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

Emotion Driven Speaker Verification

Επαλήθευση ομιλητή με χρήση συναισθήματος

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΙΟΥ ΚΑΤΣΙΡΟΥ

Επιβλέποντες: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής

Θεόδωρος Γιαννακόπουλος
Ερευνητής Β, ΕΚΕΦΕ Δημόκριτος

Αθήνα, Ιούλιος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομεας Σηματων, Ελεγχου και Ρομποτικης

Emotion Driven Speaker Verification

Επαλήθευση ομιλητή με χρήση συναισθήματος

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΙΟΥ ΚΑΤΣΙΡΟΥ

Επιβλέποντες: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής

Θεόδωρος Γιαννακόπουλος
Ερευνητής Β, ΕΚΕΦΕ Δημόκριτος

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19 Ιουλίου 2021.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Αλέξανδρος Ποταμιάνος	Θεόδωρος Γιαννακόπουλος	Κωνσταντίνος Τζαφέστας
Αναπληρωτής Καθηγητής	Ερευνητής Β, ΕΚΕΦΕ Δημόκριτος	Αναπληρωτής Καθηγητής

Αθήνα, Ιούλιος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομεας Σηματων, Ελεγχου και Ρομποτικης

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Δημήτρης Κατσίρος

19 Ιουλίου 2021

Ευχαριστίες

Ολοκληρώνοντας τις προπτυχιακές σπουδές μου θα ήθελα να ευχαριστήσω ιδιαιτέρως κάποιους ανθρώπους των οποίων ο ρόλος υπήρξε καθοριστικός όσον αφορά την πορεία μου.

Αρχικά θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου Θοδωρή Γιαννακόπουλο, με τον οποίο συνεργάστηκα καθ' όλη τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας. Οι συμβουλές του και οι συζητήσεις μας υπήρξαν καθοριστικές και καθόρισαν σε μεγάλο βαθμό την διεκπαιρέωση της παρούσας εργασίας.

Στη συνέχεια θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου Αλέξανδρο Ποταμιάνο, ο οποίος μου έδωσε γερά θεμέλια στον τομέα της Τεχνητής Νοημοσύνης μέσω των μαθημάτων του, χρήσιμες συμβουλές αλλά και την ευκαιρία να συνεργαστώ με το εργαστήριο του.

Θα ήθελα ακόμη να πω ένα μεγάλο ευχαριστώ σε όλα εκείνα τα άτομα που στάθηκαν πλάι μου τα χρόνια αυτά και μου χάρισαν αξέχαστες στιγμές. Κυρίως όμως θα ήθελα να ευχαριστήσω του γονείς μου Τάσο και Μαριάνθη και τα αδέρφια μου Άννα και Γιάννη, οι οποίοι όλα αυτά τα χρόνια υπήρξαν συνεχώς δίπλα μου και μου έδωσαν όλα τα απαραίτητα εφόδια προκειμένου να είμαι το άτομο που είμαι σήμερα.

Αθήνα, Ιούλιος 2021

Δημήτρης Κατσίρος

Περίληψη

Η Επαλήθευση Ομιλητή επιτρέπει την ταυτοποίηση μιας ισχυριζόμενης ταυτότητας από μετρήσεις σε ένα φωνητικό σήμα. Το συναίσθημα ωστόσο, ως ένας φυσικός και συχνά ακούσιος κωδικοποιητής της φωνής, διαθέτει μηχανισμούς υπεύθυνους για τη φωνητική διαμόρφωση της. Παρά την προσοχή που έχει κερδίσει το πεδίο όλα αυτά τα χρόνια, δεν έχει υπάρξει μεγάλη προσπάθεια προκειμένου να προσδιοριστούν οι σχέσεις μεταξύ των δύο αυτών αντικειμένων. Αν και φαινομενικά μακριά, το συναισθηματικό περιεχόμενο θα μπορούσε να έχει ένα τεράστιες επιπτώσεις στη διαδικασία διάκρισης ομιλητών.

Σε αυτή τη διπλωματική, διερευνούμε τη συσχέτιση μεταξύ επαλήθευσης ομιλητή και αναγνώρισης συναισθηματικού λόγου. Πρώτα απ ́όλα, δημιουργούμε διάφορα σετ συναισθηματικής αξιολόγησης, με στόχο το καθένα να παρακολουθεί διαφορετικά την επίδραση του συναισθήματος στο αντικείμενο επαλήθευσης ομιλητή. Σε μια προσπάθεια μείωσης ή ακόμη και εξάλειψης του αποτελέσματος προσπαθούμε να μεταφέρουμε συναισθηματική γνώση στο αντικείμενο. Για το σκοπό αυτό, εφαρμόζουμε τέσσερις διαφορετικές αρχιτεκτονικές, όπου η καθεμία από αυτές, χειρίζεται τη συναισθηματική πληροφορία με διαφορετικό τρόπο. Κατόπιν, εξετάζουμε την απόδοση των μοντέλων μας στα σετ συναισθηματικής αξιολόγησης.

Τα αποτελέσματά μας υποδηλώνουν ότι η συναισθηματική πληροφορία έχει καθοριστικό ρόλο στην επαλήθευση ομιλητών. Ακόμη και σε χαμηλή ένταση, το συναίσθημα τόσο στην πρόταση εγγραφής όσο και στην πρόταση επαλήθευσης μπορεί να υποβαθμίσει σημαντικά την απόδοση ενός συστήματος. Επιπλέον, τα συναισθήματα έντονης έντασης, φαίνεται να επιδεινώνουν το αποτέλεσμα οδηγώντας σε πολύ φτωχά αποτελέσματα. Μεταξύ των επτά συναισθημάτων που εξετάστηκαν, διαπιστώνουμε ότι ο *θυμός* και ο *φόβος* είχαν το πιο αξιοσημείωτο αντίκτυπο.

Σε μια προσπάθεια αντιμετώπισης των προαναφερθέντων ζητημάτων εξετάζουμε την απόδοση των αρχιτεκτονικών μας με γνώση συναισθήματος. Τα αποτελέσματά μας δείχνουν ότι με την εφαρμογή κλασικών τεχνικών μεταφοράς μάθησης, μπορούμε να παρέχουμε μοντέλα ανθεκτικά σε συναισθηματικά φορτισμένο περιεχόμενο και ταυτόχρονα να αποδίδουμε πολύ καλύτερα στην ίδια τη διαδικασία της επαλήθευσης ομιλητή. Τελος, δοκιμάζουμε την υπόθεσή μας σχετικά με την παροχή ίδιου συναισθήματος κατά την πρόταση εγγραφής και επαλήθευσης και παρατηρούμε σημαντική σχετική αύξηση περίπου 20%, ανεξάρτητα από τη συναισθηματική προ-εκπαίδευση.

Συνολικά, μπορούμε να αποτυπώσουμε μια ισχυρή σχέση μεταξύ διάκρισης ομιλητών και συναισθηματικού περιεχομένου. Υποστηρίζουμε ότι ο έλεγχος του συναισθηματικού περιεχομένου είναι απαραίτητος για την καλή απόδοση ενός μοντέλου, ειδικά για πραγματικά σενάρια, όπου το συναίσθημα είναι ενεργά παρόν. Συνεπώς, μπορούμε εφαρμόζοντας παραδοσιακές τεχνικές εκμάθησης μεταφοράς από το αντικείμενο αναγνώρισης συναισθημάτων ομιλίας στο

αντικείμενο της επαλήθευσης ομιλητή, να μειώσουμε τη συναισθηματική επιρροή και να βελτιώσουμε την συνολική αποτελεσματικότητα των μοντέλων μας .

Λέξεις Κλειδιά

Επαλήθευση Ομιλητή, Επαλήθευση Ομιλητή Ανεξαρτήτως Κειμένου , Αναγνώριση Συναισθημάτων Ομιλίας, Επαλήθευση Ομιλητή με Χρήση Συναισθήματος

Abstract

Speaker Verification (SV) enables the authentication of a claimed identity from measurements on a voice signal. Emotion as a natural and often involuntary encoder of voice, has the mechanisms responsible for vocal modulation. Despite the attention that the field has gained over the years, little effort has been made in order to identify the relations between these two subjects. Although seemingly far, emotional content could have a huge impact on speaker discrimination.

In this thesis, we investigate the correlation between speaker verification and speech emotion recognition. First of all, we create various emotional evaluation sets, each one aiming to track differently the effect of emotion on the speaker verification task. In an attempt to decrease or even eliminate the effect we try to transfer emotional knowledge to our task. For this purpose, we implement four different architectures, each one of them, handling emotional information in a different manner. Then we examine our models' performance on the emotional evaluation sets.

Our results suggest that emotional information has a crucial role on speaker verification. Even on low intensity, emotion on both on enrollment and verification can significantly degrade a system's performance. On addition, emotions on strong intensity, seem to escalate the effect and ensue in poor results. Among the seven emotions examined, we find that, anger and fear were these having the most remarkable impact.

In an endeavor to address the aforementioned issues we examine the performance of our emotion-aware architectures. Our results indicate that by applying classic fine tuning techniques, we are able provide emotion robust models and at the same time perform much better on the speaker verification task. Last but not least, we test our hypothesis on providing same-emotion utterances on evaluation phase and we observe a relative improvement around 20%, irrespective of emotional pre-training.

Overall, we can capture a strong relation between speaker discrimination and emotional content. We contend that controlling emotional content is necessary for a model's robustness, especially for real life scenarios, where emotion is present. Ultimately, we can reduce the effect and improve our models performance by applying traditional transfer learning techniques from speech emotion recognition to speaker verification.

Keywords

Speaker Verification (SV), Text-Independent Speaker Verification (TISV), Speech Emotion Recognition (SER) , Emotion Driven Speaker Verification

Contents

List of Figures

List of Tables

Κεφάλαιο 0

Εκτεταμένη Περίληψη στα Ελληνικά

0.1 Εισαγωγή

0.1.1 Κίνητρο

Παρά τις πρόσφατες εξελίξεις στον τομέα της επαλήθευσης των ομιλητή, η παραγωγή ενιαίων, συμπαγών αναπαραστάσεων για τμήματα λόγου ομιλητών που να μπορούν να χρησιμοποιηθούν αποτελεσματικά σε θορυβώδη και μη περιερισμένες συνθήκες εξακολουθούν να αποτελούν σημαντική πρόκληση. Τέτοια συστήματα είναι πιθανό να είναι επιρρεπή στην ποικιλία των εκφράσεων των ομιλητών σε καθημερινή βάση. Το συναίσθημα ως φυσικός και συχνά ακούσιος κωδικοποιητής φωνής, διαθέτει μεταξύ άλλων, τους μηχανισμούς που είναι υπεύθυνοι για τη φωνητική διαμόρφωση του λόγου. Παρά την πολυπλοκότητά του και το γεγονός ότι κυριαρχεί στην καθημερινή ομιλία, η επίδραση του συνήθως θεωρείται αμελητέα και δεν λαμβάνεται ενερά υπόψη.

Δεδομένου του ότι τα περισσότερα συστήματα επαλήθευσης ομιλητή δεν λαμβάνουν υπόψη το συναισθηματικό περιεχόμενο, εγείρονται ερωτήματα σχετικά με τις αδυναμίες τέτοιων συστημάτων. Θα ήταν πολύ ενδιαφέρον να εμβαθύνουμε στην επίδραση του συναισθηματικού περιεχομένου στη διάκριση ομιλητών και στο εάν ορισμένα συναισθήματα έχουν ως αποτέλεσμα χειρότερη επίδοση συγκριτικά με άλλα.

Η Αναγνώριση Ομιλητή με Χρήση Συναισθήματος (στην αγγλική ορολογία Emotion Driven Speaker Verification) είναι ακριβώς η προσπάθεια χαρτογράφησης, για πρώτη φορά, του συναισθηματικού περιεχομένου σε ένα σύστημα επαλήθευσης ομιλητή. Πραγματοποιείται μία προσπάθεια προκειμένου να διευκρινιστούν οι σχέσεις μεταξύ των αντικειμένων της Επαλήθευσης Ομιλητή και της Αναγνώρισης Συναισθημάτων Ομιλίας. Το πρόβλημα είναι πολύ απαιτητικό λόγω των ακόλουθων δύο ιδιοτήτων. Πρώτον, τα χαρακτηριστικά για τη διάκριση ομιλητών δεν είναι μονοσήμαντα ορισμένα. Δεύτερον, η αναγνώριση συναισθημάτων είναι ένα δύσκολο έργο ως αυτό καθ΄ αυτό, καθώς τα συναισθήματα διαφέρουν από άνθρωπο σε άνθρωπο.

0.1.2 Προσέγγιση και Συνεισφορά

Σε αυτήν τη διπλωματική εργασία διερευνούμε πώς τα διαφορετικά συναισθήματα επηρεάζουν ένα σύστημα επαλήθευσης ομιλητή ανεξαρτήτως κειμένου. Πιο συγκεκριμένα διεξάγουμε πολλαπλά πειράματα για να κατανοήσουμε πώς το συναισθηματικό περιεχόμενο επηρεάζει

την επαλήθευση ομιλητή, πώς επηρεάζει κάθε συναίσθημα χωριστά σε αυτήν τη διαδικασία, και πώς θα μπορούσαμε να αντιμετωπίσουμε αυτή την επιρροή, χρησιμοποιώντας τη συναισθηματική γνώση υπέρ μας. Πρώτα απ΄όλα, εκπαιδεύουμε ένα μοντέλο αναγνώρισης συναισθημάτων ομιλίας (στην αγγλική ορολογία speech emotion recognition SER) στο σύνολο δεδομένων του IEMOCAP. Στη συνέχεια, δημιουργούμε τέσσερις διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων. Κάθε μία απο αυτές προσπαθεί να λύσει διαφορετικά, το πρόβλημα της αποτελεσματικής μεταφοράς συναισθηματικής γνώσης από το μοντέλο αναγνώρισης συναισθημάτων ομιλίας. Θεωρούμε την πρώτη μας αρχιτεκτονική ως το βασικό μοντέλο το οποίο δεν διαθέτει καμία συναισθηματική γνώση κατα την εκπαίδευση του. Η δεύτερη αρχιτεκτονική μας στοχεύει στο να αποκτήσει συναισθηματική γνώση μέσω της προσεκτικής επανεκπαίδευσης απο το αντικείμενο της αναγνώρισης συναισθημάτων. Το τρίτο μας μοντέλο προσπαθεί να μεταφέρει συναισθηματική γνώση χωρίς πλήρη επανεκπαίδευση, αλλά κρατώντας σταθερά τα σημεία όπου το μοντέλο αναγνώρισης συναισθημάτων επικεντρωνόταν στο αρχικό σήμα. Η τέταρτη μας αρχιτεκτονική πρόκειται για ένα μοντέλο σύντηξης (στην αγγλική ορολογία fusion) και χωρίζεται σε δύο μέρη. Το πρώτο απο αυτά είναι ένα προεκπαιδευμένο δίκτυο αναγνώρισης συναισθημάτων, το οποίο στόχο έχει να προσφέρει τις συναισθηματικές αναπαραστάσεις που παράγει στο δεύτερο μοντέλο, που εκπαιδεύεται καθαρά στο αντικείμενο της επαλήθευσης ομιλητή. Θα πρέπει να αναφέρουμε ότι όλα τα μοντέλα μας αξιοποιούν το σύνολο δεδομένων του VoxCeleb προκειμένου να έχουν μια αποδοτική εκπαίδευση.

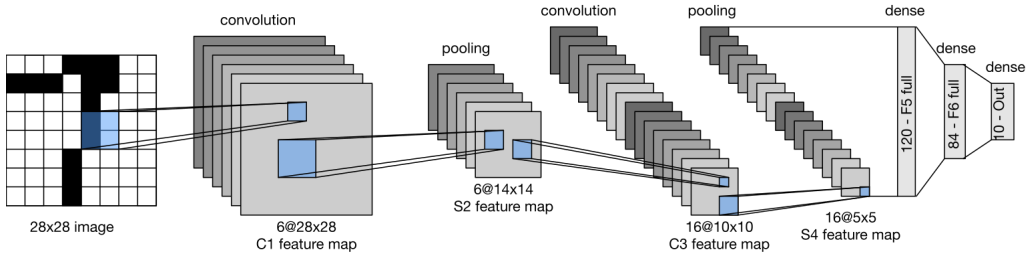
Το επόμενο μας βήμα είναι η δημιουργία ενός συνόλου απο συναισθηματικά πειράματα. Τα πειράματα αυτά αποτελούνται απο μία πολύ συγκεκριμένη και προσεκτική οργάνωση, προκειμένου πρώτα να ποιοτικοποιήσουμε και στη συνέχεια να ποσοτικοποιήσουμε την επίδραση του συναισθηματικού περιεχομένου στο λόγο.

Στη συνέχεια εξετάζουμε την απόδοση των μοντέλων μας στα προαναφερθέντα συναισθηματικά σύνολα αξιολόγησης και εκτιμούμε την αντοχή τους. Για κάθε πείραμα αξιολογούμε κάθε μοντέλο χωριστά και στο τέλος συγκρίνουμε κάθε ένα με τα αποτελέσματα του μοντέλου χωρίς συναισθηματική γνώση.

Τα αποτελέσματα μας υποδεικνύουν ότι το συναίσθημα έχει ένα σοβαρότατο ρόλο στην επαλήθευση ομιλητή χωρίς προκαθορισμένες προτάσεις. Αρχικά, αποτυπώνουμε ότι το έντονο συναίσθημα μπορεί να οδηγήσει σε τεράστιες μειώσεις της απόδοσης, ανεξάρτητα απο το συναίσθημα στην πρόταση εγγραφής. Παρατηρούμε ότι διαφορετικά συναισθήματα στην πρόταση εγγραφής και επαλήθευσης μπορούν να μεγενθύνουν την επίδραση και να φτάσουν μέχρι και 30% σε EER. Μετά απο προσεκτικό έλεγχο της συνεισφοράς του κάθε συναισθήματος χωριστά στα αποτελέσματα, παρατηρούμε ότι ο *φόβος* και ο *θυμός* τείνουν να αποδίδουν χειρότερα απο όλα τα υπόλοιπα. Πιο συγκεκριμένα, το λάθος στα μοντέλα χωρίς συναισθηματική γνώση, πλησιάζει εκείνο της τυχαίας πρόβλεψης με EER κοντά στο 40%. Ακολούθως, εξετάσαμε εάν η παρουσία ίδιου συναισθήματος κατά την στην πρόταση εγγραφής και επαλήθευσης μπορούν να βοηθήσουν προκειμένου τα μοντέλα μας να είναι πιο ικανά στη διάκριση ομιλητών. Τα αποτελέσματα μας υποδεικνύουν ότι η διασφάλιση κοινού συναισθήματος, μπορεί να βελτιώσει δραστικά την απόδοση του συστήματος ανεξάρτητα απο το αν έχει προηγηθεί κάποιου είδους συναισθηματική εκπαίδευση. Τέλος, επιδεικνύουμε ότι αξιοποιώντας παραδοσιακές τεχνικές μεταφοράς μάθησης, είναι δυνατόν να δημιουργήσουμε μοντέλα με επίγνωση συναισθήματος που υπερτερούν των κλασικών συστημάτων.

0.2 Θεωρητικό Υπόβαθρο

Συνελικτικά Νευρωνικά Δίκτυα



Σχήμα 1: *Ένα Συνελικτικό Νευρωνικό Δίκτυο [1]*

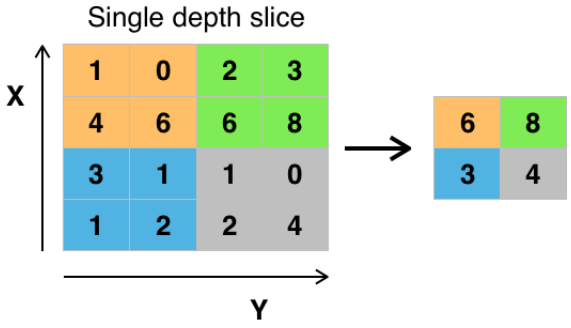
Τα Συνελικτικά Νευρωνικά Δίκτυα (στην αγγλική ορολογία convolutional neural networks ή για συντομία CNN) είναι μια συγκεκριμένη κατηγορία βαθιών νευρωνικών δικτύων, που είναι εξειδικευμένη στην επεξεργασία δεδομένων που έχουν τοπολογία πλέγματος, όπως για παράδειγμα οι εικόνες. Τα δίκτυα αυτά θεωρούνται μια από τις μεγαλύτερες βιολογικές εμπνεύσεις στην τεχνητή νοημοσύνη, καθώς οι βασικές έννοιες σχεδιασμού τους δανείζονται στοιχεία από τη νευροεπιστήμη και ειδικά από την οργάνωση του ανθρώπινου οπτικού φλοιού. Ένα τέτοιο δίκτυο έχει σχεδιαστεί προκειμένου να μιμείται τον ανθρώπινο οπτικό φλοιό, μέσω της εφαρμογής συνελικτικών λειτουργιών στην εικόνα εισόδου, με χρήση πολλαπλών φίλτρων χαμηλής διάστασης.

Τα Συνελικτικά Νευρωνικά Δίκτυα είναι ομαλοποιημένες εκδόσεις των Νευρωνικών Δικτύων Εμπρόσθιας Τροφοδότησης (στην αγγλική ορολογία feed forward neural networks ή για συντομία FFNN. Τα FFNN συνήθως σημαίνουν ένα πλήρως συνδεδεμένο δίκτυο, δηλαδή κάθε νευρώνας σε ένα στρώμα συνδέεται με όλους τους νευρώνες στο επόμενο στρώμα. Η πλήρης συνδεσιμότητα αυτών των δικτύων τα καθιστά επιρρεπή σε υπερβολική εφαρμογή (στην αγγλική ορολογία overfitting) σε δεδομένα. Τα ΝΝ παίρνουν μία διαφορετική προσέγγιση για τη ομαλοποίηση των δεδομένων και τη μείωση της πολυπλοκότητας. Πιο συγκεκριμένα, επωφελούνται της ύπαρξης ιεραρχικών προτύπων στα δεδομένα και σχηματίζουν μοτίβα αυξανόμενης πολυπλοκότητας χρησιμοποιώντας μικρότερα και απλούστερα μοτίβα στα φίλτρα τους. Αυτό τους προσφέρει επίσης ανθεκτικότητα στις χωρικές μετατοπίσεων και στην χαρτογράφηση στόχων. Επομένως, σε κλίμακα συνδεσιμότητας και πολυπλοκότητας, τα CNN αποτελούν την καλύτερη λύση.

Οι πιο σημαντικές πτυχές της αρχιτεκτονικής των Συνελικτικών Νευρωνικών Δικτύων είναι τα ακόλουθα επίπεδα:

- Τα συνελικτικά επίπεδα (στην αγγιλική ορολογία convolutional layers) σε ένα ΝΝ εφαρμόζουν συστηματικά φίλτρα που έχουν μάθει στις εικόνες εισόδου για να δημιουργήσουν χάρτες χαρακτηριστικών που συνοψίζουν την παρουσία αυτών των χαρακτηριστικών.

Αυτό μπορεί να αποδειχθεί πολύ αποτελεσματικό, καθώς η στοίβασή τους σε βαθιά μοντέλα επιτρέπει στο κλείσιμο των επιπέδων στην είσοδο για να μάθουν χαρακτηριστικά χαμηλού επιπέδου (π.χ. γραμμές), ενώ επίπεδα βαθύτερα στο μοντέλο μαθαίνουν υψηλής ποιότητας ή πιο αφηρημένα χαρακτηριστικά, όπως σχήματα ή συγκεκριμένα αντικείμενα.



Σχήμα 2: *Παράδειγμα ενός Max Pooling Layer [2]*

- Τα συγκεντρωτικά στρώματα (στην αγγλική ορολογία pooling layers σχηματίζουν ένα νέο στρώμα που προστίθεται μετά από κάθε συνελικτικό. Ο ρόλος τους είναι η μείωση του χωρικού μεγέθους των χαρακτηριστικών, ενώ ταυτόχρονα η μετατροπή τους σε χωρικά ανεξάρτητα. Τα πιο συχνά χρησιμοποιούμενα είναι τα λεγόμενα max polling και average pooling , τα οποία διαφέρουν ως προς τον τρόπο εφαρμογής της μείωσης διαστάσεων.

Αναγνώριση Συναισθήματος σε Ομιλία

Τα συναισθήματα έχουν εξαιρετικά σημαντικό ρόλο στην ψυχική ζωή του σύγχρονου ανθρώπου. Είναι ένα μέσο έκφρασης της προοπτικής ή της ψυχικής κατάστασης κάποιου, στους άλλους. Η Αναγνώριση Συναισθήματος σε Ομιλία (στην αγγλική ορολογία Speech Emotion Recognition ή για συντομία SER) μπορεί να οριστεί ως η εξαγωγή της συναισθηματικής κατάστασης του ομιλητή από το σήμα ομιλίας του. Υπάρχουν λίγα καθολικά συναισθήματα, όπως Ουδέτερο, Θυμός, Ευτυχία, Θλίψη, όπου οποιοδήποτε έξυπνο σύστημα με πεπερασμένους υπολογιστικούς πόρους μπορεί να εκπαιδευτεί για να αναγνωρίσει ή να συνθέσει όπως απαιτείται.

Αναγνώριση Ομιλητή

Η αναγνώριση ομιλητή είναι η διαδικασία ταυτοποίησης ενός ατόμου από χαρακτηριστικά της φωνής του. Το συγκεκριμένο αντικείμενο, έχει ιστορία που χρονολογείται περίπου τέσσερις δεκαετίες από το 2021. Χρησιμοποιεί τα ακουστικά χαρακτηριστικά της ομιλίας που έχει διαπιστωθεί ότι διαφέρουν μεταξύ ατόμων. Αυτά τα ακουστικά μοτίβα αντικατοπτρίζουν τόσο την ανατομία όσο και τα διάφορα πρότυπα συμπεριφοράς.

Αξίζει να σημειωθεί ότι υπάρχουν δύο μεγάλες υποκατηγορίες στο πεδίο αναγνώρισης ομιλητή: **η επαλήθευση και η αναγνώριση**. Στην περίπτωση όπου ο ομιλητής ισχυρίζεται

ότι έχει μία συγκεκριμένη ταυτότητα και η φωνή του χρησιμοποιείται για την επιβεβαίωση ή την απόρριψη αυτού του ισχυρισμού, πρόκειται για επαλήθευση ομιλητή (στα αγγλικά speaker verification ή SV). Αντίθετα, η αναγνώριση ομιλιτή (στα αγγλικά speaker identification ή SI) πρόκειται για τον προσδιορισμό της ταυτότητας ενός άγνωστου ομιλητή ανάμεσα σε πολλούς. Κατά μία έννοια, η επαλήθευση ομιλητή είναι μία 1:1 αντιστοιχία ενώ η αναγνώριση ομιλητή είναι μία αντιστοίχιση 1:N, καθώς η φωνή συγκρίνεται με πρότυπα πολλαπλών ατόμων.

Όσον αφορά την επαλήθευση ομιλητή, το πεδίο χωρίζεται σε δύο υποκατηγορίες: την εξαρτώμενη και τη μη εξαρτώμενη απο το κείμενο. Στην πρώτη κατηγορία το σύστημα καλείται να επαληθεύσει τον ομιλητή μέσω μίας γνωστής και προκαθορισμένης πρότασης. Αντίθετα, στη μη εξαρτώμενη επαλήθευση ο ομιλητής μπορεί να εισηγάγει οποιαδήποτε πρόταση στο σύστημα.

Αναφορικά με τη λειτουργία ενός συστήματος επαλήθευσης ομιλητή πρόκειται για ένα σύστημα που δέχεται ως είσοδο δύο προτάσεις. Η πρώτη πρόταση πρόκειται για την **πρόταση εγγραφής** (γνωστή στην αγγλική ορολογία και ως enrollment utterance).Η δεύτερη πρόταση αποτελεί την **πρόταση επαλήθευσης** (γνωστή στην αγγλική ορολογία και ως verification utterance). Ένα σύστημα αναγνώρισης ομιλητή, ανεξάρτητο κειμένου, δέχεται τις δυο αυτές προτάσεις και προσπαθεί να αποφανθεί εάν οι δύο αυτές προτάσεις προέρχονται απο τον ίδιο ομιλητή. Το αποτέλεσμα ενός τέτοιου δικτύου συνήθως πρόκειται για έναν αριθμό εμπιστοσύνης, στο εύρος [0,1], όπου 1 είναι η σίγουρη αποδοχή, ενώ 0 η σίγουρη απόρριψη.

Εξαγωγή Χαρακτηριστικών

Ένα φασματογράφημα (στην αγγλική ορολογία Spectrogram) είναι μια οπτική αναπαράσταση του φάσματος των συχνοτήτων ενός σήματος ως συνάρτηση του χρόνου. Τα φασματογραφήματα χρησιμοποιούνται εκτενώς στους τομείς της μουσικής, της γλωσσολογίας, σόναρ, ραντάρ, επεξεργασία ομιλίας, σεισμολογία αλλά και άλλα. Φασματογραφήματα ήχου μπορούν να χρησιμοποιηθούν ακόμη στον εντοπισμό προφορικών λέξεων φωνητικά και στην ανάλυση διαφόρων ομιλιών ζώων. Τα φασματογράμματα μπορούν να δημιουργηθούν από ένα χρονικό σήμα με έναν από τους δύο τρόπους: προσεγγιστικά ως τράπεζα φίλτρων που προκύπτει από μια σειρά ζωνοπερατών φίλτρων (αυτός συγκεκριμένα ήταν ο μόνος τρόπος πριν από την έλευση της σύγχρονης επεξεργασίας ψηφιακού σήματος), ή εναλλακτικά από το σήμα χρόνου χρησιμοποιώντας τον μετασχηματισμό Fourier. Αυτές οι δύο μέθοδοι σχηματίζουν πραγματικά δύο διαφορετικές χρονικές αναπαραστάσεις στο πεδίο της συχνότητας, αλλά είναι ισοδύναμες υπό ορισμένες προϋποθέσεις. Μετά την εφαρμογή μιας κλίμακας Mel στο φασματογράφημα, υπολογίζεται το λεγόμενο Mel Spectrogram.

0.3 Επαλήθευση Ομιλιτή με Χρήση Συναισθήματος

0.3.1 Διαίσθηση

Το κύριο σημείο της διπλωματικής, είναι να εντοπίσουμε τις εξαρτήσεις μεταξύ των διακριτικών χαρακτηριστικών των ομιλητών και συναισθήματος. Υποθέτοντας ότι η φωνή των ομιλητών εμπεριέχει πληροφορίες όπως η ηλικία, το φύλο και το συναίσθημα πέρα από τα γλωσσικά χαρακτηριστικά, είναι εύκολα αντιληπτό ότι η φωνητική ταυτότητα δεν μπορεί να

ορισθεί μονοσήμαντα. Αυτό δημιουργεί ένα ερώτημα σχετικά με το πώς όλα αυτά τα χαρακτηριστικά επιδρούν στη φωνή. Μια λογική υπόθεση θα ήταν ότι, εάν αυτά τα χαρακτηριστικά τροποποιούν τη φωνή με κάποιον πολύπλοκο τρόπο, θα ήταν πολύ δύσκολο για ένα σύστημα επαλήθευσης ομιλητή να ταξινομήσει σωστά έναν χρήστη. Σε αυτή τη διπλωματική, διερευνούμε την επίδραση των συναισθημάτων σε ένα τέτοιο σενάριο.

Για να μπορέσουμε να διερευνήσουμε την επίδραση του συναισθηματικού περιεχομένου στην ομιλία, πρέπει πρώτα να ορίσουμε τα βασικά μοντέλα μας, τόσο για την αναγνώριση συναισθήματος στην ομιλία όσο και για την επαλήθευση ομιλητή. Κατόπιν, δημιουργούμε μια σειρά πειραμάτων, προκειμένου να εξερευνήσουμε τα αποτελέσματα του συναισθηματικού λόγου.

0.3.2 Προσέγγιση

Σε αυτή τη διπλωματική, διερευνούμε την επίδραση του συναισθηματικού περιεχομένου στην ομιλία με τέτοιο τρόπο, ώστε να αντιμετωπιστούν τόσο κεντρικά ζητήματα δεν έχουν διευθετηθεί σε προηγούμενες έρευνες όσο και την ενίσχυση σε παραδοσιακά μοντέλα μηχανικής μάθησης.

Προσπαθούμε αρχικά, να αποδείξουμε την ύπαρξη του προβλήματος και στη συνέχεια να το ποσοτικοποιήσουμε. Έπειτα προτείνουμε κάποιες παραδοσιακές μεθόδους μηχανικής μάθησης για μεταφορά γνώσης προκειμένου να μειώσουμε ή ακόμη και να εξαλείψουμε την επίδραση αυτή. Ακριβέστερα, χρησιμοποιούμε ένα προεκπαιδευμένο συναισθηματικό μοντέλο και προσπαθούμε να μεταφέρουμε συναισθηματική πληροφορία σε ένα μοντέλο επαλήθευσης ομιλητή. Τέλος, δημιουργούμε ένα σύνολο πειραμάτων προσπαθώντας να δώσουμε απαντήσεις στα παρακάτω ερωτήματα:

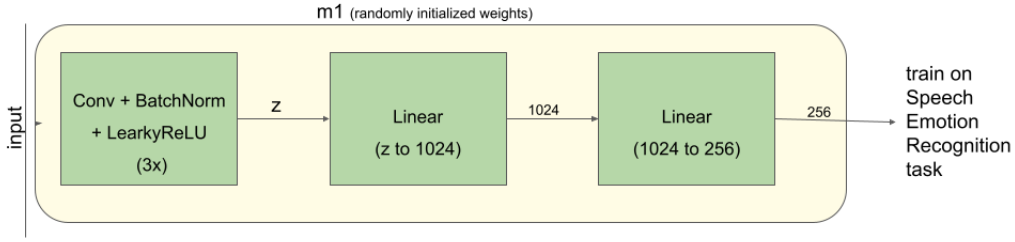
- Μπορεί η συναισθηματική γνώση να βελτιώσει ένα μοντέλο επαλήθευσης ομιλητή ;
- Ποιά είναι η επίδραση του συναισθηματικού περιεχομένου στο αντικείμενο της αναγνώρισης ομιλητή;
- Με ποιόν τρόπο επηρεάζει η ένταση του συναισθήματος τη διαδικασία ;
- Πως επιδρά κάθε συναίσθημα χωριστά στο αντικείμενο της επαλήθευσης ομιλητή;
- Θα μπορούσε μια πρόταση με συναίσθημα στην πρόταση εγγραφής να επιφέρει καλύτερα αποτελέσματα απο μία συναισθηματικά ουδέτερα πρόταση, δεδομένου ότι η πρόταση επιβεβαίωσης ήταν συναισθηματικά φορτισμένη;

0.3.3 Αρχιτεκτονικές Μοντέλων

Πρωτού τρέξουμε οποιοδήποτε πείραμα κατασκευάζουμε δύο μοντέλα. Το πρώτο μοντέλο $m1$ πρόκειται για το μοντέλο το οποίο εκπαιδεύεται στην αναγνώριση συναισθήματος στην ομιλία. Αντιμετωπίζουμε το μοντέλο αυτό σαν τον **συναισθηματικό εγκέφαλο** στο σύστημα μας. Το δεύτερο μοντέλο $m2$ πρόκειται για ένα σύστημα επαλήθευσης ομιλητή. Ο στόχος μας είναι να μεταφέρουμε γνώση απο το $m1$ στο $m2$ και να εξετάσουμε την απόδοση του σε διάφορα σενάρια.

Προς αυτή την κατεύθυνση λοιπόν, κατασκευάζουμε και αξιολογούμε διαφορετικές αρχιτεκτονικές, όπου όλες έχουν ως στόχο την αποτελεσματικότερη μεταφορά γνώσης απο το ένα πεδίο στο άλλο. Χωρίζουμε λοιπόν παρακάτω την ανάλυση μας σε δυο υποκατηγορίες, μία για το συναισθηματικό μοντέλο και μία για το μοντέλο επαλήθευσης ομιλητή. Για κάθε ένα προσπαθούμε να βρούμε την αρχιτεκτονική που θα δώσει τα καλύτερα δυνατά αποτελέσματα.

Μοντέλο Αναγνώρισης Συναισθήματος



Σχήμα 3: *Το μοντέλο m1: Εκπαιδευμένο στην αναγνώριση συναισθήματος στο πρόβλημα τεσσάρων συναισθηματων του IEMOCAP.*

Το μοντέλο που χρησιμοποιούμε για την αναγνώριση συναισθήματος πρόκειται για ένα Συνελικτικό Νευρωνικό Δίκτυο. Πιο συγκεκριμένα το δίκτυο μας, αποτελείται απο 3 συνεχόμενα συνελικτικά στρώματα, 32 καναλιών και πυρήνα μεγέθους 5. Μεταξύ αυτών μεσολαβούν κανονικοποίηση, συνάρτηση ενεργοποίησης LeakyReLU και δύο διαστάσεων max pooling, ανά batch Τα στρώματα αυτά ακολουθούνται απο δύο πλήρως συνδεδεμένα γραμμικά στρώματα προβάλλοντας την είσοδο σε 1024 και στη συνέχεια σε 256. Επιπλέον κάνουμε χρήση του Cross Entropy Loss και χρησιμοποιούμε για βελτιστοποίηση το SGD. Το μοντέλο αυτο εκπαιδεύεται για περίπου 50 εποχές στην αναγνώριση τεσσάρων συναισθημάτων στο IEMOCAP. Τα συναισθήματα αυτά είναι: *Ουδέτερο, Θυμός, Χαρά και Λύπη.*

Τα αποτελέσματα μας στο συγκεκριμένο πείραμα αξιολογούνται με βάση δύο μετρικές, την ακρίβεια (στην αγγλική ορολογία accuracy) και το f1-macro score.

Μοντέλο	Ακρίβεια %	f1-macro %
συναισθηματικός εγκέφαλος $m1$	56.7	55.94

Πίνακας 1: *Τα αποτελέσματα της εκπαίδευσης του μοντέλου* $m1$.

Μοντέλο Επαλήθευσης Ομιλητή

Στην προσπάθεια μας να αποκτήσουμε όσον το δυνατόν καλύτερη μεταφορά γνώσης κατασκευάζουμε 4 διαφορετικές αρχιτεκτονικές. Κάθε μία από αυτές αξιολογήθηκε σε κάθε συναισθηματικό πείραμα. Επιπλέον ότι χρησιμοποιήθηκαν διαφορετικοί ρυθμοί εκμάθησης, διασφαλίζοντας ότι θα βρεθεί το χαμηλότερο σημείο EER. Οι αρχιτεκτονικές ήταν οι εξής:

- Μοντέλο $t0$: Αυτό το μοντέλο θα χρησιμοποιηθεί ως μοντέλο χωρίς γνώση συναισθημάτων στα πειράματα μας. Αυτό σημαίνει ότι, δεν θα διαθέτει καμία συναισθηματική επεξήγηση κατα της διάρκεια της εκπαίδευσης του. Αποτελείται μόνο από το μοντέλο $m2$ αγνοώντας εντελώς την επιρροή του συναισθήματος στην ομιλία. Η αρχιτεκτονική αυτού του μοντέλου απεικονίζεται στο Σχήμα3.3.

- Μοντέλο $t1$: Αυτό το μοντέλο στοχεύει στη μεταφορά γνώσης από το $m1$ στο $m2$ εφαρμόζοντας πλήρη επανεκπαίδευση (στην αγγλική ορολογία finetuning). Πιο συγκεκριμένα, χρησιμοποιούμε τα εκπαιδευμένα βάρη του $m1$ ως αφετηρία για το $m2$. Στη συνέχεια, το $m2$ εκπαιδεύεται πλήρως στην επαλήθευση ομιλητή. Σε αυτήν τη διαδικασία, χρησιμοποιούνται διαφορετικά ποσοστά εκμάθησης, τα καλύτερα από τα οποία διατηρούνται ως το τελικό μοντέλο **με γνώση συναισθήματος** $t1$. Η αρχιτεκτονική αυτού του μοντέλου απεικονίζεται στο Σχήμα 3.4.

- Μοντέλο $t2$: Αυτό το μοντέλο στοχεύει στη μεταφορά γνώσης από το $m1$ στο $m2$ μέσω της επανεκπαίδευσης του $m2$ με έναν πολύ συγκεκριμένο τρόπο. Στόχος μας είναι να κρατήσουμε σταθερά τα σημεία όπου το αρχικό δίκτυο επικεντρωνόταν στο ηχητικό σήμα. Χρησιμοποιούμε τα προεκπαιδευμένα βάρη του $m1$ ως σημείο εκκίνησης για το $m2$. Εξαναγκάζουμε, στη συνέχεια, κάθε βάρος στα συνελικτικά στρώματα να μην εκπαιδευτεί περαιτέρω. Αυτό σημαίνει ότι δεν μπορεί να ενημερωθεί κανένα βάρος σε αυτά τα επίπεδα, κατά τη διάρκεια της διαδικασίας εκπαίδευσης του $m2$. Στη συνέχεια, το $m2$ εκπαιδεύεται στο αντικείμενο της επαλήθευσης ομιλητή. Τέλος, χρησιμοποιούνται διαφορετικά ποσοστά εκμάθησης, το καλύτερα από τα οποία διατηρείται ως το τελικό μοντέλο **με γνώση συναισθήματος** $t2$. Η αρχιτεκτονική αυτού του μοντέλου απεικονίζεται στο Σχήμα 3.5.

- Μοντέλο $t3$: Αυτό το μοντέλο στοχεύει στη μεταφορά γνώσεων από $m1$ σε $m2$ χρησιμοποιώντας μια αρχιτεκτονική σύντηξης. Πιο συγκεκριμένα, χρησιμοποιούμε προκαθορισμένα βάρη από το $m1$ ως σημείο εκκίνησης για ένα υποδίκτυο στο $m2$. Έτσι το νευρωνικό δίκτυο χωρίζεται σε δύο μέρη. Στον «συναισθηματικό εγκέφαλο», καθώς το $m1$ είναι πλέον μέρος του μεγαλύτερου δικτύου και στο υποδίκτυο διάκρισης ομιλητών. Κατά τη διάρκεια της εκπαιδευτικής διαδικασίας, το συναισθηματικό μέρος δεν επανεκπαιδεύεται. Η συμβολή του στη συνολική διαδικασία είναι η παροχή ισχυρών συναισθηματικών ενσωματώσεων (στην αγγλική ορολογία embeddings) στο δίκτυο. Αυτές οι ενσωματώσεις συνδυάζονται με τις αντίστοιχες του συστήματος επαλήθευσης ομιλητή στα τελικά στρώματα του δικτύου σύντηξης. Καθώς εκπαιδεύαμε το συγκεκριμένο μοντέλο στη διαδικασία επαλήθευσης ομιλητών, χρησιμοποιήθηκαν διαφορετικά ποσοστά εκμάθησης, εκ των οποίων τα καλύτερα, διατηρούνται ως το μοντέλο **με γνώση**

συναισθήματος $t3$. Η αρχιτεκτονική αυτού του μοντέλου απεικονίζεται στο Σχήμα 3.6.

Το μοντέλο επαλήθευσης ομιλητή έχει πολύ παρόμοια αρχιτεκτονική με το μοντέλο της αναγνώρισης συναισθημάτων. Αυτό γίνεται συνειδητά, προκειμένου να διευκολύνεται η διαδικασία μεταφοράς γνώσης. Εκπαιδεύουμε κάθε μοντέλο για 1000 εποχές με ενεργοποιημένη την πρόωρη διακοπή (στην αγγλική ορολογία early stopping). Η εκπαιδευτική διαδικασία επιτυγχάνεται αξιοποιώντας το σύνολο δεδομένων VoxCeleb 1. Τροφοδοτούμε το μοντέλο μας με ζεύγη πολλαπλών εκφωνητών και υπολογίζουμε αντίστοιχα το GE2E Loss.

0.3.4 Σύνολα Αξιολόγησης

Βασική προϋπόθεση για να μπορέσουμε να πραγματοποιήσουμε επιτυχώς τα πειράματα στο Κεφάλαιο 4, ήταν ο ορισμός των συνόλων αξιολόγησής μας. Αυτά ήταν με άλλα λόγια τα σετ δοκιμών, τα οποία χρησιμοποιήσαμε. Για καθένα από τα πειράματά μας, δημιουργήσαμε ένα διαφορετικό σύνολο δοκιμών και στη συνέχεια ελέγξαμε την απόδοση κάθε μοντέλου ξεχωριστά.

Όλα τα σύνολα προέρχονται από το σύνολο δεδομένων RAVDESS. Ένα σύνολο αποτελείται ουσιαστικά από έναν αριθμό προσεκτικά επιλεγμένων πλειάδων. Κάθε πλειάδα περιέχει μια πρόταση εγγραφής, μια πρόταση επαλήθευσης και μια ετικέτα. Η ετικέτα υποδεικνύει εάν οι δύο προτάσεις ανήκουν στο ίδιο άτομο ή όχι.

Για να ορίσουμε αυστηρά την πειραματική μας ρύθμιση, πρέπει πρώτα να εξηγήσουμε τη δομή του συνόλου δεδομένων RAVDESS. Πρώτα απ ΄όλα χρησιμοποιήσαμε μόνο τον ήχο και επίσης απορρίψαμε όλες τις προτάσεις που περιείχαν τραγούδι.

Στη συνέχεια κατασκευάσαμε τέσσερα σετ πλειάδων όπως περιγράφεται αναλυτικά στο 3.2.4. Κάθε πλειάδα απο αυτές χρησιμοποιήθηκε στο αντίστοιχο πείραμα. Στα πειράματα 1, 2 και 3 έγινε επιπρόσθετος διαχωρισμός των πλειάδων σε μικρότερα σύνολα ανάλογα με την ένταση του συναισθήματος. Η έντασεις ήταν: **ήπια και έντονη**. Στο πείραμα 3 εφαρμόστηκε επιπλέον διαχωρισμός προκειμένου τα σύνολα να διαχωριστούν ανα συναίσθημα. Τέλος στο πείραμα 4, ο διαχωρισμός έγινε ανα συναίσθημα και ανά την συναισθηματική γνώση ή οχι στην πρόταση εγγραφής. Η σκέψη πίσω απο τους συγκεκριμένους διαχωρισμούς των προτάσεων των ομιλητών επεξηγείται αναλυτικά, παρακάτω στο 0.4.

0.4 Πειράματα

Σε αυτό το σημείο θα περιγράψουμε εν συντομία τα πειράματά μας, και κάποια συμπεράσματα που εξάγουμε βάση αυτών, ενώ στη συνέχεια θα δούμε και τα αποτελέσματα της ποιοτικής μελέτης που διεξήγαμε. Σε κάθε πείραμα παρακάτω αξιολογούμε τα μοντέλα μας $t1, t2$ και $t3$ που *κατέχουν συναισθηματική γνώση*, με το μοντέλο $t0$ όπου δεν έχει γνώση συναισθήματος.

model	EER (%)	s^2	statistical significance
$t0$	19.7	0.74	0.04
$t1$	**16.35**	0.43	0.02
$t2$	20.66	0.79	0.03
$t3$	**18.28**	0.51	0.04

Πίνακας 2: *The results on the speaker verification task, on VoxCeleb's evaluation set.*

0.4.1 Βασική Αξιολόγηση στην Επαλήθευση Ομιλητή

Το αρχικό αντικείμενο της έρευνας μας, αποτέλεσε η μελέτη της απόδοσης των μοντέλων μας σε ένα κλασικό πείραμα επαλήθευσης ομιλητή. Εκεί αξιολογήθηκαν τα μοντέλα μας στο σύνολο πλειάδων που αντιστοιχεί στην αξιολόγηση πάνω στο VoxCeleb, όπου και εκπαιδεύτηκαν.

Σε αυτό μας το πείραμα παρατηρήσαμε μία σχετική βελτίωση της τάξης του 7% για το μοντέλο μας $t3$ και μία του 17% για το μοντέλο $t1$ σε σχέση με το $t0$. Αντίθετα, το μοντέλο $t2$ φαίνεται να πήγε χειρότερα.

Παρατηρούμε λοιπόν ότι η μεταφορά συναισθηματικής γνώσης στο αντικείμενο της επαλήθευσης ομιλητή βελτιώνει εμφανώς τα αποτελέσματα μας στο σύνολο αξιολόγησης του αντικειμένου μας.

0.4.2 Η Επίδραση του Συναισθηματικού Περιεχομένου στην Επαλήθευση Ομιλητή

Σκοπός αυτού του πειράματος υπήρξε, ο εντοπισμός συσχετήσεων μεταξύ συναισθηματικού περιεχομένου στις προτάσεις και της απόδοσης των μοντέλων μας. Χωρίζουμε το πείραμα σε 2 μέρη. Πρώτα εξετάζουμε πώς μία φορτισμένη πρόταση επαλήθευσης μπορεί να επηρεάσει ένα σύστημα επαλήθευσης ομιλητή δεδομένου ότι ο χρήστης είχε δηλωθεί με ουδέτερο συναίσθημα. Κατόπιν, εισάγουμε και στην πρόταση εγγραφής συναίσθηματικη φόρτιση και εξετάζουμε τη συμπεριφορά των μοντέλων μας.

Πιο συγκεκριμένα, για την αξιολόγηση του πειράματος, κατασκευάσαμε πλειάδες. Κάθε πλειάδα περιείχε μία συναισθηματικά **Ουδέτερη** πρόταση ως εγγραφή και μία **συναισθηματικά φορτισμένη** ως πρόταση επαλήθευσης. Επιπρόσθετα, οι προτάσεις διαχωρίστηκαν με βάση την ένταση των συναισθημάτων.

Exp	model	VoxCeleb eval.	RAVDESS weak emotion	RAVDESS strong emotion
1.1	$t0$	19.7	**16.37**	30.65
1.2	$t1$	**16.35**	16.74	**27.23**
1.3	$t2$	20.66	17.51	**30.51**
1.4	$t3$	**18.28**	19.49	**27.53**

Πίνακας 3: *The effect of emotion on speaker verification task, in the case of a neutral enrollment is followed by an emotional verification utterance.*

Όπως παρατηρούμε στον Πίνακα 4.2 τα μοντέλα μας συμπεριφέρονται αρκετά καλά στο ήπιο συναίσθημα. Αντίθετα όμως, παρατηρούμε μια τεράστια εκτόξευση του EER το οποίο σχεδόν διπλασιάζεται για το έντονο συναίσθημα. Προφανώς λοιπόν το έντονο συναίσθημα

επηρεάζει άμεσα τα σύστηματα μας και ιδιαίτερα το $t0$.

Παρατηρώντας την συμπεριφορά των υπόλοιπων μοντέλων, εύκολα αναγνωρίζουμε την υπεροχή τους στο έντονο συναίσθημα ενταντίον του $t0$. Όλα τα μοντέλα μας είναι ικανά να διαχειριστούν καλύτερα τις έντονες συναισθηματικά φορτισμένες προτάσεις στο στάδιο της επαλήθευσης. Αντιλαμβανόμαστε λοιπόν πως η εισαγωγή κάποιας γνώση συναισθήματος κατα την εκπαίδευση του μοντέλου μπορεί να βοηθήσει στην καλύτερη αντιμετώπιση του σε ένα πραγματικό σενάριο.

Προκειμένου να μεγενθύνουμε την επίδραση του συναισθήματος, στο επόμενο πείραμα εισάγαγαμε συναίσθημα και στην πρόταση εγγραφής. Με αυτόν τον τρόπο παρουσιάζουμε ένα πιο έντονο σενάριο όπου και οι δύο προτάσεις να είναι συναισθηματικά φορτισμένες.

Exp	model	VoxCeleb eval.	RAVDESS weak emotion	RAVDESS strong emotion
2.1	$t0$	19.7	21.88	32.64
2.2	$t1$	**16.35**	**21.38**	**31.37**
2.3	$t2$	20.66	**20.78**	**31.13**
2.4	$t3$	18.28	23.29	**30.78**

Πίνακας 4: *The effect of emotion on SV task, when different emotions occur both during enrollment and verification phase.*

Όπως παρατηρούμε στην Πίνακα 4.3, το φαινόμενο πλέον επιδεινώνεται. Εύκολα αναγνωρίζει κανείς ότι σε αυτή την περίπτωση τα μοντέλα χειροτερεύουν την απόδοση τους ακόμη και στο ήπιο συναίσθημα. Μια αύξηση του EER της τάξης του 25% σε σχέση με το προηγούμενο πείραμα, πρακτικά σημαίνει ότι τα μοντέλα μας δεν έχουν τη δυνατότητα να αναγνωρίσουν δύο προτάσεις οι οποίες είναι μακριά συναισθηματικά. Παράλληλα βλέπουμε μια σταθερά κακή απόδοση στο έντονο συναίσθημα. Σε αυτό, τα μοντέλα μας φαίνεται πως συμπεριφέρονται παρόμοια με το προηγούμενο πείραμα.

Αξίζει να παρατηρήσει κανείς την σταθερή βελτίωση των συναισθηματικών μοντέλων. Πιο συγκεκριμένα, γίνεται προφανές ότι τα μοντέλα μας ξεπερνούν το μοντέλο χωρίς γνώση $t0$ σε απόδοση, ιδιαίτερα στα έντονα συναισθήματα. Το μοντέλο $t1$, ειδικά, για μία ακόμη φορά ξεπερνά το $t0$ και στους δύο βαθμούς συναισθήματος.

Προφανώς, το συναισθηματικό περιεχόμενο εμπλέκεται με την ικανότητα ενός συστήματος να διακρίνει ομιλητές. Κάτι τέτοιο εγείρει ερωτήματα σχετικά με τα χαρακτηριστικά που επιλέγει ένα τέτοιο νευρωνικό δίκτυο και την ευαισθησία τους στις ανθρώπινες εκφράσεις και συναισθήματα.

0.4.3 Η Επίδραση κάθε Συναισθήματος στην Επαλήθευση Ομιλητή

Μετά τη διεξαγωγή του προηγούμενου πειράματος, εγείρονται ερωτήματα σχετικά με το ποιά είναι εκείνα τα συναισθήματα που προκαλούν αυτή την μείωση στην απόδοση των μοντέλων μας. Στο πείραμα αυτό, θα προσπαθήσουμε να αναγνωρίσουμε εαν υπάρχουν τέτοια συναισθήματα και να σκιαγραφήσουμε την επιδρασή τους.

Παρατηρώντας τον Πίνακα 5, βλέπουμε ότι το συναίσθημα της *Αηδίας* είναι εκείνο που συμπεριφέρεται χειρότερα απο όλα στα ήπια συναισθήματα. Αντίθετα τα συναισθήματα του *Φόβου* και του *Εκνευρισμού* φαίνεται να είναι τα αντίστοιχα χειρότερα για τα έντονα. Παράλληλα, το συναίσθημα *Ηρεμία* φαίνεται να συμπεριφέρεται καλύτερα απο όλα.

Exp	Συναίσθημα	RAVDESS Ήπιο	RAVDESS Έντονο
3.1	Ηρεμία	9.38	15.62
3.2	Χαρά	15.1	32.29
3.3	Στεναχώρια	12.5	31.77
3.4	Θυμός	17.71	39.58
3.5	Φόβος	18.23	38.54
3.6	Αηδία	24.48	20.31
3.7	Έκπληξη	17.71	27.6

Πίνακας 5: *Η επίδραση των διαφορετικών συναισθημάτων στο t0.*

Στους πίνακες 4.5, 4.6 και 4.7 εμφανίζονται οι αντίστοιχες συμπεριφορές των μοντέλων με γνώση συναισθήματος. Δίπλα από κάθε στήλη εμφανίζεται και η αντίστοιχη βελτίωση, συγκριτικά με το μοντέλο $t0$.

Παρατηρούμε ότι το συναίσθημα της *Ηρεμίας* και αυτό της *Αηδίας* συμπεριφέρονται χειρότερα από όλα. Πιο αναλυτικά σε αυτά τα δύο φαίνεται να έχουμε τεράστια χειροτέρευση του EER σε όλα μας τα μοντέλα και σε όλους τους βαθμούς συναισθήματος. Αυτό σχετίζεται φυσικά με την απουσία συναισθημάτων κατα την εκπαίδευση των μοντέλων. Αναγνωρίζουμε έυκολα, ότι σχεδόν όλα τα συναισθήματα που ήταν παρόντα κατά τη διάρκεια της εκπαίδευσης του μοντέλου $t1$ φαίνονται να πήγαν καλύτερα σε σχέση με το $t0$. Δυστυχώς το ίδιο δεν συμβαίνει για τα μοντέλα $t2$ και $t3$. Το πρώτο φαίνεται να πηγαίνει πολύ χειρότερα στα ήπια συναισθήματα ακόμη και στα προεκπαιδευμένα. Αντίθετα το δεύτερο έχει πολύ έντονες αποκλίσεις απο το μοντέλο $t0$ με είτε πολύ μεγάλες βελτιώσεις σε κάποια είτε πολύ μεγάλες επιδεινώσεις, καθιστώντας το ασταθές.

Συνολικά λοιπόν συμπεραίνουμε ότι τα συναισθήματα του *Φόβου* και του *Εκνευρισμού* φαίνεται να δυσκολεύουν όλα τα μοντέλα περισσότερο απο τα άλλα συναισθήματα. Παράλληλα το συναίσθημα της *Αηδίας* φαίνεται να έχει μεγάλη επίδραση ξεχωρίζοντας ανάμεσα στα ήπια συναισθήματα. Τέλος η έλλειψη συναισθηματικής γνώσης φαίνεται να βοηθά το μοντέλο $t0$ στο συναίσθημα της *Ηρεμίας*.

0.4.4 Η Επίδραση του Κοινού Συναισθήματος σε πρόταση Εγγραφής και Επαλήθευσης

Σε αυτό το πείραμα, καλούμαστε να διευκρινήσουμε πώς επιδρά η επιβολή κοινού συναισθήματος στην πρόταση εγγραφής και στην πρόταση επαλήθευσης. Πιο συγκεκριμένα προσπαθούμε να διαπιστώσουμε, εάν ένα τέτοιο σενάριο επιδρά θετικά τόσο στη συνολική απόδοση των μοντέλων όσο και στην ανα συναίσθημα απόδοση.

Προκειμένου να διεξάγουμε το πείραμα χωρίζουμε το διαθέσιμο σύνολο σε δύο σύνολα και ανα συναίσθημα. Το πρώτο σύνολο θεωρείται αυτό που δεν έχει γνώση συναισθήματος κατα τη διάρκεια της απόφασης των μοντέλων ενώ το δεύτερο έχει. Η γνώση αυτή δεν σχετίζεται σε καμία περίπτωση με την εκπαίδευση των μοντέλων παρέχοντας συναισθηματική γνώση.

Όπως παρατηρούμε στον Πίνακα 6 το μοντέλο $t0$, χωρίς καμία γνώση συναισθήματος κατά την εκπαίδευση τυγχάνει μίας βελτίωση της τάξης του 21%. Στην ανα συναίσθημα ανάλυση φαίνεται καθαρά αύξηση σε όλα τα συναισθήματα, εκτός αυτού της λύπης. Πιο συγκεκριμένα όλα τα υπόλοιπα συναισθήματα αντιμετωπίζουν μια αύξηση απο 12% έως και 38%. Ένα τέτοιο

Exp	Συναισθήματα	EER (%) σε Άγνοια	EER (%) σε Γνώση	Σχετική Βελτίωση (%)
4.1	calm	9.9	7.66	22.63
4.2	happy	23.44	17.41	25.73
4.3	sad	20.83	22.92	−10.03
4.4	angry	33.33	20.46	38.61
4.5	fearful	28.39	24.93	12.19
4.6	disgust	19.01	14.58	23.3
4.7	surprised	17.71	11.38	35.74
4.8	average	21.80	17.05	21.17

Πίνακας 6: *Τα αποτελέσματα της Συναισθηματικής Άγνοιας και της Συναισθηματικής Γνώσης ανά συναίσθημα για το μοντέλο* $t0$

αποτέλεσμα είναι πολύ σημαντικό καθώς φαίνεται έντονα η σύνδεση συναισθηματικής γνώσης στη βελτίωση της συνολικής απόδοσης. Αυτό πρακτικά σημαίνει ότι, σε ένα πραγματικό σύστημα, εάν με κάποιο τρόπο αναγνωρίζαμε το συναίσθημα στην πρόταση επιβεβαίωσης, θα μπορούσαμε να επιλέγαμε την κατάλληλη συναισθηματικά πρόταση επιβεβαίωσης. Κατά συνέπεια η βελτίωση στον τομέα της αναγνώρισης συναισθήματος θα μπορούσε απευθείας να συμβάλει στη βελτίωση και της επιβεβαίωσης ομιλητή σε πραγματικές εφαρμογές.

Αντίστοιχα, στους πίνακες 4.9, 4.10 και 4.11 παρατηρούμε ακριβώς την ίδια συμπεριφορά. Φαίνεται καθαρά ότι μία τέτοια επιλογή βελτιώνει τις αποκλίσεις που υπήρχαν στα προηγούμενα πειράματα. Πιο συγκεκριμένα το συναίσθημα της *Ηρεμίας* ενώ στο προηγούμενο πείραμα φάνηκε να επηρεάζει έντονα, εδώ πετυχαίνουμε μια βελτίωση της τάξης του $40 - 52\%$ στα μοντέλα $t1$, $t2$ και $t3$. Παράλληλα το μοντέλο $t1$ φαίνεται να πηγαίνει καλύτερα και στο συναίσθημα της λύπης.

Συμπερασματικά λοιπόν, σε ένα πραγματικό σύστημα επαλήθευσης ομιλητή, η εισαγωγή γνώσης από το πεδίο της αναγνώρισης συναισθήματος θα μπορούσε να αποτελέσει μεγάλη αρωγή στην συνολική απόδοση του συστήματος.

Τέλος στον Πίνακα 4.12 παρουσιάζονται οι σχετικές αποδόσεις των μοντέλων $t1, t2$ και $t3$ συγκριτικά με το $t0$. Παρατηρούμε ότι για ακόμη μία φορά το μοντέλο μας $t1$ συμπεριφέρεται καλύτερα από όλα.

0.4.5 Συνεισφορά

Μέσω της διπλωματικής αυτής εργασίας καταλήξαμε σε μερικά πολύ σημαντικά αποτελέσματα. Τα αποτελέσματα αυτά συνοψίζονται συνοπτικά παρακάτω:

- Δείξαμε ότι αξιοποιώντας την μεταφορά γνώσης απο ένα προεκπαιδευμένο μοντέλο αναγνώρισης συναισθημάτων στο αντικείμενο της επαλήθευσης ομιλητή μπορεί να βελτιωθεί σημαντικά η απόδοση, σε σχέση με ένα μοντέλο εκπαιδευμένο απευθείας στην επαλήθευση ομιλητή. Πιο συγκεκριμένα πετύχαμε μια βελτίωση της τάξης του 7% για την αρχιτεκτονική $t3$ και 17% για αυτήν του $t1$.

- Δείξαμε ότι το συναισθηματικό περιεχόμενο στο πεδίο της επαλήθευσης ομιλητή μπορεί να επηρεάσει σε τεράστιο βαθμό την απόδοση του συστήματος, εξαρτώμενο πλήρως απο το βαθμό της συναισθηματικής φόρτισης. Κάτι τέτοιο έχει ως αποτέλεσμα, μο-

ντέλα ευάλωτα στη συναισθηματική διάθεση του ομιλητή. Παράλληλα, τα διαφορετικά συναισθήματα στην πρόταση εγγραφής και επαλήθευσης μπορούν να μεγενθύνουν το φαινόμενο, καταλήγοντας σε πολύ φτωχά αποτελέσματα. Αυτό υποδεικνύει ότι οι αρχιτεκτονικές εκπαιδευμένες απευθείας στο πεδίο της επαλήθευσης ομιλητή, δεν έχουν διαίσθηση για το τους τρόπους με τους οποίους το συναίσθημα διαμορφώνει το λόγο.

- Δείξαμε ότι ο *Θυμός* και ο *Φόβος* είναι τα συναισθήματα που επηρεάζουν περισσότερο τη διαδικασία επαλήθευσης ομιλητή. Μετά απο προσεκτική εξέταση των συναισθημάτων που παρέχονται απο το RAVDESS, τα πειράματα μας υποδεικνύουν ότι αυτά τα δύο έχουν το μεγαλύτερο αντίκτυπο τη διαδικασία. Τόσο ο *Θυμός* όσο και ο *Φόβος* ανεβάζουν δραστικά το EER επηρεάζοντας έντονα τη διαδικασία διαχώρισης ομιλητών. Συγκεκριμένα τα μοντέλα χωρίς γνώση συναισθήματος έχουν μια πολύ χαμηλή απόδοση κοντά στο 40% EER.

- Κάναμε επίδειξη του πώς μπορούμε μέσω κλασικών μεθόδων μεταφοράς μάθησης απο το πεδίο της αναγνώρισης συναισθημάτων σε ομιλία στο πεδίο της επαλήθευσης ομιλητή να εκπαιδεύσουμε αποτελεσματικά μοντέλα με συναισθηματική επίγνωση. Παρατηρήσαμε ότι τα μοντέλα αυτά είναι ικανά να μειώσουν την επίδραση του συναισθηματικού λόγου και να βελτιώσουν τα συνολικά μας αποτελέσματα. Πιο συγκεκριμένα η αρχιτεκτονική μας $t1$, υπερτερεί ενός κλασικού μοντέλου χωρίς επίγνωση συναισθήματος όπως είναι το $t0$, τόσο σε ήπια όσο και σε έντονα συναισθηματικά επίπεδα.

- Δείξαμε ότι η γνώση του συναισθήματος στην πρόταση επαλήθευσης, μπορεί να βελτιώσει την απόδοση ενός συστήματος κοντά στο $17 - 24\%$, ακόμη και σε μοντέλα χωρίς κάποια επίγνωση συναισθήματος. Αυτό είναι πολύ σημαντικό δεδομένου του ότι βελτίωση στο πεδίο της αναγνώρισης συναισθημάτων μπορεί να συμβάλει στην βελτίωση στο πεδίο της επαλήθευσης ομιλητή. Μία τέτοια παρατήρηση θα μπορούσε να έχει εφαρμογή σε πραγματικά συστήματα καθώς θα μπορούσαμε πρακτικά να παρέχουμε συγκεκριμένες προτάσεις επαλήθευσης, ανάλογα με την έξοδο ενός συστήματος κατηγοριοποίησης συναισθημάτων.

0.5 Συμπεράσματα και Μελλοντικές Κατευθύνσεις

0.5.1 Συμπεράσματα

Σε αυτή τη διπλωματική εργασία, μελετήσαμε την επίδραση του συναισθήματος στην επαλήθευση των ομιλητών. Πειραματιστήκαμε με διαφορετικές τεχνικές, προκειμένου να κατανοήσουμε καλύτερα πώς τα διαφορετικά συναισθήματα εμπλέκονται με τα μοντέλα βαθιάς μηχανικής μάθησης.

Πρώτα απ΄όλα, εξετάσαμε πώς το συναισθηματικό περιεχόμενο επηρεάζει το αντικείμενο επαλήθευσης ομιλητή και δείξαμε ότι η συναισθηματική ομιλία μπορεί να υποβαθμίσει σημαντικά την απόδοση ενός τέτοιου συστήματος, ανάλογα με τη συναισθηματική ένταση. Διαπιστώσαμε ότι η ισχυρή συναισθηματική φόρτιση συνολικά μπορεί να κάνει το μοντέλο επαλήθευσης ομιλητή να έχει πολύ κακή απόδοση. Εξετάσαμε πώς το ουδέτερο συναίσθημα επηρεάζει τη διαδικασία και πώς το EER μεταβάλλεται εάν προσθέσουμε συναισθηματικό περιεχόμενο και

στη δήλωση εγγραφής, εκτός από αυτήν της επαλήθευσης. Τα αποτελέσματα δείχνουν ότι διαφορετικά ζεύγη συναισθημάτων σε αυτές τις δύο εκφωνήσεις έχουν ως αποτέλεσμα τη χειρότερη δυνατή απόδοση, καθιστώντας τα βασικά μοντέλα μας, πρακτικά ανίκανα να αναγνωρίσουν σωστά ομιλητές.

Στη συνέχεια, διερευνήσαμε πώς κάθε συναίσθημα επηρεάζει ξεχωριστά τη διαδικασία και εντοπίσαμε αυτά που εμπλέκονται περισσότερο. Διαπιστώσαμε ότι ο θυμός και ο φόβος είναι τα συναισθήματα που είναι πιο πιθανό να κάνουν τα μοντέλα μας πιο επιρρεπή σε λανθασμένες απορρίψεις χρηστών ή λανθασμένες αποδοχές. Πιο συγκεκριμένα, παρατηρήσαμε ότι η απόλυτη τιμή του EER έφτασε τα 40% όταν αυτά τα συναισθήματα ήταν παρόντα. Ο θυμός είναι ένα συναίσθημα που έχει μια απροσδόκητη συμπεριφορά σε πολλές περιπτώσεις.

Επιπλέον, ελέγξαμε εάν ένα σύστημα επαλήθευσης ομιλητή μπορεί να βελτιωθεί εισάγοντας μια πρόταση ίδιου συναισθήματος για πρόταση εγγραφής στ φάση αξιολόγησης. Η επιλογή μας αυτή, έδειξε μεγάλη βελτίωση στην απόδοση του συστήματος, ακόμα και όταν τα μοντέλα δεν είχαν συναισθηματική πληροφορία κατά τη διάρκεια της εκπαίδευσης τους. Καταγράφουμε μια σχετική αύξηση περίπου 20% για όλα τα μοντέλα μας και για σχεδόν όλα τα συναισθήματα. Τα αποτελέσματά μας επισημαίνουν την ανάγκη να ληφθεί υπόψη η επίδραση του συναισθήματος στο πεδίο επαλήθευσης ομιλητή και της σχεδίασης πιο περίπλοκων αρχιτεκτονικών, όπου ένας ταξινομητής συναισθημάτων θα μπορούσε να αλληλεπιδράσει με τη δήλωση εγγραφής και επαλήθευσης. Με αυτόν τον τρόπο, η διαδικασία θα μπορούσε να γίνει λιγότερο ευάλωτη στο συναισθηματικό λόγο.

Τέλος, καταλήξαμε σε τρεις αρχιτεκτονικές που στόχευαν στη μείωση της συναισθηματικής επίδρασης στην επαλήθευση ομιλητών, καθεμία με διαφορετικό τρόπο. Δείξαμε ότι κάθε μία από αυτές υπερέχει του μοντέλου μας χωρίς γνώση συναισθημάτων σε κάποιες συνθήκες. Το πιο σημαντικό είναι ότι η αρχιτεκτονική μας $t1$, ξεπέρασε σημαντικά το βασικό μας μοντέλο χωρίς γνώση συναισθημάτων σε όλα τα σενάρια που εξετάσαμε. Επομένως, τα πειράματά μας δείχνουν ότι εφαρμόζοντας παραδοσιακές μεθόδους μεταφοράς γνώσης, μπορούμε να μεταφέρουμε αποτελεσματικά τη συναισθηματική γνώση στο πεδίο της επαλήθευσης ομιλητή, βελτιώνοντας την συνολική απόδοση των μοντέλων, ακόμα και όταν υπάρχει συναισθηματική ομιλία.

Όπως καταλαβαίνει κανείς εύκολα, το συναίσθημα έχει καίριο ρόλο και εμπλέκεται σε μεγάλο βαθμό με ένα σύστημα επαλήθευσης ομιλητή. Σε αυτή τη διπλωματική εργασία, εντοπίσαμε μερικές από αυτές τις σχέσεις και προσπαθήσαμε να τις ξεπεράσουμε, εφαρμόζοντας τεχνικές μεταφοράς συναισθηματικής γνώσης.

0.5.2 Μελλοντικές Κατευθύνσεις

Αυτή η διπλωματική εργασία θα μπορούσε να έχει πολλές ενδιαφέρουσες μελλοντικές επεκτάσεις. Σε αυτήν την υποενότητα προσπαθούμε να παρουσιάσουμε ορισμένες απο αυτές.

Πρώτα απ ΄όλα, μια κατεύθυνση θα ήταν να μελετήσουμε την επίδραση της γνώσης επαλήθευσης των ομιλητών στο πεδίο της αναγνώρισης συναισθημάτων ομιλίας και τον τρόπο με τον οποίο τα χαρακτηριστικά διάκρισης των ομιλητών σχετίζονται με τις συναισθηματικές εκφράσεις τους. Πιο συγκεκριμένα, καθώς αυτά τα δύο πεδία συσχετίζονται, θα μπορούσαμε να προσπαθήσουμε να εισάγουμε συγκεκριμένη γνώση προσαρμοσμένη ανα ομιλητή, στη διαδικα-

σία αναγνώρισης συναισθημάτων. Καθώς τα συναισθήματα δεν είναι μονοσήμαντα ορισμένα, θα ήταν ενδιαφέρον να μελετήσουμε εάν τα ειδικά αυτά χαρακτηριστικά ανα ομιλητή μπορούν να βελτιώσουν τη συνολική απόδοση ενός τέτοιου συστήματος.

Μια άλλη ενδιαφέρουσα πτυχή για μελέτη, θα ήταν η δημιουργία μιας προσαρμοσμένης συνάρτησης απώλειας (στην αγγλική ορολογία loss function). Γνωρίζουμε αυτήν τη στιγμή, ότι η εκπαίδευση ενός μοντέλου με συνηθισμένες loss function στην επαλήθευσης ομιλητή, όπως η GE2E για παράδειγμα, δεν λαμβάνουν υπόψη τις διάφορους πιθανούς παράγοντες όπως το συναίσθημα. Αυτό αποτελεί ένα μεγάλο ελάττωμα γιατί ουσιαστικά αγνοούμε εντελώς τη συναισθηματική πληροφορία. Ως αποτέλεσμα, εστιάζουμε τυφλά στη διάκριση των ομιλητών χωρίς να προσδιορίσουμε την επίδραση της συναισθηματικής φόρτισης και το πώς εκείνη διαμορφώνει την ομιλία. Όπως φαίνεται στα πρόσφατα έργα [15], [16], τα loss functions ειδικά προσαρμοσμένα ανα αντικείμενο, μπορούν να βελτιώσουν σημαντικά τη συνολική απόδοση. Έχουν γίνει προσπάθειες ακόμη και στο πεδίο της αναγνώρισης συναισθήματος [17], προκειμένου να διαχωριστούν οι βασικοί παράγοντες που επηρεάζουν στην παρουσία συναισθηματικού λόγου. Θα μπορούσαμε να εξετάσουμε πώς να ενσωματώσουμε συναισθηματική πληροφορία σε μια προσαρμοσμένη συνάρτηση απώλειας και να ελέγξουμε εάν η απόδοση βελτιώνεται τόσο στην εργασία επαλήθευσης ομιλητών αυτή καθ΄ αυτή, όσο και στην επαλήθευση ομιλητών όταν το συναίσθημα είναι έντονα παρόν.

Κεφάλαιο 1

Introduction

1.1 Motivation

Despite recent advances in the field of speaker verification, producing single, compact representations for speaker segments that can be used efficiently under noisy and unconstrained conditions is still a significant challenge. Such systems, in everyday usage, are likely to be prone to the variety of speakers expressions. Emotion as a natural and often involuntary encoder of voice, has the mechanisms responsible for vocal modulation. Despite its complexity and the fact that emotion is dominating common speech, its effect is usually considered negligible.

Given the fact that most speaker verification systems do not take emotional content into consideration, some questions arise about the vulnerability of such systems. It would be interesting to enlighten the effect of emotion on speaker discrimination and whether some emotions cause worse performance than other. If that is the case, then it would be a logical continuation to try to eliminate the emotional effect. A system like that, would be much more robust in real life scenarios.

Emotion Driven Speaker Verification is a attempt to exactly explore the emotional content on a speaker verification system. We make an effort to clarify the connections between Speaker Verification (SV) and Speech Emotion Recognition (SER). The problem is highly challenging due to two properties. Firstly, speaker characteristics are not unambiguous settled. Secondly, emotion recognition is a difficult task itself, as emotions differs from person to person and has no formal definition.

1.2 Approach and Contribution

In this Diploma Thesis we investigate how do different emotions affect a text-independent speaker verification system. More specifically we conduct multiple experiments in order to understand how does emotional content affect speaker verification, how does each emotion contribute to this process, and how could we overcome this, utilizing emotional knowledge in our favor.

First of all, we train a speech emotion recognition (SER) model on the IEMOCAP dataset. After that, we create four different architectures. Each one tries to solve differently, the problem of efficiently transferring emotional knowledge from the SER model. We

consider our first architecture as our baseline and which contains no emotional information during training. Our second architecture aims at discovering emotional knowledge, after a careful fine tuning from the SER task. Our third model tries to transfer knowledge without fully fine tuning, in order to keep track of the points where the original SER model focused on the input signal. Our fourth architecture consists of a fusion model and is further separated to two parts. Its first part, a pretrained emotion classifier aims at providing emotional information to the second part, which is trained strictly on the speaker verification task. We should note that all our models utilize the VoxCeleb dataset for efficiently training on the speaker verification task.

The next step is to we construct a set of emotional experiments. These experiments consist of a specific and careful setup as a mean to first qualify and then quantify the emotional effect on speech. We utilize the RAVDESS dataset, due to the multiple speakers and intensities that it provides. From that, we derive four different evaluation sets, each one looking at emotional effect from different perspective.

Afterwards, we examine our models' performance on these emotionally injected evaluation sets and assess their robustness. For each experiment, we test each model individually and compare each one with the emotion unaware model's results.

Our findings suggest that emotion has a crucial role on text-independent speaker verification. First of all, we capture that strong emotion can lead to huge degrades in performance, regardless of the emotion of the enrollment utterance. We notice that different emotions at enrollment and verification phase can magnify the effect and reach as low as 30% on Equal Error Rate. After a specific check on each emotion's individual contribution to the overall drop of performance we observe that *anger* and *fear* tend to perform worse. In more detail their error in an emotion-unaware model almost reaches that of a random guess with close to 40% equal error rate. Last but not least, we examined whether same-emotion pairs on enrollment and verification could help our model to better discriminate speakers. Our results indicate that same-emotion pairs improve drastically the systems' performance, regardless of emotional pretraining. Finally, we demonstrate that by applying traditional transfer learning techniques, it is possible to create an emotion-aware model that outperforms a classical speaker verification system.

1.3 Thesis Structure

In Chapter 2 we provide the theoretical background of our work. Specifically, we first introduce the basic concepts of machine learning and then focus on the deep learning techniques, that are most relevant to our work. We also explain the basic concepts of speech emotion recognition, speaker verification as well as the metrics used to evaluate our results.

In Chapter 3, we formulate the problem that we try to solve. We explain our approach and the questions that drove us to conduct the experiments. We first provide a section of related work and then proceed to the problem definition. In the rest of the chapter, we analyze our speech emotion recognition model, the different speaker verification models' architectures and the evaluation sets used for our emotional experiments.

In Chapter 4 we present our experiments and compare the results with our baseline methods.

In Chapter 5 we conclude our thesis, providing a summary of our work, and some future research ideas.

Κεφάλαιο 2

Theoretical Background

In this chapter, we provide the basic knowledge that is prerequisite for the reader to be able to understand the contribution of this thesis. We start by presenting the fundamentals of machine learning, the most famous deep learning architectures and some basic metrics for evaluation. After we analyze how traditional methods were used for speaker verification and speech emotion recognition and how these were replaced by deep learning techniques. Last but not least, we include a detailed description of all datasets used for this work.

2.1 A brief history of Machine Learning

It was in 1940s when the first manually operated computer system, ENIAC, was invented. At that time the word "computer" was being used as a name for a human with intensive numerical computation capabilities, therefore, ENIAC was called a numerical computing machine. From the beginning, the idea was to build a machine able to emulate human thinking and learning. In the 1950s, there is the first computer game program claiming to be able to beat the checkers world champion. This program helped checkers players a lot in improving their skills! Around the same time, Frank Rosenblatt invented the Perceptron, which at that time, it was a real breakthrough. Then we see several years of stagnation of the neural network field due to its difficulties in solving certain problems.

Thanks to statistics, machine learning became very famous in 1990s. The intersection of computer science and statistics gave birth to probabilistic approaches in AI. This shifted the field further toward data-driven approaches. Having large-scale data available, scientists started to build intelligent systems that were able to analyze and learn from large amounts of data. As a highlight, IBM's Deep Blue system beat the world champion of chess, the grand master Garry Kasparov.

2.2 Introduction to Machine Learning

Machine Learning (ML) is a subfield of AI. It enables computers to learn from data and even improve themselves without being explicitly programmed. The basic premise of ML is to build algorithms that can receive input data and use statistical analysis to improve themselves on predicting an output. It differs in this regard from other computational approaches within Computer Science, where algorithms are given explicit instructions on

how to solve problems and everything has to be accounted for by the programmer. In ML, learning from data results in the algorithms learning by themselves what to account for and how to deal with every hurdle. In recent years, AI has experienced a resurgence due to a subfield of ML, Deep Learning (DL). DL utilizes copious amounts of data and models with the ability to memorize a plethora of rules and high-level concepts, all encoded in their parameters.

Supervised Learning

Supervised learning (SL) is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.

So, for example, given a set of N training examples a supervised learning algorithm would seek to learn the function that maps the features of the input data to the labels corresponding to them. In math notation, we denote the input features as x^i and the labels as y^i. A data set D contains many data points, so $D = \{(x^i, y^i); i = 1, ..., n\}$, where n stands for the number of training data. We denote X as the input space and Y as the output space. Our goal is to learn a function $f : X \mapsto Y$ utilizing the dataset D in order $f(x)$ to correctly determine the label y of x.

Supervised algorithms are split in two main categories, based on the desired output, classification and regression.

- **Classification** problems have categorical output values. For instance classification problems are email spam detection [18] and image classification [19] .
- **Regression** problems have output values that are real numbers, such as stock price prediction [20].

Unsupervised Learning

The unsupervised learning problem involves learning input patterns without given output values (labels). Also known as self-organization, unsupervised learning allows for modeling of probability densities over inputs.

Reinforcement Learning

Reinforcement Learning is an approach to Machine Learning concerned with how systems have to take decisions in order to maximize a cumulative reward. Reinforcement learning differs from supervised learning in not needing labelled input/output pairs to be presented and any sub-optimal actions to be explicitly corrected. The problem setting is typically stated in the form of a Markov decision process, as many reinforcement learning algorithms utilize dynamic programming techniques. Reinforcement Learning algorithms

differ from classical dynamic programming methods in that the first do not assume knowledge of an exact mathematical model of the Markov decision process and they target such processes where exact methods become impracticable.

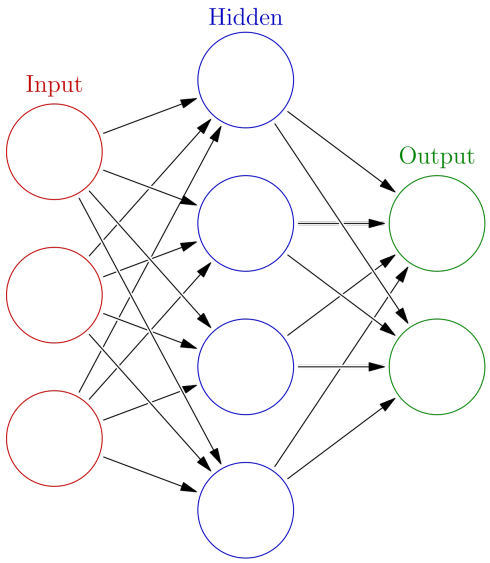
2.3 The Deep Learning Era

Nowadays, deep learning has seen a huge progress and has achieved major breakthroughs in many fields, including computer vision, natural language processing and drug discovery. These achievements became possible mainly due to the following reasons:

- High availability of massive data sets.
- Increased performance of computer processors, GPUs and TPUs.
- More complex neural network architectures.

However, before the deep learning breakthrough, it was common sense for scientists to use handcrafted features. For example in speaker verification the state of the art (SOTA) required manual extraction of voice frequencies, through a method called Mel-frequency Cepstral Coefficients (MFCCs). On the contrary, deep learning neural networks have automated the feature extraction procedure. Their complexity allows them to learn complex functions and achieve new possibilities. In this section we will describe common deep learning architectures.

2.3.1 Feed Forward Neural Networks



Σχήμα 2.1: *Feed Forward Neural Network with 1 hidden layer. Source [3]*

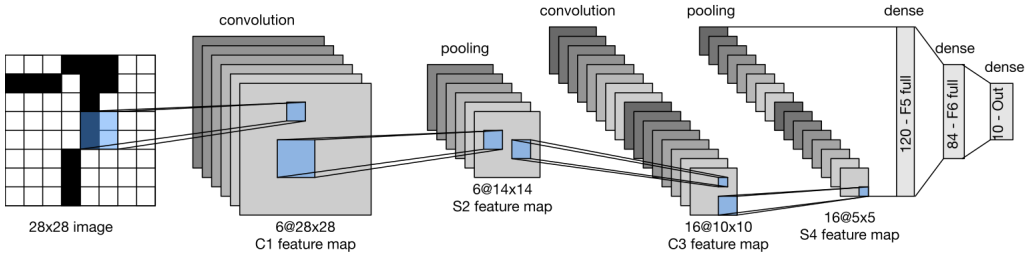
The most typical example of deep learning models is the Feed Forward Neural Network (FFNN) or alternatively multilayer perceptron (MLP). This class of networks consists of multiple layers of computational units, usually interconnected in a feed-forward way. Each

neuron in one layer has directed connections to the neurons of the subsequent layer. The information moves in only one direction —forward— from the input nodes, through the hidden nodes (if any) and to the output nodes. This flow of information from the input to the output is called forward propagation. More generally, any directed acyclic graph may be used for a feed forward network, with some nodes (with no parents) designated as inputs, and some nodes (with no children) designated as outputs.

FFNNs are applied with great success to many problem settings, either alone or as part of a more complex network. Apart from their experimental success they have also theoretical guarantees. The universal approximation theorem[1] states that every continuous function that maps intervals of real numbers to some output interval of real numbers can be approximated, arbitrarily closely, by a MLP with just one hidden layer and a sufficient number of neurons. This result requires a suitable activation function, but it holds for most that are used, one of the first proofs was for the sigmoid [21].

So the multiple neurons structure a "network", which can be represented as a collection of different functions in the form of an directed acylic graph. A FFNN is essentially a mapping $y = f(x;\theta)$ that tries to learn the parameters ϑ, that result in the best possible approximation of the real function. That is to say, if a MLP represents a function f, an alternative representation can be a chain of functions $f(x) = f^{(n)}(f^{(n-1)}(...(f^{(1)}(x))))$, where $f^{(i)}, i \in 1, ..., n$ are generally different functions. $f^{(1)}$ is called the first or input layer, $f^{(2)}$ the second, ... , and $f^{(n)}$ is the output layer. n is the depth of the network, and the deep learning terminology arose from increasing the depth of such networks when data and computational power became ample. When training the FFNN to learn to predict y_i based on the input example x_i, the only value the network computes whose correctness is dictated by the training set is the output y. Therefore, all layers but the output layer (that outputs that y) are called hidden layers. Finally it becomes evident that parameters ϑ are split between layers.

2.3.2 Convolutional Neural Networks



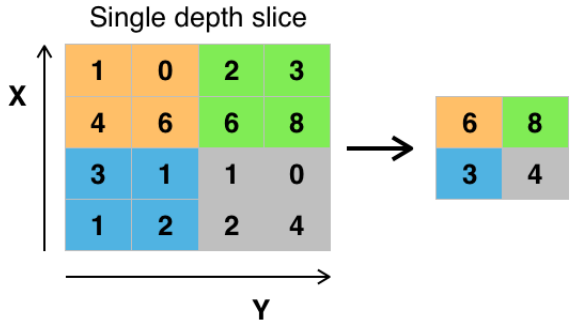
Σχήμα 2.2: *Illustration of CNN [1]*

Another success story of neuroscientific influence on ML are Convolutional Neural Net-

[1] Universal Approximation Theorem Wikipedia

works (CNN). Convolutional neural networks (CNNs) are a specific class of deep neural networks, that is specialized for processing data that have a grid topology, such as images. These networks are considered as one of the greatest biological inspirations in artificial intelligence, as their key design concepts are borrowed from neuroscience, and especially from the organization of the human visual cortex. A convolution network layer is designed to mimic the human cortex, by applying convolutional [2] operations on the input image with many low dimensional filters.

CNNs are regularized versions of FFNNs. FFNNs usually mean fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The "full connectivity" of these networks make them prone to over-fitting data. CNNs take a different approach in order to regularize data and reduce complexity. They take advantage of the hierarchical pattern in data and assemble patterns of increasing complexity using smaller and simpler patterns embossed in their filters. That also offers them ignorance of posotional shifts and target translations. Therefore, on a scale of connectivity and complexity, CNNs are on the lower extreme.



Σχήμα 2.3: *Max Pooling Layer [2]*

The most important aspects of the CNN architecture are following layers:

- **Convolutional layers** in a CNN systematically apply learned filters to input images in order to create feature maps that summarize the presence of those features in the input. They prove very effective, as stacking them in deep models allows layers close to the input to learn low-level features (e.g. lines) and layers deeper in the model to learn high-order or more abstract features, like shapes or specific objects.

- **Pooling layers** form a new layer added after each convolutional layer. Their role is to reduce the spatial size of the features while at the same time transforming them to position invariant. The most frequently used are the max pooling or the average pooling, which differ in the way the apply the dimensionality reduction.

2.3.3 Activation Functions

In order to classify non-linearly separable data points, it is essential for us to introduce non-linearities. This introduction is achieved through activation functions and allow us to

[2]Mathematically, convolution is an operation on two functions, f and g that produces a third function $f * g$. In mathematical notation $(f * g) := \int_{-\infty}^{+\infty} f(\tau)g(t-\tau)d\tau$.

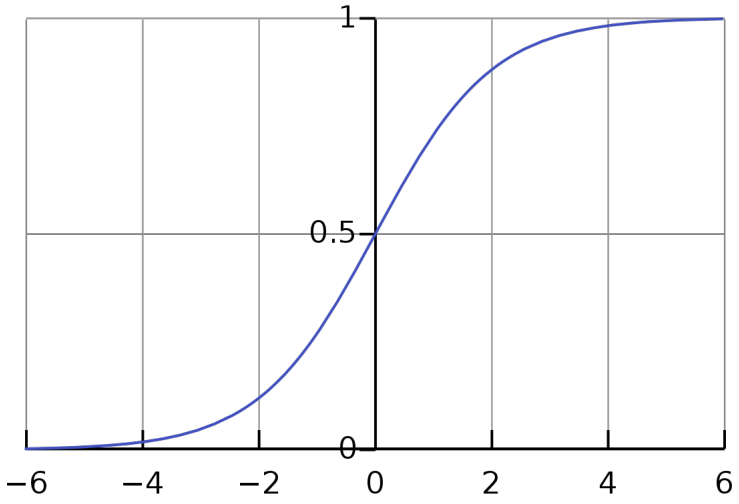
approximate arbitrarily complex functions. Activation functions take as input the output of one node, then apply a function f to produce the final output, making a non-linear decision.

Some common activation functions are the following:

Sigmoid

The sigmoid function is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Σχήμα 2.4: *Sigmoid Curve*

A sigmoid function is a mathematical function having the characteristic shape as shown in 2.4. It is basically a "S"-shaped curve or sigmoid curve. In general, a sigmoid function is monotonic and its first derivative is bell shaped. It takes a real-valued number and maps them to the range $[0, 1]$. Nowadays, the sigmoid is rarely used as it has two major drawbacks. First of all, for values near 0 or 1 the gradient is close to 0 resulting in the so called "vanishing gradient" effect [3]. Secondly, the inputs of the next neurons are always positive as the sigmoid is not zero-centered.

Hyperbolic Tangent (tanh)

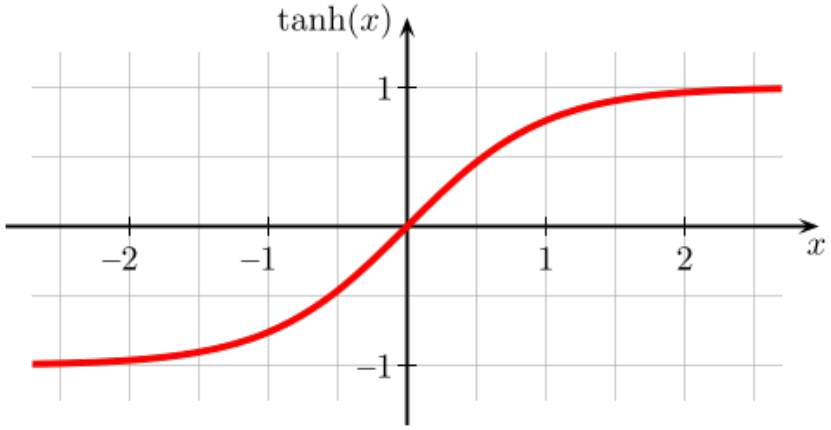
The hyperbolic tangent or tanh function is defined as:

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Similarly to the sigmoid, the hyperbolic tangent is a real function taking every real number as input and mapping them input to the range $[-1, 1]$. It is also bounded and differentiable.

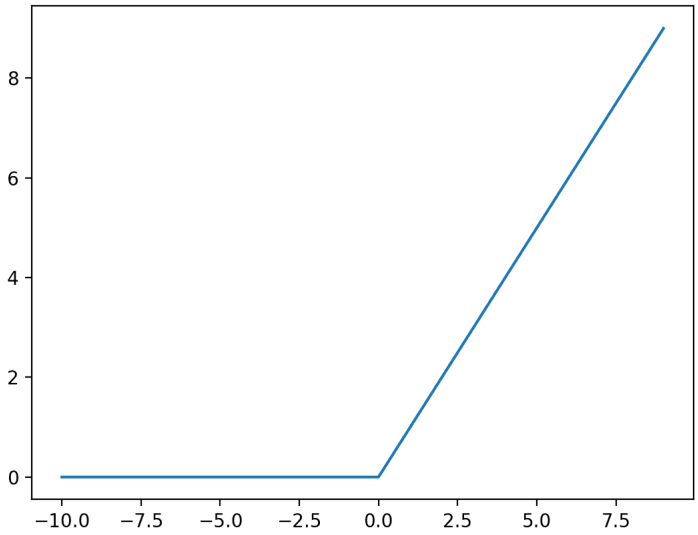
[3] https://en.wikipedia.org/wiki/Vanishing_gradient_problem

Its derivative is also non negative at each point. Unfortunately, it suffers from the vanishing gradient effect exactly as the sigmoid function did, except this one is zero-centered.



Σχήμα 2.5: *Hyperbolic Tangent. Source [4]*

Rectified Linear Unit (ReLU)



Σχήμα 2.6: *ReLU for x in range [-10,10]. Source [5]*

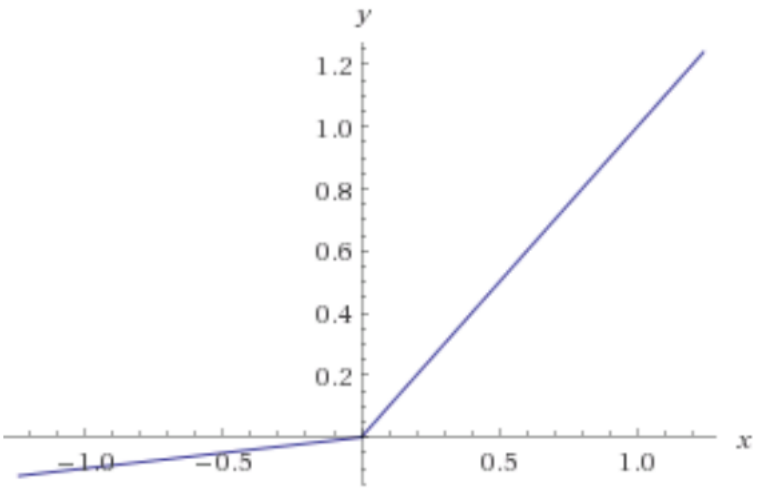
The ReLU activation function is defined as:

$$f(x) = max(0, x)$$

The ReLU activation function is common practice nowadays. The ReLU (otherwise known as ramp function) basically thesholds the input at zero, allowing only positive inputs to pass through. Moreover it accelerates convergance while at the same time, it is not computationally expensive as the aforementioned functions. Its main disadvantage is

the "dying ReLU" problem. This derives from the fact that ReLU is not differentiable at zero and it is also not zero centered and bounded. This results in some neurons becoming inactive and only giving output 0 for any input.

Leaky ReLU



Σχήμα 2.7: *Leaky ReLU. Source [6]*

The Leaky ReLU's activation function is defined in mathematical form as:

$$f(z) = f(x) = \begin{cases} x & \text{if } x > 0, \\ 0.01x & \text{otherwise.} \end{cases}$$

This activation function is an attempt to fix the "dying ReLU" problem that was referenced above. It only allows a small, positive gradient when the unit is not active.

Softmax

The softmax function is basically a generalization of the logistic function to multiple dimensions. It is used in multinomial logistic regression and is often used as the last activation function of a neural network. This is usually done in order to normalize the output of a network to a probability distribution over the predicted output classes. Generally, it takes a vector x of K real numbers as input , and normalizes it into a probability distribution consisting of K probabilities proportional to the input exponentials. Therefore, given an input vector x and a weighting vector w, we have:

$$P(y = j \mid \mathbf{x}) = \frac{e^{\mathbf{x}^\mathsf{T}\mathbf{w}_j}}{\sum_{k=1}^{K} e^{\mathbf{x}^\mathsf{T}\mathbf{w}_k}}$$

2.3.4 Training Pipeline

In order for the neural network to learn efficiently the data distribution, the architecture engineer should also pick an optimizer and a loss function that fits to their needs. Different

loss functions focus on different aspects of learning and different optimizers utilize different algorithms in order to converge.

Loss Functions

A Loss function is a way to measure the performance of a supervised learning task. It's a method to evaluate how the algorithm models the data. It corresponds to a non negative value that measures the error between the predicted and the target output. In mathematical notation, if we denote our model as a function f with parameters w then for a data pair (x_i, y_i) the loss is computed as $L(y_i, f(x_i; w))$. In order to quantify the total loss, we denote an objective function $J(w)$ as empirical risk, which we try to minimize over the entire dataset of size N. That is:

$$J(w) = \frac{1}{n} \sum_{i=1}^{N} L(y_i, f(x_i; w))$$

In this subsection we should also discuss some useful loss functions. The first is **binary cross entropy (BCE)** loss which is suitable for binary classification tasks. If we denote $f(x_i, w)$ as $\hat{y}$ and the ground truth label as y, it is defined as:

$$BCE(y, \hat{y}) = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})]$$

Another important loss function for multi-class classification tasks is the generalization of the binary cross entropy loss named **cross-entropy**. If we denote the probability vector as y (usually with only one element equal to 1 and all others equal to 0) and the output vector of the softmax output function as $\hat{y}$, it is defined as:

$$L(y, \hat{y}) = -\sum_{i} y_i \log \hat{y}_i$$

Optimization

Model training is an optimization process, in which we try to configure the model parameters w that minimize the loss function $J(w)$. An optimizer is a method or algorithm to update the various parameters that can reduce the loss in much less effort. In this subsection we will discuss some frequently used techniques for optimization.

Gradient Descent is a simple optimization algorithm that provides the means to iteratively find close to optimal arguments / parameters of a differentiable function with access only to the derivative at the current configuration of parameters. In detail, given a differentiable function R, we can express the function by its Taylor series approximation:

$$R(w[t] + \Delta w) \approx R(w[t]) + \nabla R(w)|_{w=w[t]} \Delta w$$

So, given parameters $w[t]$, if we are looking for the value of Δw to modify it so as to lower the value of R, then a sensible choice is to select a Δw in the opposite direction

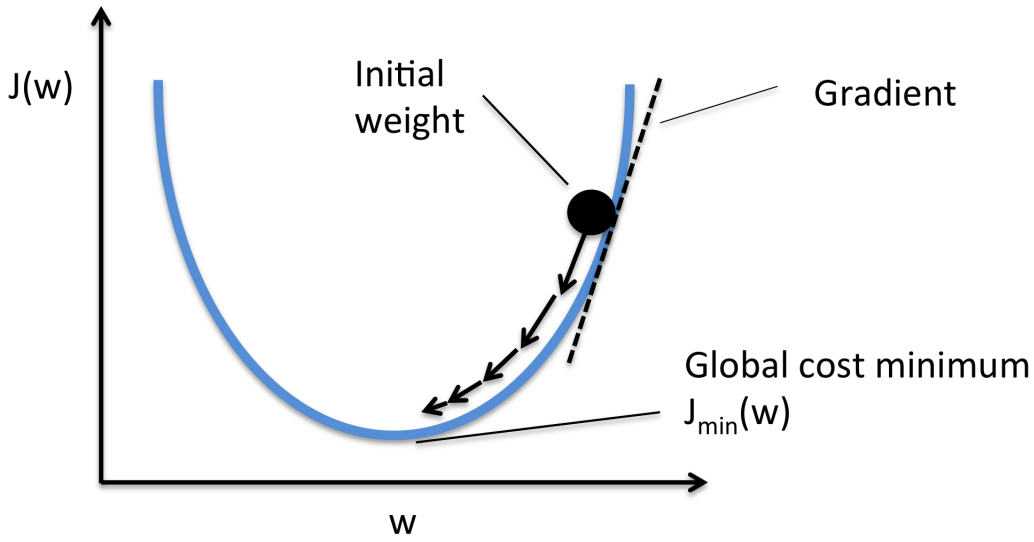
$\nabla R(w)|_{w=w[t]}$, i.e. $\Delta w = -\alpha \nabla R(w)|_{w=w[t]}, \alpha > 0$. So in the next step $w[t+1]$ will be equal to:

$$w[t+1] = w[t] - \alpha \nabla R(x)|_{x=w[t]}$$

and

$$R(w[t+1]) = R(w[t] + \Delta w) \approx R(w[t]) - \alpha \|\nabla R(w)|_{w=w[t]}\|^2 \leq R(w[t])$$

The process is repeated until the desired convergence critirion is met. Is is also important to note that Taylor series approximation holds for small Δw, so a careful choice of α, the **learning rate**, must be made.



Σχήμα 2.8: *Illustration of Gradient Descend on for 1 parameter. Source [7]*

The Gradient Descent algorithm is very simple and effective, but it has a major drawback. Computing the loss over the entire dataset at each iteration is computationally expensive and inefficient, especially for large datasets.

Stochastic Gradient Descent (SGD) is a variant of Gradient Descent that solves the aforementioned problem. It computes the gradient of the loss function over a subset of samples, not for the whole dataset. It makes an estimation of the gradient, instead of computing the true gradient using all samples, therefore the name "stochastic".

In its simplest form, the gradient is computed over each unique training example, but this can lead to very noisy gradients and cause the loss function to fluctuate. For this reason, a variation called mini-batch SGD is commonly used in practice. A mini batch of B training data points is picked and the average gradient over those B points is calculated and used:

$$w[t+1] = w[t] - \alpha \frac{1}{B} \sum_{b=1}^{B} \nabla_w R_b(x)|_{x=w[t]}$$

This method is still fast to compute and gives a much better estimate of the true gradient. The larger the batch size, the more accurate the estimation of the gradient, which leads to smoother convergence and allows for larger learning rates.

One important thing to notice, is that when we train a deep neural network, the loss

surface becomes non-convex because of the introduced non-linearities. This means that there is no guarantee that a gradient based method will converge to a global minimum, where the loss function is zero.

Backpropagation

In machine learning, backpropagation is a widely used algorithm for training feedforward neural networks. In fitting a neural network, backpropagation efficiently computes the gradient of the loss function with respect to the weights of the network for a single input–output example. The backpropagation algorithm works by computing the gradient of the loss function with respect to each weight by the chain rule, computing the gradient one layer at a time, iterating backward from the last layer to avoid redundant calculations of intermediate terms in the chain rule[4]; This idea is essentially an example of dynamic programming.

Normalization

Each layer of a neural network has inputs with a corresponding distribution, which is affected during the training process by the randomness in the parameter initialization and the randomness in the input data. The effect of these sources of randomness on the distribution of the inputs to internal layers during training is described as internal covariate shift. Although a clear-cut precise definition seems to be missing, the phenomenon observed in experiments is the change on means and variances of the inputs to internal layers during training. As all layers are changed during an update, the update procedure is forever chasing a moving target. For example, the weights of a layer are updated given an expectation that the prior layer outputs values with a given distribution. This distribution is likely changed after the weights of the prior layer are updated. Furthermore, internal covariate shift results in slower learning rates and careful parameter initialization making training hard.

A method that addresses this problem is called **batch normalization**. This method performs a normalization over the entire mini-batch. Utilizing batch normalization during training, fixes the mean and the variance of each input layer and at the same time acts as a regularizer.

Dropout

Dropout is a useful regularization method for neural network training. It helps in reducing overfitting by preventing complex co-adaptations on training data. The term *dropout* refers to randomly "dropping out", or omitting, units (both hidden and visible) during the training process of a neural network.

[4] https://en.wikipedia.org/wiki/Chain_rule

2.4 Deep Learning for Speech Emotion Recognition (SER)

2.4.1 Overview of the Field

Emotion recognition has seen huge improvement over the recent years and aims at better communication between man and machine. There are many channels of communication between people: the content of speech, the nods, face and body movements and emotions. In order to make human-computer interaction more clear and comfortable, the computer utilize emotional knowledge to understand and react appropriately to human emotions. It has even been supported that emotional intelligence in computers is more important than computing and verbal, to make applications more human-friendly. Emotional intelligence is necessary to determine people's preferences and to adapt computers to the characteristics of each person individually. Some real world examples of speech emotion recognition models applications are:

- Applications in **psychology**. The automatic detection of emotionally charged moments would facilitate psychologists in their work.
- Automation in **call centers**. In automated call centers many people find it difficult to communicate with the voice recognition machine and lose their patience, as a result of which their request is not fulfilled. If there was automatic recognition of anger, frustration and resentment, the customer could be led to a human representative without suffering.

Automatic speech emotion recognition (SER) is achieved by the development of methodologies based on digital signal processing and machine learning. The journey of research in this field is three decades-long; still, the results are not good enough to be applied in natural environments with high accuracy. There is a multitude of information present in the speech signal. A speech signal contains lexical contents (what has been spoken), speaker (who is the speaker), emotions (how it has been spoken), and language (in which language it has been spoken). If one has to recognize particular information in speech, then ideally the effect of other information should be nullified. For example, if one has to recognize emotion from speech, then the effect of the speaker, lexical content, and language should ideally be nullified to generalize the SER system. This is the primary reason why automatic SER systems don't work very well for real-life applications. This problem occurs due to mismatch of speaker, text, language, and culture – collectively referred to as 'environment' – in the training and testing data. As a result, the accuracy significantly decreases in the case of real-life applications or 'natural environment'. Here, 'natural' refers to the variation of speakers, text, language, culture, surroundings, etc., within and across the development and deployment environments of SER systems.

2.4.2 Basic Emotions

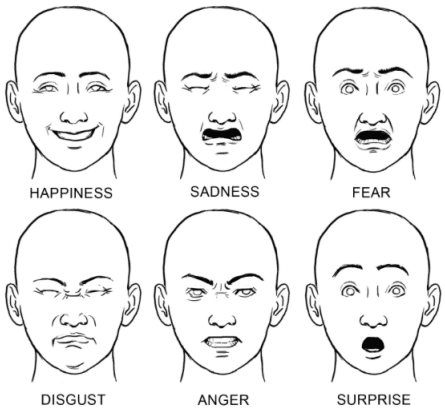
Emotions are psychological states brought on by neurophysiological changes, variously associated with thoughts, feelings, behavioural responses, and a degree of pleasure or dis-

pleasure.There is currently no scientific consensus on a definition. Emotions are often intertwined with mood, temperament, personality, disposition, creativity, and motivation.

Emotions are complex. There are various theories on the question of whether or not emotions cause changes in our behaviour. On the one hand, the physiology of emotion is closely linked to arousal of the nervous system. Emotion is also linked to behavioral tendency. Extroverted people are more likely to be social and express their emotions, while introverted people are more likely to be more socially withdrawn and conceal their emotions. Emotion is often the driving force behind motivation. On the other hand, emotions are not causal forces but simply syndromes of components, which might include motivation, feeling, behaviour, and physiological changes, but none of these components is the emotion. Nor is the emotion an entity that causes these components.

Classification

For more than 40 years, Paul Ekman[5] has supported the view that emotions are discrete, measurable, and physiologically distinct. Ekman's most influential work revolved around the finding that certain emotions appeared to be universally recognized, even in cultures that were preliterate and could not have learned associations for facial expressions through media. Another classic study found that when participants contorted their facial muscles into distinct facial expressions (for example, disgust), they reported subjective and physiological experiences that matched the distinct facial expressions. Ekman's facial-expression research examined six basic emotions: anger, disgust, fear, happiness, sadness and surprise.

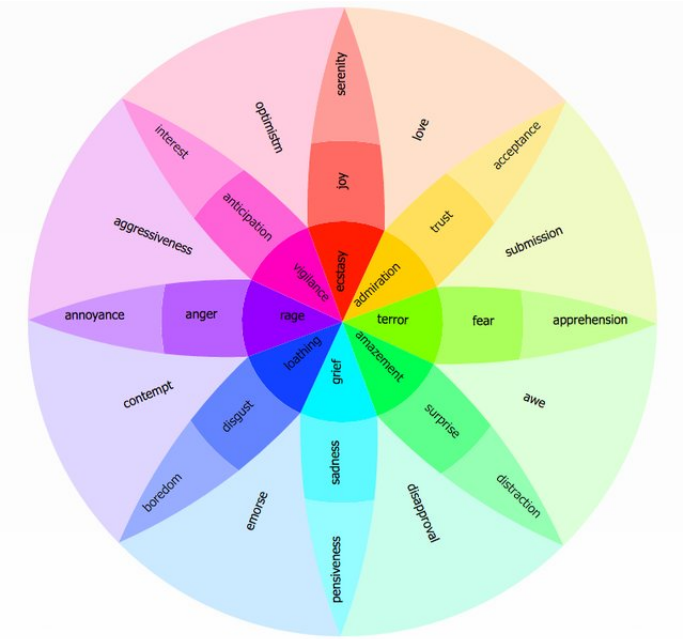


Σχήμα 2.9: *Basic emotions as Ekman firstly defined them. Source [8]*

Later in his career, Ekman theorized that other universal emotions may exist beyond these six. In light of this, recent cross-cultural studies led by Daniel Cordaro and Dacher Keltner, both former students of Ekman, extended the list of universal emotions. In addition to the original six, these studies provided evidence for amusement, awe, contentment, desire, embarrassment, pain, relief, and sympathy in both facial and vocal expressions. They also found evidence for boredom, confusion, interest, pride, and shame facial expressions, as well as contempt, relief, and triumph vocal expressions.

[5] https://en.wikipedia.org/wiki/Paul_Ekman

Robert Plutchik agreed with Ekman's biologically driven perspective but developed the "wheel of emotions", suggesting eight primary emotions grouped on a positive or negative basis: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation. Some basic emotions can be modified to form complex emotions. The complex emotions could arise from cultural conditioning or association combined with the basic emotions. Alternatively, similar to the way primary colors combine, primary emotions could blend to form the full spectrum of human emotional experience. For example, interpersonal anger and disgust could blend to form contempt. Relationships exist between basic emotions, resulting in positive or negative influences.

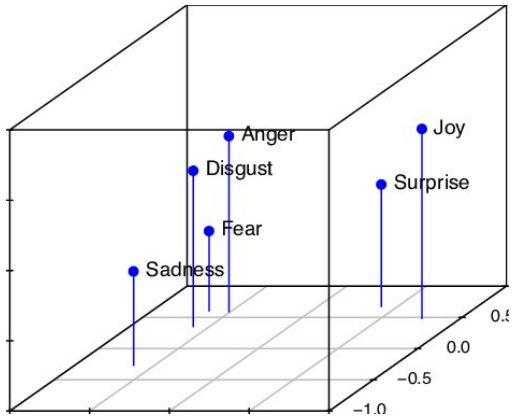


Σχήμα 2.10: *The "wheel of emotions" as developed by Robert Plutchik. Source [9]*

Psychologists have used methods such as factor analysis to attempt to map emotion-related responses onto a more limited number of dimensions. Such methods attempt to boil emotions down to underlying dimensions that capture the similarities and differences between experiences. Often, the first two dimensions uncovered by factor analysis are valence (how negative or positive the experience feels) and arousal (how energized or enervated the experience feels). These two dimensions can be depicted on a 2D coordinate map. This two-dimensional map has been theorized to capture one important component of emotion called core affect. Modern research suggests that a three-dimensional mapping can be used as well and capture an extra dimension, called dominance.

2.4.3 Feature Extraction

In machine learning, pattern recognition, and image processing, feature extraction starts from an initial set of measured data and builds derived values (*features*). Its main goal is to reduce the dimensionality of the data by extracting informative and non-redundant interpretations.



Σχήμα 2.11: *Valence-Arousal-Dominance (VAD) model. For illustration, the position of Ekman's six Basic Emotions are included. Source [10]*

As it is for speech signal processing, raw waveform is frequently replaced by more dense representations. In this subsection, we shall discuss some of them.

Mel-Frequency Cepstral Coefficients (MFCCs)

The mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. This log scale concept derives from the human auditory system's response, as humans are much more capable in distinguishing low frequencies rather than high, thus approximating it more closely than the linearly-spaced frequency bands used in the normal spectrum.

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. In order to compute the MFCC of a sound signal, the process requires:

1. Applying the fourier transform[6] to the signal.

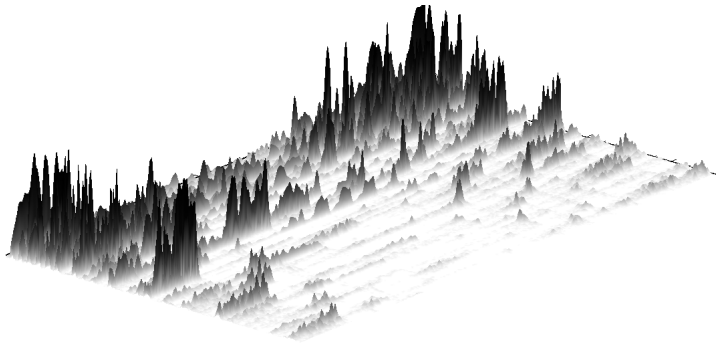
2. Mapping the powers of the spectrum obtained above onto the mel scale. This mapping is accomplished, using triangular overlapping windows or alternatively, cosine overlapping windows.

3. Taking the logs of the powers at each of the mel frequencies.

4. Taking the discrete cosine transform of the list of mel log powers, as if it were a signal.

5. Finally, the MFCCs are the amplitudes of the resulting spectrum.

One main disadvantage of the MFCC values, is that they are not very robust in the presence of additive noise. As a result, it is common to normalise their values to lessen the influence of noise.



Σχήμα 2.12: *Three-dimensional spectrogram of a part from a music piece.*

Mel Spectrogram

A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. Spectrograms are used extensively in the fields of music, linguistics, sonar, radar, speech processing, seismology, and others. Spectrograms of audio can be used to identify spoken words phonetically, and to analyse the various calls of animals.

Spectrograms may be created from a time-domain signal in one of two ways: approximated as a filterbank that results from a series of band-pass filters (this was the only way before the advent of modern digital signal processing), or calculated from the time signal using the Fourier transform. These two methods actually form two different time–frequency representations, but are equivalent under some conditions.

After applying a Mel-scale on the Spectrogram, Mel Spectrogram is computed.

2.4.4 Datasets

IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [22] database is an acted, multimodal and multispeaker database. The IEMOCAP dataset consists of 151 videos of recorded dialogues, with 2 speakers per session for a total of 302 videos across the dataset. Each segment is annotated for the presence of 9 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral) as well as valence, arousal and dominance (VAD). The dataset is recorded across 5 sessions with 5 pairs of speakers.

RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [23] contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

[6] Fourier Transform Wikipedia

2.5 Deep Learning for Speaker Recognition (SR)

2.5.1 Overview of the Field

Speaker recognition is the identification of a person from characteristics of voices. It has a history dating back some four decades as of 2021 and uses the acoustic features of speech that have been found to differ between individuals. These acoustic patterns reflect both anatomy and learned behavioral patterns. Some real world examples of a speaker recognition model applications are:

- Applications in **security**. Many modern security systems utilize speaker recognition models as a means to authorize a person for accessing important documents.

- Applications in **speech recognition**. Speaker recognition technology can be used to reduce speaker variability in speech recognition systems by speaker adaptation.

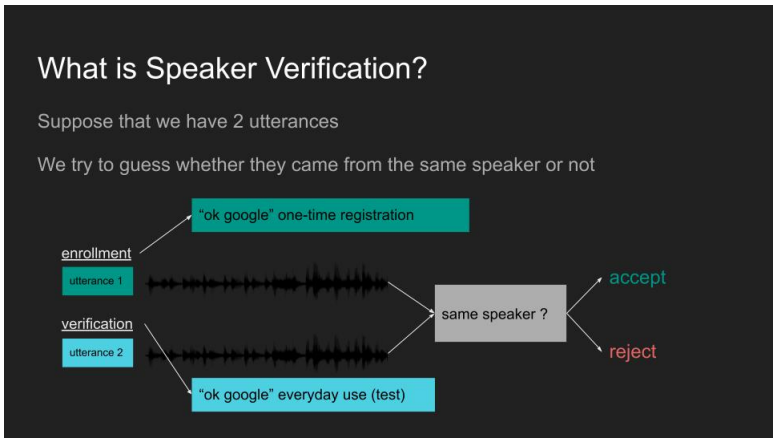
Verification versus Identification

There are two major applications of speaker recognition technologies and methodologies. If the speaker claims to be of a certain identity and the voice is used to verify this claim, this is called verification or authentication. On the other hand, identification is the task of determining an unknown speaker's identity. In a sense, speaker verification is a 1:1 match where one speaker's voice is matched to a particular template whereas speaker identification is a 1:N match where the voice is compared against multiple templates.

From a security perspective, identification is different from verification. Speaker verification is usually employed as a "gatekeeper" in order to provide access to a secure system. These systems operate with the users' knowledge and typically require their cooperation. Speaker identification systems can also be implemented covertly without the user's knowledge to identify talkers in a discussion, alert automated systems of speaker changes, check if a user is already enrolled in a system, etc.

In forensic applications, it is common to first perform a speaker identification process to create a list of "best matches" and then perform a series of verification processes to determine a conclusive match. Working to match the samples from the speaker to the list of best matches helps figure out if they are the same person based on the amount of similarities or differences. The prosecution and defense use this as evidence to determine if the suspect is actually the offender.

Each speaker recognition system has two phases: **enrollment** and **verification**. During enrollment, the speaker's voice is recorded and typically a number of features are extracted to form a voice print, template, or model. In the verification phase, a speech sample or "utterance" is compared against a previously created voice print. For identification systems, the utterance is compared against multiple voice prints in order to determine the best match(es) while verification systems compare an utterance against a single voice print. Because of the process involved, verification is faster than identification.



Σχήμα 2.13: *Enrollment and Verification phase in a real world example*

Text Dependent versus Text Independent

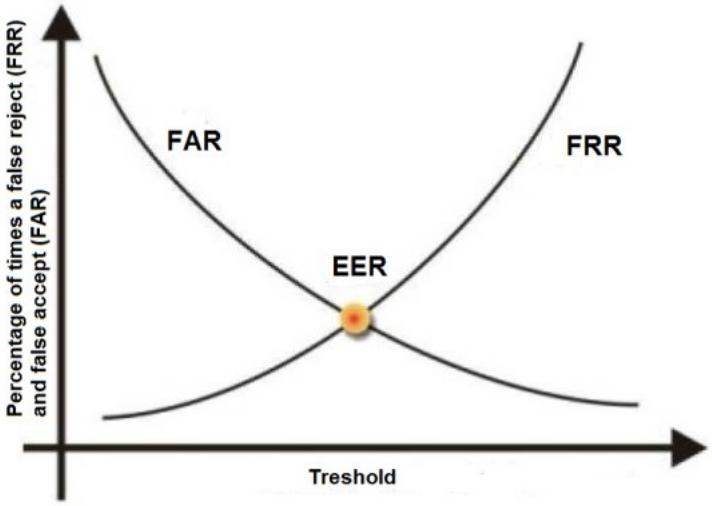
Speaker recognition systems fall into two categories:

- Text-Dependent (TD-SV). If the text must be the same for enrollment and verification this is called text-dependent recognition. In a text-dependent system, prompts can either be common across all speakers (e.g. a common pass phrase) or unique. In addition, the use of shared-secrets (e.g.: passwords and PINs) or knowledge-based information can be employed in order to create a multi-factor authentication scenario.
- Text-Independent (TI-SV). Text-independent systems are most often used for speaker identification as they require very little if any cooperation by the speaker. In this case the text during enrollment and test is different. In fact, the enrollment may happen without the user's knowledge, as in the case for many forensic applications. As text-independent technologies do not compare what was said at enrollment and verification, verification applications tend to also employ speech recognition to determine what the user is saying at the point of authentication. In text independent systems both acoustics and speech analysis techniques are used.

2.5.2 History and commonly used methods

Equal Error Rate (EER)

In order to evaluate a SV system, we take advantage of EER. An SV system predicts whether a person is an authenticated user with a probability p. Then a decision should be taken for whether to accept or reject the user based on a threshold θ. A low threshold would result in accepting all users but at the same time accepting many impostors. On the other hand, a high threshold would not allow impostors, with the risk of rejecting true clients themselves. This trade off between rejections in clients and impostors is well described using **false acceptance rate (FAR)** and **false rejection rate (FRR)**. FAR depicts the rate in which the system falsely accepts impostors. FRR depicts the rate in which the system falsely rejects an authenticated users. EER is the point where these two curves intersect.



Σχήμα 2.14: *EER as the point where FAR and FRR curves intersect. Source [11]*

In order to define EER we should first define FAR and FRR. If we denote

$$FAR = \frac{\#false\ acceptance}{\#identification\ attempts}$$

and

$$FRR = \frac{\#false\ rejection}{\#identification\ attempts}$$

then

$$EER = min_\theta\ (\mid FAR(\theta) - FRR(\theta)\ \mid)$$

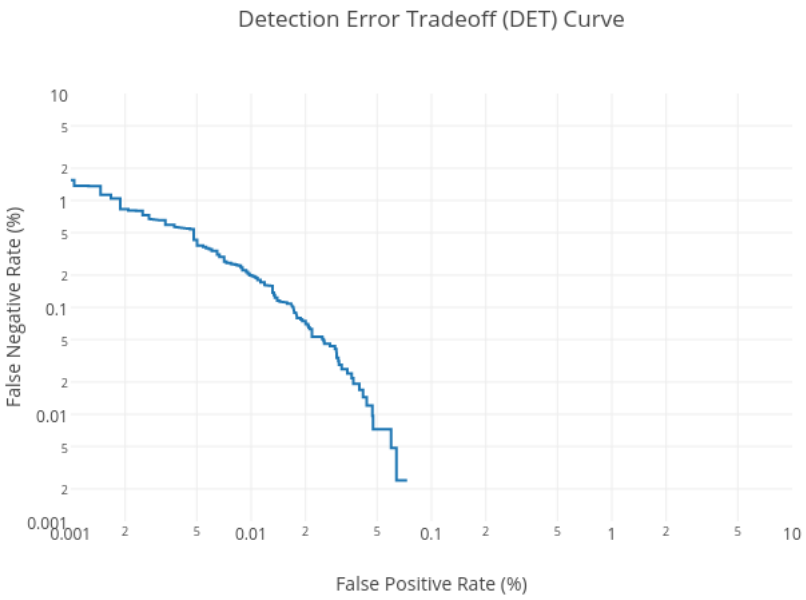
DET Curve

The two error rates on the speaker verification field are functions of the decision threshold. As a result, it is common to represent the performance of such a system by plotting *False Acceptance Rate* as a function of *False Rejecetion Rate*. This curve is monotonus, decreasing and called system operating characteristic plot.

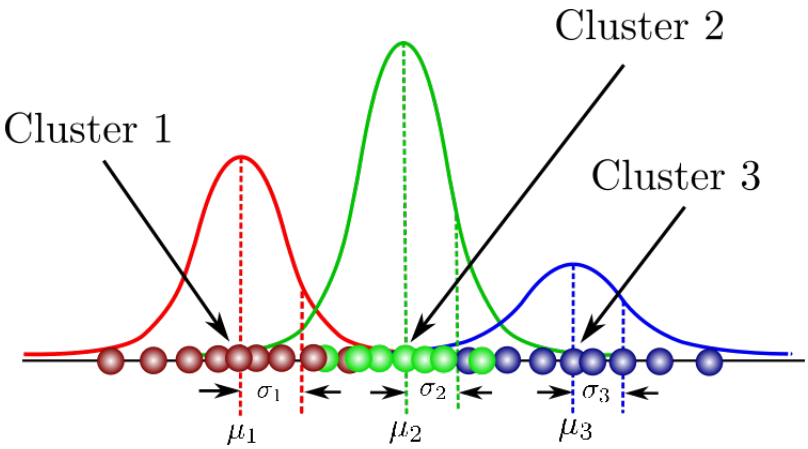
If we plot the error curve on a normal deviate scale **??** in which case the curve is known as the detection error trade-offs (DETs) curve. The closer to the origin the curve appears, the better the system is. In real scenarios, the score distributions are not exactly Gaussians but are quite close to it. Therefore, the DET curve representation is more clearly readable and allows for a comparison of the system's performances on a large range of operating conditions. Figure 2.15 shows a typical example of a DET curve.

Gaussian Mixture Model

A Gaussian Mixture is a function that is comprised of several Gaussians, each identified by $k \in 1, \ldots, K$, where K is the number of clusters of our dataset. Each Gaussian k in the mixture is comprised of the following parameters:



Σχήμα 2.15: *A detection error trade-off curve [12]*



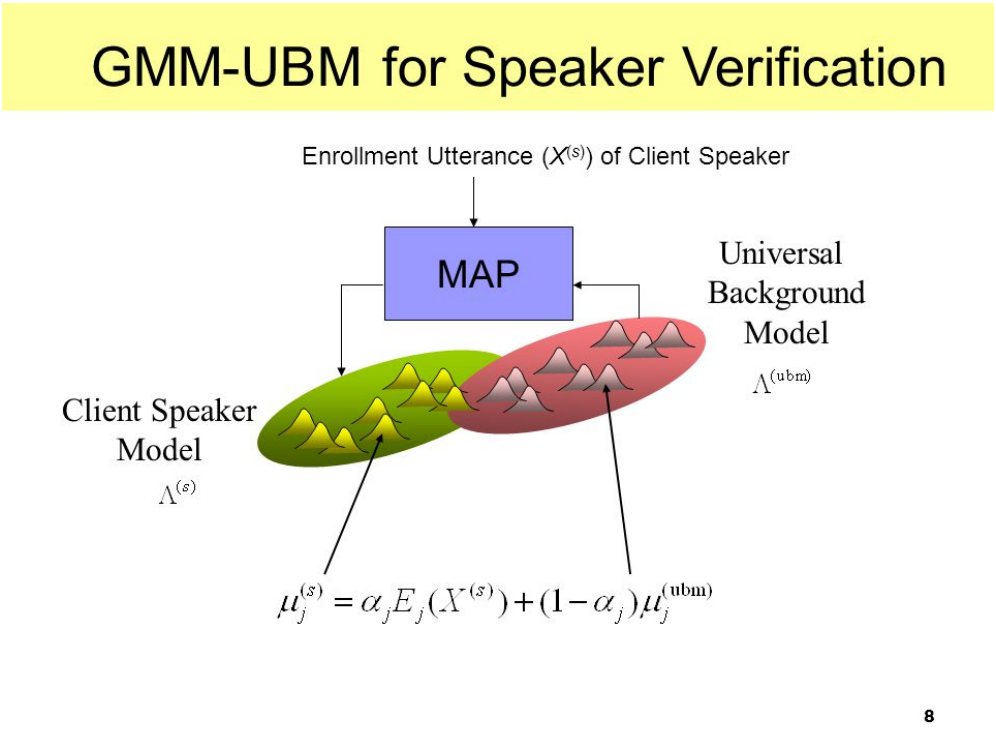
Σχήμα 2.16: *Gaussian Mixture Model for three clusters in two-dimensional space. [13]*

- A mean μ that defines its centre.
- A covariance Σ that defines its width.
- A mixing probability π that defines how big or small the Gaussian function will be.

As part of the "fitting" process, in order for each distribution to match a cluster, a widely used algorithm for optimization problems, where the objective function has complexities is used. It is called **Expectation-Maximization (EM)**[7].

As stated in [24] , the GMM can be viewed as a hybrid between a parametric and nonparametric density model. Like a parametric model it has structure and parameters that control the behavior of the density in known ways, but without constraints that the data must be of a specific distribution type, such as Gaussian or Laplacian. Like a

[7] EM-algorithm



Σχήμα 2.17: *A universal background model with a client-speaker model [14]*

nonparametric model, the GMM has many degrees of freedom to allow arbitrary density modeling, without undue computation and storage demands.

Universal Background Model (UBM)

In the early stages of speaker verification, Universal Background Model was the state of the art method for discriminating different speakers. This model was used as a mean to represent person-independent feature characteristics, which would later be compared against a model of person-specific feature characteristics. This comparison provided the final answer of whether we should accept or reject that person.

UBM is actually a Gaussian Mixture Model (GMM) which was trained with speech samples from a large set of speakers to represent general speech characteristics. Providing a speaker-specific GMM, which was trained on particular enrollment utterances from a specific speaker, we can compute a likelihood ratio between the match score of that model and the UBM. The UBM may also be used while training the speaker-specific model by acting as a the prior model in Maximum A Posteriori (MAP) parameter estimation. For the two-class hypothesis problem:

- $H_0 : X^{(c)}$ comes from the speaker
- $H_1 : X^{(c)}$ comes from an impostor

$$score = \log \frac{P(X^{(c)}|H_0)}{P(X^{(c)}|H_1)} = \log P(X^{(c)}|H_0) - \log P(X^{(c)}|H_1)$$

Afterwards, a threshold θ is applied and if the score is greater than that, the speaker is accepted. Otherwise he is considered an impostor and is rejected.

$$decision = \begin{cases} accept & , \text{if } score \geq \theta \\ reject & , \text{if } score < \theta \end{cases}$$

Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is a method used in statistics and other fields, to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

LDA is closely related to principal component analysis (PCA) [8] in that they both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA, in contrast, does not take into account any difference in class.

Discriminant analysis is used when groups are known a priori (unlike in cluster analysis). Each case must have a score on one or more quantitative predictor measures, and a score on a group measure. In simple terms, discriminant function analysis is classification - the act of distributing things into groups, classes or categories of the same type.

Loss Functions

Over the last years, severall different loss functions have been proposed in order to train speaker verification models effeciently. In this subsection, we shall explain some of those:

- **Triplet Loss**: This loss samples mini-batch of triplets. If we denote X_α our anchor point, X_p our positive example and X_n our negative example we construct a tuple as follows. Each one of them is a $T = (X_\alpha\ , X_p\ , X_n)$ and the aim is to push away the negative point and bring as near as possible the positive point.

- **N-pair Loss**: In order to fix the traditional triplet loss issue, where it only payed attention to the information of one negative sample in each optimization, in [25] the authors introduce N-pair loss. This loss is actually a generalization of the triplet loss, when $N = 2$. This loss optimizes the identification of a possible example from $N - 1$ negative examples.

2.5.3 Datasets

VoxCeleb

As introduced in [26], VoxCeleb is a large scale dataset that consists over $100,000$ utterances for $1,251$ celebrities. These utterances are extracted from videos uploaded on YouTube via a fully automated pipeline based on computer vision techniques. The data is

[8] Principal Component Analysis

Σχήμα 2.18: *Frames from VoxCeleb 1 Dataset*

mostly gender balanced (males comprise of 55%). The celebrities span a diverse range of accents, professions, and age. There is no overlap between the development and test sets.

VoxCeleb is consider classic nowadays for speaker recognition tasks. Due to the large amount of speakers, it is excellent for such applications. Authors later produced an updated version of VoxCeleb, named VoxCeleb2 [27], which contains over 1 million utterances for 6,112 celebrities. It is common for speaker recognition tasks to train on VoxCeleb 2 and then test on VoxCeleb 1. Nonetheless, VoxCeleb1 includes its own evaluation file, which we used in this thesis.

Κεφάλαιο 3

Emotion Driven Speaker Verification

This chapter contains the main contribution of our thesis. First, we provide a detailed related work section. Then we proceed with the problem definition and our methodology.

3.1 Related Work

I-Vectors

I-Vectors have been the state of the art model for many years in the speaker verification field. This model, introduced in [28], instead of separating speaker and channel variability subspaces, models them both in a "total variability space". This idea derives from the fact that speaker variability was originally attributed to channel effects, ignoring intra-speaker and phonetic variation effects involved. For each segment, this model creates a low-dimensional representation, also called identity vector (for short i-vector).

The main idea is that session- and channel dependent supervectors of contatenated GMM means can be modeled as:

$$s = m + Tw$$

In the equation above, m is the session- and channel-independent component of the mean supervector, T is a matrix of bases spanning the subspace covering the important variability (both speaker- and session-specific) in the supervector space, and w is a standard normally distributed latent variable. For each utterance (observation), i-vector is computed as the Maximum A Posteriori (MAP) [1] point estimate of the latent variable w.

X-Vectors

X-vectors, take advantage of deep neural networks (DNN), so as to capture speaker discriminative characteristics. As introduced in [29], x-vectors utilize a time-delay acoustic model[2] with p-norm nonlinearities and LDA for dimensionality reduction. At the final layer the representations are length-normalized and modeled by PLDA.

Combining best of both worlds, x-vectors had a strong contender for next-generation representations for speaker recognition and became state of the art, replacing i-vectors.

[1] Maximum A Posteriori
[2] Time Delay Neural Network

X-vectors meet emotions: A study on dependencies between emotion and speaker recognition

In this work, authors came to the conclusion that knowledge learned for speaker recognition can be reused for emotion recognition through transfer learning. As stated in [30] by fine-tuning, they obtained 30.40%, 7.99%, and 8.61% absolute improvement on IEMOCAP, MSP-Podcast, and Crema-D respectively over baseline model with no pre-training. Finally they tested the effect of emotion on the performance of the speaker verification system by creating speaker verification trials by comparing every utterance against each other. These cross-emotion and same-emotion trials were then tested on speaker verification task and obtained that neutral pairs performed best in IEMOCAP (2.4.4) and MSP-Podcast[3].

Generalized End to End Loss (GE2E Loss)

As introduced in [31] by Google, GE2E loss is an efficient loss function for training speaker verification systems. Its architecture constructs tuples from input sequences of various lengths in a more efficient way, leading to a significant boost of performance and training speed for both TD-SV and TI-SV.

GE2E loss exploits the network output and constructs a compact representation, called **centroid** for every speaker in a batch. A centroid C_j is computed for each speaker in the batch, as a linear combination of all the utterances e_{ji} of the speaker j. Then, it calculates a similarity matrix $S_{ji,k}$ that holds the cosine similarity between a utterance and a centroid, thus representing the effectiveness of the centroids. Ideally, after training, a neural network should map different speakers far apart in the latent space, whereas utterances from the same speaker should fall near each other.

3.2 Problem Setup

3.2.1 Intuition

The main point of our thesis, is to identify dependencies between speaker discriminative voice characteristics and emotion. Assuming that a speakers voice contains information such as age, gender and emotion on top of the linguistic attributes, it is easy to see that voice identity can not be unambiguously settled. This arises a question about how do all those characteristics alternate voice. A logical assumption would be, that, if these characteristics modify voice in a complex manner, it would be difficult for a SV system to correctly classify a user. In this thesis, we explore the effect of emotions in such a scenario.

In order for us, to be able to explore the effect of emotional content on speech, we should first define our baseline models for both the emotion and the speaker verification task. Then we should conduct a list of experiments, so as to explore emotional's speech results.

[3] MSP-Podcast

3.2.2 Approach

In this thesis, we explore the effect of emotional content on speech in a way that addresses both pivotal issues that we have identified with previous work and the reliance on traditional machine learning models.

We attempt to prove and quantify the problem, and then suggest some traditional transfer learning methods, towards reducing or even eliminating the effect. Namely we use a pretrained **emotional model** and try to transfer emotional information to a speaker verification model. Then we create a set of experiments in an effort to address the following questions:

- Does emotional pre-knowledge improve the speaker verification task ?
- Does emotional content have an effect on speaker verification task ?
- Does emotional intensity affect the speaker verification task ?
- How does each emotion affect the speaker verification task ?
- Do same-emotion pairs on verification and enrollment perform better ?

3.2.3 Models Architecture

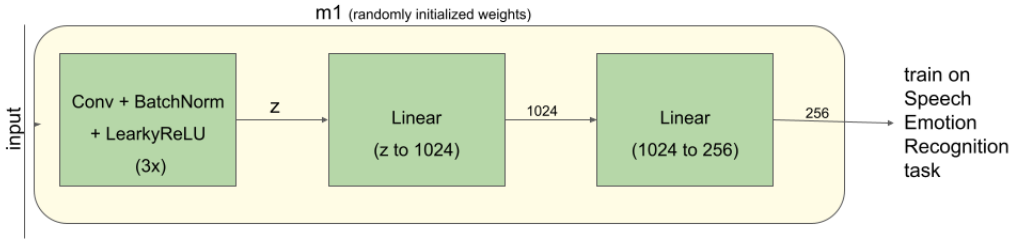
Before running any experiment, we construct two models. Our first model $m1$ is trained on the speech emotion recognition (SER) task. We treat this model as the *emotional brain* of our system. Our second model $m2$ is a speaker verification (SV) system. Our goal is to transfer knowledge from $m1$ to $m2$ and then examine their performance on different scenarios.

In this direction, we assemble and evaluate different architectures, all of which aim at discovering the most efficient way of transferring emotional knowledge into the speaker verification task. We split our setup below, into two subsections: the emotional and the speaker discriminative. For each one we try to discover the best standalone architecture, but at the same time the most efficient ensemble.

Emotional Model

Our emotional model consists of a Convolutional Neural Network (CNN) with linear layer at its final layers. More specifically, our model consists of 3 sequential convolutional followed by 2 linear layers. Each convolution uses 32 channels, and a kernel of size 5 followed by batch normalization, a LeakyReLU unit and a max pooling (2-dimensional). The linear layers project the output to 1024 and 256, making use of dropout with values 0.75 and 0.5 respectively. We made use of Cross Entropy Loss, set the optimizer to SGD and trained for around 50 epochs on the 4-class task on IEMOCAP. Ideally, we would want our model to correctly classify all the following emotions: *neutral, anger, happiness and sadness*.

Our metric for these experiments is the f1-macro score and the accuracy. The results are presented in the table 3.1. The confusion matrix is presented on Figure 3.2.

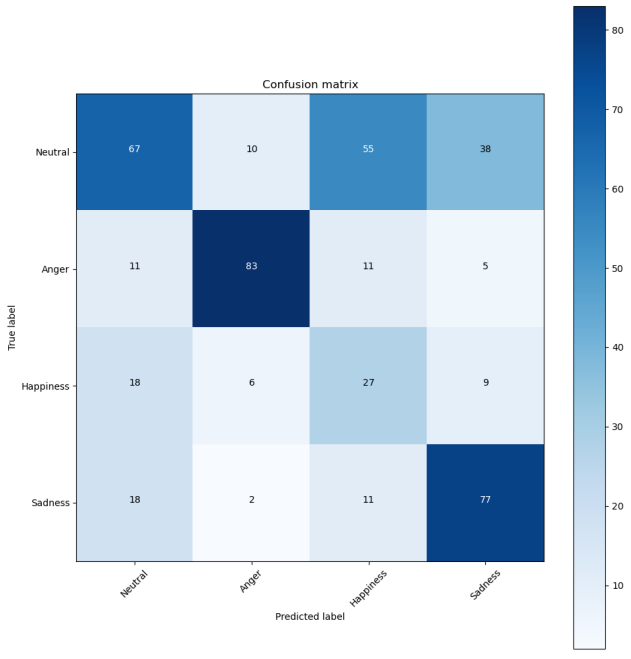


Σχήμα 3.1: *Model m1: trained on SER and utilized as an emotional brain for experiments later on.*

model	accuracy %	f1-macro %
emotional brain $m1$	56.7	55.94

Πίνακας 3.1: *The training results of model m1.*

It is important to notice, that $m1$ is trained as the emotion aware part from which we shall derive "emotional knowledge" later on. Therefore, it is crucial for the whole procedure to correctly classify the emotions of the 4-class task. If that is not the case, problems could arise, as it could disorientate model $m2$.

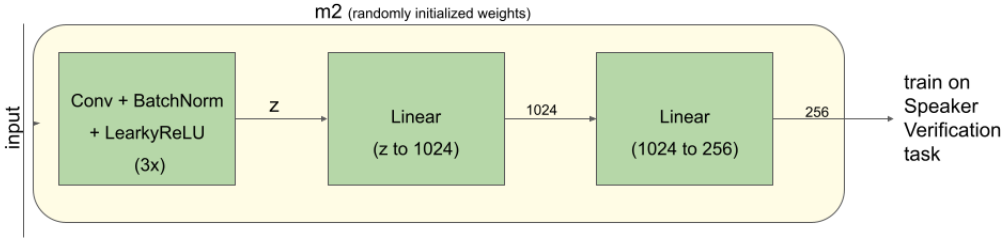


Σχήμα 3.2: *The confusion matrix on the 4 classes task of IEMOCAP dataset.*

We observe that our model is sensitive to neutral class classification. This means that it easily misclassifies neutral with happiness or sadness. Emotion misclassification could become a problem later on, when speaker verification utterances are grouped by their emotional content.

Verification Model

In anticipation of better capturing emotional knowledge, we construct 4 different architectures. Each one of them was evaluated on every emotional experiment. On top of that, different learning rates were used, assuring that the lower EER point will be found. The architectures were the following:

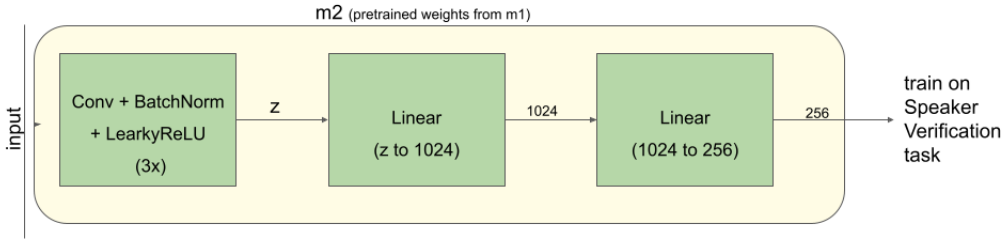


Σχήμα 3.3: *t0 model: It does not contain any emotional information.*

- *model* $t0$: This model will be utilized as our **emotion unaware** model in our experiments. This means that, no emotional information should be specified during its training. It only consists of the model $m2$ completely ignoring the emotional effect on speech. This model's architecture is visualized on 3.3.

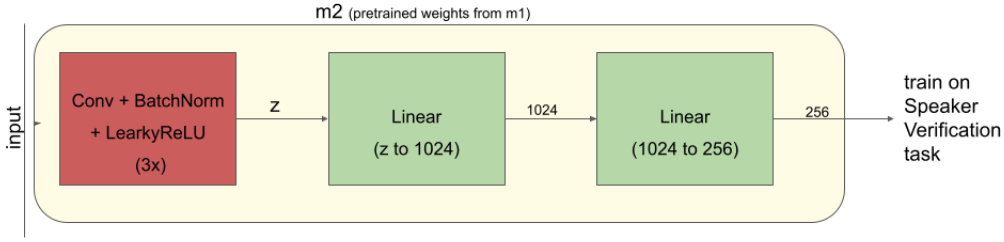
- *model* $t1$: This model aims to transfer knowledge from $m1$ to $m2$ by applying fully finetuning. More specifically, we use $m1$ pretrained weights and biases as a starting point for $m2$. Then $m2$ is fully trained on the SV task. In this process, different learning rates are used, the best of which is kept as the **emotion-aware** $t1$ model. This model's architecture is visualized on 3.4.

- *model* $t2$: This model aims to transfer knowledge from $m1$ to $m2$ through fine tuning in a unique way. Our goal is to keep the points where the original network focused on the audio signal constant. We make use of $m1$ pretrained weights and biases as a starting point for $m2$. We force every weight and bias on the convolutional layers to be **frozen**. This means that no weight or bias in these layers can be updated, during $m2$ training process. Then $m2$ is trained on the SV task. Finally, different



Σχήμα 3.4: *t1 model: Its weights and are initialized as the m1's model. The parts of the network tuned are painted with green.*

.

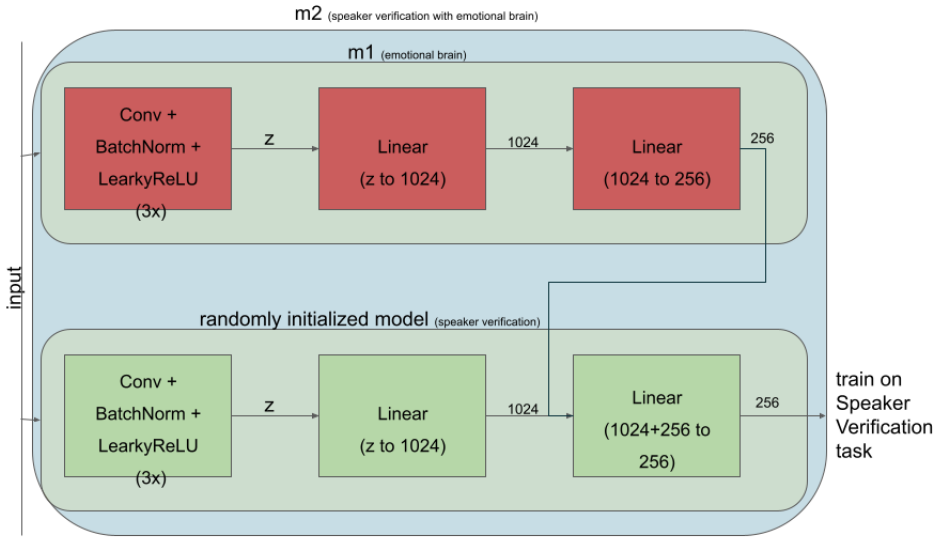


Σχήμα 3.5: *t2 model: Its weights and are initialized as the m1's model. The parts of the network tuned are painted with green.*

.

learning rates are used, the best of which is kept as the **emotion-aware** $t2$ model. This model's architecture is visualized on 3.5.

- *model* $t3$: This model aims to transfer knowledge from $m1$ to $m2$ utilizing a **fusion** architecture. More specifically, we use $m1$ pretrained weights and biases as a starting point for a sub-network in $m2$. That splits the neural network into two parts. The "emotional brain", as $m1$ is now a part of the bigger network, and the speaker discriminative part. During the training procedure, the emotional part is not trained. Its contribution in the overall process is to provide robust emotional embeddings to



Σχήμα 3.6: *t3 model: The network m2 contains m1 as a subnetwork providing emotional embeddings. The parts of the network tuned are painted with green.*

.

the network. These embeddings are concatenated with then SV embeddings, in the final layers of the fusion network. While training for the speaker verification task, different learning rates are used, the best of which is kept as the **emotion-aware** $t3$ model. This model's architecture is visualized on 3.6.

Our speaker verification model has very similar architecture to the SER's model. This is done consciously, so as to facilitate the transfer learning process. We train each model for 1000 epochs with early stopping enabled. The training procedure is achieved exploiting the VoxCeleb 1 dataset. We feed our model with pairs of multiple speaker utterances and compute the GE2E loss accordingly.

3.2.4 Evaluation Sets

A basic prerequisite for us to be able successfully conduct the experiments on Chapter 4, was the definition of our evaluation sets. These were in other words the test sets, which we utilized. For each one of our experiments, we created a different test set and then checked the performance of each model separately.

All the sets derive from the RAVDESS dataset. A set actually consists of a number of carefully selected tuples. Each tuple contains an enrollment utterance, a verification utterance and a label. The label indicates whether the two utterances belong to the same person or not.

In order to strictly define, our experimental setup, we should first explain the RAVDESS dataset's structure. First of all we only made use of the audio modality and the speech vocal channel. We denote the *emotions* as $e \in E = \{neutral, calm, happy, sad, angry, fearful, disgust, surprised\}$. In order to exclude *neutral* from some experiments, we as denote $e^{\dagger} \in E^{\dagger} = \{E \setminus \{neutral\}\}$. We denote the emotional *intensity* as $i \in I = \{0, 1\}$ for "*weak*" and "*strong*" respectively. At this point, we should note that there is no strong intensity

for the *neutral* emotion. Then, we denote the *different statements* as $st \in ST = \{1, 2\}$, the *repetition* as $r \in R = \{1, 2\}$ and the *actor* as $a \in A = \{a \in \mathbb{N}; 1 \leq a \leq 24\}$. Finally, we denote as X the *samples* which can be uniquely identified, using the different properties described above as indexes and y the *labels*, where $y \in Y = \{0, 1\}$. We list our different evaluation sets below:

- S_1: In this evaluation file, we create tuples, where the enrollment is neutral, whereas the verification utterance contains emotion. For each possible combination, we add one pair for same-speaker tuple and one pair for different-speaker tuple. For the experiment's purpose, we further separate our set, based on the emotional intensities, into two subsets. These are S_{1_weak} and S_{1_strong}:

$$S_{1\_weak} = \{(1, X(a, e, st, i = 0, r), X(a, neutral, st, i = 0, r)))\} \cup$$

$$\cup\{(0, X(a, e^{\dagger}, st, i = 0, r), X(a' \neq a, neutral, st, i = 0, r)))\}$$

$$, \forall a \forall st \forall r \forall e^{\dagger}$$

  and

$$S_{1\_strong} = \{(1, X(a, e^{\dagger}, st, i = 1, r), X(a, neutral, st, i = 0, r)))\} \cup$$

$$\cup\{(0, X(a, e^{\dagger}, st, i = 1, r), X(a' \neq a, neutral, st, i = 0, r)))\}$$

$$, \forall a \forall st \forall r \forall e^{\dagger}$$

- S_2: In this evaluation file, we create tuples where the enrollment and the verification contain emotion. For each possible combination, we add one pair for same-speaker tuple and one pair for different-speaker tuple. We intensively do not include same-emotion pairs. For the experiment purpose, we separate emotional intensities to two subsets. No duplicates were included. That is S_{2_weak} and S_{2_strong}

$$S_{2\_weak} = \{(1, X(a, e, st, i = 0, r), X(a, e' \neq e, st, i = 0, r)))\} \cup$$

$$\cup\{(0, X(a, e, st, i = 0, r), X(a' \neq a, e' \neq e, st, i = 0, r)))\}$$

$$, \forall a \forall st \forall r \forall e$$

  and

$$S_{2\_strong} = \{(1, X(a, e, st, i = 1, r), X(a, e' \neq e, st, i = 0, r)))\} \cup$$

$$\cup\{(0, X(a, e, st, i = 1, r), X(a' \neq a, e' \neq e, st, i = 0, r)))\}$$

$$, \forall a \forall st \forall r \forall e$$

- S_3: In this evaluation file, we create tuples where the enrollment is neutral, whereas the verification utterance contains emotion. For each possible combination, we add

one pair for same-speaker tuple and one pair for different-speaker tuple. We intensively do not include same-emotion pairs. For the experiment's purpose, we separate the set to 7 subsets S_{3_e}, each one corresponding to a specific emotion, except neutral. No duplicates were included. Finally, each emotion set S_{3_e} is further split into weak and strong emotional intensity. Therefore each set $S_{3\_emotion\_intensity}$ derives from the following formula:

$$S_{3\_e\_weak} = \{(1, X(a, e^{\dagger}, st, i = 0, r), X(a, neutral, st, i = 0, r)))\} \cup$$
$$\cup\{(0, X(a, e^{\dagger}, st, i = 0, r), X(a' \neq a, neutral, st, i = 0, r)))\}$$
$$, \forall a \forall st \forall r \forall e^{\dagger}$$

and

$$S_{3\_e\_strong} = \{(1, X(a, e^{\dagger}, st, i = 1, r), X(a, neutral, st, i = 0, r)))\} \cup$$
$$\cup\{(0, X(a, e^{\dagger}, st, i = 1, r), X(a' \neq a, neutral, st, i = 0, r)))\}$$
$$, \forall a \forall st \forall r \forall e^{\dagger}$$

- S_4: For the needs of the 4th experiment we construct two evaluation files (sets) each one representing a different cause. $S_{4\_ignorance}$ stands for an evaluation file where an neutral enrollment is followed by an emotional verification phase. That is similar to S_1, except here we are combining the two different intensities. $S_{4\_knowledge}$ stands for a verification file with an emotion, if it has preceded an enrollment with the same emotion. Trivial pairs containing the same samples are rejected. Therefore our sets derive from the following formula:

$$S_{4\_e\_ignorance} = \{(1, X(a, e^{\dagger}, st, i, r), X(a, neutral, st, i, r)))\} \cup$$
$$\cup\{(0, X(a, e^{\dagger}, st, i = 0, r), X(a' \neq a, neutral, st, i = 0, r)))\}$$
$$, \forall a \forall st \forall r \forall e^{\dagger} \forall i$$

$$S_{4\_e\_knowledge} = \{(1, X(a, e^{\dagger}, st, i, r), X(a, e^{\dagger}, st', i', r')))\} \cup$$
$$\cup\{(0, X(a, e^{\dagger}, st, i = 0, r), X(a' \neq a, e^{\dagger}, st', i', r')))\}$$
$$, \forall a \forall st \forall r \forall e^{\dagger} \forall i \ \& \ (st, i, r) \neq (st', i', r')$$

Κεφάλαιο 4

Experiments

In this chapter, we present our experiments. In order to have a fair comparison, we first create and train a model $t0$ on the speaker verification task. This model is consider "emotion unaware" and is used in each experiment as baseline. Then we train our emotion-aware models $t1, t2$ and $t3$. Afterwards each model is evaluated on each one of our experiments.

4.1 Datasets

We take advantage of the following datasets:

1. IEMOCAP: This dataset was used for the training of the emotional model. Any emotional aware model is pretrained on this dataset. More details are included in the 2.4.4.

2. VoxCeleb: This dataset was used for training our models into the speaker verification task. VoxCeleb's default evaluation file **is used as our baseline** for the SV task. More details for the dataset are included in the 2.5.3.

3. RAVDESS: This dataset is used for the emotional experiments setup. It provides multiple utterances from different users, with different emotions. Most of all, it provides different emotional intensities, classifying emotions as "weak" or "strong". For each one of the following experiments, we utilize our predesigned evaluation sets as described in 3.2.4. More details for the dataset can be found in the 2.4.4.

4.2 Emotion Driven Speaker Verification

4.2.1 Baseline SV Evaluation

In this subsection, we present the results that we derived while training our models directly on the speaker verification task. The point of this experiment is to ascertain whether emotional knowledge can improve a speaker verification model. The results are presented on the 4.1.

We observe that both models $t1$ and $t3$ outperform our emotion unaware model $t0$. More specifically we capture a 7.2% relative improvement in fusion model and a 17% in

model	EER (%)	s^2	statistical significance
$t0$	19.7	0.74	0.04
$t1$	**16.35**	0.43	0.02
$t2$	20.66	0.79	0.03
$t3$	**18.28**	0.51	0.04

Πίνακας 4.1: *The results on the speaker verification task, on VoxCeleb's evaluation set.*

our fully finetuned model relatively to $t0$.

4.2.2 The effect of emotional content on SV task

The point of this experiment is to ascertain how does emotional content affect speaker verification. In more detail, we construct tuples of **neutral versus emotional** content. We separate emotional content into two classes: "*weak*" and "*strong*". Each one indicates the emotional intensity of the emotional verification utterance that we feed our models with. On the other hand, enrollment utterance remains neutral. These evaluation sets are well described in 3.2.4 as S_{1_weak} and S_{1_strong}. A real world a scenario, would be someone enrolling a service being emotionally neutral and evaluate being angry.

Exp	model	VoxCeleb eval.	RAVDESS weak emotion	RAVDESS strong emotion
1.1	$t0$	19.7	**16.37**	30.65
1.2	$t1$	**16.35**	16.74	**27.23**
1.3	$t2$	20.66	17.51	**30.51**
1.4	$t3$	**18.28**	19.49	**27.53**

Πίνακας 4.2: *The effect of emotion on speaker verification task, in the case of a neutral enrollment is followed by an emotional verification utterance.*

Observing Table 4.2, we notice the effect of emotion on our models. Despite weak emotion performing relatively well, we capture a significant EER increase with the presence of strong emotional intensity. Taking a closer look, EER on weak emotion seems to be near the baseline (VoxCeleb's evaluation) for most of our models. This may suggest that our training dataset for speaker verification task, contains some emotional information. On contrast, in the case of strong emotion, we capture a massive relative increase 55.5% on EER for model $t0$.

Examining our emotion aware models $t1, t2$ and $t3$, we got some interesting remarks. All emotion aware models seem to perform better than $t0$. Most significantly, our model $t1$, achieves the lowest EER, both on baseline and strong emotional intensity. At the same time, on weak emotion its EER is very close to that of $t0$, indicating a general emotional robustness. Model $t2$ shows an improvement from the results of the baseline testing but performs worse than the other models in general. Finally our fusion model $t3$ performs relatively well especially in strong intensities but worse in weak.

It is interesting to notice, that lower baseline EER, does not come up with lower EER on emotional context. If that was the case, we would expect models that performed better on VoxCeleb's evaluation set (baseline) to perform relatively well on different sets (emotional

experiments). On the contrast, we capture, for example, that even though model $t2$ was much worse on baseline than $t0$, it actually outperformed it on strong emotional intensity.

From the results above, we can deduce that strong emotional information can drop a system's performance drastically, degrading its overall performance. We can safely reject the hypothesis that greater EER derives from domain mismatch, as on weak emotion all models perform relatively well. On the contrast, strong emotion seems to affect all of our models. Nevertheless, it is clear, that our emotion aware models can handle emotion better that $t0$. Therefore we can suggest that emotional information during the training procedure has affected the overall performance in a positive way.

Another interesting aspect, of emotional speech affecting speaker verification task, would be to examine how do different emotional states both on enrollment and verification affect speaker verification. For this purpose, we construct two sets of **emotional versus emotional** content, as described in 3.2.4. That means, that both enrollment and verification utterances will be "emotionally-injected". These sets are further split to *weak* and *strong* emotional intensity. Nevertheless, we set the following constraint; no utterance shall exist both in enrollment and verification, at the same time with the same emotion. Providing that, we anticipated a greater EER increase in respect with the previous experiment, as the effect of emotional content has already been identified. This experiment interprets how do emotional expressions of speakers alternate their ability to be distinguished in a SV system. This scenario is aspired from real world cases, as it is not far from reality. It is difficult for humans to enroll or verify being emotionally neutral due to the fact that it is difficult to uniquely define emotions. Different people come up with different definitions and understandings of *neutral* thus affecting the procedure.

Exp	model	VoxCeleb eval.	RAVDESS weak emotion	RAVDESS strong emotion
2.1	$t0$	19.7	21.88	32.64
2.2	$t1$	**16.35**	**21.38**	**31.37**
2.3	$t2$	20.66	**20.78**	**31.13**
2.4	$t3$	18.28	23.29	**30.78**

Πίναχας 4.3: *The effect of emotion on SV task, when different emotions occur both during enrollment and verification phase.*

After observing Table 4.3, it becomes obvious, that both emotional intensities perform much worse than the VoxCeleb's evaluation set. First of all, we capture a relative increase around 10% on weak intensity relatively to the baseline. At the same time, emotionally strong utterances, increase EER around 40%. These results indicate, that **different emotions affect significantly the speaker verification task, even at weak intensity**.

Comparing Table 4.2 and Table 4.3, we capture that weak emotional intensities increased EER around 25%. At the same time, strong emotional intensities increased EER around $6-15\%$. It is important to notice, that our emotional models outperform $t0$ again, especially $t1$, which scores the lowest EER on baseline, weak and strong.

Our results above indicate that emotional information on both utterances (enrollment and verification) can be a catalyst for very poor results. We can safely say, that all of our models are strongly affected by emotion, this time even on weak. We notice that all of our

emotional models once again outperform our emotion unaware model $t0$. On addition, we can recognise that our fine tune model $t1$ outperforms $t0$ once again. Obviously, emotional content implicates with the capability of a SV system to robustly discriminate speakers. This arises questions about the characteristics that a SV neural network captures and their sensibility to human expressions.

4.2.3 The effect of each emotion on SV task

After conducting the experiments on 4.2.2, the question arises as to how does each emotion affect the speaker verification task individually? Are there any specific emotions that magnify the effect? Is there any correlation between the errors from the "emotional brain" (SER model) and these that occur under the presence of emotional utterances (SV model)? In this subsection, we try to answer these questions.

In order to set up this experiment, we had to construct an evaluation file, where each emotion would be examined independently, as explained in 3.2.4. We used the RAVDESS dataset and create two custom test sets, in which, each emotion in the verification utterance is tested against a neutral enrollment. These tests are further separated by emotional intensity, to end up with a table of 7 emotions $\times$ 2 intensities. Then, we evaluate our emotion unaware model $t0$. Finally, we examine our emotion aware models performance and check their relative performance based on model $t0$ evaluation.

Exp	emotion	RAVDESS weak	RAVDESS strong
3.1	calm	9.38	15.62
3.2	happy	15.1	32.29
3.3	sad	12.5	31.77
3.4	angry	17.71	39.58
3.5	fearful	18.23	38.54
3.6	disgust	24.48	20.31
3.7	surprised	17.71	27.6

Πίνακας 4.4: *The effect of different emotions on model t0.*

First of all, we can see that *calm* performs better than all the other emotions. This could be explained with the similarity it has with *neutral*. On the other hand, all the other emotions increase EER drastically. *Angry, fearful and happy and sad* have the higher EER on strong emotion and *disgust* on weak. As we can see, happy angry sad and fearful are among the most diffucult emotions, significantly decreasing EER on model $t0$.

On tables 4.5, 4.6 and 4.7 we evaluate on the same task, our emotion aware models, $t1$, $t2$ and $t3$ respectively. Each table contains two columns with the relative performance to $t0$.

We capture a relative improvement in almost all pretrained emotions (*happy, sad, angry*). We suggest that, the emotions which where present during the training of the "emotional brain", resulted in a more emotionally robust model after fine tuning. We should note that *calm, fearful, disgust and surprised* were emotions not taught to our model. Improvement in these may suggest that our model tracks generally emotional content in a more efficient way. We also note, that utterances with the unseen emotions of

Exp	emotion	RAVDESS weak	relative to t0 (%)	RAVDESS strong	relative to t0 (%)
3.1	calm	10.42	−11.09	17.71	−13.38
3.2	happy	14.06	6.89	27.08	16.14
3.3	sad	11.46	8.32	30.73	3.27
3.4	angry	20.83	−17.62	34.9	11.82
3.5	fearful	18.23	0.0	32.81	14.87
3.6	disgust	25.52	−4.25	23.96	−17.97
3.7	surprised	16.15	8.81	18.23	33.95

Πίνακας 4.5: *The effect of different emotions on model t1.*

disgust and calm perform much worse than $t0$.

Exp	emotion	RAVDESS weak	relative to t0 (%)	RAVDESS strong	relative to t0 (%)
3.1	calm	14.06	−49.89	16.67	−6.72
3.2	happy	17.71	−17.28	32.81	−1.61
3.3	sad	13.02	−4.16	28.12	11.49
3.4	angry	20.83	−17.62	36.98	6.57
3.5	fearful	17.71	2.85	40.62	−5.4
3.6	disgust	27.6	−12.75	29.17	−43.62
3.7	surprised	21.88	−23.55	26.04	5.65

Πίνακας 4.6: *The effect of different emotions on model t2.*

In the table 4.6, we observe that our model $t2$ performs much worse than our baseline $t0$ especially on weak intensity emotional utterances. On the other hand, our model seems to perform better on strong emotional content on *sad, angry* and *surprised*. Unfortunately, at the same time, the emotions of *calm, fearful* and *disgust* undermine the model's total performance. We should note that EER on *disgust* and *calm* worsens significantly, relatively to $t0$.

Exp	emotion	RAVDESS weak	relative to t0 (%)	RAVDESS strong	relative to t0 (%)
3.1	calm	12.5	−33.26	19.27	−23.37
3.2	happy	16.67	−10.4	31.25	3.22
3.3	sad	11.46	8.32	25.0	21.31
3.4	angry	22.92	−29.42	38.54	2.63
3.5	fearful	21.35	−17.11	29.69	22.96
3.6	disgust	24.48	0.0	25.52	−25.65
3.7	surprised	21.88	−23.55	25.0	9.42

Πίνακας 4.7: *The effect of different emotions on model t3.*

Observing the Table 4.7 we can capture similarities with the Table 4.5. *Disgust* and *calm* seem to confuse the model $t3$, just as $t1$; this time on a greater degree though. We capture that weak emotions perform much worse than $t0$, while almost all have not negligible relative increases. On the other hand, $t3$ performs much better than $t0$ on strong emotional content, with outstanding improvement on most emotions. Unfortunately, *calm* and *disgust* drop our model's performance, making it unstable and degrading its overall performance.

In summary, we should note that *angry* and *fearful* are the most difficult emotions affecting drastically the overall speaker verification procedure. That is by shouting up EER's absolute value near to 40%. We should note that our model $t1$ manages to perform relatively well even on these emotions. Additionally, on weak intensity, the worse EER is spotted on *disgust* with EER's absolute value to fluctuate around 25%. Last but not least, we should mention that *calm* emotion does not perform very well in all our emotion aware models.

4.2.4 The effect of same-emotion utterances both on enrollment and verification

In this subsection, we explore the relations between emotion in the enrollment and emotion in the verification phase. More specifically, given an emotional utterance during the verification phase, we try to understand, whether a SV system performs better, having a same-emotion enrollment. A real world scenario of such an extension, would require the system to correctly classify the emotion of the utterance during evaluation and then compare it with the corresponding same-emotion enrollment. This leads to a system that registers its users with different emotional states on enrollment stage.

In order to set up this experiment, we create an evaluation file with some specific properties. For each emotion, we construct tuples that fall onto two categories. *Emotional ignorance* and *emotional knowledge*. We consider each tuple of neutral enrollment versus an emotional verification as emotional ignorance in our system. Nevertheless, we consider each tuple of the same emotion both on enrollment and verification phase to be an emotional knowledge. This segregation is only considered during our model's testing and has no relation with our emotional training and our emotion aware models.

In the following tables, we list the performance of our models for each emotion and each one of *emotional ignorance* and *emotional knowledge*. On the last column we capture the relative performance increase, for each emotion. On the final row, we present the mean values for each column.

Exp	emotion	ignorance EER (%)	knowledge EER (%)	relative improvement (%)
4.1	calm	9.9	7.66	22.63
4.2	happy	23.44	17.41	25.73
4.3	sad	20.83	22.92	−10.03
4.4	angry	33.33	20.46	38.61
4.5	fearful	28.39	24.93	12.19
4.6	disgust	19.01	14.58	23.3
4.7	surprised	17.71	11.38	35.74
4.8	average	21.80	17.05	21.17

Πίναχας 4.8: *Emotional ignorance versus emotional knowledge for model t0*

For our model $t0$, we capture that all emotions except sadness, perform better than emotional ignorance while having emotional knowledge. We spot an improvement of $\sim 12-38\%$ with an average of $\sim 21.17\%$. This result can also be interpreted as following; a system suffering from emotional ignorance during testing can perform up to $\sim 21\%$ worse.

Exp	emotion	ignorance EER (%)	knowledge EER (%)	relative improvement (%)
4.1	calm	14.32	6.85	52.16
4.2	happy	20.83	19.49	6.43
4.3	sad	20.05	21.43	−6.88
4.4	angry	29.43	21.06	28.44
4.5	fearful	26.56	25.0	5.87
4.6	disgust	19.27	16.0	16.97
4.7	surprised	13.8	12.28	11.01
4.8	average	20.61	16.65	17.34

Πίναχας 4.9: *Emotional ignorance versus emotional knowledge for model t1*

After examining our model $t1$, we capture similar results. First of all, we capture a degrade on *sadness* error, therefore being less sensitive to emotional ignorance. Overall all emotions get a $\sim 5-52\%$ increase, with an average of 17.34%. This result, indicates a lower emotional dependence of our emotion aware model $t1$. At the same time the average scores both on emotional ignorance and knowledge outperform those of model $t0$. Most importantly, we capture a strong increase on *calm* emotion. This indicates that poor performance on emotions spotted on 4.5 seems to fix straight forward.

Exp	emotion	ignorance EER (%)	knowledge EER (%)	relative improvement (%)
4.1	calm	11.72	6.7	42.83
4.2	happy	23.44	16.82	28.24
4.3	sad	18.75	21.35	−13.87
4.4	angry	31.25	23.51	24.77
4.5	fearful	27.6	23.88	13.48
4.6	disgust	20.83	15.33	26.4
4.7	surprised	19.79	9.82	50.38
4.8	average	21.91	16.77	24.6

Πίναχας 4.10: *Emotional ignorance versus emotional knowledge for model t2*

Observing Table 4.10, we capture a huge improvement through emotional knowledge. In more detail, there is a relative improvement of $\sim 13-50\%$ for all the emotions, except sadness, with a mean of 24.6%. We should point out, that *calm* emotion which performed poor on 4.6 improves around 43%. Finally, the average emotional knowledge's EER, outperforms $t0$ model.

After reviewing Table 4.11 we can detect many similarities with the previous tables. That is *sadness* being the only emotion on which emotional knowledge degrades performance. In other respects, all emotions face a relative increase $\sim 6-44\%$ with an average of 22.49%. Despite the general improvement in performance, on the downside, $t0$ performed better than $t3$ both on emotional ignorance and emotional knowledge.

In Table 4.12, we can see how model $t1$ outperforms once again model $t0$. First of all, we capture a $\sim 5.5\%$ improvement on emotional ignorance. That means that model $t1$ is more likely to be robust in emotional content when no emotional clue is given during evaluation. At the same time, when an emotional enrollment is provided, model $t1$ once

Exp	emotion	ignorance EER (%)	knowledge EER (%)	relative improvement (%)
4.1	calm	18.75	10.34	44.85
4.2	happy	24.22	18.15	25.06
4.3	sad	21.35	24.48	−14.66
4.4	angry	33.33	25.0	24.99
4.5	fearful	32.81	24.4	25.63
4.6	disgust	22.66	21.13	6.75
4.7	surprised	26.82	14.81	44.78
4.8	average	25.71	19.76	22.49

Πίνακας 4.11: *Emotional ignorance versus emotional knowledge for model t3*

model	relat. improvement ignorance (%)	relat. improvement knowledge (%)
t1	**5.47**	**2.32**
t2	−0.5	1.62
t3	−17.91	−15.9

Πίνακας 4.12: *Relative performance of models t1, t2 and t3 to t0 on emotional ignorance and emotional knowledge*

again outperforms model $t0$. At last, model $t2$ performs relatively well, while $t3$ does not.

4.3 Discussion

In this section, we summarize the main contribution of this thesis, as after conducting the experiments above, we came to some interesting conclusions:

- We showed that a pre-trained SER model, fine tuned on SV can outperform a model trained directly on SV. More specifically, we observed that two out of three emotion-aware architectures outperform the emotion-unaware model on VoxCeleb's evaluation set. We capture a 7% and a 17% increase in performance, relatively to our emotion-unaware model.

- We demonstrated that emotional content on speaker verification can drastically degrade a system's performance, depending on the emotional intensity of the speaker. This results on prone models easily effected by user's emotional state. Furthermore different emotions on the enrollment and the verification phase can magnify the effect ending in very poor results. This indicates that architectures trained directly on speaker verification, have no intuition on the ways that emotion modulates speech.

- We showed that *anger* and *fear* are the emotions affecting the most the speaker discrimination procedure. After carefully examining the emotions provided by the RAVDESS dataset, our experiments suggest that these two have the greater impact on the procedure. Both *anger* and *fear* drastically increase EER, affecting heavily the discrimination procedure. Emotion unaware models reach a poor performance nearly 40% EER.

- We demonstrated that by applying traditional transfer learning methods from SER

to SV, we can effectively train emotion-aware models. We observed that these models can reduce the effect of emotional content on speech and improve our overall results. In more detail, our architecture $t1$ outperforms one traditional emotion-unaware model, such as $t0$, both on weak and strong emotional intensity.

- We showed that *emotional knowledge* during evaluation phase, can improve a system's performance nearly $17 - 24\%$, even on an emotion unaware model. This is crucial, considering that improvement on the SER can lead to improvement on the SV task. This observation could have applications in real world systems, as we could provide specific enrollment utterances, depending on the output of an emotional classifier.

Κεφάλαιο 5

Conclusion

5.1 Summary

In this thesis, we studied the effect of emotion on speaker verification. We experimented with different techniques, in order to better understand how different emotions implicate with deep learning models.

First of all, we examined how emotional content affects the speaker verification task and showed that emotional speech can significantly degrade an SV system's performance, depending on the emotional intensity. We found that strong intensity overall can make a speaker verification model to perform very poorly. We examined how neutral emotion affects the procedure and how does equal error rate change if we add emotional content on the enrollment utterance besides the verification. The results indicate that different emotion-pairs on these two utterances result in the worst possible performance making our baseline models, practically incapable of correctly recognising speakers.

Then, we explored how does each emotion individually affect the procedure, and identified these which implicate the most. We found that *anger* and *fear* are the emotions that are most likely to make our models more prone to false user rejections or false impostor acceptances. More specifically, we noticed that equal error rate's absolute value reached 40% when these emotions where present. *Anger* is an emotion that has an unexpected behaviour in many cases.

Furthermore, we inspected, whether a SV system can be improved by inserting an emotional utterance in the evaluation phase. Our suggestion showed a great improvement on system's performance, even when the models had not seen emotional information during their training. We capture a relative increase about 20% for all of our models and for almost all the emotions. Our results point out the need of taking into account the effect of emotion on SV and design more complex architectures, where a SER classifier could interact with the enrollment and verification utterance. In this way, SV process could become more robust.

Last but not least, we came up with three architectures that aimed at reducing the emotional effect on speaker verification, each one in a different manner. We showed that each one of them excels our emotion-unaware model on some tasks. Most significantly, our fine tune architecture, remarkably outperformed our baseline emotion-unaware model on all the tasks that we examined. Therefore, our experiments demonstrate that by applying

traditional transfer learning methods, we can efficiently transfer emotional knowledge to the speaker verification task, improving the models' robustness, even when emotional speech is present.

As one easily understands, emotion has a crucial role and implicates with a speaker verification system in a great degree. In this thesis, we identified some of these relations and tried to overcome them by transferring emotional knowledge.

5.2 Future Work

This thesis could have many interesting future extensions. In this subsection we try to address some of those.

First of all, one direction would be to study the effect of speaker verification on speech emotion recognition and how speaker discriminative features relate with speaker emotional expressions. More specifically, as these two tasks are correlated, we could try inserting speaker specific knowledge, into the emotion recognition procedure. As emotions are not unambiguously settled, it would be interesting to study whether speaker-specific characteristics improve the overall performance.

Another interesting aspect to study, would be the implementation of a custom loss function. We know by this time, that training a model with common speaker verification losses such as GE2E loss, does not take different modalities such as emotion into consideration. This is a major flaw because essentially we completely ignore emotional information. As a result we blindly focus on speaker discrimination without identifying the effect of emotional modulation on speech. As shown in recent works [15], [16], loss functions specifically customized for the task, can significantly improve the overall performance. Attempts have been made even on the SER field [17] for splitting different modalities on the presence of emotional speech. We suggest that we could examine how to integrate emotional information into a custom loss function and inspect whether performance improves both on pure speaker verification task and to emotion-injected speaker verification.

Bibliography

[1] Mu Li Alex J. Smola Aston Zhang, Zack C. Lipton. *Dive Into Deep Learning*. 2018.

[2] *Max Pooling*. https://commons.wikimedia.org/wiki/File:Max_pooling.png. Accessed: 2021-05-22.

[3] *Feed Forward Neural Networks*. https://automaticaddison.com/artificial-feedforward-neural-network-with-backpropagation-from-scratch/. Accessed: 2021-05-22.

[4] *Tanh*. https://commons.wikimedia.org/wiki/File:Hyperbolic_Tangent.svg. Accessed: 2021-05-23.

[5] *Rectified linear activation function for deep learning neural networks*. https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/. Accessed: 2021-05-23.

[6] *Activation Functions; So why do we need Activation functions in our neural networks?* https://bit.ly/2SPU8yL. Accessed: 2021-05-23.

[7] *Gradient Descend*. http://rasbt.github.io/mlxtend/user_guide/general_concepts/gradient-optimization_files/ball.png. Accessed: 2021-05-24.

[8] *Basic emotions and how they affect us*. https://themindsjournal.com/basic-emotions-and-how-they-affect-us/. Accessed: 2021-05-25.

[9] Radoslaw Nielek, Miroslaw Ciastek και Wiesław Kopeć. 2017.

[10] Sven Buechel και Udo Hahn. *Readers vs. Writers vs. Texts: Coping with Different Perspectives of Text Understanding in Emotion Annotation*. 2017.

[11] *Equal Error Rate (EER)*. https://i1.wp.com/wentzwu.com/wp-content/uploads/2019/05/hqufd.jpg?resize=584%2C399&ssl=1. Accessed: 2021-05-26.

[12] Jeremy Karnowski. *Detection Error Tradeoff (DET)*. https://jeremykarnowski.wordpress.com/2015/08/07/detection-error-tradeoff-det-curves/. Accessed: 2021-07-11.

[13] *Gaussian Mixture Models Explained ;From intuition to implementation*. https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95?gi=e5e472af0742. Accessed: 2021-05-28.

[14] *Universal Background Models*. https://slideplayer.com/slide/3991874/. Accessed: 2021-07-11.

[15] Amina Asif, Muhammad Dawood και Fayyazul Amir Afsar Minhas. *A Generalized Meta-loss function for regression and classification using privileged information*, 2019.

[16] Jiwei Xu, Xinggang Wang, Bin Feng και Wenyu Liu. *Deep multi-metric learning for text-independent speaker verification*. *Neurocomputing*, 410:394–400, 2020.

[17] Srinivas Parthasarathy, Viktor Rozgic, Ming Sun και Chao Wang. *Improving Emotion Classification through Variational Inference of Latent Variables*. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 7410–7414, 2019.

[18] Oluwafemi Osho Shafi'i Muhammad Abdulhamid, Maryam Shuaib. *Comparative Analysis of Classification Algorithms for Email Spam Detection*. *I. J. Computer Network and Information Security*, 2018.

[19] Hieu Pham, Zihang Dai, Qizhe Xie, Minh Thang Luong και Quoc V. Le. *Meta Pseudo Labels*, 2020.

[20] Mr. Krishna charlapally Dr. P. K. Sahoo. *Stock Price Prediction Using Regression Analysis*. *International Journal of Scientific Engineering Research*, 6(3), 2015.

[21] G. Cybenko. *Approximation by superpositions of a sigmoidal function*. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989.

[22] C.C. Lee A. Kazemzadeh E. Mower S. Kim J.N. Chang S. Lee C. Busso, M. Bulut και S.S. Narayanan. *IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation*. *Journal of Language Resources and Evaluation*, (4):335–359, 2008.

[23] Steven R. Livingstone και Frank A. Russo. *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English*. *PLOS ONE*, 13(5):1–35, 2018.

[24] Quatieri Thomas F. Reynolds, Douglas A. και Robert B. Dunn. *Speaker Verification Using Adapted Gaussian Mixture Models*. *Digital Signal Processing*, 2:19–41, 2000.

[25] Kihyuk Sohn. *Improved Deep Metric Learning with Multi-class N-pair Loss Objective*. *Advances in Neural Information Processing Systems*D. Lee, M. Sugiyama, U. Luxburg, I. Guyon και R. Garnett, επιμελητές. Curran Associates, Inc.

[26] Arsha Nagrani, Joon Son Chung και Andrew Zisserman. *VoxCeleb: A Large-Scale Speaker Identification Dataset*. *Interspeech 2017*, 2017.

[27] Joon Son Chung, Arsha Nagrani και Andrew Zisserman. *VoxCeleb2: Deep Speaker Recognition*. *Interspeech 2018*, 2018.

[28] Reda Dehak Pierre Dumouchel Najim Dehak, Patrick Kenny και Pierre Ouellet. *Front-end factor analysis for speaker verification. IEEE Trans. on Audio, Speech, and Language Processing*, 19:788–798, 2011.

[29] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey και Sanjeev Khudanpur. *X-Vectors: Robust DNN Embeddings for Speaker Recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[30] Raghavendra Pappagari, Tianzi Wang, Jesus Villalba, Nanxin Chen και Najim Dehak. *x-vectors meet emotions: A study on dependencies between emotion and speaker recognition*, 2020.

[31] Li Wan, Quan Wang, Alan Papir και Ignacio Lopez Moreno. *Generalized End-to-End Loss for Speaker Verification*, 2020.