimodal saliency models for videos," in *From Human Attention to Computational Attention: A Multidisciplinary Approach*, M. Mancas, V. P. Ferrera, N. Riche, and J. G. Taylor, Eds. Springer New York, 2016, pp. 291–304.

[184] G. Song, "Effect of sound in videos on gaze: Contribution to audio-visual saliency modeling," Ph.D. dissertation, Universite de Grenoble, 2013.

[185] X. Min, G. Zhai, C. Hu, and K. Gu, "Fixation prediction through multimodal analysis," in *Proc. IEEE Int. Conf. on Visual Communications and Image Processing*, 2015.

[186] X. Min, G. Zhai, K. Gu, and X. Yang, "Fixation prediction through multimodal analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 13, no. 1, 2017.

[187] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. on Computer Vision*, 2017, pp. 609–617.

[188] ——, "Objects that sound," in *Proc. European Conf. on Computer Vision*, 2018, pp. 435–451.

[189] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon, "Learning to localize sound source in visual scenes," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 4358–4366.

[190] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. European Conf. on Computer Vision*, 2018, pp. 631–648.

[191] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. European Conf. on Computer Vision*, 2018, pp. 247–263.

[192] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. European Conf. on Computer Vision*, 2018, pp. 570–586.

[193] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in *Proc. IEEE Int. Conf. on Computer Vision*, 2019, pp. 1735–1744.

[194] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proc. Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[195] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *Advances in Neural Information Processing Systems*, 2018, pp. 7763–7774.

[196] M. Chan, E. Campo, D. Estève, and J.-Y. Fourniols, "Smart homes – current features and future perspectives," *Maturitas*, vol. 64, no. 2, pp. 90–97, 2009.

[197] M. P. Poland, C. D. Nugent, H. Wang, and L. Chen, "Smart home research: Projects and issues," *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 1, no. 4, pp. 32–45, 2009.

[198] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects," *Personal and Ubiquitous Computing*, vol. 17, pp. 127–144, 2013.

[199] P. Naylor and N. Gaubitch, Eds., *Speech Dereverberation.* Springer, 2010.

[200] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications.* Springer, 2001.

[201] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech and Audio Process.*, vol. 4, no. 2, pp. 148–152, 1996.

[202] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 6, pp. 791–803, 2003.

[203] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Speech and Audio Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[204] C. Zhang, D. Florêncio, and Z. Zhang, "Why does phat work well in low noise, reverberative environments?" in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2008, pp. 2565 – 2568.

[205] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing.* Springer, 2008, vol. 1.

[206] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech and Audio Process.*, vol. 5, no. 3, pp. 288–292, 1997.

[207] M. Brandstein, J. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Computer Speech and Language*, vol. 9, no. 2, pp. 153–169, 1995.

[208] I. Rodomagoulakis, P. Giannoulis, Z.-I. Skordilis, P. Maragos, and G. Potamianos, "Experiments on far-field multichannel speech processing in smart homes," in *Int. Conf. on Digital Signal Processing*, vol. 1. IEEE, 2013, pp. 121–124.

[209] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Processing*, vol. 59, no. 3, pp. 253–266, 1997.

[210] E. Jan and J. Flanagan, "Sound source localization in reverberant environments using an outlier elimination algorithm," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 3. IEEE, 1996, pp. 1321–1324.

[211] "The DIRHA (Distance-speech Interaction for Robust Home Applications) EU project," [Online] Availble at: `http://dirha.fbk.eu/`.

[212] A. Brutti, "Speech acoustic scene analysis and interpretation: Data collections," fBK. [Online].

[213] A. Brutti, M. Omologo, P. Svaizer, and C. Zieger, "Classification of acoustic maps to determine speaker position and orientation from a distributed microphone network," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, vol. 4, 2007, pp. 493–496.

[214] J. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[215] H. Do, H. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, vol. 1, 2007, pp. 121–124.

[216] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[217] I. Gebru, C. Evers, P. Naylor, and R. Horaud, "Audio-visual tracking by density approximation in a sequential Bayesian filtering framework," in *Proc. Hands-free Speech Communication and Microphone Arrays Workshop*, 2017, pp. 71–75.

[218] G. Garau, A. Dielmann, and H. Bourlard, "Audio-visual synchronisation for speaker diarisation," in *Proc. Int. Conf. on Speech Communication and Technology*, 2010, pp. 2654–2657.

[219] V. Minotto, C. Jung, and B. Lee, "Multimodal multi-channel on-line speaker diarization using sensor fusion through svm," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1694–1705, 2015.

[220] B. V. Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 5, pp. 4–24, 1988.

[221] H. Van Trees, *Optimum Array Processing.* Wiley, 2002.

[222] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Communication*, vol. 49, no. 7, pp. 657–666, 2007.

[223] K. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer, 2001, pp. 39–60.

[224] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Proc. Sensor Array and Multichannel Signal Processing Workshop*, 2002, pp. 209–213.

[225] Y. Ephraim and D. Mallah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[226] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[227] J. Neumann Jr. and K. Gabriel, "A fully-integrated CMOS-MEMS audio microphone," in *Proc. Int. Conf. on Transducers, Solid-State Sensors, Actuators and Microsystems*, vol. 1, 2003, pp. 230–233.

[228] E. Zwyssig, M. Lincoln, and S. Renals, "A digital microphone array for distant speech recognition," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2010, pp. 5106–5109.

[229] *MP34DT01 MEMS audio sensor omnidirectional digital microphone datasheet*, MP34DT01, STMicroelectronics, 2013, [Online]. Available: http://www.st.com/web/en/resource/technical/document/datasheet/DM00039779.pdf.

[230] A. Tsiami, I. Rodomagoulakis, P. Giannoulis, A. Katsamanis, G. Potamianos, and P. Maragos, "ATHENA: A Greek multi-sensory database for home automation control," in *Proc. Interspeech*, 2014, pp. 1608–1612.

[231] "Robot Operating System (ROS)," http://www.ros.org/about-ros/.

[232] C. Ishi, S. Matsuda, T. Kanda, T. Jitsuhiro, H. Ishiguro, S. Nakamura, and N. Hagita, "A robust speech recognition system for communication robots in noisy environments," *IEEE Trans. Robot.*, vol. 24, no. 3, pp. 759–763, 2008.

[233] M. Wölfel and J. McDonough, *Distant speech recognition.* John Wiley & Sons, 2009.

[234] A. Katsamanis, I. Rodomagoulakis, G. Potamianos, P. Maragos, and A. Tsiami, "Robust far-field spoken command recognition for home automation combining adaptation and multichannel processing," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2014, pp. 5547–5551.

[235] I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, and P. Maragos, "Room-localized spoken command recognition in multi-room, multi-microphone environments," *Computer Speech and Language*, vol. 46, pp. 419–443, 2017.

[236] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. Audio Engineering Society Convention*, 2000.

[237] S. Young, "Using pomdps for dialog management," in *Spoken Language Technology Workshop.* IEEE, 2006, pp. 8–13.

[238] "Google cloud speech-to-text," https://cloud.google.com/speech-to-text/.

[239] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open source language understanding and dialogue management," in *Proc. Advances in Neural Information Processing Systems*, 2017.

[240] "RASA," http://rasa.com.

[241] "RASA," http://https://github.com/RasaHQ.

[242] "DIAPLASIS Rehabilitation Center," https://www.diaplasis.eu/.

[243] A. Temko, D. Macho, C. Nadeu, and C. Segura, "UPC-TALP database of isolated acoustic events," *Internal UPC report*, 2005.

[244] A. Fleury, M. Vacher, F. Portet, P. Chahuara, and N. Noury, "A multimodal corpus recorded in a health smart home," in *Proc. Language Resources and Evaluation Conf.*, 2010, pp. 99–105.

[245] "ELAN (EUDICO Linguistic Annotator)," [Online] Available at: http://tla.mpi.nl/tools/tla-tools/elan/.

[246] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. European Conf. on Computer Vision*, 2014, pp. 505–520.

[247] G. Schillaci, S. Bodiroža, and V. V. Hafner, "Evaluating the effect of saliency detection and attention manipulation in human-robot interaction," *Int. Journal of Social Robotics*, vol. 5, no. 1, pp. 139–152, 2013.

[248] A. Tsiami, A. Katsamanis, P. Maragos, and A. Vatakis, "Towards a behaviorally-validated computational audiovisual saliency model," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2016, pp. 2847–2851.

[249] C. Bordier, F. Puja, and E. Macaluso, "Sensory processing during viewing of cinematographic material: Computational modeling and functional neuroimaging," *Neuroimage*, vol. 67, pp. 213–226, 2013.

[250] J. W. Gebhard and G. H. Mowbray, "On discriminating the rate of visual flicker and auditory flutter," *The American Journal of psychology*, vol. 72, no. 4, pp. 521–529, 1959.

[251] S. Shimojo and L. Shams, "Sensory modalities are not separate modalities: plasticity and interactions," *Current Opinion in Neurobiology*, vol. 11, no. 4, pp. 505–509, 2001.

[252] Y. Wada, N. Kitagawa, and K. Noguchi, "Audio–visual integration in temporal perception," *Int. journal of psychophysiology*, vol. 50, no. 1, pp. 117–124, 2003.

[253] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[254] L. Shams, Y. Kamitani, and S. Shimojo, "What you see is what you hear." *Nature*, vol. 408, p. 788, 2000.

[255] R. B. Welch, L. D. DutionHurt, and D. H. Warren, "Contributions of audition and vision to temporal rate perception," *Perception & Psychophysics*, vol. 39, no. 4, pp. 294–300, 1986.

[256] R. Sekuler, A. B. Sekuler, and R. Lau, "Sound alters visual motion perception." *Nature*, vol. 385, no. 6614, p. 308, 1997.

[257] C. V. Parise, V. Harrar, M. O. Ernst, and C. Spence, "Cross-correlation between auditory and visual signals promotes multisensory integration," *Multisensory Research*, vol. 26, no. 3, pp. 307–316, 2013.

[258] M. Rolf, M. Hanheide, and K. J. Rohlfing, "Attention via synchrony: Making use of multimodal cues in social learning," *IEEE Trans. Autonomous Mental Development*, vol. 1, no. 1, pp. 55–67, 2009.

[259] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," *Advances in Neural Information Processing Systems*, vol. 12, pp. 813–819, 2000.

[260] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, 2013.

[261] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *arXiv preprint arXiv:1604.03605*, 2016.

[262] L. Zhaoping and L. Zhe, "Primary visual cortex as a saliency map: A parameter-free prediction and its test by behavioral data," *PLOS Computational Biology*, vol. 11, no. 10, pp. 1–39, 2015.

[263] H. J. Müller and P. M. Rabbitt, "Reflexive and voluntary orienting of visual attention: Time course of activation and resistance to interruption." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, no. 2, pp. 315–330, 1989.

[264] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. European Conf. on Computer Vision*, 2016, pp. 20–36.

[265] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Joint Pattern Recognition Symposium.* Springer, 2007, pp. 214–223.

[266] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2013, pp. 1147–1154.

[267] P. K. Mital, T. J. Smith, R. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cognitive Computation*, vol. 3, no. 1, pp. 5–24, 2011.

[268] A. Tsiami, P. Koutras, A. Katsamanis, A. Vatakis, and P. Maragos, "A behaviorally inspired fusion approach for computational audiovisual saliency modeling," *Signal Processing: Image Communication*, vol. 76, pp. 186 – 200, 2019.

[269] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," *arXiv preprint arXiv:1801.07424*, 2018.

[270] A. Coutrot, N. Guyader, G. Ionescu, and A. Caplier, "Video viewing: do auditory salient events capture visual attention?" *Annals of Telecommunications*, vol. 69, no. 1, pp. 89–97, 2014.

[271] T. Liu, X. Hu, X. Li, M. Chen, J. Han, and L. Guo, "Merging neuroimaging and multimedia: methods, opportunities and challenges," *IEEE Trans. Human-Machine Systems*, vol. 44, no. 2, pp. 270–280, 2014.

[272] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, 1979.

[273] P. Koutras, A. Katsamanis, and P. Maragos, "Predicting eyes' fixations in movie videos: Visual saliency experiments on a new eye-tracking database," in *International Conference on Engineering Psychology and Cognitive Ergonomics*, 2014, pp. 183–194.

[274] P. Koutras, G. Panagiotaropoulou, A. Tsiami, and P. Maragos, "Audio-visual temporal saliency modeling validated by fmri data," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition Workshop*, 2018, pp. 2000–2010.

[275] G. Panagiotaropoulou, P. Koutras, A. Katsamanis, P. Maragos, A. Zlatintsi, A. Protopapas, E. Karavasilis, and N. Smyrnis, "fMRI-based perceptual validation of a computational model for visual and auditory saliency in videos," in *Proc. IEEE Int. Conf. on Image Processing*, 2016, pp. 699–703.

[276] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. on Learning Representations (ICLR)*, 2015.

[277] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[278] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. on Computer Vision*, 2017, pp. 2961–2969.

[279] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[280] J. Ray, H. Wang, D. Tran, Y. Wang, M. Feiszli, L. Torresani, and M. Paluri, "Scenes-objects-actions: A multi-task, multi-label video dataset," in *Proc. European Conf. on Computer Vision*, 2018, pp. 635–651.

[281] P. Koutras and P. Maragos, "Susinet: See, understand and summarize it," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition Workshop*, 2019.

[282] P. Primus, H. Eghbal-zadeh, D. Eitelsebner, K. Koutini, A. Arzt, and G. Widmer, "Exploiting parallel audio recordings to enforce device invariance in cnn-based acoustic scene classification," *arXiv preprint arXiv:1909.02869*, 2019.

[283] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2013, pp. 6645–6649.

[284] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2018, pp. 4774–4778.

[285] H. R. Tavakoli, A. Borji, E. Rahtu, and J. Kannala, "Dave: A deep audio-visual embedding for dynamic saliency prediction," *arXiv preprint arXiv:1905.10693*, 2019.

[286] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[287] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. on Computer Vision*, 2015, pp. 1395–1403.

[288] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 221–230.

[289] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," *Computer Vision and Image Understanding*, pp. 1–9, 2018.

Παράρτημα Α΄

Λίστα δημοσιεύσεων

Δημοσιεύσεις σε διεθνή περιοδικά με κριτές

[1] **A. Tsiami**, P. Koutras, A. Katsamanis, A. Vatakis and P. Maragos. "A behaviorally inspired fusion approach for computational audiovisual saliency modeling". In Signal Processing: Image Communication. vol. 76, pp.186-200, Aug. 2019.

[2] I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, **A. Tsiami** and P. Maragos. "Room-localized spoken command recognition in multi-room, multi-microphone environments". In Computer Speech & Language, vol.46, pp.419-443 , Nov.2017.

[3] N. Efthymiou, P.P. Filntisis, P. Koutras, **A. Tsiami**, J. Hadfield, G. Potamianos and P. Maragos, "ChildBot: Multi-Robot Perception and Interaction with Children". International Journal of Social Robotics (under review)

Δημοσιεύσεις σε διεθνή συνέδρια με κριτές

[1] A. Katsamanis, I. Rodomagoulakis, G. Potamianos, P. Maragos and **A. Tsiami**. "Robust Far-Field Spoken Command Recognition for Home Automation Combining Adaptation and Multichannel Processing", In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2014, Florence, Italy.

[2] P. Giannoulis, **A. Tsiami**, I. Rodomagoulakis, A. Katsamanis, G. Potamianos and P. Maragos. "The Athena-RC System for Speech Activity Detection and Speaker Localization in the DIRHA Smart Home", In Proc. of 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA) 2014, Nancy, France.

[3] **A. Tsiami**, A. Katsamanis, P. Maragos and G. Potamianos. "Experiments in Acoustic Source Localization Using Sparse Arrays in Adverse Indoors Environments". In Proc. of 22nd European Signal Processing Conference (EUSIPCO) 2014, Lisbon, Portugal.

[4] **A. Tsiami**, I. Rodomagoulakis, P. Giannoulis, A. Katsamanis, G. Potamianos and P. Maragos. "ATHENA: A Greek Multi-Sensory Database for Home Automation Control". In Proc.

Interspeech 2014, Singapore.

[**5**] Z. I. Skordilis, **A. Tsiami**, P. Maragos, G. Potamianos, L. Spelgatti and R. Sannino. "Multi-channel Speech Enhancement Using MEMS Microphones". In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2015, Brisbane, Australia.

[**6**] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, **A. Tsiami** and P. Maragos. "Multimodal Human Action Recognition in Assistive Human-Robot Interaction". In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2016, Shanghai, China.

[**7**] **A. Tsiami**, A. Katsamanis, P. Maragos and A. Vatakis. "Towards a behaviorally-validated computational audiovisual saliency model". In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2016, Shanghai, China.

[**8**] A. Brutti, **A. Tsiami**, A. Katsamanis and P. Maragos. "A Phase-Based Time-Frequency Masking for Multi-Channel Speech Enhancement in Domestic Environments". In Proc. Interspeech 2016, San Fransisco, USA.

[**9**] A. Dometios, **A. Tsiami**, A. Arvanitakis, P. Giannoulis, X. Papageorgiou, C. Tzafestas and P. Maragos. "Integrated Speech-based Perception System for User Adaptive Robot Motion Planning in Assistive Bath Scenarios". In Proc. of 25th European Signal Processing Conference (EUSIPCO) - Workshop:"MultiLearn 2017 - Multimodal processing, modeling and learning for human-computer/robot interaction applications", 2017, Kos, Greece.

[**10**] **A. Tsiami**, P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos and P. Maragos. "Far-field Audio-Visual Scene Perception of Multi-party Human-Robot Interaction for Children and Adults". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018, Calgary, Canada.

[**11**] **A. Tsiami**, P. Koutras, N. Efthymiou, P. P. Filntisis, G. Potamianos and P. Maragos. "Multi3: Multi-sensory Perception System for Multi-modal Child Interaction with Multiple Robots". In Proc. of IEEE International Conference on Robotics and Automation (ICRA) 2018, Brisbane, Australia.

[**12**] P. Koutras, G. Panagiotaropoulou, **A. Tsiami** and P. Maragos. "Audio-Visual Temporal Saliency Modeling Validated by fMRI Data". In Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2018, Salt Lake City, USA.

[**13**] **A. Tsiami**, P. Koutras and P. Maragos, "STAViS: Spatio-Temporal AudioVisual Saliency Network", submitted in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2020) (to appear)

[**14**] G. Chalvatzaki*, P. Koutras*, **A. Tsiami***, C.S. Tzafestas and P. Maragos. "I-Walk Intelligent Assessment System: Activity, Mobility, Intention, Communication", submitted in Proc. of the International Conference on Robotics and Automation (IROS-2020) (under review)