ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

«Αναγνώριση των Συναισθημάτων από την Ομιλία»

Εξαμηνιαία Εργασία

στο μάθημα «Μετρήσεις και Έλεγχοι στη Βιοϊατρική Τεχνολογία»

των φοιτητών

Στεφανάκης Γεώργιος, Α.Μ.: 03118436

Παπαζαφειρόπουλος Αναστάσιος, Α.Μ.: 03118079

Μπότσας Χαράλαμπος Σπυρίδων, Α.Μ.: 03118121

Καρανίκας Ραφαήλ, Α.Μ.: 03118826

Υπεύθυνη Εργασίας: Ρανώ Μαντά

Διδάσκοντες: Δ. Κουτσούρης, Ο. Πετροπούλου

Αθήνα, Φεβρουάριος 2022

Περίληψη

Η θεωρία των συναισθημάτων σημειώνει μεγάλη ανάπτυξη τα τελευταία χρόνια και δεν θα ήταν εφικτό να απουσιάζει από την επιστήμη τους πληροφορικής, αφού η επικοινωνία ανθρώπου μηχανής βασίζεται σε ένα πολύ μεγάλο ποσοστό, αν όχι εξ ολοκλήρου, σε αυτή. Τα ανθρώπινα συναισθήματα αναγνωρίζονται, ερμηνεύονται και επεξεργάζονται από το πεδίο της συναισθηματικής υπολογιστικής που αναπτύσσει συστήματα και συσκευές με σκοπό την κατάλληλη ερμηνεία και επεξεργασία. Η αναγνώριση των συναισθημάτων γίνεται συνήθως με τρεις τρόπους, από τα φυσιολογικά σήματα μέσω αισθητήρων, από την ομιλία, και από τις εκφράσεις του προσώπου. Η ομιλία αποτελεί έναν από τους πιο συνήθεις τρόπους αλληλεπίδρασης μεταξύ ανθρώπου και μηχανής. Οπότε, η αναγνώριση συναισθημάτων μέσω ομιλίας (Speech Emotion Recognition, SER) αποσκοπεί να αναλύσει την συναισθηματική κατάσταση μέσω σημάτων ομιλίας κι έχει πρόσφατα προσελκύσει το ενδιαφέρον συνεχώς αυξανόμενου πλήθους ερευνητών. Ωστόσο, παραμένει ένας απαιτητικός τομέας για τον οποίο το πώς θα εξαχθούν με αποδοτικό τρόπο τα συναισθηματικά χαρακτηριστικά είναι ένα ανοιχτό ερώτημα. Οι ραγδαίες εξελίξεις στις τεχνικές ψηφιακής επεξεργασίας σήματος και στη μηχανική μάθηση, κυρίως με τα νευρωνικά δίκτυα έχουν πυροδοτήσει μια σειρά εξελίξεων ως προς τις μεθόδους έρευνας των προβλημάτων SER, καθώς ο τρόπος εξαγωγής χαρακτηριστικών από την ομιλία, άρα και τα αποτελέσματα των ερευνών βελτιώνονται συνεχώς. Σκοπός αυτής της εργασίας είναι να γίνει μια περιγραφή των κύριων προσεγγίσεων του συγκεκριμένου προβλήματος με χρήση τεχνικών ψηφιακής επεξεργασίας σήματος ή μηχανικής μάθησης στη διεθνή βιβλιογραφία, με παράλληλη αναφορά στις χρησιμοποιούμενες τεχνολογίες, και τελικά να γίνει μια συγκριτική αποτίμηση των αποτελεσμάτων με την πάροδο του χρόνου.

Λέξεις Κλειδιά: Αναγνώριση Συναισθημάτων, Ψηφιακή Επεξεργασία Σήματος, Μηχανική Μάθηση, Αναγνώριση Συναισθημάτων μέσω Ομιλίας, Νευρωνικά Δίκτυα, Βαθιά Μάθηση, Αλληλεπίδραση Ανθρώπου – Υπολογιστή

Πίνακας περιεχομένων

1. Εισαγωγή

Η ομιλία είναι ένα σύνθετο σήμα που περιέχει πληροφορίες σχετικά με το μήνυμα, τον ομιλητή, τη γλώσσα, το συναίσθημα κ.ο.κ.. Το σήμα της ομιλίας τροφοδοτείται σε συστήματα ομιλίας, τα περισσότερα εκ των οποίων επεξεργάζονται ηχογραφημένη φυσική ομιλία από στούντιο με αποτελεσματικό τρόπο. Ωστόσο, η απόδοσή τους είναι ανεπαρκής στην περίπτωση της ομιλίας με συναισθηματικό τόνο. Αυτό οφείλεται στη δυσκολία μοντελοποίησης και χαρακτηρισμού των συναισθημάτων που υπάρχουν στην ομιλία.

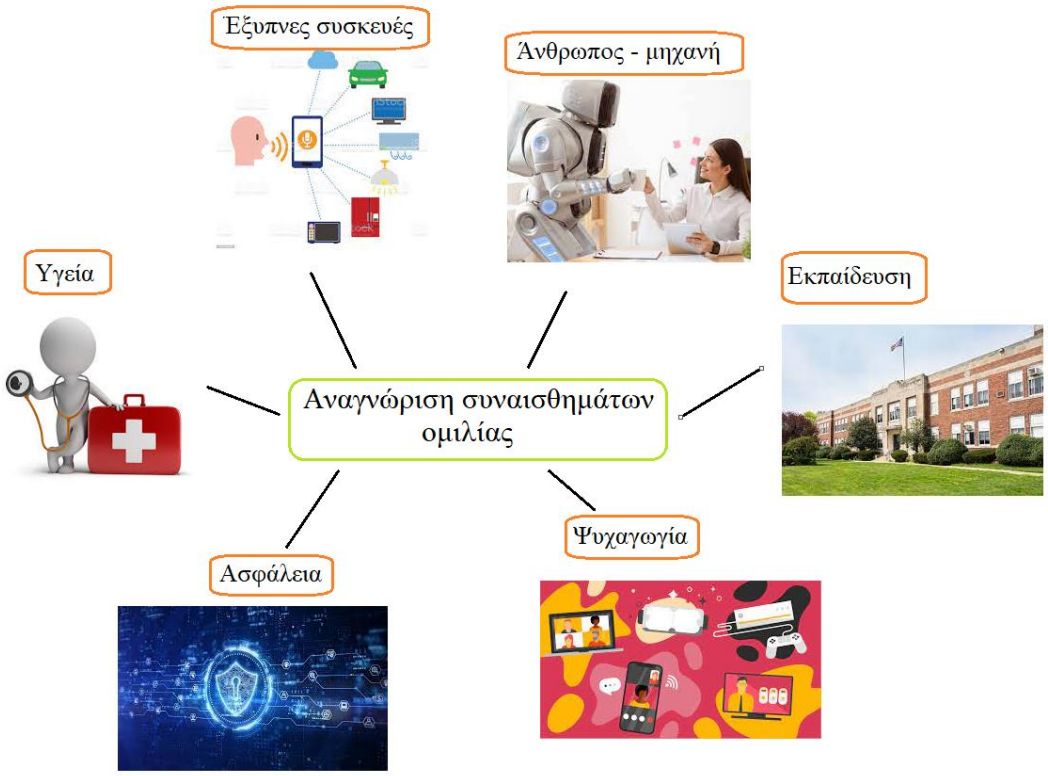
Η παρουσία συναισθημάτων κάνει την ομιλία πιο φυσική. Για την κατανόησή της έχουμε τόσο το λεκτικό νόημα, που προέρχεται από το σημασιολογικό περιεχόμενο των λέξεων στον λόγο, όσο και το μη λεκτικό, δηλαδή τον τρόπο ή τόνο με τον οποίο εκφέρονται οι λέξεις. Ο τρόπος εκφοράς, μάλιστα, μπορεί να αποτελεί παράμετρο διαφοροποίησης των γλωσσών. Μερικές γλώσσες είναι τονικές (λ.χ. κάποιες Ασιατικές), δηλαδή αξιοποιούν διαφορετικά τονικά ύψη για την διαφοροποίηση του νοήματος, όπως για παράδειγμα η Ιαπωνική. Άλλες, μη τονικές, όπως η Ελληνική, χρησιμοποιούν απλώς τόνους στις συλλαβές τους. Είναι επόμενο, λοιπόν, ο τρόπος έκφρασης των συναισθημάτων στον λόγο από γλώσσα σε γλώσσα να ποικίλει.

Εκτός από τις ιδιομορφίες που εμφανίζει η εκάστοτε γλώσσα σε μια συνομιλία, η μη λεκτική επικοινωνία φέρει σημαντικές πληροφορίες, όπως για παράδειγμα την πρόθεση του ομιλητή. Πέρα από το μήνυμα που μεταδίδεται μέσω του κειμένου, ο τρόπος με τον οποίο εκφωνούνται οι λέξεις μεταφέρει ουσιαστικές αλλά μη γλωσσικές πληροφορίες. Το ίδιο κειμενικό μήνυμα θα μπορούσε να έχει διαφορετική σημασία (νόημα), ενσωματώνοντας κατάλληλο ύφος. Ιδιαίτερα ο προφορικός λόγος μπορεί να έχει διάφορες ερμηνείες, ανάλογα με τον τρόπο που εκφέρεται. Για παράδειγμα, η λέξη «okay» στα αγγλικά, χρησιμοποιείται για να εκφράσει θαυμασμό, δυσπιστία, συγκατάθεση, αδιαφορία ή ισχυρισμό. Ένα αντίστοιχο παράδειγμα είναι αυτό της λέξης «μάλιστα» που επίσης έχει ποικίλες χρήσεις στα ελληνικά. Επομένως, η κατανόηση της σημασιολογίας από μόνη της δεν αρκεί για την ερμηνεία μιας έκφρασης.

Οι μη γλωσσικές πληροφορίες μπορούν να παρατηρηθούν οπτικά μέσω μορφασμών και εκφράσεων του προσώπου, ηχητικά μέσω επιτονισμού, επιφωνημάτων κ.α. στην περίπτωση του προφορικού λόγου, και σημείων στίξης στην περίπτωση γραπτού κειμένου. Η συζήτηση στην παρούσα ανάλυση περιορίζεται σε συναισθήματα ή εκφράσεις που σχετίζονται με την ομιλία. Βασικοί στόχοι της επεξεργασίας του συναισθηματικού λόγου είναι: (α) η κατανόηση των συναισθημάτων που εμφανίζονται στην ομιλία και (β) η σύνθεση των επιθυμητών συναισθημάτων στην ομιλία σύμφωνα με το επιδιωκόμενο μήνυμα. Από τη σκοπιά της μηχανής, η κατανόηση των συναισθημάτων της ομιλίας μπορεί να θεωρηθεί ως ταξινόμηση ή διάκριση των συναισθημάτων. Η σύνθεση των συναισθημάτων μπορεί να θεωρηθεί ότι ενσωματώνει ειδικές γνώσεις συναισθημάτων κατά τη σύνθεση ομιλίας [1].

Η αναγνώριση συναισθημάτων μέσω της ομιλίας έχει πλήθος σημαντικών εφαρμογών σε ποικίλους τομείς. Ενισχύει την ποιότητα της αλληλεπίδρασης του ανθρώπου με τη μηχανή, κάνοντας την τελευταία πιο φυσιολογική και φιλική για τον άνθρωπο. Επίσης, η αναγνώριση συναισθημάτων μπορεί να παίξει καίριο ρόλο σε ένα σύστημα οδήγησης αυτοκινήτου καθώς μπορεί να επεξεργαστεί τη ψυχική κατάσταση του οδηγού και να τον κρατήσει σε εγρήγορση κατά τη διάρκεια της οδήγησης αποτρέποντας πιθανά ατυχήματα. Μια εξίσου αξιόλογη εφαρμογή είναι στις συνομιλίες τηλεφωνικών

κέντρων, όπου μπορεί να χρησιμοποιηθεί για την ανάλυση της συμπεριφοράς στις τηλεφωνικές κλήσεις υπαλλήλων - πελατών και να συμβάλει στη βελτίωση της ποιότητας της υπηρεσίας του τηλεφωνικού κέντρου. Παρακάτω απεικονίζονται μερικοί τομείς που χρησιμοποιείται η αναγνώριση συναισθημάτων μέσω ομιλίας:



Εικόνα 1: Εφαρμογές της αναγνώρισης συναισθημάτων μέσω ομιλίας σε διάφορους τομείς

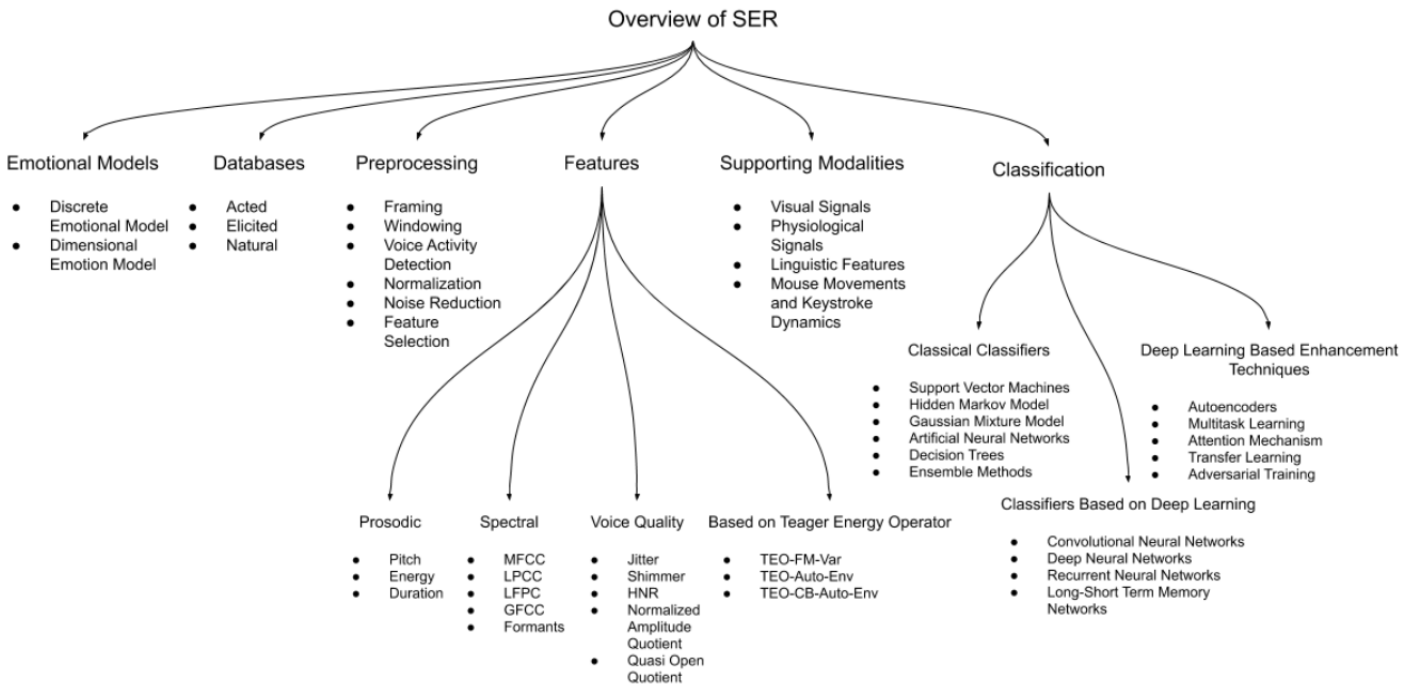
Εφαρμογές του SER εντοπίζουμε και σε τομείς που βρίσκονται ήδη στην αιχμή της τεχνολογίας όπως στον βιοϊατρικό τομέα και τον τομέα της ασφάλειας. Στον πρώτο, οι γιατροί μπορούν να χρησιμοποιούν συναισθηματικά περιεχόμενα της ομιλίας ενός ασθενούς ως ένα διαγνωστικό εργαλείο για διάφορες διαταραχές. Ενώ στον χώρο της ασφάλειας σημαντικό ρόλο μπορεί να παίξει η ανάλυση συναισθημάτων της τηλεφωνικής συνομιλίας μεταξύ εγκληματιών για την αρωγή της έρευνας του τμήματος διερεύνησης εγκλημάτων.

Επιπλέον εφαρμογές υπάρχουν στην επιστήμη της ρομποτικής και της αυτόματης μετάφρασης. Η συνομιλία με ρομποτικά κατοικίδια και ανθρωπόμορφους συνεργάτες θα ήταν πιο ρεαλιστική και ενδεχομένως πιο ευχάριστη, αν είναι σε θέση να κατανοήσουν και να εκφράσουν συναισθήματα όπως οι άνθρωποι (ψυχαγωγία, αυτοματισμοί, οικιακές χρήσεις κλπ). Ενώ, σε συστήματα αυτόματης μετάφρασης ομιλίας σε ομιλία (automatic speech to speech translation systems), όπου η ομιλία στη γλώσσα Χ μεταφράζεται σε άλλη γλώσσα Υ από το μηχάνημα, χρησιμοποιείται τόσο η αναγνώριση συναισθημάτων όσο και η σύνθεση. Τα συναισθήματα που υπάρχουν στην ομιλία πηγής (γλώσσα Χ) πρέπει να αναγνωρίζονται και τα ίδια συναισθήματα πρέπει να συντίθενται στην μεταφρασμένη ομιλία (γλώσσα Υ), καθώς η μεταφρασμένη ομιλία αναμένεται ως ένα βαθμό να αντιπροσωπεύει και τη συναισθηματική κατάσταση του αρχικού ομιλητή.

Η αναγνώριση συναισθημάτων από ηχητικά σήματα ομιλίας αποτελεί, λοιπόν, ένα πολύ σημαντικό επιστημονικό αντικείμενο τα τελευταία χρόνια. Εν γένει είναι ένα περίπλοκο πρόβλημα καθώς η κατανόηση συναισθημάτων είναι μία ιδιότροπη διαδικασία και διαφέρει από άνθρωπο σε άνθρωπο και από πολιτισμό σε πολιτισμό. Επιπλέον, υπάρχει ακόμα συζήτηση σχετικά με τον ορισμό βασικών κατηγοριών συναισθημάτων και ποια είναι τα κριτήρια με τα οποία επιλέγουμε τόσο αυτές τις κατηγορίες, όσο και τις κατάλληλες ακουστικές ιδιότητες που θα μας οδηγήσουν σε γόνιμα αποτελέσματα. Για αυτό το λόγο, έχουν τεθεί σε χρήση διάφορες παραδοσιακές τεχνικές μηχανικής μάθησης όπως Hidden Markov Models (HMM), Gaussian Mixture Models (GMM) και Support Vector Machines (SVM) σε συνδυασμό με καλή προεπεξεργασία των ηχητικών δειγμάτων και feature engineering [2]. Όλα αυτά, σε συνδυασμό με την πληθώρα βάσεων δεδομένων που περιέχουν ηχητικά δείγματα για την εκπαίδευση και την δοκιμή των διατάξεων που έχουν κατασκευαστεί, οδήγησαν σε μοντέλα ικανοποιητικής ακρίβειας. Παρόλα αυτά, σύγχρονες εφαρμογές αλληλεπίδρασης ανθρώπου – υπολογιστή αλλά και εφαρμογές στην υγεία και την ψυχαγωγία καθιστούν επιτακτική την έρευνα πάνω σε αναδυόμενες τεχνολογίες βαθιάς μάθησης και νευρωνικών δικτύων [3].

Η υφιστάμενη κατάσταση της τεχνολογίας (state-of-the-art) στην αναγνώριση συναισθημάτων μέσω ομιλίας (SER) διατυπώνεται ως ένα πρόβλημα ταξινόμησης που εκμεταλλεύεται τυπικούς αλγορίθμους μηχανικής μάθησης αλλά και βαθιά νευρωνικά δίκτυα για την κατασκευή ικανοποιητικών μοντέλων αναγνώρισης συναισθημάτων μέσω ομιλίας [4]. Η τυπική μεθοδολογία που ακολουθείται ξεκινάει από την κατάτμηση των δειγμάτων σε μικρά χρονικά πλαίσια και την εξαγωγή ιδιοτήτων χαμηλού επιπέδου, όπως τη θεμελιώδη συχνότητα F_0, Mel-Frequency Cepstral συντελεστές (MFCC's) κ.ά. [5]. Στη συνέχεια, γίνεται επιλογή κάποιων ιδιοτήτων (features) και συναλήθευση αυτών με χρήση στατιστικών μεθόδων, ώστε να γίνει αναγωγή σε πρόβλημα λιγότερων διαστάσεων απ' το αρχικό (dimensionality reduction). Όλη αυτή η προεπεξεργασία που γίνεται στα ακατέργαστα δεδομένα όμως, οριοθετεί την ακρίβεια των εξαγόμενων μοντέλων, ενώ καθιστά δύσκολη τη γενίκευσή τους σε πραγματικές εφαρμογές που απαιτούν αποτελέσματα σε πραγματικό χρόνο. Για αυτό το λόγο, σήμερα γίνονται προσπάθειες ώστε να αντικατασταθεί η τυπική μεθοδολογία με αλγορίθμους end-to-end μηχανικής μάθησης, ο οποίοι να μπορούν να προσαρμοστούν σε ιδιότητες χαμηλού επιπέδου και να μπορέσουν να εξάγουν από αυτές νέες ιδιότητες υψηλού επιπέδου. Με αυτόν τον τρόπο δε θα υπάρχει υπερεξάρτηση των μοντέλων από την επιλογή των χαρακτηριστικών και από άλλα βήματα προεπεξεργασίας, πράγμα που αποτελεί και το στόχο της βαθιάς μάθησης [4].

2. Αναγνώριση Συναισθημάτων μέσω Ομιλίας – Επισκόπηση



Εικόνα 2: Επισκόπηση των συστημάτων αναγνώρισης συναισθημάτων από ομιλία [10]

2.1 Μοντέλα Συναισθημάτων

2.1.1 Πολυδιάστατα Μοντέλα Συναισθημάτων

Προκειμένου να κατατάξουμε τα συναισθήματα με χρήση υπολογιστή, απαιτείται η ύπαρξη ενός μαθηματικού μοντέλου που τα περιγράφει. Μία εκ των βασικών προσεγγίσεων είναι εκείνη που χρησιμοποιεί έναν μικρό αριθμό διαστάσεων για να χαρακτηρίσει τα συναισθήματα. Τέτοιες διαστάσεις είναι η δυναμικότητα (valence), η δραστηριοποίηση (activation), η κυριαρχία (dominance) κ.ά.. Σύμφωνα με αυτή την προσέγγιση, τα συναισθήματα δεν είναι ανεξάρτητα αλλά ανάλογα μεταξύ τους. Ένα συνηθισμένο πολυδιάστατο μοντέλο είναι ένα δισδιάστατο μοντέλο που τοποθετεί τη δραστηριοποίηση στον ένα άξονα και τη δυναμικότητα στον άλλο. Η δυναμικότητα περιγράφει το κατά πόσο ένα συναίσθημα είναι θετικό ή αρνητικό και κυμαίνεται μεταξύ ευχάριστου και δυσάρεστου. Η δραστηριοποίηση ορίζει την ένταση του επικείμενου συναισθήματος και άρα κινείται μεταξύ απάθειας και απόλυτου ενθουσιασμού. Το τρισδιάστατο μοντέλο περιλαμβάνει και τη διάσταση της κυριαρχίας, η οποία αναφέρεται στη φαινομενική δύναμη του ατόμου. Έτσι, η τρίτη διάσταση διαφοροποιεί τον θυμό από τον φόβο, θεωρώντας το άτομο δυνατό και αδύναμο, αντίστοιχα. Τέτοια μοντέλα επιτρέπουν την περιγραφή σχεδόν όλων των συναισθημάτων, ωστόσο δεν είναι επαρκώς διαισθητικά, γεγονός που δυσκολεύει την τιτλοφόρηση των δεδομένων. Επιπλέον, ένα τέτοιο σύστημα θα ήταν αρκετά σύνθετο για να υλοποιηθεί με χρήση μηχανικής μάθησης. Για αυτό το λόγο, στις μελέτες που σχετίζονται με

μηχανική μάθηση τα δείγματα χωρίζονται σε διακριτές κατηγορίες, σύμφωνα με κάποιο διακριτό μοντέλο συναισθημάτων [2].

2.1.2 Διακριτά Μοντέλα Συναισθημάτων

Τα διακριτά μοντέλα συναισθημάτων, αντί να παρέχουν ένα συνεχές φάσμα συναισθημάτων με βάση ορισμένες διαστάσεις, τα κατατάσσουν σε ένα πεπερασμένο πλήθος κατηγοριών. Ένα από τα πιο δημοφιλή μοντέλα, με εκτεταμένη χρήση και στην αναγνώριση συναισθημάτων από την ομιλία, θεωρείται εκείνο του Paul Ekman [6], που χωρίζει τα συναισθήματα σε έξι βασικές κατηγορίες: λύπη, χαρά, φόβο, θυμό, αποστροφή και έκπληξη. Αυτά τα εγγενή και ανεξάρτητα από το πολιτισμικό υπόβαθρο συναισθήματα, βιώνονται για σύντομο χρονικό διάστημα, και επίσης ένα πλήθος άλλων συναισθημάτων μπορούν να προκύψουν ως συνδυασμός των παραπάνω. Η πλειοψηφία των συστημάτων που αναπτύσσονται για την αναγνώριση συναισθημάτων από την ομιλία, εστιάζουν σε αυτές τις κατηγορίες συναισθημάτων. Ωστόσο, τα διακριτά αυτά μοντέλα αδυνατούν να περιγράψουν τις σύνθετες συναισθηματικές καταστάσεις που παρατηρούνται στην καθημερινή επικοινωνία.

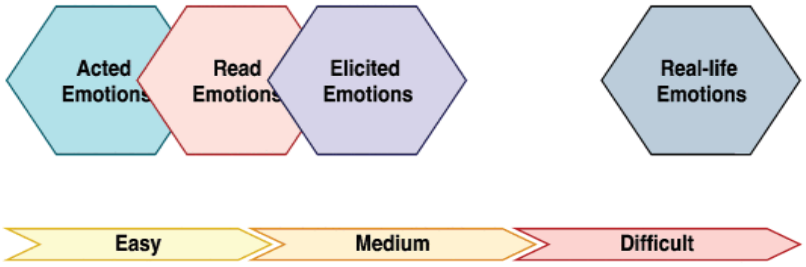
2.2 Σύνολα Δεδομένων

Για να χρησιμοποιηθεί οποιαδήποτε μέθοδος μηχανικής μάθησης χρειαζόμαστε αρχικά ένα σύνολο δειγμάτων ούτως ώστε να εκπαιδεύσουμε το μοντέλο και να ελέγξουμε κατά πόσο ταξινομεί σωστά τα δείγματα στα οποία δεν γνωρίζει εκ των προτέρων τη σωστή απάντηση. Για να κατασκευαστεί μία καλά θεμελιωμένη βάση δεδομένων για SER, απαιτείται πρώτα η ετικετοποίηση των δειγμάτων (labeling) χειροκίνητα με ανθρώπινη κρίση. Λόγω της διφορούμενης φύσης αυτής της διαδικασίας αξιοποιούνται περισσότεροι του ενός άνθρωποι και επιλέγονται τα δείγματα στα οποία υπάρχει περισσότερη εμπιστοσύνη στην επισήμανση που έγινε [2]. Οι τρείς βασικές κατηγοριοποιήσεις των συνόλων δεδομένων που περιγράφηκαν είναι οι εξής: Προσομοιωμένα (simulated), ημι-φυσικά (semi-natural ή induced) και φυσικά (natural). Τα προσομοιωμένα σύνολα δεδομένων κατασκευάζονται από εκπαιδευμένους ομιλητές – ηθοποιούς φωνής οι οποίοι διαβάζουν το ίδιο κείμενο με διαφορετικό συναίσθημα κάθε φορά. Τα ημι-φυσικά σύνολα δεδομένων κατασκευάζονται προσλαμβάνοντας απλούς ανθρώπους ή ηθοποιούς να διαβάσουν ένα σενάριο συνομιλίας που περιέχει διάφορα συναισθήματα. Τέλος, τα φυσικά σύνολα δεδομένων είναι ηχητικά δείγματα που εξάγονται από πραγματικές συνομιλίες στις οποίες υπάρχει πραγματική εξωτερίκευση συναισθημάτων από τους ομιλητές και ταξινομούνται από ανθρώπους ακροατές.

Λόγω της χρονοβόρας και απαιτητικής φύσης της ετικετοποίησης συναισθημάτων σε ηχητικά αποσπάσματα, οι βάσεις δεδομένων SER με μεγάλο αριθμό δειγμάτων είναι αρκετά περιορισμένες. Επιπλέον, τα πιο σημαντικά δημοφιλή σύνολα δεδομένων κατασκευάζονται σε ελεγχόμενο περιβάλλον που περιέχει δεδομένα απαλλαγμένα από θόρυβο τα οποία μπορεί να μην αντικατοπτρίζουν συνθήκες του πραγματικού κόσμου. Γίνεται προ-επεξεργασία των ηχητικών δειγμάτων όπως επαύξηση κάποιων χαρακτηριστικών τους με διάφορους μετασχηματισμούς, λόγου χάριν την επέκταση του χρόνου, την αύξηση ή μείωση της έντασης κάποιων συχνοτήτων και την προσθήκη θορύβου.

Είναι προφανές πως η πολυπλοκότητα των δειγμάτων και η εξαγωγή χαρακτηριστικών από αυτά αυξάνεται καθώς αυξάνεται και ο ρεαλισμός των δειγμάτων. Συγκεκριμένα, για τα προσομοιωμένα σύνολα δεδομένων, διακρίνουμε πολύ απλοϊκά δείγματα στα οποία είναι πολύ προφανές το συναίσθημα που υπονοείται. Σχεδόν το 60% των βάσεων δεδομένων ομιλίας κατατάσσεται σε αυτή την κατηγορία [3]. Όσον αφορά τα ημι-φυσικά σύνολα δεδομένων, συγκριτικά με τα προσομοιωμένα, προσφέρονται πιο φυσικά δείγματα. Πρέπει όμως να αναφερθεί πως η εξαγωγή των δειγμάτων γίνεται

χωρίς να γνωρίζει ο ομιλητής ότι καταγράφεται και συνεπώς τίθεται ένα πρόβλημα ηθικής φύσεως, κατά πόσο συναινεί ο ομιλητής στην καταγραφή για ερευνητικές δραστηριότητες. Τα φυσικά σύνολα δεδομένων είναι και τα πιο ρεαλιστικά, υστερούν όμως στην ποιότητα του ήχου καθώς καταγράφονται σε δημόσια περιβάλλοντα, τηλεφωνικά κέντρα και εκδηλώσεις. Για αυτό το λόγο απαιτούν πολύ προ-επεξεργασία και (αφαίρεση θορύβου και ενίσχυση) ενώ συχνά δεν είναι ξεκάθαρα τα συναισθήματα που υπαινίσσονται.



Εικόνα 3: Βάσεις δεδομένων αναγνώρισης συναισθημάτων και επίπεδο δυσκολίας [3]

Στη συνέχεια παρουσιάζονται οι πιο σημαντικές βάσεις δεδομένων αναγνώρισης συναισθημάτων για τις τρεις κατηγορίες που προαναφέρθηκαν.

2.2.1 Berlin Database of Emotional Speech (EMO-DB)

Μία από τις πιο ευρέως γνωστές προσομοιωμένες βάσεις δεδομένων που χρησιμοποιείται για SER. Είναι στα γερμανικά και αποτελείται από δέκα προτάσεις, πέντε μεγάλες και πέντε μικρές τις οποίες προφέρουν με διαφορετικά συναισθήματα πέντε γυναίκες και πέντε άνδρες ομιλητές. Τα συναισθήματα στα οποία είναι ταξινομημένα τα δείγματα είναι: ουδετερότητα, θυμός, φόβος, χαρά, λύπη, αποστροφή, ανία. Επιπλέον, με κάθε καταγραφή, ανιχνεύθηκε και το ηλεκτρογλωττογράφημα (EGG) των ομιλητών για την εξαγωγή προσωδιακών και φωνητικών χαρακτηριστικών, χαρακτηριστικά πολύ σημαντικά για την αυτοματοποιημένη ταξινόμηση των συναισθημάτων.

2.2.2 Danish Emotional Database (DES)

Πολύ γνωστή Δανική προσομοιωμένη βάση δεδομένων που αποτελείται από δύο γυναίκες και δύο άνδρες ηθοποιούς. Καθένας από τους ηθοποιούς διάβασε δύο λέξεις, εννιά προτάσεις και δύο μακροσκελή εδάφια, καταδεικνύοντας ουδετερότητα, έκπληξη, χαρά, λύπη και θυμό σε καθένα από αυτά. Τα αποτελέσματα ταξινομήθηκαν από είκοσι ακροατές όσον αφορά το συναίσθημα και την επιτυχία τους.

2.2.3 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

Μια από τις μεγαλύτερες βάσεις που περιγράφονται, καθώς αποτελείται από 2496 δείγματα ταξινομημένα στα συναισθήματα: χαρά, λύπη, θυμός φόβος, έκπληξη, μεταστροφή, ηρεμία και ουδετερότητα. Η πιο σημαντική διαφορά της όμως σε σχέση με τις υπόλοιπες βάσεις είναι ότι κάθε πρόταση προφέρεται και με κανονικό αλλά και με τραγουδιστό τρόπο. Επιπλέον, τα δείγματα προφέρονται και με Αγγλική αλλά και με Βορειοαμερικανική προφορά, γεγονός που προσφέρει επιπλέον πληροφορία στον ταξινομητή που κατασκευάζουμε.

2.2.4 The eNTERFACE Audio – Visual Emotion Database

Η συγκεκριμένη οπτικοακουστική βάση αποτελεί σημαντικό κομμάτι αυτής της αναφοράς καθώς χρησιμοποιείται σε ένα πείραμα που περιγράφεται στη συνέχεια. Είναι μία πολυποίκιλη βάση δεδομένων αφού περιέχει 42 ομιλητές με διαφορετικές καταγωγές από πολλές χώρες, ενώ το 81% των ομιλητών είναι άνδρες και το 19% γυναίκες [7]. Περιέχει 1296 διαφορετικά δείγματα στην αγγλική γλώσσα τα οποία έχουν ταξινομηθεί στις εξής έξι κατηγορίες: θυμός, αποστροφή, φόβος, χαρά, λύπη και έκπληξη. Η επισήμανση έγινε κατά την διήγηση μίας ιστορίας που προκαλούσε κάποιο συναίσθημα. Στη συνέχεια παρατίθεται μία μελέτη στην οποία χρησιμοποιείται η συγκεκριμένη βάση σε μία διάταξη ενός βαθιού νευρωνικού δικτύου, το οποίο ορίζει αποσπάσματα ενός δευτερολέπτου στα ηχητικά δείγματα της βάσης, και εκμεταλλεύεται το φασματογράφημά τους, ώστε να εξάγει πληροφορία σχετική τόσο με ακουστικά όσο και με σημασιολογικά χαρακτηριστικά [4].

2.2.5 Surrey Audio-Visual Expressed Emotion Database (SAVEE)

Όπως η προηγούμενη, και η συγκεκριμένη βάση δεδομένων χρησιμοποιήθηκε στη μελέτη που προαναφέρθηκε με σκοπό να συγκριθούν τα εξαγόμενα αποτελέσματα και να αποφανθούμε σχετικά με την δυνατότητα γενίκευσης της μεθόδου που χρησιμοποιήθηκε [4]. Το συγκεκριμένο σύνολο δεδομένων περιέχει 480 δείγματα στην αγγλική γλώσσα από άνδρες ομιλητές και τα συναισθήματα που περιέχει είναι τα εξής: θυμός, αποστροφή, φόβος, χαρά, λύπη, έκπληξη, και ουδετερότητα.

2.2.6 Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D)

Η μεγαλύτερη βάση δεδομένων με 7.442 δείγματα, από 91 ηθοποιούς και με ευρύ φάσμα πολιτισμικών ταυτοτήτων. Καλύπτει έξι συναισθήματα: θυμό, χαρά, λύπη, φόβο, αποστροφή και ουδετερότητα, τα οποία έχουν βαθμολογηθεί με πληθοπορισμό (crowdsourcing) από 2.443 βαθμολογητές. Η ταξινόμηση που έγινε από το κοινό διασταυρώθηκε με την αρχικά προοριζόμενη ταξινόμηση και με χρήση του στατιστικού συντελεστή Krippendorff's alpha έγινε διαλογή των καλύτερων δειγμάτων. Επιπλέον, σε κάθε κλάση συναισθήματος περιέχονται περίπου ίδιος αριθμός δειγμάτων.

2.2.7 Interactive Emotional Dyadic Motion Capture Database (IEMOCAP)

Η βάση IEMOCAP είναι ένα ημι-φυσικό σύνολο δεδομένων στην Αγγλική γλώσσα και περιέχει τόσο ηχητικά όσο και οπτικά δείγματα. Αποτελείται από δέκα ομιλητές, πέντε γυναίκες και πέντε άνδρες και έχει 1.150 δείγματα. Τα συναισθήματα που εμφανίζονται είναι: χαρά, λύπη, θυμός, εκνευρισμός. Επιπλέον, τα δείγματα έχουν ετικετοποιηθεί με τα χαρακτηριστικά κυριαρχία (dominance), δυναμικότητα (valence) και δραστηριοποίηση (activation) αξιοποιώντας έτσι και το τρισδιάστατο ντετερμινιστικό μοντέλο αναγνώρισης συναισθημάτων που χρησιμοποιείται από ψυχαναλυτές. Το συγκεκριμένο σύνολο δεδομένων περιέχει πολύ πιο φυσικούς διαλόγους σε σχέση με τα προηγούμενα. Κατά μέσο όρο, οι συνομιλίες διαρκούν πέντε λεπτά, πράγμα που τις καθιστά κατάλληλες για εφαρμογές βαθιάς μάθησης.

2.2.8 Vera am Mittag Database (VAM)

Ένα φυσικό οπτικοακουστικό σύνολο δεδομένων, βασισμένο στους διαλόγους του ομότιτλου Γερμανικού τηλεοπτικού σόου. Περιέχει τρισδιάστατη πληροφορία όπως εκείνη του IEMOCAP και 47 ομιλητές με 1.018 δείγματα. Λόγω του γεγονότος ότι είναι φυσικό σύνολο δεδομένων, παρατηρούνται πολλές γρήγορες εναλλαγές συναισθημάτων, ενώ τα ίδια τα συναισθήματα δεν είναι ξεκάθαρα και εστιασμένα.

Παρακάτω παρατίθενται τα χαρακτηριστικά των προαναφερθεισών βάσεων, καθώς επίσης και κάποιων άλλων σημαντικών πηγών.

Πίνακας 1: Χαρακτηριστικά γνωστών συνόλων δεδομένων [2], [4], [7]

	EMO-DB	**DES**	**RAVDESS**	**eNTERFACE**	**SAVEE**	**CREMA-D**	**IEMOCAP**	**VAM**
Συναισθήματα	7	5	8	6	7	6	7	3
Δείγματα	700	210	2496	1296	480	7442	1150	1018
Ομιλητές	10	4	24	42	4	91	10	47
Μέσο μέγεθος δείγματος	2.8 s	2.7 s	3.7 s	-	-	2.5 s	5 m	3.0 s
Θυμός	•	•	•	•	•	•	•	
Χαρά	•	•	•	•	•	•	•	
Λύπη	•	•	•	•	•	•	•	
Ουδετερότητα	•	•	•		•	•	•	
Έκπληξη		•	•	•	•		•	
Φόβος	•		•	•	•	•	•	
Αποστροφή	•		•	•	•	•	•	
Ανία	•							
Ηρεμία			•					
Απογοήτευση							•	
Ενθουσιασμός							•	
Σθένος							•	•
Δραστηριοποίηση							•	•
Κυριαρχία							•	•

2.3 Επιλογή Χαρακτηριστικών

Η αναγνώριση συναισθημάτων βασίζεται σε μεγάλο βαθμό στην αποτελεσματικότητα των χαρακτηριστικών που θα επιλεγούν για τη διαδικασία της ταξινόμησης. Για το λόγο αυτό, σημαντική έρευνα έχει γίνει για την εύρεση εκείνων των στοιχείων της ομιλίας που αντιπροσωπεύουν καλύτερα διαφορετικές συναισθηματικές καταστάσεις. Ωστόσο, δεν υπάρχει κάποιο καθολικά αποδεκτό σύνολο χαρακτηριστικών που να οδηγεί σε ξεκάθαρο διαχωρισμό.

Τα ποικίλα χαρακτηριστικά που μπορούν να εξαχθούν από ένα σήμα φωνής και βρίσκουν εφαρμογή σε συστήματα SER, μπορούν να χωριστούν σε τρεις κατηγορίες:

- Προσωδικά χαρακτηριστικά (Prosodic features)
- Φασματικά χαρακτηριστικά (Spectral features)
- Χαρακτηριστικά ποιότητας φωνής (Voice Quality features)

2.3.1 Προσωδικά Χαρακτηριστικά

Τα προσωδικά χαρακτηριστικά είναι εκείνα που γίνονται αντιληπτά από τους ανθρώπους, όπως ο ρυθμός, ο επιτονισμός και η ένταση και έχει βρεθεί ότι είναι αντιπροσωπευτικά του συναισθηματικού περιεχομένου του λόγου [8]. Τα πιο διαδεδομένα προσωδικά χαρακτηριστικά στη βιβλιογραφία είναι η θεμελιώδης συχνότητα (F_0), η ενέργεια και η διάρκεια. Η θεμελιώδης συχνότητα είναι αποτέλεσμα της

δόνησης των φωνητικών χορδών και σχετίζεται με τα τονικά και ρυθμικά χαρακτηριστικά του λόγου. Η ενέργεια ενός σήματος φωνής, ενίοτε αναφερόμενη και ως όγκος ή ένταση, παρέχει μια αναπαράσταση που αντανακλά τις διακυμάνσεις πλάτους του σήματος στο πέρασμα του χρόνου. Έρευνες συνιστούν ότι συναισθήματα υψηλής δραστηριοποίησης όπως ο θυμός, η χαρά και η έκπληξη συνοδεύονται από υψηλή ενέργεια, ενώ συναισθήματα όπως η αποστροφή και η λύπη από μειωμένη ενέργεια. Η διάρκεια αφορά στον χρόνο που χρειάζεται για την εκφορά γλωσσικών δομών και συνήθως μετράται με χαρακτηριστικά σχετικά με το ρυθμό ομιλίας, όπως ο αριθμός λέξεων ανά λεπτό, η μέση διάρκεια παύσεων κ.ά. [9]. Γενικά, τα προσωδικά χαρακτηριστικά υστερούν στο διαχωρισμό έντονων συναισθημάτων, όπως ο θυμός και η χαρά, ωστόσο μαζί με τα φασματικά χαρακτηριστικά είναι από τα πιο συχνά χρησιμοποιούμενα χαρακτηριστικά.

2.3.2 Φασματικά Χαρακτηριστικά

Τα φασματικά χαρακτηριστικά έχουν μελετηθεί εκτενώς για την ανίχνευση συναισθημάτων στην ομιλία. Το σημαντικότερο προτέρημά τους έναντι των προσωδικών χαρακτηριστικών είναι ότι μπορούν εύκολα να ξεχωρίσουν τα έντονα συναισθήματα. Ορισμένα φασματικά χαρακτηριστικά που χρησιμοποιούνται σε εφαρμογές SER είναι οι Mel-Frequency Cepstral Coefficients (MFCC), οι Linear Prediction Cepstral Coefficients (LPCC), οι Log-Frequency Power Coefficients (LFPC), οι Gammatone Frequency Cepstral Coefficients (GFCC) και τα Formants. Παρακάτω αναλύονται οι MFCC ως πιο διαδεδομένοι.

2.3.2.1 Mel-Frequency Cepstral Coefficients (MFCCs)

Οι συντελεστές MFCCs συνιστούν ένα από τα βασικότερα και πιο αποτελεσματικά σύνολα χαρακτηριστικών στην αναγνώριση ομιλίας. Προκύπτουν από μια πιο εξειδικευμένη ανάλυση τύπου Cepstrum που σε πρώτη φάση επιδίωκε να μιμηθεί τον τρόπο με τον οποίο το ανθρώπινο ακουστικό σύστημα επεξεργάζεται τον ήχο. Σε αντίθεση με την κλασική ανάλυση Cepstrum που τοποθετεί ίδια βάρη σε κάθε περιοχή συχνοτήτων, το Mel-Frequency Cepstrum αυξάνει τα βάρη των χαμηλών συχνοτήτων. Αυτό έρχεται σε συμφωνία με την ευρέως γνωστή εξάρτηση της ευαισθησίας της ανθρώπινης ακοής από τη συχνότητα. Πράγματι, η ακουστική ικανότητα του ανθρώπου γενικά κινείται μεταξύ 20-20kHz, όμως η ικανότητά του να διακρίνει το τονικό ύψος είναι σαφώς καλύτερη στις χαμηλές συχνότητες. Οι MFCCs παράγονται εφαρμόζοντας ένα μετασχηματισμό συνημιτόνου στον λογάριθμο του φάσματος βραχέος χρόνου εκφρασμένου σε Mel.

2.3.3 Χαρακτηριστικά Ποιότητας Φωνής

Τα χαρακτηριστικά ποιότητας φωνής συνήθως υπερέχουν στην αναγνώριση συναισθημάτων από τον ίδιο ομιλητή. Ωστόσο, επειδή καθορίζονται από τις φυσικές ιδιότητες της φωνητικής οδού, διαφέρουν από ομιλητή σε ομιλητή και αυτό δυσχεραίνει τη χρήση τους σε συνθήκες όπου υπάρχουν διαφορετικοί ομιλητές. Ακούσιες αλλαγές στη φωνητική οδό οδηγούν σε διαφοροποιήσεις στο φωνητικό σήμα που προκύπτει, οι οποίες συχνά συνδέονται με τα συναισθήματα του ομιλητή. Ορισμένες μετρικές που χρησιμοποιούνται είναι το jitter, το shimmer και ο λόγος αρμονικών προς θόρυβο (harmonics to noise ratio – HNR). Το jitter αφορά στη μεταβλητότητα της θεμελιώδους συχνότητας ανάμεσα σε διαδοχικές δονήσεις, ενώ το shimmer αφορά στην αντίστοιχη μεταβλητότητα του πλάτους. Συνεπώς, το jitter δίνει ένα μέτρο της αστάθειας της συχνότητας και το shimmer της αστάθειας του πλάτους. Τέλος, το HNR δηλώνει την ύπαρξη ή όχι θορύβου στο φάσμα των φωνημάτων και είναι ο λόγος μεταξύ περιοδικών και απεριοδικών συνιστωσών ενός φωνήματος.

Πίνακας 2: Συνοπτικές ακουστικές ιδιότητες των συναισθημάτων [3]

Emotions	Pitch	Intensity	Speaking rate	Voice quality
Anger	abrupt on stress	much higher	marginally faster	breathy, chest
Disgust	wide, downward inflections	lower	very much faster	grumble chest tone
Fear	wide, normal	lower	much faster	irregular voicing
Happiness	much wider, upward inflections	higher	faster/slower	breathy, blaring tone
Joy	high mean, wide range	higher	faster	breathy; blaring timbre
Sadness	slightly narrower	downward inflections	lower	resonant

2.4 Μέθοδοι Ταξινόμησης

2.4.1 Κλασικές Μέθοδοι Ταξινόμησης

Στα πρώτα στάδια της ανάπτυξής του, το πεδίο της αναγνώρισης συναισθημάτων από την ομιλία υιοθέτησε τις υπάρχουσες προσεγγίσεις για την αναγνώριση φωνής (Automatic Speech Recognition – ASR). Έτσι, σε πρώτη φάση χρησιμοποιήθηκαν μέθοδοι όπως SVMs, GMMs και HMMs, οι οποίες παραμένουν διαδεδομένες μέχρι και σήμερα. Ένα μειονέκτημα που παρουσιάζουν αυτές οι μέθοδοι είναι ότι απαιτούν εκτενές feature engineering και βαθιά κατανόηση του αντικειμένου για την εύρεση των καταλληλότερων χαρακτηριστικών. Μεταξύ των τεχνικών ταξινόμησης που έχουν χρησιμοποιηθεί συγκαταλέγονται επίσης τα δέντρα απόφασης (Decision Trees – DT), ο ταξινομητής k κοντινότερων γειτόνων (k-NN), ο Naive Bayes και η ανάλυση πρωτευουσών συνιστωσών (PCA) [3], [10], [11].

2.4.1.1 Support Vector Machines (SVMs)

Τα Support Vector Machines είναι μία οικογένεια αλγορίθμων επιβλεπόμενης μάθησης που ταξινομούν κατασκευάζοντας ένα υπερ-επίπεδο που διαχωρίζει γραμμικά τα δεδομένα σε κλάσεις με τον βέλτιστο τρόπο. Εάν δεν είναι εφικτός ο γραμμικός διαχωρισμός, χρησιμοποιείται μία συνάρτηση kernel η οποία απεικονίζει τα αρχικά δεδομένα σε ένα νέο γεωμετρικό χώρο και στη συνέχεια αναζητείται η βέλτιστη ταξινόμηση στον χώρο αυτό. Θεωρούνται από τους διαδεδομένους και ακριβείς ταξινομητές στην αναγνώριση συναισθημάτων από την ομιλία [12], επιτυγχάνοντας αντίστοιχες ή και καλύτερες επιδόσεις με πιο εξελιγμένες προσεγγίσεις. Ενδεικτικά, έχει βρεθεί ότι εξάγοντας χαρακτηριστικά μέσω βαθιάς μάθησης και τροφοδοτώντας με αυτά ένα SVM, μπορεί να επιτευχθεί κορυφαία επίδοση [13].

Στη δημοσίευση των Chavhan κ.ά. [14], υπολογίζονται οι συντελεστές MFCC και Mel-Energy spectrum Dynamic Coefficients (MEDC) από τα αρχεία .wav των σημάτων ομιλίας του συνόλου δεδομένων EMO-DB και στη συνέχεια εφαρμόζονται σε ένα SVM. Τα αποτελέσματα δείχνουν ότι ένα τέτοιο σύστημα παρουσιάζει καλές επιδόσεις ανεξαρτήτως ομιλητή και συγκεκριμένα χρησιμοποιώντας ως συνάρτηση kernel την Radial Basis Function (RBF) και την πολυωνυμική, τα αποτελέσματα ήταν 93.75% και 96.25%, αντίστοιχα.

2.4.1.2 Hidden Markov Models (HMMs)

Τα κρυφά Μαρκοβιανά μοντέλα είναι μία μέθοδος που χρησιμοποιείται ευρέως στην αναγνώριση φωνής – ήδη από τη δεκαετία του '60 – και η οποία έχει επεκταθεί επιτυχώς για την αναγνώριση συναισθημάτων. Όπως δηλώνει η ονομασία, ένα κρυφό Μαρκοβιανό μοντέλο βασίζεται στην ιδιότητα Markov, σύμφωνα με την οποία η κατάσταση μίας στοχαστικής διαδικασίας τη χρονική στιγμή t, εξαρτάται αποκλειστικά από την κατάσταση που είχε τη χρονική στιγμή t-1 και όχι από προγενέστερες καταστάσεις. Ο όρος "κρυφό" καταδεικνύει την αδυναμία παρατήρησης της διαδικασίας που προκαλεί την κατάσταση κατά τη στιγμή t. Το μοντέλο επιχειρεί να προβλέψει τη μελλοντική κατάσταση παρατηρώντας την τωρινή κατάσταση του συστήματος.

2.4.1.3 Gaussian Mixture Models (GMMs)

Τα μοντέλα μειγμάτων Γκαουσιανών είναι πιθανοτικά μοντέλα που θεωρούν ότι το σύνολο των δεδομένων έχει προκύψει από τον συνδυασμό ενός πεπερασμένου αριθμού Γκαουσιανών κατανομών με άγνωστες παραμέτρους. Έχουν μελετηθεί εκτενώς στα πλαίσια της αναγνώρισης συναισθημάτων από ομιλία.

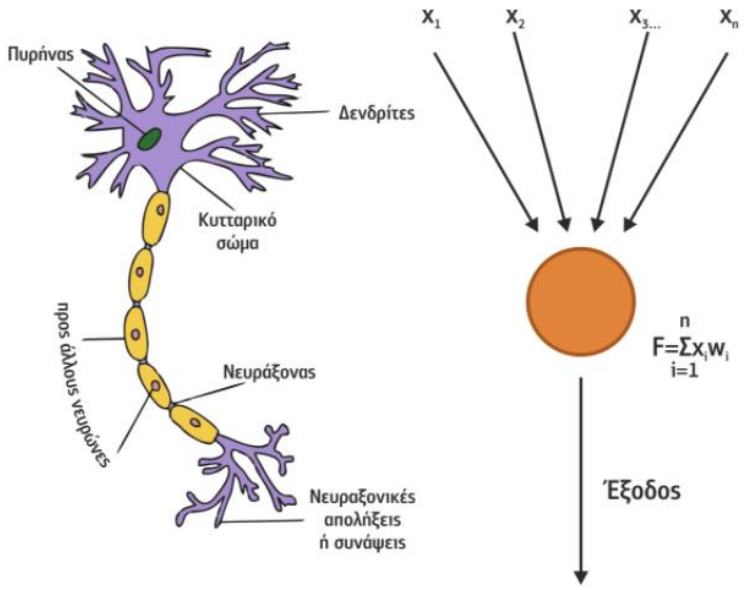
2.4.2 Νευρωνικά Δίκτυα

Η τεχνητή νοημοσύνη και τα νευρωνικά δίκτυα γνώρισαν μεγάλη ανάπτυξη από την αρχή της δεκαετίας του 1980 και μετά, με αποτέλεσμα πολλοί ερευνητές να αρχίσουν να εντάσσουν αυτές τις τεχνολογίες στο πεδίο έρευνάς τους. Αρχικά, εξαιτίας των περιορισμών στην μνήμη και την ταχύτητα των επεξεργαστών του υπολογιστή δεν είχαν πετύχει εντυπωσιακά αποτελέσματα. Αυτό με την πάροδο του χρόνου και με την εξέλιξη των CPU και GPU άλλαξε και πλέον θεωρούνται state-of-the-art μέθοδοι για την αντιμετώπιση προβλημάτων αναγνώρισης προτύπων και ταξινόμησης, όπως το SER. Οπότε, οι HMM και SVM αρχιτεκτονικές αντικαταστάθηκαν από Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks, ANNs), Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks, CNNs), Δίκτυα Μακράς Βραχύχρονης Μνήμης (Long Short-Term Memory, LSTM) και Παραγωγικά Αντιπαραθετικά Δίκτυα (Generative Adversarial Networks, GANs).

2.4.2.1 Artificial Neural Networks (ANNs)

Στη μηχανική μάθηση και στην επιστήμη των υπολογιστών πολλά προβλήματα δε μπορούν να μετασχηματιστούν έτσι ώστε να λύνονται με έναν σαφή αλγόριθμο. Η λύση σε τέτοιου είδους προβλήματα χρειάζεται να είναι δυναμική και να προσαρμόζεται για κάθε περίπτωση που εφαρμόζεται ο αλγόριθμος. Η προσαρμοστικότητα είναι ένα απαραίτητο χαρακτηριστικό του ανθρώπινου εγκεφάλου, αλλά στην περίπτωση των υπολογιστών δεν υπάρχει εύκολος τρόπος να γραφτεί «προσαρμοστικός» κώδικας. Ο ανθρώπινος εγκέφαλος έχει τη δυνατότητα να γενικεύει που τον βοηθάει να σκέφτεται επαγωγικά και να μαθαίνει. Εμπνευσμένα από τους νευρώνες του Κεντρικού Νευρικού Συστήματος, τα ANN είναι μια αποτελεσματική λύση για τη μάθηση. Έχουν τη δυνατότητα να μαθαίνουν και να χειρίζονται πολύπλοκες μη γραμμικές σχέσεις μεταξύ των εισόδων και των παραγόμενων εξόδων. Πιο συγκεκριμένα, τα ANN επεξεργάζονται πληροφορίες ενώ ανταποκρίνονται δυναμικά σε εξωτερικά ερεθίσματα (είσοδοι). Κάθε τεχνητός νευρώνας αποτελείται από πολλές εισόδους x_i και μία έξοδο y. Σε κάθε είσοδο x_i αντιστοιχεί ένα βάρος w_i και η συνάρτηση αθροίσματος F δίνει τα αποτελέσματα. Για να δώσει ο τεχνητός νευρώνας έξοδο μέσω της συνάρτησης μετάβασης πρέπει το ζυγισμένο άθροισμα των εισόδων να είναι μεγαλύτερο από μια ορισμένη τιμή κατωφλίου θ. Αν υποτεθεί ότι τα βάρη διασύνδεσης του τεχνητού νευρώνα σχηματίζουν τα ηλεκτρικά χαρακτηριστικά της επαφής της σύναψης του φυσικού νευρώνα, ενώ η τιμή του κατωφλίου προσομοιώνει τη συμπεριφορά κορεσμού του τότε γίνεται σαφές ότι ένας τεχνητός νευρώνας αποτελεί

μια απλοποιημένη μοντελοποίηση ενός φυσικού. Αυτό παρουσιάζεται σχηματικά παρακάτω, όπου αριστερά εμφανίζεται ο φυσικός νευρώνας και δεξιά ο τεχνητός:



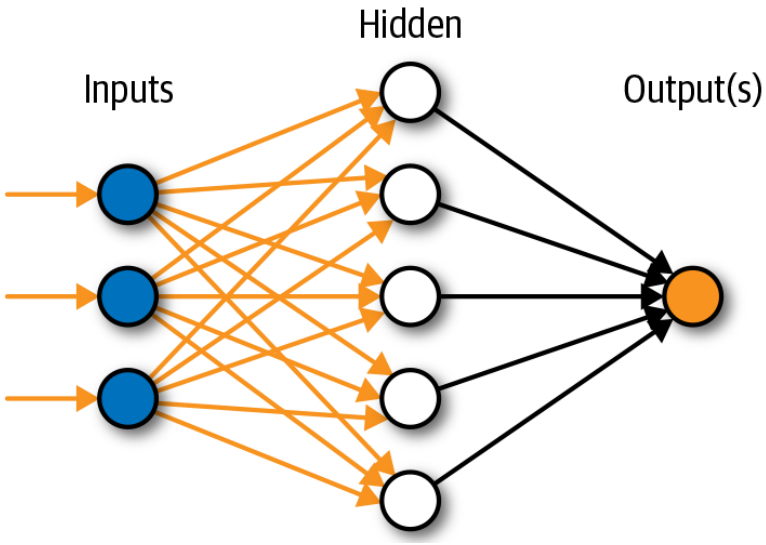
Εικόνα 4: Παρομοίωση φυσικού νευρώνα με τεχνητό[1]

Τα βασικά χαρακτηριστικά ενός νευρωνικού δικτύου είναι τα εξής:

- Είναι οργανωμένο σε επίπεδα (layers), ή αλλιώς στρώματα. Τα ενδιάμεσα επίπεδα ονομάζονται κρυμμένα επίπεδα και δεν είναι απαραίτητη η ύπαρξή τους.
- Κάθε επίπεδο αποτελείται από έναν αριθμό κόμβων (nodes), ή αλλιώς μονάδες, που συνδέονται έτσι ώστε ένας κόμβος να έχει συνδέσμους με πολλούς άλλους του ίδιου ή διαφορετικού επιπέδου.
- Οι κόμβοι αλληλεπιδρούν μεταξύ τους διεγείροντας ή αναστέλλοντας την ενεργοποίησή τους. Για το σκοπό αυτό ο κόμβος λαμβάνει το σταθμισμένο άθροισμα όλων των εισόδων μέσω των συνδέσμων που καταλήγουν σε αυτόν και μέσω της συνάρτησης μετάβασης παράγει μοναδική έξοδο, εάν το άθροισμα αυτό υπερβαίνει την τιμή κατωφλίου.
- Οι είσοδοι βρίσκονται στο επίπεδο εισόδου (input layer) το οποίο επικοινωνεί με ένα ή περισσότερα κρυμμένα επίπεδα (hidden layers). Αυτά συνδέονται με το επίπεδο εξόδου (output layer) από το οποίο εξάγεται η απόκριση του συστήματος.

[1] http://repfiles.kallipos.gr/html_books/93/04a-main.html

Παρακάτω απεικονίζεται ένα απλό παράδειγμα ANN:



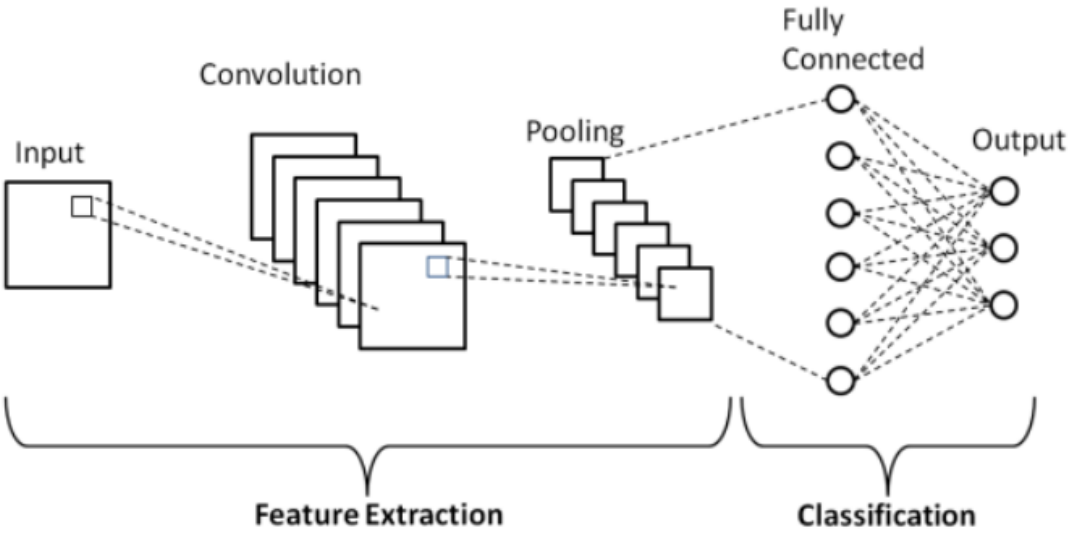
Εικόνα 5: Παράδειγμα Τεχνητού Νευρωνικού Δικτύου[2]

Στην έρευνά τους, ο Shaw και οι υπόλοιποι ερευνητές [15] δημιούργησαν ένα σύστημα αναγνώρισης χρησιμοποιώντας ANN για να αναγνωρίσουν τέσσερις κλάσεις συναισθημάτων: χαρά, λύπη, θυμός και ουδέτερο. Στην υλοποίηση του συστήματός τους συμπεριέλαβαν τα προσωδικά και τα φασματικά χαρακτηριστικά των σημάτων. Το δίκτυό τους, είχε ένα επίπεδο εισόδου, ένα κρυμμένο επίπεδο και την έξοδο των τεσσάρων κλάσεων. Το αποτέλεσμα που επέτυχαν ήταν 81% ορθότητα για όλες τις κλάσεις. Αργότερα, το 2018 οι Darekar και Dhande [16] παρουσίασαν ένα σύστημα βασισμένο σε τεχνητά νευρωνικά δίκτυα και στην ιδέα των Bhatnagar και Gupta [17]. Το μοντέλο αυτό χρησιμοποιούσε ένα ANN με δύο κρυμμένα επίπεδα και για να εκπαιδευτεί χρησιμοποιήθηκε ένας feedforward PSO αλγόριθμος. Εφαρμόστηκε στο σύνολο δεδομένων RAVDESS και επιτεύχθηκε κατά 10.85% βελτιωμένη ορθότητα από την αρχική μέθοδο. Έπειτα, αντικαταστάθηκε το ANN με έναν καλύτερο ταξινομητή και η ακρίβεια βελτιώθηκε ακόμα περισσότερο. Η υλοποίηση και η εκπαίδευση του ANN γίνεται γρηγορότερα από τα υπόλοιπα είδη νευρωνικών δικτύων, αλλά ένα ANN ενός μόνο επιπέδου δε μπορεί να λύσει πολύ πολύπλοκα και μη γραμμικά προβλήματα. Αυτό είναι το άνω φράγμα, στο οποίο το ANN σταματά. Δηλαδή, το ANN μπορεί να είναι γρήγορο, έχει όμως και περιορισμούς ως προς τις δυνατότητές του.

[2] https://www.innoarchitech.com/blog/artificial-intelligence-deep-learning-neural-networks-explained

2.4.2.2 Convolutional Neural Networks (CNN)

Τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs) είναι συγκεκριμένου τύπου νευρωνικά δίκτυα, όπου στο κρυμμένο τους επίπεδο έχουν διαφορετικά φίλτρα ή περιοχές που αποκρίνονται σε ένα συγκεκριμένο χαρακτηριστικό του σήματος εισόδου. Ένα μεγάλο πλεονέκτημα των συνελικτικών νευρωνικών δικτύων είναι η ικανότητά τους να μαθαίνουν χαρακτηριστικά από πολυδιάστατα δεδομένα εισόδου, ενώ ακόμη μαθαίνει την εμφάνιση μικρών παραλλαγών και διαστρεβλώσεων, και άρα προϋποθέτει την ύπαρξη κατάλληλου αποθηκευτικού χώρου την ώρα που τρέχει. Άρα, στα CNNs συνήθως υπάρχει ένα επίπεδο συνέλιξης ακολουθούμενο από έναν δειγματοληπτικό μηχανισμό. Το συνελικτικό επίπεδο διαθέτει ποικίλα φίλτρα, με τα οποία τα βάρη του ρυθμίζονται συχνά μέσω της οπισθοδιάδοσης [18]. Παρακάτω απεικονίζεται μια βασική δομή ενός συνελικτικού νευρωνικού δικτύου:



Εικόνα 6: Δομή ενός απλού CNN[3]

Σε μια έρευνα που πραγματοποίησαν οι D.Bertero και P.Fung [19] παρουσίασαν ένα CNN ικανό να ανιχνεύσει θυμό, χαρά και λύπη με 66.1% ποσοστό ορθότητας. Για να μπορέσουν να εκπαιδεύσουν και να επαληθεύσουν τη μέθοδό τους χρησιμοποίησαν ένα πλήθος από TED ομιλίες. Υλοποίησαν το CNN χρησιμοποιώντας το Theano toolkit. Για σύγκριση, εκπαίδευσαν μία γραμμική SVM με το σύνολο δεδομένων INTERSPEECH 2009. Ανέφεραν ότι το συνελικτικό νευρωνικό δίκτυο μπόρεσε να ανιχνεύσει τρεις κλάσεις: θυμό, χαρά και λύπη με ποσοστό ορθότητας 66.1%. Επίσης, έδειξαν ότι η νευρωνική δραστηριότητα συγκεντρώνεται γύρω από θεμελιώδεις συχνότητες συσχετισμένες με τα συναισθήματα. Το 2020, πραγματοποιήθηκε ακόμη μία μελέτη [20] με ένα μονοδιάστατο συνελικτικό νευρωνικό δίκτυο και αναφέρθηκε ορθότητα 96.60% στην ταξινόμηση αρνητικών συναισθημάτων από σύνολα δεδομένων της γλώσσας Tai. Σε αυτή τη μελέτη η αναπτυσσόμενη μέθοδος εφαρμόστηκε ακόμη στα σύνολα δεδομένων: SAVEE, RAVDESS, TESS και CREMA-D.

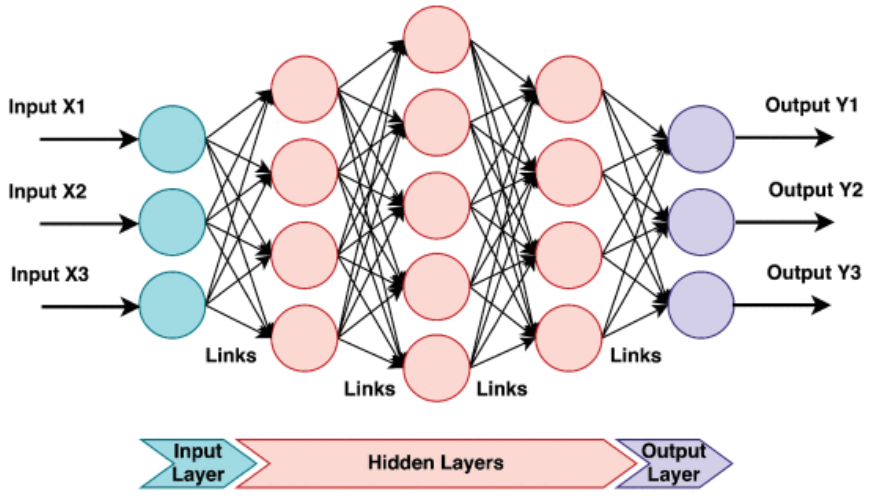
[3] https://www.researchgate.net/figure/Schematic-diagram-of-a-basic-convolutional-neural-network-CNN-architecture-26_fig1_336805909

2.4.3 Βαθιά Μάθηση με Νευρωνικά Δίκτυα

Η Βαθιά Μάθηση είναι μία κλάση από τεχνικές μηχανικής μάθησης που χαρακτηρίζονται από ιεραρχική πολυεπίπεδη αρχιτεκτονική, η οποία σκοπεύει στην εκμάθηση από ένα σύνολο δεδομένων εισόδου, με πολλαπλά επίπεδα αφαιρετικότητας [4]. Ο όρος «βαθιά» προκύπτει από τον αριθμό των κρυμμένων επιπέδων που μπορεί να φτάνει τις εκατοντάδες σε αντίθεση με τα «παραδοσιακά» νευρωνικά δίκτυα, στα οποία συνήθως υπάρχουν δύο με τρία επίπεδα. Θεωρείται σήμερα αναδυόμενη τεχνολογία λόγω των μεγάλων εξελίξεων που έχουν επιτευχθεί στα πεδία της επιβλεπόμενης και μη επιβλεπόμενης μάθησης, καθώς επίσης και στην βελτίωση του hardware και την αύξηση των υπολογιστικών πόρων. Χρησιμοποιείται ευρέως και σε άλλες υπολογιστικά απαιτητικές εφαρμογές όπως την όραση υπολογιστών, αναγνώριση φωνής και άλλα έργα που απαιτούν τεχνητή νοημοσύνη [5]. Ένα από τα πιο διάσημα παραδείγματα χρήσης βαθιάς μάθησης που απασχόλησαν ευρέως την επιστημονική κοινότητα ήταν η εμφάνιση του AlexNet το 2012 [21], το οποίο ήταν ένα Συνελικτικό Νευρωνικό Δίκτυο πολλαπλών επιπέδων που είχε εκπαιδευτεί στο σύνολο δεδομένων ImageNet το 2010 για να αναγνωρίζει 1.000 διαφορετικές κλάσεις και επιτύγχανε εξαιρετικά αποτελέσματα. Μετά από αυτό ακολούθησε ένα «κύμα» εξέλιξης και διαφορετικών αρχιτεκτονικών σε αυτό το πεδίο. Το 2011 πραγματοποιήθηκε μια μελέτη [22] που χρησιμοποιούσε βαθύ νευρωνικό δίκτυο για να δημιουργήσει την κατανομή πιθανότητας για διαφορετικά συναισθήματα δοσμένα σε τμήματα. Για να συγκρίνουν τα αποτελέσματά τους με άλλες μεθόδους χρησιμοποίησαν επίσης ένα σύστημα αναγνώρισης βασισμένο σε HMM, ένα σε SVM, ένα σε DNN-HMM και ένα σε DNN-SVM (DNN: Deep Neural Network). Έδειξαν ότι η ακρίβεια της προσέγγισής τους είναι θεωρητικά (5-20%) υψηλότερη συγκριτικά με τις άλλες μεθόδους. Τα επόμενα χρόνια ακολούθησαν πολλές ακόμη μελέτες στον τομέα SER.

2.4.3.1 Deep Neural Networks (DNNs)

Ως βαθιά νευρωνικά δίκτυα μπορούν να οριστούν τα δίκτυα με περισσότερα από ένα κρυμμένα επίπεδα. Με αυτό τον ορισμό καλύπτονται όλες οι βαθιές δομές, όπως τα συνελικτικά, τα Μακράς Βραχυπρόθεσμης Μνήμης, και τα παραγωγικά αντιπαραθετικά δίκτυα [2]. Ένα βαθύ νευρωνικό δίκτυο είναι ένα νευρωνικό δίκτυο με πολλαπλά κρυφά επίπεδα. Παρακάτω απεικονίζεται η βασική του δομή:



Εικόνα 7: Απλό παράδειγμα DNN [3]

Η πληροφορία τροφοδοτείται από την είσοδο του πρώτου επιπέδου προς μία κατεύθυνση. Κάθε επίπεδο του δικτύου περιέχει ένα πλήθος νευρώνων τα οποία δέχονται ένα βεβαρυμμένο άθροισμα των εξόδων του προηγούμενου επιπέδου και επιπλέον έναν αυθαίρετο όρο ως είσοδο, η οποία ανακατευθύνεται σε μία μη γραμμική συνάρτηση.

$$z^{(l)} = W^{(l)} y^{(l-1)} + b^{(l)}$$

$$y^{(l)} = f\left(z^{(l)}\right)$$

Όπου $l \in \{1,2,\dots,L-1\}$, L είναι το πλήθος των επιπέδων του δικτύου, $z^{(l)}$ είναι το διάνυσμα εισόδου στο επίπεδο l, $W^{(l)}$ και $b^{(l)}$ είναι ο πίνακας βαρών και το διάνυσμα αυθαίρετων όρων αντίστοιχα. Η είσοδος είναι το διάνυσμα $y^{(0)}$ και η f είναι η μη γραμμική συνάρτηση που εφαρμόζεται στην έξοδο κάθε νευρώνα.

Σημαντικές μη γραμμικές συναρτήσεις εξόδου αποτελούν τα σιγμοειδή, και η υπερβολική εφαπτομένη $\tan h$ συνάρτηση. Τελευταία όμως χρησιμοποιείται και η συνάρτηση πυροδότησης Rectified Linear Unit (ReLU) η οποία έχει πολλά πλεονεκτήματα σε σχέση με τις προηγούμενες όπως γρηγορότερη σύγκλιση εκμάθησης και πιο απλή βελτιστοποίηση. Η συνάρτηση αυτή έχει την εξής μορφή:

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

Για ένα πρόβλημα ταξινόμησης σε k κατηγορίες, με $k > 2$, η ύστερη πιθανότητα κάθε κλάσης υπολογίζεται τοποθετώντας σαν τελευταίο επίπεδο ένα επίπεδο softmax, για το οποίο ισχύουν τα εξής:

$$z^{(L)} = W^{(L)} y^{(L-1)} + b^{(L)}$$

$$y_k^{(L)} = \frac{\exp\left(z_k^{(L)}\right)}{\sum_{k=1}^{K} \exp\left(z_k^{(L)}\right)}$$

Όπου k ο δείκτης μίας συγκεκριμένης κλάσης, K το πλήθος των κλάσεων, $z^{(L)}$ το διάνυσμα εισόδου στο επίπεδο L, το οποίο είναι μήκους K και περιέχει τις πιθανότητες η είσοδος να ταξινομείται σε κάθε μία από τις K κλάσεις.

Στη συνέχεια, ο κλασικός τρόπος με τον οποίο εκπαιδεύεται το δίκτυο ώστε να διορθωθεί ο πίνακας βαρών και οι αυθαίρετοι όροι και να πετύχουμε καλύτερη ακρίβεια χρησιμοποιώντας ένα δείγμα τη φορά, ονομάζεται Stochastic Gradient Descent. Ουσιαστικά, είναι αλγόριθμος οπισθοδρόμησης βελτιστοποίησης μίας συνάρτησης κόστους. Η συνάρτηση αυτή για το τελικό softmax επίπεδο έχει την εξής μορφή:

$$C = -\sum_{k=1}^{K} Y_k \log\left(y_k^{(L)}\right)$$

Όπου Y_k είναι ένας πίνακας επιλογέας μεγέθους K που δείχνει σε ποια κλάση ανήκει πραγματικά το εξεταζόμενο δείγμα. Η παραπάνω συνάρτηση ονομάζεται και Cross Entropy Cost function.

Στη συνέχεια υπολογίζεται η παράγωγος της συνάρτησης κόστους και γίνεται οπισθοδρόμηση χρησιμοποιώντας κανόνα αλυσίδας για να υπολογιστούν και οι κλίσεις των προηγούμενων επιπέδων.

Λόγω του ότι τα βαθιά νευρωνικά δίκτυα είναι επιρρεπή στο overfitting (το δίκτυο έχει πολύ μεγάλη ακρίβεια μόνο στο σύνολο εκπαίδευσης), επιστρατεύονται στη συνέχεια μέθοδοι εξομάλυνσης του δικτύου) [5].

Το 2015, οι Fayek, Lech και Cavedon [4], επιχείρησαν να κατασκευάσουν μία διάταξη ενός βαθιού νευρωνικού δικτύου το οποίο όμως θα μπορούσε να παραγάγει αποτελέσματα σε πραγματικό χρόνο, πράγμα πολύ σημαντικό για σύγχρονες εφαρμογές.

Η πειραματική διάταξη που ακολουθήθηκε ήταν αυτή που περιγράφεται παρακάτω. Από τις οπτικοακουστικές βάσεις eNTERFACE και SAVEE, χρησιμοποιήθηκαν μόνο ακουστικά δείγματα. Τα δεδομένα διασπάστηκαν σε 70% δεδομένα εκπαίδευσης, 15% δεδομένα επικύρωσης και 15% δεδομένα ελέγχου. Τα σύνολα αυτά ήταν μεταξύ τους ξένα. Στη συνέχεια πραγματοποιήθηκε επαύξηση των δεδομένων, ένα σημαντικό βήμα στην εκμάθηση του δικτύου που σκοπεύει να ελαττώσει την υπερ-προσαρμογή που προαναφέρθηκε και βοηθά στη γενίκευση του μοντέλου. Αναφέρεται ότι αυτή η διαδικασία βελτίωσε σημαντικά τη γενίκευση της μεθοδολογίας που περιγράφεται. Τα δεδομένα του συνόλου εκπαίδευσης δειγματολήφθηκαν σε τέσσερις διαφορετικές συχνότητες δειγματοληψίας: 0,8, 0,9, 1,1, και 1,2 φορές επί την αρχική συχνότητα δειγματοληψίας. Έτσι πολλαπλασιάστηκε το πλήθος των δεδομένων του training set. Επιχειρήθηκε επιπλέον η χρήση και άλλων μεθόδων επαύξησης των δεδομένων όπως προσθήκη Gaussian noise, αλλά δεν φάνηκαν τόσο επιτυχείς όσο η προαναφερθείσα μέθοδος. Στη συνέχεια, τα δεδομένα δειγματολήφθηκαν ξανά με τη μικρότερη συχνότητα των 8kHz, και χρησιμοποιήθηκε αλγόριθμος εντοπισμού σιωπηλών αποσπασμάτων, τα οποία αφαιρέθηκαν. Στη συνέχεια, για την κατασκευή των φασματογραφημάτων των δειγμάτων, τα δείγματα αναλύθηκαν χρησιμοποιώντας πλαίσιο Hamming (διάσπαση του δείγματος σε παράθυρα μικρού μήκους) των 25ms ενώ το stride (απόσταση μεταξύ των αρχών δύο διαδοχικών πλαισίων) επιλέχθηκε να είναι 15ms. Τα φασματογραφήματα κατασκευάστηκαν χρησιμοποιώντας 41 γραμμικά τοποθετημένα, λογαριθμικά φίλτρα βασισμένα στον μετασχηματισμό Fourier.

Τα δεδομένα κανονικοποιήθηκαν ώστε να έχουν μηδενική μέση τιμή και διακύμανση ίση με τη μονάδα σε όλο το σύνολο δεδομένων. Αυτές οι τιμές υπολογίσθηκαν χρησιμοποιώντας το training set και μόνο τότε χρησιμοποιήθηκαν για να κανονικοποιηθεί η εκπαίδευση, η επικύρωση και το testing. Δεν εφαρμόστηκαν διαφορετικές μέθοδοι για διαφορετικό φύλο ομιλητή, καθώς θεωρήθηκε ότι κατά τη διάρκεια της δοκιμής του μοντέλου, δεν υπήρχαν πληροφορίες σχετικές με αυτόν.

Η διάσταση του διανύσματος εισόδου του δικτύου ήταν ίση με 2.624 ενώ το πλήθος των επιπέδων και των νευρώνων ήταν μεταβλητό. Για απλότητα, κάθε κρυμμένο επίπεδο είχε ίσο αριθμό νευρώνων. Όλα τα επίπεδα εκτός του τελευταίου χρησιμοποιούσαν συνάρτηση πυροδότησης με ReLU’s. Στην περίπτωση του eNTERFACE, η διάσταση του επιπέδου εξόδου ήταν έξι ενώ στο σύνολο δεδομένων SAVEE, ήταν επτά, όσες δηλαδή και οι κλάσεις συναισθημάτων για κάθε βάση δεδομένων. Τα βάρη του δικτύου αρχικοποιήθηκαν τυχαία και χρησιμοποιήθηκε Mini – batch SGD με πλήθος δειγμάτων για ένα πέρασμα ίσο με 128. Ο ρυθμός εκμάθησης (learning rate) αρχικοποιήθηκε σε 0,05 και αυξανόταν κατά το ήμισυ κάθε πέντε εποχές, το ποσοστό απόσυρσης (dropout rate) σε 0,5 για τα κρυφά επίπεδα και 0,9 για το επίπεδο εισόδου ενώ όλες οι έξοδοι των νευρώνων είχαν μέγιστο όριο την τιμή 1.

Όσον αφορά τα αποτελέσματα του πειράματος, στη μελέτη, ερευνήθηκαν τα αποτελέσματα του δικτύου για τα δύο σύνολα δεδομένων που προαναφέρθηκαν, καθώς επίσης και η εξάρτηση της απόδοσης του δικτύου από τον αριθμό των νευρώνων και το πλήθος των επιπέδων που επιστρατεύει. Χρησιμοποιήθηκε και στα δύο σύνολα δεδομένων η ίδια μέθοδος εκμάθησης.

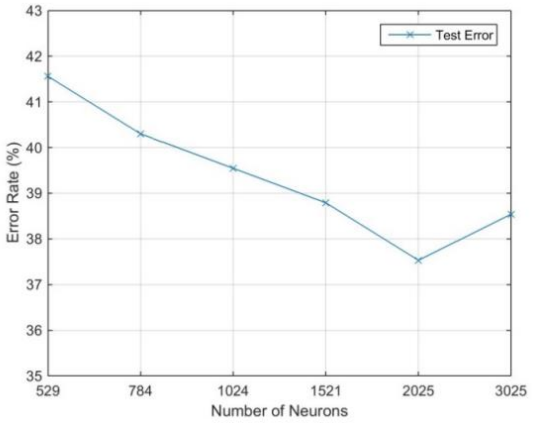
Τα αποτελέσματα φαίνονται παρακάτω.

Πίνακας 3: Επίδοση στο σύνολο δεδομένων eNTERFACE

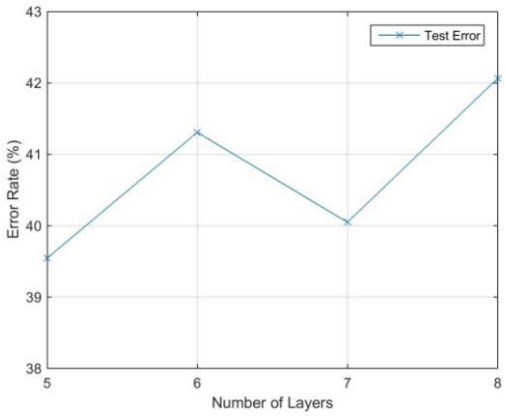
Συναίσθημα	Ακρίβεια	Ευαισθησία	F-score	Ορθότητα
Θυμός	0.63	0.7	0.66	-
Αποστροφή	0.65	0.52	0.57	-
Φόβος	0.51	0.58	0.55	-
Χαρά	0.68	0.46	0.55	-
Λύπη	0.57	0.76	0.65	-
Έκπληξη	0.64	0.57	0.6	-
Σύνολο	-	-	-	60.53%

Πίνακας 4: Επίδοση στο σύνολο δεδομένων SAVEE

Συναίσθημα	Ακρίβεια	Ευαισθησία	F-score	Ορθότητα
Θυμός	0.72	0.76	0.74	-
Αποστροφή	0.41	0.44	0.42	-
Φόβος	0.33	0.45	0.38	-
Χαρά	0.55	0.26	0.35	-
Ουδετερότητα	0.76	0.89	0.82	
Λύπη	0.71	0.71	0.71	-
Έκπληξη	0.5	0.53	0.51	-
Σύνολο	-	-	-	59.7%



Εικόνα 8: Ποσοστό σφαλμάτων για διάφορα πλήθη νευρώνων και συγκεκριμένο αριθμό κρυφών επιπέδων [4]



Εικόνα 9: Ποσοστό σφαλμάτων για διάφορα πλήθη κρυφών επιπέδων και συγκεκριμένο αριθμό νευρώνων ανά επίπεδο [4]

Όπως παρατηρήσαμε και στις παραπάνω εικόνες που δείχνουν το ποσοστό σφαλμάτων για διάφορα πλήθη νευρώνων και κρυφών επιπέδων, το ποσοστό σφαλμάτων μειώθηκε αυξάνοντας τους νευρώνες, μέχρι τους 2.025 νευρώνες όπου υπήρξε μικρή μείωση. Επιπλέον, η αύξηση των νευρώνων οδήγησε και στην αύξηση του χρόνου των υπολογισμών. Όσον αφορά το πλήθος των κρυμμένων επιπέδων, για σταθερό αριθμό νευρώνων ανά επίπεδο, δεν ήταν προφανές κάποιο μοτίβο.

Στους παραπάνω πίνακες παρατίθενται, για κάθε κλάση συναισθήματος, οι μετρικές:

- **Ακρίβεια**: Πλήθος σωστών θετικών προβλέψεων ανά πλήθος θετικών προβλέψεων $\frac{TP}{TP+FP}$
- **Ευαισθησία**: Πλήθος σωστών θετικών προβλέψεων ανά πλήθος πραγματικών θετικών δειγμάτων $\frac{TP}{TP+FN}$
- **F-score:** Μετρική που ισορροπεί τις δύο προηγούμενες $\mathrm{F}_1 = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision+Recall}}$

Όπου TP το πλήθος των δειγμάτων που σωστά ταξινομήθηκαν σε μία κλάση, FP το πλήθος των δειγμάτων που λανθασμένα ταξινομήθηκαν στην κλάση, FN τα δείγματα που λανθασμένα δεν ταξινομήθηκαν στην κλάση.

Παρατηρούμε πολύ καλά αποτελέσματα δεδομένης της απλότητας του δικτύου που κατασκευάστηκε. Και στις δύο βάσεις, παρατηρήσαμε συνολική ακρίβεια περίπου ίση με 60%. Στο σύνολο SAVEE, παρατηρήθηκαν λίγο χαμηλότερα F-score για τα συναισθήματα φόβος και χαρά, παρόλα αυτά η συνολική ακρίβεια ήταν ίση με του eNTERFACE. Αυτό μάλιστα είναι πολύ ενδιαφέρον καθώς οι παράμετροι του νευρωνικού επιλέχθηκαν για τη καλύτερη απόδοση στο σύνολο eNTERFACE και όχι στο SAVEE. Επιπλέον, στη βάση SAVEE υπήρχε και άλλο ένα συναίσθημα (neutral) για το οποίο το μοντέλο πέτυχε πολύ καλό F-score, ίσο με 0,82. Σημαντικά ενδιαφέρον ήταν επίσης το γεγονός ότι η δεύτερη βάση είναι περίπου τρεις φορές μικρότερη από την πρώτη.

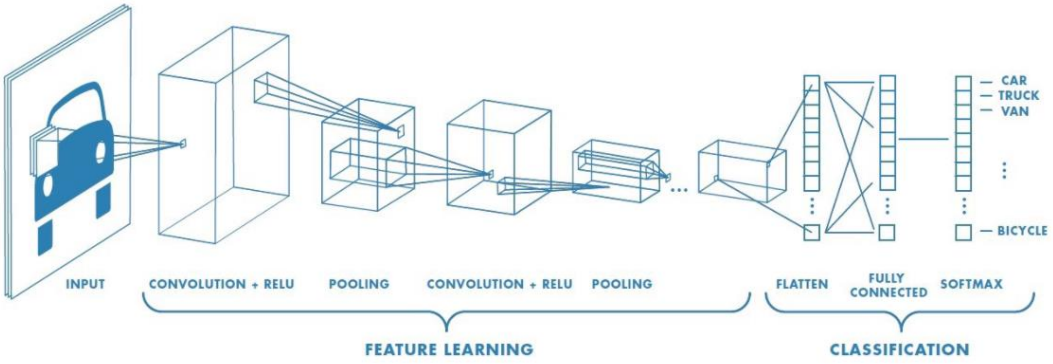
Λέγεται ότι η επίδοση του μοντέλου θα μπορούσε να βελτιωθεί ακόμα περισσότερο καθώς κάθε πλαίσιο ενός δευτερολέπτου κάθε δείγματος είχε το ίδιο label με όλο το δείγμα. Αυτό καταλαβαίνουμε πως δεν αντικατοπτρίζει πλήρως την πραγματικότητα, καθώς σε ένα απόσπασμα ομιλίας μπορούμε να συναντήσουμε διακυμάνσεις στα συναισθήματα του ομιλητή, πόσο μάλλον όταν πρόκειται για

συνομιλία μεταξύ δύο ή και παραπάνω ατόμων. Συνεπώς, εάν χειροκίνητα κάποιος μετονόμαζε κάθε πλαίσιο του δείγματος στο πραγματικό συναίσθημα που υπονοείται, το δίκτυο θα αποκτούσε περισσότερη και περιεκτικότερη πληροφορία. Τέλος, το γεγονός ότι οι συμμετέχοντες των συνόλων δεδομένων δεν ήταν επαγγελματίες ηθοποιοί, μπορεί να αποτέλεσε επιπλέον εμπόδιο στην εκμάθηση του μοντέλου.

Καθώς σύγχρονες τεχνολογικές εφαρμογές απαιτούν υπολογισμούς σε πραγματικό χρόνο ώστε ο υπολογιστής να ανταποκρίνεται άμεσα, δημιουργείται η ανάγκη για πιο απλές μεθόδους τεχνητής νοημοσύνης σε σχέση με τις τυπικές μεθοδολογίες SER. Έτσι, η προσέγγιση που αναλύθηκε καθίσταται ισχυρή υποψήφια για εφαρμογές πραγματικού χρόνου.

2.4.3.2 Deep Convolutional Neural Networks (DCNN)

Τα Βαθιά Συνελικτικά Νευρωνικά Δίκτυα συνήθως αποτελούνται από πολλαπλά επίπεδα συνελικτικών κόμβων, ακολουθούμενα από ένα ή περισσότερα πλήρως συνδεδεμένα επίπεδα για να ολοκληρώσουν τη διαδικασία ταξινόμησης. Παρακάτω απεικονίζεται η βασική δομή ενός DCNN:



Εικόνα 10: Απλό παράδειγμα δομής DCNN[4]

Έχουν γίνει πολλές μελέτες για αναγνώριση συναισθημάτων από ομιλία μέσω DCNN , από τις οποίες θα αναφερθούν κάποιες από τις πιο πρόσφατες. Μία από αυτές είναι η μελέτη των Harar, Burge, Kishore Dutta [23]. Σε αυτή υλοποίησαν το σύστημα πάνω στο Berlin Database of Emotional Speech. Για να είναι δυνατή η σύγκριση με προηγούμενες έρευνες περιόρισαν τις κλάσεις σε θυμό, ουδέτερο και λύπη. Στο σύστημά τους κατήργησαν τη σιωπή από τα σήματά τους και κατάτμησαν τα αρχεία σε κομμάτια 20ms χωρίς επικάλυψη. Στο δίκτυό τους, πριν από την επιλογή χαρακτηριστικών, είχαν έξι επίπεδα συνέλιξης με επιτυχημένα dropout επίπεδα με τιμές p ίσες με 0.1 και ακολουθούσε ένα δικτυωτό (πλέγμα) από δύο παράλληλους επιλογείς χαρακτηριστικών και τέλος σειρές από πλήρως συνδεδεμένα επίπεδα. Η τμηματική ορθότητα του συστήματος ήταν 77.51%, αλλά το κάθε αρχείο σημείωνε ορθότητα 96.97%, με διάστημα εμπιστοσύνης 69.55%. Παρά το υψηλό ποσοστό ορθότητας του συστήματος σε επίπεδο αρχείου, στην πραγματικότητα δεν υπήρχε ένδειξη σημείου διαχωρισμού

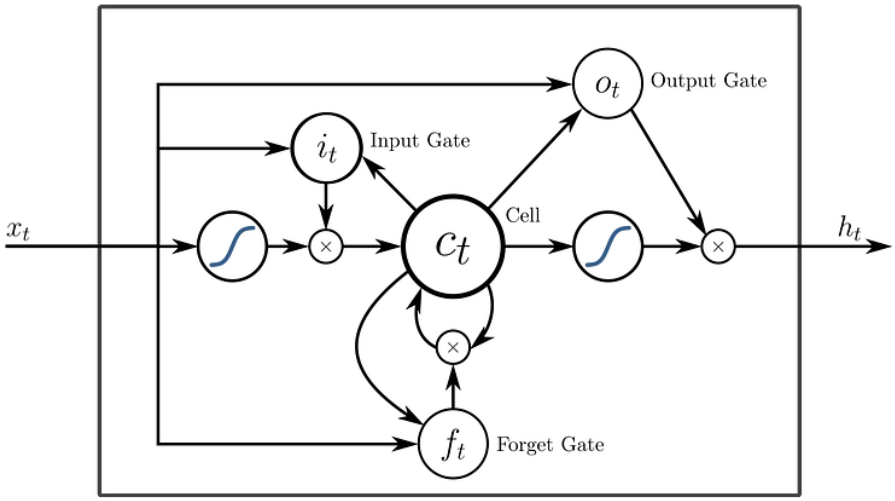
[4] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

της ομιλίας και το σύστημα χρειάζεται βελτίωση στην ανεξάρτητη ανίχνευση. Το 2018 δημοσιεύτηκε άλλη μία έρευνα [24] ενός συστήματος αναγνώρισης συναισθήματος βασισμένη σε DCNN, σχεδιασμένη για τον διαγωνισμό ImageNet LSVRC-2010. Αυτό το δίκτυο, AlexNet, είναι προ-εκπαιδευμένο με ένα σύνολο δεδομένων 1.2 εκατομμυρίων εικόνων και μετά έγινε fine-tune χρησιμοποιώντας δείγματα από το EMO-DB. Το σύστημα αυτό μπορούσε να αναγνωρίσει τέσσερις κλάσεις συναισθημάτων: θυμό, λύπη, χαρά και ουδέτερο. Επιπλέον, έδειξαν ότι το σύστημα τους μπορούσε να φτάσει επίπεδα ορθότητας πάνω από 80% με EMO-DB, περίπου 20% από το πρότυπο SVM. Επίσης, εφάρμοσαν τη μέθοδο αυτή στις βάσεις δεδομένων: EML, eNTERFACE05 και Baum-1s με αποτελέσματα ακόμη καλύτερα από τη βασική μέθοδο. Σε αυτό το σύστημα η προσοχή ήταν στραμμένη στο πως η αυτόματη επιλογή χαρακτηριστικών στα DCNNs μπορεί να αποδώσει καλύτερα από την επιλογή χαρακτηριστικών στα CNNs. Τα Βαθιά Συνελικτικά Νευρωνικά Δίκτυα είναι ισχυρά στη μοντελοποίηση των μικρότερων μεταβολών του σήματος. Ωστόσο, αυτή η ικανότητα έρχεται μαζί με κόστος εκθετικά περισσότερων μεταβλητών προσαρμογής, άρα και ανάγκη περισσότερων δειγμάτων για την εκπαίδευση του συστήματος. Για παράδειγμα σε περιπτώσεις εφαρμογών εικόνας, αυτά τα δίκτυα εκπαιδεύονται με εκατομμύρια δείγματα. Ωστόσο, στην αναγνώριση συναισθημάτων μέσω ομιλίας συνήθως οι αριθμοί των δειγμάτων περιορίζονται σε χιλιάδες. Αυτό, όπως αναφέρθηκε και παραπάνω, καθιστά τα βαθιά συνελικτικά δίκτυα περισσότερο επιρρεπή στην υπερπροσαρμογή.

2.4.3.3 Long Short – Term Memory Networks (LSTMs)

Τα επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks, RNNs) αποτελούν μια κατηγορία τεχνητών νευρωνικών δικτύων όπου οι συνδέσεις μεταξύ κόμβων σχηματίζουν ένα κατευθυνόμενο ή μη κατευθυνόμενο γράφημα κατά μήκος μιας χρονικής ακολουθίας. Μπορούν να μαθαίνουν και να αντιδρούν σε χρονικά συμβάντα χωρίς να αλλάζουν τα αργά διαμορφωμένα βάρη χάρη στη ανάδραση του, που ενεργοποιείται βραχυπρόθεσμα για πρόσφατα συμβάντα. Αυτό το χαρακτηριστικό είναι πολύ σημαντικό για εφαρμογές στις οποίες ο χρόνος είναι ένα απαραίτητο χαρακτηριστικό, όπως η επεξεργασία της ομιλίας (Speech Processing), η σύνθεση μουσικής και η περιγραφή video. Παρ' όλα αυτά όσο εκπαιδεύονταν τα επαναλαμβανόμενα νευρωνικά δίκτυα μέσω οπισθοδιάδοσης στο χρόνο, εσφαλμένα σήματα που έτρεχαν ανάποδα στο χρόνο μπορούσαν να είναι είτε όλο και μεγαλύτερα είτε να εξαφανίζονταν σε σχέση με το μέγεθος των βαρών. Αυτό μπορούσε να προκαλέσει είτε ταλάντωση βαρών, είτε να καταστήσει το δίκτυο αργό ως προς την εκπαίδευση και τη σύγκλιση. Το 1997 προκειμένου να μπορέσουν να συμπεριλάβουν τη βραχυπρόθεσμη προσαρμογή των RNNs και να αποφύγουν τα προαναφερθέντα προβλήματα οι Hochreiter και Schmidhuber [25] παρουσίασαν μια νέα αρχιτεκτονική που την ονόμασαν Long Short-Term Memory (LSTM). Τα δίκτυα LSTM είχαν τη δυνατότητα να «γεφυρώσουν» χρονικά κενά μεγαλύτερα από 1.000 βήματα, ακόμα και όταν η ακολουθία εισόδου είναι ασυμπίεστη και θορυβώδης. Ενσωμάτωσαν έναν αλγόριθμο κλίσης που εξανάγκαζε τη ροή σφαλμάτων να γίνει προς εξατομικευμένες μονάδες, ειδικά σχεδιασμένες για τη διαχείριση του Short-Term. Δηλαδή, κατάφεραν να κόψουν τους υπολογισμούς κλίσης σε ένα καθορισμένο σημείο χωρίς να επηρεαστούν οι Long-Term ενέργειες.

Παρακάτω απεικονίζεται η δομή ενός απλού δικτύου LSTM:



Εικόνα 11: Απλό παράδειγμα δικτύου LSTM. Αποτελείται από τρία δομικά στοιχεία ή θύρες (Input, Output, Forget Gate)[5]

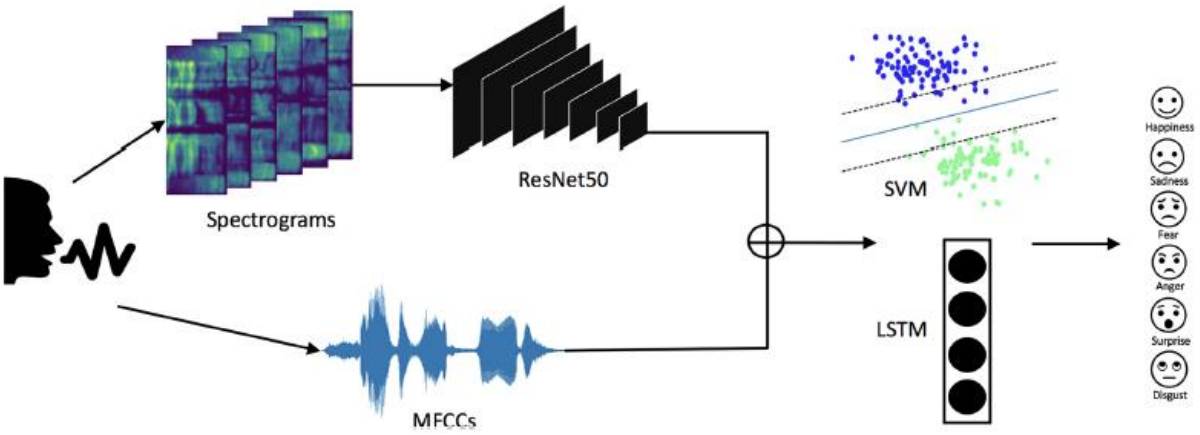
Τα τελευταία χρόνια τα LSTM δίκτυα βρίσκονταν στο επίκεντρο προσοχής των ερευνητών για πολλές εφαρμογές που περιλάμβαναν χρονικές σειρές γεγονότων, όπως η επεξεργασία ομιλίας και η αναγνώριση συναισθήματος μέσω ομιλίας (SER). Μία από τις πρώτες προτάσεις για χρήση LSTM δικτύων έγινε το 2013 με την έρευνα του Martin Wöllmer και άλλων επιστημόνων [26]. Πρότειναν ένα δίκτυο ταξινόμησης βασισμένο σε LSTM τεχνολογία που εκμεταλλευόταν ακουστική, γλωσσική και οπτική πληροφορία. Σε αυτή τη μελέτη σύγκριναν τα αποτελέσματά τους χρησιμοποιώντας «μονοκατευθυνόμενα» και «αμφικατευθυνόμενα» (εκπαίδευση δεδομένων εισόδου δύο φορές, μία με κανονική κατεύθυνση και μία με ανάποδη) δίκτυα LSTM. Επίσης, σύγκριναν τα αποτελέσματά τους με το AVEC 2011 Audio/ Visual Emotion Challege [27]. Σε αυτή τη μελέτη εξήγαγαν 1941 ηχητικά χαρακτηριστικά αποτελούμενα από προσωδικά, φασματικά και χαρακτηριστικά ποιότητας φωνής, γλωσσικό σε επίπεδο λέξεων περιεχόμενο και όλα τα χαρακτηριστικά video που είχαν εξαχθεί εφαρμόζοντας τη μέθοδο Viola-Jones. Μετά όλα τα χαρακτηριστικά τροφοδότησαν το «μονοκατευθυνόμενο» (unidirectional) και το «αμφικατευθυνόμενο» (bidirectional) LSTM δίκτυο.

Το 2016, ο Trigeorgis και οι συνεργάτες του [28], ακολουθώντας την προσέγγιση χρήσης νευρωνικών δικτύων απ'άκρη σ'άκρη, πρότειναν ένα context-aware σύστημα που συνδυάζει ένα δίκτυο CNN, ακολουθούμενο από ένα LSTM. Συγκεκριμένα, το CNN χρησιμοποιείται για την εξαγωγή των πιο κατάλληλων χαρακτηριστικών του σήματος φωνής, τα οποία στη συνέχεια οδηγούνται στο LSTM, δίνοντας σημαντικά καλύτερες επιδόσεις από άλλα συστήματα που προεπιλέγουν τα χαρακτηριστικά πριν την εκπαίδευση. Η συγκεκριμένη οπτική έρχεται σε συμφωνία με την τάση που υπάρχει στην κοινότητα της μηχανικής μάθησης για εξαγωγή μιας αναπαράστασης της εισόδου από μη-επεξεργασμένα δεδομένα («raw» input), παρά από «χειροκίνητα» επιλεγμένα χαρακτηριστικά που απαιτούν ανθρώπινη γνώση. Μια τέτοιου είδους αναπαράσταση αναμένεται ότι θα προσαρμόζεται στις εκάστοτε συνθήκες εφαρμογής, οδηγώντας έτσι σε καλύτερα, context-aware αποτελέσματα.

[5] https://www.unite.ai/what-are-rnns-and-lstms-in-deep-learning/

Ενδιαφέροντα αποτελέσματα παρουσιάζει και η δημοσίευση των Zhao κ.ά. [29], οι οποίοι χρησιμοποίησαν ένα CNN-LSTM δίκτυο μίας διάστασης και ένα CNN-LSTM δύο διαστάσεων για την εξαγωγή τοπικών και συνολικών χαρακτηριστικών από σήματα φωνής και λογαριθμικά φασματογραφήματα Mel, αντίστοιχα. Αμφότερα τα δίκτυα έχουν παρόμοια δομή, καθώς και τα δύο αποτελούνται από τέσσερα block εκμάθησης τοπικών χαρακτηριστικών (local feature learning blocks – LFLBs) ακολουθούμενα από ένα επίπεδο LSTM. Τα LFLBs, που κατά κύριο λόγο περιέχουν ένα συνελικτικό επίπεδο και ένα επίπεδο max-pooling για μείωση διαστάσεων, αποσκοπούν στην ανεύρεση τοπικών συσχετίσεων, ενώ το επίπεδο LSTM χρησιμεύει στην ανίχνευση πιο ευρύτερων χρονικά εξαρτήσεων μεταξύ των χαρακτηριστικών. Τα δύο διαφορετικών διαστάσεων δίκτυα που σχεδιάστηκαν, εφαρμόστηκαν σε δύο γνωστά σύνολα δεδομένων, το EMO-DB και το IEMOCAP, και δείχνουν να συνδυάζουν τα πλεονεκτήματα και να ξεπερνούν τα μειονεκτήματα που θα είχε ένα CNN και ένα LSTM ξεχωριστά. Τα αποτελέσματα καταδεικνύουν υπεροχή έναντι των καθιερωμένων προσεγγίσεων, με το δισδιάστατο δίκτυο – που απέδωσε καλύτερα – να επιτυγχάνει συνολική ορθότητα 95.89% και 52.14% στο EMO-DB και στο IEMOCAP, αντίστοιχα, σε ανεξάρτητα του ομιλητή (speaker-independent) πειράματα.

Πρόσφατα, η Araño και οι συνεργάτες της [5], επιχείρησαν να συνδυάσουν τους συνηθισμένους συντελεστές MFCC με χαρακτηριστικά εικόνας εξαγόμενα από φασματογραφήματα με τη χρήση του προεκπαιδευμένου συνελικτικού δικτύου Resnet50, αποφεύγοντας έτσι την κοπιώδη και απαιτητική σε υπολογιστικούς πόρους διαδικασία της εκπαίδευσης ενός νευρωνικού δικτύου. Η γενική δομή της προτεινόμενης προσέγγισης απεικονίζεται στο παρακάτω διάγραμμα. Πιο συγκεκριμένα, οι συγγραφείς ανέπτυξαν συνολικά έξι μοντέλα, χρησιμοποιώντας ως χαρακτηριστικά είτε μόνο τους MFCCs, είτε μόνο τα εξαγόμενα μέσω βαθιάς μάθησης χαρακτηριστικά, είτε τον συνδυασμό τους («Hybrid») και ως ταξινομητή ένα SVM ή ένα LSTM. Τα μοντέλα δοκιμάστηκαν στο σύνολο δεδομένων RAVDESS, με την καλύτερη ορθότητα να προκύπτει από το σύστημα MFCC-LSTM με 73,5%, ακολουθούμενο από το Hybrid-SVM με 71,3%. Αυτό που κατέδειξε η έρευνα είναι ότι οι MFCC, παρότι συμβατικοί, δύνανται να παρέχουν state-of-the-art επιδόσεις, ενώ υπενθυμίζεται και η αποτελεσματικότητα των LSTM σε εφαρμογές όπου παίζει ρόλο ο χρόνος, όπως η αναγνώριση φωνής. Επιπλέον, με βάση το μοντέλο MFCC-LSTM αναπτύχθηκε μία εφαρμογή πραγματικού χρόνου, με το πλεονέκτημα του μειωμένου χρόνου επεξεργασίας, καθώς η εξαγωγή και η ανάλυση των MFCC απαιτεί ελάχιστη υπολογιστική ισχύ. Παρακάτω απεικονίζεται διαγραμματικά το προτεινόμενο σύστημα:



Εικόνα 12: Διάγραμμα του προτεινόμενου συστήματος στο [28]

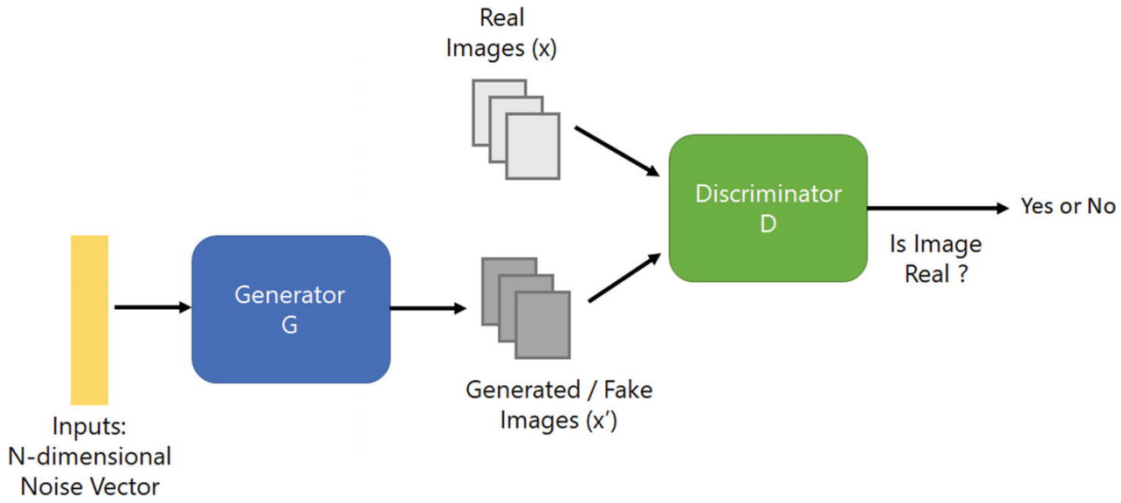
2.4.3.4 Generative Adversarial Networks (GANs)

Τα Παραγωγικά Αντιπαραθετικά Δίκτυα (Generative Adversarial Networks) είναι μια κατηγορία συστημάτων μηχανικής μάθησης και θεωρούνται από τα σημαντικότερα εργαλεία για ενίσχυση και αναπαράσταση δεδομένων και για αποθορυβοποίηση. Εφευρέθηκαν από τον I.Goodfellow και τους συναδέλφους του το 2014 [30]και βασίζονται στη λογική της αντιπαραθετικής μάθησης. Δηλαδή, δύο νευρωνικά δίκτυα διαγωνίζονται σε ένα παίγνιο. Το παραγωγικό δίκτυο(G) δημιουργεί υποψηφίους και το διαχωριστικό δίκτυο (D) τους αξιολογεί. Πιο συγκεκριμένα, ο γεννήτορας G(z) παίρνει μια είσοδο z, ένα δείγμα κατανομής πιθανότητας P(z) και παράγει συνθετικά δεδομένα. Από την άλλη, ο «διαχωριστής» παίρνει την πληροφορία και καθορίζει ποια δεδομένα εισόδου είναι πραγματικά και ποια παραγόμενα. Τελικά, τα δίκτυα φτάνουν σε μια ισορροπία κατά τρόπο που υπάρχει μια συνάρτηση τιμής, την οποία ο ένας θέλει να μεγιστοποιήσει και ο άλλος να ελαχιστοποιήσει, όπως φαίνεται στην παρακάτω σχέση:

$$min_G max_D V(D,G) = E_{x\sim P(x)}[\log D(x)] + E_{z\sim P(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right]$$

όπου D(x) και D(G(z)) είναι οι πιθανότητες το x και το G(z) να αναγνωριστούν ως πραγματικά δείγματα από τον διαχωριστή.

Παρακάτω απεικονίζεται η δομή ενός απλού GAN:



Εικόνα 13: Παράδειγμα απλού δικτύου GAN[6]

Σε μία δημοσίευση του 2019, η Chatziagapi κ.ά. [31] αντιμετώπισαν το πρόβλημα των μη-ισορροπημένων δεδομένων, όπου δηλαδή τα μη-ουδέτερα συναισθήματα εμφανίζονται λιγότερο συχνά στα σύνολα δεδομένων, οδηγώντας σε ανεπαρκείς επιδόσεις. Προσάρμοσαν και βελτίωσαν ένα conditional Generative Adversarial Network (GAN) ώστε να δημιουργήσουν τεχνητά

[6] https://link.springer.com/chapter/10.1007/978-1-4842-3679-6_8

φασματογραφήματα για τα υποεκπροσωπούμενα συναισθήματα και σύγκριναν τα αποτελέσματα με άλλες πιο συμβατικές μεθόδους επαύξησης δεδομένων. Ειδικότερα, η προτεινόμενη μέθοδος αξιολογήθηκε στα σύνολα δεδομένων IEMOCAP και FEEL-25k, πετυχαίνοντας αυξημένη επίδοση κατά 10% και 5%, αντίστοιχα, σε σύγκριση με τις άλλες μεθόδους.

Ένα από τα σημαντικότερα μειονεκτήματα αυτής της προσέγγισης είναι ότι η σύγκλιση εξαρτάται σε μεγάλο βαθμό από τα δεδομένα και την αρχικοποίηση. Ωστόσο, για να προσπεραστεί αυτός ο περιορισμός και για γρηγορότερη σύγκλιση, το GAN αρχικοποιείται χρησιμοποιώντας έναν προ-εκπαιδευμένο αυτόματο κωδικοποιητή (autoencoder), τύπος τεχνητού νευρωνικού δικτύου που χρησιμοποιείται για την εκμάθηση αποτελεσματικών κωδικοποιήσεων μη επισημασμένων δεδομένων) [2].

3. Συμπεράσματα

Σε αυτήν την ενότητα, συνοψίζονται οι μεθοδολογίες και τα σύνολα δεδομένων που μελετήθηκαν παραπάνω, προκειμένου να κατασκευαστεί μία έκθεση σύγκρισης μεταξύ των τεχνολογιών που υπάρχουν και να φανεί η αποτελεσματικότητά της κάθε περίπτωσης. Αρχικά, γίνεται αναφορά στα σύνολα δεδομένων που αξιολογήθηκαν στη μελέτη και παρουσιάζονται κάποια συμπεράσματα για τις μελέτες που προηγήθηκαν. Στη συνέχεια, αναφέρονται οι δυσκολίες που αντιμετωπίζονται στον κλάδο γενικότερα και τέλος παρατίθενται κατευθύνσεις που πιθανότατα θα απασχολήσουν μελλοντικές έρευνες. Υπογραμμίζεται ότι στην εργασία χρησιμοποιήθηκε ως κύριο μέτρο απόδοσης και αξιολόγησης των ερευνών η τιμή ορθότητας που κατάφεραν να πετύχουν, αλλά στην πραγματικότητα δε μπορεί μόνο αυτή η τιμή να καθορίσει την επιτυχία ή αποτυχία μιας έρευνας

3.1 Ανασκόπηση Αποτελεσμάτων

Όπως είναι φανερό και στον πίνακα παρακάτω, γενικά η EMO-DB ήταν η βάση δεδομένων που χρησιμοποιήθηκε περισσότερο, ενώ ακολουθούν οι RAVDESS, IEMOCAP, eNTERFACE και SAVEE. Παρά ταύτα, η IEMOCAP είναι εκείνη που προτείνεται στις πιο νέες μεθόδους καθώς έχει το μεγαλύτερο πλήθος δειγμάτων, πράγμα που συντελεί σημαντικά στην εκπαίδευση πιο περίπλοκων αρχιτεκτονικών.

Επιπλέον, παρατίθεται και μία συνολική σύγκριση των μεθόδων που ακολουθήθηκαν οργανωμένες ως προς τη βιβλιογραφία, τη μέθοδο που ακολούθησαν, τα features που εκμεταλλεύτηκαν, τα σύνολα δεδομένων που χρησιμοποιήθηκαν και τη συνολική ακρίβεια των μοντέλων. Για τη σύγκριση των εκθέσεων που μελετήθηκαν, ιδιαίτερο ενδιαφέρον συγκεντρώνουν μη βεβαρυμμένα μεγέθη που συνοψίζουν την επίδοσή τους, όπως τη μετρική F_1. Επιπλέον, κάποιες από αυτές τις εκθέσεις περιείχαν πολλά μεγέθη ορθότητας και για αυτό το λόγο αποφασίστηκε να παρουσιαστούν μόνο οι καλύτερες τιμές ορθότητας για κάθε σύνολο δεδομένων που χρησιμοποιήθηκε.

Όπως φαίνεται και παρακάτω, το χαρακτηριστικό που χρησιμοποιήθηκε περισσότερο ήταν ο συντελεστής MFCC.

Είναι επιπλέον προφανές ότι οι παραδοσιακές μέθοδοι SER (SVM, HMM κ.λπ.) βασίζονταν πάρα πολύ στην επεξεργασία των ηχητικών σημάτων, ενώ οι πιο πρόσφατες έρευνες εστιάζουν στη βαθιά μάθηση και στη βελτιστοποίηση των νευρωνικών δικτύων, πράγμα που συνδέεται άρρηκτα με την εξέλιξη των

υπολογιστικών πόρων σήμερα. Με την πρόοδο που έχει επιτευχθεί στο λογισμικό και το hardware των ηλεκτρονικών υπολογιστών, οι ερευνητές έχουν καταφέρει να επιστρατεύσουν πολύ περίπλοκα δίκτυα όπως LSTMs, GANs, και VAEs. Παράλληλα, η μεγάλη εκτόξευση της έρευνας στον τομέα των CNNs δείχνει ότι τα αναπτυσσόμενα συστήματα βελτιώνουν την ικανότητά τους στο να εντοπίζουν μικροδιαφορές. Η ενσωμάτωση και των αρχιτεκτονικών βαθιά συνελικτικών LSTM έδωσε στο δίκτυο μακροπρόθεσμη μνήμη, και άρα τη δυνατότητα να μπορεί να ταυτοποιήσει τα μακροπρόθεσμα παραγλωσσικά σχήματα.

Πίνακας 5: Μία συνολική συγκριτική αποτίμηση των ερευνών που αναλύθηκαν

Τίτλος δημοσίευσης	**Μέθοδος**	**Χαρακτηριστικά**	**Σύνολο δεδομένων και αποτελέσματα**
Speech Emotion Recognition Using Support Vector Machines, Chavhan et al., 2010 [14]	• SVM	• MFCC • MEDC	• EMO-DB: 93.75%
Emotion Recognition and Classification in Speech using Artificial Neural Networks, Shaw et al., 2016 [15]	• ANN	• Ενέργεια, Τόνος • 20 MFCCs	• 86.87%
Emotion recognition from Marathi speech database using adaptive artificial neural network, Darekar, and Dhande, 2018 [16]	• ANN • PSO-FF	• MFCC • NMF • Τόνος	• RAVDESS: 88.7%
A First Look Into A Convolutional Neural Network For Speech Emotion Detection, Bertero, and Fung, 2017 [19]	• CNN	• PCM	• TEDLIUM2: 66.1%
Deep Neural Networks for Acoustic Emotion Recognition: Raising The Benchmarks, Stuhlsatz et al., 2011 [22]	• GerDA • RBM	• ZCS • F_0(Hz) • Πιθανότητα ηχηροποίσης • Διάφορες ζώνες φασματικής ενέργειας 25%, 50%, 75%, 90% roll of point • MFCCs	• EMO-DB: 85.6% • eNTERFACE: 72.4% • ABC: 61.5% • SUSAS: 56.5% • AVIC: 79.1% • DES: 60.1% • SAL: 34.3% • SmartKom: 59.5% • VAM: 68.0%
Towards real-time Speech Emotion Recognition using deep neural networks [4]	• DNN	• Φασματογραφικά χαρακτηριστικά	• eNTERFACE: 60,53% • SAVEE: 59,7%
Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching, Zhang et al., 2018 [24]	• DCNN (Alex-Net) • DTPM • SVM	• LMS • Delta • Delta Delta	• EMO-DB: 87.31% •]RML: 75.34% • eNTERFACE05:79.25% • BAUM-1s: 44.61%
LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework, Wöllmer et al., 2013 [26]	• LSTM • BLSTM	• Ένταση, ZCR, Ενέργεια εντός 250-600Hz, 1-4kHz	• RECOLA: 68.4%

		• 25%, 50%, 75% και 90% φασματικών roll of point, εντροπία, διασπορά • Ψυχοακουστική οξύτητα, αρμονία, 10 MFCCs • F_0 • Ηχηρότητα, jitter • logHNR (Logarithmic Harmonics-to-Noise Ratio)	
Adieu Features? End-To-End Speech Emotion Recognition Using A Deep Convolutional Recurrent Network, Trigeorgis et al., 2016 [28]	• DCNN, LSTM	• PCM	• RECOLA: 68.4%
Speech Emotion Recognition using deep 1D & 2D CNN LSTM networks, Zhao et al., 2019 [29]	• DCNN, LSTM	• PCM • Log-Mel Spectrogram	• EMO-DB: 95.33% • IEMOCAP: 86.16%
When Old Meets New: Emotion Recognition from Speech Signals, Araño et al., 2021 [5]	• SVM • LSTM	• MFCC • Χαρακτηριστικά εξαγόμενα μέσω DCNN από φασματογραφήματα • Συνδυασμός των παραπάνω	• RAVDESS: 73.5%
Data Augmentation Using GANs for Speech Emotion Recognition, Chatziagapi et al., 2019 [31]	• DCNN, GAN	• 128 MFCCs	• IEMOCAP: 53.6% • Feel-25k: 54.6%

Αξιολογώντας τα παραπάνω αποτελέσματα, γίνεται αντιληπτό πως κάποια από αυτά ξεπερνούν πολύ μεγάλα ποσοστά ορθότητας, της τάξης του 90%. Αυτές οι μελέτες χρησιμοποιούν παλαιότερες βάσεις δεδομένων όπως την EMO-DB και τη DES, οι οποίες υστερούν σε πλήθος δειγμάτων. Αυτό πιθανώς σημαίνει ότι έχει γίνει υπερπροσαρμογή (overfitting) σε αυτά τα μοντέλα. Στην περίπτωση των Zhao κ.ά. [29], το σύστημα έχει περίπου 2.500.000 μεταβλητές που πρέπει να ρυθμιστούν βασισμένες σε 535 προτάσεις της EMO-DB. Το κενό είναι πολύ μεγάλο για να αποφευχθεί η υπερπροσαρμογή.

Επιπλέον, συγκρίνοντας τις ορθότητες των μεθόδων βαθιάς μάθησης που μελετήθηκαν, βασισμένες σε EMO-DB και IEMOCAP, υπάρχει σημαντική διαφορά. Αυτό μπορεί να οφείλεται και πάλι στο πολύ μικρότερο πλήθος δειγμάτων που υπάρχει στην EMO-DB σε σχέση με την IEMOCAP, πράγμα που είναι πολύ περιοριστικό στην εκμάθηση ενός βαθιού νευρωνικού δικτύου.

Επιπλέον, οι σχετικά παλιότερες μέθοδοι βαθιάς μάθησης, ως αποτέλεσμα της χρήσης ταξινομητών που χρησιμοποιούν εξαγωγή χαρακτηριστικών και ψηφιακή επεξεργασία σήματος, έχουν χαμηλότερες ορθότητες. Ωστόσο, στις τελευταίες έρευνες, η μέση ορθότητα έχει αυξηθεί. Το πρόβλημα της υπερπροσαρμογής και της ευαισθησίας σε θόρυβο παραμένει και, τα πιο πρόσφατα έτη, έχουν γίνει μελέτες για να αντιμετωπιστεί και αυτό το ζήτημα.

Άλλη μία σημαντική παρατήρηση είναι ότι απ' ότι φαίνεται δεν υπάρχει κάποια σχέση που να συνδέει την πολυπλοκότητα του συνόλου των χαρακτηριστικών και τις ορθότητες που αναφέρονται, και οι μέθοδοι που προτείνονται έχουν σημαντικό αντίκτυπο στα αποτελέσματα. Στη μελέτη των Harar κ.ά. [23], χρησιμοποιώντας EMO-DB, το σύνολο των χαρακτηριστικών είναι απλά δείγματα PCM των αρχείων ήχου, και η ορθότητα είναι 96,97%. Αντιθέτως, στη μελέτη των Song κ.ά. [32], υπάρχει ένα πιο περίπλοκο σύνολο χαρακτηριστικών και η ορθότητα του μοντέλου ήταν 59,8%. Στο παρόμοιο μοντέλο των Zhao κ.ά. [29], η πολύ μεγαλύτερη βάση IEMOCAP κατάφερε ορθότητα ίση με 86,16% ενώ στη μελέτη των Eskimez κ.ά. [33], ορθότητα 71,2%.

3.2 Προκλήσεις

Παρά το γεγονός ότι παρατηρείται μεγάλη πρόοδος στις μεθόδους που χρησιμοποιούνται, υπάρχουν ακόμα μερικοί περιορισμοί που πρέπει να ξεπεραστούν προκειμένου να υπάρξει ακόμα μεγαλύτερη βελτίωση των μοντέλων στο μέλλον. Η πιο σημαντική από αυτές είναι η διαθεσιμότητα συνόλων δεδομένων σχεδιασμένα για βαθιά μάθηση, καθώς τα υπάρχοντα σύνολα δεν διαθέτουν αρκετά μεγάλο πλήθος δειγμάτων για την εκπαίδευση βαθιών αρχιτεκτονικών. Η αναγνώριση συναισθημάτων μέσω ομιλίας υστερεί ακόμα σε σύνολα δεδομένων σε σχέση με άλλα συναφή πεδία όπως την αναγνώριση φωνής και την αναγνώριση εικόνας, τα οποία διαθέτουν βάσεις δεδομένων, όπως το ImageNet με 14 εκατομμύρια δείγματα και το Google Audioset με 2,1 εκατομμύρια δείγματα, αντίστοιχα. Επιπλέον, τα περισσότερα σύγχρονα μοντέλα που κατασκευάζονται χρησιμοποιούν ημι-φυσικά και προσομοιωμένα σύνολα δεδομένων, τα οποία είναι σκηνοθετημένα και απαλλαγμένα από θορύβους, γεγονός που τα καθιστά μη ρεαλιστικά. Συνεπώς, τα μοντέλα που εκπαιδεύονται με αυτά τα σύνολα δεν μπορούν να έχουν μεγάλο ποσοστό επιτυχίας σε πραγματικά σενάρια. Αν και υπάρχουν πραγματικά σύνολα δεδομένων, των οποίων τα δείγματα έχουν εξαχθεί από τηλεοπτικές μεταδόσεις και τηλεφωνικά κέντρα, λόγω του ότι καταγράφονται σε αυτά τα πλαίσια, δεν περιέχουν αρκετές κατηγορίες συναισθημάτων.

Ένα άλλο μεγάλο εμπόδιο στο αντικείμενο αποτελεί η εξάρτηση των μοντέλων από τον πολιτισμό και τη γλώσσα, δύο παράγοντες που καθορίζουν τόσο το συναίσθημα που υπονοείται όσο και τον τρόπο που εκλαμβάνεται αυτό από τον συνομιλητή. Ένα ικανοποιητικό μοντέλο αναγνώρισης συναισθημάτων χρειάζεται χαρακτηριστικά που δεν εξαρτώνται από αυτούς τους παράγοντες και τα εξαγόμενα χαρακτηριστικά των μοντέλων που μελετήθηκαν μπορεί να μην είναι αρκετά. Επιπλέον, η επισήμανση των δειγμάτων είναι εν γένει μία αβέβαιη διαδικασία και βασίζεται στην υποκειμενική κρίση του ατόμου που τα ταξινομεί. Αυτό δεν συμβαίνει σε άλλα πεδία όπως την αναγνώριση εικόνας καθώς τις περισσότερες φορές η αναπαράσταση ενός αντικειμένου σε μία εικόνα είναι ξεκάθαρη και δεν εξαρτάται από το πώς την εκλαμβάνει ο παρατηρητής. Αυτή η υποκειμενικότητα που χαρακτηρίζει τη διαδικασία της ετικετοποίησης κάνει την διαδικασία πιο περίπλοκη και περιορίζει την πιθανότητα συνένωσης των συνόλων δεδομένων προς κατασκευή μεγαλύτερων συνόλων δεδομένων που αφορούν τα συναισθήματα [2].

Επιπλέον, ο πραγματικός προφορικός λόγος είναι συνεχόμενος περιέχει εναλλαγές στη ροή του, ενώ υπάρχουν επικαλύψεις των προτάσεων μεταξύ των ομιλητών. Τα σύνολα που διατίθενται απ' την άλλη, περιέχουν μεμονωμένες προτάσεις λόγου, πολύ ξεκάθαρες, χωρίς θόρυβο και εναλλαγές στο συναισθηματισμό του ομιλητή. Για αυτό το λόγο, πρέπει τα μοντέλα που σχεδιάζονται για πραγματική χρήση να μπορούν να χειριστούν και την εναλλαγή διαφόρων συναισθημάτων σε μία φράση.

3.3 Μελλοντικές Κατευθύνσεις

Προκειμένου να επιλυθεί το πρόβλημα της αναγνώρισης συναισθημάτων από ομιλία, πρέπει να αντιμετωπιστούν οι προκλήσεις που αναφέρθηκαν στην προηγούμενη ενότητα. Όσον αφορά το βασικό ζήτημα του μεγέθους των υφιστάμενων συνόλων δεδομένων, η προφανής λύση είναι η δημιουργία ενός νέου συνόλου, με σημαντικά μεγαλύτερο πλήθος δειγμάτων, που θα διευκολύνει την εκπαίδευση περίπλοκων αρχιτεκτονικών βαθιάς μάθησης. Ωστόσο, το κόστος ενός τέτοιου εγχειρήματος καθιστά τη λύση αυτή λιγότερο συμφέρουσα. Μία άλλη πρόταση θα ήταν ο συνδυασμός ορισμένων από τα τωρινά σύνολα δεδομένων, με σκοπό την απόκτηση ενός μεγαλύτερου με τη δυσκολία όμως να εντοπίζεται στις διαφορετικές μεθόδους και τεχνικές που χρησιμοποιήθηκαν για τη δημιουργία κάθε συνόλου, γεγονός που δυσχεραίνει την απόπειρα συνδυασμού. Από την άλλη, η κατασκευή ενός εξ ολοκλήρου συνθετικού συνόλου δεδομένων αξιοποιώντας παραγωγικές μεθόδους όπως τα GANs, συνιστά μια ενδιαφέρουσα λύση που δεν έχει εξερευνηθεί.

Άλλο ένα ζήτημα προς αντιμετώπιση, είναι οι διαφορές στην έκφραση συναισθημάτων σε διαφορετικές γλώσσες. Ενδεχομένως, με τη χρήση transformers μπορεί να αναπτυχθεί κάποιο μοντέλο που θα επιτυγχάνει ικανοποιητικές επιδόσεις ανεξάρτητα από το γλωσσικό περιβάλλον.

Επιπρόσθετα, όπως προαναφέρθηκε πολλά από τα δεδομένα που χρησιμοποιούνται είναι προσομοιωμένα και άρα διαφέρουν σημαντικά από τις πραγματικές συνθήκες, που χαρακτηρίζονται από έντονο θόρυβο. Όπως έχει επιχειρηθεί και παλιότερα, μπορεί να γίνει προσθήκη θορύβου στα δείγματα, με στόχο την ανάπτυξη ενός πιο ανθεκτικού σε πραγματικές συνθήκες συστήματος.

Επιπρόσθετα, μελλοντικό στόχο αποτελεί η επιτυχής ταξινόμηση συναισθημάτων σε κατάσταση συνεχούς ομιλίας. Για το σκοπό αυτό, θα μπορούσε να γίνεται διαχωρισμός του σήματος με κάποιο κυλιόμενο παράθυρο και να εξετάζεται το συναισθηματικό περιεχόμενο σε κάθε προκύπτον τμήμα [2].

4. Βιβλιογραφία

[1] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: A review,” *International Journal of Speech Technology*, vol. 15, no. 2. 2012. doi: 10.1007/s10772-011-9125-1.

[2] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, “Deep learning techniques for speech emotion recognition, from databases to models,” *Sensors (Switzerland)*, vol. 21, no. 4. MDPI AG, pp. 1–27, Feb. 02, 2021. doi: 10.3390/s21041249.

[3] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, “Speech Emotion Recognition Using Deep Learning Techniques: A Review,” *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2936124.

[4] H. M. Fayek, M. Lech, and L. Cavedon, “Towards real-time Speech Emotion Recognition using deep neural networks,” 2015. doi: 10.1109/ICSPCS.2015.7391796.

[5] K. A. Araño, P. Gloor, C. Orsenigo, and C. Vercellis, “When Old Meets New: Emotion Recognition from Speech Signals,” *Cognitive Computation*, vol. 13, no. 3, 2021, doi: 10.1007/s12559-021-09865-2.

[6] P. Ekman, “‘Basic Emotions’, Handbook of Cognition and Emotion.,” *T. Dalgleish and M. Power (Eds.). Sussex, U.K.: John Wiley & Sons, Ltd.*, vol. 39, no. 1, 1999.

[7] O. Martin, I. Kotsia, B. Macq, and I. Pitas, “The eNTERFACE’05 Audio-Visual emotion database,” 2006. doi: 10.1109/ICDEW.2006.145.

[8] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, 2009, doi: 10.1109/TPAMI.2008.52.

[9] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, “Analysis of emotional speech—a review,” in *Intelligent Systems Reference Library*, vol. 105, 2016. doi: 10.1007/978-3-319-31056-5_11.

[10] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Communication*, vol. 116. 2020. doi: 10.1016/j.specom.2019.12.001.

[11] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, “A review on emotion recognition using speech,” 2017. doi: 10.1109/ICICCT.2017.7975169.

[12] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” 2017. doi: 10.1109/ICASSP.2017.7952552.

[13] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” 2014. doi: 10.21437/interspeech.2014-57.

[14] Y. Chavhan, M. L. Dhore, and P. Yesaware, “Speech Emotion Recognition using Support Vector Machine,” *International Journal of Computer Applications*, vol. 1, no. 20, 2010, doi: 10.5120/431-636.

[15] A. Shaw, R. Kumar, and S. Saxena, “Emotion Recognition and Classification in Speech using Artificial Neural Networks,” *International Journal of Computer Applications*, vol. 145, no. 8, 2016, doi: 10.5120/ijca2016910710.

[16] R. V. Darekar and A. P. Dhande, “Emotion recognition from Marathi speech database using adaptive artificial neural network,” *Biologically Inspired Cognitive Architectures*, vol. 23, 2018, doi: 10.1016/j.bica.2018.01.002.

[17] K. Bhatnagar and S. C. Gupta, “Extending the neural model to study the impact of effective area of optical fiber on laser intensity,” *International Journal of Intelligent Engineering and Systems*, vol. 10, no. 4, 2017, doi: 10.22266/ijies2017.0831.29.

[18] J. Weng, N. Ahuja, and T. S. Huang, “Cresceptron: a self-organizing neural network which grows adaptively,” 2003. doi: 10.1109/ijcnn.1992.287150.

[19] D. Bertero and P. Fung, “A first look into a Convolutional Neural Network for speech emotion detection,” 2017. doi: 10.1109/ICASSP.2017.7953131.

[20] S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, “Negative Emotion Recognition using Deep Learning for Thai Language,” 2020. doi: 10.1109/ECTIDAMTNCON48261.2020.9090768.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, 2017, doi: 10.1145/3065386.

[22] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, “Deep neural networks for acoustic emotion recognition: Raising the benchmarks,” 2011. doi: 10.1109/ICASSP.2011.5947651.

[23] P. Harar, R. Burget, and M. K. Dutta, “Speech emotion recognition with deep learning,” 2017. doi: 10.1109/SPIN.2017.8049931.

[24] S. Zhang, S. Zhang, T. Huang, and W. Gao, “Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching,” *IEEE Transactions on Multimedia*, vol. 20, no. 6, 2018, doi: 10.1109/TMM.2017.2766843.

[25] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, 1997, doi: 10.1162/neco.1997.9.8.1735.

[26] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, “LSTM-modeling of continuous emotions in an audiovisual affect recognition framework,” *Image and Vision Computing*, vol. 31, no. 2, 2013, doi: 10.1016/j.imavis.2012.03.001.

[27] M. Gnjatović and D. Rösner, “Inducing genuine emotions in simulated speech-Based human-Machine interaction: The NIMITEK corpus,” *IEEE Transactions on Affective Computing*, vol. 1, no. 2, 2010, doi: 10.1109/T-AFFC.2010.14.

[28] G. Trigeorgis *et al.*, “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2016, vol. 2016-May. doi: 10.1109/ICASSP.2016.7472669.

[29] J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks,” *Biomedical Signal Processing and Control*, vol. 47, 2019, doi: 10.1016/j.bspc.2018.08.035.

[30] I. J. Goodfellow *et al.*, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, 2014, vol. 3, no. January. doi: 10.3156/jsoft.29.5_177_2.

[31] A. Chatziagapi *et al.*, “Data augmentation using GANs for speech emotion recognition,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, vol. 2019-September. doi: 10.21437/Interspeech.2019-2561.

[32] P. Song, Y. Jin, L. Zhao, and M. Xin, “Speech emotion recognition using transfer learning,” in *IEICE Transactions on Information and Systems*, 2014, vol. E97-D, no. 9. doi: 10.1587/transinf.2014EDL8038.

[33] S. E. Eskimez, Z. Duan, and W. Heinzelman, “Unsupervised learning approach to feature analysis for automatic speech emotion recognition,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018, vol. 2018-April. doi: 10.1109/ICASSP.2018.8462685.