



National Technical University of Athens
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF SIGNALS, CONTROL AND ROBOTICS

**Multimodal Deep Learning for Emotion
Recognition and Expression Synthesis with
Applications in Human-Robot Interaction**

Ph.D. Dissertation

PANAGIOTIS PARASKEVAS FILNTISIS

Dipl.-Ing. in Electrical and Computer Engineering, NTUA

Supervisor: Prof. Petros Maragos

Athens, November 2022



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΟΡΑΣΗΣ ΥΠΟΛΟΓΙΣΤΩΝ, ΕΠΙΚΟΙΝΩΝΙΑΣ ΛΟΓΟΥ ΚΑΙ
ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΩΝ

Πολυτροπική βαθιά μάθηση για την αναγνώριση
συναισθημάτων και τη σύνθεση εκφράσεων
προσώπου με εφαρμογές στην αλληλεπίδραση
ανθρώπου-ρομπότ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΤΟΥ

Παναγιώτη Παρασκευά Φιλντίση

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

Επιβλέπων: Πέτρος Μαραγκός, Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2022



Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Πολυτροπική βαθιά μάθηση για την αναγνώριση
συναισθημάτων και τη σύνθεση εκφράσεων
προσώπου με εφαρμογές στην αλληλεπίδραση
ανθρώπου-ρομπότ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

Παναγιώτη Παρασκευά Φιλντίση

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

Συμβουλευτική Επιτροπή: Πέτρος Μαραγκός, Καθηγητής
Αλέξανδρος Ποταμιάνος, Αναπληρωτής Καθηγητής
Γεράσιμος Ποταμιάνος, Αναπληρωτής Καθηγητής

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 7η Νοεμβρίου 2022.

(Υπογραφή)

(Υπογραφή)

(Υπογραφή)

(Υπογραφή)

.....
Πέτρος Μαραγκός
Καθηγητής
Ε.Μ.Π.

(Υπογραφή)

.....
Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής
Ε.Μ.Π.

(Υπογραφή)

.....
Γεράσιμος Ποταμιάνος
Αναπληρωτής Καθηγητής
Πανεπιστήμιο Θεσσαλίας

(Υπογραφή)

.....
Κωνσταντίνος Τζαφέστας
Αναπληρωτής Καθηγητής
Ε.Μ.Π.

.....
Αθανάσιος Κατσαμάνης
Ερευνητής Β'
Ερευνητικό Κέντρο ΑΘΗΝΑ

.....
Στέφανος Κόλλιας
Καθηγητής
Ε.Μ.Π.

.....
Αναστάσιος Ρούσσος
Ερευνητής Β'
I.T.E.

Αθήνα, Νοέμβριος 2022

(Υπογραφή)

.....
Παναγιώτης Παρασκευάς Φιλντίσης

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

Copyright ©–All rights reserved Παναγιώτης Παρασκευάς Φιλντίσης, 2022.

Με επιφύλαξη παντός δικαιώματος.

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 το συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Ο χώρος του “affective computing” είναι ένας συναρπαστικός καινούργιος τομέας έρευνας που έχει ως στόχο να παρέχει στους υπολογιστές και τα ρομπότ τη δυνατότητα αναγνώρισης, έκφρασης, μοντελοποίησης αλλά και «αίσθησης» συναισθημάτων. Το διεπιστημονικό αυτό πεδίο του “affective computing” αντλεί πόρους από την επιστήμη των υπολογιστών, τα μαθηματικά, τις γνωσιακές επιστήμες και την ψυχολογία. Σε αυτή τη διατριβή, η οποία χωρίζεται σε δύο κύρια μέρη, διερευνούμε δύο πτυχές του συγκεκριμένου πεδίου: η πρώτη πτυχή είναι η «αναγνώριση συναισθήματος» και η δεύτερη είναι η «σύνθεση της έκφρασης του συναισθήματος». Οι δύο αυτές κατευθύνσεις αποτελούν τις πιο σημαντικές πτυχές που πρέπει να εξετάσει κανείς όταν χτίζει συστήματα αλληλεπίδρασης ανθρώπου-ρομπότ. Για το σκοπό αυτό, στο πρώτο μέρος, διερευνούμε και μελετάμε διάφορα κανάλια που περιέχουν πολύτιμες πληροφορίες για την αναγνώριση των συναισθημάτων ενός ανθρώπου και σχεδιάζουμε αρχιτεκτονικές βασισμένες σε βαθιά μάθηση, που μπορούν να συνδυάσουν αποτελεσματικά πληροφορίες από τα κανάλια αυτά, με απώτερο στόχο την ανάπτυξη ενός συστήματος για σεναρία αλληλεπίδρασης ανθρώπου/παιδιού με ρομπότ. Ενώ η αναγνώριση συναισθημάτων στο παρελθόν έχει ως επί το πλείστον επικεντρωθεί στις εκφράσεις του προσώπου και στην ομιλία, κατά τη διατριβή αυτή λαμβάνουμε υπόψη τη γλώσσα του σώματος του ανθρώπου, τη σκηνή στην οποία βρίσκεται, καθώς και τη σημασιολογική έννοια των συναισθημάτων. Στο δεύτερο μέρος, αρχικά ενισχύουμε τις υπάρχουσες μεθόδους για τη σύνθεση της οπτικοακουστικής ομιλίας, δίνοντάς τους τη δυνατότητα να συνδυάζουν και να εκφράζουν συναισθήματα σε διαφορετικά επίπεδα έντασης. Στη συνέχεια, σχεδιάζουμε ένα μοντέλο βαθιάς μάθησης με σκοπό τη σύνθεση οπτικοακουστικής ομιλίας που επιτυγχάνει υψηλό επίπεδο ρεαλισμού και εκφραστικότητας, ξεπερνώντας τις προηγούμενες μεθόδους. Τέλος, παρουσιάζουμε την πρώτη μέθοδο για τρισδιάστατη ανακατασκευή προσώπων από βίντεο μίας όψης, με έμφαση στα χαρακτηριστικά και τη γεωμετρία του στόματος κατά την ομιλία. Η μέθοδος αυτή παρακάμπτει τη συνηθισμένη απαίτηση για κοπιώδη συλλογή μεγάλου πλήθους τρισδιάστατων δεδομένων υψηλής πιστότητας, προσφέροντας έναν εύκολο και καινοτόμο τρόπο για την απόκτηση τρισδιάστατων δεδομένων εκφραστικών προσώπων από βίντεο.

Λέξεις Κλειδιά: Αναγνώριση Συναισθήματος, Γλώσσα Σώματος, Σκηνή, Σύνθεση Συναισθήματος, Οπτικοακουστικός, Βαθιά Μάθηση, Τρισδιάστατη Ανακατασκευή, 3D Morphable Models, Οπτική Ομιλία

Abstract

Affective computing is an exciting new research area with the goal of equipping computers and robots with the capability of recognizing, expressing, modeling, and even “feeling” emotions. An interdisciplinary field, affective computing draws resources from computer science, mathematics, cognitive sciences, and psychology. In this thesis, which is split into two major parts, we explore two aspects of affective computing; namely “emotion recognition” and “expression synthesis” since they constitute the most important aspects one needs to consider when building human-robot interaction systems. To this end, in the first part, we explore and study various information streams that contain valuable information for recognizing the emotions of a human, and design architectures based on deep learning, that can efficiently combine information from these streams, with the ultimate goal of deploying the system for human-robot interaction, with an emphasis in child-robot interaction scenarios. While traditional approaches for emotion recognition have mostly focused on facial expressions and speech, we take into account the body language the context, and also employ embeddings that accurately capture the semantic distances of discrete emotions. In the second part, we first enhance existing methods for audiovisual speech synthesis, by giving them the capabilities to both combine, and express emotions in different intensity levels. Then, we design a deep learning-based architecture for expressive audiovisual speech synthesis which achieves a high level of realism and expressiveness, outperforming previous methods. Lastly, we present the first method for visual speech aware monocular perceptual 3D reconstruction in the wild. This work tackles the traditional bottleneck of data collection for high-fidelity 3D ground truth data and offers the field of affective computing a way for easier acquisition of expressive 3D facial data data from monocular videos.

Keywords: Emotion Recognition, Affect, Body Language, Context, Expression Synthesis, AudioVisual, Deep Learning, 3D Reconstruction, Visual Speech, 3D Morphable Models

Πρόλογος

Με τη διατριβή αυτή ολοκληρώνεται μια μακρά πορεία επτά χρόνων από την οποία εξέρχομαι ένα εντελώς διαφορετικό άτομο, όχι μόνο σε επίπεδο γνώσεων, αλλά κυρίως σε επίπεδο εμπειριών. Κοιτώντας πίσω νιώθω πλήρης και ευγνώμων για όλα αυτά που έζησα στη διαδρομή αυτή: φιλίες, γέλια και άγχη, χαρές και απογοητεύσεις, αβεβαιότητα και σιγουριά.

Πρωτίστως, θα ήθελα να ευχαριστήσω τον Καθ. Πέτρο Μαραγκό, για την ευκαιρία που μου έδωσε να εκπονήσω αρχικά τη διπλωματική μου εργασία, και στη συνέχεια τη διδακτορική μου διατριβή υπό την επίβλεψη του στο εργαστήριο IRAL. Η υποστήριξη του καθ'όλη τη διάρκεια του διδακτορικού και η καθοδήγηση του αποδείχθηκαν καθοριστικές για την ερευνητική μου εξέλιξη. Παράλληλα, θα ήθελα να ευχαριστήσω το Νάσο Κατσαμάνη, ο οποίος με βοήθησε εκτενώς στα πρώτα μου βήματα στο τομέα της έρευνας, και χάρη στον οποίο βρήκα το αντικείμενο με το οποίο καταπιάστηκα στη διατριβή αυτή. Πολυ σημαντική ήταν επίσης η συμβολή του Καθ. Γεράσιμου Ποταμιάνου στα πρώτα χρόνια της διατριβής μου. Τέλος, ιδιαίτερης αξίας τα τελευταία δύο χρόνια ήταν η παρουσία του Τάσσου Ρούσσου, η συνεργασία με τον οποίο βοήθησε σημαντικά στην ερευνητική μου πορεία.

Βέβαια, το σημαντικότερο πράγμα που αποκομίζω στο τέλος της μακράς αυτής διαδρομής είναι οι ανθρώπινες σχέσεις που άνοιξαν κατά τη διάρκεια της. Το εργαστήριο υπήρξε όχι μόνο χώρος ανάπτυξης επαγγελματικών σχέσεων και ώσμωσης ιδεών, αλλά χώρος άκμασης δυνατών φιλιών που συνεχίζουν να αντέχουν στο πέρασμα του χρόνου. Πρώτη και σημαντικότερη η Νίκη, με την οποία ξεκινήσαμε και πορευτήκαμε μαζί, και ήταν και είναι επαγγελματικό, και κυρίως αναντικατάστατο ψυχολογικό στήριγμα. Στη συνέχεια, ο Γιώργος, ο Πέτρος, και η Αντιγόνη, με τους οποίους μοιραστήκαμε όχι μόνο ξενύχτια πάνω από ρομπότ, υπολογιστές, δημοσιεύσεις, αλλά και πολλά γέλια, ανησυχίες και σχέψεις. Επιπλέον η Ξανθή, η Γεωργία, ο Πάρης, ο Παναγιώτης, η Τζο, ο Μάνος, ο Νίκος, ο Χρήστος, η Δάφνη, με τους όποιες έχουμε ζήσει και ελπίζω να συνεχίσουμε να ζούμε πολλές στιγμές, εντός και εκτός εργαστηρίου. Και φυσικά ο Θανάσης, η Άμη, η Βίκυ, η Φωτεινή, η Δέσποινα, η Νάνσυ. Πραγματικά συνεχίζει να με εντυπωσιάζει ακόμα και τώρα το πλήθος των φιλιών που μπόρεσαν να αναπτυχθούν μέσα στο επαγγελματικό περιβάλλον του εργαστηρίου.

Ξεχωριστά θέλω να ευχαριστήσω τους φίλους μου που μου έχουν σταθεί όλα αυτά τα χρόνια: τον Γιώργο, τον Κωνσταντίνο, την Άννα, τον Φώτη, τη Στελίνα, τον Δημήτρη, τον Λευτέρη, τον Γιώργο, τη Νικολέτα, τη Χριστιάννα, την Κατερίνα, τον Στέλιο, τον Μαρίνο.

Τίποτα από όλα αυτά δε θα ήταν εφικτό χωρίς την αμέριστη και απύθμενη στήριξη και αγάπη των γονέων μου, του πατέρα μου Επαμεινώνδα και της μητέρα μου Παναγιώτας, καθώς και του αδερφού μου Αναστάσιου, ο οποίος πάντα είναι εκεί για μένα και του έλαχε το φορτίο του να με προσέχει. Η διατριβή αυτή είναι αφιερωμένη σε αυτούς.

στους γονείς μου, Επαμεινώνδα και Παναγιώτα,
και στον αδερφό μου Αναστάσιο

Εκτεταμένη Περίληψη

Εισαγωγή

Τα συναισθήματα και η ικανότητα έκφρασής τους αποτελούν θεμελιώδη πτυχή της ανθρώπινης συμπεριφοράς. Για χιλιάδες χρόνια, τα συναισθήματα έχουν εξελιχθεί μαζί με την ανθρωπότητα και έχουν τόσο ενδοπροσωπικές όσο και διαπροσωπικές λειτουργίες. Βοηθούν στην ταχεία λήψη αποφάσεων, επηρεάζουν τις σκέψεις και είναι συνυφασμένα με τη γνωστική διαδικασία. Διαπροσωπικά, τα συναισθήματα μας βοηθούν να κατανοήσουμε καλύτερα τους άλλους, χτίζοντας ισχυρότερους κοινωνικούς δεσμούς μέσω της ενσυναίσθησης και της συναισθηματικής εκφραστικότητας.

Ο βρόχος της επικοινωνίας μεταξύ ενός ρομπότ/υπολογιστή και ενός ανθρώπου έχει δύο πλευρές. Αναφορικά με την πρώτη, το ρομπότ πρέπει να είναι σε θέση να μοντελοποιεί και να αντιλαμβάνεται τα ανθρώπινα συναισθήματα ενώ η δεύτερη αφορά την ικανότητά του να εκφράζει συναισθήματα.

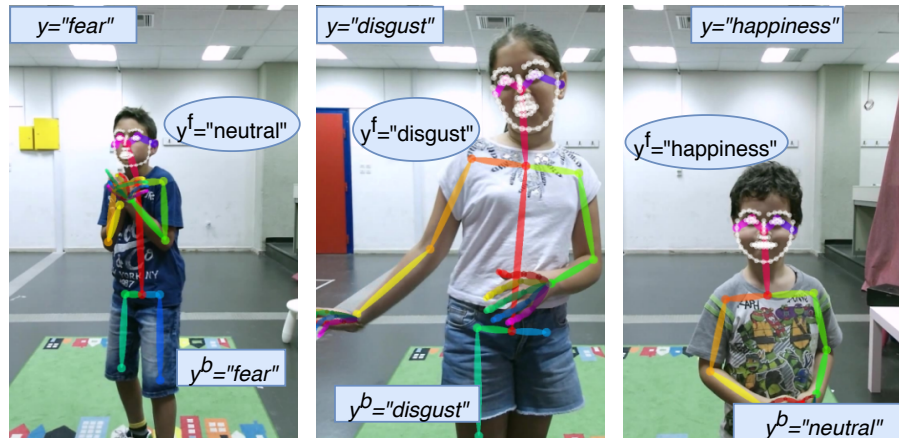
Στόχος της παρούσας διατριβής είναι η ανάπτυξη μεθόδων που να αποδίδουν στους ρομποτικούς πράκτορες τις εξής ικανότητες: 1) να αντιλαμβάνονται τις ανθρώπινες εκφράσεις και συναισθήματα και 2) να εκφράζονται με τρόπο που ομοιάζει άνθρωπο. Κατά συνέπεια, η διατριβή χωρίζεται σε δύο βασικά μέρη:

1. Αναγνώριση Συναισθημάτων: Στο πρώτο μέρος, ο ρομποτικός πράκτορας αναλαμβάνει το ρόλο του δέκτη στον βρόχο επικοινωνίας μεταξύ ανθρώπου και ρομπότ. Στην περίπτωση αυτή, αναζητούμε και ανακαλύπτουμε πρόσθετες ροές πληροφοριών που μπορούν να αξιοποιηθούν ώστε να ενισχυθεί η αυτόματη αναγνώριση συναισθημάτων.

2. Σύνθεση Εκφράσεων: Στο δεύτερο μέρος είναι η σειρά του ρομποτικού πράκτορα να λάβει ενεργητικό ρόλο στο βρόχο της επικοινωνίας. Εδώ μελετάμε την έκφραση συναισθημάτων υπό το πρίσμα του οπτικοακουστικού λόγου. Η δουλειά μας σε αυτό το μέρος μπορεί να χωριστεί σε δύο διαφορετικές κατευθύνσεις: στην πρώτη κατεύθυνση, σχεδιάζουμε μεθόδους για **εκφραστική** σύνθεση οπτικοακουστικού λόγου. Στη δεύτερη κατεύθυνση, αντιμετωπίζουμε ένα διαφορετικό πρόβλημα: την τρισδιάστατη ανακατασκευή προσώπων από βίντεο μίας όψης, με έμφαση στη διατήρηση των χαρακτηριστικών του στόματος κατά τη διάρκεια της ομιλίας του.

Συνδυασμός Γλώσσας Σώματος και Εκφράσεων Προσώπου για την Αναγνώριση Συναισθημάτων

Στο πρώτο κεφάλαιο του Μέρους 1 αυτής της διατριβής, μελετάμε τη σημασία της γλώσσας του σώματος κατά την έκφραση συναισθημάτων και την επίδρασή της στην ευρωστία και την ακρίβεια των συστημάτων αυτόματης αναγνώρισης συναισθημάτων. Ενώ παραδοσιακά το κυρίαρχο μέσο μελέτης στην έρευνα των συναισθημάτων ήταν το πρόσωπο [Ekman and Friesen, 1967], πιο πρόσφατες μελέτες έχουν τονίσει τη σημασία της γλώσσας του σώματος, υποδηλώνοντας ότι το συναίσθημα μεταδίδεται, στις περισσότερες περιπτώσεις, εξίσου μέσω των

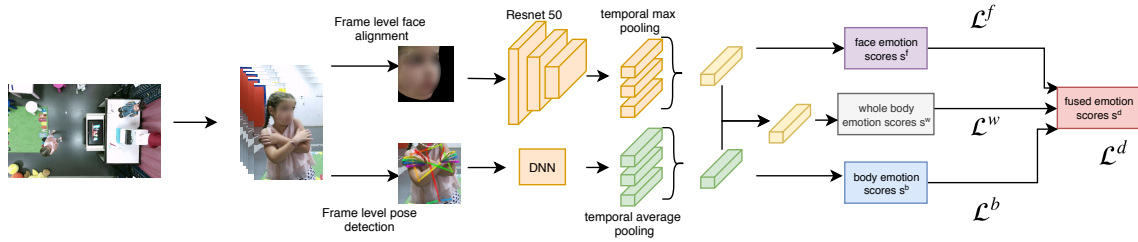


Σχήμα 1: Ιεραρχικές πολλαπλές ετικέτες για την αναγνώριση συναισθημάτων από το σώμα και το πρόσωπο. Το y υποδηλώνει την ετικέτα των συναισθημάτων ολόκληρου του σώματος, το y^f την ετικέτα της έκφρασης του προσώπου και το y^b την ετικέτα έκφρασης του σώματος.

εκφράσεων του σώματος και του προσώπου [De Gelder, 2009, Wallbott, 1998], ενώ τόσο η στατική στάση του σώματος όσο και η δυναμική [Atkinson et al., 2004, Calvo et al., 2015] συμβάλλουν στην αντίληψη των συναισθημάτων. Ωστόσο, μέχρι σήμερα, οι περισσότερες έρευνες για την αυτόματη αναγνώριση συναισθημάτων έχουν επικεντρωθεί κυρίως στις εκφράσεις του προσώπου [Jung et al., 2015, Kuo et al., 2018], με λίγες μόνο να περιλαμβάνουν συναισθηματικές εκφράσεις σώματος [De Gelder, 2009]. Με κίνητρο την προηγούμενη ανάλυση, σε αυτό το κεφάλαιο χτίζουμε μια μέθοδο αυτόματης αναγνώρισης του συναισθήματος σε σενάρια αλληλεπίδρασης παιδιού-ρομπότ, η οποία αξιοποιεί τη στάση του σώματος μαζί με τις εκφράσεις του προσώπου, για αυξημένη απόδοση και ευρωστία.

Οι συνεισφορές μας σε αυτό το κεφάλαιο μπορούν να συνοψιστούν ως εξής:

- Προτείνουμε μια μέθοδο που βασίζεται στα Βαθιά Νευρωνικά Δίκτυα (DNN) που συγχωνεύει τις πληροφορίες της στάσης του σώματος με τις εκφράσεις του προσώπου για την αυτόματη αναγνώριση των συναισθημάτων.
- Χρησιμοποιούμε ιεραρχικές ετικέτες (Εικόνα 1), που περιγράφουν όχι μόνο το συναισθηματικό του ατόμου στο σύνολό του, αλλά και τις ξεχωριστές εκφράσεις του σώματος και του προσώπου. Αυτοί οι σχολιασμοί μας επιτρέπουν να εκπαιδεύσουμε, είτε από κοινού είτε χωριστά, την ιεραρχική μας μέθοδο πολλαπλών ετικετών, παρέχοντάς υπολογιστικά μοντέλα για τις διαφορετικές μορφές εκφράσεων καθώς και τη συγχώνευση τους.
- Αναπτύσσουμε και αναλύουμε μια βάση δεδομένων που περιέχει παρακρινούμενες και αυθόρμητες συναισθηματικές εκφράσεις παιδιών που συμμετέχουν σε ένα παιχνίδι με ένα ρομπότ CRI και συζητάμε τις προκλήσεις της δημιουργίας ενός συστήματος αυτόματης αναγνώρισης συναισθημάτων για παιδιά. Η βάση δεδομένων περιέχει συναισθηματικές εκφράσεις τόσο στο πρόσωπο όσο και στη στάση του σώματος, επιτρέποντάς μας να παρατηρούμε και να αναγνωρίζουμε μοτίβα σωματικών συναισθηματικών εκφράσεων παιδιών σε διάφορες ηλικίες.



Σχήμα 2: Αναγνώριση συναισθημάτων με χρήση ιεραρχικών ετικετών.

Αναγνώριση συναισθημάτων ολόκληρου του σώματος με χρήση της βαθιάς μάθησης

Ένα πρόβλημα που προκύπτει όταν ασχολούμαστε με την αυτόματη αναγνώριση συναισθημάτων είναι το γεγονός ότι διαφορετικά άτομα εκφράζονται με διαφορετικούς τρόπους (σώμα, πρόσωπο, φωνή). Αυτό το γεγονός είναι επιβλαβές για αλγόριθμους εποπτευόμενης μάθησης, π.χ., σε δείγματα όπου μια επισημείωση συναισθήματος αντιστοιχεί μόνο στην έκφραση του προσώπου και όχι στο σώμα, πράγμα που σημαίνει ότι το εν λόγω άτομο προτίμησε να χρησιμοποιήσει μόνο το πρόσωπο ενώ το σώμα παρέμεινε σε ουδέτερη έκφραση. Σε τέτοια δεδομένα, ένας τρόπος για να μετριάσει αυτό το ζήτημα είναι να συμπεριληφθούν ιεραρχικές επισημειώσεις, οι οποίες πρώτα δηλώνουν τους διαφορετικούς τρόπους έκφρασης. Πιο συγκεκριμένα, υποθέτουμε ότι έχουμε την επισημείωση για ολόκληρο το σώμα y , καθώς και τις ιεραρχικές ετικέτες y^f για το πρόσωπο και y^b για το σώμα. Με βάση την προαναφερθείσα ανάλυση, το Σχήμα 2 παρουσιάζει την αρχιτεκτονική μας DNN για αυτόματη αναγνώριση συναισθημάτων χρησιμοποιώντας ιεραρχική εκπαίδευση πολλαπλών επισημειώσεων (HMT). Το δίκτυο αρχικά αποτελείται από δύο διαφορετικούς κλάδους, με έναν κλάδο να εστιάζει στις εκφράσεις του προσώπου και έναν κλάδο να εστιάζει στη στάση του σώματος. Στη συνέχεια, οι δύο κλάδοι συνδυάζονται σε μεταγενέστερο στάδιο για να σχηματίσουν τον κλάδο αναγνώρισης έκφρασης ολόκληρου του σώματος που λαμβάνει υπόψη και τις δύο πηγές πληροφοριών. Αυτός ο σχεδιασμός επιτρέπει τη ρύθμιση διαφορετικών κριτηρίων σε διαφορετικά στάδια του δικτύου με βάση τις ιεραρχικές ετικέτες, προσφέροντας αυστηρότερη επίβλεψη κατά τη διάρκεια της εκπαίδευσης. Η έξοδος του δικτύου είναι η εκτιμώμενη συναισθηματική κατάσταση του ατόμου που ανιχνεύεται στο βίντεο εισόδου.

Ο κλάδος αναγνώρισης των εκφράσεων προσώπου του δικτύου είναι υπεύθυνος για την αναγνώριση των συναισθημάτων αποκωδικοποιώντας τις εκφράσεις του προσώπου με χρήση ενός συνελκτικού νευρωνικού (Convolutional Neural Network-CNN) μοντέλου. Στον δεύτερο κλάδο, για κάθε καρέ του βίντεο εισόδου $I_i |_{i=1, \dots, N}$, εφαρμόζουμε μια μέθοδο ανίχνευσης πόζας 2D για να λάβουμε τον σκελετό $J_i \in \mathbb{R}^{K \times 2}$, όπου K είναι ο αριθμός των αρθρώσεων στον ανιχνευμένο σκελετό. Ο σκελετός στη συνέχεια εισάγεται ως διάνυσμα σε ένα (Deep Neural Network-DNN) προκειμένου να ληφθεί μια αναπαράσταση $H_i^b |_{i=1, \dots, N}$. Τέλος, εφαρμόζουμε χρονικό μέσο όρο (GTAP) σε ολόκληρη την ακολουθία εισόδου.

Οι βαθμολογίες για το συναίσθημα του σώματος s^b λαμβάνονται περνώντας την αναπαράσταση πόζας του βίντεο H^b πάνω από ένα Fully Connected (FC) επίπεδο. Το κριτήριο σε αυτόν τον κλάδο είναι ίσο με την cross entropy (Εξ. 2.1) μεταξύ των επισημειώσεων σώματος y^b και των πιθανοτήτων \tilde{s}^b , $\mathcal{L}^b(y^b, \tilde{s}^b)$.

Κλάδος αναγνώρισης έκφρασης ολόκληρου του σώματος Για να λάβουμε βαθμολογίες για την αναγνώριση των συναισθημάτων από ολόκληρο το σώμα \tilde{s}^w , ενώνουμε τα H^f και H^b και τα τροφοδοτούμε μέσω ενός άλλου FC. Στη συνέχεια χρησιμοποιούμε τις

Emotion	% using facial exp.	% using body exp.
Happiness	100%	20%
Sadness	86%	49%
Surprise	100%	43%
Fear	42%	98%
Disgust	98%	42%
Anger	85%	70%

Πίνακας 1: Ιεραρχικές ετικέτες στην BabyRobot Emotion Dataset (BRED) που απεικονίζουν τη χρήση εκφράσεων σώματος και προσώπου για κάθε συναίσθημα.

	Label	y (6 classes)		y^f (7 classes)		y^b (7 classes)	
		F1	ACC	F1	ACC	F1	ACC
SEP	Body br.	0.30 (0.29)	0.35 (0.33)	-	-	0.34 (0.48)	0.37 (0.46)
	Face br.	0.60 (0.62)	0.65 (0.65)	0.54(0.61)	0.59 (0.63)	-	-
	Sum Fusion	0.62 (0.64)	0.65 (0.66)	-	-	-	-
	Joint-1L	0.66 (0.67)	0.67 (0.67)	-	-	-	-
HMT-3a	Body br.	0.30 (0.30)	0.34 (0.33)	-	-	0.32 (0.44)	0.36 (0.44)
	Face br.	0.58 (0.61)	0.65 (0.66)	0.53 (0.59)	0.60 (0.64)	-	-
	Fusion	0.67 (0.69)	0.69 (0.70)	-	-	-	-
HMT-3b	Body br.	0.29 (0.29)	0.33 (0.32)	-	-	0.35 (0.47)	0.38(0.46)
	Face br.	0.57 (0.60)	0.64 (0.66)	0.54 (0.59)	0.60 (0.65)	-	-
	Whole body br.	0.65 (0.67)	0.68 (0.69)	-	-	-	-
HMT-4	Body br.	0.30 (0.30)	0.34 (0.32)	-	-	0.32 (0.44)	0.36(0.44)
	Face br.	0.57 (0.60)	0.64 (0.66)	0.53 (0.59)	0.59 (0.64)	-	-
	Fusion	0.70 (0.71)	0.72 (0.72)	-	-	-	-

Πίνακας 2: Αναλυτικά αποτελέσματα στη βάση δεδομένων BRED για διάφορες διαμορφώσεις του δικτύου HMT. Οι αριθμοί εκτός παρένθεσης αναφέρουν ισορροπημένες βαθμολογίες και εντός παρενθέσεων μη ισορροπημένες βαθμολογίες.

ετικέτες συναισθημάτων ολόκληρου του σώματος y για να λάβουμε την cross entropy loss ολόκληρου του σώματος μεταξύ των επισημειώσεων y ολόκληρου σώματος και τις πιθανότητες \tilde{s}^w , $\mathcal{L}^w(y, \tilde{s}^w)$. Τέλος, χρησιμοποιούμε ένα σχήμα σύμμειξης ως εξής: συνενώνουμε τις βαθμολογίες \tilde{s}^f , \tilde{s}^b και \tilde{s}^w και χρησιμοποιούμε ένα τελικό FC για να λάβουμε τις συγχωνευμένες βαθμολογίες \tilde{s}^d . Με αυτόν τον τρόπο παίρνουμε μια τελική απώλεια $\mathcal{L}^d(y, \tilde{s}^d)$ που είναι η cross entropy, μεταξύ των ετικετών ολόκληρου του σώματος y και του \tilde{s}^d .

Κατά τη διάρκεια της εκπαίδευσης, το κριτήριο εκπαίδευσης του δικτύου είναι:

$$\mathcal{L} = \mathcal{L}^f(y^f, \tilde{s}^f) + \mathcal{L}^b(y^b, \tilde{s}^b) + \mathcal{L}^w(y, \tilde{s}^w) + \mathcal{L}^d(y, \tilde{s}^d) \quad (1)$$

Η τελική πρόβλεψη του δικτύου για την ανθρώπινη δράση στο βίντεο λαμβάνεται μέσω τη βαθμολογίας σύμμειξης \tilde{s}^d .

Η βάση BabyRobot Emotion Database

Για να αξιολογήσουμε τη μέθοδό μας, συλλέξαμε μια βάση δεδομένων που περιλαμβάνει πολυτροπικές καταγραφές παιδιών που αλληλεπιδρούν με δύο διαφορετικά ρομπότ (Zeno [Robokind, 2022], Furhat [Furhat, 2022]), σε ένα εργαστηριακό περιβάλλον που έχει διακοσμηθεί για να μοιάζει με παιδικό δωμάτιο. Ο Πίνακας 1 περιέχει στατιστικά στοιχεία για τις ιεραρχικές ετικέτες της βάσης, η οποία περιλαμβάνει συνολικά 215 ακολουθίες συναισθημάτων.

Πειράματα

Ο Πίνακας 2 περιέχει αποτελέσματα πέντε διαφορετικών μεθόδων που εφαρμόστηκαν στη βάση BRED. *SEP* υποδηλώνει την ανεξάρτητη εκπαίδευση του κλάδου του σώματος και του προσώπου χρησιμοποιώντας τις αντίστοιχες ετικέτες τους. Το *Joint-1L* υποδηλώνει την εκπαίδευση του κλάδου συναισθημάτων ολόκληρου του σώματος και χρησιμοποιεί μόνο το κριτήριο \mathcal{L}^w . Το *HMT-3a* υποδηλώνει την κοινή εκπαίδευση του ιεραρχικού δικτύου εκπαίδευσης πολλαπλών επισημειώσεων, παραλείποντας τον κλάδο της αναγνώρισης συναισθημάτων ολόκληρου του σώματος, δηλαδή με τα κριτήρια \mathcal{L}^d , \mathcal{L}^f και \mathcal{L}^b . Το *HMT-3b* υποδηλώνει την κοινή εκπαίδευση με τα τρία κριτήρια: \mathcal{L}^b , \mathcal{L}^f και \mathcal{L}^w , παραλείποντας τη τελική σύμμειξη. Τέλος, το *HMT-4* υποδηλώνει την κοινή εκπαίδευση και με τα τέσσερα κριτήρια του δικτύου HMT.

Η αρχική μας παρατήρηση είναι το γεγονός ότι ο συνδυασμός της στάσης του σώματος και της έκφρασης του προσώπου οδηγεί σε σημαντική βελτίωση για όλες τις διαφορετικές μεθόδους. Δεύτερον, βλέπουμε ότι το HMT-4 επιτυγχάνει τις υψηλότερες βαθμολογίες για όλες τις μετρήσεις (0.70 F1 και 0.72 ακρίβεια), σε όλες τις μεθόδους, όσον αφορά την επισημείωση συναισθημάτων ολόκληρου του σώματος.

Στο Σχήμα 3 παρουσιάζουμε διάφορα αποτελέσματα, περιλαμβάνοντας εξίσου σωστές και λανθασμένες αναγνώσεις της μεθόδου μας.

Πολυτροπική αναγνώριση συναισθημάτων με χρήση ήχου και εκφράσεων προσώπου σε παιδιά

Στην προηγούμενη ενότητα τονίσαμε τη σημασία της γλώσσας του σώματος στην αυτόματη αναγνώριση του συναισθήματος. Σε αυτή την ενότητα, παρουσιάζουμε μια βελτιωμένη προσέγγιση, η οποία εξετάζει πολλαπλές τροπικότητες καθώς και οπτικές αναπαράστασεις.

Η αρχιτεκτονική του προτεινόμενου συστήματος αναγνώρισης συναισθημάτων φαίνεται στο Σχήμα 4. Το σύστημα αποτελείται από διαφορετικούς κλάδους, ο καθένας από τους οποίους εστιάζει σε ένα διαφορετικό κανάλι εισόδου/τροπικότητα. Ο οπτικός κλάδος βασίζεται στα Temporal Segments Network (TSN) [Wang et al., 2016]. Κατά τη διάρκεια της εκπαίδευσης, το βίντεο εισόδου χωρίζεται σε K διαφορετικά τμήματα ίσης διάρκειας M και στο επόμενο βήμα, ένα απόσπασμα μήκους $N < M$ διαδοχικών καρέ λαμβάνεται τυχαία από κάθε τμήμα, με αποτέλεσμα τη δημιουργία K αποσπασμάτων T_k . Στη συνέχεια, κάθε απόσπασμα τροφοδοτεί ένα CNN, αποδίδοντας βαθμολογίες κλάσης S_k για κάθε απόσπασμα. Στο τελευταίο βήμα, οι βαθμολογίες των διαφορετικών αποσπασμάτων συγχωνεύονται (μέσος όρος). Όπως και με τα TSN για την αναγνώριση ενεργειών, χρησιμοποιούμε και την οπτική ροή σαν είσοδο. Στον κλάδο ήχου, λαμβάνοντας υπόψη την κυματομορφή εισόδου του βίντεο, εξάγουμε πρώτα την αναπαράσταση φασματογραφήματος και, στη συνέχεια, εφαρμόζουμε ένα CNN $F_a(\mathbf{W}_a)$ σε αυτό για να εξαγάγουμε την ηχητική αναπαράσταση. Μετέπειτα, όπως και με την οπτική μέθοδο, χρησιμοποιείται ένα FC για να ληφθούν οι τελικές βαθμολογίες των συναισθημάτων. Προκειμένου να συγχωνευτούν πληροφορίες από τις οπτικές και ηχητικές λειτουργίες, εξετάζουμε δύο διαφορετικούς τύπους σύμμειξης μεταξύ RGB-ήχου, καθώς και οπτικής ροής και ήχου: σύμμειξη χαρακτηριστικών και σύμμειξη βαθμολογιών, καθώς και δύο σχήματα εκπαίδευσης: ανεξάρτητη εκπαίδευση και συνδυασμένη εκπαίδευση.



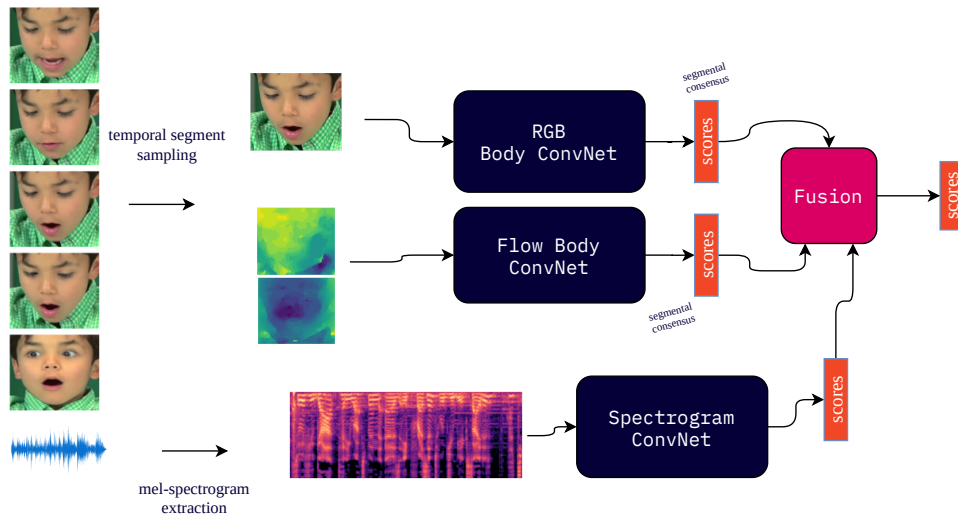
Σχήμα 3: Παραδείγματα αναγνώρισης συναισθήματος με χρήση ολόκληρου του σώματος. Οι λεζάντες στην κορυφή κάθε εικόνας υποδηλώνουν τις τελικές προβλέψεις, ενώ τα οβάλ σχήματα υποδηλώνουν προβλέψεις του κλάδου του προσώπου. Τα ορθογώνια σχήματα στο κάτω μέρος της εικόνας υποδηλώνουν προβλέψεις του κλάδου του σώματος. Τα σχήματα πράσινου χρώματος υποδηλώνουν μια σωστή πρόβλεψη, ενώ τα σχήματα κόκκινου χρώματος υποδηλώνουν μια εσφαλμένη πρόβλεψη. Εάν η τελική προβλεπόμενη επισήμειωση είναι λάθος, τότε μέσα στην παρένθεση συμπεριλαμβάνουμε τη σωστή επισήμειωση.

Πειραματικό πλαίσιο και αποτελέσματα

Το σύνολο δεδομένων EmoReact

Το σύνολο δεδομένων που χρησιμοποιούμε είναι το σύνολο δεδομένων EmoReact που περιέχει βίντεο με αντιδράσεις 63 παιδιών (32 F, 31 M, ηλικίας 4 έως 14 ετών) σε διαφορετικά θέματα και έχει συλλεχθεί από το κανάλι YouTube React. Ο αριθμός όλων των βίντεο είναι 1102. Κάθε βίντεο περιέχει ένα ή περισσότερα από τα ακόλουθα συναισθήματα: Περιέργεια, Αβεβαιότητα, Ενθουσιασμός, Ευτυχία, Έκπληξη, Αηδία, Φόβος και Απογοήτευση.

Παρουσιάζουμε τα τελικά αποτελέσματα του συστήματος αναγνώρισης συναισθημάτων στον Πίνακα 3.3, όπου έχουμε επίσης προσθέσει το αποτέλεσμα της σύμμειξης με απλό μέσο όρο των τριών διαφορετικών τρόπων εκπαίδευσης (χρησιμοποιώντας ανεξάρτητη εκπαίδευση). Στο κλάδο ήχου η αρχιτεκτονική μας επιτυγχάνει σημαντικά καλύτερη AUC ROC από το [Nojavanasghari et al., 2016], καθώς και παρόμοια αποτελέσματα με τους Nagarajan et



Σχήμα 4: Η προτεινόμενη πολυτροπική αρχιτεκτονική αναγνώρισης συναισθημάτων για την αλληλεπίδραση παιδιού-ρομπότ.

al. [Nagarajan and Oruganti, 2019]. Ωστόσο, η προσέγγισή μας είναι end-to-end και απλή στην υλοποίηση, ενώ οι Nagarajan et al. χρησιμοποίησαν ένα περίπλοκο σχήμα που περιλάμβανε πολλαπλές αρχιτεκτονικές AlexNet για την εξαγωγή των χαρακτηριστικών συνδυασμένες με ένα SVM. Στον οπτικό κλάδο, η αρχιτεκτονική μας RGB TSN βελτιώνει σημαντικά το καλύτερο προηγούμενο δημοσιευμένο αποτέλεσμα, το οποίο χρησιμοποιούσε χαρακτηριστικά που εξήχθησαν από τον αλγόριθμο OpenFace σε συνδυασμό με ένα SVM [Nojavanasghari et al., 2016].

Τέλος, το οπτικοακουστικό μας σχήμα συγχώνευσης αυξάνει περαιτέρω το ROC-AUC έως 0.754.

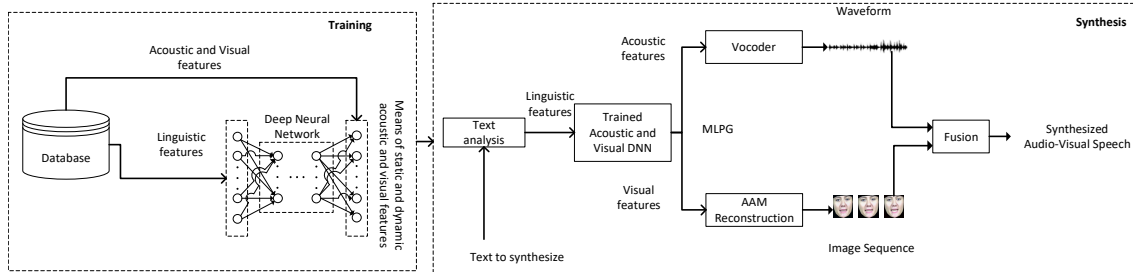
Οπτικοακουστική Σύνθεση Φωνής με Βαθεία Νευρωνικά Δίκτυα

Στο δεύτερο μέρος αυτής της διατριβής, μελετάμε τη δεύτερη πλευρά της αλληλεπίδρασης ανθρώπου-ρομπότ. Εδώ, η μηχανή έχει τον ενεργό ρόλο της ομιλίας και της μετάδοσης των συναισθημάτων, ενώ ο άνθρωπος είναι πλέον ο αποδέκτης. Με κίνητρο τις πρόσφατες εξελίξεις στη σύνθεση ομιλίας με χρήση βαθιάς μάθησης [Ling et al., 2015], προτείνουμε δύο διαφορετικές αρχιτεκτονικές βαθύων νευρωνικών δικτύων (DNN) για εκφραστική οπτικοακουστική σύνθεση ομιλίας (EAVTTS) και εξετάζουμε το επίπεδο στο οποίο αυτές οι αρχιτεκτονικές είναι σε θέση να μοντελοποιήσουν τα ειδικά χαρακτηριστικά και την πλήρη έκταση της έκφρασης της ομιλίας σε οπτικοακουστικό πλαίσιο. Αυτό επιτυγχάνεται μέσω άμεσης σύγκρισης τους με την παραδοσιακή παραμετρική προσέγγιση του EAVTTS που βασίζεται σε HMM, καθώς και με ένα σύστημα συνδυαστικής επιλογής μονάδων (Unit-Selection) EAVTTS, τόσο στον ρεαλισμό όσο και στον την εκφραστικότητα του παραγόμενου ομιλώντος προσώπου, όταν τα συστήματα εκπαιδεύονται στη βάση δεδομένων CVSP-EAV.

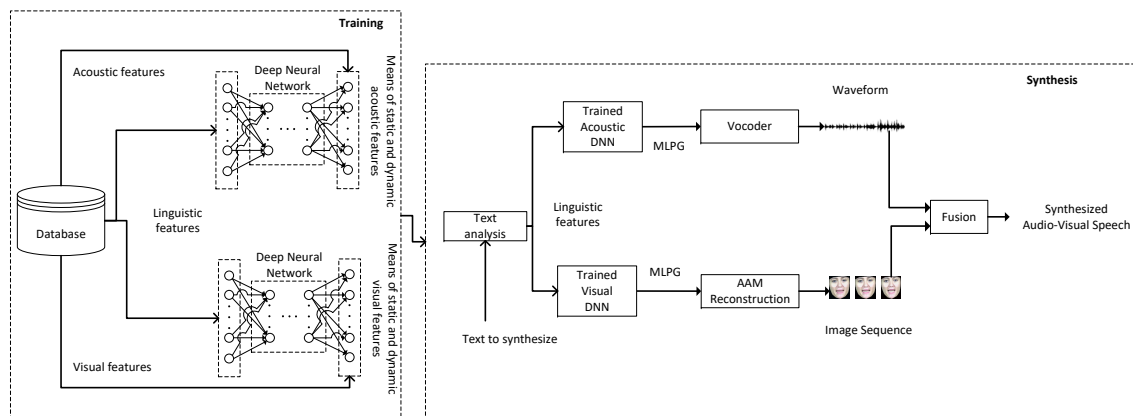
Μέθοδος

Στις δύο προτεινόμενες αρχιτεκτονικές, που βασίζονται σε DNN, για εκφραστική οπτικοακουστική σύνθεση ομιλίας κάθε συναίσθημα μοντελοποιείται ξεχωριστά από διαφορετικό μοντέλο.

Αυτά τα υποσυστήματα ακολουθούν μία από τις δύο αρχιτεκτονικές που φαίνονται στα Σχήματα 5 και 6. Τα ακουστικά, οπτικά και γλωσσικά χαρακτηριστικά εξάγονται από μία οπτικοακουστική βάση και στη συνέχεια χρησιμοποιούνται για την εκπαίδευση των υποσυστημάτων νευρωνικών δικτύων κάθε αρχιτεκτονικής. Αρχικά θα περιγράψουμε τα χαρακτηριστικά που χρησιμοποιούμε για την οπτικοακουστική μοντελοποίηση και, στη συνέχεια, θα περιγράψουμε τα δύο διαφορετικά μοντέλα DNN.



Σχήμα 5: Σύνθεση οπτικοακουστικής ομιλίας με DNN με συνδυαστική μοντελοποίηση ακουστικών και οπτικών χαρακτηριστικών.



Σχήμα 6: Σύνθεση οπτικοακουστικής ομιλίας που βασίζεται σε DNN με ξεχωριστή μοντελοποίηση ακουστικών και οπτικών χαρακτηριστικών.

Οι δύο προτεινόμενες αρχιτεκτονικές διαφέρουν στο γεγονός ότι στο Σχήμα 5, το νευρωνικό δίκτυο αντιστοιχίζει γλωσσικά χαρακτηριστικά σε ακουστικά και οπτικά χαρακτηριστικά ταυτόχρονα, ενώ στο σχήμα 6, αυτή η αντιστοίχιση γίνεται χωριστά για τα ακουστικά και οπτικά χαρακτηριστικά μέσω δύο διαφορετικών δικτύων. Τα γλωσσικά χαρακτηριστικά περιέχουν πληροφορίες σχετικές με το λεξιλογικό πλαίσιο του τρέχοντος φωνήματος και μπορεί να αποτελούνται είτε από απαντήσεις σε δυαδικές ερωτήσεις (π.χ. «το τρέχον φώνημα είναι φωνήεν») είτε από αριθμητικές τιμές (π.χ. τον αριθμό των συλλαβών σε μια λέξη). Τα χαρακτηριστικά εξόδου (ακουστικά, οπτικά ή κοινά οπτικοακουστικά) περιλαμβάνουν επίσης δυναμικά χαρακτηριστικά (πρώτη και δεύτερη παράγωγο).

Κατά τη φάση της εκπαίδευσης, γλωσσικά χαρακτηριστικά που εξάγονται από τη βάση δεδομένων, μαζί με ακουστικά και οπτικά χαρακτηριστικά, χρησιμοποιούνται για την εκπαίδευση των δικτύων. Η αντιστοίχιση από γλωσσικά χαρακτηριστικά σε ακουστικά, οπτικά ή κοινά οπτικοακουστικά χαρακτηριστικά αποτελεί πρόβλημα παλινδρόμησης και η ακόλουθη συνάρτηση μέσου τετραγώνου σφάλματος ελαχιστοποιείται από το δίκτυο

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (2)$$

όπου N είναι ο αριθμός των χαρακτηριστικών εξόδου, το \hat{y}_i είναι το i -οστό προβλεπόμενο χαρακτηριστικό και το y_i είναι το πραγματικό αποτέλεσμα.

Στο στάδιο της σύνθεσης, μετά την ανάλυση του προς σύνθεση κειμένου και την εξαγωγή της αναπαράστασης των γλωσσικών χαρακτηριστικών του, τα νευρωνικά δίκτυα προβλέπουν τα ακουστικά, οπτικά ή κοινά οπτικοακουστικά χαρακτηριστικά εξόδου. Οι έξοδοι των νευρωνικών δικτύων θεωρούνται ότι είναι το μέσο διάνυσμα των ακουστικών και οπτικών χαρακτηριστικών ενώ οι συνδιακυμάνσεις υπολογίζονται εκ των προτέρων από τα δεδομένα εκπαίδευσης. Στη συνέχεια, χρησιμοποιείται ο αλγόριθμος δημιουργίας παραμέτρων μέγιστης πιθανοφάνειας [Tokuda et al., 2000] για την παραγωγή ομαλών τροχιών ακουστικών και οπτικών χαρακτηριστικών. Αυτό το βήμα είναι επιβεβλημένο προκειμένου να μετριάσει το γεγονός ότι τα DNN δεν έχουν μνήμη ή λαμβάνουν υπόψη παρακείμενα πλαίσια κατά την εκπαίδευση [Zen, 2015].

Πειραματικά αποτελέσματα

Αξιολόγηση του ρεαλισμού και της εκφραστικότητας των μεθόδων EAVTTS

Για να αξιολογήσουμε τις μεθόδους σχεδιάσαμε και αναπτύξαμε ένα διαδικτυακό ερωτηματολόγιο που περιέχει πολλούς τύπους ερωτήσεων. Οι μέθοδοι που αξιολογούμε είναι οι ακόλουθες:

1. EAVTTS που βασίζεται σε HMM (HMM)
2. EAVTTS που βασίζεται σε DNN με κοινή μοντελοποίηση ακουστικών και οπτικών χαρακτηριστικών (DNN-J)
3. EAVTTS που βασίζεται σε DNN με ξεχωριστή μοντελοποίηση ακουστικών και οπτικών χαρακτηριστικών (DNN-S)
4. Unit-Selection EAVTTS

Στόχος μας είναι η σύγκριση των μεθόδων τόσο ως προς το ρεαλισμό όσο και την εκφραστικότητα του συνθετικού ομιλούντος προσώπου.

Για να αξιολογηθεί ο ρεαλισμός του ομιλούντος προσώπου κάθε μεθόδου, παρουσιάσαμε σε χρήστες ζεύγη βίντεο που απεικονίζουν το συνθετικό πρόσωπο να μιλάει και να να εκφέρει την ίδια πρόταση, με το ίδιο συναίσθημα και τους ζητήσαμε να επιλέξουν το πιο ρεαλιστικό βίντεο. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 3.

Από τον πίνακα μπορούμε να δούμε ότι και οι δύο αρχιτεκτονικές DNN προτιμώνται σημαντικά στο επίπεδο $p < 0.01$ έναντι των μεθόδων HMM και US, ενώ το HMM προτιμάται επίσης σημαντικά έναντι της US σε επίπεδο $p < 0,01$. Μεταξύ των δύο αρχιτεκτονικών DNN βλέπουμε ότι οι βαθμολογίες των προτιμήσεων είναι πολύ κοντά και δεν υπάρχει σημαντική διαφορά.

Εκτίμηση της εκφραστικότητας

Η εκφραστικότητα αξιολογήθηκε με τον ίδιο τρόπο όπως ο ρεαλισμός. Δεν συμπεριλήφθηκαν βίντεο με το ουδέτερο συναίσθημα. Τα αποτελέσματα φαίνονται στον Πίνακα 4. Βλέπουμε ότι η αρχιτεκτονική DNN-S προτιμάται σημαντικά από τις μεθόδους HMM και US. Η αρχιτεκτονική

DNN-S	DNN-J	HMM	US	N/P
25.0	22.22	-	-	52.78
51.11	-	15.56	-	33.33
75.56	-	-	18,89	5,55
-	43.33	22.22	-	34.44
-	72.22	-	22.78	5.0
-	-	63.89	27.78	8.33

Πίνακας 3: Αποτελέσματα (%) υποκειμενικών προτιμήσεων ανά ζεύγη στον οπτικοακουστικό ρεαλισμό ομιλίας. Η έντονη γραμματοσειρά υποδεικνύει σημαντική προτίμηση σε επίπεδο $p < 0.01$.

DNN-S	DNN-J	HMM	US	N/P
23.08	23.08	-	-	53.85
50.64		15.38	-	33.97
70.51		-	23.07	6.41
-	42,3	26,28	-	31,41
-	66.67	-	27.56	5.77
-	-	57.69	36.54	5.77

Πίνακας 4: Αποτελέσματα (%) υποκειμενικών προτιμήσεων ανά ζεύγη σχετικά με την οπτικοακουστική εκφραστικότητα του λόγου. Η έντονη γραμματοσειρά υποδεικνύει σημαντική προτίμηση σε επίπεδο $p < 0.01$.

DNN-J προτιμάται σημαντικά έναντι του US, και παρόλο που προτιμάται έναντι του HMM, δεν είναι στατιστικά σημαντική. Μεταξύ των HMM και US, προτιμάται το πρώτο, αν και το αποτέλεσμα και πάλι δεν είναι στατιστικά σημαντικό. Επίσης, είναι εμφανής η συσχέτιση μεταξύ ρεαλισμού και εκφραστικότητας, αφού παρατηρείται ότι τα αποτελέσματα ακολουθούν μια παρόμοια πορεία με την αξιολόγηση του οπτικοακουστικού ρεαλισμού.

Συμπεράσματα

Προτείναμε δύο διαφορετικές αρχιτεκτονικές για την εκφραστική οπτικοακουστική σύνθεση ομιλίας που βασίζονται σε DNN και κάναμε μια άμεση σύγκριση με εκφραστικά οπτικοακουστικά συστήματα σύνθεσης ομιλίας που βασίζονται σε HMM και συνδυαστικά συστήματα επιλογής μονάδων σχετικά με τον ρεαλισμό του παραγόμενου ομιλώντας κεφάλι, και για τη συναισθηματική δύναμη που συλλαμβάνεται από κάθε σύστημα όταν εκπαιδεύεται σε ένα συναισθηματικό σώμα.

Τα αποτελέσματά μας δείχνουν ότι και οι δύο αρχιτεκτονικές που βασίζονται σε DNN ξεπερνούν σημαντικά τις άλλες δύο μεθόδους όσον αφορά τον οπτικοακουστικό ρεαλισμό του συνθετικού ομιλώντος προσώπου. Η αρχιτεκτονική που βασίζεται σε DNN και χρησιμοποιεί ξεχωριστή μοντελοποίηση αρχιτεκτονικής ακουστικών και οπτικών χαρακτηριστικών (DNN-S) ξεπερνά επίσης σημαντικά τις μεθόδους HMM και US ως προς την εκφραστικότητα. Επιπλέον, το DNN-S πέτυχε σημαντικά καλύτερα αποτελέσματα σε σχέση με όλες τις άλλες αρχιτεκτονικές όταν εξετάστηκε μόνο η ακουστική ομιλία.

Τρισδιάστατη ανακατασκευή προσώπων από βίντεο μίας όψης, με έμφαση στη διατήρηση του ομιλούντος στόματος

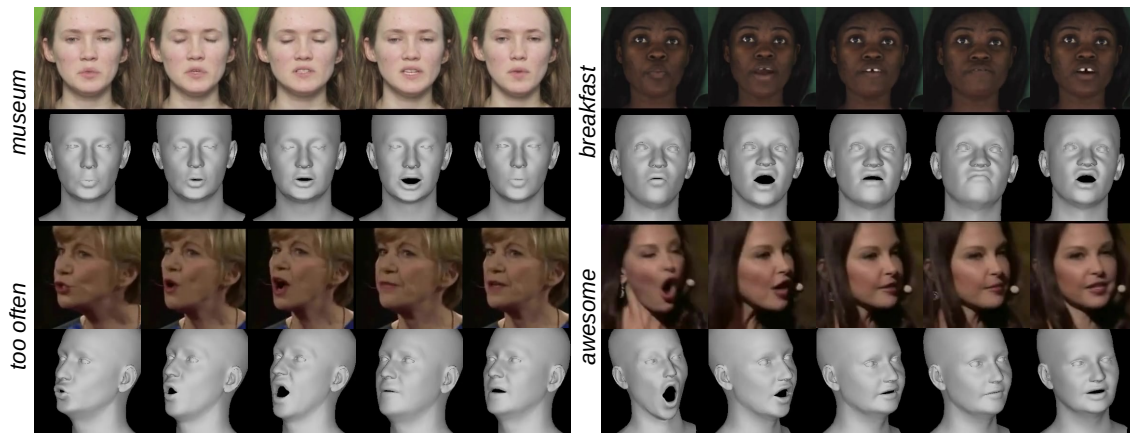
Σε αυτό το κεφάλαιο, ανακατευθύνουμε την εστίασή μας στην τρισδιάστατη ανακατασκευή προσώπων από βίντεο μίας όψης, με έμφαση στη διατήρηση του ομιλούντος στόματος. Αναμφίβολα, η πρόσφατη τεχνολογία της 3D ανακατασκευής προσώπου από δεδομένα εικόνας έχει κάνει μερικές εντυπωσιακές προόδους χάρη στην έλευση του Deep Learning και είναι σε θέση να ανακατασκευάσει λεπτές λεπτομέρειες της τρισδιάστατης γεωμετρίας του προσώπου καθώς και να αποδώσει μια αξιόπιστη εκτίμηση της ανατομίας του. Ωστόσο, έχει επικεντρωθεί ως επί το πλείστον στην είσοδο που προέρχεται από μια μεμονωμένη εικόνα RGB, παραβλέποντας τους ακόλουθους σημαντικούς παράγοντες: α) Σήμερα, η συντριπτική πλειονότητα των δεδομένων που περιλαμβάνουν πρόσωπα δεν προέρχονται από μεμονωμένες εικόνες αλλά από βίντεο, τα οποία περιέχουν πλούσιες δυναμικές πληροφορίες. β) Επιπλέον, αυτά τα βίντεο συνήθως απαντανάζουν άτομα σε κάποια μορφή λεκτικής επικοινωνίας (δημόσιες συνομιλίες, τηλεδιασκέψεις, οπτικοακουστικές αλληλεπιδράσεις ανθρώπου-υπολογιστή, συνεντεύξεις, μολόγοι/διάλογοι σε ταινίες κ.λπ.). Όταν εφαρμόζονται υπάρχουσες μέθοδοι ανακατασκευής προσώπου σε τέτοια βίντεο, τα λάθη στην αναδόμηση του σχήματος και της κίνησης της περιοχής του στόματος είναι συχνά σοβαρά, καθώς δεν ταιριάζουν καλά με τον ήχο της ομιλίας. Αυτό με τη σειρά του, εμποδίζει οποιαδήποτε χρήση της μονάδας ανακατασκευής 3D προσώπου σε άλλες εφαρμογές, όπως αυτή που είδαμε στην προηγούμενη ενότητα, ή π.χ. στη δημιουργία τρισδιάστατων avatars.

Αναμφισβήτητα, ένας κρίσιμος παράγοντας για τους περιορισμούς των υφιστάμενων μεθόδων είναι το γεγονός ότι οι περισσότερες μέθοδοι χρησιμοποιούν αδύναμη εποπτεία δισδιάστατων σημείων που βρίσκονται αυτόματα από μεθόδους ευθυγράμμισης προσώπου π.χ. [Saito et al., 2016, Thies et al., 2018, Thies et al., 2016, Jackson et al., 2017, Booth et al., 2018a, Tewari et al., 2017, Feng et al., 2021, Yang et al., 2020]. Ενώ αυτά τα σημεία μπορούν να δώσουν μια χονδρική εκτίμηση του σχήματος του προσώπου, αποτυγχάνουν στο να παρέχουν μια ακριβή αναπαράσταση των εκφραστικών λεπτομερειών μιας εξαιρετικά παραμορφώσιμης περιοχής, αυτής του στόματος.

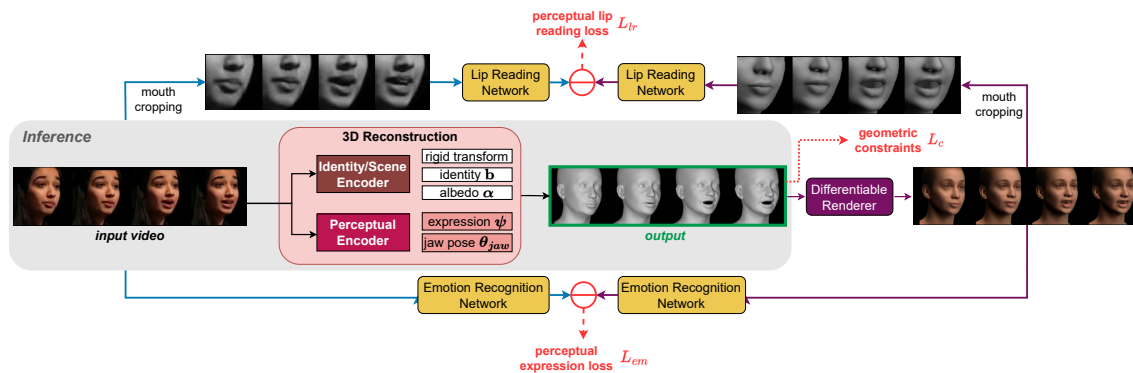
Είναι επίσης σημαντικό να σημειωθεί ότι τα σχήματα του ανθρώπινου στόματος συσχετίζονται με την ομιλία και ο ρεαλισμός ενός τρισδιάστατου ομιλούντος κεφαλιού είναι στενά συνδεδεμένος με την εκφωνηθείσα πρόταση. Ως αποτέλεσμα, ένα τρισδιάστατο μοντέλο που μιλάει χωρίς τα χείλη να κλείνουν όταν προφέρει τα διχειλικά σύμφωνα (δηλαδή, /m/, /p/, /b/), ή χωρίς στρογγυλότητα των χειλιών όταν εκφωνεί ένα στρογγυλεμένο φωνήεν (όπως το /o/ /u/) έχει κακή φυσικότητα.

Συμπεραίνουμε ότι, αν και η αντίληψη ομιλίας από ανακατασκευασμένα τρισδιάστατα πρόσωπα είναι σημαντική για διάφορες εφαρμογές (π.χ. επαυξημένη και εικονική πραγματικότητα, γαμινγ, συναισθηματικά είδωλα κ.λπ.) [Hofer et al., 2020, Marín-Morales et al., 2020, Stuart et al., 2022], είναι μια παράμετρος που συνήθως παραβλέπεται στην υπάρχουσα βιβλιογραφία. Αξίζει να σημειωθεί ότι η κύρια μέτρηση αξιολόγησης που χρησιμοποιείται από τις περισσότερες υπάρχουσες μεθόδους είναι η απόσταση των προβλεπόμενων σημείων του μοντέλου από τα πραγματικά. Ωστόσο, τα γεωμετρικά σφάλματα των εκφράσεων του προσώπου/στοματικών εκφράσεων δεν συσχετίζονται απαραίτητα με την ανθρώπινη αντίληψη [Daněček et al., 2022, Mori et al., 2012, Garrido et al., 2016a].

Για να ξεπεραστούν οι περιορισμοί της υπάρχουσας βιβλιογραφίας, το κεφάλαιο αυτό πραγματεύεται το πρόβλημα της 3D ανακατασκευής προσώπου από ένα βίντεο, με ιδιαίτερη έμφαση στην περιοχή του στόματος, τις εκφράσεις και τις κινήσεις του που είναι άρρηκτα συνδεδεμένες με την άρθρωση του λόγου. Οι κύριες συνεισφορές μας μπορούν να συνοψιστούν ως εξής:



Σχήμα 7: Η μέθοδος SPECTRE εκτελεί τρισδιάστατη ανακατασκευή με έκφραση στην οπτική ομιλία, έτσι ώστε η αντίληψη της ομιλίας από το αρχικό βίντεο να διατηρείται στην ανακατασκευή. Στα αριστερά συμπεριλαμβάνουμε τη λέξη/φράση που λέγεται για κάθε παράδειγμα.



Σχήμα 8: Επισκόπηση του SPECTRE, της αρχιτεκτονικής μας για αντιληπτική τρισδιάστατη ανακατασκευή.

- Σχεδιάζουμε και εφαρμόζουμε την πρώτη (από όσο γνωρίζουμε) μέθοδο, για αντιληπτική τρισδιάστατη ανακατασκευή ανθρώπινων προσώπων με επίκεντρο την ομιλία **χωρίς την ανάγκη επισημειωμένου κειμένου ή του αντίστοιχου ήχου**.
- Επισοούμε ένα κριτήριο που βασίζεται στο διάβασμα των χειλιών, το οποίο καθοδηγεί την εκπαίδευση έτσι ώστε το ανακατασκευασμένο πρόσωπο και ιδιαίτερα η περιοχή του στόματος να προκαλεί παρόμοια αντίληψη στον θεατή και να φαίνεται πιο ρεαλιστική όταν συνδυάζεται με τον αντίστοιχο ήχο.
- Διεξάγουμε εκτενή αντικειμενική και υποκειμενική αξιολόγηση που αποδεικνύει τη σημαντική αύξηση στην αντίληψη του ανακατασκευασμένου ομιλούντος κεφαλιού.

Μέθοδος

Προκαταρκτικά

Η εργασία μας βασίζεται στο DECA [Feng et al., 2021]. Δεδομένης μιας εικόνας εισόδου I , ένα ResNet50 CNN προβλέπει τις παραμέτρους ταυτότητας $\beta \in \mathbb{R}^{100}$, τη στάση του λαιμού και τη γνάθο $\theta \in \mathbb{R}^6$, τις παραμέτρους έκφρασης $\psi \in \mathbb{R}^{50}$, albedo $\alpha \in \mathbb{R}^{50}$, το φωτισμό

$I \in \mathbb{R}^{27}$ και την κάμερα $c \in \mathbb{R}^3$. Στη συνέχεια, αυτές οι παράμετροι χρησιμοποιούνται για τη δημιουργία του προβλεπόμενου τρισδιάστατου προσώπου.

Αρχιτεκτονική

Μια επισκόπηση της αρχιτεκτονικής φαίνεται στο Σχήμα 8. Δεδομένης μιας ακολουθίας εικόνων K RGB που ελήφθησαν από ένα βίντεο εισόδου V , η μέθοδός μας αναδομεί για κάθε καρέ I το 3D πλέγμα του προσώπου στην τοπολογία FLAME, έτσι ώστε οι κινήσεις του στόματος και οι γενικές εκφράσεις του προσώπου να διατηρούνται αντιληπτικά. Ακολουθώντας την ονοματολογία του μοντέλου προσώπου 3D FLAME, διαχωρίζουμε τις εκτιμώμενες παραμέτρους σε δύο διακριτά σύνολα:

Παράμετροι Ταυτότητας και Κάμερας: Δανειζόμαστε τον κωδικοποιητή CNN από το DECA για να προβλέψουμε ανεξάρτητα για κάθε εικόνα I στην ακολουθία εισόδου την ταυτότητα β , στάση λαιμού θ_{neck} , albedo $\alpha \in \mathbb{R}^{50}$, φωτισμό $l \in \mathbb{R}^{27}$ και κάμερα c . Όπως το EMOCA [Daněček et al., 2022], διατηρούμε αυτό το δίκτυο σταθερό κατά τη διάρκεια της εκπαίδευσης.

Παράμετροι έκφρασης: Η έκφραση ψ και η παράμετρος σαγονιού θ_{jaw} υπολογίζονται από ένα ξεχωριστό CNN. Αυτές οι παράμετροι ελέγχουν ρητά τις εκφράσεις και τις κινήσεις του στόματος. Χρησιμοποιούμε μια αρχιτεκτονική MobileNet v2, αλλά εισάγουμε επίσης έναν πυρήνα συνέλιξης στην έξοδο του, προκειμένου να μοντελοποιήσουμε το χρονική δυναμική των κινήσεων του στόματος και των εκφράσεων του προσώπου στην ακολουθία εισόδου.

Απώλειες εκπαίδευσης

Για να εκπαιδεύσουμε τον CNN έκφρασης, χρησιμοποιούμε δύο κριτήρια για την καθοδήγηση της ανακατασκευής, μαζί με γεωμετρικούς περιορισμούς. Η έξοδος του κωδικοποιητή χρησιμοποιείται μαζί με τις προβλέψεις ταυτότητας, albedo, κάμερας και φωτισμού προκειμένου να γίνει ανακατασκευή του προσώπου. Στη συνέχεια, το βίντεο εισόδου και το ανακατασκευασμένο τρισδιάστατο πρόσωπο τροφοδοτούνται σε ένα δίκτυο αναγνώρισης συναισθημάτων (δανεισμένο από το EMOCA [Daněček et al., 2022]) και λαμβάνονται δύο ακολουθίες διανυσμάτων χαρακτηριστικών. Ακολουθώντας, εφαρμόζουμε μια απώλεια αντιληπτικής έκφρασης L_{em} , προσπαθώντας να ελαχιστοποιήσουμε την απόσταση μεταξύ των δύο ακολουθιών διανυσμάτων χαρακτηριστικών. Αντί να βασιζόμαστε μόνο σε μια γεωμετρική απώλεια με αδύναμη επίβλεψη χρησιμοποιώντας 2D σημεία, χρησιμοποιούμε μια πρόσθετη αντιληπτική απώλεια, που καθοδηγεί τους συντελεστές έκφρασης για να διατηρήσουν τις περιπλοκές των κινήσεων του στόματος. Για το σκοπό αυτό χρησιμοποιούμε ένα δίκτυο που έχει εκπαιδευτεί στο σύνολο δεδομένων LRS3 (Lip Reading in the Wild 3) [Ma et al., 2022]. Το δίκτυο ανάγνωσης χειλιών είναι το προεκπαιδευμένο μοντέλο που παρέχεται από τους Ma et al. [Ma et al., 2022] και στόχος μας εδώ είναι να ελαχιστοποιήσουμε την αντιληπτική απόσταση των κινήσεων της ομιλίας μεταξύ του αρχικού και τελικού βίντεο.

Πειράματα

Αξιολογούμε τη μέθοδό μας τόσο ποιοτικά όσο και ποσοτικά, ακολουθώντας παρόμοια διαδικασία αξιολόγησης με το [Daněček et al., 2022]. Για αξιολόγηση χρησιμοποιούμε τα ακόλουθα σύνολα δεδομένων:

LRS3 [Afouras et al., 2018]: Το δοκιμαστικό σύνολο του LRS3.

	LRS3		TCD-TIMIT		MEAD	
	CER ↓	VER ↓	CER ↓	VER ↓	CER ↓	VER ↓
Original vid	24.9	22.0	35.7	29.6	49.7	42.8
DECA	77.5	70.8	84.2	75.8	84.8	77.8
EMOCA	83.3	76.3	86.4	79.2	85.1	77.9
3DDFAv2	97.5	95.3	101.8	98	94.5	90.2
DAD	84.1	78.2	87.3	81	86.0	79.9
SPECTRE	67.5	60.9	78.1	69.6	78.5	71.1

Table 5: Αποτελέσματα διαβάσματος χειλιών. Για όλες τις μετρήσεις, το χαμηλότερο ποσοστό σφάλματος είναι το πιο επιθυμητό. Η μέθοδός μας ξεπερνά σημαντικά όλες τις άλλες μεθόδους τρισδιάστατης ανακατασκευής.

	DECA	EMOCA	3DDFAv2	DAD
SPECTRE	201/37	185/53	218/20	150/88

Table 6: Αποτελέσματα προτίμησης της πρώτης υποκειμενικής μελέτης. Η μέθοδός μας είναι **σημαντικά** ($p < 0.01$) πιο ρεαλιστική όσον αφορά τις κινήσεις του στόματος και την άρθρωση.

MEAD: Αυτό είναι ένα πρόσφατο σύνολο δεδομένων [Wang et al., 2020a] που περιέχει 48 ηθοποιούς (28M, 20F) που προφέρουν προτάσεις σε 7 βασικά συναισθήματα συν μια ουδέτερη έκφραση και σε τρία διαφορετικά επίπεδα έντασης. Το σύνολο των δεδομένων περιλαμβάνει 31.059 προτάσεις. Πραγματοποιήσαμε τυχαία δειγματοληψία 2.000 προτάσεων για να δημιουργήσουμε ένα σύνολο δοκιμών.

TCD-TIMIT [Harte and Gillen, 2015]: Αυτό το σώμα περιλαμβάνει 62 Άγγλους ηθοποιούς που διαβάζουν 6913 προτάσεις από την TIMIT [Garofolo et al., 1993].

Συγκρίνουμε τη μέθοδό μας με τις ακόλουθες πρόσφατες μεθόδους τελευταίας τεχνολογίας για την ανακατασκευή 3D προσώπου: **DECA** [Feng et al., 2021], **EMOCA** [Daněček et al., 2022], **3DDFAv2** [Guo et al., 2020] και **DAD-3DHeads**.

Ποσοτικά Αποτελέσματα

Αξιολογούμε τις μεθόδους αντικειμενικά όσον αφορά τις μετρήσεις ανάγνωσης χειλιών εφαρμόζοντας ένα προεξπαιδευμένο δίκτυο διαβάσματος χειλιών. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 5. Η μέθοδός μας επιτυγχάνει πολύ χαμηλότερες βαθμολογίες CER, VER σε σύγκριση με τις άλλες μεθόδους.

Υποκειμενικά αποτελέσματα

Για να αξιολογήσουμε το ρεαλισμό και την αντίληψη του τρισδιάστατου ανακατασκευασμένου προσώπου σε ανθρώπους σχεδιάσαμε και πραγματοποιήσαμε δύο μελέτες. Στη μία μελέτη που περιγράφουμε εδώ, δείξαμε στους χρήστες ζεύγη τρισδιάστατων ανακατασκευασμένων προσώπων, μαζί με το αρχικό βίντεο, και ζητώντας τους να επιλέξουν το πιο ρεαλιστικό όσον αφορά τις κινήσεις του στόματος και την άρθρωση. Τα αποτελέσματα αυτής της μελέτης εμφανίζονται στον Πίνακα 6. Μπορούμε να δούμε ότι η μέθοδός μας προτιμάται σημαντικά από όλες τις άλλες μεθόδους ($p < 0.01$).

Συνεισφορά

Σε αυτή τη διατριβή μελετήσαμε πτυχές του affective computing υπό από την ομπρέλα της αλληλεπίδρασης ανθρώπου-μηχανής. Στο πρώτο μέρος, οι συνεισφορές μας μπορούν να συνοψιστούν ως εξής:

1. Εισαγάγαμε και μελετήσαμε πολλαπλές ροές πληροφοριών που μπορούν να χρησιμοποιηθούν προκειμένου να βελτιωθούν τα συστήματα αναγνώρισης συναισθημάτων.
2. Σχεδιάσαμε και αναπτύξαμε αρχιτεκτονικές βασισμένες σε βαθιά νευρωνικά δίκτυα και μελετήσαμε τη βέλτιστη σύμμειξη των διαφορετικών ροών, προκειμένου να επιτύχουμε πιο ισχυρά αποτελέσματα και να αυξήσουμε την απόδοση.
3. Συλλέξαμε και διαθέσαμε το BabyRobot Emotion Dataset, ένα μεσαίου μεγέθους σύνολο δεδομένων παιδιών που εκτελούν συναισθήματα ενώ αλληλεπιδρούν με δύο διαφορετικά ρομπότ.

και στο δεύτερο μέρος:

1. Μελετήσαμε την οπτικοακουστική σύνθεση του λόγου και προτείνουμε μεθόδους που προσδίδουν σε τέτοια συστήματα δυνατότητες συναισθηματικής έκφρασης
2. Προτείνουμε δύο συστήματα οπτικοακουστικής σύνθεσης εκφραστικής ομιλίας βασισμένα σε βαθιά νευρωνικά δίκτυα, τα οποία ξεπέρασαν σημαντικά τις παραδοσιακές εναλλακτικές λύσεις.
3. Καταγράψαμε και αναπτύξαμε το σύνολο δεδομένων CVSP-EAV, το οποίο περιλαμβάνει συνολικά 3.600 συναισθηματικές προτάσεις σε τέσσερα συναισθήματα στα ελληνικά.
4. Σχεδιάσαμε και εφαρμόσαμε την πρώτη μέθοδο τρισδιάστατης ανακατασκευής της έκφρασης του προσώπου με επίκεντρο τη διατήρηση της οπτικής ομιλίας.

Contents

Contents	34
List of Figures	35
List of Tables	43
1 Introduction and Background	45
1.1 Goals and Contributions	46
1.2 Theoretical and Computational Models of Emotion	47
1.2.1 Theories of Emotion	47
1.2.2 Emotion Modeling in Affective Computing	48
1.3 MultiModal Emotion Recognition	49
1.3.1 Facial Expressions	49
1.3.2 Body Expressions	50
1.3.3 Context	50
1.3.4 Speech	51
1.3.5 Emotional Embeddings	51
1.4 Synthesis of Facial Expressions	51
1.4.1 AudioVisual Speech Synthesis and 3DMMs	52
1.4.2 Adding Expressions to AudioVisual Speech Synthesis	53
1.4.3 3D Model Fitting and Data Availability	53
1.4.4 Model-Free generation	54
1.5 Enabling Technologies	54
1.5.1 Body Pose Estimation	54
1.5.2 Face Alignment	55
1.6 Thesis Outline	56
I Emotion Recognition	59
2 Fusing Body Posture with Facial Expressions for Emotion Recognition in Children and Adults	63
2.1 Introduction	63
2.2 Related Work	65
2.3 Whole Body Emotion Recognition using Deep Learning	67
2.3.1 Hierarchical Emotional Labels	67
2.3.2 Method	67
2.4 The BabyRobot Emotion Database	69
2.5 Experiments	71
2.5.1 Network Setup and Initialization	72

2.5.2	Exploratory Results on the GEMEP Database	73
2.5.3	Results on the BabyRobot Emotion Database	74
2.6	Chapter Conclusions	76
3	Multi-Modal Emotion Recognition using Audio and Facial Expressions in Children	79
3.1	Introduction	79
3.2	Related Work	80
3.3	Method	81
3.3.1	Visual Branch	81
3.3.2	Audio Branch	82
3.3.3	Training and Audiovisual Fusion	82
3.4	Experimental Framework and Results	82
3.4.1	The EmoReact Dataset	82
3.4.2	Implementation Details	82
3.4.3	Results	84
3.5	Chapter Conclusions	86
4	Using Body, Context, and Emotional Label Embeddings for Emotion Recognition in the Wild	89
4.1	Introduction	89
4.2	Related Work	89
4.3	Model Architecture	90
4.4	Dataset	92
4.5	Experimental Results	93
4.6	Chapter Conclusions	94
II	Expression Synthesis	97
5	AudioVisual Speech Synthesis with Compound Emotions	101
5.1	Introduction	101
5.1.1	Desired capabilities of expressive agents	102
5.2	Related Work	103
5.3	Features for Parametric AudioVisual Speech Synthesis	104
5.3.1	Acoustic Features	104
5.3.2	Visual Features	104
5.4	HMM-based expressive audio-visual speech synthesis	105
5.4.1	Overview of HMM audio-visual speech synthesis	106
5.5	Adaptation for HMM-based EAVTTS	107
5.6	Interpolation for HMM-based EAVTTS	107
5.7	The CVSP-Expressive Audio-Visual Speech Corpus	108
5.7.1	Recording of the Corpus	108
5.7.2	Processing of the Corpus	109
5.7.3	Feature Extraction	110
5.8	Experimental Results	113
5.8.1	Evaluation Procedure	113
5.8.2	Evaluation of HMM Adaptation	113
5.8.3	Evaluation of HMM Interpolation	115
5.9	Chapter Conclusions	118

6	DNN-Based Expressive Speech Synthesis	121
6.1	Introduction	121
6.2	Method	121
6.3	Unit Selection audio-visual speech synthesis	123
6.4	Experimental Results	124
6.4.1	Evaluation Procedure	124
6.4.2	Evaluation of realism and expressiveness of the EAVTTS methods	125
6.5	Chapter Conclusions	129
7	Visual Speech-Informed Perceptual 3D Reconstruction and Manipulation of Facial Expressions	133
7.1	Introduction	133
7.2	Related Work	134
7.2.1	Method	136
7.2.2	Two-stage Training	140
7.3	Experiments	140
7.3.1	Quantitative Evaluation	141
7.3.2	User Studies	143
7.3.3	Visual Comparisons and Ablations	144
7.3.4	Failure Cases	147
7.4	Discussion	148
7.5	Visual Speech-Informed Semantic Control of Facial Expressions	149
7.5.1	Method	150
7.5.2	Experimental Results	153
7.6	Chapter Conclusions	155
8	Contributions & Future Directions	159
8.1	Contributions	159
8.1.1	Emotion Recognition	159
8.1.2	Expression Synthesis	159
8.2	Future Directions	160
8.2.1	Emotion Recognition	160
8.2.2	Expression Synthesis	160
A	Glossary	187
B	Wearable-based behavior analysis of patients with psychotic disorders	191
B.1	Introduction	191
B.2	Experimental Protocol and Data Collection	193
B.2.1	Experimental Protocol	193
B.2.2	Method & Data Collection	193
B.3	Feature Extraction and Data PreProcessing	195
B.4	Experimental Results	198
B.4.1	Wakefulness comparison	199
B.4.2	Sleep comparison	199
B.4.3	Sleep-wake ratio and total steps	200
B.5	Discussion	200
B.5.1	Conclusion	201
C	List of Publications	205

D List of this Thesis' Open Source Codes	211
E List of this Thesis' Released Datasets	213
F List of Awards and Grants	215

List of Figures

1.1	The human-robot communication loop. The robotic agent should be able to perceive human feelings and expressions, while at the same time have the ability to act expressively.	46
1.3	A data driven set of emotions created through clustering of verbal affective states [Kosti et al., 2017a].	49
1.4	All sources of information that can be used to infer the emotion of Jack Sparrow: body posture, facial expressions, scene context, speech, and emotional embeddings.	51
1.5	The FLAME [Li et al., 2017] 3D morphable model.	54
1.6	The OpenPose and MediaPipe BlazePose methods for body pose estimation along with an example skeletal topology.	55
1.7	Face landmarks detected by FAN [Bulat and Tzimiropoulos, 2017].	56
2.1	The importance of body expressions in affect recognition.	64
2.2	Hierarchical multi-labels for affect recognition via body and face cues, where y denotes the whole body emotion label, y^f the facial expression label, and y^b the body expression one.	67
2.3	Hierarchical multi-label training for recognition of affect from multiple visual cues in CRI.	68
2.4	The experimental setup of the BRED Database and snapshots showing children playing the “Express the feeling” game.	70
2.5	Confusion matrices for the face, body and whole body (feature fusion) branches of HMT in the GEMEP corpus.	74
2.6	Example results of whole body affect recognition. Captions on top of each image denote the final predictions while oval shapes denote predictions of the face branch and rectangle shapes denote predictions of the body branch. Green color shapes denote a correct prediction, whereas red color shapes denote an incorrect prediction. If the final predicted label is wrong then inside parenthesis we include the correct label.	75
2.7	Confusion matrices of the body, face, and score fusion branches of HMT-4, against whole body labels y on the BRED database.	76
3.1	The proposed multimodal emotion recognition architecture for child-robot interaction.	80
3.2	Example images from the EmoReact dataset.	83
3.3	ROC AUC per emotion, for each different modality and their average score fusion.	85
4.1	TSN with two RGB spatial streams (body and context) and one optical flow stream. The final results are obtained using average score fusion.	90

4.2	RGB and Flow body and context	91
4.3	PCA projection of the categorical emotions GloVe word embeddings.	92
5.1	Various examples of video-realistic talking heads.	104
5.2	HMM-based audio-visual speech synthesis system architecture	106
5.3	Interpolation for HMM-based EAVTTS.	108
5.4	Sample images for each of the different emotions present in the CVSP-EAV Corpus	109
5.5	Example of the first eigenshape and the variations it causes to the mean shape between values $[-3\sqrt{\lambda_1}, +3\sqrt{\lambda_1}]$ where λ_1 is the corresponding weight, for an AAM trained on an expressive corpus.	111
5.6	Example of the first eigentexture and the variations it causes to the mean texture between values $[-3\sqrt{\lambda_1}, +3\sqrt{\lambda_1}]$ where λ_1 is the corresponding weight, for an AAM trained on an expressive corpus.	111
5.7	Facial landmarks used for building the Active Appearance Model.	112
5.8	Subjective evaluation of the level of expressiveness captured by an adapted HMM audio-visual speech synthesis system for each different emotion (and total), and for a variable number of sentences. Bold line represents the median, x represents the mean, the boxes extend between the 1st and 3rd quantile, whiskers extend to the lowest and highest datum within 1.5 times the inter-quantile range of the 1st and 3rd quartile respectively, and outliers are represented with circles.	113
5.9	Results of audio-visual synthesis (consecutive frames from the same sentence) from a neutral HMM set (a) and its adaptation to the three emotions of (b) anger, (c) happiness, and (d) sadness, using 50 adaptation sentences.	114
5.10	Results of audio-visual synthesis (consecutive frames from the same sentence) from interpolating HMM sets trained on anger and happiness (w_a : anger weight, w_h : happiness weight).	117
6.1	DNN-based audio-visual speech synthesis with joint modeling of acoustic and visual features.	121
6.2	DNN-based audio-visual speech synthesis with separate modeling of acoustic and visual features.	122
6.3	Unit selection based audio-visual speech synthesis.	124
6.4	Boxplot of the MOS test results on the audio-visual realism of the different EAVTTS methods. Bold line represents the median, x represents the mean, the boxes extend between the 1st and 3rd quantile, whiskers extend to the lowest and highest datum within 1.5 times the inter-quantile range of the 1st and 3rd quartile respectively, and outliers are represented with circles.	126
6.5	Results of the MOS test broken down for each different emotion. Bold line represents the median, x represents the mean, the boxes extend between the 1st and 3rd quantile, whiskers extend to the lowest and highest datum within 1.5 times the inter-quantile range of the 1st and 3rd quartile respectively, and outliers are represented with circles.	128
7.1	Examples of inaccuracies in 2D landmark detection in current state-of-the-art methods [Bulat and Tzimiropoulos, 2017]. Notice how especially on the right column the face alignment has not accurately predict mouth closure which is of vital important for realistic perception of bilabial consonants (/p/, /m/, /b/).	135

7.2	Our method SPECTRE performs visual-speech aware 3D reconstruction so that speech perception from the original footage is preserved in the reconstructed talking head. On the left we include the word/phrase being said for each example.	137
7.3	Overview of our architecture for perceptual 3D reconstruction. The input video is first fed into the 3D reconstruction component, where a fixed encoder detects the scene parameters (camera, lighting), identity parameters (albedo/identity) and an initial estimate of the jaw and expression parameters. Then, a Mouth/Expression encoder predicts the refined facial expression parameters and jaw pose, and a differentiable renderer renders the predicted 3D shape. Finally, the mouth area is differentially cropped in both the input and rendered image sequences and a lip reader is applied on both in order to estimate the perceptual lip reading loss between them. Similarly, a perceptual expression loss is used on the full face based on an emotion recognition network. Inference requires only the 3D reconstruction component.	138
7.4	Comparison of 3D reconstructions of the mouth area for 2 frames from an example VOCASET clip. The MAE, CER and VER errors over the clip are also reported (best result w.r.t. each metric is in bold). MAE values are scaled by $\times 10^3$. Notice the discrepancy in the ranking of the different methods between MAE and CER/VER metrics. We observe that the perceived quality of mouth reconstruction seems to have a much better correlation with CER and VER metrics, rather than MAE.	142
7.5	Three example words from the second user study (lip reading). We show our method against DAD and EMOCA. In PERFUME and NARROW, our method accurately predicts the rounded mouth formations. In the third case of PEOPLE, we see a failure case, where the bilabial consonant /p/ which corresponds to closed mouth was not predicted accurately. Note how also in PERFUME, our method accurately depicts /f/ in the third frame.	145
7.6	Visual comparison with other methods on the MEAD, TCDTIMIT, and LRS3 datasets. Note that our method is only trained on the LRS3 train test. From left to right: original footage, 3DDFAv2 [Guo et al., 2020], DAD [Martyniuk et al., 2022], DECA [Feng et al., 2021], EMOCA [Daněček et al., 2022], ours. We also highlight with red boxes some erroneous results, and with green boxes some examples of retaining the original mouth formation.	146
7.7	More visual results and comparisons with other methods on the LRS3, MEAD, and TCDTIMIT datasets. From left to right: original footage, 3DDFAv2 [Guo et al., 2020], DAD [Martyniuk et al., 2022], DECA [Feng et al., 2021], EMOCA [Daněček et al., 2022], SPECTRE.	147
7.8	Training of the perceptual encoder without (a) and with (b) geometric constraints based on 2D landmarks. Omitting geometric constraints from the rest of the face leads to the emergence of artifacts in the eyes and nose in some cases, while completely omitting 2D information from mouth landmarks can lead to failure cases in the mouth area. Please zoom in for details.	148

7.9	Two examples of adversarial attacks using the CTC loss. Middle row shows sampled frames from the original predicted sequence by DECA [Feng et al., 2021] of two sentences with starting CER (character error rate) around 0.90, while the third row shows completely distorted examples which however achieve near-perfect CER.	149
7.10	Ablation between using absolute position of mouth landmarks or relative intra-mouth distances. The first column is the initial estimate of DECA, the second column the predicted reconstruction of a model trained with an L_1 loss imposed on the mouth landmarks as well, and the third column a model trained with a more relaxed loss on the mouth using the intra-mouth distances. Strict mouth landmark losses erroneously guide the output to resemble DECA, while the relaxed constraints leave enough freedom to the perceptual loss to accurately capture the formation of lips.	150
7.11	Examples from failure cases of our model. The domain gap problem can still cause some mouth artifacts, even when guided by our geometric constraints. Note also that any failed results of the lipread network propagate to our 3D reconstruction method as well.	150
7.12	<i>Neural Emotion Director (NED)</i> can manipulate facial expressions in input videos while preserving speech, conditioned on either the semantic emotional label (top part of figure), or on an external reference style as extracted from a reference video (bottom part).	151
7.13	Overview of <i>Neural Emotion Director (NED)</i> at inference time. NED consists of three modules: first 3D Face Analysis is performed on the input video. Then, the extracted 3D parameters are translated to a target emotional domain using the <i>3D-based Emotion Manipulator</i> . Finally, a neural renderer is used in order to render the manipulated photo-realistic frames.	151
7.14	Effect of the speech-preserving loss. Without this loss (middle row), the result does not preserve the mouth movement from the input video. In contrast, enforcing this loss (bottom row) successfully translates the expression of the actor to happy without altering his mouth movements and speech	153
7.15	Visual comparison with state-of-the-art methods in the emotional “self-translation” experiment on the MEAD actors. Note that ICface [Tripathy et al., 2020] requires a tighter face cropping and padding with the background color has been used for visualization.	154
B.1	Examples of features considered in this work during one subject’s full day.	194
B.2	Steps per day and hours spent sleeping and awake during one month of a subject’s recordings.	195
B.3	Boxplots for <i>accelerometer</i> and <i>gyroscope</i> features of controls (in blue) and patients (in light brown) while (a) awake and (b) asleep. The bold line represents the median, the boxes extend between the 1st and 3rd quartile, whiskers extend to the lowest and highest datum within 1.5 times the interquartile range (IQR) of the 1st and 3rd quartile respectively, and outliers are shown as diamonds.	195

B.4	Boxplots for <i>heart rate variability</i> features of controls and patients while awake (top rows) and asleep (bottom rows). The bold line represents the median, the boxes extend between the 1st and 3rd quartile, whiskers extend to the lowest and highest datum within 1.5 times the inter-quantile range (IQR) of the 1st and 3rd quartile respectively, and outliers are shown as diamonds.	196
B.5	Boxplots of sleep/wake ratio and steps per day (mean-std).	200

List of Tables

2.1	Hierarchical multi-label annotations of the BabyRobot Emotion Dataset (BRED) depicting usage of body and facial expressions for each emotion.	70
2.2	Patterns of bodily expression of emotion in the BRED corpus and example images.	71
2.3	Accuracy results for the body, face, and whole body branch on the GEMEP database (12 classes).	72
2.4	Detailed results on the BRED database for various configurations of the HMT network. Numbers outside parentheses report balanced scores and inside parentheses unbalanced scores. The highest achieved scores when evaluating against whole body labels are shown in bold.	73
3.1	ROC AUC and average time elapsed per epoch with varying number of sampled snippets.	83
3.2	Results on the EmoReact dataset for different fusion and training schemes between the RGB-audio and Flow-audio modalities.	84
3.3	Final ROC AUC results on the EmoReact dataset.	85
4.1	Ablation experiment by training with and without \mathcal{L}_{emb}	93
4.2	Results on the validation and test set of BoLD including the RGB context stream and \mathcal{L}_{emb}	93
5.1	Statistics of the post-processed CVSP-EAV corpus.	110
5.2	Fitting results in terms of mean reconstruction error for each of the emotions in the CVSP-EAV Corpus	112
5.3	Classification of emotions in the emotion individual HMM systems (% scores).	114
5.4	Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting the neutral emotion, and the second one trained on an emotional training set depicting happiness (% scores, w_n : Neutral Weight, w_h : Happiness Weight).	115
5.5	Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting the neutral emotion, and the second one trained on an emotional training set depicting anger (% scores, w_n : Neutral Weight, w_a : Anger Weight).	115
5.6	Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting the neutral emotion, and the second one trained on an emotional training set depicting sadness (% scores, w_n : Neutral Weight, w_s : Sadness Weight).	115

5.7	Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting anger, and the second one trained on an emotional training set depicting happiness (% scores, w_a : Anger Weight, w_h : Happiness Weight).	116
5.8	Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting anger, and the second one trained on an emotional training set depicting sadness (% scores, w_a : Anger Weight, w_s : Sadness Weight).	116
5.9	Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting sadness, and the second one trained on an emotional training set depicting happiness (% scores, w_s : Sadness Weight, w_h : Happiness Weight).	116
6.1	Results (%) of subjective pairwise preference tests on audio-visual speech realism. Bold font indicates significant preference at $p < 0.01$ level.	126
6.2	Significant differences between systems, on the audio-visual realism of the generated talking head, at levels $p < 0.05$ and $p < 0.01$. Blank cell denotes no significant different.	127
6.3	Results (%) of subjective pairwise preference tests on visual speech realism. Bold font indicates significant preference at $p < 0.01$ level.	127
6.4	Results (%) of subjective pairwise preference tests on acoustic speech realism. Bold font indicates significant preference at $p < 0.01$ level.	128
6.5	Results (%) of subjective pairwise preference tests on audio-visual speech expressiveness. Bold font indicates significant preference at $p < 0.01$ level.	129
7.1	Lipreading (CER, VER) and geometric-based metrics (IMAE, vMAE, R^2 score) are reported on the VOCASET test set. Mean absolute error is calculated both on the mouth landmarks (<i>IMAE</i>) and their temporal velocity (<i>vMAE</i>). While SPECTRE achieves significantly better lipreading metrics, this result is not reflected on traditional geometric errors (MAE scaled by $\times 10^3$).	141
7.2	Lipreading results on the LRS3-test, TCD-TIMIT and MEAD datasets (network trained on LRS3-train set). For all metrics, lower is better (error rates). Our method significantly outperforms all other 3D reconstruction methods. The 1st row corresponds to results on the original videos, reported as reference.	143
7.3	Preference results of the first subjective study. The “a/b” depiction of the result means that SPECTRE (on the left) was preferred <i>a</i> times while the competing method (given as column header) was chosen <i>b</i> times out of the 238 pairs the subjects assessed. Our method is significantly more realistic ($p < 0.01$ with binomial test after adjusting for multiple comparisons) in terms of mouth movements and articulation.	144
7.4	Classification accuracy in the second user study (word-level lipreading).	144
7.5	Per-word recognition results for the second user study, including all considered SoTA methods. We report indicative cases of failure (first five columns) and success (last five columns) of our approach compared to the other methods.	145

7.6	Quantitative comparisons on MEAD in the emotional “self-translation” experiment. Bold values denote the best value for each metric (lower is better). Averaging is done over both the full set of 7 emotion labels and the set of 5 labels supported by DSM [Solanki and Roussos, 2021], for the sake of fair comparison.	155
7.7	Realism ratings (percentage of users that rated the videos with 4 or 5) and classification accuracy of the user study on MEAD.	156
7.8	Realism ratings of the user study on 6 YouTube actors. Columns 1-5 show the number of times that users gave this rating. The column “real” shows the percentage of users that rated the videos with 4 or 5. Bold values denote the most frequent user rating for each method and actor.	156
B.1	Demographics information of controls and patients at the time of recruitment, illness information, and amount of recorded data for each group during wakefulness and sleep. There were no significant differences for the recorded data (tested with Student’s t-test and Shapiro-Wilk for normality).	194
B.2	Statistical difference analysis using Mann-Whitney U-tests with BH correction in each state (wakefulness, sleeping). Bold values denote significance at the 95% confidence levels. For each group the median and the IQR (in parenthesis) is shown for each feature.	199

1

Introduction and Background

Emotions and the ability to express oneself is a fundamental aspect of human behavior. Across thousands of years, emotions have been evolved along with humanity and have both intra- and inter-personal functions. They help with rapid decision-making, influence thoughts, and are interwoven with the cognitive process. Inter-personally, emotions help us understand others better, building stronger social bonds through empathy and emotional expressiveness.

“Affective computing”, coined by Picard in her seminal work [Picard, 1995] concerns the ability of computers and robotic agents to express, recognize, model, or even “feel” emotions. Although the latter (“feeling” - emotional awareness) is probably the most important aspect and a core component of the human nature, it currently constitutes a philosophical problem. On the other hand, the first three have a direct more practical value: they enable more natural human-computer interaction. A direct application of affective computing is social robotics: a fairly new area in robotics that has been enjoying a swift rise in its applications, some of which include robot-assisted therapy in adults and children [Belpaeme et al., 2013], activities of daily living [Broadbent et al., 2009], and education [Belpaeme et al., 2018]. In applications where humans directly communicate and interact with machines and robots, recognizing and expressing emotions can elicit trust in the human party and make the interaction more life-like.

The communication loop between a robot and a human has two interwoven aspects (see Fig.1.1). In the first aspect, the robot needs to be able to model and perceive human emotions. Indeed, empathy, i.e. the capacity to correctly interpret the social cues of humans that are manifestations of their affective state, has been proven to be a critical capability of social robotics. Empathic agents are able to change their behavior and actions according to the perceived affective states and as a result establish rapport, trust, and healthy long-term interactions [Bickmore and Picard, 2005]. Especially in the field of education, empathic robot behaviors that are congruent with the child’s feelings increase trust and have a positive impact on the child-robot relationship, whereas incongruent behavior has a significantly negative effect [Leite et al., 2014].

The counterpart (and second aspect of the loop) of perceiving human emotions and expressions is the ability of robotic agents to synthesize emotions and express themselves as well. Expressive agents are much more appealing and offer “the illusion of life” [Bates, 1994], greatly enhancing the resulting social interaction [Mavridis, 2015]. In addition, expressive agents affect the emotional state of the other party [Hatfield et al., 1993, Keltner and Haidt, 1999], and motivate it to express itself as well.

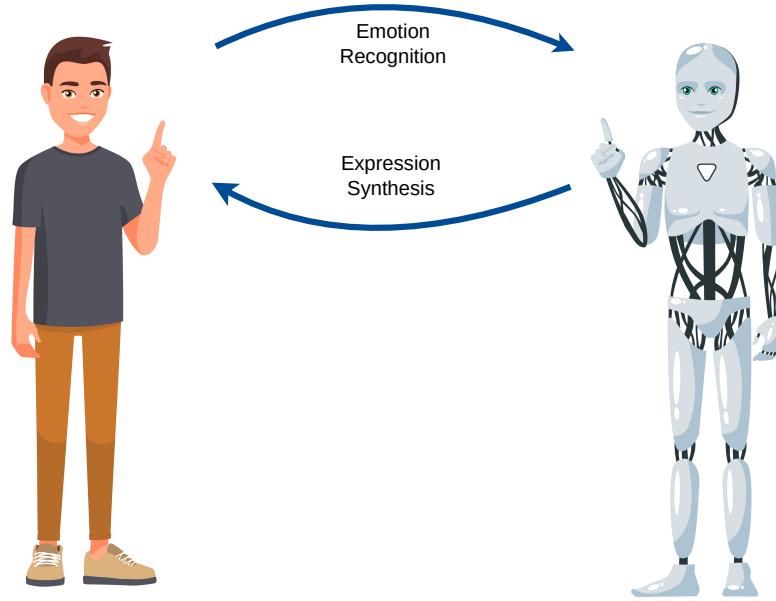


Figure 1.1: The human-robot communication loop. The robotic agent should be able to perceive human feelings and expressions, while at the same time have the ability to act expressively.

1.1 Goals and Contributions

This thesis is concerned with equipping computers and robots with the ability 1) to perceive human expressions and emotions and 2) to express themselves in a human-like way. Consequently, the thesis is split into two major parts:

Emotion Recognition In the first part, the robotic agent assumes the role of the receiving party of the human-robot communication loop. We search and discover additional streams/channels of information that can be exploited, in order to enhance automatic emotion recognition [Cowie et al., 2001]. These streams of information can be classified either as additional ways in which humans tend to express themselves (e.g., speech, facial expressions, body expressions), as well as different representations of the same underlying information stream, i.e., casting the emotion recognition problem as static (recognition from RGB images) or dynamic (recognition based on estimation of optical flow). We also find that we can benefit from implicit information streams, such as the context/scene a human resides in, or text modeling of emotional words.

Using this additional knowledge, we design and build architectures based on Deep Learning [LeCun et al., 2015], which are able to successfully encode, model, and fuse these different information streams, in order to enhance the reliability of emotion recognition systems. Some of the proposed neural networks architectures are also designed to be lightweight, finding a “sweet spot” between emotion recognition accuracy and computational footprint.

Expression Synthesis In the second part, “Expression Synthesis”, it is now the turn of the robotic agent to become the acting party of the communication loop and express itself. Here, we study expression of emotions through the prism of audio and visual speech, which is a foundational aspect of inter-personal communication and expression. Our work

in this part can be disentangled in two different directions: in the first direction, based on the fact that speech is multimodal at its core [McGurk and MacDonald, 1976, Ekman, 1984] like emotion [Darwin, 1871, Richard J. Davidson, 2002], we design methods for **expressive** synthesis of audio-visual speech. These include both adaptation of an existing audiovisual speech synthesis system to a target emotion in order to render it more life-like, as well as designing a deep neural network based expressive audiovisual speech synthesis system, capable of expressing 4 different emotions. On the other hand, the second direction addresses a different problem: facial expression manipulation and, our main contribution, perceptual 3D reconstruction of facial expressions that retains the perception of visual speech. Arguably, the utility of such a system for human-robot interaction and affective computing in general is enormous; it allows bypassing the need of collecting 3D data for model training and has immediate applications in a plethora of industries.

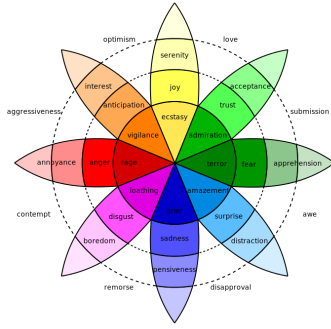
1.2 Theoretical and Computational Models of Emotion

1.2.1 Theories of Emotion

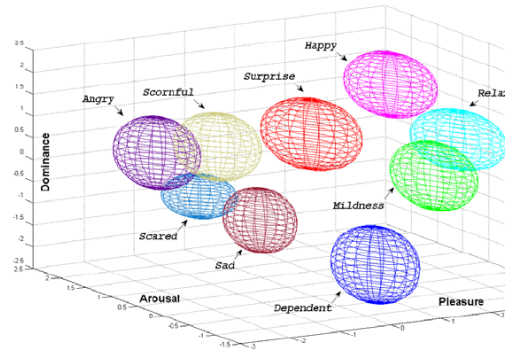
There is no scientific consensus on either the definition nor the modeling of human emotions. Most works in literature approach the subject from two different perspectives. The first perspective sees emotions as discrete entities which also manifest in specific, distinguishable ways. This study of “basic” emotions and how they manifest cross-culturally was the study of Darwin’s pioneering work in his book “The expression of the emotions in man and animals” [Darwin, 1871]. A more systematic approach and an attempt at deriving a set of “basic emotions” was presented decades later by Ekman and Friesen [Ekman, 1992], with the initial derivation of 6 basic emotions: anger, disgust, fear, happiness, sadness, and surprise. After discovering evidence of cross-culture [Ekman et al., 1969, Ekman and Friesen, 1971] constants in emotion and facial expressions, they derived the “Facial Action Coding System” (FACS) which breaks down facial expressions into individual components of muscle movements (Actions Units (AUs)) (see Figure 1.2d for examples of action units). The FACS system has subsequently associated different AUs with basic emotions [Friesen et al., 1983]. A second more complex model of categorical emotions is the wheel of Plutchik (Figure 1.2a) where 8 basic (and bipolar) emotions (joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation) can get combined to form more complex emotional states such as remorse (sadness plus disgust) or contempt (disgust plus anger). Furthermore, the model includes different intensities of emotions.

The second viewpoint sees emotions not as discrete categories, but rather as a continuous spectrum in one or more dimensions. The most famous model in this viewpoint is the VAD (Valence, Arousal, Dominance) or PAD (where valence is called pleasure) model, developed by Mehrabian and Russel [Mehrabian and Russell, 1974, Mehrabian, 1980]. Valence (or Pleasure) measures the degree of “pleasantness” of the emotion, arousal measures the “energy” of the emotional expression, and dominance the controlling/dominant nature of the emotion (e.g., fear is a submissive emotion, while anger is a dominant one). An example of the PAD model can be seen in Figure 1.2b.

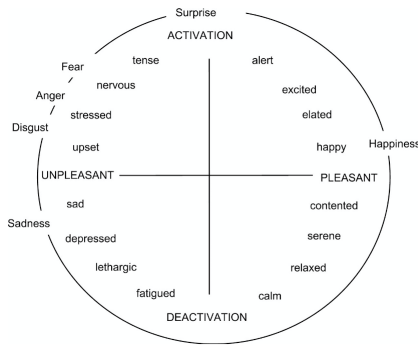
There is also a category of models that combines the above two representations, called the “circumplex” models. In these models, discrete emotions lie on a two-dimensional basis of valence and arousal. Models of this type were popularized by Russell [Russell, 1980] (see Figure 1.2c).



(a) Plutchik's wheel of emotions. More complex emotions are formed from combinations of others.



(b) The PAD model of affect and an estimate of the position of categorical emotions in the it (Source: [Zhang et al., 2010]).



(c) The circumplex model of affect by Russell (Source: <https://psu.pb.unizin.org/psych425/chapter/circumplex-models/>).



(d) Examples of action units which model facial muscle movements (Figure from [Baltrusaitis et al., 2018]).

1.2.2 Emotion Modeling in Affective Computing

Following the different viewpoints in approaching the classification of emotions, in the field of affective computing, there is no established “set of emotions”; different studies have employed different sets of emotions. The most commonly used labeling scheme however is the initial 6 basic emotions proposed by Ekman, or an extended version of it with the addition of contempt. This is a reasonable consequence of the fact that the categorical model is easy to understand and includes labels abundant in our everyday life and verbal communication [Sreeja and Mahalakshmi, 2017]. As a result, it is easier to annotate large datasets using the categorical viewpoint. Following the discrete model of affect, other works include different sets of emotions with varying cardinality. Some have also created their own data-driven categorical set of emotions [Kosti et al., 2017b], by clustering a vocabulary of affective states (see Fig. 1.3).

On the other hand, while the dimensional model is harder to adopt for dataset annotation, it allows for a more fine-grained classification and separation of emotional states. Newer annotational methodologies that use cloud aggregation of labels (e.g., amazon Turk) have allowed the creation of large-scale datasets with both the categorical and the dimensional model ([Mollahosseini et al., 2017, Kosti et al., 2017b, Luo et al., 2020]).

1. Peace: well being and relaxed; no worry; having positive thoughts or sensations; satisfied
2. Affection: fond feelings; love; tenderness
3. Esteem: feelings of favorable opinion or judgment; respect; admiration; gratefulness
4. Anticipation: state of looking forward; hoping on or getting prepared for possible future events
5. Engagement: paying attention to something; absorbed into something; curious; interested
6. Confidence: feeling of being certain; conviction that an outcome will be favorable; encouraged; proud
7. Happiness: feeling delighted; feeling enjoyment or amusement
8. Pleasure: feeling of delight in the senses
9. Excitement: feeling enthusiasm; stimulated; energetic
10. Surprise: sudden discovery of something unexpected
11. Sympathy: state of sharing others emotions, goals or troubles; supportive; compassionate
12. Doubt/Confusion: difficulty to understand or decide; thinking about different options
13. Disconnection: feeling not interested in the main event of the surrounding; indifferent; bored; distracted
14. Fatigue: weariness; tiredness; sleepy
15. Embarrassment: feeling ashamed or guilty
16. Yearning: strong desire to have something; jealous; envious; lust
17. Disapproval: feeling that something is wrong or reprehensible; contempt; hostile
18. Aversion: feeling disgust, dislike, repulsion; feeling hate
19. Annoyance: bothered by something or someone; irritated; impatient; frustrated
20. Anger: intense displeasure or rage; furious; resentful
21. Sensitivity: feeling of being physically or emotionally wounded; feeling delicate or vulnerable
22. Sadness: feeling unhappy, sorrow, disappointed, or discouraged
23. Disquietment: nervous; worried; upset; anxious; tense; pressured; alarmed
24. Fear: feeling suspicious or afraid of danger, threat, evil or pain; horror
25. Pain: physical suffering
26. Suffering: psychological or emotional pain; distressed; anguished

Figure 1.3: A data driven set of emotions created through clustering of verbal affective states [Kosti et al., 2017a].

1.3 MultiModal Emotion Recognition

Emotion is manifested from various modalities:

1. speech and paralinguistic features
2. facial expressions and other non-verbal cues (e.g., body language)
3. text and semantics

This work is concerned mainly with the second modality: computer vision for emotion recognition in images and videos, while also using in some cases the first modality (speech), for achieving more robust results.

1.3.1 Facial Expressions

The first pioneering works on automatic visual emotion recognition were almost exclusively focused on the face and facial expressions [Sebe et al., 2005, Mase, 1991]. This was possibly due to the work of Darwin [Darwin, 1871] and his study of facial expression of emotions, as well as later studies on prototypical emotions by Ekman and Friesen [Ekman and Friesen, 1967, Ekman, 1992] and the association of these with standard facial muscle movements/contractions [Friesen et al., 1983]. It is also important to note that emotion

recognition from face grew in parallel to improvements in facial detection and recognition methods, with each one borrowing research from the other.

The standard pipeline for automatic emotion recognition using facial expressions thus included face detection, feature extraction, and then classification. Some of the first representations extracted included the optical flow [Mase, 1991], as well as geometric (shape) features, using active shape models (ASM). Then, more sophisticated features also have been used such as Local Binary Patterns [He et al., 2006], Gabor filters [Deng et al., 2005], in conjunction with traditional machine learning techniques such as Support Vector Machines (SVMs) or Adaboost [Luo et al., 2013, Owusu et al., 2014].

More recently, with the breakthrough of deep learning [LeCun et al., 2015] and the development of larger databases (some examples include AffectNet [Mollahosseini et al., 2017] Aff-Wild2 [Kollias and Zafeiriou, 2018], and EmotioNet [Fabian Benitez-Quiroz et al., 2016]), methods now skip completely the feature extraction procedure, and design end-to-end neural network architectures, trained with gradient descent methods that can achieve high performance in datasets that include hundreds of thousands of samples. As an example, some of the state-of-the-art methods for facial expression recognition include residual convolutional neural networks and 3D convolutional architectures.

1.3.2 Body Expressions

Nevertheless, as we know from our personal experiences, emotion is not only displayed through the human face but also through body language. This was highlighted again by Ekman and Friesen in [Ekman and Friesen, 1967], and has subsequently been the study of many works. These early works assumed that body posture was only indicative of the intensity of the emotion, and that there existed no specific body configurations and movements associated with specific emotions. However, more recent studies have discarded this, showing that both posture and movement convey distinct information about the emotional state [Dael et al., 2012a, Tracy and Robins, 2004, Atkinson et al., 2004]. Furthermore, observing that the body influences how we perceive emotional expression by the face or vocally [Van den Stock et al., 2007, Aviezer et al., 2012, Aviezer et al., 2008]. In [Dael et al., 2012b], the Body Action and Posture Coding System (BAP) was developed; an attempt at a systematic approach to identifying specific body micro-movements and configurations that can be associated with different categorical emotions.

1.3.3 Context

In a strict sense, the context (i.e., the surrounding environment/scene) differs from the two previous channels of communication. Context includes a two-way channel, a person expressing their inner emotion affects his environment (e.g., a person throwing or breaking something), but also is affected by their environment (e.g., a graveyard elicits different emotions when compared to a theme park). In this way, context is a more implicit stream of information, and for this reason, it has been largely ignored in automatic emotion recognition.

Nevertheless, recent studies have underscored that emotional perception can be heavily affected by the context/scene [Carroll and Russell, 1996, Barrett and Kensinger, 2010]. The first attempt at introducing context as additional information for visual emotion recognition was done by Kosti et al. [Kosti et al., 2017b, Kosti et al., 2017a] in 2017. They introduced the large-scale dataset EMOTIC (Emotions in Context), which contains annotated emotions of individuals in multiple scenes. Subsequently, they introduced a

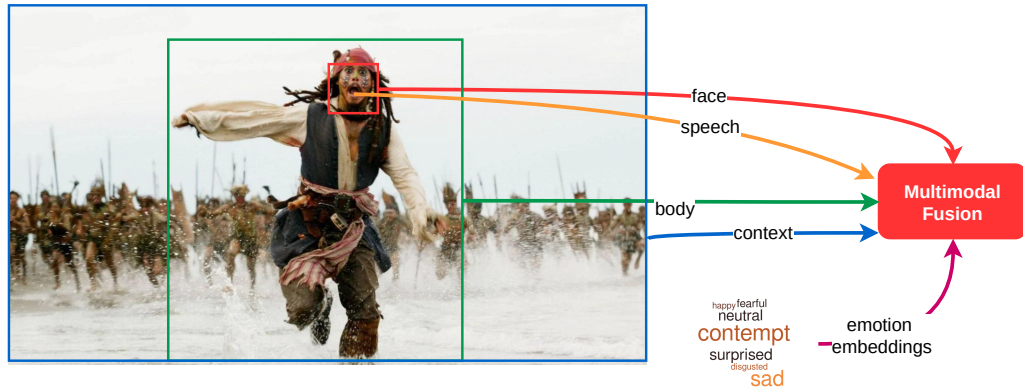


Figure 1.4: All sources of information that can be used to infer the emotion of Jack Sparrow: body posture, facial expressions, scene context, speech, and emotional embeddings.

two-way convolutional neural network architecture which fused information from both the human and the scene, in order to recognize the depicted emotion.

1.3.4 Speech

As a channel of emotion, speech and non-verbal sounds have been thoroughly studied, on par with studies on facial expressions [Scherer, 1986]. In a similar fashion as well, initial speech emotion recognition systems included feature extraction, and a great deal of effort has been put into recognizing those speech features that are more representative of emotion [Eyben et al., 2015, El Ayadi et al., 2011]. Later works leveraging deep learning have skipped the feature extraction step as well, acting upon either the raw speech waveform, or the spectrogram representation [Trigeorgis et al., 2016, Tzirakis et al., 2017a, Zhang et al., 2018b, Lim et al., 2016, Zhao et al., 2019].

1.3.5 Emotional Embeddings

As a final implicit stream of information we consider emotional embeddings. Emotion recognition, unlike other classification problems, has the special attribute that the considered classes (when using categorical modeling), do not have the same semantic distance between them. For example, happiness is semantically closer to excitement compared to anger. Consequently, in order to consider this additional information when building emotional classifiers we can resort to using word embeddings for the emotional labels. Word embeddings [Pennington et al., 2014] are text representations that cast different words into a high dimensional subspace where the previously stated semantic meaning is retained.

In Figure 1.4 we show an example in which all the above mentioned streams of information can be used to infer the depicted emotion. Note how emotional embeddings do not directly correspond to the considered scene but are used to enforce semantics on the classifier.

1.4 Synthesis of Facial Expressions

With the term “Expression Synthesis”, we refer to bestowing upon robotics agents and virtual avatars the capability of expressions while interacting with humans. We study synthesis of expressions through the prism of *speech*, which traditionally has been the

main channel of human-machine communication (apart from text in screens). However, synthesis of speech does not include only the generation of a human-like voice; speech is multimodal in nature [Mattheyses and Verhelst, 2015, Ekman, 1984] and important information is included in the visual stream of information (i.e. the human face and more generally the human head and its movements) along with the acoustic stream. Importantly, several works provide extensive evidence that the inclusion of a visual stream of information increases the intelligibility of speech, especially under noise, even when the face is a virtual talking head [Beskow, 1996, Mattheyses and Verhelst, 2015].

1.4.1 AudioVisual Speech Synthesis and 3DMMs

Before considering generation of facial expressions that correspond to emotion, we must consider audio and, our main modality, visual speech. While the first talking head models were created manually through a mesh of polygons and vertices and moved through rules based on FACS [Beskow, 1996, Le Goff and Benoît, 1996, Pelachaud et al., 1996], more recent methods have benefited from statistical techniques and deep learning in order to create more realistic models. One of the most important contributions in this domain, is the Active Appearance Model [Cootes et al., 2001] and the 3D morphable model [Blanz and Vetter, 1999] which can be built by applying statistical techniques such as PCA, on a large set of 3D scans, in order to obtain 3D models capable of representing various head shapes and identities, as well as facial expressions. Following the work of Blanz and Vetter [Blanz and Vetter, 1999], there have been proposed multiple 3D models built from scans: FLAME [Li et al., 2017](Fig. 1.5, LSFM [Booth et al., 2018b], the Basel Face Model [Paysan et al., 2009, Gerig et al., 2018], the FaceWarehouse model [Cao et al., 2013], FaceScape [Yang et al., 2020], and FaceVerse [Wang et al., 2022] models.

These models can be used for a vast number of applications: human-machine interaction, graphics and gaming, entertainment, Visual Effects, and Computer Generated Imagery. An example is visual speech synthesis where deep architectures have been proposed where an input audio is mapped to the corresponding visual speech, through predictions of the AAM/3DMM parameters. One such architecture is VOCA [Cudeiro et al., 2019a], which uses a pretrained automatic speech recognition network to extract features from audio, and then predicts the position of the approx. 5,000 vertices which comprise the FLAME model, so that the output animated 3D model is uttering the input audio. In Neural Voice Puppetry [Thies et al., 2020], a similar network predicted the 3D parameters of the BFM model that correspond to an input speech. Afterwards, the rendered 3D model was also fed into a neural renderer, in order to add photorealism and texture to the model. Other important works that map input audio to video include [Zhang et al., 2013, Song et al., 2020, Zhou et al., 2020, Fan et al., 2015].

Compared to audio input, using text as an input constitutes a hard problem since there is no apparent alignment between the input and the output. As a result, there are significantly fewer works who have tried to predict audio **and/or** visual speech through text alone. Some older examples include the work of Schabus et al. [Schabus et al., 2014], which used Hidden Markov Models(HMMs), in order to jointly synthesize audio and predict visual speech trajectories, in the form of raw 3D markers. Another example is the coupled HMM of Xie and Liu [Xie and Liu, 2007], which again predicted audiovisual speech, modeling the visual modality by mapping phonemes to visemes. In a follow-up work in [Xie et al., 2014], Xie et al. also introduced a videorealistic method of text-to-audiovisual speech, this time powered by trajectory HMMs. A recent technique powered by deep learning was based on the Tacotron 2 speech synthesis model and was presented

in [Abdelaziz et al., 2021]. There, along with the spectrogram, the Tacotron 2 model was also trained to predict the blendshapes of a 3D model. It is worth mentioning that one can cast the text-to-audiovisual speech synthesis problem in a cascaded way: first use text to synthesize high-fidelity speech, and then use speech and an audio-driven technique in order to get the final talking head [Karras et al., 2017].

1.4.2 Adding Expressions to AudioVisual Speech Synthesis

Starting with the works of Anderson et al. and Wan et al. [Anderson et al., 2013, Wan et al., 2013] who performed text driven expressive audiovisual speech synthesis with AAMs and HMMS, several works have attempted to create talking heads with expressivity. These again mostly use audio (either as features, or raw) as input, and add emotional modeling inside the network/method that predicts the parameters of the 3D model. Karras et al. [Karras et al., 2017], trained a database of latent vectors for each different audio window, in order to create an initial version of an emotion database of style vectors. After pruning out irrelevant vectors, the rest of the vectors can be used as input in their network in order to synthesize expressive audiovisual speech. In [Pham et al., 2017], Pham et al. introduced an LSTM-based approach in order to animate the face and head of the FaceWareHouse 3D model, using audio input from an emotional database. Sadoughi and Busso [Sadoughi and Busso, 2017] extracted articulatory and emotional features from speech, and then used them as conditioning, in order to generate expressive talking head animations. Regarding text input, however, only a handful of works have attempted text-driven expressive audiovisual speech synthesis. In [Shaw and Theobald, 2016], modeling of emotional expressions is achieved using AAMs and Independent Component Analysis (ICA). Another example is the work of Dahmani et al. [Dahmani et al., 2019], which used a conditional variation auto-encoder in order to predict speech vocoder features and blendshapes of a 3D model.

1.4.3 3D Model Fitting and Data Availability

A bottleneck in the process of synthesizing expressive talking heads is that of data requirements, especially when considering large deep neural network models. Currently, the most perceptually convincing talking head models require collecting dataset with stereo recordings, and in some cases using markers on the human face for easier annotation. Methods that attempt to bypass this step rely on using monocular 3DMM estimation in unconstrained databases [Zhou et al., 2020, Thies et al., 2020, Song et al., 2020], in order to obtain “ground truth” 3D annotations, which can then be used to train the network, along with the input speech. Current state-of-the-art monocular 3D estimation techniques such as DECA [Feng et al., 2021] and 3DDFA [Guo et al., 2020] are able to reconstruct fine details of the 3D facial geometry. However, when applied on videos of individuals in verbal communication, significant artifacts arise in the reconstruction of the mouth shape and motion, which negatively affect human perception. This problem creates a significant gridlock in the acquisition of robust and perceptually accurate 3D reconstruction, without relying on highly constrained stereo recordings, and subsequent annotations. As we will show in Chapter 7, an important contribution of this thesis is the introduction of the first method for perceptual 3D reconstruction of human faces during verbal communication, alleviating the aforementioned drawback, and opening up easier acquisition of 3D data from in-the-wild videos.

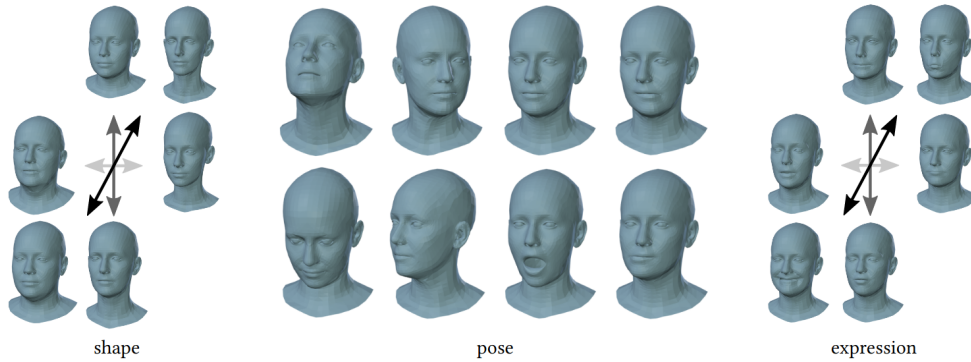


Figure 1.5: The FLAME [Li et al., 2017] 3D morphable model.

1.4.4 Model-Free generation

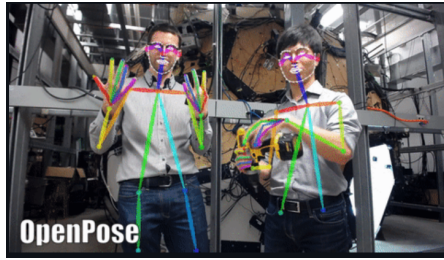
A final subset of methods we should also mention is that of model-free generation of visual speech, that is, not relying on a parametric model of the face, but directly using and rendering image pixels. Some of the initial attempts in the pre-deep learning era used unit-selection methods [Cao et al., 2005, Melenchón et al., 2009, Liu and Ostermann, 2011], where a large database of images was stored, and at inference time, according to the desired output, the method selected the corresponding frames from the database and blended them together. Newer works that used deep architectures include “You said that?” [Chung et al., 2017], where a CNN model is using MFCC features and a still image as input, and outputs a video of the image uttering the speech segment. A similar real-time approach, using the same inputs but relying on temporal Generative Adversarial Networks (GANs) was presented in [Vougioukas et al., 2019].

1.5 Enabling Technologies

Emotion recognition and expression synthesis are higher-level problems; before attempting to estimate the emotion of a person we need to be able to detect their face, mouth, and body posture. As a result, the techniques presented in this dissertation rely upon these technologies, in order to first localize and extract landmarks of the person.

1.5.1 Body Pose Estimation

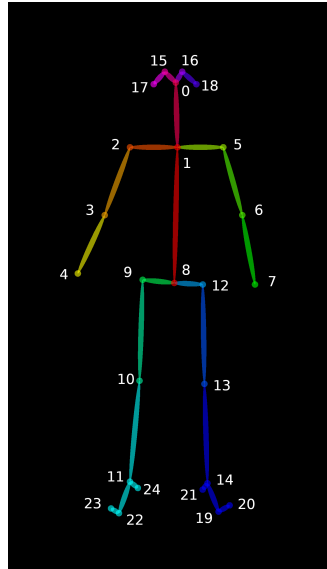
Body pose estimation refers to localizing a human body, and then extracting key landmarks from the body. An example of a body model is BODY25 [Cao et al., 2017], which is shown in Figure 1.6c. The result of a body pose estimator, which can be either the 2D/3D points, or the image cropped around the body, can then be used to perform higher level tasks, such as action recognition, intent recognition, and emotion recognition, among others. While there is a large interest in body estimation and techniques are improving with a large rate, there are some established methods in the literature, which already achieve remarkable results in this task. One such example is OpenPose [Cao et al., 2019] (Figure 1.6a), which achieves realtime multi-person 2D pose estimation, using Part Affinity Fields and bipartite matching, in order to correctly associate the parts of multiple persons in an image. Other methods include AlphaPose [Li et al., 2021b], MediaPipe [Lugaresi et al., 2019], DensePose [Güler et al., 2018], the latest one based on transformers, ViTPose [Xu et al., 2022]. In this thesis, we use these techniques as building blocks, in order to move



(a) Body pose estimation using OpenPose (Source: [Gerig et al., 2018]).



(b) Body pose estimation using MediaPipe (Source: [Lugaresi et al., 2019]).



(c) The BODY25 model which includes 25 landmarks across the entire human body.

Figure 1.6: The OpenPose and MediaPipe BlazePose methods for body pose estimation along with an example skeletal topology.

on and tackle higher level problems, i.e., emotion recognition and expression synthesis. In some cases we rely on the detected 2D landmarks to estimate emotion, while in other cases we use convolutional neural networks on the cropped body image.

1.5.2 Face Alignment

The second important building block of the techniques described in this thesis is Face Detection and Alignment. Following the seminal work of Viola and Jones for face detection [Viola and Jones, 2004], like with body pose estimation, there has been a growing line of research focused around detecting the human face, and extracting 2D and 3D landmarks. The most widespread method for face alignment is probably FAN, the work of Bulat and Tzimiropoulos [Bulat and Tzimiropoulos, 2017], which is built by stacking four HourGlass networks. Examples of estimation with this method are shown in Fig. 1.7. Another notable works in this field is the lightweight 3DDFA v1 and v2 [Guo et al., 2020], which jointly predicts not only 2D and 3D landmarks, but also parameters of the 3D Basel Face Model [Paysan et al., 2009].



Figure 1.7: Face landmarks detected by FAN [Bulat and Tzimiropoulos, 2017].

1.6 Thesis Outline

The rest of the thesis is split in two Parts, “Emotion Recognition” and “Expression Synthesis”, each with 3 chapters, organized as follows:

1. In Chapter 2, we propose a two-branch deep learning architecture with hierarchical labels and losses, which is able to jointly model body movements with facial expressions and identify emotions in children during challenging child-robot interaction (CRI) scenarios. In this chapter, we also introduce the BabyRobot Emotion dataset, which includes acted and spontaneous affective expressions of children, which fills in the gap in the literature of available children emotional datasets.
2. In Chapter 3 we further introduce an additional stream of information (speech), as well as a more advanced convolutional neural network (CNN) based architecture that takes into account explicitly the dynamic nature of emotion as well. In this chapter, we also take into account several considerations for CRI and explore the trade-offs between computational burden and system performance.
3. In the last Chapter 4 of Part I we introduce the additional implicit stream of scene and context, as useful auxiliary information for emotion recognition. Furthermore, acknowledging the fact that emotions do not have the same semantic distances with each other, we use word embeddings to introduce an additional semantic loss that increases the performance of emotion recognition in the wild.
4. Chapter 5 is the first chapter of Part II. Here, we introduce a method for adapting an expressive audiovisual speech synthesis system to new emotions using only a small amount of data. Furthermore, we use interpolation of HMMs (hidden Markov models) in order to allow the system to combine emotions and build more complex expressions. This chapter also introduces the CVSP-Expressive Audiovisual Corpus, a dataset that includes a total of 3600 sentences spoken in 4 emotions from one speaker.
5. Next, in Chapter 6, drawing tools from deep learning, we design a deep expressive audiovisual speech synthesis system, capable of expressing 4 different emotions, which outperforms both HMM synthesis, as well as unit-selection synthesis
6. The penultimate Chapter 7, includes the second direction of Part II, which is facial expression manipulation and, most importantly, perceptual 3D reconstruction of facial expressions, which retain the perception of visual speech. This is done through

our perceptual “lip read” loss, which is able to guide a 3D fitting process in order to accurately capture the diverse deformities of the mouth.

7. Finally, in Chapter 8, we conclude the thesis and its contributions and list indicative future directions.

Part I

Emotion Recognition

2

Fusing Body Posture with Facial Expressions for Emotion Recognition in Children and Adults

2.1 Introduction

In the first chapter of Part I of this thesis, we study the importance of body language, when expressing emotions and how it increases the robustness and performance of automatic emotion recognition systems. While the face has traditionally been the primary studied medium in emotion research [Ekman and Friesen, 1967], more recent studies have highlighted the importance of body language, suggesting that emotion is equally conveyed through bodily expressions and actions in most cases [De Gelder, 2009, Wallbott, 1998], while both the static body posture as well as the dynamics [Atkinson et al., 2004, Calvo et al., 2015] contribute in its perception. Furthermore, it is evident that in real-life scenarios we often decode the emotions of our interlocutor or people in our surroundings by observing their body language, especially in cases where the face of the subject in question is occluded, hidden, or far in the distance. In general, the body language can act both as a supportive modality, in which case it enforces the confidence in an already recognized emotion from the face or provides crucial missing information (e.g., in cases where the face cannot reliably identify the emotion due to its intensity [Aviezer et al., 2012]), as well as a primary modality, in which case it is the only source of information from which we can deduce the emotion.

Furthermore, there are emotions such as pride [Tracy and Robins, 2004] that are more discernible through the body rather than face observation. An also consistent finding in multiple studies is the fact that considering both body and face concurrently increases emotion recognition rates [Van den Stock et al., 2007]. Aviezer et al. also point out in [Aviezer et al., 2012] that the body can be a deciding factor in determining intense positive or negative emotions.

Figure 2.1 shows some primary examples where body language is useful for correctly decoding emotions. In Figure 2.1a we can deduce that the child expresses a negative emotion. It is also important to note that the direction of the hands shows us the source of the negative feelings, which is something that facial expressions do not reveal. Another example is presented in Figure 2.1b (Figure from [Aviezer et al., 2012]), where it can be seen that without the whole body we cannot identify whether the emotion of the person



Figure 2.1: The importance of body expressions in affect recognition.

is positive or negative, due to its intensity. In Figure 2.1c we can identify sadness by the head pose, while in Figure 2.1d the body acts as a supportive modality; we can also deduce anger just by the facial expression.

Nevertheless, to date, most research on automatic recognition of emotion has focused primarily on facial expressions [Jung et al., 2015, Kuo et al., 2018], with only a few including emotional body expressions into the recognition loop [De Gelder, 2009]. Motivated by the preceding analysis, in this chapter we build a method for automatic recognition of affect in child-robot interaction scenarios, which leverages the body posture along with facial expressions, for increased performance and robustness.

Emotion recognition in Social Robotics Social robotics is a fairly new area in robotics that has been enjoying a swift rise in its applications, some of which include robot-assisted therapy in adults and children [Belpaeme et al., 2013], activities of daily living [Broadbent et al., 2009], and education [Belpaeme et al., 2018]. A critical capability of social robots is empathy: the capacity to correctly interpret the social cues of humans that are manifestations of their affective state. Empathic agents are able to change their behavior and actions according to the perceived affective states and as a result establish rapport, trust, and healthy long-term interactions [Bickmore and Picard, 2005]. Especially in the field of education, empathic robot behaviors that are congruent with the child’s feelings increase trust and have a positive impact on the child-robot relationship, whereas incongruent behavior has a significantly negative effect [Leite et al., 2014].

An important factor in many social robot applications, and especially in child-robot interaction (CRI) [Tsiami et al., 2018], is the fact that the flow of interaction is unpredictable and constantly fluctuating [Ros et al., 2011]. Although interaction with adults

can usually be restricted and controlled, the spontaneous nature of children fails to meet this criterion and becomes a true challenge. A direct implication is the fact that robots can no longer rely only on facial expressions to recognize emotion, which is the main visual cue employed in automatic affect recognition [De Gelder, 2009], but also have to take into account body expressions that can stay visible and detectable even when the face is unobservable.

Based on the aforementioned, our contributions in this chapter can be summarized as follows:

- We propose a method based on Deep Neural Networks (DNNs) that fuses body posture skeleton information with facial expressions for automatic recognition of emotion. The networks can be trained both separately and jointly, and result in significant performance boosts when compared to facial-only expression baselines.
- We use hierarchical multi-label annotations (Figure 2.2), that describe not only the emotion of the person as a whole but also the separate body and facial expressions. These annotations allow us to train, either jointly or separately, our hierarchical multi-label method, providing us with computational models for the different modalities of expressions as well as their fusion.
- We develop and analyze a database containing acted and spontaneous affective expressions of children participating in a CRI scenario, and we discuss the challenges of building an automatic emotion recognition system for children. The database contains emotional expressions both in face and posture, allowing us to observe and automatically recognize patterns of bodily emotional expressions across children in various ages.

2.2 Related Work

As mentioned in the first two chapters, the overwhelming majority of previous works in emotion recognition from visual cues have focused on using only facial information [De Gelder, 2009]. Recent surveys however [Noroozi et al., 2018, Karg et al., 2013, Klein-smith and Bianchi-Berthouze, 2013] highlight the need for taking into account bodily expression as additional input to automatic emotion recognition systems, as well as the lack of large-scale databases for this task.

Gunes and Piccardi [Gunes and Piccardi, 2009] focused on combining handcrafted facial and body features for recognizing 12 different affective states in a subset of the FABO database [Gunes and Piccardi, 2006] that contains upper body affective recordings of 23 subjects. Barros et al. [Barros et al., 2015] used Sobel filters combined with convolutional layers on the same database, while Sun et al. [Sun et al., 2018] employed a hierarchical combination of bidirectional long short-term memory (LSTM) and convolutional layers for body-face fusion using support vector machines. Piana et al. [Piana et al., 2016] built an automatic emotion recognition system that exploits 3D human pose and movements and explored different higher level features in the context of serious games for autistic children.

Recognition of children’s affect, as we mentioned above, is crucial in creating empathic robots deployed during CRI. Thus, interesting research works have been presented for designing advanced emotion recognition modules to equip intelligent robots. Goulart et al. proposed in [Goulart et al., 2019] a computational system for estimating children’s emotion during CRI, deploying visual information from both RGB and infrared thermal cameras. The proposed system detects the facial regions of interest that are relevant

to five basic emotions. Lopez-Rincon in [Lopez-Rincon, 2019] proposed a Convolutional Neural Network (CNN) combined with a Viola-Jones face detector, trained using the AffectNet database [Mollahosseini et al., 2017], and tuned it with children data in order to recognize children facial emotional expressions. Marinou et al. [Marinoui et al., 2018] proposed an automated approach using 3d skeleton data and a CNN architecture for action and continuous emotion recognition during robot-assisted therapy sessions of children with Autism Spectrum Disorders (ASD). In [Castellano et al., 2013], a system perceived children’s affective expressions while playing chess with an iCat robot and modified the behavior of the robot to be more friendly and increase children’s engagement. Similarly, Filippini et al. [Filippini et al., 2020] classified children’s emotional states to understand their engagement level using thermal signal analysis during interaction with the Mio Amico Robot. An adaptive robot behavior based on the perceived emotional responses was also developed for a NAO robot in [Tielman et al., 2014].

Bänziger et al. [Bänziger et al., 2012] introduced the GEMEP (GENeva Multimodal Emotion Portrayal) corpus, the core set of which includes 10 actors performing 12 emotional expressions. In [Dael et al., 2012b], Dael et al. proposed a body action and posture coding system similar to the facial action coding system [Friesen and Ekman, 1978], which is used for coding human facial expressions, and subsequently utilized it in [Dael et al., 2012a] for classifying and analyzing body emotional expressions found in the GEMEP corpus.

In [Castellano et al., 2008], Castellano et al. recorded a database of 10 participants performing 8 emotions, using the same framework as the GEMEP dataset. Afterward, they fused audio, facial, and body movement features using different Bayesian classifiers for automatically recognizing the depicted emotions. In [Psaltis et al., 2016], a two-branch face-body late fusion scheme is presented by combining handcrafted features from 3D body joints and action units detection using facial landmarks.

Regarding the application of affect recognition in CRI, the necessity of empathy as a primary capability of social robots for the establishment of positive long-term human-robot interaction has been the research focus of several studies [Bickmore and Picard, 2005, Leite et al., 2014]. In [Castellano et al., 2013], Castellano et al. presented a system that learned to perceive affective expressions of children playing chess with an iCat robot and modify the behavior of the robot resulting in a more engaging and friendly interaction. An adaptive robot behavior based on the perceived emotional responses was also developed for a NAO robot in [Tielman et al., 2014]. In [Marinoui et al., 2018], 3D human pose was used for estimating the affective state of the child in the continuous arousal and valence dimensions, during the interaction of autistic children with a robot.

Compared to previous approaches, in this chapter we introduce the concept of hierarchical multi-labels, by taking into account the medium through which a person expresses its emotion (face and/or body). These labels are used in a novel neural network architecture that utilizes multi-stage losses, offering tighter supervision during training, as well as different sub-networks, each specialized in a different modality. Our method is end-to-end, uses only RGB information, and is built with the most recent ML architectures. The efficiency of the proposed framework is validated by performing extensive experimental results on two different databases, one of which includes emotions acted by children and was collected by us during the EU project BabyRobot¹.

¹More info: <http://babyrobot.eu/>

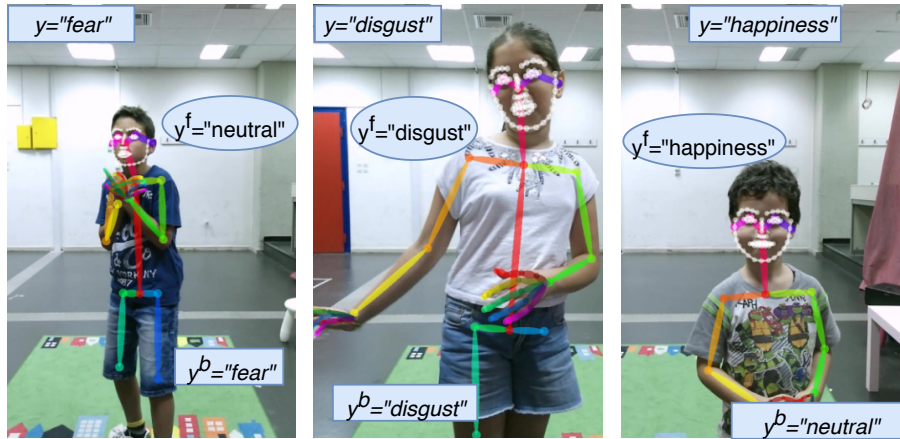


Figure 2.2: Hierarchical multi-labels for affect recognition via body and face cues, where y denotes the whole body emotion label, y^f the facial expression label, and y^b the body expression one.

2.3 Whole Body Emotion Recognition using Deep Learning

2.3.1 Hierarchical Emotional Labels

A problem that arises when dealing with spontaneous (i.e., not acted) or in-the-wild data is the fact that different individuals express themselves through different modalities, depending on which cue they prefer using (body, face, voice) [Calvo et al., 2014]. This fact is cumbersome for supervised learning algorithms, e.g., in samples where an emotion label corresponds to the facial expression only and not the body, which means that the subject in question preferred to use only the face while the body remained neutral. In such data, one way to alleviate this issue is to include hierarchical labels, which first denote the ground truth labels of the different modalities. Examples of hierarchical multi-labels are shown in Figure 2.2, where y denotes the emotion the human is expressing (which we call the “whole” body label), y^f the emotion that is conveyed through the face (i.e., $y^f = y$ if the subject uses the face to express y , else $y^f = \text{“neutral”}$), and y^b the emotion that is conveyed through the body (i.e., $y^b = y$ if the subject uses the body, else $y^b = \text{“neutral”}$).

2.3.2 Method

Based on the aforementioned analysis, Figure 2.3 presents our DNN architecture for automatic multi-cue affect recognition using hierarchical multi-label training (HMT). We assume that we have both the whole body label y , as well as the hierarchical labels y^f for the face and y^b for the body. The network initially consists of two different branches, with one branch focusing on facial expressions, and one branch focusing on body posture. The two branches are then combined at a later stage to form the whole body expression recognition branch that takes into account both sources of information. This design allows setting up different losses on different stages of the network based on the hierarchical labels, offering stricter supervision during training. The output of the network is the recognized emotional state of the person detected in the input video.

Facial Expression Recognition Branch The facial expression recognition branch of the network is responsible for recognizing emotions by decoding facial expressions.

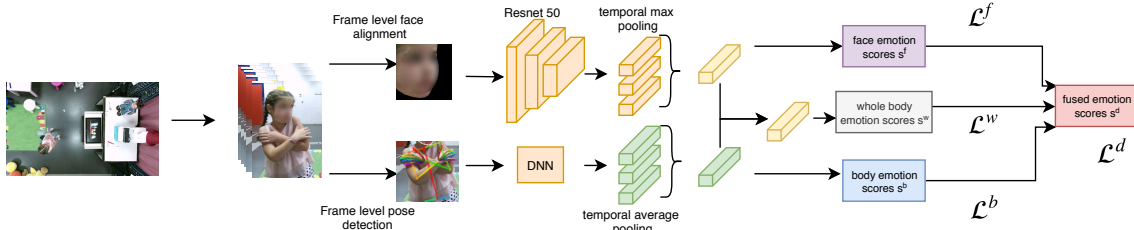


Figure 2.3: Hierarchical multi-label training for recognition of affect from multiple visual cues in CRI.

Considering a sequence of frames $I_i|_{i=1,\dots,N}$, we first apply at each one a head detection and alignment algorithm in order to obtain the cropped face image (see Section 2.5.1). This is subsequently fed into a Residual Network [He et al., 2016] CNN architecture to get a 2048-long feature vector description of each frame $H_i^f|_{i=1,\dots,N}$. Then, we apply temporal max-pooling over the video frames to obtain the representation of the facial frame sequence. By assuming that the feature map obtains its maximum values in frames where the facial expression is at peak intensity, max-pooling selects only the information regarding the facial expressions at their peak over the frame sequence. Then, we apply a fully connected (FC) layer on H^f to obtain the facial emotion scores, s^f .

We can calculate the loss obtained through this branch as the cross entropy $\mathcal{L}^f(y^f, \tilde{s}^f)$ between the face labels y^f and the probabilities of the face scores \tilde{s}^f obtained via a softmax function:

$$\mathcal{L}^f(y^f, \tilde{s}^f) = - \sum_{c=1}^C y_c^f \log \tilde{s}_c^f \quad (2.1)$$

with C denoting the number of emotion classes and y_c^f is a binary indicator denoting whether the class was correctly predicted.

Body Expression Recognition Branch In the second branch, for each frame of the input video $I_i|_{i=1,\dots,N}$, we apply a 2D pose detection method in order to get the skeleton $J_i \in \mathbb{R}^{K \times 2}$, where K is the number of joints in the detected skeleton (see Section 2.5.1). The 2D pose is then flattened and input as a vector into a DNN in order to get a representation $H_i^b|_{i=1,\dots,N}$. We then apply global temporal average pooling (GTAP) over the entire input sequence:

$$H^b = \frac{1}{N} \sum_{i=1}^N H_i^b \quad (2.2)$$

In contrast to the face branch, we use temporal average pooling for the body branch in order to capture the general pattern of the features during the temporal sequence and not completely discard temporal information. The scores for the body emotion \tilde{s}^b are obtained by passing the pose representation of the video H^b over a fully connected (FC) layer. The loss in this branch is the cross-entropy loss (Eq. 2.1) between the body labels y^b and the probabilities \tilde{s}^b , $\mathcal{L}^b(y^b, \tilde{s}^b)$.

Whole Body Expression Recognition Branch In order to obtain whole-body emotion recognition scores \tilde{s}^w , we concatenate H^f and H^b and feed them through another FC. We then use the whole body emotion labels y to obtain the whole body cross-entropy loss between the whole body labels y and the probabilities \tilde{s}^w , $\mathcal{L}^w(y, \tilde{s}^w)$. Essentially, this branch performs feature fusion.

Fusion Finally, we employ a late score fusion scheme as follows: we concatenate the scores \tilde{s}^f , \tilde{s}^b , and \tilde{s}^w and use a final FC in order to obtain the fused scores \tilde{s}^d . This way we get a final loss $\mathcal{L}^d(y, \tilde{s}^d)$ which is the cross entropy between the whole body labels y and \tilde{s}^d .

During training, the loss that is backpropagated through the network is:

$$\mathcal{L} = \mathcal{L}^f(y^f, \tilde{s}^f) + \mathcal{L}^b(y^b, \tilde{s}^b) + \mathcal{L}^w(y, \tilde{s}^w) + \mathcal{L}^d(y, \tilde{s}^d) \quad (2.3)$$

The network final prediction of the human affect in the video is obtained by the fusion score vector \tilde{s}^d .

2.4 The BabyRobot Emotion Database

In order to evaluate our method, we have collected a database that includes multimodal recordings of children interacting with two different robots (Zeno [Robokind, 2022], Furhat [Furhat, 2022]), in a laboratory setting that has been decorated in order to resemble a child’s room (Figure 2.4).

We call this dataset the BabyRobot Emotion Database (BRED). BRED includes two different kinds of recordings: *Pre-Game Recordings* during which children were asked by a human to express one of six emotions, and *Game Recordings* during which children were playing a game called “Express the feeling” with the Zeno and Furhat robots. The game was touchscreen-based, and throughout its duration children selected face-down cards, each of which represented a different emotion. After seeing the cards, the children were asked to express the emotion, and then one of the robots followed up with a facial gesture that expressed the emotion as well. A total of 30 children of ages between 6 to 12 took part in both recordings. It is important to note that we did *not* give any guidelines or any information to the children on how to express their emotions. The experimental procedure was approved by an Independent Ethics Committee from the Athena Research and Innovation Center in Athens, Greece.

The emotions included in the database are: *Anger, Happiness, Fear, Sadness, Disgust, and Surprise*, the 6 basic emotions included in Ekman and Friesen’s initial studies [Ekman and Friesen, 1971]. This categorical representation of emotion is the most commonly used in research studies of automatic emotion recognition [Gunes and Pantic, 2010], and is typically adopted across different databases of emotional depictions [Li and Deng, 2018]. When compared to dimensional approaches (e.g., valence/arousal space), the categorical emotional approach is less flexible in expressing more complex emotions, however, it is easier to annotate [Gunes and Schuller, 2013].

Hierarchical Database Annotations In total, the initial recordings included 180 samples of emotional expressions from the “Pre-Game” session and 180 samples from the “Game” session (30 children \times 6 emotions for both sessions). The annotation procedure included three different phases. In the first phase, 6 different annotators filtered out recordings where the children did not perform any emotion (due to shyness, lack of attention, or other reasons), and identified the temporal segments during which the expression of emotions takes place (starting with the onset of the emotion and ending just before the offset). In the second phase, 2 annotators validated the annotations of the previous phase. Finally in the third phase, three different annotators annotated the videos hierarchically, by indicating for each video whether the child was using the face, body, or both, to express the emotion. The final hierarchical labels were obtained using majority voting



Figure 2.4: The experimental setup of the BRED Database and snapshots showing children playing the “Express the feeling” game.

Emotion	% using facial exp.	% using body exp.
Happiness	100%	20%
Sadness	86%	49%
Surprise	100%	43%
Fear	42%	98%
Disgust	98%	42%
Anger	85%	70%

Table 2.1: Hierarchical multi-label annotations of the BabyRobot Emotion Dataset (BRED) depicting usage of body and facial expressions for each emotion.

over the three annotations. Inter-annotator agreement was also measured using Fleiss’ kappa coefficient [Fleiss and Cohen, 1973], with a value 0.48 for the face labels and 0.84 for the body labels. The values show that for the body labels we have an almost perfect agreement between the annotators, while for the face labels there are some cases where the annotators disagreed due to really slight facial expressions.

In total, the database features 215 valid emotion sequences, with an average length of 72 frames at 30FPS. The smaller number of valid sequences extracted from the 360 initial recordings shows that, when collecting data from children, attention should be paid in data validation and cleaning. Table 2.1 contains more insights regarding the database and its annotations. For each different emotion, we show the percentage of samples where the child used its face/body to depict emotion, against the number of total samples. We observe that almost all children used their body to express fear (98%), while less than half used their face. Another emotion where a large percentage of children utilized the body is anger (70%). To indicate happiness, surprise, and disgust, almost *all* children used facial expressions (100%, 100%, and 98%, respectively). Table 2.2 also contains some of the annotators’ observations regarding the bodily expression of emotion in BRED, as well as examples from the database. All images include facial landmarks (although we do not use them in any way in our method) in order to protect privacy.

The newly collected BRED dataset is very challenging as it features many intra-class variations, multiple poses, and in many cases similar body expressions for different classes. These include the similar pattern of hand crossing in anger and fear,


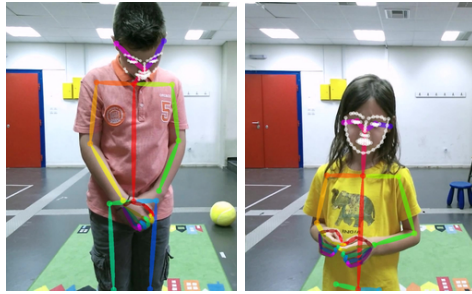
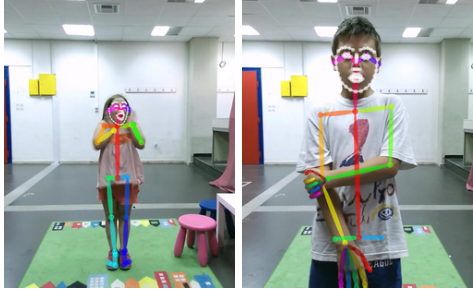

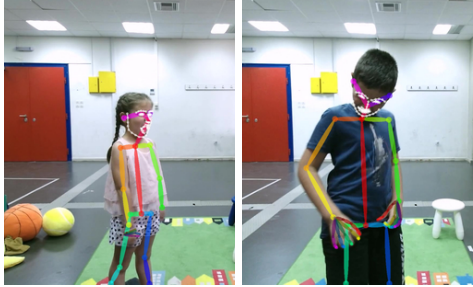

<p>happiness mainly facial, rare jumping and/or open raised hands, body erect, upright head</p> 	<p>sadness crying (hands in front on face), motionless, head looking down, contracted chest</p> 
<p>surprise quick eye gaze, weak facial expressions, arms crossed in front of body, head sink</p> 	<p>fear expanded chest, hand movement without specific patterns, either positive or negative</p> 
<p>disgust mainly with facial expression (tongue out), movement away from/hands against robot</p> 	<p>anger clenched fists, arms crossed, squared shoulders</p> 

Table 2.2: Patterns of bodily expression of emotion in the BRED corpus and example images.

and lowering of the head pose in fear and sadness. The BRED dataset is available at <https://zenodo.org/record/3233060>.

2.5 Experiments

In this section we present our experimental procedure and results ². We first perform an exploratory analysis of the different branches and pathways of the HMT architecture of Figure 2.3 on the GEMEP (GENEVA Multimodal Emotion Portrayals) database [Bänziger et al., 2012]. As far as we are aware, this is the only publicly available video database that includes annotated whole body expressions of emotions. We believe that databases of upper body depictions, such as FABO [Gunes and Piccardi, 2006] where the subjects are

²The source code is available at <https://github.com/filby89/body-face-emotion-recognition>.

Video		Frame	
Method	ACC	Method	ACC
Body br. (TCN)	0.31	Body br.	0.23
Body br. (LSTM)	0.28	Face br.	0.21
Body br. (GTAP)	0.34	Whole Body br.	0.33
Face br.	0.43		
Whole Body br.	0.51		
Human Baseline	0.47 [Bänziger et al., 2012]		

Table 2.3: Accuracy results for the body, face, and whole body branch on the GEMEP database (12 classes).

sitting, restrict body posture expression and force the subjects to focus mostly on using their hands. Our main evaluation is then conducted on BRED where we experiment with variations of the HMT network.

2.5.1 Network Setup and Initialization

In order to avoid overfitting due to the small number of sequences in both GEMEP and BRED, and especially in the facial branch which includes a large number of parameters, we pretrain the branch on the AffectNet Database [Mollahosseini et al., 2017]. The AffectNet Database contains more than 1 million images of faces collected from the internet and annotated with one of the following labels: Neutral, Happiness, Anger, Sadness, Disgust, Contempt, Fear, Surprise, None, Uncertain, and Non-face. The manually annotated images amount to 440k with about 295k falling into one of the emotion categories (neutral plus 7 emotions). The database also includes a validation set of 500 images for each class, while the test set is not yet released.

To prepare the facial branch for the subsequent feature extraction for our task, we start with a Resnet-50 CNN which has been trained using the ImageNet Database³. Next, in order to learn features that are pertinent to our task, we train again the network, this time on AffectNet by replacing the final FC layer of the network with a new FC layer with 8 output classes (the 7 emotions of AffectNet plus neutral). The network was trained for 20 epochs using a batch size of 128 and the Adam optimizer [Kingma and Ba, 2015], achieving the best accuracy on the AffectNet validation set at the 13th epoch (52.2%). As opposed to the facial branch, the body branch was not pretrained and its weights were initialized as in [LeCun et al., 2012].

For detecting, cropping, and aligning the face for each frame, we use the OpenFace 2 toolkit [Baltrusaitis et al., 2018]. We then use our pretrained facial branch to extract a 2048-dimensional feature vector which is used during training. This means that during training the parameters of the feature extraction layers of the facial branch remain fixed. Similarly, we extract the 2D pose of the subjects in each database (GEMEP and BRED) using OpenPose [Cao et al., 2017] along with the 2D hand keypoints [Simon et al., 2017]. In order to filter out badly detected keypoints, we set all keypoints with a confidence score lower than 0.1 as 0 for BRED and lower than 0.3 for the GEMEP database. These thresholds result in a percentage of approximately 70% valid joints in each database. The total size of the input vector for the body expression recognition branch is 134: 25 2D keypoints of the skeleton and 21 2D keypoints for each hand.

³These weights are provided by the PyTorch Framework. More information can be found in <https://pytorch.org/docs/stable/torchvision/models.html>.

	Label	y (6 classes)		y^f (7 classes)		y^b (7 classes)	
		F1	ACC	F1	ACC	F1	ACC
SEP	Body br.	0.30 (0.29)	0.35 (0.33)	-	-	0.34 (0.48)	0.37 (0.46)
	Face br.	0.60 (0.62)	0.65 (0.65)	0.54(0.61)	0.59 (0.63)	-	-
	Sum Fusion	0.62 (0.64)	0.65 (0.66)	-	-	-	-
	Joint-1L	0.66 (0.67)	0.67 (0.67)	-	-	-	-
HMT-3a	Body br.	0.30 (0.30)	0.34 (0.33)	-	-	0.32 (0.44)	0.36 (0.44)
	Face br.	0.58 (0.61)	0.65 (0.66)	0.53 (0.59)	0.60 (0.64)	-	-
	Fusion	0.67 (0.69)	0.69 (0.70)	-	-	-	-
HMT-3b	Body br.	0.29 (0.29)	0.33 (0.32)	-	-	0.35 (0.47)	0.38(0.46)
	Face br.	0.57 (0.60)	0.64 (0.66)	0.54 (0.59)	0.60 (0.65)	-	-
	Whole body br.	0.65 (0.67)	0.68 (0.69)	-	-	-	-
HMT-4	Body br.	0.30 (0.30)	0.34 (0.32)	-	-	0.32 (0.44)	0.36(0.44)
	Face br.	0.57 (0.60)	0.64 (0.66)	0.53 (0.59)	0.59 (0.64)	-	-
	Fusion	0.70 (0.71)	0.72 (0.72)	-	-	-	-

Table 2.4: Detailed results on the BRED database for various configurations of the HMT network. Numbers outside parentheses report balanced scores and inside parentheses unbalanced scores. The highest achieved scores when evaluating against whole body labels are shown in bold.

2.5.2 Exploratory Results on the GEMEP Database

The GEMEP database includes videos of 10 adult actors performing 17 different emotions: Admiration, Amusement, Anger, Anxiety, Contempt, Despair, Disgust, Fear, Interest, Irritation, Joy, Pleasure, Pride, Relief, Sadness, Surprise, and Tenderness. In this work we use the core set of the database that includes the first 12 of the aforementioned emotions.

We use 10-fold leave-one-subject-out cross-validation and repeat the process for 10 iterations, averaging the scores in the end. For all different evaluation setups, we train for 200 epochs, reducing the learning rate by a factor of 10 at 150 epochs. We report Top-1 accuracy for several experimental setups in Table 2.3. For the body expression recognition branch we compare three different implementations: a) the implementation with global temporal average pooling (GTAP) using a hidden FC layer of 256 neurons with ReLU activation, b) a temporal convolutional network (TCN) [Bai et al., 2018] with 8 temporal convolutional residual blocks, 128 channels and kernel size 2, and c) a bidirectional long short-term memory network (LSTM) [Hochreiter and Schmidhuber, 1997] with 100 hidden units and two layers preceded by an FC layer of 128 neurons with activation. For both TCN and LSTM we average the outputs over all time steps. In the first part of the table, we observe that GTAP (shown in bold) achieves the highest accuracy (0.34) although it’s a much simpler method. We believe that due to the small amount of data the methods focus only on certain representative postures that occur during the expression of emotions and ignore sequential information. As a result, the LSTM and TCN cannot outperform the DNN combined with GTAP, and would require a larger database in order to accurately capture temporal information. The face branch achieves a higher accuracy score (0.43) than the body branch (0.34), which is an expected result. Our main observation is the fact that the whole body emotion recognition branch (with the GTAP implementation) (shown in bold) achieves a significant improvement over the face branch baseline (an absolute 8% improvement, from 43% to 51%).

In Table 2.3 we also include experiments at the frame level, where we take only the middle frame of each video sequence and skip the temporal pooling structures in each branch. We observe that again the whole body emotion recognition branch (in bold)

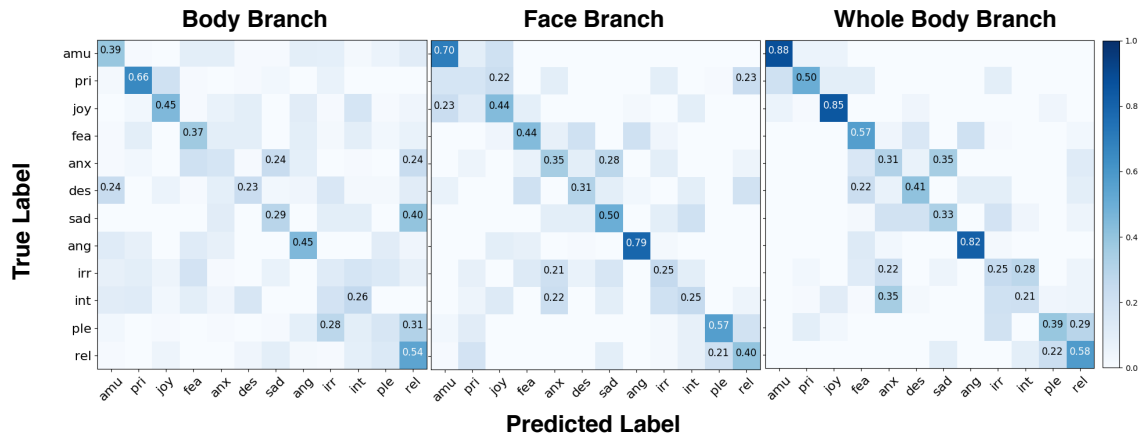


Figure 2.5: Confusion matrices for the face, body and whole body (feature fusion) branches of HMT in the GEMEP corpus.

yields a large performance boost over the facial branch (from 21% to 33%), as well as the significance of applying temporal pooling over all video frames.

Emotion specific details can be seen in the confusion matrices of Figure 2.5. We show the confusion matrices for the separately trained body, face, and whole-body branches. We can see that in cases such as pride, the body branch is much more efficient in recognizing the emotion, as opposed to the face branch, a result which is also in line with [Tracy and Robins, 2004]. In other emotions such as joy and anger, combination of face and body posture results in higher accuracy. There are also emotions for which the body branch fails to learn any patterns such as anxiety or pleasure. In these cases, the whole body branch achieves a lower accuracy than the face branch.

2.5.3 Results on the BabyRobot Emotion Database

For BRED we follow the exact same procedure as with the GEMEP database: training for 200 epochs, reducing the learning rate by a factor of 10 at 150 epochs, and 10-fold cross validation for 10 iterations. For the 10-fold cross validation, we ensure that each subject (30 in total) does not appear in both the training and test set of the same split. Because the database is highly unbalanced, especially for the body labels, we report results in balanced and unbalanced F1-score and accuracy. Due to this imbalance we also use a balanced cross entropy loss for \mathcal{L}^b , since the number of instances labeled as neutral are much larger than the emotion instances. We also note that for BRED, the annotations y^f and y^b include 7 classes (all emotions plus neutral), while the whole body annotation y includes 6 classes (all emotions).

We report our results in Table 2.4. The column labeled with y reports the metrics on the whole body labels, while columns y^f and y^b report results on the hierarchical face and body labels, respectively. For calculating the metrics of the face and body branches against y , we ignore the scores of the “neutral” label. Numbers outside parentheses report balanced scores and inside parentheses unbalanced scores. The highest achieved scores when evaluating against whole body labels are shown in bold.

Table 2.4 contains results of 5 different methods: *SEP* denotes independent training of the body and face branch using their corresponding labels. *Joint-1L* denotes training of the whole body emotion branch and only using the \mathcal{L}^w loss. *HMT-3a* denotes joint training of the hierarchical multi-label training network, if we omit the branch of the whole body emotion recognition, i.e., with the losses \mathcal{L}^d , \mathcal{L}^f , and \mathcal{L}^b . *HMT-3b* denotes joint training of the three losses: \mathcal{L}^b , \mathcal{L}^f , and \mathcal{L}^w , by omitting the final score fusion. Finally, *HMT-4*



Figure 2.6: Example results of whole body affect recognition. Captions on top of each image denote the final predictions while oval shapes denote predictions of the face branch and rectangle shapes denote predictions of the body branch. Green color shapes denote a correct prediction, whereas red color shapes denote an incorrect prediction. If the final predicted label is wrong then inside parenthesis we include the correct label.

denotes the joint training with all four losses of the HMT network. In the methods that include the fusion branch, we obtain the final prediction by the scores of the fusion s^d . In the case of HMT-3b, where we omit the final fusion, we obtain the final whole body label prediction by the whole body branch.

Our initial observation is the fact that the combination of body posture and facial expression results in a significant improvement over the facial expression baselines, for all different methods. Secondly, we see that HMT-4 achieves the highest scores for all metrics (0.70 balanced F1-score and 0.72 balanced accuracy), across all methods, as far as the whole body emotion label is concerned, while HMT-3a and HMT-3b exhibit similar performance (0.67 and 0.65 balanced F1-score, respectively) that is also comparable to the separate training of the body and face branches and their combination with post-process sum-based fusion (0.62).

We remind that y^f and y^b have one more class than y (neutral), which is why the scores appear lower for the face branch in the y^f column. This is not the case for the body branch, due to the fact that y^b and y are different by a lot more labels (99), while y^f and y differ in only 37 labels.

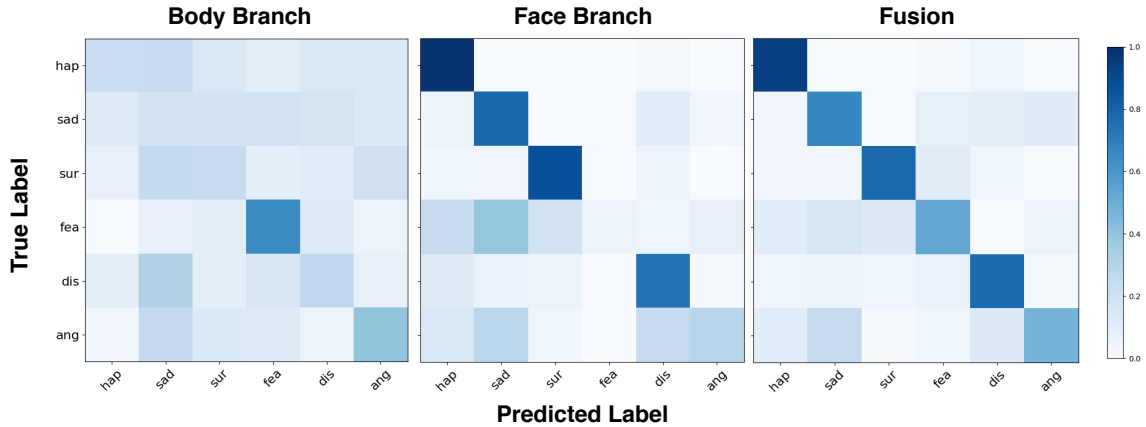


Figure 2.7: Confusion matrices of the body, face, and score fusion branches of HMT-4, against whole body labels y on the BRED database.

In Figure 2.6 we present several results (both correct and incorrect recognitions) of our method, while in Figure 2.7 we also depict the confusion matrices for the body, face, and fusion predictions when faced against the whole body labels y . We observe that generally, due to the fact that children in BRED relied more on facial expressions than bodily expressions (as it was observed in Table 2.1), only including the body branch in a system would result in low performance. We also observe that the face branch achieves low recognition rates for fear and anger. However, fusing the two using the HMT network results in a model that can reliably recognize all emotions.

2.6 Chapter Conclusions

In the first chapter of Part I, we proposed a method for automatic recognition of affect that combines whole body posture and facial expression cues in the context of CRI. CRI presents a challenging application that requires leveraging body posture for emotion recognition and cannot rely only on facial expressions. The proposed method can be trained both end-to-end, as well as individually, and leverages multiple hierarchical labels providing computational models that can be used jointly and individually.

We performed an extensive evaluation of the proposed method on the BabyRobot Emotion Database that features whole body emotional expressions of children during a CRI scenario. Our results show that fusion of body and facial expression cues can be used to significantly improve the emotion recognition baselines that are based only on facial expressions, and that 2D posture can be used with promising results for emotion recognition. We also show that hierarchical multi-label training can be exploited for improving system performance.

We believe our research shows promising results towards establishing body posture as a necessary direction for emotion recognition in human-robot interaction scenarios, and highlights the need for creating large-scale whole body emotional expression databases.

3

Multi-Modal Emotion Recognition using Audio and Facial Expressions in Children

3.1 Introduction

In Chapter 2 we highlighted the importance of body language in the automatic identification of emotion; an idea that we subsequently exploited for improving the performance of emotion recognition in child-robot interaction scenarios. In this chapter, we present an enhanced approach, which considers multiple modalities (visual plus audio) as well as visual representations (RGB and optical flow).

We mentioned before that building an emotion recognition system for children is challenging and presents many obstacles. Children not only differ from adults in their natural characteristics (e.g., voice pitch, body height) but also exhibit different behavioral patterns, which for example can result in abrupt movements and occlusions [Filtisis et al., 2019, Nojavanasghari et al., 2016]. To counter these, a robust system for child emotion recognition should leverage information from multiple modalities, exploiting the fact that different emotions can be expressed through different information channels. Further, recognition systems should be computationally efficient, especially in the context of real-life CRI scenarios. Finally, an additional challenge concerns the general lack of high-quality, large-scale children emotion datasets [Nojavanasghari et al., 2016] that are crucial in developing state-of-the-art deep learning supervised techniques. Children corpora tend to be of small size due to the fact that they are hard to obtain, one of the more important reasons being data sensitivity. We saw an example of this in Chapter 2 where a simple method of average pooling performed better than other established superior methods that leverage sequential data (LSTMs and TCNs), due to the lack of high amounts of data.

Going forward and building on the previous work, we present an audiovisual emotion recognition system aiming to address the aforementioned challenges. We investigate a different modality (audio) instead of the body skeleton, in order to tackle occlusions and increase robustness, and also use a different dataset (EmoReact), which includes both visual and aural expression of emotion, and is on a much larger scale than the BabyRobot Emotion Dataset (BRED). Furthermore, while the CNN architecture in that work considered each frame in the video separately, here we use CNN architectures that take into account multiple frames in the video using temporal sampling, as well as leverage the video dynamics using the optical flow representation. The system takes as input both the child's speech, as well as the visual channel, in the form of the raw RGB data stream, which can

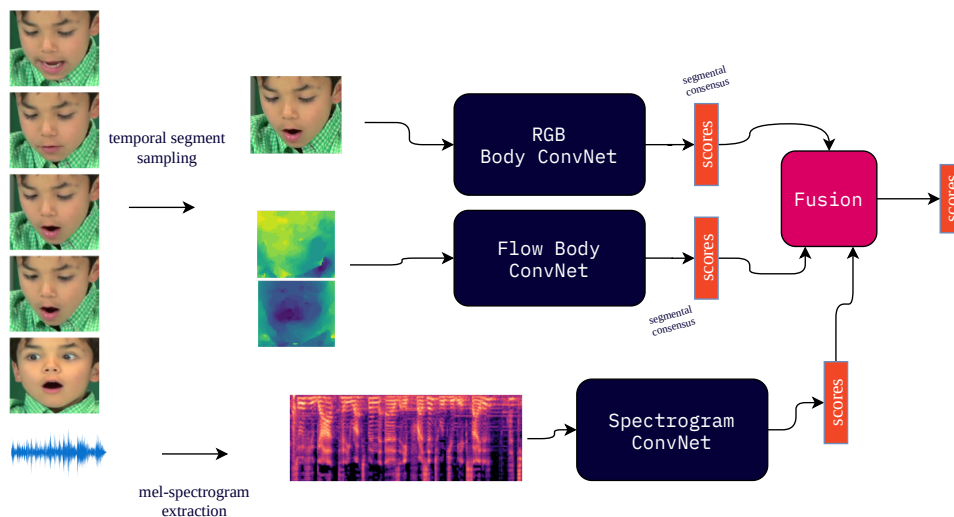


Figure 3.1: The proposed multimodal emotion recognition architecture for child-robot interaction.

be used to effectively identify static facial expressions and the optical Flow stream, which is effective in modeling the dynamics of emotions. This selection of different modalities is verified by ablation studies that analyze the contribution of each modality for the prediction of different emotions and identify the most effective fusion scheme that can be used to combine information from all channels. In addition, the deep learning based methods that we have employed allow for computationally efficient training and inference, and they can be developed on small datasets, avoiding overfitting. We perform extensive ablation studies on the EmoReact dataset, which, to the best of our knowledge, is the only dataset of children expressing emotions both verbally and visually, and establish a good trade-off between computational load and system performance. Finally, our approach is verified by comparing our system to the previous best published results on the EmoReact dataset, significantly outperforming them.

The rest of the chapter is organized as follows: Section 3.2 discusses related work on multi-modal emotion recognition for children and CRI. Section 3.3 describes in detail our proposed audiovisual emotion recognition architecture, and Section 3.4 presents our thorough experimental results on the EmoReact dataset. Finally, Section 3.5 provides our conclusions and directions for future work.

3.2 Related Work

A number of studies have emphasized the importance of leveraging multiple modalities for emotion recognition in adults [Bänziger et al., 2006, De Silva, 2004, Pantic et al., 2005]. In [Nguyen et al., 2017], 3D CNNs were used to extract deep spatiotemporal features from both the video and audio (represented as short-time Fourier transform) in order to determine emotion scores. Kim et al. [Kim et al., 2013] used deep belief networks for audiovisual feature generation, while [Tzirakis et al., 2017b] combined a two-branch feature extraction scheme with a long short-term memory network for continuous dimensional emotion recognition.

On the contrary, there is a lack of works studying multiple modalities for emotion recognition in children. Apart from the face, which is the most commonly used channel

for identifying emotion [De Gelder, 2009], there are other modalities equally powerful to reveal children affect such as speech and body movements. In [Nagarajan and Oruganti, 2019], an ensemble of AlexNet networks was applied on multiple spectrograms in order to extract deep features, which were then used by an SVM to identify emotions in the EmoReact dataset. For the same dataset, [Nojavanasghari et al., 2016] combined traditional audio features and features extracted from the OpenFace framework [Baltrusaitis et al., 2018] (action units, shape parameters, and head orientation) with an SVM for audiovisual emotion recognition.

3.3 Method

The architecture of the proposed emotion recognition system is shown in Figure 3.1. The system is composed of different branches, each one focusing on a different input channel/modality.

3.3.1 Visual Branch

The visual branch is based on the Temporal Segments Network (TSN) framework [Wang et al., 2016]. During training, the input video is split into K different segments of equal duration M , and in the next step, a snippet of length $N < M$ consecutive frames is randomly sampled from each segment, resulting in K snippets $T_k | k = 1, \dots, K$. Subsequently, each snippet is fed to a CNN, yielding class scores S_k for each snippet. In the last step, the scores of the different snippets are fused using the segmental consensus function H that is applied on the representations of all different snippets to obtain the final scores:

$$S = H(S_k |_{k=1\dots K}) = H(F_v(T_k; \mathbf{W}_v) |_{k=1\dots K}) \quad (3.1)$$

where $F_v(T_k; \mathbf{W}_v)$ denotes the application of a CNN F_v with parameters \mathbf{W}_v on the snippet T_k . The most common consensus function that can be used is averaging, while others include maximum or weighted averaging (we use simple averaging). The CNN is then trained using standard cross-entropy loss in the case of multiclass classification, or binary cross-entropy in the case of multilabel classification (which is the case of emotion recognition we consider).

Traditionally, TSNs take input from both the RGB of the input video, as well as the optical flow, with each one trained separately and then fused using average or weighted average fusion. As with TSNs for action recognition, we also use both modalities, since the optical flow can be used to model the dynamics that arise during expressions of emotion, while the RGB modality can best identify static expressions such as smiles. We also need to mention that we crop the input video (both RGB and Flow) around the child’s face, by using the facial landmarks obtained by OpenFace [Baltrusaitis et al., 2018].

The paradigm of TSNs offers several benefits to emotion recognition and CRI in particular. Considering an input video with a child expressing emotion, the archetype facial expressions and action units that correspond to each emotion are not present throughout the video, but usually only during a short period of it. As a result, temporal sampling allows the network to access several parts of the video and model its long-range temporal structure, thus being more likely to observe the corresponding facial expression. In addition, compared to processing the entire video, the sampling process ignores redundant information in consecutive video frames, helping avoid overfitting and offering a type of data augmentation, valuable for children emotion databases of small size.

Finally, since the ultimate goal of the system is its deployment in real-life CRI scenarios, it is important to consider the computational costs of training, as well as the ability to run in real-time. Due to the fact that the system does not consider the entire input video chunk, the computational load is reduced significantly, both at training, as well as during inference.

3.3.2 Audio Branch

In the audio branch, considering the input waveform of the video, we first extract its mel-spectrogram representation and then apply a CNN $F_a(\mathbf{W}_a)$ on it in order to extract the audio representation. Here, we bypass the cumbersome feature extraction methods by considering the mel-spectrogram of the waveform as an image, and applying standard computer vision techniques. Next, as with the visual modality, a fully connected layer is used in order to obtain the final emotion scores. The audio branch is susceptible to overfitting because the full spectrogram is fed to the network, contrary to the visual branch where temporal sampling is used. To counter this, we apply a more aggressive regularization scheme with high penalty for L2 regularization during training.

3.3.3 Training and Audiovisual Fusion

In order to fuse information from the visual and audio modalities, we consider two different types of fusion between both RGB-audio, as well as Flow-audio modalities: feature fusion and score fusion, and two training schemes: independent training and joint training.

During joint training, the RGB (or Flow) and audio CNN are trained concurrently, and depending of the fusion scheme, we either concatenate their feature vectors (feature fusion) before the last fully connected layer, or average the scores (score fusion) obtained after the last fully connected layer. In order to achieve feature fusion under joint training, we repeat the audio feature vector K times (where K is the number of segments/snippets), and associate each visual snippet with the audio feature vector for the whole video, through concatenation of the feature vectors. In contrast, in independent training the RGB (or Flow) and audio networks are trained separately, and we then average their emotion scores.

3.4 Experimental Framework and Results

3.4.1 The EmoReact Dataset

The dataset we use is the EmoReact dataset. The EmoReact dataset [Nojavanasghari et al., 2016] contains videos of 63 children (32F, 31M, aged 4 to 14) reactions to different topics, and has been collected from the YouTube channel React. The number of all videos across the training (432 videos), validation (303 videos), and test set (367) is 1102. Each video is annotated with one or more emotions, from a total of 8 emotion labels: Curiosity, Uncertainty, Excitement, Happiness, Surprise, Disgust, Fear, and Frustration. To the best of our knowledge, the EmoReact dataset is the only dataset of children expressing emotion, both verbally and visually.

3.4.2 Implementation Details

The CNN backbone of the visual and audio branches is a residual CNN with 50 layers (ResNet50) [He et al., 2016]. Specifically for the CNN of the visual RGB branch, we have pretrained it on the largest facial expression dataset, AffectNet [Mollahosseini et al.,



Figure 3.2: Example images from the EmoReact dataset.

Segments	ROC AUC		sec/train epoch	sec/val epoch
	Balanced	Unbalanced		
RGB				
1	0.685	0.773	11	7
3	0.713	0.786	27	20
5	0.709	0.787	40	26
10	0.715	0.788	73	51
Flow				
1	0.585	0.741	37	23
3	0.596	0.744	101	70
5	0.623	0.757	166	115
10	0.627	0.759	294	210

Table 3.1: ROC AUC and average time elapsed per epoch with varying number of sampled snippets.

2017], achieving 59.47% accuracy on the validation set (test set is not available). Because the label distribution of AffectNet is highly skewed, we employ balanced sampling so that the network sees the underrepresented classes more often. The residual networks of the audio branch and Flow modality are pretrained on ImageNet (we obtain the weights of the network as provided by the PyTorch framework).

We train all networks and modalities with stochastic gradient descent for 60 epochs, starting with a learning rate of $1e-2$, momentum 0.9, and regularization with weight decay (L2 regularization) $5e-4$. The learning rate is reduced by a factor of 10 at 20 and 40 epoch milestones¹. Training is done using binary cross-entropy loss. For evaluation, we select the epoch with the best validation area under receiver operating characteristic (ROC AUC), and apply the corresponding network on the test set, reporting class-balanced and unbalanced ROC AUC. Especially in the case of audio, we found out that a more aggressive regularization scheme is needed to avoid overfitting, and thus we increased the weight decay tenfold to $5e-3$.

Fusion	Training	ROC AUC	
		Balanced	Unbalanced
Single Modality	Audio	0.715	0.750
	Visual (RGB)	0.713	0.786
	Visual (Flow)	0.623	0.757
Score Fusion RGB-audio	Joint Training	0.720	0.756
	Independent Training	0.747	0.799
Score Fusion Flow-audio	Joint Training	0.719	0.746
	Independent Training	0.725	0.787
Feature Fusion RGB-audio	Joint Training	0.719	0.769
Feature Fusion Flow-audio	Joint Training	0.707	0.744

Table 3.2: Results on the EmoReact dataset for different fusion and training schemes between the RGB-audio and Flow-audio modalities.

3.4.3 Results

Number of segments As a first ablation study, we consider the number of segments (and as a consequence the number of snippets), which are used during training of the visual branch with the RGB and Flow modalities. We consider 4 different values: 1, 3, 5, and 10, and report in Table 3.1 the results on the ROC AUC (balanced per class and unbalanced), as well as average time taken per epoch for training and inference, on a computer with an RTX 2080 GPU.

We can see that in the case of RGB, increasing the number of segments above 3 does not result in significant performance difference, showing that even a small number of segments can achieve satisfactory performance. However, increasing the number of segments increases significantly both the training and inference times. For the Flow modality, we see that selecting 5 as a number of segments results in a balanced trade-off between performance and execution time, since the performance increase using 10 segments is minuscule. For the following experiments, we use 3 segments for RGB and 5 segments for the Flow modality.

Audiovisual fusion and training schemes Next, we experiment with the different kinds of fusion schemes that can be used to merge the RGB and audio, as well as the Flow and audio modalities: feature vs. score fusion, as well as the pretraining scheme: joint training of both networks vs. independent training. The results of this study are shown in Table 3.2. Training the networks independently and then averaging their scores achieves the best result in both cases of audiovisual fusion (RGB-audio and Flow-audio), when compared to both the single modalities, as well as their fusion using joint training. This could be attributed to the fact that while the TSN framework inherently avoids overfitting using the temporal sampling, in the case of audio this is not the case, since the full spectrogram is used, and more elaborate schemes of regularization are needed.

Emotion by modality Next, we explore the strengths and weaknesses of each different modality, by showing the different ROC AUC scores for each emotion, in Figure 3.3. We

¹We have made the code for the experiments publicly available at <https://github.com/filby89/multimodal-emotion-recognition>

	ROC AUC	
	Balanced	Unbalanced
Audio		
audio features + SVM [Nojavanasghari et al., 2016]	0.610	-
dnn ensemble + SVM [Nagarajan and Oruganti, 2019]	0.718	-
Ours (End-to-End)	0.715	0.750
Visual		
openface + SVM [Nojavanasghari et al., 2016]	0.620	-
Ours (Flow)	0.623	0.757
Ours (RGB)	0.713	0.786
AudioVisual		
[Nojavanasghari et al., 2016]	0.640	-
Ours (RGB+Audio+Flow)	0.754	0.809

Table 3.3: Final ROC AUC results on the EmoReact dataset.

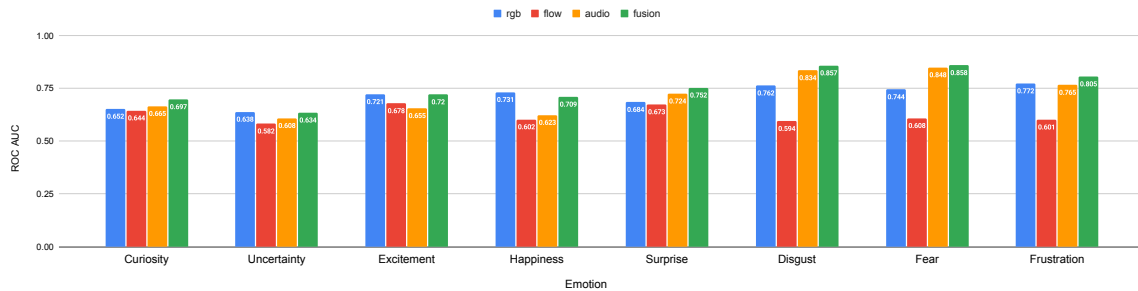


Figure 3.3: ROC AUC per emotion, for each different modality and their average score fusion.

observe that especially for Happiness, RGB is the most appropriate modality, while Fear and Disgust, are best identified through the children’s speech. Flow, in almost all cases underperforms when compared to the other modalities, however in the case of Excitement and Surprise it achieves a high score, which can be explained by the more intense movements a person does when expressing these emotions. The figure also shows the result of average score fusion using independent training for all three modalities, RGB, Flow, and audio. We can see that overall, fusion increases the total balanced and unbalanced scores, however in the case of Uncertainty, Excitement, and Happiness, it results in slightly lower score when compared to RGB only.

Final Results We present the final results of the emotion recognition system on EmoReact in Table 3.3, where we have also added the result of average score fusion between the three different modalities (using independent training), as well as the previous reported best results in the literature. For the audio modality, our architecture achieves significantly better ROC AUC than [Nojavanasghari et al., 2016], which used a carefully selected speech features set with an SVM, as well as similar results with Nagarajan et al. [Nagarajan and Oruganti, 2019]. However, our approach is end-to-end and simple to implement, while Nagarajan et al. employed an elaborate scheme involving multiple AlexNet architectures for feature extraction and an SVM on top of them to achieve the final result.

In the visual modality, our RGB TSN architecture improves significantly upon the best previous published result, which used features extracted from the OpenFace framework with an SVM [Nojavanasghari et al., 2016].

Finally, our audiovisual fusion scheme using all three modalities with independent training further increases the ROC-AUC up to 0.754, resulting in significant score improvement upon all previous studies.

3.5 Chapter Conclusions

In this chapter we proposed a novel multimodal emotion recognition system that can be used for deducing the emotion of children, with the ultimate goal being child-robot interaction scenarios. To that end, we have used deep learning methods that tackle challenges met in CRI: small datasets, real-time inference, and computationally low-cost training. We have also thoroughly explored several aspects of our architecture and identified the contribution of different parts of our network to the final outcome. We have evaluated the emotion recognition system on the EmoReact dataset of children expressing their emotions multimodally, and showed that it achieves high performance and state-of-the art results.

4

Using Body, Context, and Emotional Label Embeddings for Emotion Recognition in the Wild

4.1 Introduction

In the previous chapters, we identified several modalities and streams of auxiliary information that can be used to increase the performance of automatic emotion recognition systems, with a major focus on child-robot interaction. In this chapter we present an approach which builds on our previous results, this time for emotion recognition in the wild. The method described here won the first place in the First International Workshop on Bodily Expressed Emotion Understanding (BEEU) challenge, organized in conjunction with the ECCV2020 Conference.

4.2 Related Work

Apart from the body language and facial expressions, another auxiliary stream of information that can help in identifying emotions is the context and the surrounding environment of the person [Kosti et al., 2017b, Mittal et al., 2020]. It is apparent that both the place, as well as objects and other humans can influence a person’s emotions. Furthermore, as we also saw in Chapter 3, inherently, emotion recognition is a multi-label problem - the subject might be feeling two or more emotions. This is true, especially when considering an extended set of emotions, as in [Luo et al., 2020]. The emotions in extended sets do not have the same “semantic” distance between them. For example, anger is more close to annoyance than to happiness. Considering that previous works in literature have showed the superiority of methods that attempt to learn a joint embedding space that contains both word embeddings and visual representations [Dong et al., 2016, Frome et al., 2013, Ren et al., 2017], we believe that trying to attach a semantic meaning to the extracted visual feature is a natural way forward.

Regarding the context modality, Kosti et al. [Kosti et al., 2017b] introduced a large scale dataset for emotion recognition (EMOTIC) in different contexts (e.g., other people, places, or objects) and a convolutional neural network (CNN) based two-stream architecture that focused on the body and context of the subjects. The CAER video dataset for context-based emotion recognition was presented in [Lee et al., 2019], along with a two-stream architecture which employed adaptive-fusion to merge the two streams. In [Mittal et al., 2020], Mittal et al. designed a deep architecture with several branches, focusing

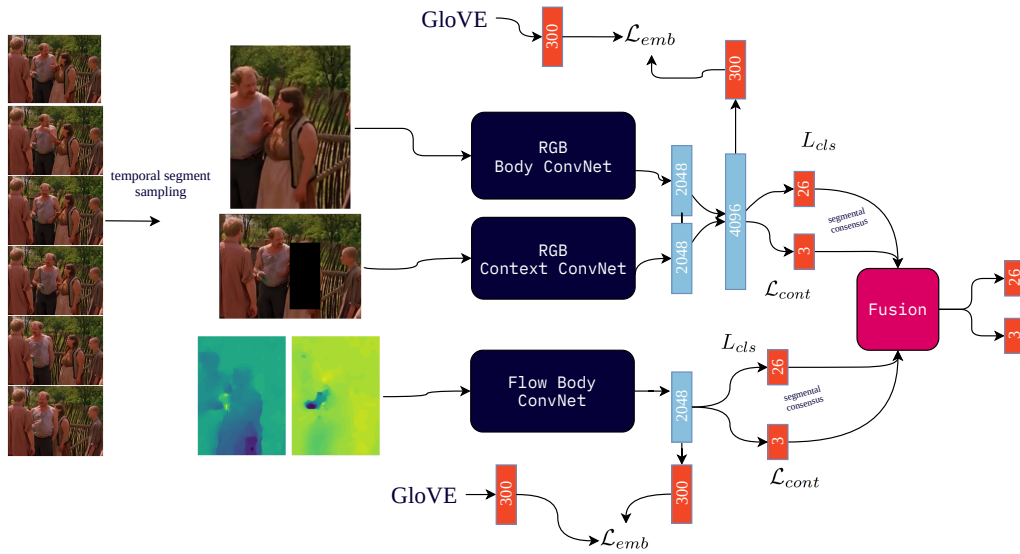


Figure 4.1: TSN with two RGB spatial streams (body and context) and one optical flow stream. The final results are obtained using average score fusion.

on different interpretations of the surrounding context (e.g., environment and interaction context) to significantly increase resulting predictions in the EMOTIC dataset.

Finally, some recent works have also focused on extracting visual representations from images that present the semantic relations found in embeddings built from words. The DeVISE embedding model [Frome et al., 2013] extracted semantically-meaningful visual representations by introducing a similarity loss between the feature vector extracted from a CNN and the word embedding from a skip-gram text model. Using a similar method, Wei et al. [Wei et al., 2020] built joint text and visual embeddings as emotion representation from web images, and in [Yeh and Li, 2020], Ye and Li built semantic embeddings for a multi-label classification problem.

The proposed method combines Temporal Segment Networks (TSNs) [Wang et al., 2016] focusing on the body, using the context in each video as an additional stream, and also uses an extra visual-semantic embedding loss, based on GloVe (Global Vectors) [Pennington et al., 2014] word embedding representations. Our experiments in the validation set verify the better performance of our method compared to the traditional TSNs, while our emotion recognition score on the test set was 0.26235, winning first place in the First International Workshop on Bodily Expressed Emotion Understanding (BEEU) challenge.

4.3 Model Architecture

Similarly with Chapter 3, we based our model on the TSN architecture [Wang et al., 2016], which has been widely used in action recognition and can be seen in Fig. 4.1.

Context Recognizing the significance of the context in emotion, we introduce one additional stream based on the context-environment surrounding the annotated human. For the RGB modality, we input the context in the network in the same way as in [Mittal et al., 2020], by masking out the instance body (we set all pixels to 0). We call this stream RGB-c, and the body streams RGB-b and Flow-b. During training, the RGB-b



Figure 4.2: RGB and Flow body and context

and RGB-c streams are combined at the feature level (RGB-bc) and are trained jointly while the Flow-b TSN is trained independently. Figure 4.2 shows an example of all 4 different streams.

Embedding Loss Our second extension is the introduction of an embedding loss on the feature vector extracted by the Convolutional Neural Network (ConvNet). This is done to exploit the fact that some emotions are closer semantically to others. This is also revealed by examining the correlation matrix of the dataset labels in [Luo et al., 2020], where some labels occur more frequently in combination with others (e.g. Happiness and Pleasure, Annoyance and Anger, etc.). Due to this result, we try to attach a semantic meaning to the feature vector extracted by the backbone image network.

To implement this, we first obtain for each one of the 26 categorical labels of BoLD their 300-dimensional GloVE word embedding [Pennington et al., 2014]. A PCA-projection of the 26 embeddings is shown in Fig. 4.3, where it is apparent that the distances between embeddings are indicative of their “semantic” distance. We then use a fully connected layer to map the feature extracted from the image to a 300-dimensional space and introduce the following mean-squared based loss:

$$\mathcal{L}_{emb} = \left\| \mathbf{W} f_v(\mathbf{x}) - \frac{1}{|K|} \sum_{y \in K} f_w(\mathbf{y}) \right\|_2 \quad (4.1)$$

where $f_v(\mathbf{x})$ is the feature vector extracted by applying the convNet on the image \mathbf{x} , \mathbf{W} is a matrix representing linear transformation from the space of the feature vector to the word embedding space, $f_w(\mathbf{y})$ is the word embedding of the label y , and K is the set of all positive labels for the image \mathbf{x} . That is, we try to reduce the Euclidean distance between the projected image feature and the arithmetic mean of the GloVE embeddings of the positive labels for image/video.

Predictions: Finally, after extracting for each sampled image its feature vector, we use two fully connected layers, one to classify to the 26 different categorical labels, and one to regress over the 3 different categorical emotions. The two TSNs are trained using the following loss:

$$\mathcal{L} = \mathcal{L}_{cls_1} + \mathcal{L}_{cls_2} + \mathcal{L}_{cont} + \mathcal{L}_{emb} \quad (4.2)$$

Specifically, since the dataset does not provide explicitly the multilabel targets, but the crowdsourced scores between 0 and 1, we include two different losses for the classification part: \mathcal{L}_{cls_1} that is the binary cross-entropy between the predicted scores and the multilabel

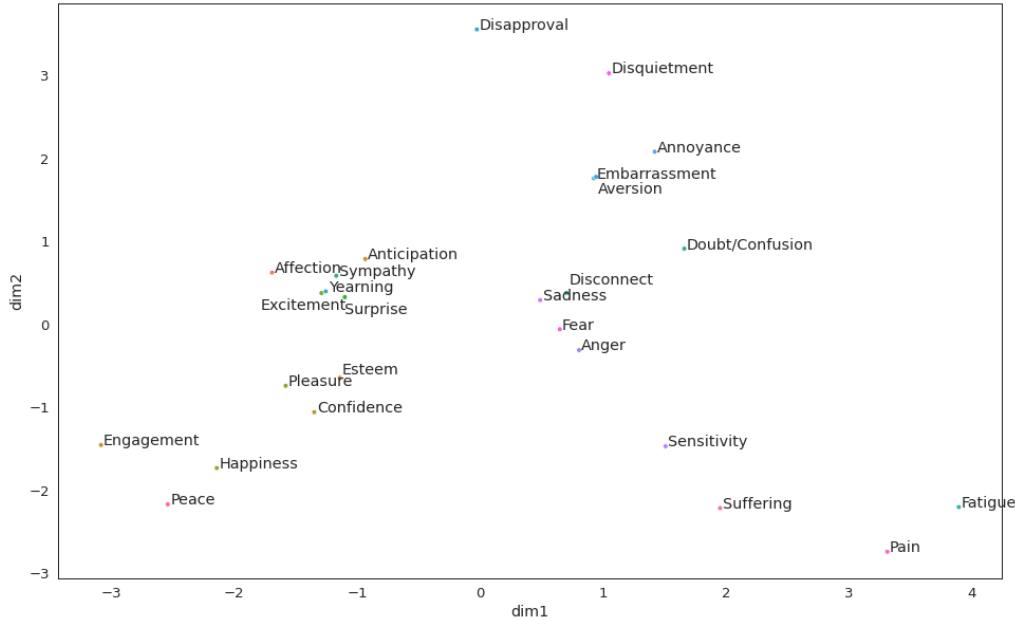


Figure 4.3: PCA projection of the categorical emotions GloVe word embeddings.

target (obtained after thresholding the multilabel scores at 0.5) and \mathcal{L}_{cls_2} that is the mean squared error between the predicted scores and the multilabel scores. We empirically found that the inclusion of \mathcal{L}_{cls_2} slightly boosted performance. For the regression part, \mathcal{L}_{cont} is the mean-squared error between the regressed values and the continuous emotions. Finally \mathcal{L}_{emb} is as in (4.1).

4.4 Dataset

The dataset used in the is the BoLD (Body Language Dataset) corpus [Luo et al., 2020] consisting of 9,876 video clips of humans expressing emotion, primarily through body movements. Each clip can contain multiple characters, yielding a total of 13,239 annotations, split into a training, validation, and test set. The dataset has been annotated by crowdsourcing employing two widely accepted categorizations of emotion. The first one is the categorical annotation with a total of 26 labels first used in [Kosti et al., 2017b], by collecting and processing an extensive affective vocabulary. The second annotation regards the continuous emotional dimensions of the VAD (Valence - Arousal - Dominance) Emotional State Model [Russell and Mehrabian, 1977]. The methods are evaluated using the following Emotion Recognition Score (ERS):

$$ERS = \frac{1}{2} \left(mR^2 + \frac{1}{2}(mAP + mRA) \right) \tag{4.3}$$

where mR^2 is the mean coefficient of determination (R^2) score for the three dimensional emotions (VAD), and mAP and mRA is the mean Average Precision and the mean area under receiver operating characteristic curve (ROC AUC) of the multilabel categorical predictions.

	Model	mAP	mRA	mR^2	ERS
without \mathcal{L}_{emb}	RGB-b	0.1567	0.6140	0.0538	0.21955
	Flow-b	0.1444	0.5914	0.0507	0.2093
	RGB-b + Flow-b	0.1623	0.6307	0.078	0.2375
with \mathcal{L}_{emb}	RGB-b	0.1564	0.6143	0.0546	0.21997
	Flow-b	0.1465	0.5947	0.0579	0.2142
	RGB-b + Flow-b	0.1637	0.6327	0.0874	0.2428

Table 4.1: Ablation experiment by training with and without \mathcal{L}_{emb} .

set	Model	mAP	mRA	mR^2	ERS
valid	RGB-c	0.1395	0.5760	0.0365	0.1971
	RGB-bc	0.1566	0.6055	0.0675	0.2243
	RGB-bc + Flow-b	0.1656	0.6266	0.0917	0.2439
test	RGB-bc + Flow-b	0.1796	0.6416	0.1141	0.26235

Table 4.2: Results on the validation and test set of BoLD including the RGB context stream and \mathcal{L}_{emb} .

4.5 Experimental Results

We train each TSN for 50 epochs using Stochastic Gradient Descent (SGD), with initial learning rate 10^{-3} which drops by a factor of 10 at 20 epochs¹. The backbone networks used is a residual network (ResNet) with 101 layers for the body convNets and a ResNet with 50 layers for the context convNet. We use the default hyperparameters of TSNs: 3 segments, 1 frame from each segment for the RGB streams, and 5 frames from each segment for the optical flow stream. The consensus used for segment fusion is averaging. For each network, we select the epoch with the best validation ERS. We have also found experimentally that the partialBN (Batch Normalization) technique used in [Wang et al., 2016] gives a nontrivial boost to the performance of the network.

First, in Table 4.1 we present two ablation experiments regarding the addition of \mathcal{L}_{emb} . We can see that adding the embedding loss increases slightly the performance in the RGB-b stream, and gives a boost to the performance of the Flow-b stream.

Then, in Table 4.2 we present our experimental results on the validation set of BoLD including the RGB context stream. From the results we can see that including the context along with the body in the RGB modality boosts the validation ERS of the architecture. We also experimented with including the context in the Flow network, but this resulted in worse performance. Our final submission for the test set was the model with the best validation score (0.2439 employing RGB-bc + Flow-b), using 25 segments instead of 3. The results of the different metrics on the test set can also be seen in Table 4.2, while the final ERS is 0.26235, improving upon the previous best result of 0.2530[Luo et al., 2020].

¹PyTorch code available at <https://github.com/filby89/NTUA-BEEU-eccv2020>

4.6 Chapter Conclusions

In this chapter we presented a method for automatic recognition of emotion which leverages the body appearance as well as more implicit channels of information, the context, and the semantic distances between emotions. Our method was submitted at the BEEU challenge, winning first place. We extended the TSN framework to include a visual-semantic embedding loss, by utilizing GloVE word embeddings, and also included an additional context stream for the RGB modality. We verified the superiority of our extensions compared to the baseline on the validation set of the challenge, and submitted the best system which achieved 0.26235 Emotion Recognition Score on the BoLD test set, surpassing the previous best result of 0.2530.

Part II

Expression Synthesis

5

AudioVisual Speech Synthesis with Compound Emotions

In the second part of this thesis, we study the second aspect of human-robot interaction - the one where the machine has the active role of speaking and conveying emotions, while the human acts as the receiving end.

5.1 Introduction

Achieving a high degree of naturalness in HCI is highly correlated with the ability of the agent to express emotions. Agents capable of expressiveness are more believable and life-like, thus have a stronger appeal to their interlocutor [Bates, 1994]. In addition, expressive behavior itself contains important information [Ambady and Rosenthal, 1992] and affects the emotional state of the other party [Hatfield et al., 1993, Keltner and Haidt, 1999] and, consequently, its decision making [Schwarz, 2000, Bechara, 2004].

Speech synthesis does not include only the generation of a human-like voice; speech is multimodal in nature [McGurk and MacDonald, 1976, Ekman, 1984] and important information is included in the visual stream of information (i.e., the human face, and more generally the human head and its movements) along with the acoustic stream. It has been shown that the inclusion of a visual stream of information increases the intelligibility of speech, especially under noise, even when the face is a virtual talking head [Sumbly and Pollack, 1954, Ouni et al., 2007].

Strongly correlated with speech, emotion is conveyed multimodally [Darwin, 1871, Richard J. Davidson, 2002], so the agent must be capable of expressing emotion multimodally as well. It is also reasonable to assume that emotions should be expressed through all the information channels simultaneously; otherwise it is possible that the receptor of the signals might become confused in regards to the emotional state of the agent, since neurological studies have shown that the perception of the acoustic and the visual streams affect each other [Skipper et al., 2007]. Under the same assumption it is also desirable that the levels of expressiveness of both information channels are correlated - i.e., a full blown facial expression of anger is accompanied with a full blown vocal expression of anger.

Audio-Visual Text-To-Speech synthesis (AVTTS) explores the generation of audio-visual speech signals [Mattheyses and Verhelst, 2015] (i.e., a talking head), and video-realistic AVTTS, more specifically, explores the generation of talking heads that highly resemble a human being as if a camera was recording it. Although naturalness in video-realistic AVTTS systems has increased greatly, the addition of expressions has proven to be a challenging task [Anderson et al., 2013, Schröder, 2009] due to the large variability

they introduce, especially in extreme expressions such as expressions of anger or happiness, in both acoustic and visual modeling. This large complexity increases the probability of introducing artifacts in the generated face and causing the “uncanny valley” effect [Seyama and Nagayama, 2007, Mori et al., 2012].

5.1.1 Desired capabilities of expressive agents

When considering synthesis of expressive speech we would ideally desire that the agent is able to express itself in the same ways a human can, in order to achieve the maximum level of naturalness. Studies on the nature of emotions have claimed both that some emotions can be defined as the combinations of others [Plutchik, 2001, Plutchik and Kellerman, 1980, Plutchik, 1980], and that each emotion has different levels of intensity which are expressed with variations between each other [Ekman, 1984, Ekman et al., 1980, Hess et al., 1997]. We believe that these studies can be proven useful when considering the problem of the number of emotional states the agent must be able to express. If we consider that the expression of emotions can also be defined as combinations of other emotions, we could instruct the agent to express more complex expressions, and speech with intermediate style between different emotions, through mixtures of emotions that it “knows” how to perform. In the same manner, considering that expressions vary with the intensity of the emotion, if we see the different intensity levels of an emotion as a combination of the emotion with the neutral expression/voice, we can express emotions in different continuous intensity levels.

Closely correlated with the number of emotional states and intensity levels we would like to be able to synthesize, is the ability of an EAVTTS system to adapt to a target emotion, given some examples of it.

Considering the two desired abilities we just stated, we argue that Unit Selection (US) synthesis is not suitable for extensions to expressive speech. US synthesis is more natural than parametric synthesis (under the assumption of a large enough training set) because the imperfect reconstruction of speech from parameters is avoided, however, immense data needs to be collected for variations in speaking style [Black, 2003]. Parametric synthesis on the other hand, appears to be suitable for the task in hand due to its flexibility that arises from the statistical modeling process [Zen et al., 2009], which allows modification of voice characteristics and speaking style. The same conclusion holds for EAVTTS as well; US suffers from low flexibility in changing facial expressions, while a parametric model allows us to do so.

In this chapter, we tackle these two problems in the context of parametric HMM (hidden Markov model) audiovisual speech synthesis. We assess the level of emotional strength acquired by an HMM-based AVTTS system that has been adapted to another emotion, by directly comparing with an HMM-based AVTTS system that has been trained on the full corpus of the respective emotion. We also employ HMM model interpolation in order to generate audio-visual speech with different intensity levels of expressiveness and more complexity. To our knowledge, there has not yet been a study to show the results of HMM adaptation and interpolation for HMM-based AVTTS.

In addition, in this chapter we present the CVSP-EAV medium-sized corpus that features expressive audio-visual speech in three emotions: anger, happiness, and sadness, plus neutral reading style, from one speaker and for the Greek language.

5.2 Related Work

Audiovisual speech synthesis (AVTTS) can be divided in two distinguishable categories based on the way the talking head is synthesized [Bailly et al., 2003]. The first category includes 2D/3D graphical models typically constructed by a mesh of polygons and vertices, and the creation of facial expressions directly involves the movement of the mesh. Examples of model-based audiovisual speech synthesis include rule-based systems [Beskow, 1996, Le Goff and Benoît, 1996, Pelachaud et al., 1996] or more complex systems simulating the movement of muscles of the human face [Sifakis et al., 2006].

The second category of image-based AVTTS is driven by a training set of image sequences, where based on recordings of a speaking person, snapshots of the person speaking arbitrary utterances not found in the recordings are generated. This category of image-based or photorealistic AVTTS can be further divided in unit selection and parametric methods.

Unit selection methods typically use raw or slightly modified images (and waveforms) in the training set in order to construct the target audiovisual sequence [Cosatto et al., 2000, Mattheyses et al., 2008, Huang et al., 2002]. In these architectures, typically the cost that is minimized by the unit selection module involves both acoustic and visual features.

Parametric methods involve the training of a statistical model controlled by a small number of parameters that can be used to reconstruct photo-realistic frames. Examples of parametric visual models in this category involve Active Appearance Models [Cootes et al., 2001] and Eigenfaces [Turk and Pentland, 1991].

A first example of parametric video-realistic audio-visual speech synthesis which is based on HMMs and uses the same pipeline as HMM-based text-to-speech synthesis (TTS) [Zen et al., 2007a] is in [Tamura et al., 1999, Sako et al., 2000], where human lips recordings are modeled through a technique similar to eigenfaces (eigenlips in this case), and the eigenlips weights are added to the observation vector along with the acoustic parameters (mel-cepstral coefficients and fundamental frequency). [Ezzat et al., 2002] employs MMMs and combines them with a custom trajectory synthesis technique for generating video-realistic speech. In [Xie et al., 2014, Xie and Liu, 2007], a lower-face Active Appearance Model combined with HMMs (and their variations) is used to generate full-face audio-visual speech by employing Poisson image stitching [Pérez et al., 2003]. [Fan et al., 2015] also used a similar technique for full-face video-realistic generation and successfully employed a Bidirectional Long Short Term Memory network in order to predict the weights of a lower-face Active Appearance Model, from a small number of linguistic features. Hybrid methods where the selection of images from the corpus is driven by HMM modeling have also been proposed [Wang et al., 2010, Mattheyses et al., 2011].

In the field of expressive AVTTS (EAVTTS) with 2D/3D models, some direct approaches have included the modeling of human facial expressions with a set of parameters (most commonly FAPs - facial animation parameters [Pandzic and Forchheimer, 2003]), and then using these parameters to drive a graphics based 3D model [Wu et al., 2006, Deng et al., 2006, Li et al., 2016]. Image-based unit selection methods for generating video-realistic expressive speech have been presented in [Cao et al., 2005, Melenchón et al., 2009, Liu and Ostermann, 2011].

Parametric video-realistic EAVTTS has not seen many studies. An example of an expressive video-realistic talking head using Active Appearance Models and HMM modeling was presented in [Anderson et al., 2013, Wan et al., 2013], where AAM modeling of the face was also extended to alleviate local facial deformations such as blinking, and remove the facial pose. This system used cluster adaptive training of HMMs in order to

model expressions of different emotions, as well as generate combinations of emotions. In [Shaw and Theobald, 2016], modeling of emotional expressions is achieved using AAMs and Independent Component Analysis (ICA). Furthermore, ICA is employed to generate mixtures of expressions.

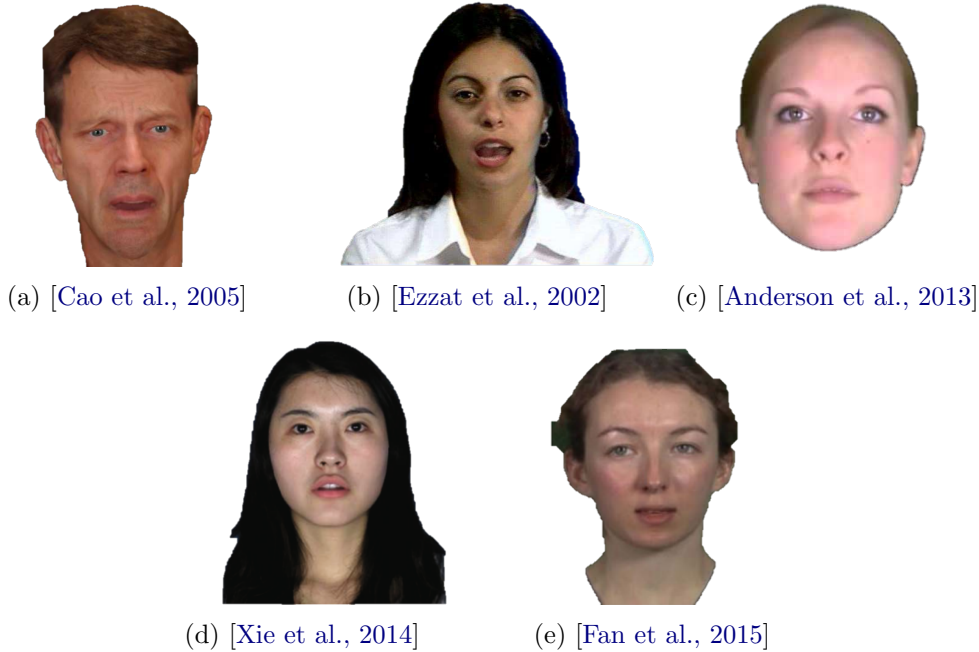


Figure 5.1: Various examples of video-realistic talking heads.

Figure 5.1 shows various examples of video-realistic talking heads from previous works.

5.3 Features for Parametric AudioVisual Speech Synthesis

5.3.1 Acoustic Features

Speech is modeled by mel-frequency cepstral coefficients (MFCCs), the logarithmic fundamental frequency, and band-aperiodicity coefficients using STRAIGHT analysis [Kawahara et al., 1999, Kawahara et al., 2001]. To reconstruct the waveform from the spectral and prosodic features, the STRAIGHT vocoder is used.

5.3.2 Visual Features

To obtain a low dimensional parametric model of the face for all of the different emotions, we employ an Active Appearance Model (AAM) [Cootes et al., 2001, Matthews and Baker, 2004]. We model the whole face and not only the lower part since emotional expressions include the upper facial half as well.

In active appearance modeling, a face (and more generally the object modeled) consists of the shape and the texture. The shape is represented by a vector \mathbf{s} , the elements of which are the coordinates of M vertices that make up the mesh of the face. For a particular snapshot (frame), the shape is expressed as the mean shape $\bar{\mathbf{s}}$ (that is the mean of the coordinates of the vertices of several frames after a Procrustes analysis is applied to them), plus a linear combination of n eigenvectors (called eigenshapes) \mathbf{s}_i that are found via employing a Principal Component Analysis (PCA) to the training meshes:

$$\mathbf{s} = \bar{\mathbf{s}} + \sum_{i=1}^n p_i \mathbf{s}_i \quad (5.1)$$

where p_i is the weight applied to the eigenshape \mathbf{s}_i .

The texture of the face is modeled in the same way as the shape, after normalizing the shape of each training mesh using an affine transformation or another method such as thin plate splines:

$$A(\mathbf{x}) = \bar{A}(\mathbf{x}) + \sum_{i=1}^n \lambda_i A_i(\mathbf{x}) \quad (5.2)$$

where $A(\mathbf{x})$ is the texture defined over all pixels \mathbf{x} that lie in the mesh of the mean shape $\bar{\mathbf{s}}$, $\bar{A}(\mathbf{x})$ is the mean texture, $A_i(\mathbf{x})$ are the eigenvectors found via PCA (called eigentextures) and λ_i is the weight applied to the eigentexture $A_i(\mathbf{x})$.

Upon obtaining the weights of the eigenshapes and the eigenvectors of a snapshot of the face, the image can be reconstructed by warping the texture $A(\mathbf{x})$ from the mean shape \mathbf{s}_0 to the computed shape \mathbf{s} based on a warp $\mathbf{W}(\mathbf{x}; \mathbf{p})$, where \mathbf{p} is the vector of the shape weights p_i .

During the fitting of an active appearance model to a new frame of the modeled object, if we denote as $I(\mathbf{x})$ the texture of the frame defined over the pixels \mathbf{x} , we seek to minimize the euclidean norm (called the reconstruction error):

$$E = |I(\mathbf{W}(\mathbf{x}; \mathbf{p})) - A(\mathbf{x})|_2^2 \quad (5.3)$$

where $I(\mathbf{W}(\mathbf{x}; \mathbf{p}))$ is the warped back image texture and $A(\mathbf{x})$ is the synthesized texture.

Building an AAM for a large database depicting different expressions (and extreme ones that arise during emotions such as happiness or anger) appears to be a challenging task. Due to the large variations introduced, minimization of the reconstruction error usually results in undesirable results. For this reason, we incorporate prior constraints [Papandreou and Maragos, 2008] during the fitting, as a means of increasing the robustness of the model.

In a model including prior constraints, the error minimized is:

$$E_p = |I(\mathbf{W}(\mathbf{x}; \mathbf{p})) - A(\mathbf{x})|_2^2 + Q(\mathbf{q}) \quad (5.4)$$

where $Q(\mathbf{q})$ is a quadratic penalty corresponding to a prior Gaussian distribution with mean \mathbf{q} . More information on the prior constraints can be found in [Papandreou and Maragos, 2008].

5.4 HMM-based expressive audio-visual speech synthesis

The HMM-based EAVTTS architecture models each emotion separately, through multiple HMM AVTTS systems. In this section we will do an overview of HMM-based audio-visual speech synthesis and then describe the methods of adaptation and interpolation that the system employs in order to adapt to new emotions, and mix known emotions.

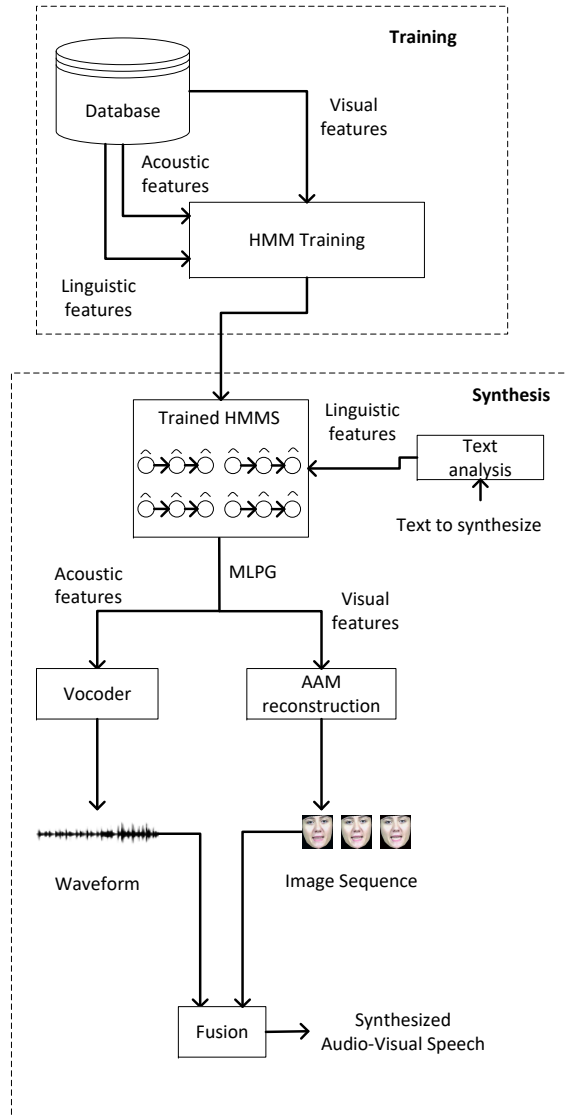


Figure 5.2: HMM-based audio-visual speech synthesis system architecture

5.4.1 Overview of HMM audio-visual speech synthesis

The architecture of an HMM-based audio-visual speech synthesis system is shown in Figure 5.2. Typically, the same pipeline with HMM-based acoustic speech synthesis [Tokuda et al., 2013] is employed.

Multi-Space probability Distribution Hidden Semi Markov Models (MSD-HSMMs) [Zen et al., 2007a, Heiga et al., 2007, Yoshimura et al., 1999] are used to model both the acoustic and visual features of speech simultaneously, so as to enforce a strong temporal alignment of the visual and acoustic streams.

The observation vector also contains dynamic features, in order to avoid discontinuities in that arise from the step-wise sequence that is generated in the synthesis stage.

To alleviate the problem of limited training data, a decision tree based context clustering method is applied [Odell, 1995]. During the decision tree clustering approach, using a predefined set of contextual questions, each node is split into two, by choosing the question that minimizes the Description Length [Shinoda and Watanabe, 1997] of the data. Upon

termination, states that belong in the same terminal node (leaf) of the tree are merged.

In the synthesis part, the maximum-likelihood parameter generation algorithm [Tokuda et al., 2000] is used to generate smooth trajectories of both the acoustic and visual parameters from the static and dynamic parameters emitted from the HSMMs. Just like in the DNN architectures, postfiltering in the cepstral domain [Yoshimura et al., 2005] is applied to acoustic features.

5.5 Adaptation for HMM-based EAVTTS

In order to tackle the data collection overhead that arises when considering expressive audio-visual speech, we use HMM adaptation in order to adapt an audio-visual HMM set that is trained a neutral training set, to a target emotion, using adaptation data that depict the emotion. The algorithm we employ is the CSMAPLR (Constrained Structural Maximum a Posteriori Linear Regression) adaptation method [Yamagishi et al., 2009, Yamagishi et al., 2007].

The CSMAPLR adaptation method combines the advantages of both SMAP [Shinoda and Lee, 2001] and CMLLR [Gales, 1998] methods, and makes use of linguistic information, through the regression tree that is used to propagate the prior information from the root of the tree, to the lower nodes. In addition, because the method is employing recursive MAP estimation [Digalakis and Neumeier, 1996], it is robust when a low amount of adaptation data is available [Lorenzo-Trueba et al., 2015]. After the CSMAPLR adaptation, an additional MAP adaptation is applied.

In the CSMAPLR adaptation, the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of a Gaussian state-output or state-duration distribution, are transformed simultaneously through the transformation matrix \mathbf{Z} and the transformation bias $\boldsymbol{\epsilon}$:

$$\bar{\boldsymbol{\mu}} = \mathbf{Z} \boldsymbol{\mu} + \boldsymbol{\epsilon} \quad (5.5)$$

$$\bar{\boldsymbol{\Sigma}} = \mathbf{Z} \boldsymbol{\Sigma} \mathbf{Z}^\top \quad (5.6)$$

5.6 Interpolation for HMM-based EAVTTS

Interpolation between emotions not only allows us to obtain emotions with various intensity levels and form more complex speaking styles and expressions, but also offers us the ability to control more formally the resulting expressions through the use of the interpolation weights, as opposed to adaptation methods.

In [Yoshimura et al., 2000] interpolation of HMM models trained on different datasets is done via maximization of the KL divergence between the output Gaussian distributions of the models. [Yamagishi et al., 2004] follow a simpler approach (interpolation between observations) for the interpolation of the HMM models, which we adopt as well for the interpolation of HMM sets trained on our four emotional training sets.

If we denote the output Gaussian distribution of an HMM state as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is the mean vector and covariance matrix of the distribution respectively, it becomes apparent that for HMM-based speech synthesis, the problem of interpolating emotions corresponds to interpolation of Gaussian distributions for respective HMM states across systems trained on a different training set which represent a speaking style.

The mean and variance of the interpolated pdf $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is:

$$\boldsymbol{\mu} = \sum_{i=1}^K \alpha_i \boldsymbol{\mu}_i \quad (5.7)$$

$$\boldsymbol{\Sigma} = \sum_{i=1}^K \alpha_i^2 \boldsymbol{\Sigma}_i \quad (5.8)$$

where K is the number of the different pdfs that will be interpolated and a_i is the weight corresponding to the i -th pdf. This interpolation is applied to the duration models as well. It is noted that the weights are chosen so that:

$$\sum_{i=1}^K |\alpha_i| = 1 \quad (5.9)$$

To deal with the fact that HMM sets trained on different datasets have a different tying structure, the interpolation of the pdfs is done on the synthesis level, after constructing a sentence HMM from each HMM set for the specific label to be synthesized. Figure 5.3 depicts this approach.

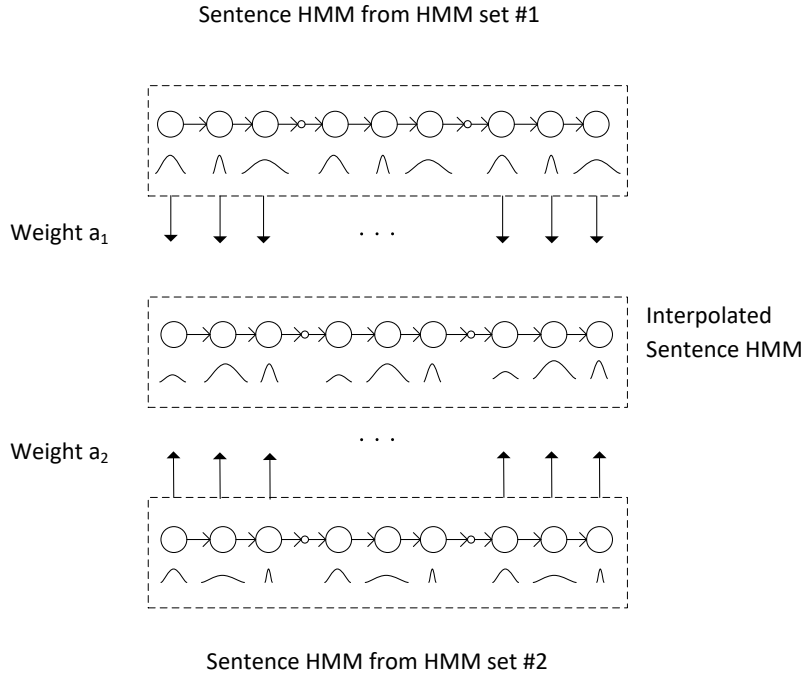


Figure 5.3: Interpolation for HMM-based EAVTTS.

5.7 The CVSP-Expressive Audio-Visual Speech Corpus

5.7.1 Recording of the Corpus

In order to evaluate the methods described in this study, we collected a medium sized corpus featuring expressive audio-visual speech in Greek for 4 emotions (neutral, anger, sadness and happiness). The database, which we call CVSP-EAV (CVSP - Expressive Audio-Visual) Corpus, was recorded in an anechoic studio at the Athena Innovation Research Center. A professional actress was hired to act the aforementioned emotions. The



Figure 5.4: Sample images for each of the different emotions present in the CVSP-EAV Corpus

actress was instructed to express the emotions in an extreme and clear manner. Although we are aware that humans rarely feel or express emotions in an extreme manner - and as such a talking head expressing extreme emotions would feel unrealistic in most HCI cases - since we are exploring modeling of emotional speech, it is reasonable to take into account the extreme case of each specific emotion. Furthermore, this allows us to correctly evaluate the synthesis of different intensity levels corresponding to each emotion which can range from reading style (neutral) to the full-blown expression of the emotion.

The actress was seated in front of a high-definition camera recording video in 1080p resolution at 29.97 frames per second in a H.264 format. A high quality microphone was used to capture the audio with a sampling rate of 44100 Hz.

The textual corpus consists of 900 sentences in Greek which were selected so that the corpus would have a phonetic distribution representative of the Greek language. Each sentence was pronounced by the actress 4 times, once for each of the four emotions: anger, sadness, happiness, and neutral, resulting in a total of 3600 sentences.

Figure 5.4 depicts a sample image from the recordings for each different emotion.

5.7.2 Processing of the Corpus

Because the recording of all the sentences in the corpus was continuous, in order to split the recorded video and audio in the different sentences the actress pronounced, we employed the sail-align toolkit [Katsamanis et al., 2011] which contains hidden Markov models (HMM) for the Greek language, trained in the Logotypografia database [Digalakis et al., 2003], featuring ~ 72 hours of speech from 125 speakers. The toolkit employs a three step speaker adaptation algorithm in order to increase the accuracy of the alignments. This alignment method was applied to the full recording of each of our four different emotions, and apart from obtaining the splits of the recordings at a sentence level, we also obtained four different speaker dependent HMM sets adapted to each of the different emotions in the corpus.

Next, the audio recorded from the high quality microphone was temporally aligned with the video using the cross-correlation between the high quality sound and the sound

	Neutral	Anger	Happiness	Sadness
Sentences	899	898	896	894
Duration (minutes)	72	71	72	86
Frames (approx.)	129,000	129,000	129,000	150,000

Table 5.1: Statistics of the post-processed CVSP-EAV corpus.

from the camera, and the frames from the video corresponding to each sentence were extracted in high quality JPEG format.

Finally, we force aligned at the phoneme level each sentence with its transcription, using the previously obtained adapted models for each different emotion. No further manual correction of the labels took place.

Due to recording and clipping problems, a few sentences from each emotion were discarded. Table 5.1 shows statistics on the post-processed corpus.¹

5.7.3 Feature Extraction

After processing the corpus, we resampled the audio at 16 kHz, and extracted 31 mel-frequency cepstral coefficients, the logarithmic fundamental frequency, and 25 band-aperiodicity coefficients with a frame shift of 5 ms using the STRAIGHT tool for MATLAB.

For the extraction of the eigenshape and eigentexture weights for each different frame in the database using the AAM modeling method described in Section 6.2, we used the following heuristic approach in order to minimize the fitting error of Equation 5.4.

1. We hand labeled a total of 981 frames with 61 facial landmarks as shown in Figure 5.7. From the 981 frames, 179 correspond to neutral expressions, 262 to angry expressions, 322 to happy expressions, and 218 to sad expressions.
2. For each different emotion in the corpus and its set of frames, we first use the face detection algorithm of [Mathias et al., 2014] and then use multivariate regression to map from the detected rectangle to an initial shape that is going to be used to obtain an initial estimate for the first frame of each sentence:

$$\mathbf{S} = \mathbf{AR} + \mathbf{b} \quad (5.10)$$

where \mathbf{S} is the shape corresponding to the detected rectangle \mathbf{R} , \mathbf{A} is the regression matrix and \mathbf{b} the intercepts. This results in four emotion-dependent regression models.

We also applied the same process to the whole corpus, in order to obtain an emotion-independent regression model. We found that generally the emotion-dependent regression models achieved a better shape initialization compared to the emotion-independent regression model

3. We then proceeded to build an AAM using all of the annotated frames.

From the found eigenshapes and eigentextures, we keep in both cases the vectors that account for 95% of the variation, a total of 9 eigenshapes and 58 eigentextures. Figures 5.5 and 5.6 show the mean shape and texture, as well as the first eigenshape and eigentexture and the variation they case on the former.

¹The CVSP-EAV corpus is available at <http://cvsp.cs.ntua.gr/research/eavtts/>.

- For each different sentence in the corpus, we get the best initial estimate of the shape of the first frame in a sentence, by using either an emotion-dependent regression or an emotion-independent one. We choose by comparing the reconstruction errors after fitting with each different estimates. For each of the subsequent frames in the sentence, we use the shape found previously as an initial estimate.

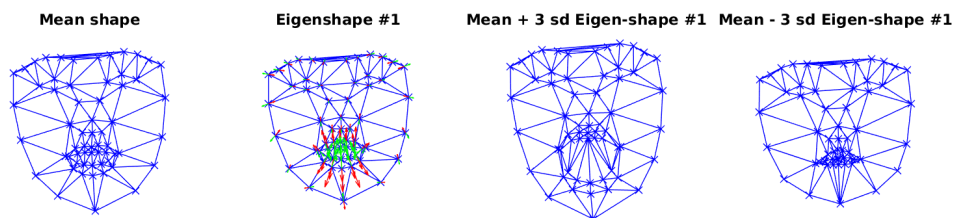


Figure 5.5: Example of the first eigenshape and the variations it causes to the mean shape between values $[-3\sqrt{\lambda_1}, +3\sqrt{\lambda_1}]$ where λ_1 is the corresponding weight, for an AAM trained on an expressive corpus.

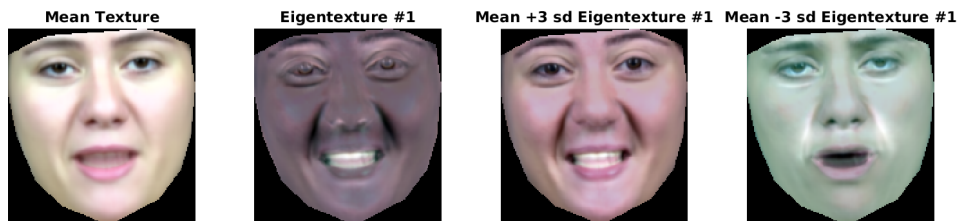


Figure 5.6: Example of the first eigentexture and the variations it causes to the mean texture between values $[-3\sqrt{\lambda_1}, +3\sqrt{\lambda_1}]$ where λ_1 is the corresponding weight, for an AAM trained on an expressive corpus.

In order to automatically exclude sentences where fitting presented artifacts, as much as possible, from the subsequent training of the systems, we discarded the sentences where the mean reconstruction error was above the threshold of 0.0018 and where more than 10 frames in the sentence had a reconstruction error of more than 0.0030. The threshold of 0.0018 was found heuristically to represent excellent fitting and reconstruction.

Table 5.2 contains the final mean reconstruction error resulting from fitting the Active Appearance Model for each different emotion and excluding the aforementioned sentences.

It is evident from both the mean reconstruction error as well as the number of sentences we had to discard, that the most difficult expression to fit is the expression of happiness which is expressed with strong variations in the human face. We would expect that the same would hold for the emotion of anger, however the number discarded sentences for anger is not on the same level as with happiness.

The final visual features extracted for each sentence, were resampled at 200 fps in order to match the previously extracted acoustic features.

In the end, in order to have a fair comparison across all emotions in the experiments, we kept for training all sentences that were common across all emotions, a total of 774 sentences.

For annotating the frames we used the am-tools software² and for building and fitting

²https://personalpages.manchester.ac.uk/staff/timothy.f.cootes/software/am_tools.doc/ind



Figure 5.7: Facial landmarks used for building the Active Appearance Model.

	Neutral	Anger	Happiness	Sadness
Mean Rec. Error	0.0013	0.0013	0.0015	0.0013
# discarded	5	11	93	6

Table 5.2: Fitting results in terms of mean reconstruction error for each of the emotions in the CVSP-EAV Corpus

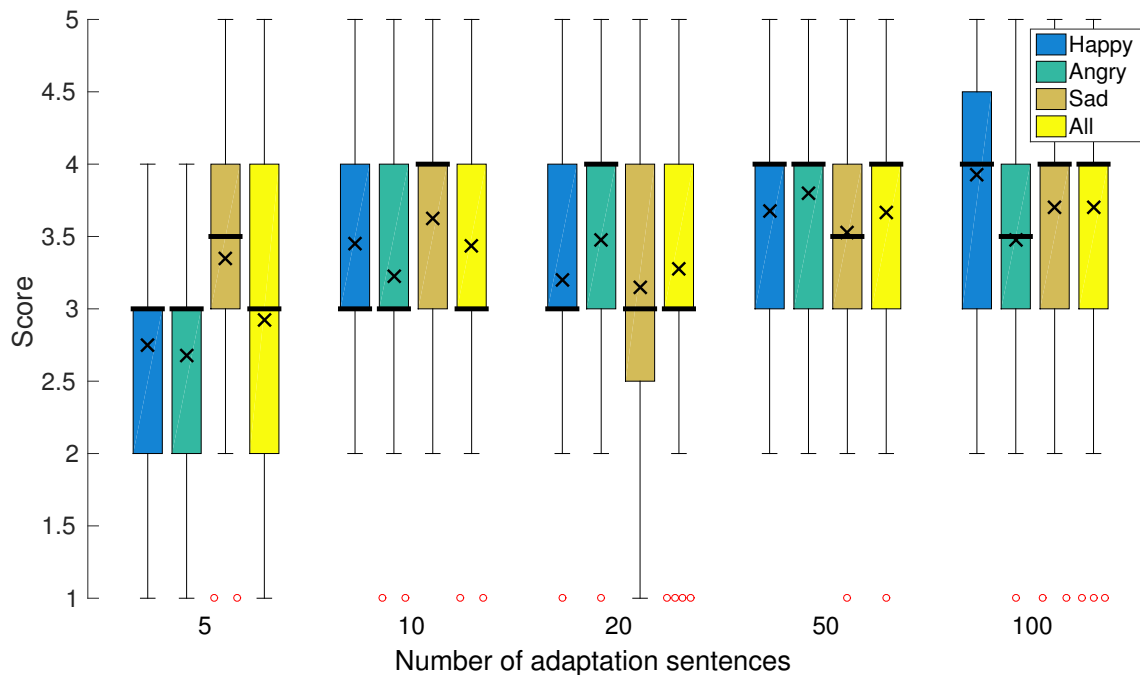


Figure 5.8: Subjective evaluation of the level of expressiveness captured by an adapted HMM audio-visual speech synthesis system for each different emotion (and total), and for a variable number of sentences. Bold line represents the median, x represents the mean, the boxes extend between the 1st and 3rd quartile, whiskers extend to the lowest and highest datum within 1.5 times the inter-quartile range of the 1st and 3rd quartile respectively, and outliers are represented with circles.

the Active Appearance Model we use the AAM-tools toolkit³ [Papandreou and Maragos, 2008].

5.8 Experimental Results

5.8.1 Evaluation Procedure

In order to assess the methods described in this chapter and the next, we designed and developed a web-based questionnaire containing multiple types of questions and tests which will be described in the following sections. Each questionnaire⁴ had a maximum of 102 random questions distributed to our different evaluations.

5.8.2 Evaluation of HMM Adaptation

To evaluate our method, we adapted an HMM-based AVTTS subsystem trained on the neutral emotion, to each of the other three emotions in the corpus, using the CSMAPLR adaptation described in 5.5, followed by a MAP adaptation. We also used a variable number of sentences for each of the above adaptations, namely 5, 10, 20, 50 and 100

ex.html

³<http://cvsp.cs.ntua.gr/software/AAMtools/>

⁴The questionnaire along with numerous videos of the talking head can be found at <http://cvsp.cs.ntua.gr/research/eavtts/>

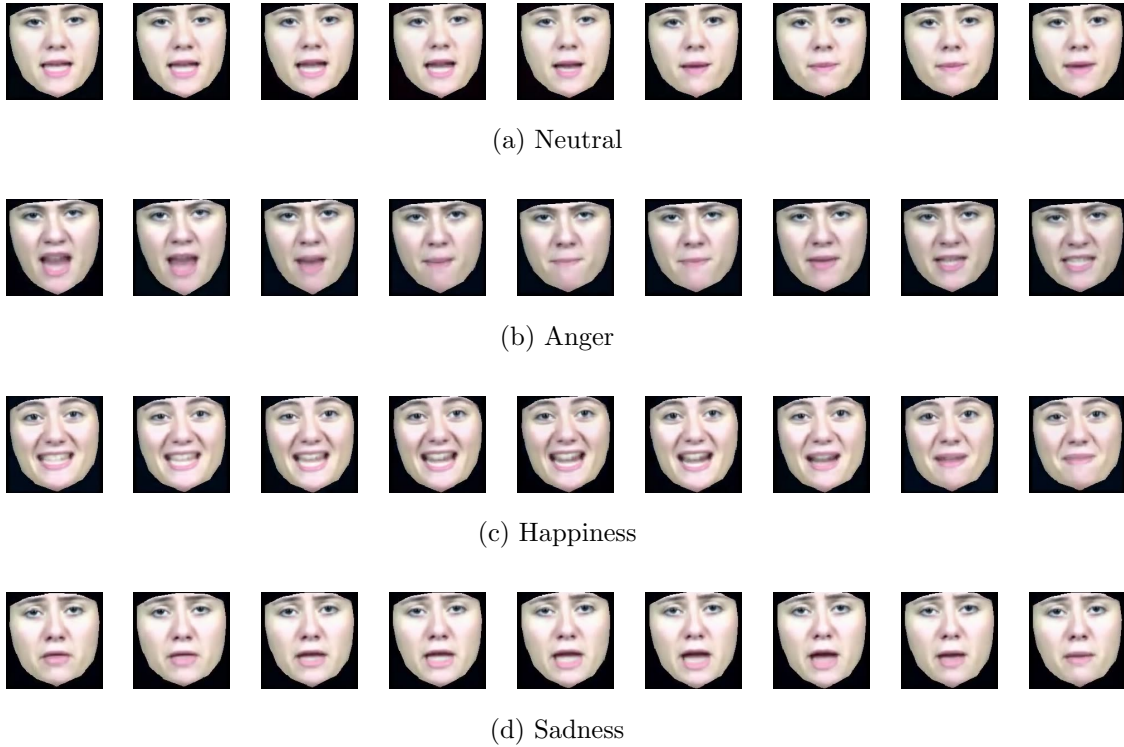


Figure 5.9: Results of audio-visual synthesis (consecutive frames from the same sentence) from a neutral HMM set (a) and its adaptation to the three emotions of (b) anger, (c) happiness, and (d) sadness, using 50 adaptation sentences.

	Neutral	Happiness	Anger	Sadness	Fear	Pride	Pity	Other
Neutral	100.0	0	0	0	0	0	0	0
Happiness	0	80	0	0	0	13.33	0	6.67
Anger	6.67	0	73.33	6.67	0	0	6.67	6.67
Sadness	6.67	0	0	80.0	6.67	0	6.67	0

Table 5.3: Classification of emotions in the emotion individual HMM systems (% scores).

sentences each time and for each of the different number of sentences, and emotions, we generated 8 unseen sentences from the test set.

In the questionnaire, each subject was presented with random videos for each different emotion (apart from neutral) and for each different number of adaptation sentences (a total of 15 videos), and were asked to evaluate the expressiveness of the talking head on an increasing scale of 1 to 5. We also included for each video, a video generated by the respective emotion individual HMM-based system built in Part 6.4.2 and advised the evaluators that this second video serves as a ground truth for the rating of 5, since we make the assumption that the adapted HMM system is capped, as far as expressiveness is concerned, by the corresponding emotion-independent HMM AVTTS system.

For each different emotion and number of adaptation sentences, 40 videos were evaluated (a total of 600 evaluations). Figure 5.8 shows the results of this subjective evaluation, for each different emotion, and for each different number of sentences used for adaptation.

We observe that the median value over all emotions increases as the number of sentences used for adaptation increase. We also observe that the emotion of sadness achieves even for 5 adaptation sentences a large score/median, compared to the other two emotions. This

(w_n, w_h)	Neutral	Happiness	Sadness	Pride	Disgust	Pity	Other
(0.1, 0.9)	0	93.33	0	6.67	0	0	0
(0.3, 0.7)	13.33	80	6.67	0	0	0	0
(0.5, 0.5)	53.33	40.00	0	6.67	0	0	0
(0.7, 0.3)	66.67	00.00	0	6.67	6.67	6.67	13.33
(0.9, 0.1)	86.67	0	0	0	0	0	13.33

Table 5.4: Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting the neutral emotion, and the second one trained on an emotional training set depicting happiness (% scores, w_n : Neutral Weight, w_h : Happiness Weight).

(w_n, w_a)	Neutral	Anger	Sadness	Pride	Disgust	Pity	Other
(0.1, 0.9)	13.33	66.67	0	6.67	6.67	0	6.67
(0.3, 0.7)	20.00	53.33	0	0	20	0	6.67
(0.5, 0.5)	46.67	33.33	0	6.67	13.33	0	0
(0.7, 0.3)	80.00	6.67	0	6.67	0	6.67	0
(0.9, 0.1)	86.67	0	6.67	6.67	0	0	0

Table 5.5: Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting the neutral emotion, and the second one trained on an emotional training set depicting anger (% scores, w_n : Neutral Weight, w_a : Anger Weight).

(w_n, w_h)	Neutral	Sadness	Anger	Fear	Pride	Disgust	Pity	Envy	Other
(0.1, 0.9)	13.33	73.33	0	6.67	0	0	6.67	0	0
(0.3, 0.7)	20.00	73.33	0	0	0	0	6.67	0	0
(0.5, 0.5)	60.00	20.00	0	0	0	13.33	6.67	0	0
(0.7, 0.3)	73.33	6.67	0	0	6.67	6.67	0	0	6.67
(0.9, 0.1)	73.33	6.67	6.67	0	0	6.67	0	6.67	0

Table 5.6: Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting the neutral emotion, and the second one trained on an emotional training set depicting sadness (% scores, w_n : Neutral Weight, w_s : Sadness Weight).

could be explained by the fact that neutral speaking style possesses a similar speaking rate to the sad speaking style, as opposed to happiness and anger, where speaking rate is generally faster. It is important to note that we observe a high degree of agreement between the evaluators, since in almost all cases the range of the boxes is only 1 point on the MOS scale. Our general consensus is that HMM adaptation can be successfully employed for HMM-based EAVTTS.

In Figure 5.9 we also show 10 consecutive frames from the same sentence, when adapting the neutral HMM set to one of the other three emotions using 50 sentences.

5.8.3 Evaluation of HMM Interpolation

Finally, our final evaluation was on the application of HMM interpolation to the emotion individual HMM-based EAVTTS systems built in the first part of this section. As preparation, for each of the 6 different HMM set pairs arising when combining the 4 different

(w_a, w_h)	Neutral	Happiness	Anger	Pride	Disgust	Envy	Other
(0.1, 0.9)	0	86.67	0	13.33	0	0	0
(0.3, 0.7)	6.67	80.00	0	6.67	0	6.67	0
(0.5, 0.5)	13.33	60.00	0	13.33	6.67	0	6.67
(0.7, 0.3)	40.00	0	33.33	6.67	6.67	6.67	6.67
(0.9, 0.1)	0.0	0	86.67	0.0	6.67	0	6.67

Table 5.7: Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting anger, and the second one trained on an emotional training set depicting happiness (% scores, w_a : Anger Weight, w_h : Happiness Weight).

(w_a, w_s)	Neutral	Anger	Sadness	Happiness	Fear	Disgust	Pity	Shame	Other
(0.1, 0.9)	6.67	0	80.00	6.67	0	6.67	0	0	0
(0.3, 0.7)	13.33	0	60	0	0	0	6.67	13.33	6.67
(0.5, 0.5)	33.33	33.33	13.33	0	6.67	0	6.67	0	6.67
(0.7, 0.3)	20.00	53.33	6.67	0	0	13.33	0	0	6.67
(0.9, 0.1)	6.67	66.67	0.0	0	0	20	0	0	6.67

Table 5.8: Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting anger, and the second one trained on an emotional training set depicting sadness (% scores, w_a : Anger Weight, w_s : Sadness Weight).

(w_s, w_h)	Neut	Hap	Sad	Anger	Fear	Pride	Disg	Pity	Shame	Envy	Other
(0.1, 0.9)	6.67	73.33	6.67	0	0	6.67	6.67	0	0	0	0
(0.3, 0.7)	6.67	26.67	26.67	0	0	0	6.67	0	0	6.67	20.0
(0.5, 0.5)	20.00	6.67	33.33	0	6.67	6.67	0	13.33	6.67	0	6.67
(0.7, 0.3)	13.33	0	60.00	6.67	6.67	0	0	6.67	0	0	6.67
(0.9, 0.1)	0.00	0	100.0	0	0	0	0	0	0	0	0

Table 5.9: Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting sadness, and the second one trained on an emotional training set depicting happiness (% scores, w_s : Sadness Weight, w_h : Happiness Weight).

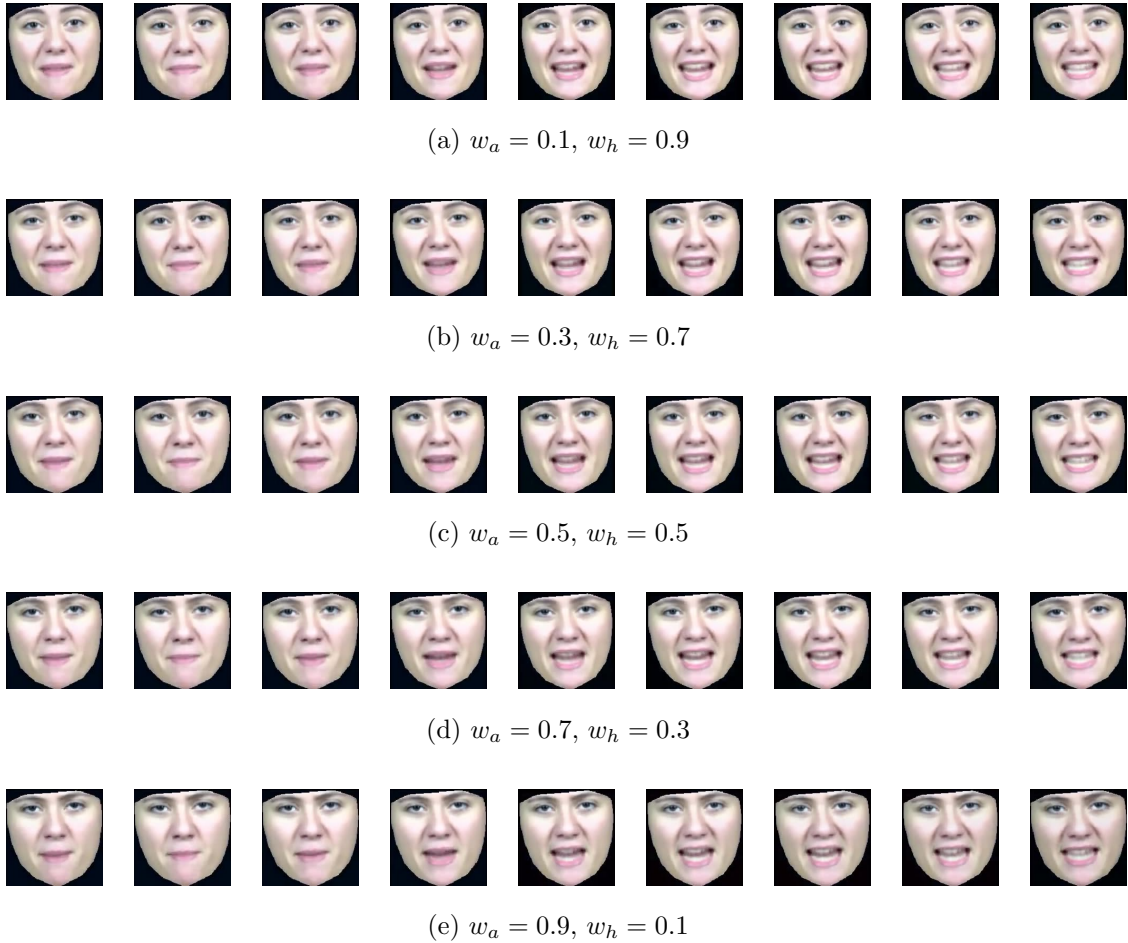


Figure 5.10: Results of audio-visual synthesis (consecutive frames from the same sentence) from interpolating HMM sets trained on anger and happiness (w_a : anger weight, w_h : happiness weight).

emotions of our corpus, we generated 6 unseen sentences from the test set, using 5 sets of interpolation weights: (0.9, 0.1), (0.7, 0.3), (0.5, 0.5), (0.3, 0.7), (0.1, 0.9). We also generated the same sentences by each emotion individual EAVTTS system.

Next, respondents were presented with the generated videos of the talking head, and were asked to recognize the emotion depicted by choosing from a list containing 11 emotions (neutral, happiness, anger, sadness, fear, pride, surprise, disgust, pity, shame, envy), plus the “other” option.

As a first evaluation, and to show that our respondents indeed recognized the emotion corresponding to each emotion independent system built in Part 6.4.2, in Table 5.3 we show the results of emotion recognition for the emotion independent systems, where we can see all emotions achieve a high classification rate, with the lowest being anger with 73.33%. We note that each different interpolation pair and emotion-independent system was evaluated 15 times.

Subsequently, we present 6 tables that show the emotion classification rate for each of the emotion pairs and interpolation pairs, in Tables 5.4 to 5.9 (in the tables we only include the emotions which were picked - that is classification rate was above zero at least in one row).

A study of Tables 5.4, 5.5 and 5.6, reveals that we can indeed achieve different levels of

the emotions by their interpolation with the neutral emotion since the emotion recognition results fluctuate mainly between the neutral emotion and the emotion under consideration in each figure. Abrupt changes in the classification scores suggest that we need to have an even smaller interpolation step, in order to control the resulting intensity level.

In Tables 5.7, 5.8 and 5.9 we can see the same trend. It is interesting to note, that the “neutral” emotion was also chosen many times. This result might suggest that the level of expressiveness at a weight of 0.5 is not strong enough, and when interpolated with another emotion at the same level the confusion causes the viewers to select the neutral stream. Several other options were also selected. We can see that for specific pairs, audiovisual speech with intermediate speaking style is generated (Anger-Sadness with respective weights (0.5, 0.5) and Sadness-Happiness with respective weights (0.3, 0.7)). We believe that a further study with more refined steps between the weights is imperative.

In Figure 5.10 we also show 10 consecutive frames from interpolating the HMM sets trained with the emotions of happiness and anger, for the weights we previously stated.

5.9 Chapter Conclusions

In this chapter, we tackled the problem of the data overhead that arises when one considers generation of expressive audio-visual speech, with the main focus being the visual modality. To that end, we aimed to equip an AVTTS system with two desired abilities: the ability to adapt to a target emotion with a small amount of adaptation data, and the ability to mix emotions in order generate audio-visual speech with multiple intensity levels and intermediate characteristics.

For the first ability, we employed HMM adaptation in order to adapt a “neutral” HMM-based AVTTS system, where visual modeling is done via an AAM, to the emotions of anger, happiness, and sadness, and showed that we can successfully adapt the expressions of the talking head, even with a small amount of adaptation data. Furthermore, the resulting expressiveness is also correlated with the nature of the target emotion that is being adapted. For the second ability, we employed HMM interpolation between two HMM-based AVTTS systems and showed that we can generate audio-visual speech with different intensity levels for an emotion and with intermediate characteristics between two emotions.

6

DNN-Based Expressive Speech Synthesis

6.1 Introduction

In this chapter, motivated by the recent advances in speech synthesis with deep learning [Ling et al., 2015], we propose two different deep neural network (DNN) pipelines for EAVTTS and examine the level at which these architectures are able to model the special characteristics and the full extent of expressive speech in an audio-visual context. This examination is achieved through direct comparison with the traditional parametric approach of HMM-based EAVTTS (presented in 5), as well as a concatenative unit selection EAVTTS system, both on the realism as well as on the expressiveness of the generated talking head, when the systems are trained on the CVSP-EAV corpus.

6.2 Method

In our two proposed DNN-based architectures for expressive audio-visual speech synthesis, each emotion is modeled separately, by a different DNN-based audio-visual synthesizer (we will use the term EAVTTS system when considering the full system that models all emotions, while the term AVTTS system denotes the subsystems). These sub-synthesizers (or subsystems) follow one of the two architectures that can be seen in Figures 6.1 and 6.2. Acoustic, visual, and linguistic features are extracted from an audio-visual corpus and then used in order to train the neural network subsystems of each architecture. We will first describe the features that we employ for audio-visual modeling, and then describe the two different DNN AVTTS models that are the components of the two different DNN-based EAVTTS architectures.

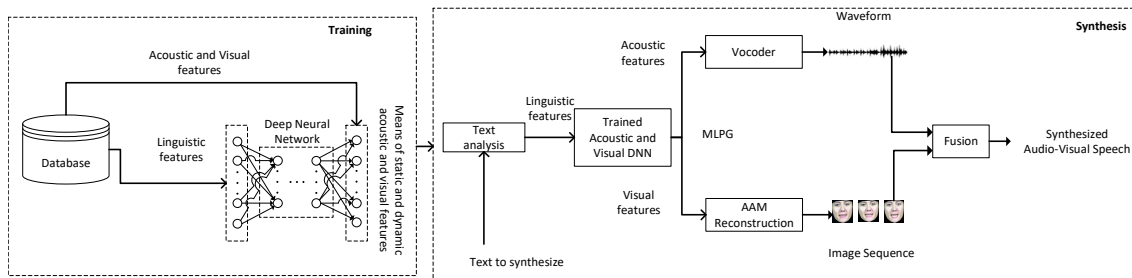


Figure 6.1: DNN-based audio-visual speech synthesis with joint modeling of acoustic and visual features.

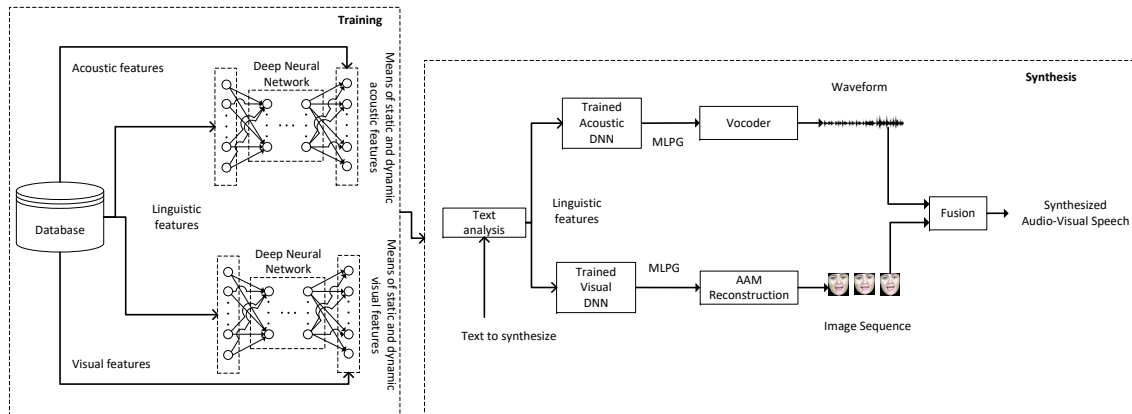


Figure 6.2: DNN-based audio-visual speech synthesis with separate modeling of acoustic and visual features.

The two DNN-based AVTTS synthesizers of each EAVTTS architecture differ in the fact that in Figure 6.1, the neural network maps linguistic features to acoustic and visual features at the same time, whilst in Figure 6.2, this mapping is done separately for the acoustic and visual features by two different neural networks. The linguistic features contain information about the lexicological context of the current phoneme and can consist of either answers to binary questions (e.g., “is the current phoneme a vowel?”) or numerical values (e.g., the number of syllables in a word). Within-phone positional features such as position of the acoustic/visual frame within the state of the phone (in an HMM context), phone and state duration, and state position within the phone [Zen et al., 2013, Wu et al., 2016] are also included. The output (acoustic, visual or joint audio-visual) features also include dynamic features (first and second derivatives of static features).

In both AVTTS synthesizers, a neural network (not shown in the figures) is employed for predicting the duration of speech. In the network responsible for duration modeling, an input vector containing frame-level linguistic features is mapped to durations of either the phoneme considered, or the phoneme states (in an HMM context).

During the training phase, linguistic features extracted from the database, along with acoustic and visual features, are used to train the networks. The mapping from linguistic features to acoustic, visual or joint audio-visual features constitutes a regression problem, and the following mean squared error function is minimized by the network with a Backpropagation procedure [Williams and Hinton, 1986]:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (6.1)$$

where N is the number of output features, \hat{y}_i is the i -th predicted feature and y_i is the real outcome.

In the synthesis stage, after analyzing the text to be synthesized and extracting its linguistic feature representation, the neural networks predict the output acoustic, visual, or joint audio-visual features. The outputs of the neural networks are considered to be the mean vector of the acoustic and visual features while the covariances are pre-calculated from the training data. The Maximum-Likelihood Parameter Generation algorithm [Tokuda et al., 2000] is then used in order to produce smooth trajectories of acoustic and visual features. This step is imperative in order to alleviate the fact that DNNs do not have memory or take into account adjacent frames during training [Zen,

2015]. Postfiltering in the cepstral domain [Yoshimura et al., 2005] is applied to acoustic features.

In general, DNN-based synthesis (both audio and audio-visual) possesses important advantageous properties as opposed to HMM-based synthesis [Zen et al., 2013, Watts et al., 2016, Qian et al., 2014]:

1. Deep layered architectures can represent highly complex function transformations compactly.
2. In DNN-based AVTTS, contrary to HMM-based AVTTS where predictions take place on a state level, predictions take place on an acoustic frame level.
3. Decision trees are incapable of modeling complex dependencies between input features whereas DNNs can compactly model these dependences. Furthermore, decision trees perform a hard split of the linguistic space which results in inferior generalization to DNN-based modeling where the weights of the network are trained using the whole training set.
4. Linguistic features can also hold numerical and not only binary values. [Zen et al., 2013] found out in experimental results that numerical values perform better and more efficiently.

6.3 Unit Selection audio-visual speech synthesis

In this section, we describe the unit selection video-realistic EAVTTS system, which employs multiple US AVTTS subsystems to model each different emotion. The subsystems are based on the unit selection acoustic speech system described in [Raptis et al., 2016, Chalamandaris et al., 2013] and were modified to include the visual modality as well. We did that in order to have a direct comparison of a concatenative EAVTTS system against our parametric systems. The system utilizes its own front end (as opposed to our previous methods) and its architecture is shown in Figure 6.3.

The subsystems follow a typical concatenative unit selection architecture, split into two components:

The NLP (natural language processing) component which is responsible for extracting all relevant information from the input text and transforming it into an intermediate format. This component comprises of the following modules: a word- and sentence- tokenization module, a text normalizer, a letter-to-sound module and a prosody generator.

The DSP (digital signal processing) component consists of the unit selection module, the signal manipulation module that generates the speech waveform, and the image reconstruction module, which is essentially the same module used in the previously described methods that reconstructs the image sequence and joins it with the speech waveform.

The unit selection module of the system optimizes a cost function that consists of two terms: the target cost, which is the cost of similarity of phonetic and prosodic context between two units, and the join cost, affected by pitch continuity, spectral similarity, and visual similarity. The visual similarity is integrated into the unit selection cost function as two additional terms in the join cost function; one for each of the visual feature vectors, namely the shape and texture feature vectors. The Euclidean norm is used as the distance between the shape and texture vectors. The modified join cost function is a weighted sum of the auditory components, i.e. the pitch and spectral cost functions, and the visual

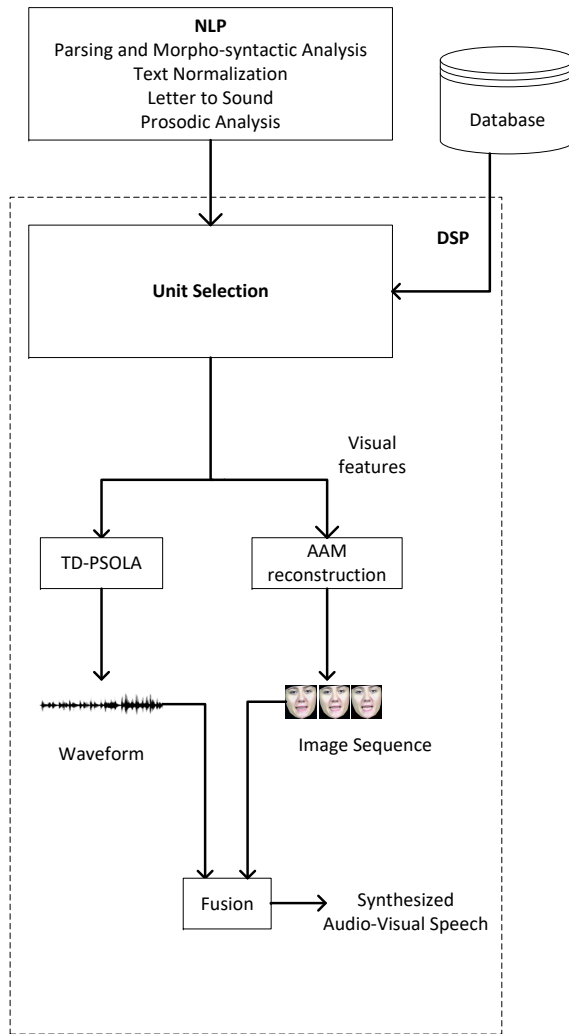


Figure 6.3: Unit selection based audio-visual speech synthesis.

components, i.e. shape and texture. The weights are chosen so that all components have equal range, i.e. assigning equal importance to both modalities¹.

The final waveform of speech is generated using a custom Time Domain Pitch Asynchronous Overlap Add (TD-PSOLA) method to concatenate the units selected by the unit selection module.

The final image sequence is generated, by concatenating the visual parameters (eigen-shape and eigentexture weights) that correspond to the audio-visual units selected by the unit selection module. During this concatenation we do not employ a smoothing technique.

6.4 Experimental Results

6.4.1 Evaluation Procedure

In order to assess the methods described in this paper we designed and developed a web-based questionnaire containing multiple types of questions and tests which will be

¹Fine-tuning the unit selection weights requires extensive listening/visual experiments which is outside the scope of this work.

described in the following sections. Each questionnaire² had a maximum of 102 random questions distributed to our different evaluations.

6.4.2 Evaluation of realism and expressiveness of the EAVTTS methods

Our evaluation of the EAVTTS methods described in this paper:

1. HMM-based EAVTTS (HMM)
2. DNN-based EAVTTS with joint modeling of acoustic and visual features (DNN-J)
3. DNN-based EAVTTS with separate modeling of acoustic and visual features (DNN-S)
4. Unit selection EAVTTS (US)

aims at comparing the methods both on the realism and expressiveness of the synthesized talking head. Furthermore, in order to gain more insight, we do not compare the methods only on the audio-visual realism, but also on the realism that is achieved by each different modality. “Realism” denotes the similarity of the talking head (or acoustic speech in case of the evaluation of the different modalities), to a human uttering the same sentence. This encapsulates both naturalness as well as intelligibility.

For each method, and for each of the four emotional training sets (neutral, anger, happiness, sadness) of the CVSP-EAV corpus we trained a subsystem (which we call from now on an emotion-independent AVTTS system). This means that, e.g., the full HMM-based EAVTTS system consists of 4 subsystems - one for each emotion. 48 test sentences, taken from the corpus, were generated from each subsystem. As a result $4 \times 48 = 192$ sentences were generated from each full EAVTTS method.

The HMM-based subsystems were built using the HTS toolkit [Zen et al., 2007b]. Five state, context-dependent MSD-HSMMs with left-to-right topology were trained and tied using a decision-tree clustering technique, using a similar set of questions with [Tokuda et al., 2002], but adapted for the Greek language. We use 29 different phonemes for the Greek language including silence.

Training of the DNN-based subsystems was implemented using the Merlin speech synthesis toolkit [Wu et al., 2016, Ronanki et al., 2016]. The input vector to the neural networks was broken down to 494 linguistic features from the almost ~ 1500 questions used for context clustering in the HMM-based systems, by exploiting the fact that non-binary linguistic features can be used as an input in the neural networks.

All networks (both networks that predict duration and networks that predict features) consisted of six hidden fully connected layers of 1024 neurons each. For training the networks with Backpropagation we use a batch size of 256 and a learning rate of 0.002. We train for a maximum of 25 epochs unless the error on the validation set (from the 774 sentences used to train each subsystem, we used 10 as a validation set) increases in more than 5 consecutive epochs after epoch 15. It is important to note at this stage that the architecture that employs separate modeling of acoustic and visual features, uses double the number of parameters than the architecture that employs joint modeling of acoustic and visual features.

The unit selection subsystems were built by modifying an existing unit selection acoustic speech synthesis system, as described in Section 6.3.

²The questionnaire along with numerous videos of the talking head can be found at <http://cvsp.cs.ntua.gr/research/eavtts/>

Table 6.1: Results (%) of subjective pairwise preference tests on audio-visual speech realism. Bold font indicates significant preference at $p < 0.01$ level.

DNN-S	DNN-J	HMM	US	N/P
25.0	22.22	-	-	52.78
51.11	-	15.56	-	33.33
75.56	-	-	18.89	5.55
-	43.33	22.22	-	34.44
-	72.22	-	22.78	5.0
-	-	63.89	27.78	8.33

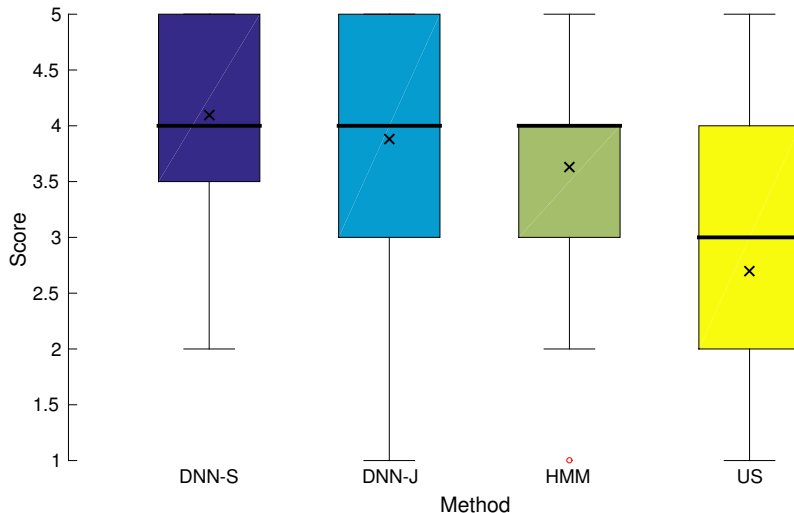


Figure 6.4: Boxplot of the MOS test results on the audio-visual realism of the different EAVTTS methods. Bold line represents the median, x represents the mean, the boxes extend between the 1st and 3rd quartile, whiskers extend to the lowest and highest datum within 1.5 times the inter-quartile range of the 1st and 3rd quartile respectively, and outliers are represented with circles.

Evaluation of audio-visual realism

To evaluate the realism of the talking head (both acoustic and visual) generated from each of the different methods respondents of the web-based questionnaire were presented with pairs of videos depicting the video-realistic talking head uttering the same sentence and in the same emotion, generated by two of the previously described four methods, and were asked to choose the most realistic video in terms of both acoustic and visual streams (with a “no preference” option available as well). The sentences were chosen randomly from the 192 sentences that were generated from each system. We also made sure that all emotions appear in the same rate. The result is a total 6 pairwise preference tests (for all different combinations of the 4 methods), with 180 pairs evaluated for each method pair (45 pairs for each emotion). The results of the preference tests are presented in Table 6.1.

Our statistical analysis of preference tests employs a sign test (ignoring ties), with Holm-Bonferroni correction over all statistical tests of this section - 30 in total.

From the table we can see that both DNN architectures are preferred significantly at the $p < 0.01$ level over the HMM and US methods, while HMM is also preferred significantly over US at the $p < 0.01$ level. Among the two DNN architectures we see that the preference scores are very close and there is not a significant difference.

Table 6.2: Significant differences between systems, on the audio-visual realism of the generated talking head, at levels $p < 0.05$ and $p < 0.01$. Blank cell denotes no significant different.

	DNN-S	DNN-J	HMM	US
DNN-S	-		$p < 0.01$	$p < 0.01$
DNN-J	-	-	$p < 0.05$	$p < 0.01$
HMM	-	-	-	$p < 0.01$
US	-	-	-	-

Table 6.3: Results (%) of subjective pairwise preference tests on visual speech realism. Bold font indicates significant preference at $p < 0.01$ level.

DNN-S	DNN-J	HMM	US	N/P
28.33	27.5	-	-	44.17
40.0	-	28.33	-	31.67
84.17	-	-	10.83	5.0
-	38.33	30.83	-	30.83
-	85.0	-	8.33	6.67
-	-	76.67	15.0	8.33

We generally observe a strong bias for the parametric approaches over the unit selection approach; a reasonable outcome considering that the size of each emotional training set is relatively low for unit selection synthesis combined with generation of unseen sentences.

A second evaluation of the audio-visual realism of the different methods was also performed, via a mean opinion score test (MOS). The respondents were presented with random videos of the talking head from each method and were asked to evaluate the realism of the talking head on a scale of 1 (poor realism) to 5 (perfect realism). Before the evaluation the respondents were also presented with samples from the original recordings and were instructed that they correspond to perfect realism. Each method was evaluated 200 times (50 for each emotion) and the results are shown in Figure 6.4.

To check for significant differences between the systems we perform pairwise Mann-Whitney U tests (with the same Holm-Bonferroni correction as before) due to the fact that Likert-type scales are inherently ordinal scales [Clark et al., 2007]. The results are shown in Table 6.2.

We can see that there is almost complete accordance of the results of the MOS test with the results obtained from the pairwise preference tests.

In Figure 6.5 we also show the MOS test results for each different emotion.

Evaluation of visual realism

Similarly with the evaluation of audio-visual realism, we conducted 6 more pairwise preference tests where respondents were presented with random pairs of muted videos and were asked to pick the most realistic video (with a “no preference” option available). Each method pair was evaluated 120 times (30 pairs for each emotion), and the results are presented in Table 6.3.

From the table we can see that statistically significant differences occur only between parametric approaches versus the unit selection one. The DNN architectures seem again to be preferred over HMM, however the result is not statistically significant.

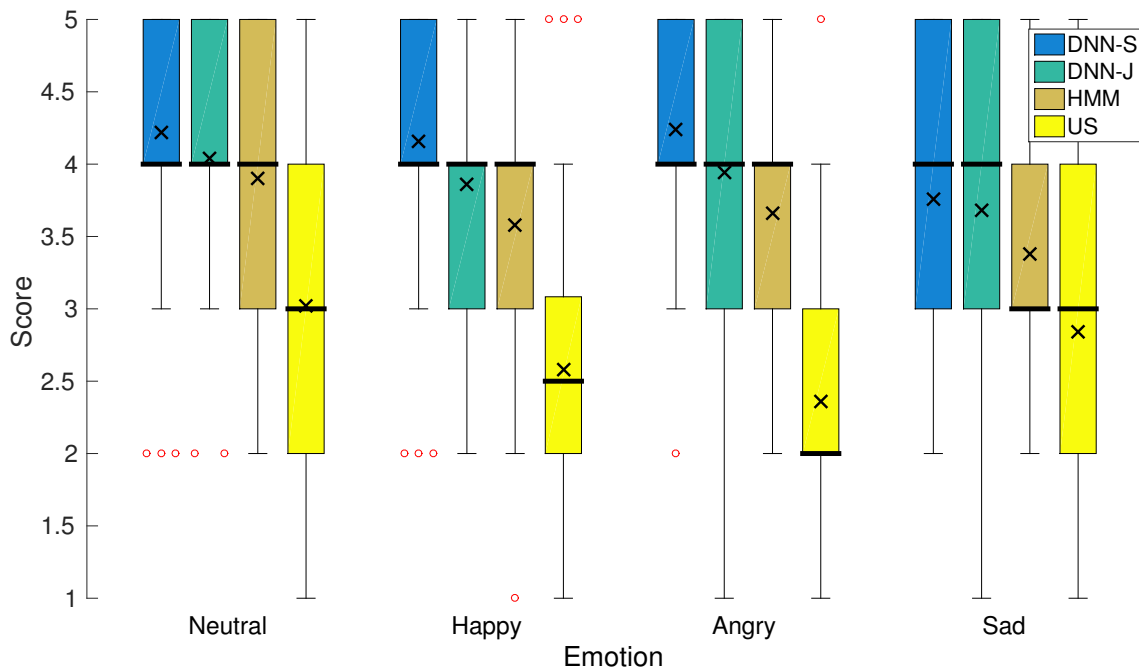


Figure 6.5: Results of the MOS test broken down for each different emotion. Bold line represents the median, x represents the mean, the boxes extend between the 1st and 3rd quartile, whiskers extend to the lowest and highest datum within 1.5 times the interquartile range of the 1st and 3rd quartile respectively, and outliers are represented with circles.

Evaluation of acoustic realism

For evaluating the acoustic speech generated, human evaluators were presented with random pairs of acoustic speech samples and were asked to pick the most realistic. Just like in the evaluation of visual realism, realism of acoustic speech was evaluated 120 times (30 pairs for each emotion) for each different method pair. The results are presented in Table 6.4.

We can see that all pairwise comparisons are significant at the $p < 0.01$ level, apart from the comparisons between DNN-J, HMM and DNN-J, US, where, although the DNN-J method is preferred, we did not observe statistical significance.

Table 6.4: Results (%) of subjective pairwise preference tests on acoustic speech realism. Bold font indicates significant preference at $p < 0.01$ level.

DNN-S	DNN-J	HMM	US	N/P
40.0	9.17	-	-	50.83
65.83	-	7.5	-	26.67
79.17	-	-	14.17	6.67
-	41.67	26.67	-	31.67
-	55.83	-	32.5	11.67
-	-	64.17	26.67	9.17

Table 6.5: Results (%) of subjective pairwise preference tests on audio-visual speech expressiveness. Bold font indicates significant preference at $p < 0.01$ level.

DNN-S	DNN-J	HMM	US	N/P
23.08	23.08	-	-	53.85
50.64		15.38	-	33.97
70.51		-	23.07	6.41
-	42.3	26.28	-	31.41
-	66.67	-	27.56	5.77
-	-	57.69	36.54	5.77

Evaluation of expressiveness

Expressiveness was evaluated in the same manner as audio-visual realism, where pairs of videos were presented and compared by human evaluators on their expressiveness. Videos of the neutral emotion were not included. The 6 pairwise preference tests on the evaluation of expressiveness were evaluated 156 (52 pairs for each emotion) times each, and we show the results in Table 6.5.

We see that the DNN-S architecture is significantly preferred over the HMM and US methods. The DNN-J architecture is significantly preferred over US, and although preferred over HMM, it is not significant in a statistical meaning. Between HMM and US, the former is preferred, though the result again is not statistically significant.

A correlation between realism and expressiveness is evident, since we can see that the results follow a resembling course with the evaluation of the audio-visual realism.

6.5 Chapter Conclusions

In this chapter we proposed two different architectures for DNN-based expressive audio-visual speech synthesis and did a direct comparison with HMM-based and concatenative unit selection expressive audio-visual speech synthesis systems on the realism of the produced talking head, and on the emotional strength that is captured by each system when it is trained on an emotional corpus.

Our results show that both DNN-based architectures significantly outperform the other two methods in terms of the audio-visual realism of the synthesized talking head, while the DNN-based architecture that uses separate modeling of acoustic and visual features architecture (DNN-S) significantly outperforms the HMM and US methods in terms of expressiveness as well. In addition, DNN-S also achieved significantly better results over all other architectures when considering acoustic speech only.

The results of the unit selection system were much worse in comparison with parametric approaches, which is to be expected when considering not only the fact that our corpus is fairly small for US synthesis, but also that the number of needed units increases when considering expressive speech.

7

Visual Speech-Informed Perceptual 3D Reconstruction and Manipulation of Facial Expressions

7.1 Introduction

In Chapters 5 and 6 we explored generation of audiovisual speech and facial expressions from text. In this chapter, we redirect our focus on a different class of problems which share the common theme of modeling and preserving the facial expressions of a person that correspond to speech. The first and main problem concerns 3D reconstruction of human faces, so that the perception of speech in the original footage is retained. The second problem we tackle is the one of modification of facial expressions of a talking person, without altering the uttered speech.

Without doubt, the recent state of the art on monocular 3D face reconstruction from image data has made some impressive advancements, thanks to the advent of Deep Learning. The current state of the art is able to robustly reconstruct fine details of the 3D facial geometry as well as yield a reliable estimation of the captured subject’s facial anatomy. This is beneficial for multiple applications, such as augmented reality, performance capture, visual effects, photo-realistic video synthesis, human-computer interaction and personalized avatars, to name but a few.

On the other hand, the vast majority of existing methods focus on 3D face reconstruction from a single RGB image, without exploiting the rich dynamic information that is inherent in humans’ faces, especially during speech. But even the few methods that include some sort of dynamics modelling to reconstruct facial videos, do not explicitly model the strong correlation between mouth motions and articulated speech. At the same time, most facial videos of interest capture individuals involved in some form of verbal communication. When existing 3D face reconstruction methods are applied in this kind of videos, the artifacts in the reconstruction of the shape and motion of the mouth area are often severe and overwhelming in terms of human perception; the movements of the mouth that correspond to speech are not captured well.

Arguably, a crucial factor for the limitations of existing methods is the fact that most methods use weak 2D supervision from landmarks predicted by face alignment methods as a form of guidance[Saito et al., 2016, Thies et al., 2018, Thies et al., 2016, Jackson et al., 2017, Booth et al., 2018a, Tewari et al., 2017, Feng et al., 2021, Yang et al., 2020]. While these landmarks can yield a coarse estimation of the facial shape, they fail to provide an accurate representation of the expressive details of a highly-deformable mouth region. It

is also important to note that the shapes of the human mouth are perceptually correlated with speech and the realism of a 3D talking head is tightly coupled with the uttered sentence. For example, a 3D model that talks without the lips closing when uttering the bi-labial consonants (i.e., /m/, /p/, and /b/), or with no lip-roundness when uttering a rounded-vowel (such as /o/ /u/) has a poor perceived naturalness.

Nonetheless, apart from inaccurate prediction of 2D landmarks in several cases (e.g., see Fig. 7.1, weak 2D supervision for dense modeling is ill-posed, especially for the lip area which can assume extremely diverse formations. Note also that while some lip landmarks (in the most commonly used template of 68 facial landmarks) such as the lip corners have a semantic meaning, intermediate lip landmarks have an intrinsic ambiguity in their definition and present significant variance across different annotators [Sagonas et al., 2016]. EMOCA [Daněček et al., 2022] made some important advancements in terms of the expressivity of the 3D reconstructed head, however the perceptual emotional consistency loss only affected the movements that correspond to facial expressions. Furthermore, the estimation of jaw articulation was not included in the model, resulting often in poor reconstructions.

We conclude that, although speech perception from reconstructed 3D faces is important for various applications (e.g., augmented and virtual reality, gaming, affective avatars etc.) [Hofer et al., 2020, Marín-Morales et al., 2020, Stuart et al., 2022], it is a commonly overlooked parameter in the existing literature.

To overcome the limitations of the existing literature, this work tackles the problem of monocular 3D face reconstruction from a video, with a strong focus on the mouth area and its expressions and movements that are connected with speech articulation. We highlight and address the fact that an accurate 3D reconstruction of a human talking in a video should retain those mouth expressions and movements that humans perceive to correspond to speech. Our method, dubbed *SPECTRE*, leverages a SoTA model of lip reading to minimize the “speech-informed perceptual” distance between the rendered and the original input video. Our main contributions can be summarized as follows:

- We design and implement the first (to our knowledge) method for perceptual 3D reconstruction of human faces focusing on speech **without the need for text transcriptions of the corresponding audio, or costly 3D annotations**.
- We propose a perceptual “lipreading” loss based on deep features, minimizing the perceptual distance of speech-related lip movements between the original and reconstructed (through a differentiable 3D face renderer) videos.
- We conduct experiments over the effectiveness of deep features against traditional geometric based metrics and showcase numerous examples where *SPECTRE* significantly outperforms other methods in speech-aligned mouth perceptibility. Our proposed system also generalizes well to other datasets, as demonstrated by our **cross-dataset experiments**.
- Finally, we also present NED, a method for facial expression manipulation in photorealistic videos, without altering the uttered speech.

7.2 Related Work

3D Models: There is extensive literature in the fields of computer vision and graphics for creating and reconstructing 3D face models from various input sources (RGB, Depth)



Figure 7.1: Examples of inaccuracies in 2D landmark detection in current state-of-the-art methods [Bulat and Tzimiropoulos, 2017]. Notice how especially on the right column the face alignment has not accurately predict mouth closure which is of vital important for realistic perception of bilabial consonants (/p/, /m/, /b/).

[Zollhöfer et al., 2018, Egger et al., 2020]. 3D Morphable Models are by far the most widely-used choice, since they offer compact representations as well as a convenient decoupling of expression and identity variation, allowing better manipulation. The traditional 3DMMs were linear, PCA-based models of 3D shape variation, but several non-linear and deep learning-based extensions have been proposed during the last years [Tran and Liu, 2018, Bagautdinov et al., 2018, Abrevaya et al., 2019, Cheng et al., 2019]. During the last decades, several 3D face models have been built from large datasets of 3D scans of human faces [Paysan et al., 2009, Gerig et al., 2018, Cao et al., 2013, Booth et al., 2018b, Li et al., 2017, Yang et al., 2020, Wang et al., 2022, Bao et al., 2021, Chai et al., 2022].

Monocular 3D Face Reconstruction: A common application of 3DMMs includes estimation of the model parameters that best fit to an RGB image. This can happen as a direct optimization procedure in an analysis-by-synthesis framework [Blanz and Vetter, 1999, Aldrian and Smith, 2012, Thies et al., 2015, Thies et al., 2018, Booth et al., 2018a]. However this is a computationally expensive procedure to run on novel images every time (e.g., the recent FaceVerse method [Wang et al., 2022] needs ~ 10 mins. for detailed refinement). Due to this reason, various methods have emerged that formulate the problem as a regression from image data, leveraging the power of Deep Learning [Tewari et al., 2017, Jourabloo and Liu, 2016, Jackson et al., 2017, Tran et al., 2018, Gecer et al., 2019, Grassal et al., 2022, Deng et al., 2019, Ruan et al., 2021, Feng et al., 2018]. Combined with a reliable facial landmarker, this can lead to accurate results, even without the need for 3D supervision.

For example, RingNet [Sanyal et al., 2019], performs 3D reconstruction using the FLAME model [Li et al., 2017], by enforcing a shape-consistency loss between images

of the shape subject, in order to decouple identity and expression. This is improved by DECA [Feng et al., 2021], which predicts the FLAME parameters jointly from a CNN, using multiple loss coefficients that tackle the lack of 3D ground truth. EMOCA [Daněček et al., 2022] focuses on the expressiveness of the reconstructed models, adding an emotion-related perceptual loss and training a CNN that predicts the expression parameters of the 3DMM on a large emotional dataset (AffectNet). ExpNet [Chang et al., 2018] generates pseudo-3DMM parameters by solving the optimization problem given an accurate 3D reconstruction of an image with a SoTA method and then training a CNN to predict them, without the need for landmarks. In 3DDFA [Guo et al., 2020, Zhu et al., 2017], face alignment and 3D reconstruction takes place concurrently, using Cascaded CNNs. MICA [Zielonka et al., 2022] focuses on accurate prediction of the identity parameters of a 3DMM, by employing a medium-scale 3D annotated dataset in conjunction with a large-scale 2D raw image dataset. DAD-3DHeads [Martyniuk et al., 2022], provides one of the first large-scale 3D head datasets, that can be used for direct supervision of 3D reconstruction. Finally, most recently Wood et al. [Wood et al., 2022] shows how synthetic data can be used for monocular 3D reconstruction which generalizes to real world footage. Some methods also try to deal with occlusions [Dey and Boddeti, 2022, Li et al., 2021a, Shang et al., 2020].

Even though the vast majority of methods reconstruct single face images or work on a frame-by-frame fashion on videos, there are a few methods that exploit the dynamic information of monocular face videos to constrain the subject’s facial shape or impose temporal coherence on the face reconstruction [Cao et al., 2015, Garrido et al., 2016b, Huber et al., 2016, Koujan and Roussos, 2018, Booth et al., 2018b].

A recent rising trend is exploiting deep features as metrics that correlate better with human perception compared to traditional metrics [Zhang et al., 2018a]. Our work is mostly similar to EMOCA [Daněček et al., 2022], in the sense that both are concerned with perceptual reconstruction. In comparison, however, EMOCA focuses on retaining affective information from images while our work focuses on accurate reconstruction of mouth and lips formations that correspond to speech production. Furthermore, EMOCA failed to accurately predict the jaw pose parameters which include opening and rotation of the mouth due to difficulties in convergence and kept the jaw pose fixed.

Mouth/Lip Reconstruction: Some of the earliest works focusing on the dynamics of mouth and lips for 3D reconstruction include the works of Basu et al. [Basu et al., 1998b, Basu et al., 1998a] which used a combined-statistical model, Gregor et al. [Kalberer and Gool, 2000] who used markers to follow the lip motions, and Cheng et al. [Cheng and Huang, 2010] who performed mouth tracking from 2D images using Adaboost and a Kalman filter. The most recent work concerned with lip tracking from video is the work of Garrido et al. [Garrido et al., 2016a], who achieved remarkable results of 3D reconstructed lips, using the ground truth shapes of a high quality 3D stereo database along with radial basis functions.

7.2.1 Method

Preliminaries

Our work is based on the state-of-the-art DECA [Feng et al., 2021] framework for monocular 3D reconstruction from static RGB images. As such we adopt the notation from the DECA paper. In the original DECA, given an input image I a coarse encoder (a ResNet50 CNN) jointly predicts the identity parameters $\beta \in \mathbb{R}^{100}$, neck pose and jaw $\theta \in \mathbb{R}^6$, expression parameters $\psi \in \mathbb{R}^{50}$, albedo $\alpha \in \mathbb{R}^{50}$, lighting $\mathbf{l} \in \mathbb{R}^{27}$, and camera (scale and

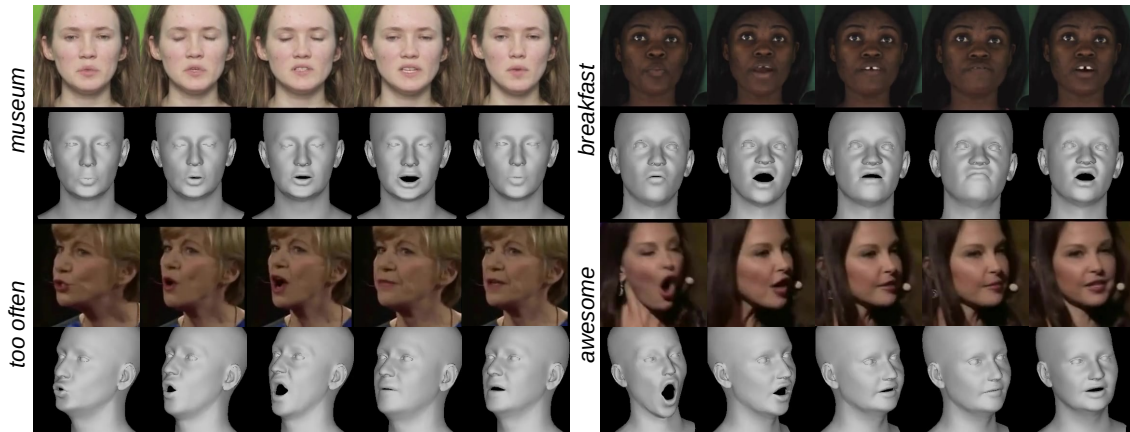


Figure 7.2: Our method SPECTRE performs visual-speech aware 3D reconstruction so that speech perception from the original footage is preserved in the reconstructed talking head. On the left we include the word/phrase being said for each example.

translation) $\mathbf{c} \in \mathbb{R}^3$. Note that these parameters are a subset of the parameters of the FLAME 3D face model. Afterwards, these parameters are used to render the predicted 3D face. DECA also included a detail encoder which predicted a latent vector associated with a UV-displacement map, that models high-frequency person-specific details such as wrinkles. More recently, EMOCA [Daněček et al., 2022] further built upon DECA by adding an extra expression encoder (ResNet50) which was used in order to predict the expression vector ψ , so that the perceived emotion of the reconstructed face is similar to that of the original image. We use these two works as starting points and focus on designing an architecture that increases the perceived expressions of the input video, concentrating on the mouth area, leading to realistic articulation movements.

Architecture

A high-level overview of the architecture is shown in Figure 7.3. Given a sequence of K RGB frames sampled from an input video V , our method reconstructs for each frame I the 3D mesh of the face in FLAME topology, such that the mouth movements and general facial expressions are perceptually preserved. Following the FLAME 3D face model nomenclature, we separate the estimated parameters into two distinct sets:

Rigid & Identity parameters We borrow the coarse encoder from DECA in order to predict independently for each image I in the input sequence the identity β , neck pose θ_{neck} , albedo $\alpha \in \mathbb{R}^{50}$, lighting $\mathbf{l} \in \mathbb{R}^{27}$, and camera \mathbf{c} . Like EMOCA [Daněček et al., 2022], we keep this network fixed through training.

Expression & Jaw parameters The expression ψ and jaw pose θ_{jaw} parameters that correspond to the input sequence is predicted by an additional “perceptual” CNN encoder. These parameters explicitly control the mouth expressions and movements under the FLAME framework and therefore should be properly estimated by our approach. We employ a lightweight MobileNet v2 architecture, but also insert a temporal convolution kernel on its output, in order to model the temporal dynamics of mouth movements and facial expressions in the input sequence. We selected the aforementioned lightweight option

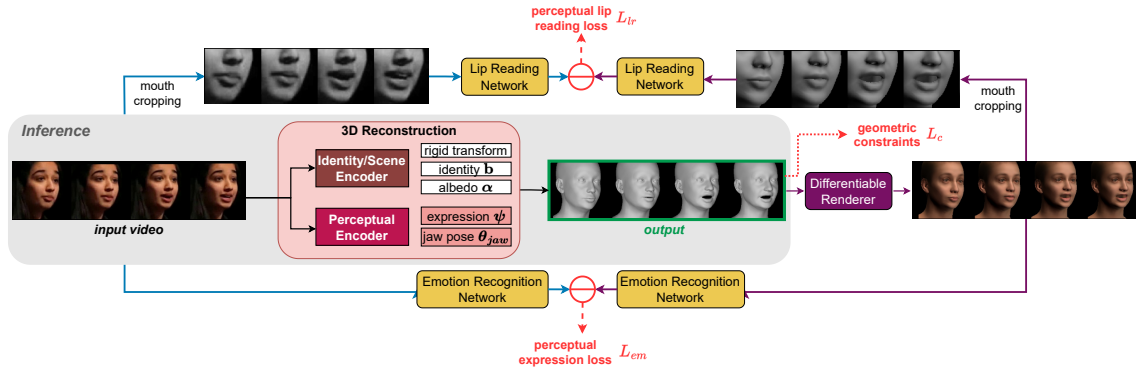


Figure 7.3: Overview of our architecture for perceptual 3D reconstruction. The input video is first fed into the 3D reconstruction component, where a fixed encoder detects the scene parameters (camera, lighting), identity parameters (albedo/identity) and an initial estimate of the jaw and expression parameters. Then, a Mouth/Expression encoder predicts the refined facial expression parameters and jaw pose, and a differentiable renderer renders the predicted 3D shape. Finally, the mouth area is differentially cropped in both the input and rendered image sequences and a lip reader is applied on both in order to estimate the perceptual lip reading loss between them. Similarly, a perceptual expression loss is used on the full face based on an emotion recognition network. Inference requires only the 3D reconstruction component.

of MobileNet to reduce the computational overhead of our system - contrary to EMOCA - since the existing DECA backbone already uses a resource-demanding ResNet50 model.

In a nutshell, we assume an architecture akin to the one introduced in EMOCA [Daněček et al., 2022], with two parallel paths of parameters as described above. Nevertheless, our focus is shifted to a very different problem and thus a set of appropriate “directions” and “constraints” should be learned through the use of the proposed set of losses, as described in the following section.

Training Losses

In order to train the *perceptual* encoder, we use two perceptual loss functions for guiding the reconstruction, along with geometric constraints.

Perceptual Expression Loss The output of the *perceptual* encoder is used along with the predictions of identity, albedo, camera, and lighting in order to differentially render a sequence of textured 3D meshes, which correspond to the original input video. Then, the input video and the reconstructed 3D mesh are fed into an emotion recognition network (borrowed from EMOCA [Daněček et al., 2022]) and two sequences of feature vectors are obtained. Then, we apply a perceptual expression loss L_{em} , by attempting to minimize the distance between the two sequences of feature vectors. Interestingly, even though the emotion recognition network is trained to predict emotions, it can faithfully retain a set of helpful facial characteristics. Therefore, such a loss is responsible for learning general facial expressions, capable to simulate emotions, which promote the realism of the derived reconstruction. Notably, this loss positively affects the eyes, leading to a more faithful estimation of eye closure, frowning actions etc.

Perceptual Lip Movements Loss The perceptual expression loss does not retain enough detailed information about the mouth, and as such, an additional mouth-related loss is needed. Instead of relying only on a geometric loss with weak supervision using 2D landmarks, we use an additional perceptual loss, that guides the output jaw and expression coefficients to capture the intricacies of mouth movements. *The necessity of such a perceptual mouth-oriented loss is further highlighted by the inaccuracies detected in the extracted 2D landmarks.*

For this purpose we use a network that has been trained on the LRS3 (Lip Reading in the Wild 3) dataset [Ma et al., 2022]. The lip-reading network is the pretrained model provided by Ma et al. [Ma et al., 2022] which takes as input sequences of grayscale images cropped around the mouth and outputs the predicted character sequence. The network has been trained with a combination of Connectionist Temporal Classification (CTC) loss with attention. The model architecture consists of a 3D convolutional kernel, followed by a 2D ResNet-18, a 12-layer conformer, and finally a transformer decoder layer which outputs the predicted sequence (for more details, see [Ma et al., 2022]). Our goal here is to minimize the perceptual distance of speech-aware movements between the original and the output image sequences. To that end, we take the differentiably rendered image sequences and subsequently crop the them around the mouth area using the predicted landmarks. Finally, we calculate the corresponding feature vectors ϵ_I and ϵ_R , from the output of the 2D ResNet-18 of the lip-reading network. We empirically found that features from the CNN output better model the spatial structure of the mouth, while features on the output of the conformer are largely influenced by the sequence context and do not preserve this much-needed spatial structure. After calculating the feature vectors, we minimize the perceptual lip reading loss between the input image sequence and the output rendered sequence $L_{lr} = \frac{1}{K} \sum^K d(\epsilon_I, \epsilon_R)$, where d is the cosine distance and K the length of the input sequence. As a sidenote, initial experiments included an explicit lip reading loss based on the CTC loss over the predicted output of the existing lip reading network, given the original transcription of the sentence. Despite its straightforward intuition, such approach had major downsides apart from the need of the video transcription. First, it had a significant computational overhead since whole sentences should be processed at once. In contrast, the proposed approach simply samples a subset of consecutive frames and tries to minimize the extracted mouth-related features. Moreover, it has proven to be ineffective in practice, suffering from the same behavior as with the features taken from the conformer’s output.

Geometric Constraints Due to the domain mismatch between the rendered and the original images, although the perceptual losses help retain the high level information on perception, they also tend to create artifacts in some cases. This is to be expected; the perceptual losses rely on pre-trained task-specific CNNs that do not guarantee in any way that the input manifold corresponds to realistic images. For example, as we report in our experiments, we can create unrealistic images of distorted facial reconstruction that produce good lip reading results - a typical problem in the adversarial examples topic [Goodfellow et al., 2014]. Thus, we guide the training process by enforcing the following geometric constraints: We regularize the expression and jaw parameters by penalizing their L_2 norm: $\|\psi\|^2$ and $\|\theta_{jaw}\|^2$.

In addition, we also apply an L_1 (average per-landmark distance) loss between the predicted and original landmarks (obtained with face alignment [Bulat and Tzimiropoulos, 2017]) of the **nose**, **eyes** and **face outline**: $L_{face} = \|\mathbf{E}_r - \mathbf{E}_{gt}\|$, where \mathbf{E}_r are the predicted and \mathbf{E}_{gt} the original landmarks.

For the mouth however, we employ a more relaxed constraint by using the intra-distances of mouth landmarks instead of the direct values: $L_{mouth} = \|\mathbf{D}_r^{mouth} - \mathbf{D}_{gt}^{mouth}\|^2$, where \mathbf{D}_r^m are the distances between pairs of the predicted mouth landmarks while \mathbf{D}_{gt}^m are the distances of pairs of original mouth landmarks. We use this more relaxed version because a straightforward loss between the absolute positions of the predicted and original landmarks is more strict and can lead to erroneous reconstructions, since perceptual losses and 2D landmark loss can be contradicting. Note that during our two-stage training scheme we switch from $L1$ to $L2$ loss. In our experiments we observed that in the case of $L1$ loss, the constraint was strict enough to not allow the appearance of artifacts, but at the same time did not allow the mouth to accurately capture some perceptually meaningful mouth formations such as lip closure or roundness. On the other hand, we observed that the $L2$ loss had the reverse behavior: better capturing of mouth formations but mouth artifacts in some rare cases.

The complete loss function is now is then:

$$L = \lambda_{lr}L_{lr} + \lambda_{em}L_{em} + \lambda_{\psi}L_{\psi} + \lambda_{\theta_{jaw}}L_{\theta_{jaw}} + \lambda_{face}L_{face} + \lambda_{mouth}L_{mouth} \quad (7.1)$$

where $\lambda_{lr} = 2$, $\lambda_{em} = 0.5$, $\lambda_{\theta_{jaw}} = 200$, $\lambda_{face} = 50$, $\lambda_m = 50$. Note that especially for the weight λ_{ψ} , we selected an adaptive weighting scheme:

$$\lambda_{\psi} = \left\{ \begin{array}{ll} 1e - 3, & \text{if } L_{\psi} < 40 \\ 2e - 3, & \text{if } L_{\psi} > 40 \end{array} \right\} \quad (7.2)$$

which we found in practice to work better than a traditional fixed weight. The motivation behind this nonlinear tweak of the regularization term is to impose stricter constraints after an empirical threshold, since we have observed that the necessity to continuously minimize the reported perceptual losses may lead to artifacts. Even though this modification does not significantly affects the procedure, we found it effectively reduced specific artifacts.

7.2.2 Two-stage Training

We follow a two stage training scheme for SPECTRE. First, we train the perceptual encoder for 10 epochs ($\sim 250,000$ iterations) on the LRS3 dataset using $L1$ loss for the function L_{mouth} , a batch size of 1 and a sequence length of $K = 20$. Then, we train for 10,000 iterations using an $L2$ loss. We found that this training scheme achieves a good trade-off, greatly mitigating visual artifacts on the mouth and also achieving better perception for the reconstructed mouth.

7.3 Experiments

We evaluate our method both qualitatively and quantitatively, following a similar evaluation procedure to [Daněček et al., 2022]. The considered datasets are the following:

- **LRS3** [Afouras et al., 2018]: We use Lip Reading Sentences 3 (LRS3) dataset [Afouras et al., 2018], which is the largest publicly available dataset for lip reading in the wild, for training and testing our system. The official *trainval* set (31,982 utterances) is used for training and validating our model, while evaluation is performed on the test set of LRS3 (1,321 utterances).

	CER ↓	VER ↓	lMAE ↓	vMAE ↓	R^2 ↑
Original Vid	42.6	32.6	-	-	-
DECA	100.9	89.9	8.52	4.82	0.62
EMOCA	100.9	90.7	9.74	6.66	0.58
DAD	92.3	86.6	6.13	5.50	0.88
SPECTRE	87.6	77.0	8.12	5.51	0.77

Table 7.1: Lipreading (CER, VER) and geometric-based metrics (lMAE, vMAE, R^2 score) are reported on the VOCASET test set. Mean absolute error is calculated both on the mouth landmarks (*lMAE*) and their temporal velocity (*vMAE*). While SPECTRE achieves significantly better lipreading metrics, this result is not reflected on traditional geometric errors (MAE scaled by $\times 10^3$).

- **MEAD**: This is a recent dataset [Wang et al., 2020b] containing 48 actors (28M, 20F) from multiple races uttering sentences from TIMIT [Garofolo et al., 1993] in 7 emotions and 3 different levels of intensity. The whole dataset includes 31,059 sentences. We randomly sampled 2,000 in order to create a test set, stratifying for subject, emotion, and intensity level.
- **TCD-TIMIT** [Harte and Gillen, 2015]: This corpus includes 62 English actors reading 6913 sentences from the TIMIT [Garofolo et al., 1993] corpus. We use the official test split for evaluation.
- **VOCASET** [Cudeiro et al., 2019b]: VOCASET includes 12 subjects speaking 40 utterances each. It is the only dataset which includes ground truth registered vertices in the FLAME mesh topology, enabling evaluation with geometric-based metrics. We use the official test split for evaluation.

Comparisons: We compare our method to the following recent state-of-the-art methods on 3D facial reconstruction: **DECA** [Feng et al., 2021], **EMOCA** [Daněček et al., 2022], **3DDFAv2** [Guo et al., 2020], and **DAD-3DHeads**, which uses direct 3D supervision from the large-scale annotated DAD-3DHeads [Martyniuk et al., 2022] dataset. Note that these methods, as almost all recent methods for visual reconstruction of the 3D face geometry, are using a single RGB image as input. Therefore, in order to reconstruct the entire input video, we apply them in every frame of the video. Especially for 3DDFAv2, we apply temporal smoothing as provided by the official implementation. For all methods we use the official implementation.

7.3.1 Quantitative Evaluation

In this section, we seek to quantify speech-related perceptual cues. A straightforward way is to evaluate the compared methods objectively in terms of lip reading metrics by applying a pretrained lipreading network on the output rendered images. To remove bias, we use a *different architecture and pretrained lipreading model* for evaluation than the one used for the lipreading loss, which is based on the Hubert transformer architecture, called AV-HuBERT [Shi et al., 2022a, Shi et al., 2022b]. The following lipreading metrics are considered: Character Error Rate (CER), as well as Viseme Error Rate (VER), obtained by converting the predicted and ground truth transcriptions to visemes using the Amazon Polly phoneme-to-viseme mapping [Amazon, 2015].

We first present results on the VOCASET dataset, which contains ground truth 3D reconstruction, in Table 7.1. Apart from the lipreading metrics (CER, VER), we also report

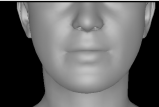

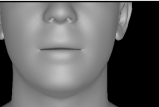
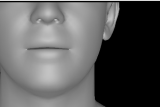

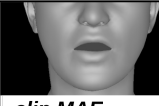

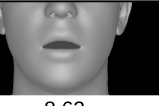


Ground Truth	DAD	DECA	EMOCA	SPECTRE
				
				
<i>clip MAE</i>	5.42	8.62	13.2	7.55
<i>clip CER</i>	93.3	97.3	128.9	71.7
<i>clip VER</i>	87.4	88.5	114.2	68.5

Figure 7.4: Comparison of 3D reconstructions of the mouth area for 2 frames from an example VOCASET clip. The MAE, CER and VER errors over the clip are also reported (best result w.r.t. each metric is in bold). MAE values are scaled by $\times 10^3$. Notice the discrepancy in the ranking of the different methods between MAE and CER/VER metrics. We observe that the perceived quality of mouth reconstruction seems to have a much better correlation with CER and VER metrics, rather than MAE.

the mean absolute error between the predicted and the ground-truth mouth landmarks ($lMAE$) and between their “velocity” - calculated as the difference of consecutive frames - ($vMAE$), as well as the R^2 score. Note how lipreading results are not correlated with the latter set of geometric errors/scores. Specifically, our approach leads to significantly improved lipreading metrics compared to the other reconstruction methods, as expected, while DAD achieves the best $lMAE/R^2$ scores, powered by its detailed 3D supervision approach. Contrary to DECA and EMOCA, our method has a high R^2 value despite being trained on the proposed lipreading loss and thus expecting loose geometric correspondence. Moreover, the $vMAE$ metric highlights the movements of the lips, where SPECTRE and DAD have similar performance hinting towards a desired over-articulation [Aldeneh et al., 2022], while EMOCA tends to be excessively active to capture intense emotions. For further validation of these behaviors, we show two example snapshots from VOCASET in Fig. 7.4 where it can be seen that the landmark MAE error does not represent well the mouth formation and is not representative of the perceived 3D reconstruction quality. This result has been highlighted by various previous works, which have pointed out that geometric errors of facial/mouth expressions do not correlate well with human perception [Daněček et al., 2022, Mori et al., 2012, Garrido et al., 2016a, Websdale et al., 2022, Aldeneh et al., 2022, Theobald and Matthews, 2012].

For the rest of datasets we do not have ground truth landmarks, and predicted ones tend to not capture mouth formations well. As a result, we evaluate only the aforementioned lipreading metrics. Results are presented in Table 7.2. Our method achieves considerably lower CER and VER scores compared to the other methods, both in the LRS3 test set, as well as in the cross-dataset evaluations of TCDTIMIT and MEAD. In the same Table we also include results on the original video footage, which showcase the domain gap “problem” (more information about this in Discussion section) of the used lip reading systems: the pre-trained models have been trained to the initial images without the possible visual degradation introduced by the rendering procedure. Nonetheless, our method provides notable boost in lip reading performance, despite missing key features such as tongue and teeth, by properly encoding speech-related features.

	LRS3		TCD-TIMIT		MEAD	
	CER ↓	VER ↓	CER ↓	VER ↓	CER ↓	VER ↓
Original vid	24.9	22.0	35.7	29.6	49.7	42.8
DECA	77.5	70.8	84.2	75.8	84.8	77.8
EMOCA	83.3	76.3	86.4	79.2	85.1	77.9
3DDFAv2	97.5	95.3	101.8	98	94.5	90.2
DAD	84.1	78.2	87.3	81	86.0	79.9
SPECTRE	67.5	60.9	78.1	69.6	78.5	71.1

Table 7.2: Lipreading results on the LRS3-test, TCD-TIMIT and MEAD datasets (network trained on LRS3-train set). For all metrics, lower is better (error rates). Our method significantly outperforms all other 3D reconstruction methods. The 1st row corresponds to results on the original videos, reported as reference.

7.3.2 User Studies

The quantitative evaluation highlighted the difficulty to pin down well-received perceptual cues into a concrete geometric error. In fact, introducing a realistic, non-excessive over-articulation should be favorable with respect to human perception despite the expected deviation from geometric errors, as pointed out in [Aldeneh et al., 2022]. Arguably, the ultimate goal of a talking head is for humans to perceive it as natural and as realistic as possible. To assess the realism and perception of the 3D reconstructed faces by humans, we have designed and conducted two web user studies [Kritsis et al., 2022]. In order to mitigate any intra-dataset bias that might arise from training on the LRS3 trainset and showing users video from its test set, for these studies, we used only videos from the MEAD and TCD-TIMIT datasets.

First Study: Realism of Articulation. For this study, we selected a preference test design, by showing users pairs of videos with 3D face reconstruction results, alongside the original footage, and asking them to select the most realistic one in terms of mouth movements and articulation. We created a question bank consisting of 30 videos from the MEAD dataset (21 emotional videos for each level of intensity and emotion and 9 neutral), and 10 videos from the TCD-TIMIT dataset and performed 3D reconstruction using the previously stated 5 methods (DAD, DECA, EMOCA, 3DDFAv2 and ours). Then, users were presented with two randomly ordered reconstructed faces, alongside the original footage, and were asked to choose the most realistic one in terms of mouth movements and articulation. Each user answered 28 randomly sampled questions from the bank (7 questions for each pair - ours vs the others), and a total of 34 users completed this study.

The results of this study can be seen in Table 7.3. We can see that our method is significantly preferred to all other methods ($p < 0.01$ with binomial test, adjusting for multiple comparisons using the Bonferroni method). 3DDFAv2 [Guo et al., 2020] was the least preferred method, with DECA and EMOCA following. The results clearly highlight the importance of the proposed method from the speech-aware perspective and how humans favorably perceive the reconstructed mouth movements as more realistic in SPECTRE, compared to the other methods.

Second Study: Lip Reading. In the second study, users (disjoint set of participants compared to the first study) were presented with a muted video of a person uttering a specific single word in the form of a 3D talking head reconstructed from one of the compared methods and then were asked to select the correct word among 4 different

	DECA	EMOCA	3DDFAv2	DAD
SPECTRE	201/37	185/53	218/20	150/88

Table 7.3: Preference results of the first subjective study. The “a/b” depiction of the result means that SPECTRE (on the left) was preferred a times while the competing method (given as column header) was chosen b times out of the 238 pairs the subjects assessed. Our method is **significantly** more realistic ($p < 0.01$ with binomial test after adjusting for multiple comparisons) in terms of mouth movements and articulation.

alternatives (multiple choice). For this, we cropped 40 single words from the MEAD and TCD-TIMIT datasets, covering different visemes, and presented each user with a random subset of 30 words (6 words for each method in each questionnaire). A total of 31 users completed this study. Classification results are shown in Table 7.4. In a similar fashion with the objective results, SPECTRE outperforms other methods in terms of word classification. An interesting result is the fact that EMOCA achieves a relatively high result compared to objective results. This could be due to the fact that in some cases, e.g., unrealistically exaggerated expressions as seen in EMOCA, can be sufficient for distinguish specific words.

We perform more in depth evaluation of the results of the second user study by showing a per-word analysis. Specifically, we report the recognition accuracy in Table 7.5 for five indicative cases where our method under-performs and also five cases where our method outperforms competition. In addition, in Figure 7.5 we also show example video reconstructions of three words: PERFUME, NARROW, and PEOPLE. In the first two words our method had a significantly higher recognition accuracy, while in the last one, DAD performed better. As we can see from the visual comparison in 7.5(c), our method in this specific case failed to accurately capture the closed mouth formations that correspond to the bilabial consonants /p/. On the other hand, in the the first two words our method accurately captures the mouth formations for the rounded vowels /o/ and /u/ in contrast with EMOCA and DAD. Note how also in PERFUME, our method accurately depicts /f/ in the third frame.

In the study there were also cases where the majority of the methods perform well due to the very distinct pronunciation of the words (e.g. “BALEFUL” and “UMBRELLA”) and cases where all methods considerably under-perform (e.g., “GREASY”, “SURRENDER”) due to subpar reconstruction and “difficult” alternative words (e.g., “SURRENDER” was mostly confused with the alternative choice “SURROUNDED”).

SPECTRE	DECA	EMOCA	3DDFAv2	DAD
47.56%	39.83%	45.12%	23.17%	45.12%

Table 7.4: Classification accuracy in the second user study (word-level lipreading).

7.3.3 Visual Comparisons and Ablations

In Figures 7.6 and 7.7 we present multiple visual comparisons from the 3 datasets with the four other methods.

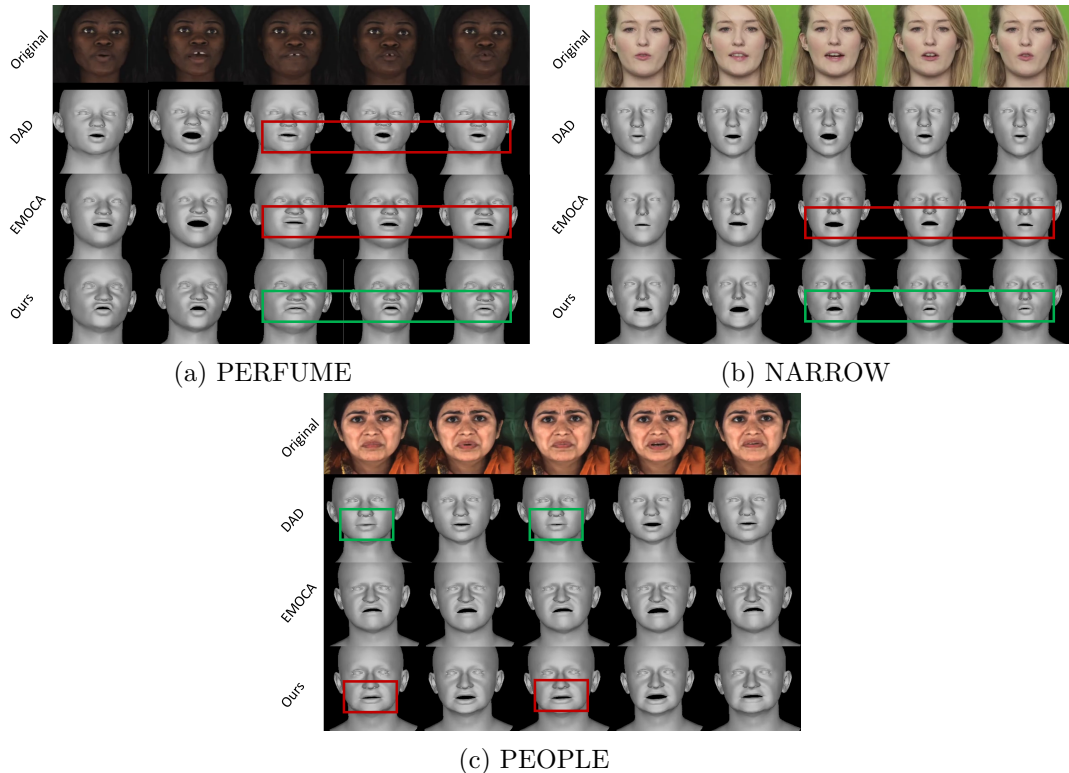


Figure 7.5: Three example words from the second user study (lip reading). We show our method against DAD and EMOCA. In PERFUME and NARROW, our method accurately predicts the rounded mouth formations. In the third case of PEOPLE, we see a failure case, where the bilabial consonant /p/ which corresponds to closed mouth was not predicted accurately. Note how also in PERFUME, our method accurately depicts /f/ in the third frame.

Ablation study on geometric constraints

In Fig. 7.8 we also show results of training the network with and without the geometric constraints from landmarks. We can see that in some cases, completely removing geometric constraints and training only with perceptual losses leads to artifacts around the eyes, nose and mouth shape.

Ablation study on lipreading features and CTC loss

ResNet18 vs Conformer features As mentioned in Section 3.2 of the main text, we selected features from the ResNet18 output of the lipreading network instead of latter

	PLACE	PEOPLE	WITHDRAW	AROUND	CONSIDERABLE	OVERALLS	WHATEVER	NARROW	AUTHORIZED	PERFUME
DAD	67	100	75	80	50	50	50	78	67	40
EMOCA	50	43	100	14	100	25	50	75	0	71
DECA	33	50	25	67	38	33	50	0	33	60
3DDFAv2	33	0	17	40	0	67	33	33	30	43
SPECTRE	50	60	29	40	67	100	78	83	100	100

Table 7.5: Per-word recognition results for the second user study, including all considered SoTA methods. We report indicative cases of failure (first five columns) and success (last five columns) of our approach compared to the other methods.

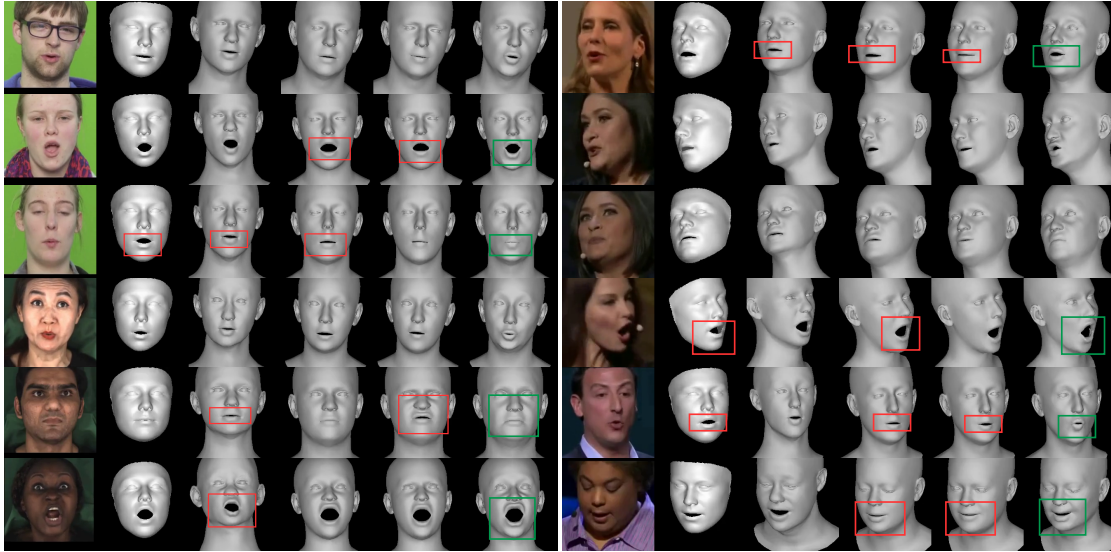


Figure 7.6: Visual comparison with other methods on the MEAD, TCDTIMIT, and LRS3 datasets. Note that our method is only trained on the LRS3 train test. From left to right: original footage, 3DDFAv2 [Guo et al., 2020], DAD [Martyniuk et al., 2022], DECA [Feng et al., 2021], EMOCA [Daněček et al., 2022], ours. We also highlight with red boxes some erroneous results, and with green boxes some examples of retaining the original mouth formation.

features from the output of the conformer. We present here an ablation study between these two features. For this ablation study, in order to study the immediate effect of different features, we directly optimize the initial estimation of DECA [Feng et al., 2021] expression ψ and jaw pose θ_{jaw} parameters using the lipread and regularization losses: $L = \lambda_{lr}L_{lr} + \lambda_{\psi}L_{\psi} + \lambda_{\theta_{jaw}}L_{\theta_{jaw}}$ where $\lambda_{lr} = 4$, $\lambda_{\psi} = 1e - 3$, and $\lambda_{\theta_{jaw}} = 200$. We avoided using the relaxed geometric loss from landmarks in this study in order to see the full effect of the different features.

The results of this ablation are in Figure 7.8. For each image sequence sample in the Figure we show in the top row the original footage, in the 2nd row the initial estimate of DECA [Feng et al., 2021], in the 3rd row the result of optimizing the lipread loss using Conformer features, and the last row optimizing the lipread loss using ResNet18 features. Although the conformer preserves useful information for the mouth area, there is not a strict visual correspondence with the original images, because the features are largely affected by the sequence context. On the other hand, features from ResNet18 retain the spatial structure and strict correspondence and are more suited to use for the perceptual lipread loss.

CTC loss and adversarial examples We also considered in our initial experiments leveraging text transcriptions and enforced a Connectionist Temporal Classification (CTC) [Graves et al., 2006] loss on the text prediction of the lipreader. However, we decided to neglect this choice, since it had several drawbacks. First of all, it has the obvious downsides of the increased computational overhead required to process whole sentences at once, as well as requirements of text transcriptions. In addition and more importantly, this loss did not retain any spatial structure, and also resulted to completely distorted facial reconstructions that achieved a perfect lip reading recognition - a common phenomenon found in adversarial attacks [Goodfellow et al., 2014, Akhtar and Mian, 2018]. We showcase

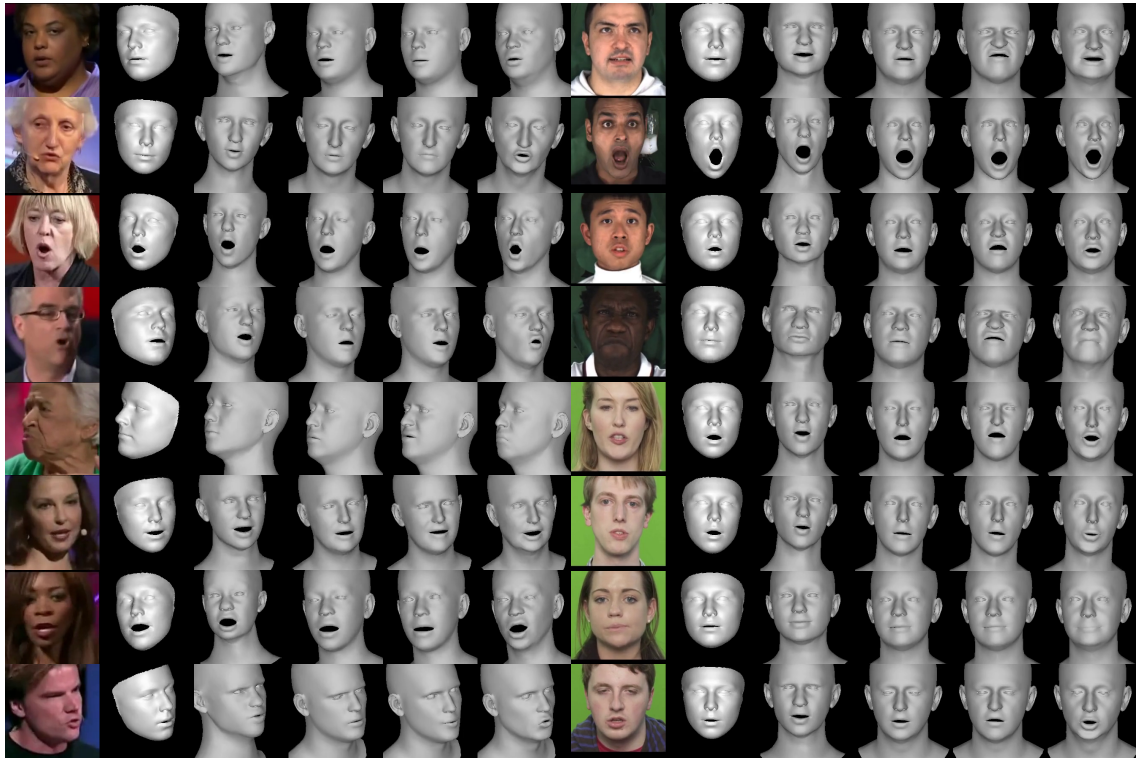


Figure 7.7: More visual results and comparisons with other methods on the LRS3, MEAD, and TCDTIMIT datasets. From left to right: original footage, 3DDFAv2 [Guo et al., 2020], DAD [Martyniuk et al., 2022], DECA [Feng et al., 2021], EMOCA [Daněček et al., 2022], SPECTRE.

this behavior in Figure 7.9.

Ablation on absolute vs relative mouth landmarks losses

In Fig. 7.10 we show an ablation example on using the absolute vs the relative (distances) losses for the geometric constraints. In this example, the left column shows the initial estimate of DECA, the middle column the predicted reconstruction of a model trained with an L_1 loss imposed on the mouth landmarks as well, and the third column a model trained with a more relaxed loss on the mouth using the intra-mouth distances. As it can be seen, strict landmark losses guide the result to resemble DECA. On the other hand, relative losses are less strict, and the model accurately predicts the mouth structure.

7.3.4 Failure Cases

Finally, we also include in Figure 7.11 two examples of erroneous mouth reconstructions of our model. In the first example there is an artifact in the mouth area, while in the second example, the reconstructed 3D shape has erroneously an open mouth. We believe that there are two major factors which can negatively affect our method. First, while our geometric relative constraints have greatly alleviated the domain gap problem in the perceptual losses, we can still find samples where this problem has created some minor artifacts. Second, since the perceptual loss itself originates from a neural network, failure cases of the lipread loss propagate to our 3D reconstruction model.

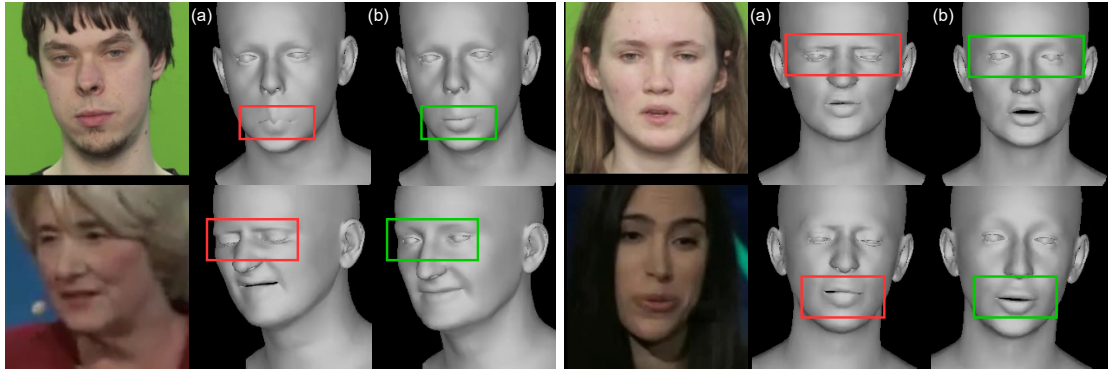


Figure 7.8: Training of the perceptual encoder without (a) and with (b) geometric constraints based on 2D landmarks. Omitting geometric constraints from the rest of the face leads to the emergence of artifacts in the eyes and nose in some cases, while completely omitting 2D information from mouth landmarks can lead to failure cases in the mouth area. Please zoom in for details.

7.4 Discussion

Our method has introduced a significant step towards creating truly realistic 3D talking heads, as it has been shown by our extensive objective and subjective evaluation against other SoTA methods. It is important to note that our method even outperforms DAD in terms of realism, which was trained with 3D annotated data on a large-scale dataset. Even though DAD was shown to achieve a geometrically accurate 3D shape, the lack of perceptual losses rendered the result less realistic, compared to SPECTRE. It should also be pointed out, as it is also evident in Figure 7.6 that the lipreading loss, not only retains the motions and shape of the mouth, but it also makes it more distinct in the rendered mesh. It becomes apparent that in order to achieve realism in terms of speech, we need to opt for more perceptual losses. This has also been done in previous methods regarding emotional expression [Daněček et al., 2022] as well as 3D shape [Feng et al., 2021, Zielonka et al., 2022]. Note also that training with our lipreading loss does not require text transcriptions or the corresponding audio.

Limitations We point out that the results of the objective evaluation on CER and WER, remain much higher compared to the original footage. This is of course, among others a problem of the different domain of the rendered images as compared to the ground truth. The absence of teeth and tongue is also important, since they play a large role in the detection of specific types of phonemes/visemes such as alveolar and dental consonants. This domain adaptation problem has not been addressed in this work, since our approach works well in practice, but it remains a hindrance to unleashing the full potential of the described losses. This domain problem also affects the perceptual losses. Both perceptual losses make the assumption that the original images and the rendered ones belong to the same visual “domain”. Nonetheless, there is indeed a realism/domain gap between these two feature spaces that may lead to inconsistencies; this is why we needed to have relative landmarks. As a result, the geometric loss and the lipreading loss sometimes compete against each other: on one hand, lip reading tries to improve the perception of the talking head while landmarks, if not detected accurately, tend to reduce the realism. On the other hand, we have observed that below a certain threshold, reduction of lip reading loss tends to create artifacts; which is why we need the constrains from landmarks to retain

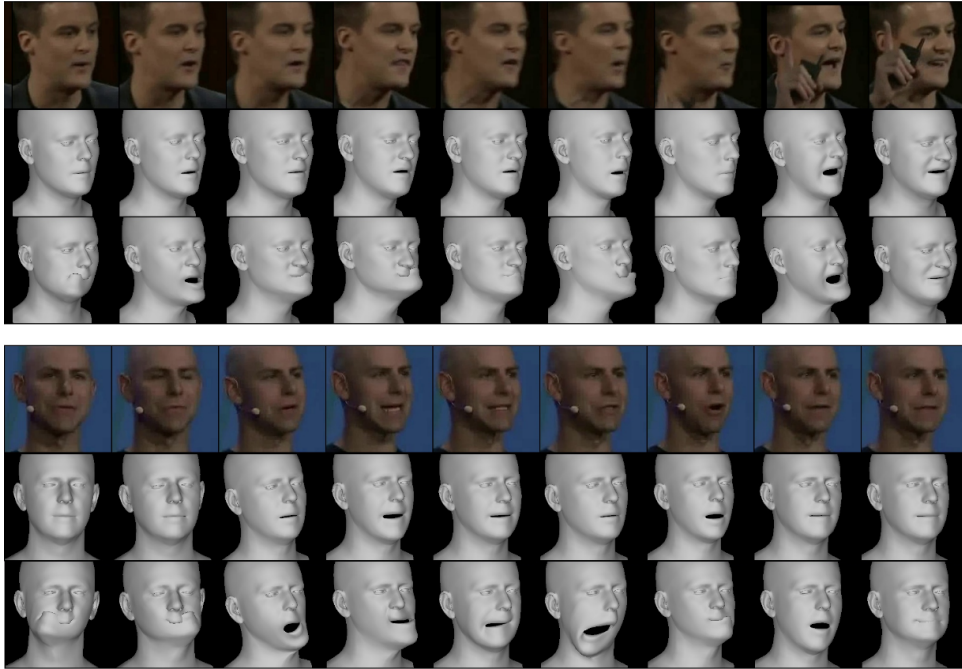


Figure 7.9: Two examples of adversarial attacks using the CTC loss. Middle row shows sampled frames from the original predicted sequence by DECA [Feng et al., 2021] of two sentences with starting CER (character error rate) around 0.90, while the third row shows completely distorted examples which however achieve near-perfect CER.

the realism of the facial shape. In addition, although our method includes a loss borrowed from EMOCA [Daněček et al., 2022], in order to retain the facial expressions outside the mouth (e.g. in eyes), since it was trained only on the LRS3 dataset (which does not include emotional samples) the results in some cases tend to not include the intensity of emotion present in EMOCA. Finally, while as we have already stated our method does not need text transcriptions or audio, we believe that these modalities, if present in the dataset, could be leveraged in order to improve the total perception, or, bypass problems such as visual occlusions.

7.5 Visual Speech-Informed Semantic Control of Facial Expressions

In this section we present Neural Emotion Director (NED) [Paraperas Papantoniou et al., 2022], a hybrid method for semantically controlling the facial expressions of a person in a source video *without altering the uttered speech*. Previous attempts at this problem have included the computationally burden procedure of having the person act the same speech in different emotions, allowing switching and blending between the different videos [Malleon et al., 2015]. Here, we build upon latest works in *image-to-image translation* [Isola et al., 2017, Zhu et al., 2017] (which have also been used in editing of facial expressions [Choi et al., 2018, d’Apolito et al., 2021]), and *face reenactment* [Kim et al., 2018, Doukas et al., 2021, Zakharov et al., 2019], and we propose a hybrid method which can alter the emotional state of actors in “in-the-wild” videos while retaining the original mouth motion (see Fig. 7.12). This can be done either by using a reference video of another person, or through semantic labels which correspond to the 6 basic emotions (angry, happy, surprise,

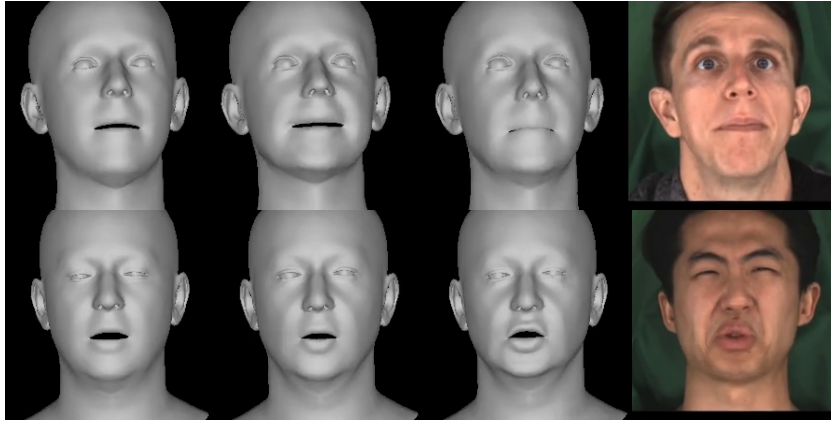


Figure 7.10: Ablation between using absolute position of mouth landmarks or relative intra-mouth distances. The first column is the initial estimate of DECA, the second column the predicted reconstruction of a model trained with an L_1 loss imposed on the mouth landmarks as well, and the third column a model trained with a more relaxed loss on the mouth using the intra-mouth distances. Strict mouth landmark losses erroneously guide the output to resemble DECA, while the relaxed constraints leave enough freedom to the perceptual loss to accurately capture the formation of lips.



Figure 7.11: Examples from failure cases of our model. The domain gap problem can still cause some mouth artifacts, even when guided by our geometric constraints. Note also that any failed results of the lipread network propagate to our 3D reconstruction method as well.

fear, disgust, sadness) plus neutral.

7.5.1 Method

Figure 7.13 presents the proposed method at inference time. As it can be seen, *Neural Emotion Director (NED)* consists of three modules: (*3D Face Analysis*, *3D-based Emotion Manipulator* and *Photo-realistic Synthesis “in the wild”*).

3D Face Analysis

The first module takes on the task of analyzing the input face and extracting 3D information. More specifically, we first apply face detection [Zhang et al., 2016] on the input video, and then segment the face using FSGAN [Nirkin et al., 2019]. Afterwards, we perform 3D face reconstruction using the latest state-of-the-art method of DECA [Feng et al., 2021]. Given an input image I DECA jointly predicts the identity parameters $\beta \in \mathbb{R}^{100}$, neck pose and jaw $\theta \in \mathbb{R}^6$, expression parameters $\psi \in \mathbb{R}^{50}$, albedo $\alpha \in \mathbb{R}^{50}$, lighting $\mathbf{I} \in \mathbb{R}^{27}$, and camera (scale and translation) $\mathbf{c} \in \mathbb{R}^3$. Note that these parameters are a subset of the



Figure 7.12: *Neural Emotion Director (NED)* can manipulate facial expressions in input videos while preserving speech, conditioned on either the semantic emotional label (top part of figure), or on an external reference style as extracted from a reference video (bottom part).

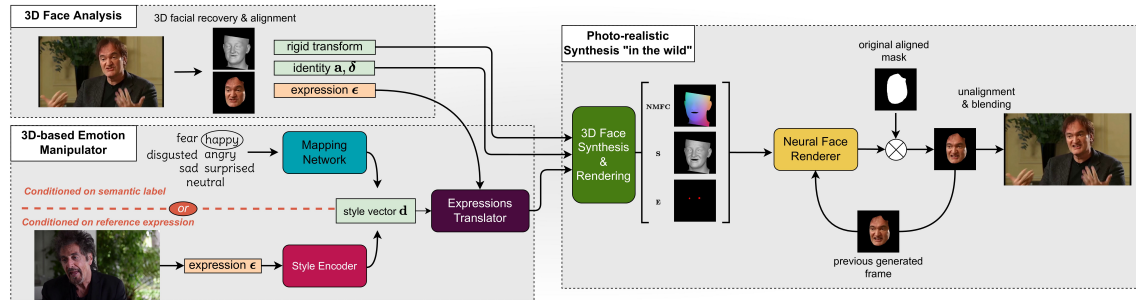


Figure 7.13: Overview of *Neural Emotion Director (NED)* at inference time. NED consists of three modules: first 3D Face Analysis is performed on the input video. Then, the extracted 3D parameters are translated to a target emotional domain using the *3D-based Emotion Manipulator*. Finally, a neural renderer is used in order to render the manipulated photo-realistic frames.

parameters of the FLAME 3D face model. Apart from the above coarse estimation, DECA also predicts a latent vector code, which is associated with a UV-displacement map, modeling high-frequency person-specific details such as wrinkles. In the next step in the 3D Face Analysis, we use FAN [Bulat and Tzimiropoulos, 2017] to obtain 68 facial landmarks for each frame, and estimate eye pupil coordinates [Doukas et al., 2021], in order to create eye images. Finally, all extracted faces are aligned using Procrustes analysis.

3D-based Emotion Manipulator

Upon completion of 3D Face Analysis, we obtain a sequence of vectors for each frame in the input sequence, describing the corresponding parameters of the FLAME model. Keeping all other parameters intact, we now wish to manipulate the 50 expression parameters ψ and 3 jaw pose parameters θ_{jaw} , in order to semantically switch the facial expression, albeit without altering the visual speech content. This is done through a four-submodule network, which is based on the StarGAN v2 [Choi et al., 2020] framework.

1. **Expressions Translator:** The translator G takes as input a sequence of expressions

\mathbf{s} and a style vector $\mathbf{d} \in \mathbb{R}^{16}$ and translates \mathbf{s} into an output sequence of expression vectors $G(\mathbf{s}, \mathbf{d}) \in \mathcal{S}$ that reflects the speaking style encoded in \mathbf{d} . To inject \mathbf{d} into G , we concatenate \mathbf{d} with each of the N vectors of the sequence.

2. **Style encoder:** Our style encoder E extracts the emotion-related style vector $\mathbf{d} = E(\mathbf{s})$ of an input sequence \mathbf{s} and, thus, enables the translator G to translate a given sequence according to the speaking style extracted from a reference sequence. In contrast to [Choi et al., 2020], our style encoder does not require any knowledge about the ground truth emotion label y of the reference sequence \mathbf{s} .
3. **Mapping network:** The mapping network M learns to generate style vectors $\mathbf{d} = M_y(\mathbf{z}) \in \mathbb{R}^{16}$ related to a target emotion $y \in \mathcal{Y}$, by transforming a latent code $\mathbf{z} \in \mathbb{R}^4$ sampled from a normal distribution. Here, $M_y(\cdot)$ denotes the output branch of M that corresponds to the emotion y . This network allows the translator to translate a sequence of expressions to a target emotion, by merely sampling random noise, and specifying the desired semantic emotion label.
4. **Expressions Discriminator:** Our discriminator D has $c = 7$ branches (similarly to M) and learns to discriminate between real \mathbf{s} and fake $G(\mathbf{s}, \mathbf{d})$ sequences of each domain y by outputting a scalar value $D_y(\mathbf{s})$ for each branch.

The network M follows a simple MLP architecture, whereas G, E and D use recurrent architectures with LSTM units [Hochreiter and Schmidhuber, 1997].

The 3D-based Emotion Manipulator is now trained using the following objective:

$$\mathcal{L}^{G,E,M} = \mathcal{L}_{adv}^G + \lambda_{sty} \mathcal{L}_{sty} + \lambda_{cyc} \mathcal{L}_{cyc} + \lambda_{mouth} \mathcal{L}_{mouth}$$

where L_{adv} is the LSGAN [Mao et al., 2017] adversarial criterion with $b = c = 1$ and $a = 0$, L_{sty} is the style reconstruction loss $\mathcal{L}_{sty} = \mathbb{E}_{\mathbf{s}, \tilde{\mathbf{d}}} [\|\tilde{\mathbf{d}} - E(G(\mathbf{s}, \tilde{\mathbf{d}}))\|_1]$, L_{cyc} is the cycle consistency loss [Zhu et al., 2017], and L_{mouth} is the speech preserving loss, which corresponds to the negative pearson correlation of the 1st jaw articulation parameter.

The *3D-based Emotion Manipulator* is trained on two video databases with annotations of the 6 basic emotions plus neutral: the Aff-Wild2 database [Kollias et al., 2019a, Kollias and Zafeiriou, 2019, Kollias et al., 2019b, Kollias et al., 2020, Kollias and Zafeiriou, 2021a, Kollias and Zafeiriou, 2021b, Kollias et al., 2021, Zafeiriou et al., 2017] of “in-the-wild” videos and the MEAD database [Wang et al., 2020a] (we exclude *contempt* for MEAD to match the emotions in Aff-Wild2). To get the best of the two databases, we pre-train our networks in Aff-Wild2 and then fine-tune them on a subset of MEAD.

3D Face Synthesis & Rendering

In the final module of the method, after translating the original 3D face parameters to a target expressive domain, we build a neural renderer which takes as input the concatenated 3D detailed rendered face, the *Normalized Mean Face Coordinate (NMFC)* image [Doukas et al., 2021], and the eye image \mathbf{E} . The neural renderer is built recurrently, and along with the previous inputs, the two previously generated images are fed into it, in order to enforce temporal consistency. Finally, after the photorealistic rendering of the 3D face, the face alignment is reversed and the face is blended with the original background.

Note, that as we will see in the experimental results, for each actor in the datasets a person-specific neural renderer has to be trained. However, we found out that in some cases the footage of the actor did not include sufficient information for the renderer to



Figure 7.14: Effect of the speech-preserving loss. Without this loss (middle row), the result does not preserve the mouth movement from the input video. In contrast, enforcing this loss (bottom row) successfully translates the expression of the actor to happy without altering his mouth movements and speech

reconstruct the expressions. To that end, we first trained a single meta-renderer, using a mix of the footage of the actors from all datasets, and then fine-tuned the meta-renderer in each actor separately. This way, in the first stage the neural renderer was able to learn a diverse range of facial expressions, and then retain them, while adding actor-specific details.

7.5.2 Experimental Results

We present objective and subjective experimental results of NED, comparing with the following state-of-the-art methods **GANmut** [d’Apolito et al., 2021], **ICface** [Tripathy et al., 2020] and **DSM** [Solanki and Roussos, 2021]:

Datasets

In order to assess the performance of NED, we use both a small in-house collected dataset from 6 YouTube videos that included facial videos of 6 actors during film scenes, TV shows and interviews, and the MEAD dataset [Wang et al., 2020a], from which we have chosen 3 actors. For every actor, we selected 30 videos for each of the 6 basic emotions (happy, angry, surprised, fear, sad, disgusted) plus neutral, resulting to a total of 630 videos from MEAD.

Quantitative Comparisons

The methods are assessed objectively using the emotional self-reenactment task [Paraperas Papantoniou et al., 2022], i.e., we translate each video of MEAD which corresponds to an emotional label to the same emotion and calculate the **Face Average Pixel Distance (FAPD)** and **Fréchet inception distance (FID)** [Heusel et al., 2017]. As it can be seen in Table 7.6, NED outperforms the baselines in both metrics overall and exhibits superior performance in almost all 7 emotions individually. This shows the higher realism of our synthesized videos as well as the better expression transferability in terms of identity preservation (see Fig. 7.15 for artifacts produced by GANmut and ICface).

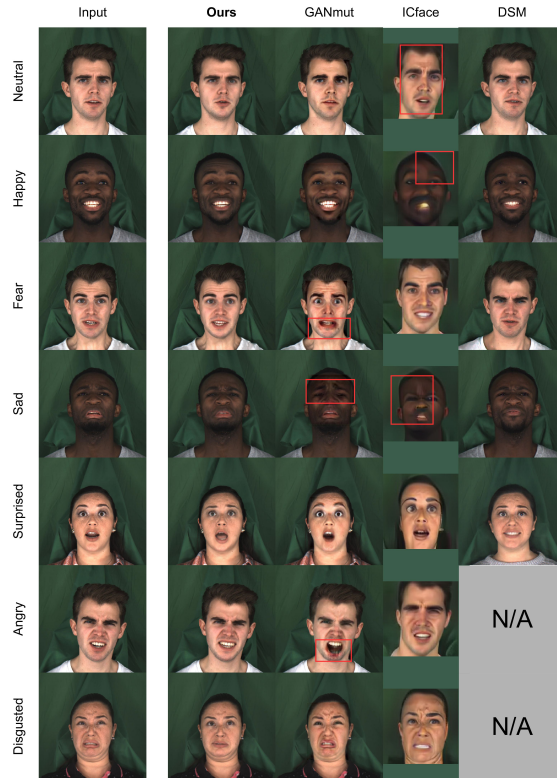


Figure 7.15: Visual comparison with state-of-the-art methods in the emotional “self-translation” experiment on the MEAD actors. Note that ICface [Tripathy et al., 2020] requires a tighter face cropping and padding with the background color has been used for visualization.

User Studies

We also include two web-based user studies for subjective comparison of NED against other methods:

Emotion Recognition and Realism on MEAD Database: In the first user study, participants were shown randomly shuffled manipulated videos of 3 actors from MEAD database in all 6 basic emotions and were asked to rate the realism of the footage on a Likert 5-point scale, as well as recognize the emotion shown (from a drop-down list including all 6 emotions). Apart from NED, the questionnaire showed to the participants manipulated videos by GANmut [d’Apolito et al., 2021], DSM [Solanki and Roussos, 2021], as well as the original real videos from MEAD. In total the questionnaire included 66 videos and 20 participants completed it. The results can be seen in Tab. 7.7 where we observe that all methods have relatively low realism scores. This can be attributed to the fact that the real videos in MEAD include particularly intense expressions, which probably resulted to an overall low frequency of ratings of 4 or 5 (even for real videos) and to an increased tendency to use these ratings (whenever they were used) more exclusively for real videos. However, we see that NED achieves significantly higher realism scores than the other methods, consistently across all 6 emotions. In terms of emotion recognition accuracy, we observe that our method synthesized videos with manipulated emotions that were consistently easier to be recognised by the participants, in comparison to DSM. However, this is not the case when we compare our method with GANmut: The synthetic videos

	NED		GANmut		DSM		ICface	
	FAPD	FID	FAPD	FID	FAPD	FID	FAPD	FID
neutral	14.9	2.1	16.8	2.9	22.1	5.5	40.0	45.8
happy	17.8	3.4	15.0	2.6	27.0	10.3	43.7	50.5
fear	18.4	3.0	20.5	4.3	28.0	8.5	43.0	46.6
sad	19.0	3.0	18.1	4.3	24.5	8.8	38.6	47.4
surprised	18.9	2.9	19.1	7.2	27.3	11.8	46.1	45.0
angry	18.4	3.0	22.4	4.2	-	-	51.7	53.5
disgusted	18.1	3.1	15.1	4.5	-	-	39.7	54.6
avg. (7)	17.9	2.9	18.1	4.3	-	-	43.3	49.0
avg. (5)	17.8	2.9	17.9	4.3	25.8	9.0	42.3	47.1

Table 7.6: Quantitative comparisons on MEAD in the emotional “self-translation” experiment. Bold values denote the best value for each metric (lower is better). Averaging is done over both the full set of 7 emotion labels and the set of 5 labels supported by DSM [Solanki and Roussos, 2021], for the sake of fair comparison.

of GANmut achieved a very high accuracy rate, which is even higher than the accuracy for real videos. This, in combination with the low realism score of GANmut, reflects the fact that GANmut synthesizes intense expressions that typically look fake but are easily recognizable.

Realism on YouTube Actors: In the second study, users were presented with manipulated videos (including the original audio) of 6 YouTube actors in all 6 basic emotions and asked them to rate the realism of the footage, following the same protocol as in the first study. We did not evaluate emotion recognition in this study since ground truth emotion annotations do not exist for this dataset (to compare with). The study included a random shuffling of videos manipulated by our method and GANmut, as well as the original videos. For some indicative frames of these videos, please refer to Fig. 7.15. DSM was not used for this study, since it cannot handle dynamic backgrounds such as those found in YouTube videos. The questionnaire included 54 videos in total and was completed by 50 participants. The ratings obtained can be seen in Tab. 7.8. We observe that the realism scores for both methods are relatively low, which can be attributed to the highly challenging task of manipulating the emotions in videos, especially under “in-the-wild” conditions as is the case for the YouTube Actors dataset. However, our method achieves a better score than GANmut and for example succeeds in synthesizing realistic videos more than 20% of the times for 3 out of 6 actors, which is a promising result that shows the potential of our approach. Furthermore, in terms of the most frequent rating, we see that our method is consistently better than GANmut, as it yields a rating of 3 or 2 as the most frequent answer for almost all actors, in contrast to GANmut that yields the rating of 1 as the most frequent.

7.6 Chapter Conclusions

We presented the first method for visual speech-informed perceptual reconstruction of 3D talking heads. Our method does not rely on text transcriptions or audio; on the contrary we employ a “lipreading” loss, in order to increase the perception of mouth. Our subjective and objective evaluations have verified that the results of 3D reconstruction are significantly preferred to counterpart methods which rely only on geometric losses for the mouth movements, as well as to methods that use direct 3D supervision. This is an important step towards reconstructing truly realistic talking heads, by focusing not only on

	Realism				Accuracy			
	NED	GANmut	DSM	Real Videos	Ours	GANmut	DSM	Real Videos
happy	17%	3%	8%	80%	63%	90%	42%	90%
fear	32%	7%	10%	67%	33%	75%	13%	25%
sad	30%	18%	12%	55%	13%	78%	25%	65%
surprised	22%	8%	7%	82%	17%	82%	5%	82%
angry	25%	10%	-	78%	50%	98%	-	80%
disgusted	40%	20%	-	67%	33%	40%	-	60%
avg.	28%	11%	9%	71%	35%	77%	21%	67%

Table 7.7: Realism ratings (percentage of users that rated the videos with 4 or 5) and classification accuracy of the user study on MEAD.

	NED						GANmut						Real Videos					
	1	2	3	4	5	'real'	1	2	3	4	5	'real'	1	2	3	4	5	'real'
McDormand	32	32	52	21	13	23%	59	23	16	28	24	35%	0	2	3	21	124	97%
Pacino	19	45	53	25	8	22%	40	41	26	24	19	29%	0	3	6	29	112	94%
Tarantino	70	29	23	17	11	19%	72	20	26	19	13	21%	1	5	9	43	92	90%
McConaughey	37	63	33	13	4	11%	88	43	12	7	0	5%	0	4	18	33	93	85%
Roberts	34	60	39	12	5	11%	88	27	17	13	5	12%	0	0	3	24	123	98%
Foxx	26	35	39	34	15	33%	79	43	18	6	4	7%	0	0	7	31	111	95%
avg.	36	44	40	20	9	20%	71	33	19	16	13	18%	1	4	8	30	109	93%

Table 7.8: Realism ratings of the user study on 6 YouTube actors. Columns 1-5 show the number of times that users gave this rating. The column “real” shows the percentage of users that rated the videos with 4 or 5. Bold values denote the most frequent user rating for each method and actor.

the purely geometric-based aspects, but also on human perception of speech articulation. We also presented NED, a novel method for photo-realistic manipulation of emotion of actors in videos. Compared to previous alternative NED retains the speech-related context of facial expressions and faithfully synthesizes the target actor’s face and composites it onto the original video. Note that NED is one of the use cases where SPECTRE could be applied as a 3D analysis module.

8

Contributions & Future Directions

In this Ph.D. dissertation we have studied aspects of “affective computing” under the umbrella of human-machine interaction. More precisely, we have decoupled our research into two interwoven parts, which constitute the “communication loop” we established in the introductory chapter.

8.1 Contributions

8.1.1 Emotion Recognition

In the first part, “Emotion Recognition”, our contributions can be summarised as follows:

1. We introduced and studied multiple streams of information which can be used in order to improve emotion recognition systems. These included both explicit (body, face, speech) and implicit (emotional embeddings, context) channels. We also showed that different representations of the same underlying source can act beneficially in cases, by considering the problem in both its static and dynamic nature.
2. We designed and developed multi-stream deep neural network based architectures and studied the optimal fusion of the different streams, in order to achieve most robust results and increase the performance. In addition, we considered social robotics applications, and resorted to more lightweight architectures that can be used to build real-time systems.
3. We collected and made available the BabyRobot Emotion Dataset, a medium-sized dataset of children performing emotions while interacting with two different robots. While analyzing the dataset we also discovered patterns of bodily expression of emotion, and how children express their emotion with facial expressions and/or body expressions.

8.1.2 Expression Synthesis

Next we summarize our contributions on the second part, “Expression Synthesis”:

1. We studied audio-visual synthesis of speech and proposed methods that grants such systems emotional expression capabilities. In addition, considering the limited amount of publicly available datasets for expressive audiovisual synthesis, we introduced emotional interpolation and adaptation with limited data, that can transform a neutral audiovisual speech synthesis system into an expressive one.

2. We proposed two deep neural network based expressive audiovisual speech synthesis systems which significantly outperformed traditional HMM-based and Unit Selection alternatives. This result was confirmed through exhaustive subjective evaluations and web user studies.
3. We recorded and developed the CVSP-EAV dataset, which includes a total of 3,600 emotional sentences in four emotions in Greek. This is the first publicly available dataset for expressive audiovisual speech synthesis in Greek.
4. We designed and implemented SPECTRE, the first method for perceptual 3D reconstruction of facial expression focusing on preserving visual speech without the need for text transcriptions or audio. The method used our devised “lip read loss”, which guides the 3D fitting process so that the reconstructed facial expressions and most importantly the mouth, retain the speech perception of the original footage. SPECTRE can be employed as a module in a variety of applications, an example of which was shown in the novel NED method for manipulation of facial expressions while preserving visual speech.

8.2 Future Directions

This thesis opens up a number of different directions, both for affect recognition and expression synthesis, as well as for affective computing in general. A straightforward application of the two presented directions (recognition and synthesis) is to merge them into a complete system, that implements the communication loop between computers and humans. Nevertheless, we can also highlight multiple important research directions for each different part.

8.2.1 Emotion Recognition

The culmination of the various modalities and representations that were studied in the first part of the thesis would be a system that leverages information from all the modalities and information streams we have explored: facial expression, body posture and movements, context and scene characteristics, speech, and emotional embeddings. Training such a system would require a large scale dataset annotated on all channels. Furthermore, fusing information from all these streams is a non-trivial problem that requires extensive experimental analysis. Nevertheless, the resulting system would, without doubt, achieve high performance of emotion recognition, tackling the challenges of human-robot interaction and emotion recognition in the wild.

A second important direction is the addition of a final stream of information we have not considered. This stream is the textual information derived from the speech of the user and includes valuable information, considering the recent interest in the creation of extensive affective vocabularies [Malandrakis et al., 2011, Carvalho et al., 2018]. Analysis of affect from textual information (also known in the literature as “sentiment analysis”) is an exciting field with a vast amount of possible applications (e.g., social media monitoring, customer feedback, chatbots, call centers, etc.).

8.2.2 Expression Synthesis

In Chapter 7 we proposed SPECTRE, the first method for perceptual 3D reconstruction of facial expressions focusing on retaining visual speech information. This method can serve

as a module for multiple affective computing applications. Currently, creating hyper-realistic 3D talking heads and avatars requires time consuming stereo 3D recordings, in order to effectively capture the intricacies of speech. Through our method, it is possible to bypass the cumbersome 3D data collection and directly acquire accurate visual speech-aware 3D reconstructions. As a result, a direct step forward would be to apply SPECTRE on existing expressive audiovisual speech datasets (such as MEAD [Wang et al., 2020a]) in order to acquire ground truth data, and then use these in order to create an emotional audio-to-visual talking head. Similarly, this method can be used to also create text-to-audiovisual avatars, following the methods established in Chapters 5 and 6. Finally, SPECTRE can be used to improve existing facial expression manipulation systems such as the one we showed in Chapter 7.

Bibliography

- [Abdelaziz et al., 2021] Abdelaziz, A. H., Kumar, A. P., Seivwright, C., Fanelli, G., Binder, J., Stylianou, Y., and Kajarekar, S. (2021). Audiovisual Speech Synthesis using Tacotron2. In *Proc. ICMI*.
- [Abrevaya et al., 2019] Abrevaya, V. F., Boukhayma, A., Wuhrer, S., and Boyer, E. (2019). A decoupled 3D facial shape model by adversarial training. In *Proc. ICCV*.
- [Adler et al., 2020] Adler, D. A., Ben-Zeev, D., Tseng, V. W., Kane, J. M., Brian, R., Campbell, A. T., Hauser, M., Scherer, E. A., and Choudhury, T. (2020). Predicting early warning signs of psychotic relapse from passive sensing data: an approach using encoder-decoder neural networks. *JMIR*.
- [Afouras et al., 2018] Afouras, T., Chung, J. S., and Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv:1809.00496*.
- [Akhtar and Mian, 2018] Akhtar, N. and Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*.
- [Al-Nuaimi et al., 2017] Al-Nuaimi, A. H., Jammeh, E., Sun, L., and Ifeachor, E. (2017). Higuchi fractal dimension of the electroencephalogram as a biomarker for early detection of Alzheimer’s disease. In *Proc. EMBC*.
- [Aldeneh et al., 2022] Aldeneh, Z., Fedzechkina, M., Seto, S., Metcalf, K., Sarabia, M., Apostoloff, N., and Theobald, B.-J. (2022). Towards a Perceptual Model for Estimating the Quality of Visual Speech.
- [Aldrian and Smith, 2012] Aldrian, O. and Smith, W. A. (2012). Inverse rendering of faces with a 3D morphable model. *IEEE Trans. Pattern Analysis Machine Intelligence*.
- [Allen, 2007] Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*.
- [Amazon, 2015] Amazon (2015). Amazon Polly Developer Guide.
- [Ambady and Rosenthal, 1992] Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychol. Bull.*
- [Anderson et al., 2013] Anderson, R. et al. (2013). Expressive Visual Text-to-Speech Using Active Appearance Models. In *Proc. CVPR*.
- [Atkinson et al., 2004] Atkinson, A. P., Dittrich, W. H., Gemmell, A. J., and Young, A. W. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*.

- [Aung et al., 2017] Aung, M. H., Matthews, M., and Choudhury, T. (2017). Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies. *Depression and Anxiety*.
- [Aviezer et al., 2008] Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., and Bentin, S. (2008). Angry, disgusted, or afraid? Studies on the malleability of emotion perception. *Psychological science*.
- [Aviezer et al., 2012] Aviezer, H., Trope, Y., and Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*.
- [Bagautdinov et al., 2018] Bagautdinov, T., Wu, C., Saragih, J., Fua, P., and Sheikh, Y. (2018). Modeling facial geometry using compositional vaes. In *Proc. CVPR*.
- [Bai et al., 2018] Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*.
- [Bailly et al., 2003] Bailly, G., Bérar, M., Elisei, F., and Odisio, M. (2003). Audiovisual speech synthesis. *Intl. J. Speech Technol.*
- [Baltrusaitis et al., 2018] Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. In *Proc. FG*.
- [Bänziger et al., 2006] Bänziger, T., Pirker, H., and Scherer, K. (2006). GEMEP-GENEVA Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions. In *Proc. LREC*.
- [Bänziger et al., 2012] Bänziger, T., Mortillaro, M., and Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion*.
- [Bao et al., 2021] Bao, L., Lin, X., Chen, Y., Zhang, H., Wang, S., Zhe, X., Kang, D., Huang, H., Jiang, X., Wang, J., Yu, D., and Zhang, Z. (2021). High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics*.
- [Barnett et al., 2018] Barnett, I., Torous, J., Staples, P., et al. (2018). Relapse Prediction in Schizophrenia through Digital Phenotyping: A Pilot Study. *Neuropsychopharmacology*.
- [Barrett and Kensinger, 2010] Barrett, L. F. and Kensinger, E. A. (2010). Context is routinely encoded during emotion perception. *Psychological science*.
- [Barros et al., 2015] Barros, P., Jirak, D., Weber, C., and Wermter, S. (2015). Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks*.
- [Basu et al., 1998a] Basu, S., Oliver, N., and Pentland, A. (1998a). 3D lip shapes from video: A combined physical–statistical model. *Speech Commun.*
- [Basu et al., 1998b] Basu, S., Oliver, N., and Pentland, A. (1998b). 3D modeling and tracking of human lip motions. In *Proc. ECCV*.
- [Bates, 1994] Bates, J. (1994). The role of emotion in believable agents. *Communic. ACM*.

- [Bechara, 2004] Bechara, A. (2004). The role of emotion in decision-making: evidence from neurological patients with orbitofrontal damage. *Brain Cogn.*
- [Belpaeme et al., 2013] Belpaeme, T., Baxter, P., De Greeff, J., Kennedy, J., Read, R., Looije, R., Neerinx, M., Baroni, I., and Zelati, M. C. (2013). Child-robot interaction: Perspectives and challenges. In *Proc. ICSR*.
- [Belpaeme et al., 2018] Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: A review. *Science Robotics*.
- [Ben-Zeev et al., 2017] Ben-Zeev, D., Brian, R., Wang, M., et al. (2017). CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric Rehabilitation J.*
- [Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. the Royal statistical society: Series B (Methodological)*.
- [Beskow, 1996] Beskow, J. (1996). Talking Heads – Communication, Articulation and Animation. In *Proc. Fonetik*.
- [Bickmore and Picard, 2005] Bickmore, T. W. and Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Trans. Computer-Human Interaction*.
- [Black, 2003] Black, A. W. (2003). Unit selection and emotional speech. In *Proc. Inter-speech*.
- [Blanz and Vetter, 1999] Blanz, V. and Vetter, T. (1999). A Morphable Model for the Synthesis of 3D Faces. In *Proc. CGIT*.
- [Boletsis et al., 2015] Boletsis, C., McCallum, S., and Landmark, B. F. (2015). The use of smartwatches for health monitoring in home-based dementia care. In *Proc. Int’l Conf. Human Aspects IT Aged Population*.
- [Booth et al., 2018a] Booth, J., Roussos, A., Ververas, E., Antonakos, E., Ploumpis, S., Panagakis, Y., and Zafeiriou, S. (2018a). 3D reconstruction of “in-the-wild” faces in images and videos. *IEEE Trans. Pattern Analysis Machine Intelligence*.
- [Booth et al., 2018b] Booth, J., Roussos, A., Ponniah, A., Dunaway, D., and Zafeiriou, S. (2018b). Large scale 3d morphable models. *Intl. J. Computer Vision*.
- [Brennan et al., 2001] Brennan, M., Palaniswami, M., and Kamen, P. (2001). Do existing measures of Poincare plot geometry reflect nonlinear features of heart rate variability? *IEEE Trans. Biomedical Engineering*.
- [Broadbent et al., 2009] Broadbent, E., Stafford, R., and MacDonald, B. (2009). Acceptance of healthcare robots for the older population: Review and future directions. *Intl. J. Social Robotics*.
- [Bulat and Tzimiropoulos, 2017] Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proc. ICCV*.

- [Calvo et al., 2014] Calvo, R., D’Mello, S., Gratch, J., Kappas, A., Lhommet, M., and Marsella, S. C. (2014). *Expressing Emotion Through Posture and Gesture*.
- [Calvo et al., 2015] Calvo, R. A., D’Mello, S., Gratch, J., and Kappas, A. (2015). *The Oxford Handbook of Affective Computing*.
- [Cao et al., 2013] Cao, C., Weng, Y., Zhou, S., Tong, Y., and Zhou, K. (2013). Facewarehouse: A 3d facial expression database for visual computing. *IEEE Trans. Visualization and Computer Graphics*.
- [Cao et al., 2015] Cao, C., Bradley, D., Zhou, K., and Beeler, T. (2015). Real-time high-fidelity facial performance capture. *ACM Trans. Graphics*.
- [Cao et al., 2005] Cao, Y., Tien, W. C., Faloutsos, P., and Pighin, F. (2005). Expressive speech-driven facial animation. *ACM Trans. Graph.*
- [Cao et al., 2017] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proc. CVPR*.
- [Cao et al., 2019] Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Trans. Pattern Analysis Machine Intelligence*.
- [Carroll and Russell, 1996] Carroll, J. M. and Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *J. personality and social psychology*.
- [Carvalho et al., 2018] Carvalho, F., Santos, G., and Guedes, G. P. (2018). AffectPT-br: an Affective Lexicon based on LIWC 2015. In *Proc. SCCC*.
- [Castellano et al., 2008] Castellano, G., Kessous, L., and Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In *Peter C., Beale R. (eds) Affect and Emotion in Human-Computer Interaction*.
- [Castellano et al., 2013] Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., and McOwan, P. W. (2013). Multimodal affect modeling and recognition for empathic robot companions. *Intl. J. of Humanoid Robotics*.
- [Cella et al., 2018] Cella, M., Okruszek, L., Lawrence, M., Zarlenga, V., He, Z., and Wykes, T. (2018). Using wearable technology to detect the autonomic signature of illness severity in schizophrenia. *Schizophrenia Research*.
- [Chai et al., 2022] Chai, Z., Zhang, H., Ren, J., Kang, D., Xu, Z., Zhe, X., Yuan, C., and Bao, L. (2022). Realy: Rethinking the evaluation of 3d face reconstruction. In *Proc. ECCV*.
- [Chalamandaris et al., 2013] Chalamandaris, A., Tsiakoulis, P., Karabetsos, S., Raptis, S., and LTD, I. (2013). The ILSP/INNOETICS text-to-speech system for the Blizzard Challenge 2013. In *Proc. Blizzard Challenge*.
- [Chang et al., 2018] Chang, F.-J., Tran, A. T., Hassner, T., Masi, I., Nevatia, R., and Medioni, G. (2018). Expnet: Landmark-free, deep, 3d facial expressions. In *Proc. FG*.
- [Chen et al., 2017] Chen, S.-Y., Feng, Z., and Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *J. Thoracic Disease*.

- [Cheng and Huang, 2010] Cheng, J. and Huang, P. (2010). Real-time mouth tracking and 3d reconstruction. In *Intl. Congress Image Signal Processing*.
- [Cheng et al., 2019] Cheng, S., Bronstein, M., Zhou, Y., Kotsia, I., Pantic, M., and Zafeiriou, S. (2019). Meshgan: Non-linear 3d morphable models of faces. *arXiv:1903.10384*.
- [Choi et al., 2018] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proc. CVPR*.
- [Choi et al., 2020] Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. (2020). StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *Proc. CVPR*.
- [Chung et al., 2017] Chung, J. S., Jamaludin, A., and Zisserman, A. (2017). You said that? *arXiv:1705.02966*.
- [Clark et al., 2007] Clark, R. A., Podsiadlo, M., Fraser, M., Mayo, C., and King, S. (2007). Statistical analysis of the Blizzard Challenge 2007 listening test results. In *Proc. ISCA SSW6*.
- [Cootes et al., 2001] Cootes, T. F., Edwards, G. J., Taylor, C. J., et al. (2001). Active appearance models. *IEEE Trans. Pattern Analysis Machine Intelligence*.
- [Corrigan et al., 1990] Corrigan, P. W., Liberman, R. P., and Engel, J. D. (1990). From noncompliance to collaboration in the treatment of schizophrenia. *Psychiatric Services*.
- [Cosatto et al., 2000] Cosatto, E., Potamianos, G., and Graf, H. P. (2000). Audio-visual unit selection for the synthesis of photo-realistic talking-heads. In *Proc. ICME*.
- [Cowie et al., 2001] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*.
- [Cudeiro et al., 2019a] Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., and Black, M. J. (2019a). Capture, Learning, and Synthesis of 3D Speaking Styles. In *Proc. CVPR*.
- [Cudeiro et al., 2019b] Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., and Black, M. J. (2019b). Capture, learning, and synthesis of 3d speaking styles. In *Proc. CVPR*.
- [Dael et al., 2012a] Dael, N., Mortillaro, M., and Scherer, K. R. (2012a). Emotion expression in body action and posture. *Emotion*.
- [Dael et al., 2012b] Dael, N., Mortillaro, M., and Scherer, K. R. (2012b). The body action and posture coding system (BAP): Development and reliability. *J. Nonverbal Behavior*.
- [Dahmani et al., 2019] Dahmani, S., Colotte, V., Girard, V., and Ouni, S. (2019). Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis. In *Proc. Interspeech*.
- [Daněček et al., 2022] Daněček, R., Black, M. J., and Bolkart, T. (2022). EMOCA: Emotion Driven Monocular Face Capture and Animation. In *Proc. CVPR*.
- [d’Apolito et al., 2021] d’Apolito, S., Paudel, D. P., Huang, Z., Romero, A., and Van Gool, L. (2021). GANmut: Learning Interpretable Conditional Space for Gamut of Emotions. In *Proc. CVPR*.

- [Darwin, 1871] Darwin, C. (1871). *The Expression of the Emotions in Man and Animals*.
- [De Gelder, 2009] De Gelder, B. (2009). Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Trans. of the Royal Society of London B: Biological Sciences*.
- [De Silva, 2004] De Silva, L. C. (2004). Audiovisual emotion recognition. In *Proc. SMC*.
- [Deng et al., 2005] Deng, H.-B., Jin, L.-W., Zhen, L.-X., Huang, J.-C., et al. (2005). A new facial expression recognition method based on local Gabor filter bank and PCA plus LDA. *Intl. J. Information Technology*.
- [Deng et al., 2019] Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., and Tong, X. (2019). Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set. In *Proc. CVPR*.
- [Deng et al., 2006] Deng, Z., Neumann, U., Lewis, J. P., Kim, T.-Y., Bulut, M., and Narayanan, S. (2006). Expressive facial animation synthesis by learning speech coarticulation and expression spaces. *IEEE Trans. Vis. Comput. Graphics*.
- [Dey and Boddeti, 2022] Dey, R. and Boddeti, V. N. (2022). Generating Diverse 3D Reconstructions from a Single Occluded Face Image. In *Proc. CVPR*.
- [Digalakis et al., 2003] Digalakis, V., Oikonomidis, D., Pratsolis, D., Tsourakis, N., Vounidis, C., Chatzichrisafis, N., and Diakouloukas, V. (2003). Large vocabulary continuous speech recognition in greek: corpus and an automatic dictation system. In *Proc. Interspeech*.
- [Digalakis and Neumeyer, 1996] Digalakis, V. V. and Neumeyer, L. G. (1996). Speaker adaptation using combined transformation and Bayesian methods. *IEEE Trans. Speech Audio Process*.
- [Dong et al., 2016] Dong, J., Li, X., and Snoek, C. G. (2016). Word2visualvec: Image and video to sentence matching by visual feature prediction. *arXiv:1604.06838*.
- [Doukas et al., 2021] Doukas, M. C., Koujan, M. R., Sharmanska, V., Roussos, A., and Zafeiriou, S. (2021). Head2Head++: Deep Facial Attributes Re-Targeting. *IEEE Trans. Biometrics, Behavior, and Identity Science*.
- [Egger et al., 2020] Egger, B., Smith, W. A. P., Tewari, A., Wuhler, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., Theobalt, C., Blanz, V., and Vetter, T. (2020). 3D Morphable Face Models—Past, Present, and Future. *ACM Trans. Graph.*
- [Ekman and Friesen, 1967] Ekman, P. and Friesen, W. V. (1967). Head and body cues in the judgment of emotion: A reformulation. *Perceptual and Motor Skills*.
- [Ekman et al., 1969] Ekman, P., Sorenson, E. R., and Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*.
- [Ekman and Friesen, 1971] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *J. of Personality and Social Psychology*.
- [Ekman et al., 1980] Ekman, P., Friesen, W. V., and Ancoli, S. (1980). Facial signs of emotional experience. *J. Personal. Soc. Psychol.*

- [Ekman, 1984] Ekman, P. (1984). Expression and the nature of emotion. *Approaches to emotion*.
- [Ekman, 1992] Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*.
- [El Ayadi et al., 2011] El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*.
- [Eyben et al., 2015] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affective Computing*.
- [Ezzat et al., 2002] Ezzat, T., Geiger, G., and Poggio, T. (2002). Trainable videorealistic speech animation. In *Proc. SIGGRAPH*.
- [Fabian Benitez-Quiroz et al., 2016] Fabian Benitez-Quiroz, C., Srinivasan, R., and Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proc. CVPR*.
- [Falconer, 2004] Falconer, K. (2004). *Fractal geometry: mathematical foundations and applications*.
- [Fan et al., 2015] Fan, B., Xie, L., Yang, S., Wang, L., and Soong, F. K. (2015). A deep bidirectional LSTM approach for video-realistic talking head. *Multimed. Tools Appl.*
- [Feng et al., 2018] Feng, Y., Wu, F., Shao, X., Wang, Y., and Zhou, X. (2018). Joint 3d face reconstruction and dense alignment with position map regression network. In *Proc. ECCV*.
- [Feng et al., 2021] Feng, Y., Feng, H., Black, M. J., and Bolkart, T. (2021). Learning an animatable detailed 3D face model from in-the-wild images. *ACM Trans. Graphics*.
- [Filippini et al., 2020] Filippini, C., Spadolini, E., Cardone, D., Bianchi, D., Preziuso, M., Sciarretta, C., del Cimmuto, V., Lisciani, D., and Merla, A. (2020). Facilitating the Child–Robot Interaction by Endowing the Robot with the Capability of Understanding the Child Engagement: The Case of Mio Amico Robot. *Intl. J. Social Robotics*.
- [Filntisis et al., 2019] Filntisis, P. P., Efthymiou, N., Koutras, P., Potamianos, G., and Maragos, P. (2019). Fusing Body Posture With Facial Expressions for Joint Recognition of Affect in Child–Robot Interaction. *IEEE Robotics Automation Letters*.
- [Fleiss and Cohen, 1973] Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*.
- [Friesen and Ekman, 1978] Friesen, E. and Ekman, P. (1978). *Facial action coding system: a technique for the measurement of facial movement*.
- [Friesen et al., 1983] Friesen, W. V., Ekman, P., et al. (1983). EMFACS-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*.

- [Frome et al., 2013] Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. (2013). DEVISe: A deep visual-semantic embedding model. In *Proc. NeurIPS*.
- [Furhat, 2022] Furhat (2022). Furhat Robotics. <http://furhatrobotics.com>.
- [Gaebel et al., 1993] Gaebel, W., Frick, U., Köpcke, W., Linden, M., Müller, P., Müller-Spahn, F., Pietzcker, A., and Tegeler, J. (1993). Early neuroleptic intervention in schizophrenia: are prodromal symptoms valid predictors of relapse? *The British J. Psychiatry*.
- [Gales, 1998] Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Comput. Speech & Lang.*
- [Garofolo et al., 1993] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon. Technical Report*.
- [Garrido et al., 2016a] Garrido, P., Zollhöfer, M., Wu, C., Bradley, D., Pérez, P., Beeler, T., and Theobalt, C. (2016a). Corrective 3D reconstruction of lips from monocular video. *ACM Trans. Graph.*
- [Garrido et al., 2016b] Garrido, P., Zollhöfer, M., Casas, D., Valgaerts, L., Varanasi, K., Pérez, P., and Theobalt, C. (2016b). Reconstruction of personalized 3D face rigs from monocular video. *ACM Trans. Graphics*.
- [Gecer et al., 2019] Gecer, B., Ploumpis, S., Kotsia, I., and Zafeiriou, S. (2019). Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proc. CVPR*.
- [Gerig et al., 2018] Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schoenborn, S., and Vetter, T. (2018). Morphable Face Models - An Open Framework. In *Proc. FG*.
- [Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Proc. NeurIPS*.
- [Goulart et al., 2019] Goulart, C., Valadao, C., Delisle-Rodriguez, D., Funayama, D., Favarato, A., Baldo, G., Binotte, V., Caldeira, E., and Bastos-Filho, T. (2019). Visual and Thermal Image Processing for Facial Specific Landmark Detection to Infer Emotions in a Child-Robot Interaction. *Sensors*.
- [Grassal et al., 2022] Grassal, P.-W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., and Thies, J. (2022). Neural Head Avatars from Monocular RGB Videos. *Proc. CVPR*.
- [Graves et al., 2006] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*.
- [Güler et al., 2018] Güler, R. A., Neverova, N., and Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. In *Proc. CVPR*.

- [Gunes and Piccardi, 2006] Gunes, H. and Piccardi, M. (2006). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *Proc. ICPR*.
- [Gunes and Piccardi, 2009] Gunes, H. and Piccardi, M. (2009). Automatic temporal segment detection and affect recognition from face and body display. *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)*.
- [Gunes and Pantic, 2010] Gunes, H. and Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. *Intl. J. Synthetic Emotions*.
- [Gunes and Schuller, 2013] Gunes, H. and Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*.
- [Guo et al., 2020] Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., and Li, S. Z. (2020). Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proc. ECCV*.
- [Harte and Gillen, 2015] Harte, N. and Gillen, E. (2015). TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Trans. Multimedia*.
- [Hatfield et al., 1993] Hatfield, E., Cacioppo, J. T., and Rapson, R. L. (1993). Emotional contagion. *Curr. Dir. Psychol. Sci.*
- [He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proc. CVPR*.
- [He et al., 2006] He, L., Zou, C., Zhao, L., and Hu, D. (2006). An enhanced LBP feature based on facial expression recognition. In *Proc. EMBC*.
- [Heiga et al., 2007] Heiga, Z., Tokuda, K., Masuko, T., Kobayasih, T., and Kitamura, T. (2007). A hidden semi-Markov model-based speech synthesis system. *IEICE Trans. Inf. Syst.*
- [Hess et al., 1997] Hess, U., Blairy, S., and Kleck, R. E. (1997). The intensity of emotional facial expressions and decoding accuracy. *J. Nonverbal Behav.*
- [Heusel et al., 2017] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. NeurIPS*.
- [Higuchi, 1988] Higuchi, T. (1988). Approach to an irregular time series on the basis of the fractal theory. *Physica D*.
- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*.
- [Hofer et al., 2020] Hofer, M., Hartmann, T., Eden, A., Ratan, R., and Hahn, L. (2020). The role of plausibility in the experience of spatial presence in virtual environments. *Frontiers in Virtual Reality*.
- [Huang et al., 2002] Huang, F. J., Cosatto, E., and Graf, H. P. (2002). Triphone based unit selection for concatenative visual speech synthesis. In *Proc. ICASSP*.

- [Huber et al., 2016] Huber, P., Kopp, P., Christmas, W., Räscher, M., and Kittler, J. (2016). Real-time 3D face fitting and texture fusion on in-the-wild videos. *IEEE Signal Processing Letters*.
- [Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. In *Proc. CVPR*.
- [Jackson et al., 2017] Jackson, A. S., Bulat, A., Argyriou, V., and Tzimiropoulos, G. (2017). Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proc. ICCV*.
- [Jourabloo and Liu, 2016] Jourabloo, A. and Liu, X. (2016). Large-pose face alignment via CNN-based dense 3D model fitting. In *Proc. CVPR*.
- [Jung et al., 2015] Jung, H., Lee, S., Yim, J., Park, S., and Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proc. CVPR*.
- [Kalberer and Gool, 2000] Kalberer, G. A. and Gool, L. J. V. (2000). Lip animation based on observed 3D speech dynamics. In *Videometrics and Optical Methods for 3D Shape Measurement*.
- [Karg et al., 2013] Karg, M., Samadani, A., Gorbet, R., Kühnlenz, K., Hoey, J., and Kulić, D. (2013). Body Movements for Affective Expression: A Survey of Automatic Recognition and Generation. *IEEE Trans. Affective Computing*.
- [Karras et al., 2017] Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graphics*.
- [Katsamanis et al., 2011] Katsamanis, A., Black, M., Georgiou, P. G., Goldstein, L., and Narayanan, S. (2011). SailAlign: Robust long speech-text alignment. In *Proc. VLSP*.
- [Kawahara et al., 1999] Kawahara, H., Masuda-Katsuse, I., and de Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Commun.*
- [Kawahara et al., 2001] Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *Proc. MAVEBA*.
- [Keltner and Haidt, 1999] Keltner, D. and Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition & emotion*.
- [Kesić and Spasić, 2016] Kesić, S. and Spasić, S. Z. (2016). Application of Higuchi’s fractal dimension from basic to clinical neurophysiology: a review. *Computer methods and programs in biomedicine*.
- [Khoa et al., 2012] Khoa, T. Q. D., Ha, V. Q., and Toi, V. V. (2012). Higuchi fractal properties of onset epilepsy electroencephalogram. *Computational and mathematical methods in medicine*.

- [Kim et al., 2018] Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollöfer, M., and Theobalt, C. (2018). Deep Video Portraits. *ACM Trans. Graphics*.
- [Kim et al., 2013] Kim, Y., Lee, H., and Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *Proc. ICASSP*.
- [Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proc. ICLR*.
- [Kleinsmith and Bianchi-Berthouze, 2013] Kleinsmith, A. and Bianchi-Berthouze, N. (2013). Affective Body Expression Perception and Recognition: A Survey. *IEEE Trans. Affective Computing*.
- [Kollias and Zafeiriou, 2018] Kollias, D. and Zafeiriou, S. (2018). Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv:1811.07770*.
- [Kollias et al., 2019a] Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., and Zafeiriou, S. (2019a). Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *Intl. J. Computer Vision*.
- [Kollias and Zafeiriou, 2019] Kollias, D. and Zafeiriou, S. (2019). Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. *arXiv:1910.04855*.
- [Kollias et al., 2019b] Kollias, D., Sharmanska, V., and Zafeiriou, S. (2019b). Face Behavior a la carte: Expressions, Affect and Action Units in a Single Network. *arXiv:1910.11111*.
- [Kollias et al., 2020] Kollias, D., Schulc, A., Hajiyeve, E., and Zafeiriou, S. (2020). Analysing Affective Behavior in the First ABAW 2020 Competition. In *Proc. FG*.
- [Kollias and Zafeiriou, 2021a] Kollias, D. and Zafeiriou, S. (2021a). Affect Analysis in-the-wild: Valence-Arousal, Expressions, Action Units and a Unified Framework. *arXiv:2103.15792*.
- [Kollias and Zafeiriou, 2021b] Kollias, D. and Zafeiriou, S. (2021b). Analysing affective behavior in the second abaw2 competition. In *Proc. ICCV*.
- [Kollias et al., 2021] Kollias, D., Sharmanska, V., and Zafeiriou, S. (2021). Distribution Matching for Heterogeneous Multi-Task Learning: a Large-scale Face Study. *arXiv:2105.03790*.
- [Kosti et al., 2017a] Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. (2017a). EMOTIC: Emotions in Context dataset. In *Proc. CVPRW*.
- [Kosti et al., 2017b] Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. (2017b). Emotion recognition in context. In *Proc. CVPR*.
- [Koujan and Roussos, 2018] Koujan, M. R. and Roussos, A. (2018). Combining dense nonrigid structure from motion and 3d morphable models for monocular 4d face reconstruction. In *Proc. SIGGRAPH (Europe)*.

- [Koutsouleris et al., 2011] Koutsouleris, N., Davatzikos, C., Bottlender, R., Patschurck-Kliche, K., Scheuerecker, J., et al. (2011). Early recognition and disease prediction in the at-risk mental states for psychosis using neurocognitive pattern classification. *Schizophrenia Bulletin*.
- [Kritsis et al., 2022] Kritsis, K., Gkiokas, A., Pikrakis, A., and Katsouros, V. (2022). DanceConv: Dance Motion Generation With Convolutional Networks. *IEEE Access*.
- [Kuo et al., 2018] Kuo, C.-M., Lai, S.-H., and Sarkis, M. (2018). A Compact Deep Learning Model for Robust Facial Expression Recognition. In *Proc. CVPRW*.
- [Le Goff and Benoît, 1996] Le Goff, B. and Benoît, C. (1996). A text-to-audiovisual-speech synthesizer for french. In *Proc. ICSLP*.
- [LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*.
- [LeCun et al., 2012] LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient backprop. In *Neural Networks: Tricks of the Trade*.
- [Lee et al., 2019] Lee, J., Kim, S., Kim, S., Park, J., and Sohn, K. (2019). Context-aware emotion recognition networks. In *Proc. ICCV*.
- [Leite et al., 2014] Leite, I., Castellano, G., Pereira, A., Martinho, C., and Paiva, A. (2014). Empathic robots for long-term interaction. *Intl. J. Social Robotics*.
- [Li et al., 2021a] Li, C., Morel-Forster, A., Vetter, T., Egger, B., and Kortylewski, A. (2021a). To fit or not to fit: Model-based face reconstruction and occlusion segmentation from weak supervision. *arXiv preprint arXiv:2106.09614*.
- [Li et al., 2021b] Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., and Lu, C. (2021b). Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proc. CVPR*.
- [Li and Deng, 2018] Li, S. and Deng, W. (2018). Deep Facial Expression Recognition: A Survey. *arXiv:1804.08348*.
- [Li et al., 2017] Li, T., Bolkart, T., Black, M. J., Li, H., and Romero, J. (2017). Learning a model of facial shape and expression from 4D scans. *Proc. SIGGRAPH (Asia)*.
- [Li et al., 2016] Li, X., Wu, Z., Meng, H., Jia, J., Lou, X., and Cai, L. (2016). Expressive Speech Driven Talking Avatar Synthesis with DBLSTM using Limited Amount of Emotional Bimodal Data. In *Proc. Interspeech*.
- [Lim et al., 2016] Lim, W., Jang, D., and Lee, T. (2016). Speech emotion recognition using convolutional and recurrent neural networks. In *Proc. APSIPA*.
- [Ling et al., 2015] Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., Meng, H. M., and Deng, L. (2015). Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Process. Mag.*
- [Liu and Ostermann, 2011] Liu, K. and Ostermann, J. (2011). Realistic facial expression synthesis for an image-based talking head. In *Proc. ICME*.

- [Lopez-Rincon, 2019] Lopez-Rincon, A. (2019). Emotion recognition using facial expressions in children using the NAO Robot. In *Proc. CONIELECOMP*.
- [Lorenzo-Trueba et al., 2015] Lorenzo-Trueba, J., Barra-Chicote, R., San-Segundo, R., Ferreiros, J., Yamagishi, J., and Montero, J. M. (2015). Emotion transplantation through adaptation in HMM-based speech synthesis. *Comput. Speech & Lang.*
- [Lugaresi et al., 2019] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., et al. (2019). Mediapipe: A framework for building perception pipelines. *arXiv:1906.08172*.
- [Luo et al., 2013] Luo, Y., Wu, C.-m., and Zhang, Y. (2013). Facial expression recognition based on fusion feature of PCA and LBP with SVM. *Optik-Intl. Journal for Light and Electron Optics*.
- [Luo et al., 2020] Luo, Y., Ye, J., Adams, R. B., Li, J., Newman, M. G., and Wang, J. Z. (2020). ARBEE: Towards automated recognition of bodily expression of emotion in the wild. *Intl. J. Computer Vision*.
- [Ma et al., 2022] Ma, P., Petridis, S., and Pantic, M. (2022). Visual Speech Recognition for Multiple Languages in the Wild. *arXiv Preprint: 2202.13084*.
- [Maglogiannis et al., 2020] Maglogiannis, I., Zlatintsi, A., Menychtas, A., Papadimitos, D., Filntisis, P. P., Efthymiou, N., Retsinas, G., Tsanakas, P., and Maragos, P. (2020). An Intelligent Cloud-Based Platform for Effective Monitoring of Patients with Psychotic Disorders. In *Proc. Int'l Conf. on Artificial Intelligence Applic. and Innovation*.
- [Malandrakis et al., 2011] Malandrakis, N., Potamianos, A., Iosif, E., and Narayanan, S. (2011). Kernel models for affective lexicon creation. In *Proc. Interspeech*.
- [Malleison et al., 2015] Malleison, C., Bazin, J. C., Wang, O., Bradley, D., Beeler, T., Hilton, A., and Sorkine-Hornung, A. (2015). FaceDirector: Continuous Control of Facial Performance in Video. In *Proc. ICCV*.
- [Mann and Whitney, 1947] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of Mathematical Statistics*.
- [Mao et al., 2017] Mao, X., Li, Q., Xie, H., Lau, R. K., Wang, Z., and Smolley, S. (2017). Least Squares Generative Adversarial Networks. In *Proc. ICCV*.
- [Maragos, 1994] Maragos, P. (1994). Fractal signal analysis using mathematical morphology. In *Advances in electronics and electron physics*.
- [Marín-Morales et al., 2020] Marín-Morales, J., Llinares, C., Guixeres, J., and Alcañiz, M. (2020). Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors*.
- [Marinoiu et al., 2018] Marinoiu, E., Zanfir, M., Olaru, V., and Sminchisescu, C. (2018). 3D Human Sensing, Action and Emotion Recognition in Robot Assisted Therapy of Children with Autism. In *Proc. CVPR*.
- [Martyniuk et al., 2022] Martyniuk, T., Kupyn, O., Kurlyak, Y., Krashenyi, I., Matas, J., and Sharmanska, V. (2022). DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image. In *Proc. CVPR*.

- [Mase, 1991] Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Trans. Information and Systems*.
- [Mathias et al., 2014] Mathias, M., Benenson, R., Pedersoli, M., and Van Gool, L. (2014). Face detection without bells and whistles. In *Proc. ECCV*.
- [Matthews and Baker, 2004] Matthews, I. and Baker, S. (2004). Active Appearance Models Revisited. *Int. J. Comput. Vis.*
- [Mattheyses et al., 2008] Mattheyses, W., Latacz, L., Verhelst, W., and Sahli, H. (2008). Multimodal unit selection for 2D audiovisual text-to-speech synthesis. In *Proc. MLMI*.
- [Mattheyses et al., 2011] Mattheyses, W., Latacz, L., and Verhelst, W. (2011). Auditory and photo-realistic audiovisual speech synthesis for Dutch. In *Proc. AVSP*.
- [Mattheyses and Verhelst, 2015] Mattheyses, W. and Verhelst, W. (2015). Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Commun.*
- [Mavridis, 2015] Mavridis, N. (2015). A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*.
- [McCandless-Glimcher et al., 1986] McCandless-Glimcher, L., McKnight, S., Hamera, E., Smith, B. L., Peterson, K. A., and Plumlee, A. A. (1986). Use of symptoms by schizophrenics to monitor and regulate their illness. *Psychiatric Services*.
- [McGorry et al., 2014] McGorry, P., Keshavan, M., Goldstone, S., Amminger, P., Allott, K., Berk, M., Lavoie, S., Pantelis, C., Yung, A., Wood, S., et al. (2014). Biomarkers and clinical staging in psychiatry. *World Psychiatry*.
- [McGurk and MacDonald, 1976] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*.
- [Mehrabian and Russell, 1974] Mehrabian, A. and Russell, J. A. (1974). *An approach to environmental psychology*.
- [Mehrabian, 1980] Mehrabian, A. (1980). *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies*.
- [Melenchón et al., 2009] Melenchón, J., Martínez, E., De La Torre, F., and Montero, J. A. (2009). Emphatic visual speech synthesis. *IEEE Trans. Audio, Speech, Language Process.*
- [Mittal et al., 2020] Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D. (2020). EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege’s Principle. In *Proc. CVPR*.
- [Mollahosseini et al., 2017] Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affective Computing*.
- [Mori et al., 2012] Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robot. Autom. Mag.*
- [Mukhopadhyay, 2014] Mukhopadhyay, S. C. (2014). Wearable sensors for human activity monitoring: A review. *IEEE Sensors J.*

- [Nagarajan and Oruganti, 2019] Nagarajan, B. and Oruganti, V. R. (2019). Cross-Domain Transfer Learning for Complex Emotion Recognition. In *Proc. TENSYPMP*.
- [Nguyen et al., 2017] Nguyen, D., Nguyen, K., Sridharan, S., Ghasemi, A., Dean, D., and Fookes, C. (2017). Deep spatio-temporal features for multimodal emotion recognition. In *Proc. WACV*.
- [Nirkin et al., 2019] Nirkin, Y., Keller, Y., and Hassner, T. (2019). FSGAN: Subject agnostic face swapping and reenactment. In *Proc. ICCV*.
- [Nojavanasghari et al., 2016] Nojavanasghari, B., Baltrušaitis, T., Hughes, C. E., and Morency, L.-P. (2016). EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proc. ICMI*.
- [Noroozi et al., 2018] Noroozi, F., Corneanu, C. A., Kamińska, D., Sapiński, T., Escalera, S., and Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *arXiv:1801.07481*.
- [Odell, 1995] Odell, J. J. (1995). *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge.
- [Ouni et al., 2007] Ouni, S., Cohen, M. M., Ishak, H., and Massaro, D. W. (2007). Visual contribution to speech perception: measuring the intelligibility of animated talking heads. *EURASIP J. Audio, Speech, Music Process.*
- [Owusu et al., 2014] Owusu, E., Zhan, Y., and Mao, Q. R. (2014). An SVM-AdaBoost facial expression recognition system. *Applied intelligence*.
- [Pandzic and Forchheimer, 2003] Pandzic, I. S. and Forchheimer, R. (2003). *MPEG-4 facial animation: the standard, implementation and applications*.
- [Pantic et al., 2005] Pantic, M., Sebe, N., Cohn, J. F., and Huang, T. (2005). Affective multimodal human-computer interaction. In *Proc. Int. Conf. on Multimedia*.
- [Papandreou and Maragos, 2008] Papandreou, G. and Maragos, P. (2008). Adaptive and Constrained Algorithms for Inverse Compositional Active Appearance Model Fitting. In *Proc. CVPR*.
- [Paraperas Papantoniou et al., 2022] Paraperas Papantoniou, F., Filntisis, P. P., Maragos, P., and Roussos, A. (2022). Neural Emotion Director: Speech-preserving semantic control of facial expressions in "in-the-wild" videos. In *Proc. CVPR*.
- [Patel et al., 2012] Patel, S., Park, H., Bonato, P., Chan, L., and Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *J. Neuro-engineering and Rehabilitation*.
- [Paysan et al., 2009] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. In *Proc. AVSS*.
- [Pelachaud et al., 1996] Pelachaud, C., Badler, N. I., and Steedman, M. (1996). Generating facial expressions for speech. *Cognitive science*.
- [Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). GloVE: Global vectors for word representation. In *Proc. EMNLP*.

- [Pérez et al., 2003] Pérez, P., Gangnet, M., and Blake, A. (2003). Poisson image editing. *ACM Trans. Graph.*
- [Pham et al., 2017] Pham, H. X., Cheung, S., and Pavlovic, V. (2017). Speech-Driven 3D Facial Animation with Implicit Emotional Awareness: A Deep Learning Approach. In *Proc. CVPRW*.
- [Piana et al., 2016] Piana, S., Staglianò, A., Odone, F., and Camurri, A. (2016). Adaptive Body Gesture Representation for Automatic Emotion Recognition. *ACM Trans. Interact. Intell. Syst.*
- [Picard, 1995] Picard, R. W. (1995). Affective computing.
- [Plutchik, 1980] Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*.
- [Plutchik and Kellerman, 1980] Plutchik, R. and Kellerman, H. (1980). *Emotion, Theory, Research, and Experience: Theory, Research and Experience*.
- [Plutchik, 2001] Plutchik, R. (2001). The Nature of Emotions Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.*
- [Psaltis et al., 2016] Psaltis, A., Kaza, K., Stefanidis, K., Thermos, S., Apostolakis, K. C., Dimitropoulos, K., and Daras, P. (2016). Multimodal affective state recognition in serious games applications. In *Proc. IST*.
- [Qian et al., 2014] Qian, Y., Fan, Y., Hu, W., and Soong, F. K. (2014). On the training aspects of deep neural network (DNN) for parametric TTS synthesis. In *Proc. ICASSP*.
- [Raptis et al., 2016] Raptis, S., Tsiakoulis, P., Chalamandaris, A., and Karabetsos, S. (2016). Expressive Speech Synthesis for Storytelling: The INNOETICS'Entry to the Blizzard Challenge 2016. In *Proc. Blizzard Challenge*.
- [Ren et al., 2017] Ren, Z., Jin, H., Lin, Z., Fang, C., and Yuille, A. L. (2017). Multiple Instance Visual-Semantic Embedding. In *Proc. BMVC*.
- [Retsinas et al., 2020] Retsinas, G., Filntisis, P. P., Efthymiou, N., Theodosis, E., Zlatintsi, A., and Maragos, P. (2020). Person Identification Using Deep Convolutional Neural Networks on Short-Term Signals from Wearable Sensors. In *Proc. ICASSP*.
- [Reyes-Ortiz et al., 2014] Reyes-Ortiz, J. L., Oneto, L., Ghio, A., Samá, A., Anguita, D., and Parra, X. (2014). Human activity recognition on smartphones with awareness of basic activities and postural transitions. In *Proc. ANN*.
- [Richard J. Davidson, 2002] Richard J. Davidson, Klaus R. Scherer, H. H. G. (2002). *Handbook of Affective Sciences*.
- [Richman and Moorman, 2000] Richman, J. S. and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American J. Physiology-Heart and Circulatory Physiology*.
- [Robokind, 2022] Robokind (2022). Robokind. Advanced Social Robots. <http://robokind.com/>.

- [Robotham et al., 2016] Robotham, D., Satkunanathan, S., Doughty, L., and Wykes, T. (2016). Do we still have a digital divide in mental health? A five-year survey follow-up. *J. Medical Internet Research*.
- [Ronanki et al., 2016] Ronanki, S., Wu, Z., Watts, O., and King, S. (2016). A Demonstration of the Merlin Open Source Neural Network Speech Synthesis System. In *Proc. ISCA SSW9*.
- [Ros et al., 2011] Ros, R., Nalin, M., Wood, R., Baxter, P., Looije, R., Demiris, Y., Belpaeme, T., Giusti, A., and Pozzi, C. (2011). Child-robot interaction in the wild: advice to the aspiring experimenter. In *Proc. ICMI*.
- [Ruan et al., 2021] Ruan, Z., Zou, C., Wu, L., Wu, G., and Wang, L. (2021). Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. *IEEE Trans. Image Processing*.
- [Russell and Mehrabian, 1977] Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *J. of Research in Personality*.
- [Russell, 1980] Russell, J. A. (1980). A circumplex model of affect. *J. personality and social psychology*.
- [Sadoughi and Busso, 2017] Sadoughi, N. and Busso, C. (2017). Joint learning of speech-driven facial motion with bidirectional long-short term memory. In *Proc. IVA*.
- [Sagonas et al., 2016] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and vision computing*.
- [Saito et al., 2016] Saito, S., Li, T., and Li, H. (2016). Real-time facial segmentation and performance capture from rgb input. In *Proc. ECCV*.
- [Sako et al., 2000] Sako, S., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). HMM-based text-to-audio-visual speech synthesis. In *Proc. ICLSP*.
- [Sanyal et al., 2019] Sanyal, S., Bolkart, T., Feng, H., and Black, M. (2019). Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision. In *Proc. CVPR*.
- [Scargle, 1982] Scargle, J. D. (1982). Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*.
- [Schabus et al., 2014] Schabus, D., Pucher, M., and Hofer, G. (2014). Joint audiovisual hidden semi-markov model-based speech synthesis. *IEEE J. Sel. Topics Signal Process.*
- [Scherer, 1986] Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological bulletin*.
- [Schröder, 2009] Schröder, M. (2009). Expressive speech synthesis: Past, present, and possible futures. In *Affective information processing*.
- [Schwarz, 2000] Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition & Emotion*.

- [Sebe et al., 2005] Sebe, N., Cohen, I., and Huang, T. S. (2005). Multimodal emotion recognition. In *Handbook of pattern recognition and computer vision*.
- [Seyama and Nagayama, 2007] Seyama, J. and Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleop. Virt. Env.*
- [Shaffer and Ginsberg, 2017] Shaffer, F. and Ginsberg, J. (2017). An overview of heart rate variability metrics and norms. *Frontiers Public Health*.
- [Shang et al., 2020] Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., and Quan, L. (2020). Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. *arXiv preprint arXiv:2007.12494*.
- [Shapiro and Wilk, 1965] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*.
- [Shaw and Theobald, 2016] Shaw, F. and Theobald, B.-J. (2016). Expressive Modulation of Neutral Visual Speech. *IEEE MultiMedia*.
- [Shi et al., 2022a] Shi, B., Hsu, W.-N., Lakhota, K., and Mohamed, A. (2022a). Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. *arXiv:2201.02184*.
- [Shi et al., 2022b] Shi, B., Hsu, W.-N., and Mohamed, A. (2022b). Robust Self-Supervised Audio-Visual Speech Recognition. *arXiv:2201.01763*.
- [Shinoda and Watanabe, 1997] Shinoda, K. and Watanabe, T. (1997). Acoustic Modelling Based on the MDL Principle for Speech Recognition. In *Proc. EUROSPEECH*.
- [Shinoda and Lee, 2001] Shinoda, K. and Lee, C.-H. (2001). A structural Bayes approach to speaker adaptation. *IEEE Trans. Speech Audio Process*.
- [Sifakis et al., 2006] Sifakis, E., Selle, A., Robinson-Mosher, A., and Fedkiw, R. (2006). Simulating speech with a physics-based facial muscle model. In *Proc. SIGGRAPH*.
- [Simon et al., 2017] Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand Key-point Detection in Single Images using Multiview Bootstrapping. In *Proc. CVPR*.
- [Skipper et al., 2007] Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex*.
- [Solanki and Roussos, 2021] Solanki, G. K. and Roussos, A. (2021). Deep Semantic Manipulation of Facial Videos. *arXiv:2111.07902*.
- [Song et al., 2020] Song, L., Wu, W., Qian, C., He, R., and Loy, C. C. (2020). Everybody’s Talkin’: Let Me Talk as You Want. *arXiv:2001.05201*.
- [Sreeja and Mahalakshmi, 2017] Sreeja, P. and Mahalakshmi, G. (2017). Emotion models: a review. *Intl. J. Control Theory and Applications*.
- [Staples et al., 2017] Staples, P., Torous, J., Barnett, I., et al. (2017). A comparison of passive and active estimates of sleep in a cohort with schizophrenia. *NPJ schizophrenia*.

- [Stuart et al., 2022] Stuart, J., Aul, K., Stephen, A., Bumbach, M. D., and Lok, B. (2022). The Effect of Virtual Human Rendering Style on User Perceptions of Visual Cues. *Frontiers in Virtual Reality*.
- [Sumbly and Pollack, 1954] Sumbly, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.*
- [Sun et al., 2018] Sun, B., Cao, S., He, J., and Yu, L. (2018). Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy. *Neural Networks*.
- [Tamura et al., 1999] Tamura, M., Kondo, S., Masuko, T., and Kobayashi, T. (1999). Text-to-audiovisual speech synthesis based on parameter generation from HMM. In *Proc. Eurospeech*.
- [Tewari et al., 2017] Tewari, A., Zollöfer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., and Theobalt, C. (2017). MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *Proc. ICCV*.
- [Theobald and Matthews, 2012] Theobald, B.-J. and Matthews, I. (2012). Relating Objective and Subjective Performance Measures for AAM-Based Visual Speech Synthesis. *IEEE Trans. Audio, Speech, and Language Processing*.
- [Thies et al., 2015] Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., and Theobalt, C. (2015). Real-Time Expression Transfer for Facial Reenactment. *ACM Trans. Graph.*
- [Thies et al., 2016] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Facevr: Real-time facial reenactment and eye gaze control in virtual reality. *arXiv:1610.03151*.
- [Thies et al., 2018] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2018). Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. *Commun. ACM*.
- [Thies et al., 2020] Thies, J., Elgharib, M., Tewari, A., Theobalt, C., and Nießner, M. (2020). Neural Voice Puppetry: Audio-driven Facial Reenactment.
- [Tielman et al., 2014] Tielman, M., Neerincx, M., Meyer, J.-J., and Looije, R. (2014). Adaptive emotional expression in robot-child interaction. In *Proc. HRI*.
- [Tokuda et al., 2000] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*.
- [Tokuda et al., 2002] Tokuda, K., Zen, H., and Black, A. W. (2002). An HMM-based speech synthesis system applied to English. In *Proc. IEEE SSW*.
- [Tokuda et al., 2013] Tokuda, K., Toda, T., and Yamagishi, J. (2013). Speech Synthesis Based on Hidden Markov Models. *Proc. IEEE*.
- [Torous et al., 2016] Torous, J., Kiang, M. V., Lorme, J., and Onnela, J.-P. (2016). New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health*.

- [Tracy and Robins, 2004] Tracy, J. L. and Robins, R. W. (2004). Show your pride: Evidence for a discrete emotion expression. *Psychological Science*.
- [Tran et al., 2018] Tran, A. T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., and Medioni, G. (2018). Extreme 3d face reconstruction: Seeing through occlusions. In *Proc. CVPR*.
- [Tran and Liu, 2018] Tran, L. and Liu, X. (2018). Nonlinear 3d face morphable model. In *Proc. CVPR*.
- [Trigeorgis et al., 2016] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proc. ICASSP*.
- [Tripathy et al., 2020] Tripathy, S., Kannala, J., and Rahtu, E. (2020). ICface: Interpretable and Controllable Face Reenactment Using GANs. In *Proc. WACV*.
- [Tsiami et al., 2018] Tsiami, A., Koutras, P., Efthymiou, N., Filntisis, P. P., Potamianos, G., and Maragos, P. (2018). Multi3: Multi-Sensory Perception System for Multi-Modal Child Interaction with Multiple Robots. In *Proc. ICRA*.
- [Turk and Pentland, 1991] Turk, M. A. and Pentland, A. P. (1991). Face recognition using eigenfaces. In *Proc. CVPR*.
- [Tzirakis et al., 2017a] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017a). End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. Selected Topics in Signal Processing*.
- [Tzirakis et al., 2017b] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017b). End-to-end multimodal emotion recognition using deep neural networks. *J. Selected Topics Signal Processing*.
- [Valenza et al., 2014] Valenza, M., Nardelli, A., Lanat'a, C., Gentili, G., Bertschy, R., Paradiso, and Scilingo, E. P. (2014). Wearable Monitoring for Mood Recognition in Bipolar Disorder Based on History-Dependent Long-Term Heart Rate Variability Analysis. *IEEE J. of Biomedical and Health Informatics*.
- [Van den Stock et al., 2007] Van den Stock, J., Righart, R., and De Gelder, B. (2007). Body expressions influence recognition of emotions in the face and voice. *Emotion*.
- [Vantuch, 2018] Vantuch, T. (2018). Analysis of time series data.
- [Viola and Jones, 2004] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *Intl. J. computer vision*.
- [Vougioukas et al., 2019] Vougioukas, K., Petridis, S., and Pantic, M. (2019). End-to-End Speech-Driven Facial Animation with Temporal GANs. *arXiv:1805.09313*.
- [Wallbott, 1998] Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology*.
- [Wan et al., 2013] Wan, V., Anderson, R., Blokland, A., Braunschweiler, N., Chen, L., Kolluru, B., Latorre, J., Maia, R., Stenger, B., Yanagisawa, K., et al. (2013). Photo-realistic expressive text to talking head synthesis. In *Proc. Interspeech*.

- [Wang et al., 2020a] Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., and Loy, C. C. (2020a). Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Proc. ECCV*.
- [Wang et al., 2020b] Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., and Loy, C. C. (2020b). Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Proc. ECCV*.
- [Wang et al., 2010] Wang, L., Qian, X., Han, W., and Soong, F. K. (2010). Photo-real lips synthesis with trajectory-guided sample selection. In *Proc. ISCA SSW7*.
- [Wang et al., 2016] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *Proc. ECCV*.
- [Wang et al., 2022] Wang, L., Chen, Z., Yu, T., Ma, C., Li, L., and Liu, Y. (2022). FaceVerse: a Fine-grained and Detail-controllable 3D Face Morphable Model from a Hybrid Dataset. In *Proc. CVPR*.
- [Watts et al., 2016] Watts, O., Henter, G. E., Merritt, T., Wu, Z., and King, S. (2016). From HMMs to DNNs: where do the improvements come from? In *Proc. ICASSP*.
- [Websdale et al., 2022] Websdale, D., Taylor, S., and Milner, B. (2022). Speaker-Independent Speech Animation Using Perceptual Loss Functions and Synthetic Data. *IEEE Trans. Multimedia*.
- [Wei et al., 2020] Wei, Z., Zhang, J., Lin, Z., Lee, J.-Y., Balasubramanian, N., Hoai, M., and Samaras, D. (2020). Learning Visual Emotion Representations From Web Data. In *Proc. CVPR*.
- [Wiersma et al., 1995] Wiersma, D., Nienhuis, F. J., Slooff, C. J., and Giel, R. (1995). Prodromes and precursors: Epidemiologic data for primary prevention of disorders with slow onset. *The American J. psychiatry*.
- [Wiersma et al., 1998] Wiersma, D., Nienhuis, F. J., Slooff, C. J., and Giel, R. (1998). Natural course of schizophrenic disorders: a 15-year follow up of a Dutch incidence cohort. *Schizophrenia bulletin*.
- [Williams and Hinton, 1986] Williams, D. and Hinton, G. (1986). Learning representations by back-propagating errors. *Nature*.
- [Wood et al., 2022] Wood, E., Baltrusaitis, T., Hewitt, C., Johnson, M., Shen, J., Milosavljevic, N., Wilde, D., Garbin, S., Raman, C., Shotton, J., Sharp, T., Stojiljkovic, I., Cashman, T., and Valentin, J. (2022). 3d face reconstruction with dense landmarks. In *Proc. ECCV*.
- [Wu et al., 2006] Wu, Z., Zhang, S., Cai, L., and Meng, H. M. (2006). Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar. In *Proc. Interspeech*.
- [Wu et al., 2016] Wu, Z., Watts, O., and King, S. (2016). Merlin: An open source neural network speech synthesis system. In *Proc. ISCA SSW9*.
- [Xie and Liu, 2007] Xie, L. and Liu, Z.-Q. (2007). A coupled HMM approach to video-realistic speech animation. *Pattern Recognit.*

- [Xie et al., 2014] Xie, L., Sun, N., and Fan, B. (2014). A statistical parametric approach to video-realistic text-driven talking avatar. *Multimed. Tools Appl.*
- [Xu et al., 2022] Xu, Y., Zhang, J., Zhang, Q., and Tao, D. (2022). ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *arxiv:2204.12484*.
- [Yamagishi et al., 2004] Yamagishi, J., Masuko, T., and Kobayashi, T. (2004). HMM-based expressive speech synthesis - Towards TTS with arbitrary speaking styles and emotions. In *Proc. of Special Workshop in Maui*.
- [Yamagishi et al., 2007] Yamagishi, J., Zen, H., Toda, T., and Tokuda, K. (2007). Speaker-Independent HMM-based Speech Synthesis System: HTS-2007 System for the Blizzard Challenge 2007. In *Proc. Blizzard Challenge*.
- [Yamagishi et al., 2009] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. Audio, Speech, Lang. Process.*
- [Yang et al., 2020] Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., and Cao, X. (2020). FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *Proc. CVPR*.
- [Yeh and Li, 2020] Yeh, M.-C. and Li, Y.-N. (2020). Multilabel deep visual-semantic embedding. *IEEE Trans. Pattern Analysis Machine Intelligence*.
- [Yoshimura et al., 1999] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. Eurospeech*.
- [Yoshimura et al., 2000] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speaker interpolation for HMM-based speech synthesis system. *Acoust. Sci. Technol.*
- [Yoshimura et al., 2005] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2005). Incorporating a mixed excitation model and postfilter into HMM-based text-to-speech synthesis. *Syst. Comput. Jpn.*
- [Zafeiriou et al., 2017] Zafeiriou, S., Kollias, D., Nicolaou, M. A., Papaioannou, A., Zhao, G., and Kotsia, I. (2017). Aff-wild: Valence and arousal ‘in-the-wild’ challenge. In *Proc. CVPRW*.
- [Zakharov et al., 2019] Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. (2019). Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *Proc. ICCV*.
- [Zen et al., 2007a] Zen, H., Tokuda, K., Masuko, T., Kobayasih, T., and Kitamura, T. (2007a). A Hidden Semi-Markov Model-Based Speech Synthesis System. *IEICE - Trans. Inf. Syst.*
- [Zen et al., 2007b] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., and Tokuda, K. (2007b). The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. ISCA SSW6*.
- [Zen et al., 2009] Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Commun.*

- [Zen et al., 2013] Zen, H., Senior, A., and Schuster, M. (2013). Statistical Parametric Speech Synthesis Using Deep Neural Networks. In *Proc. ICASSP*.
- [Zen, 2015] Zen, H. (2015). Acoustic modeling in statistical parametric speech synthesis—from HMM to LSTM-RNN. *Proc. MLSLP*.
- [Zhang et al., 2016] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*.
- [Zhang et al., 2018a] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018a). The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*.
- [Zhang et al., 2010] Zhang, S., Wu, Z., Meng, H. M., and Cai, L. (2010). Facial expression synthesis based on emotion dimensions for affective talking avatar. In *Modeling machine emotions for realizing intelligence*.
- [Zhang et al., 2018b] Zhang, S., Zhang, S., Huang, T., and Gao, W. (2018b). Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Trans. Multimedia*.
- [Zhang et al., 2013] Zhang, X., Wang, L., Li, G., Seide, F., and Soong, F. K. (2013). A new language independent, photo-realistic talking head driven by voice only. In *Proc. Interspeech*.
- [Zhao et al., 2019] Zhao, J., Mao, X., and Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*.
- [Zhou et al., 2020] Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., and Li, D. (2020). MakeltTalk: speaker-aware talking-head animation. *ACM Trans. Graphics*.
- [Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proc. ICCV*.
- [Zhu et al., 2017] Zhu, X., Liu, X., Lei, Z., and Li, S. Z. (2017). Face alignment in full pose range: A 3d total solution. *IEEE Trans. Pattern Analysis Machine Intelligence*.
- [Zielonka et al., 2022] Zielonka, W., Bolkart, T., and Thies, J. (2022). Towards Metrical Reconstruction of Human Faces. In *Proc. ECCV*.
- [Zollhöfer et al., 2018] Zollhöfer, M., Thies, J., Bradley, D., Garrido, P., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., and Theobalt, C. (2018). State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications.

Appendix A

Glossary

AVTTS	Audio-Visual Speech Synthesis	Οπτικοακουστική Σύνθεση Φωνής
AUC	Area Under Curve	Περιοχή Κάτωθεν της Καμπύλης
BN	Batch Normalization	Κανονικοποίηση σε Τμήματα
CNN	Convolutional Neural Network	Συνελικτικό Νευρωνικό Δίκτυο
CRI	Child-Robot Interaction	Αλληλεπίδραση Παιδιού-Ρομπότ
DNN	Deep Neural Network	Βαθύ Νευρωνικό Δίκτυο
EAVTTS	Expressive Audio-Visual Speech Synthesis	Εκφραστική Οπτικοακουστική Σύνθεση Φωνής
FC	Fully-Connected	Πλήρως συνδεδεμένο
FFT	Fast Fourier Transform	Γρήγορος Μετασχηματισμός Fourier
HMM	Hidden Markov Model	Κρυφό Μαρκοβιανό Μοντέλο
HRI	Human-Robot Interaction	Αλληλεπίδραση Ανθρώπου-Ρομπότ
MLE	Maximum Likelihood Estimation	Εκτίμηση Μεγίστης Πιθανοφάνειας
MSE	Mean-Squared Error	Μέσο Τετραγωνικό Σφάλμα
MOS	Mean Opinion Score	Μέση Βαθμολογία Γνώμης
PCA	Principal Component Analysis	Ανάλυση Κυρίων Συνιστωσών
RNN	Recurrent Neural Network	Επαναληπτικό Νευρωνικό Δίκτυο
SGD	Stochastic Gradient Descent	Στοχαστική Κάθοδος Βασισμένη στη Κλίση
TCN	Temporal Convolutional Network	Χρονικό Συνελικτικό Δίκτυο
TSN	Temporal Segment Network	Δίκτυο Μοντελοποίησης Χρονικών Τμημάτων
TTS	Text-To-Speech	Σύνθεση Φωνής από Κείμενο
US	Unit Selection	Παράθεση Ακουστικών Μονάδων
VAD	Valence, Arousal, Dominance	Πολικότητα, Ένταση, Κυριαρχία

Appendix B

Wearable-based behavior analysis of patients with psychotic disorders

In this appendix we present a study on analyzing the behavior of patients with psychotic disorders using wearables. This work follows a different approach to the main text - it involves statistical analyses on data extracted from wearables, with the goal of identifying behavioral differences in patients with psychotic disorders, compared to controls. Nevertheless, this work remains relevant to the main concept of this thesis, since emotion is a core aspect of human behavior.

Here, we present a rigorous statistical analysis of various descriptive representations extracted from wearable data during long-term continuous monitoring of patients with psychotic disorders and healthy control counterparts. Wearable technologies and digital phenotyping foster unique opportunities for designing intelligent e-Health services that could deal with various health and well-being issues in patients with mental disorders, offering the potential to revolutionize psychiatry and its clinical practice. Our novel analysis in conjunction with our large-scale dataset collected during the course of more than a year, identifies features in both time and frequency domain using either linear or more novel nonlinear techniques, which fluctuate significantly between the two groups. Furthermore, the analysis offers substantial insights on several factors that differentiate between controls and patients with psychosis that could be leveraged in the future for relapse prevention and individualized assistance, as well as provide new diagnostic methods for clinical practitioners.

B.1 Introduction

Wearable consumer products, such as smartwatches and fitness trackers, are gaining popularity every day and the enormous technological advances made in recent years have enabled the reliable, unobtrusive, and remote personalized collection of numerous behavioral and biometric signals through their sensors [Patel et al., 2012, Boletsis et al., 2015].

This so-called “digital phenotyping” [Torous et al., 2016] has enabled significant advances in wearables for health purposes, leading to the fact that next-generation wearable technologies are about to help transform nowadays hospital-centered healthcare practice to proactive, individualized care. Behavioral and biometric indexes have already been used in general medicine, and sports and nowadays, the evidence indicates that they could also be introduced into clinical psychiatry [Aung et al., 2017]. Despite extensive research over

the last 60 years in neurobiology and neurophysiology of psychotic disorders, their cause remains unclear, and reliable biometric indexes for the diagnosis and prediction of the course of psychotic symptomatology have not yet been found. The use of such signals for the detection of early diagnosis and prevention of psychotic relapses is now one of the major research areas in psychiatry [Wiersma et al., 1995, Koutsouleris et al., 2011, McGorry et al., 2014].

The e-Prevention project¹ is an ongoing research and development project with the goal of collecting long-term continuous recordings of biometric and behavioral signals through non-intrusive commercial wearable sensors (i.e., smartwatches), in order to develop innovative, advanced and valuable tools. Such e-Health tools would facilitate the effective monitoring, the prediction of clinical symptoms, and the identification of biomarkers, which correlate with behavioral changes in patients with psychosis to support the prevention of a possible future relapse. Timely detection of such relapses is in fact of major importance, not only for the clinicians; since patients do not often present themselves when the symptoms begin to re-emerge or worsen [Corrigan et al., 1990], but it could also assist in reducing the severity of the relapses or even prevent their occurrences.

In contrast with previous works, which have lasted from some hours to a few weeks [Valenza et al., 2014, Barnett et al., 2018, Cella et al., 2018], with some exceptions [Adler et al., 2020], our continuous research study has been going on for more than one year, intending to achieve two years of continuous monitoring. According to the literature and previous studies, as well as the medical practitioners, the clinical image and burden of patients with psychosis vary greatly over the course of time. As a result, the long-term aspect of our work is of utter importance, since it is already known that the process of psychosis is continuous and relapse is a biological process evolving over time [McCandless-Glimcher et al., 1986, Gaebel et al., 1993, Wiersma et al., 1998]; thus, we could indeed claim and agree upon that only long term data could assist such an ambitious and challenging study.

In addition, previous works have mostly used smartphones [Reyes-Ortiz et al., 2014] and focused mainly on social features such as text messages, call duration, or others, such as location data, screen on/off time, and sleep duration. [Barnett et al., 2018, Benzeev et al., 2017, Adler et al., 2020]. Compared to smartphones, wearable sensors are unobtrusive, lightweight, and can be used for monitoring while the subjects perform all kinds of daily activities [Mukhopadhyay, 2014], ensuring this way a safe and sound living environment. Additionally, it has been already shown that people with psychotic illnesses are comfortable, able, and willing to integrate personal digital devices to monitor outcomes in their daily life. This supports the fact that by using wearable sensors, we could go beyond feasibility and underscore the novel physiological and activity data that can be easily collected with low cost [Robotham et al., 2016, Staples et al., 2017].

In this work, we employ a commercial off-the-shelf smartwatch aimed to have minuscule intrusion in the subject’s life and to be worn 24/7 (except during charging). The nature of our long-term study asks for a different data processing approach than previous studies. Inspired by traditional signal processing techniques, we extract common and more complex features using short-time analysis and study them through their descriptive statistics in order to obtain a rough estimate of how they differentiate between healthy controls and patients with psychotic disorders. The experimental evaluation shows that both the more common but also some of the novel nonlinear features examined are powerful in discriminating between the two groups. Our rigorous statistical analysis, one of the main goals of this work in order to identify the most suitable representations, in conjunction with our

¹More info about the e-Prevention project can be found at: <http://eprevention.gr>

large dataset, offers a high degree of certainty that these features differ extensively and are thus of major importance to clinical practitioners. Therefore, the analysis conducted in this work is a vital step towards developing a method that can leverage informative and interpretable physiological and behavioral data from sensors that could act as diagnostic tools with the aim to timely predict relapses or adverse drug reactions.

B.2 Experimental Protocol and Data Collection

B.2.1 Experimental Protocol

Twenty-three (23) healthy control volunteers and 24 patients with a disorder in the psychotic spectrum (12 with Schizophrenia, 8 with Bipolar Disorder I, 2 with Schizoaffective disorder, 1 with Brief Psychotic Episode, and 1 with Schizophreniform Disorder) were recruited at the University Mental Health, Neurosciences and Precision Medicine Research Institute “Costas Stefanis” (UMHRI) in Athens, Greece. All volunteers signed written consent for their participation after being fully informed about the project and its goals, while written permission for the use of their personal data (anonymized) was also given, in accordance with the provisions of the General Regulation (EU) 2016/679. Additionally, all protocols of the e-Prevention research project have been approved by the Ethics Committee of the Institution.

Initially, the controls underwent a clinical evaluation to ensure there was no history of mental disorders or toxic substance usage, while for the recruitment of the patients, the clinicians met with the participants to conduct an assessment of symptoms and functioning. At recruitment, patients were in active treatment and stable. The clinical team also conducted follow-up assessments with patients once every month of the study to administer various reliable rating scales (i.e., PANSS - Positive and Negative Syndrome Scale among others), which measure various psychiatric symptoms associated with their condition.

Table B.1 contains information on the demographics of the two groups, as well as the collected data (described in Sec. B.3) at the time of writing this paper. We also include the BMI (Body-Mass Index) and the PANSS scale rating at the time of recruitment for the two groups (note that PANSS is only applicable to patients).

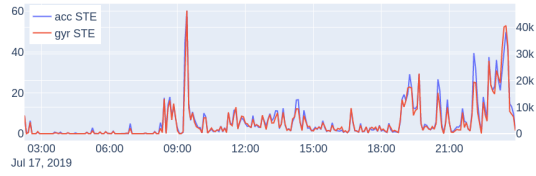
B.2.2 Method & Data Collection

The subjects were provided with a Samsung Gear S3 smartwatch that continuously monitored acceleration (*acc*), angular velocity (*gyr*), and heart rate (via Photoplethysmography [Allen, 2007]). Due to limitations on the number of available devices, each subject was recruited at a different date – controls were recruited between June 2019 and October 2019, while patients have been continuously recruited from November 2019 up to now (March 2021). Controls were continuously monitored for at least 90 days and then returned the watches, while the monitoring of patients has been an ongoing process. In the analysis presented in this paper, we select data up to September 2019 so that the collected data for each group are approximately balanced. Furthermore, to mitigate the effect of the CoVID-19 Pandemic quarantine lockdown in Greece (15/03/2020–10/05/2020), since only patients were monitored at the time, we exclude data collected during this period.

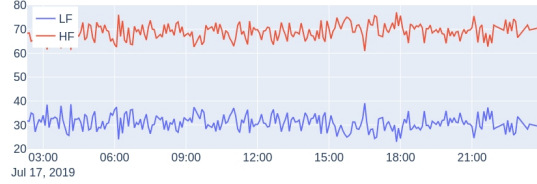
Data were collected using an in-house developed application and uploaded every day to a secure cloud infrastructure [Maglogiannis et al., 2020]. Accelerometer (*acc*) and gyroscope data (*gyr*) were collected at a frequency of 20Hz, while the heart rate and the heart rate variability (RR intervals – time intervals between two successive heart pulses)

	Controls	Patients
Demographics		
Male/Female	12/11	16/8
Age (years)	27.8 ± 3.9	30.8 ± 6.56
Education (years)	16.9 ± 1.8	13.88 ± 2.27
Smoker/Non-smoker	4/19	14/10
Illness dur. (years)	-	7.42 ± 5.63
BMI	22.9 ± 3.2	28.25 ± 5.13
PANSS (overall)	-	57.08 ± 14.10
Recorded Data		
# Weeks Recorded	84.3 ± 30.9	68.5 ± 41.7
# 5 min. mov (awake)	15746 ± 4837	13210 ± 6908
# 5 min HRV (awake)	12909 ± 3589	12221 ± 6656
# 5 min. mov (sleep)	7670 ± 2606	8865 ± 4767
# 5 hour HRV (sleep)	6924 ± 2331	8578 ± 4741

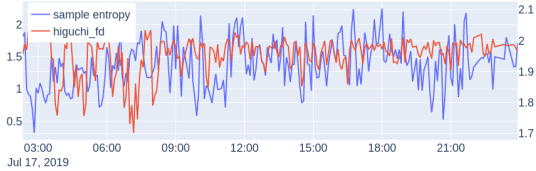
Table B.1: Demographics information of controls and patients at the time of recruitment, illness information, and amount of recorded data for each group during wakefulness and sleep. There were no significant differences for the recorded data (tested with Student’s t-test and Shapiro-Wilk for normality).



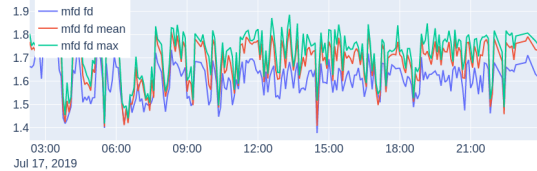
(a) Energy (STE) of the euclidean norm of *acc* and *gyr*.



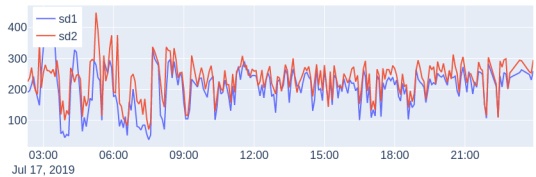
(b) Normalized LF and HF power of HRV.



(c) Sample entropy and Higuchi fractal dimension of HRV.



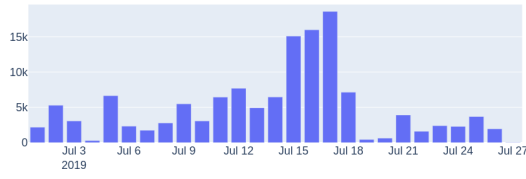
(d) MFD fractal dimension, mean MFD and max MFD of HRV.



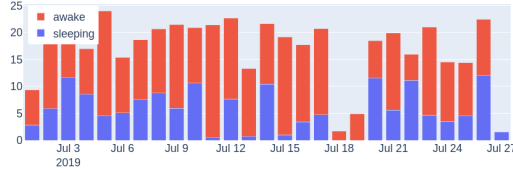
(e) Poincare SD1 and SD2 of HRV.

Figure B.1: Examples of features considered in this work during one subject’s full day.

were collected at a rate of 5Hz (if a new beat is not detected the watch duplicates the last obtained value). Using the Tizen API provided by the smartwatch, we also collected information about the sleep schedule of the subjects and their steps at aggregated intervals of 10 minutes.

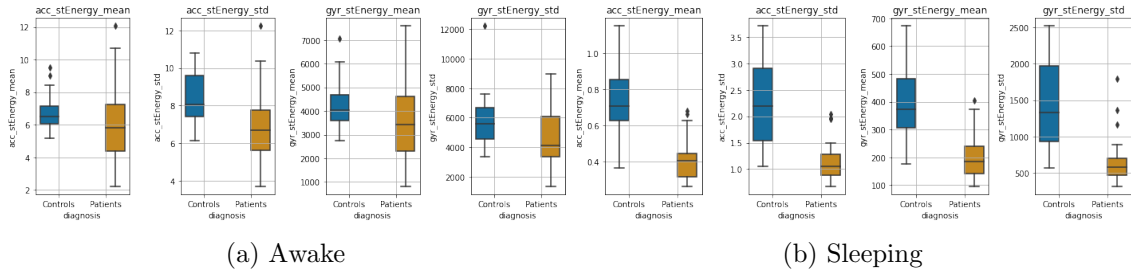


(a) Steps per day.



(b) Hours spent sleeping and awake.

Figure B.2: Steps per day and hours spent sleeping and awake during one month of a subject’s recordings.



(a) Awake

(b) Sleeping

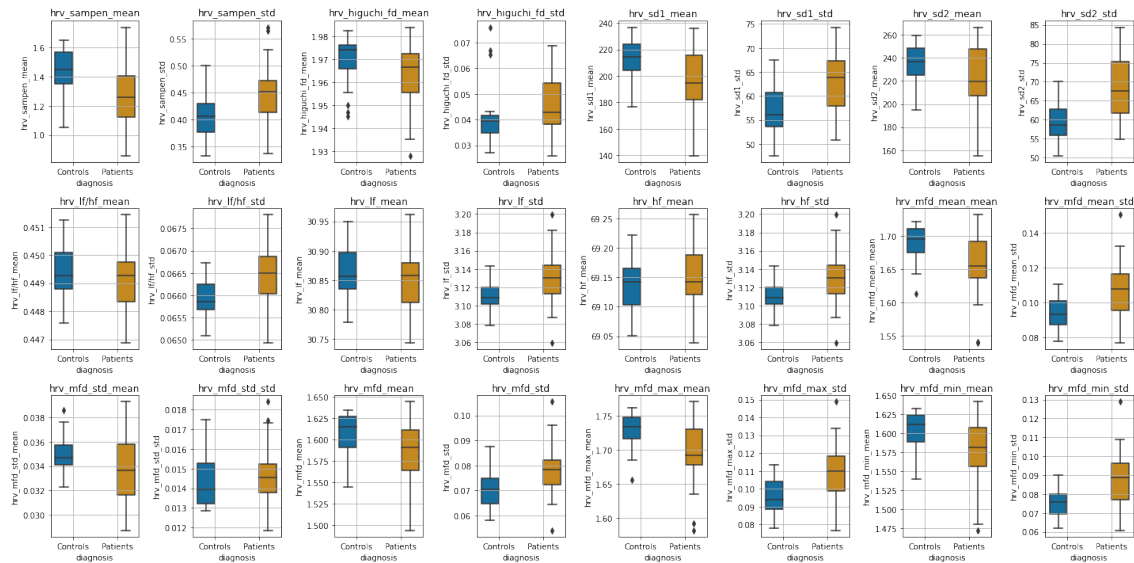
Figure B.3: Boxplots for *accelerometer* and *gyroscope* features of controls (in blue) and patients (in light brown) while (a) awake and (b) asleep. The bold line represents the median, the boxes extend between the 1st and 3rd quartile, whiskers extend to the lowest and highest datum within 1.5 times the inter-quartile range (IQR) of the 1st and 3rd quartile respectively, and outliers are shown as diamonds.

B.3 Feature Extraction and Data PreProcessing

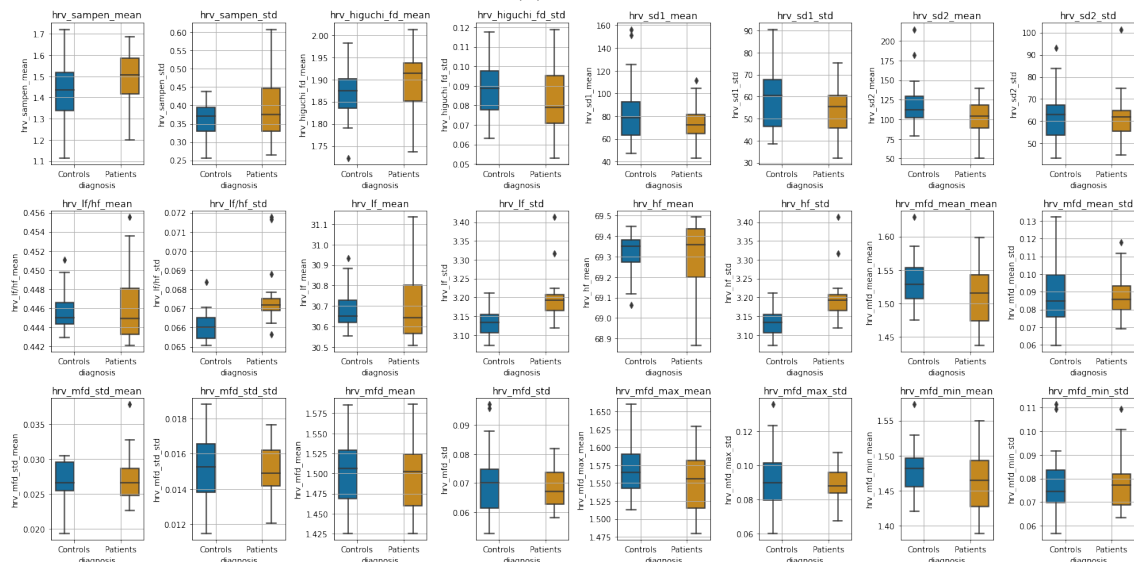
Data Preprocessing

Short-time analysis of signals using windowing is a traditional signal processing method. In the short-time analysis, we assume the process under which the data are generated to be stationary. Drawing power from these techniques, but largely increasing the time scale, we proceeded to perform “short-time” analysis in windows of 5 minutes for both movement and HRV data. Five (5) minutes intervals have been found to hold important information for distinguishing short-term patterns in a previous study [Retsinas et al., 2020].

Preprocessing of Heart-Rate Variability The heart rate variability (HRV) sequence from the 5Hz signal was obtained by dropping identical consecutive values. We also removed RR intervals larger than 2000ms and smaller than 300ms, considered as artifacts, and replaced possible non-detected pulses with linear interpolation. Finally, from the preprocessed interval, we retain the first 90% values the first 4.5 minutes (90%) of RR intervals for feature extraction, in order to mitigate the effect of different percentages of valid measurements across different intervals.



(a) Awake



(b) Sleeping

Figure B.4: Boxplots for *heart rate variability* features of controls and patients while awake (top rows) and asleep (bottom rows). The bold line represents the median, the boxes extend between the 1st and 3rd quartile, whiskers extend to the lowest and highest datum within 1.5 times the inter-quartile range (IQR) of the 1st and 3rd quartile respectively, and outliers are shown as diamonds.

Preprocessing of Accelerometer & Gyroscope In the accelerometer and gyroscope sensors, we replace missing values with the nearest interpolation and consider for feature extractions intervals with no more than 50 missing values (for comparison, a 5-minute interval has ideally 6000 values). As with the RR intervals, we retain the first 5940 (99%) samples from each preprocessed interval. We also mitigate the inherent noise in these sensors by applying high-frequency wavelet denoising [Vantuch, 2018]. The mean and standard deviation of the number of intervals for each user is reported in Table B.1.

Feature Extraction

We consider the following features; examples are shown in Fig. 1 (during one day of monitoring a subject):

Energy The energy (STE) of the euclidean norm of *acc* and *gyr* is extracted (since they are measured triaxially). We use these features as an objective measure of physical activity and general movement behavior.

Spectral features Power Spectral Density (PSD) is a common and powerful frequency-domain method for analysis of HRV describing the relative energy of the signal's cyclic fluctuations, managing to decompose the HRV signal to the sum of its sine and cosine components; allowing this way superimposed periodicities to be unraveled. Medical studies split the HRV spectrum in four frequency bands: ultra-low-frequency (ULF ≤ 0.003 Hz), very-low-frequency (VLF 0.0033–0.04 Hz), low-frequency (LF 0.04–0.15 Hz), and high-frequency (HF 0.15–0.40 Hz)[Shaffer and Ginsberg, 2017]. Since HRV is, by definition, a non-uniformly sampled signal, we perform spectral analysis using the Lomb-Scargle (LS) periodogram [Scargle, 1982].

The Lomb-Scargle periodogram is a method of power spectrum estimation that can be directly applied to non-uniformly sampled signals, and as a result, it is appropriate for HRV measurements. The periodogram is defined as:

$$P_{LS}(\Omega) = \frac{1}{2} \left\{ \frac{\left[\sum_{n=0}^{N-1} x[n] \cos(\Omega(t_n - \tau)) \right]^2}{\sum_{n=0}^{N-1} \cos^2(\Omega(t_n - \tau))} + \frac{\left[\sum_{n=0}^{N-1} x[n] \sin(\Omega(t_n - \tau)) \right]^2}{\sum_{n=0}^{N-1} \sin^2(\Omega(t_n - \tau))} \right\}, \quad (\text{B.1})$$

where τ is given by:

$$\tau = \frac{1}{2\Omega} \tan^{-1} \left(\frac{\sum_{n=0}^{N-1} \sin(2\Omega t_n)}{\sum_{n=0}^{N-1} \cos(2\Omega t_n)} \right), \quad (\text{B.2})$$

and Ω are the angular frequencies (*rad/sec*), t_n the time (*sec*) at which the signal was sampled, and $x[n]$ the value of the signal at time t_n . Using the LS periodogram, we extract for each interval the normalized power in two bands: LF and HF, as well as the ratio LF-to-HF.

Sample Entropy Nonlinear methods treat the extracted time series as the output of a nonlinear system. A typical characteristic of a nonlinear system is its complexity. The first measure of complexity we consider is the sample entropy (SampEn). Sample entropy is a measure of the rate of information generated by the system, and it has been considered to be an improved version of the approximate entropy [Richman and Moorman, 2000], due to its unbiased nature.

Higuchi Fractal Dimension Multiple algorithms have been proposed for measuring the fractal dimension of time series. In this work, we use the Higuchi fractal dimension [Higuchi, 1988], which has been used extensively in neurophysiology due to its simplicity and speed [Al-Nuaimi et al., 2017, Kesić and Spasić, 2016, Khoa et al., 2012].

Multiscale Fractal Dimension (MFD) is an efficient algorithm [Maragos, 1994] that measures the short-time fractal dimension, based on the Minkowski-Bouligand dimension [Falconer, 2004]. The algorithm is measuring the short time fractal dimension

using nonlinear multiscale morphological filters that can create geometrical covers around the graph of a signal, whose fractal dimension D can be found by:

$$D = \lim_{s \rightarrow 0} \frac{\log[\text{Area of dilated graph by disks of radius } s/s^2]}{\log(1/s)}. \quad (\text{B.3})$$

As known, D is between 1 and 2 for one-dimensional signals, and the larger the D is, the larger the amount of geometrical fragmentation of the signal. In practice, real-world signals do not have the same structure over different scales; hence D is computed by fitting a line to the log-log data of Eq. B.3 over a small scale window that can move along the s axis and thus create a profile of local *multiscale fractal dimensions (MFDs)* $D(s, t)$ at each time location t of the signal frame. By measuring the MFD we are able to examine the complexity and fragmentation of the signals at multiple scales, thus creating a profile of local MFDs at each time location. In general, the short-time fractal dimension at the smallest discrete scale ($s = 1$) has been found to provide some discrimination among various events. At higher scales, the MFD profile can also offer additional information that could help further the discrimination; more details about the algorithm can be found in [Maragos, 1994]. For this reason, we summarized the short-time measured MFD profiles by taking the following statistics: fd[1] (the fractal dimension), min, max, mean, and std for the 5 minutes HRV data.

Poincare plot measures The Poincare plot [Brennan et al., 2001] is a kind of recurrence plot where each sample of a time series is plotted against the previous, and then an ellipse is fitted on the scatter plot. The width of the ellipse (SD1) is a measure of short-term HRV, while the length (SD2) is a measure of long-term HRV.

Feature Aggregation Using the information from the sleep schedule of each subject, we split the intervals into two groups; one corresponding to intervals during sleep and one during wakefulness. We then calculated the mean and standard deviation (std) overall individual intervals, resulting in two values for each subject and feature type; and a total of 28 features. Significance tests using the Student's t-test showed no significant differences between the recorded movement and HRV intervals for each group and state (i.e., sleeping and awake). Normality was tested with Shapiro-Wilk test [Shapiro and Wilk, 1965].

Sleep/Wake Ratio and Steps In addition to the above features, we also extracted for each subject the mean and standard deviation of his/her sleep/wake ratio and the mean number of steps per day. Since the number of recorded hours each day fluctuates, for these features, we use only days with at least 20 recorded hours (no significant difference found between the number of days for controls and patients using Mann-Whitney U testing [Mann and Whitney, 1947], since the normality assumption was violated).

Figure B.1 shows examples of the extracted features, during one day of monitoring a subject, while Figure B.2 shows the steps per day and sleep/wake cycle during one month of monitoring.

B.4 Experimental Results

Figure B.3 shows boxplots of the features extracted from the accelerometer and gyroscope data during wakefulness and sleeping, while in Fig. B.4 boxplots of the hrv features are presented for the two states, respectively. Due to the differences observed perceptually between the distributions in most features, we tested for significant difference between distributions (the null hypothesis being that the two distributions are the same) using two-tailed non-parametric Mann-Whitney U tests [Mann and Whitney, 1947]. We adjusted for p-values using the Benjamini-Hochberg (BH) procedure [Benjamini and Hochberg,

	feature	Wakefulness			Sleeping		
		Controls	Patients	p value	Controls	Patients	p value
acc	STE mean	6.517 (1.058)	5.832 (2.902)	0.15	0.708 (0.228)	0.406 (0.129)	< 0.001
	STE std	8.065 (2.174)	6.694 (2.147)	0.02	2.199 (1.375)	1.057 (0.393)	< 0.001
gyr	STE mean	4045 (1080)	3431 (2313)	0.14	372 (177.542)	185.166 (97.117)	< 0.001
	STE std	5572 (2125)	4110 (2728)	0.05	1324 (1032)	578 (235.108)	< 0.001
hrv	sampen mean	1.446 (0.217)	1.260 (0.281)	0.03	1.435 (0.180)	1.505 (0.169)	0.14
	sampen std	0.407 (0.052)	0.452 (0.059)	0.03	0.370 (0.063)	0.376 (0.117)	0.61
	higuchi mean	1.974 (0.010)	1.966 (0.017)	0.07	1.875 (0.067)	1.915 (0.086)	0.18
	higuchi std	0.040 (0.007)	0.043 (0.016)	0.17	0.089 (0.020)	0.079 (0.024)	0.41
	sd1 mean	214.040 (20.128)	194.322 (33.829)	0.08	78.814 (29.462)	72.437 (16.211)	0.61
	sd1 std	56.058 (7.166)	63.894 (9.414)	0.02	60.625 (21.202)	55.604 (14.806)	0.45
	sd2 mean	237.053 (23.944)	219.853 (41.005)	0.14	112.232 (26.737)	104.269 (29.907)	0.32
	sd2 std	58.511 (6.954)	67.642 (13.689)	0.01	63.169 (13.386)	61.827 (8.968)	0.94
	lf/hf mean	0.449 (0.001)	0.449 (0.001)	0.48	0.445 (0.002)	0.445 (0.005)	0.93
	lf/hf std	0.066 (0.001)	0.067 (0.001)	0.02	0.066 (0.001)	0.067 (0.001)	< 0.001
	lf mean	30.857 (0.062)	30.858 (0.067)	0.43	30.652 (0.107)	30.644 (0.236)	0.84
	lf std	3.109 (0.018)	3.130 (0.031)	0.02	3.134 (0.047)	3.192 (0.043)	< 0.001
	hf mean	69.143 (0.062)	69.142 (0.067)	0.43	69.348 (0.107)	69.356 (0.236)	0.84
	hf std	3.109 (0.018)	3.130 (0.031)	0.02	3.134 (0.047)	3.192 (0.043)	< 0.001
	mfd mean mean	1.696 (0.035)	1.655 (0.055)	0.05	1.529 (0.046)	1.516 (0.069)	0.26
	mfd mean std	0.093 (0.014)	0.108 (0.021)	0.01	0.085 (0.023)	0.086 (0.013)	0.93
	mfd std mean	0.035 (0.002)	0.034 (0.004)	0.08	0.027 (0.004)	0.027 (0.004)	0.97
	mfd std std	0.014 (0.002)	0.015 (0.001)	0.34	0.015 (0.003)	0.015 (0.002)	0.93
	mfd mean	1.614 (0.036)	1.590 (0.047)	0.08	1.506 (0.060)	1.502 (0.065)	0.94
	mfd std	0.071 (0.010)	0.078 (0.010)	0.03	0.070 (0.014)	0.067 (0.011)	0.93
mfd max mean	1.734 (0.032)	1.692 (0.053)	0.06	1.565 (0.047)	1.557 (0.066)	0.36	
mfd max std	0.094 (0.016)	0.110 (0.020)	0.02	0.090 (0.022)	0.088 (0.012)	0.80	
mfd min mean	1.612 (0.035)	1.582 (0.051)	0.05	1.481 (0.040)	1.465 (0.066)	0.41	
mfd min std	0.076 (0.010)	0.089 (0.020)	0.01	0.075 (0.014)	0.077 (0.013)	0.97	
walk	steps mean	7054 (2358)	3960 (2928)	0.01	–	–	–
	steps std	3513 (1505)	2755 (756)	0.05	–	–	–
sleep	ratio mean	–	–	–	0.579 (0.107)	0.886 (0.471)	< 0.001
	ratio std	–	–	–	0.240 (0.149)	0.389 (0.304)	0.01

Table B.2: Statistical difference analysis using Mann-Whitney U-tests with BH correction in each state (wakefulness, sleeping). Bold values denote significance at the 95% confidence levels. For each group the median and the IQR (in parenthesis) is shown for each feature.

1995]. Due to the nature of our exploratory study, BH was preferred over the more strict Family-Wise Error Rate methods [Chen et al., 2017]. Table B.2 shows the results of Mann-Whitney U tests for all features examined in this study.

B.4.1 Wakefulness comparison

During wakefulness, the features that pertain to movements appear to present more variability in the patient group when compared to the controls, as shown in Fig. B.3a. The same appears to be also valid for some nonlinear HRV features, as for instance SampEn mean, Higuchi mean and std, SD1 and SD2, various statistics extracted from the MDF profile, and some of the frequency domain features, as can be seen in Fig. B.4a. Additionally, the significance testing, presented in Table B.2 showed significant distribution differences in the standard deviation of *acc* and *gyr* short-time energy, the *mean* and *std* of SampEn, the *std* of SD1 and SD2, the *std* of LF, HF and LF-to-HF ratio as well as the MFD statistics related to *standard deviation* as well as to *mean*, *max* and *min std*. The other features failed to reject the null hypothesis.

B.4.2 Sleep comparison

Similarly, Fig. B.3b presents the accelerometer and gyroscope feature distributions for each group during sleeping, while Fig. B.4b shows the distributions of HRV features. It is evident that especially the movement-related features present a significant difference,

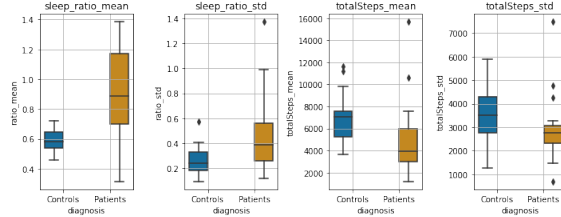


Figure B.5: Boxplots of sleep/wake ratio and steps per day (mean-std).

which is also verified in the Mann-Whitney U test results shown in Table B.2. The mean of the sample entropy among others also appears to be different (large variations), however the null hypothesis could not be rejected, possibly due to p-value adjustments for multiple testing. From the rest of the features, the *std* of LF, HF, and their ratio were found to differ significantly.

B.4.3 Sleep-wake ratio and total steps

Finally, Fig. B.5 shows the boxplots of the statistics of steps per day and sleep/wake ratio for the two groups. We observe a large significant difference between both the distributions of the *mean* and *std* of the sleep/wake ratio ($p < 0.001$ and $p = 0.01$, respectively), as well as the mean and std of total steps per day ($p = 0.01$ and $p = 0.05$ respectively).

B.5 Discussion

Our goal with the statistical analysis in this work is to exploit various traditional but also less-known and, at the same time, more novel signal processing techniques to identify common markers/features that differ drastically when a person has a psychotic disorder. These markers could prove useful in predicting potential relapses in these patients.

Our findings have shown that patients tend to behave with greater variability and present large outliers – some behave close to controls, while others might show extreme values. During wakefulness, even though the mean energy did not differ when compared to controls, the standard deviation showed a significant difference, indicating that patients tend to depict large variations in their movement behavior. On the contrary, during sleeping the patients presented a small mean and standard deviation of the energy in each of their sleeping intervals compared to the controls. We should note however that the observed differences in sleep between the two groups could be attributed to medication administered to patients, which possibly causes variability in sleep duration as well.

Some of the nonlinear features that were measured for the HRV data showed significant differences in the distributions between controls and patients, i.e., during wakefulness, as seen in Table B.2, such features are the mean and standard deviation of the sample entropy, as well as various statistics derived from the MFD analysis. Furthermore, the standard deviation of the normalized low and high-frequency bands of the HRV, as well as their ratio, were found to differ significantly both during wakefulness and sleeping. During sleeping, we did not find any other measurements of HRV to differ significantly. Finally, the sleep ratio of the two groups, as well as the mean and std of the number of steps per day, presented significant variation between the two groups.

The main merits of our work are two-fold: First, compared to previous similar studies, which have mostly lasted for a few weeks, our study has already been going on for more than a year with the goal to obtain two years of continuous monitoring of patients with

psychotic disorders. To do this, we employ a commercial off-the-shelf smartwatch that has been acknowledged by our volunteers to be comfortable, while patients are willing to insert it into their daily lives routine. Second, we show how traditional short-time-analysis combined with common but also more complex and novel features, such as the MFD features that depicted significant differences in wakefulness data, can be employed to identify biomarkers and present large inter-group variabilities between healthy controls and patients, paving a way towards both acquiring clinical insights on psychotic disorders, but also exploring the capabilities of these markers to predict relapses.

B.5.1 Conclusion

In this appendix, we identified markers that differentiate between healthy controls and people with psychotic disorders. To this end, we have specifically collected a large amount of physical activity and autonomic function data from wearable devices. Statistical analysis between the two groups, through their descriptive statistics, indicated significant differences regarding the movement behavior and in some markers of cardiac function during wakefulness and sleeping. In future analyses, we also intend to account for the effects of antipsychotics and/or other medications administered to patients, as well as other factors that differ in the two samples, such as smoker/non-smokers percentages. Finally, we aim to explore the capabilities of such markers to predict psychotic relapses and adverse drug effects.

Appendix C

List of Publications

Publications in Peer-reviewed International Journals

- [1] N. Efthymiou, P. P. Filntisis, P. Koutras, A. Tsiami, J. Hadfield, G. Potamianos, and P. Maragos, “[Childbot: Multi-robot perception and interaction with children](#),” *Robotics and Autonomous Systems*, vol. 150, p. 103975, 2022.
- [2] N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos, “[Visual Robotic Perception System with Incremental Learning for Child–Robot Interaction Scenarios](#),” *Technologies*, vol. 9, no. 4, p. 86, 2021.
- [3] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, “[Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction](#),” *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4011–4018, 2019.
- [4] P. P. Filntisis, A. Katsamanis, P. Tsiakoulis, and P. Maragos, “[Video-realistic expressive audio-visual speech synthesis for the Greek language](#),” *Speech Communication*, vol. 95, pp. 137–152, 2017.

Publications in Peer-reviewed International Conferences

- [1] P. P. Filntisis, G. Retsinas, F. Paraperas-Papantoniou, A. Katsamanis, A. Roussos, and P. Maragos, “[SPECTRE: Visual Speech-Informed Perceptual 3D Facial Expression Reconstruction from Videos](#),” in *Proc. IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) (under submission)*, 2023.
- [2] F. Paraperas Papantoniou, P. P. Filntisis, P. Maragos, and A. Roussos, “[Neural Emotion Director: Speech-preserving semantic control of facial expressions in ”in-the-wild” videos](#),” in *Proc. IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.
- [3] G. Retsinas, P. P. Filntisis, N. Kardaris, and P. Maragos, “[Attribute-based Gesture Recognition: Generalization to Unseen Classes](#),” in *Proc. Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, 2022, pp. 1–5.
- [4] C. O. Tze, P. P. Filntisis, A. Roussos, and P. Maragos, “[Cartoonized Anonymization of Sign Language Videos](#),” in *Proc. Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, 2022, pp. 1–5.

- [5] C. Garoufis, A. Zlatintsi, P. P. Filntisis, N. Efthymiou, E. Kalisperakis, V. Garyfalli, M. Lazaridi, N. Smyrnis, and P. Maragos, “Towards unsupervised subject-independent speech-based relapse detection in patients with psychosis using variational autoencoders,” in *Proc. European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1–5.
- [6] M. Panagiotou, A. Zlatintsi, P. P. Filntisis, A. Roumeliotis, N. Efthymiou, and P. Maragos, “A comparative study of autoencoder architectures for mental health analysis using wearable sensors data,” in *Proc. European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1–5.
- [7] P. P. Filntisis, N. Efthymiou, G. Potamianos, and P. Maragos, “[An Audiovisual Child Emotion Recognition System for Child-Robot Interaction Applications](#),” in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 791–795.
- [8] I. Pikoulis, P. P. Filntisis, and P. Maragos, “[Leveraging semantic scene characteristics and multi-stream convolutional architectures in a contextual approach for video-based visual emotion recognition in the wild](#),” in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2021, pp. 01–08.
- [9] P. Antoniadis, P. P. Filntisis, and P. Maragos, “[Exploiting emotional dependencies with graph convolutional networks for facial expression recognition](#),” in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2021, pp. 1–8.
- [10] P. P. Filntisis, N. Efthymiou, G. Potamianos, and P. Maragos, “[Emotion understanding in videos through body, context, and visual-semantic embedding loss](#),” in *Proc. European Conference on Computer Vision Workshops (ECCVW)*, 2020, pp. 747–755.
- [11] P. Antoniadis, I. Pikoulis, P. P. Filntisis, and P. Maragos, “[An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild](#),” in *Proc. IEEE / CVF Computer Vision and Pattern Recognition Conference Workshops (CVPRW)*, 2021, pp. 3645–3651.
- [12] N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos, “[A robotic edutainment framework for designing child-robot interaction scenarios](#),” in *Proc. Pervasive Technologies Related to Assistive Environments (PETRA)*, 2021, pp. 160–166.
- [13] G. Retsinas, P. P. Filntisis, N. Efthymiou, E. Theodosis, A. Zlatintsi, and P. Maragos, “[Person identification using deep convolutional neural networks on short-term signals from wearable sensors](#),” in *Proc. International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, 2020, pp. 3657–3661.
- [14] I. Maglogiannis, A. Zlatintsi, A. Menychtas, D. Papadimitos, P. P. Filntisis, N. Efthymiou, G. Retsinas, P. Tsanakas, and P. Maragos, “[An intelligent cloud-based platform for effective monitoring of patients with psychotic disorders](#),” in *Proc. International Federation for Information Processing (IFIP)*, 2020, pp. 293–307.
- [15] C. Garoufis, A. Zlatintsi, K. Kritsis, P. P. Filntisis, V. Katsouros, and P. Maragos, “[An environment for gestural interaction with 3d virtual musical instruments as an educational tool](#),” in *Proc. European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

- [16] 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. International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, 2018, pp. 6568–6572.
- [17] N. Efthymiou, P. Koutras, P. P. Filntisis, G. Potamianos, and P. Maragos, “[Multi-view fusion for action recognition in child-robot interaction](#),” in *Proc. International Conference on Image Processing (ICIP)*, 2018, pp. 455–459.
- [18] 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. International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4585–4592.
- [19] A. Zlatintsi, P. P. Filntisis, C. Garoufis, A. Tsiami, K. Kritsis, M. A. Kaliakatsos-Papakostas, A. Gkiokas, V. Katsouros, and P. Maragos, “[A web-based real-time kinect application for gestural interaction with virtual musical instruments](#),” in *Proc. AudioMostly*, 2018, pp. 1–6.
- [20] P. P. Filntisis, A. Katsamanis, and P. Maragos, “[Photorealistic adaptation and interpolation of facial expressions using HMMS and AAMS for audio-visual speech synthesis](#),” in *Proc. International Conference on Image Processing (ICIP)*, 2017, pp. 2941–2945.

Appendix D

List of this Thesis' Open Source Codes

- P. P. Filntisis, Code for the paper “Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction”, <https://github.com/filby89/body-face-emotion-recognition>. Source code for Chapter 2.
- P. P. Filntisis, PyTorch implementation of Multimodal Emotion Recognition, <https://github.com/filby89/multimodal-emotion-recognition>. Source code for Chapter 3.
- P. P. Filntisis, PyTorch code for the NTUA BEEU ECCV20220 Solution, <https://github.com/filby89/NTUA-BEEU-eccv2020>. Source code for Chapter 4.
- P. P. Filntisis, Code for the paper “Expressive Audiovisual Speech Synthesis for the Greek Language”, <https://github.com/filby89/expressive-audiovisual-speech-synthesis-GR>. Source code for Chapters 5 and 6.
- P. P. Filntisis, Pytorch implementation of “SPECTRE: Visual Speech-Informed Perceptual 3D Facial Expression Reconstruction from Videos”, <https://github.com/filby89/spectre>. Source code for Chapter 7 (SPECTRE).
- F. Paraperas-Papantoniou, Neural Emotion Director (NED) - Official Pytorch Implementation, <https://github.com/foivospar/NED>. Source code for Chapter 7 (NED).

Appendix E

List of this Thesis' Released Datasets

- P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, P. Maragos, BabyRobot Emotion Database (BRED), <https://zenodo.org/record/3233060>
- P. P. Filntisis, A. Katsamanis, P. Maragos, CVSP-Expressive AudioVisual Speech Synthesis Database (CVSP-EAV), <http://cvsp.cs.ntua.gr/research/eavtts/>

Appendix F

List of Awards and Grants

- Best Paper Finalist in Proc. IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) 2022 with "Neural Emotion Director: Speech-preserving semantic control of facial expressions in "in-the-wild" videos".
- First Place (Challenge Winner) in the Bodily Expressed Emotion Understanding Workshop of the European Conference in Computer Vision (ECCVW) with "Emotion Understanding in Videos Through Body, Context, and Visual-Semantic Embedding Loss".
- Travel Grant Awardee at the International Conference on Intelligent Robots and Systems (IROS) 2019 for "Fusing body posture with facial expressions for joint recognition of affect in child-robot interaction".
- Travel Grant Awardee at the International Conference on Image Processing (ICIP) 2017 for "Photorealistic adaptation and interpolation of facial expressions using HMMS and AAMS for audio-visual speech synthesis".

