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

Αναγνώριση Άγχους σε Σήματα Φωνής

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΓΕΩΡΓΙΟΣ ΜΙΧΑΗΛ ΠΑΝΤΑΖΟΠΟΥΛΟΣ

Επιβλέπων : Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής

Αθήνα, Ιούνιος 2018

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

Αναγνώριση Άγχους σε Σήματα Φωνής

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΓΕΩΡΓΙΟΣ ΜΙΧΑΗΛ ΠΑΝΤΑΖΟΠΟΥΛΟΣ

Επιβλέπων : Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 25η Ιουνίου 2018.

Αλέξανδρος Ποταμιάνος Αναπληρωτής Καθηγητής	Γεώργιος Στάμου Αναπληρωτής Καθηγητής	Κωσταντίνος Τζαφέστας Επίκουρος Καθηγητής

Αθήνα, Ιούνιος 2018

Γεώργιος Μιχαήλ Πανταζόπουλος

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

Στη γιαγιά μου, Θεοδώρα και στους αδελφούς μου, Αλέξανδρο και Ηλία.

Περίληψη

Το συναίσθημα θεωρείται αναπόσπαστο κομμάτι της ανθρώπινης επικοινωνίας καθώς καθορίζει σε μεγάλο βαθμό την αντίληψη της μεταδιδόμενης πληροφορίας από τον πομπό στον δέκτη. Μέχρι σήμερα έχει σημειωθεί σημαντική πρόοδος στον τομέα της αλληλεπίδρασης μεταξύ ανθρώπου - μηχανής, προσομοιώνοντας την διεπαφή μεταξύ των ανθρώπων. Έτσι δεδομένου της σημαντικότητας του συναισθήματος στην ανθρώπινη επικοινωνία, κρίνεται αναγκαία η έρευνα για την υπολογιστική Αναγνώριση Συναισθήματος. Στην εργασία αυτή εξετάζεται η Αναγνώριση Άγχους μέσω Σημάτων Φωνής με σκοπό την δημιουργία αναπαραστάσεων ικανών να περιγράψουν ένα σήμα φωνής και στη συνέχεια η μοντελοποίηση συστημάτων για την ορθή ταξινόμησή τους.

Στον ευρύτερο κλάδο της Αναγνώρισης Συναισθήματος εξάγονται ακουστικά χαρακτηριστικά από το σήμα φωνής με στόχο την κατηγοριοποίησή του σε μία δυνατή κλάση. Με την πάροδο του χρόνου έχει αναπτυχθεί πληθώρα υπολογιστικών μοντέλων για την κατηγοριοποίηση των σημάτων. Σε παλαιότερες μελέτες θεωρείται πως το συναίσθημα εκφράζεται μονοσήμαντα σε ολόκληρο το μήκος του σήματος. Ωστόσο συχνά στην ανθρώπινη επικοινωνία το συναίσθημα εντοπίζεται σε μεμονωμένα τμήματα του σήματος φωνής. Κατά συνέπεια ενδέχεται η Αναγνώριση Συναισθήματος να επωφελείται από την ανάλυση των σημάτων μέσω μικρότερων τμημάτων. Επιπλέον το σύνολο των εξαγόμενων χαρακτηριστικών συμπυκνώνει την εμπειρική γνώση του ανθρώπου για την μοντελοποίηση των χρήσιμων ιδιοτήτων του σήματος προς κατηγοριοποίηση. Δεδομένου της απουσίας του μαθηματικού φορμαλισμού των γνώσεων του ανθρώπου για το συναίσθημα, υπάρχει έμφυτη αμφιβολία στην επίδοση μιας υπολογιστικής μηχανής.

Στο πρώτο σκέλος της εργασίας εξετάζεται η κατηγοριοποίηση των σημάτων μέσω κατακερματισμού σε μικρότερα τμήματα. Αρχικά ακολουθείται η παλαιότερη προσέγγιση εξάγοντας ακουστικά χαρακτηριστικά σε ολόκληρο το σήμα φωνής και έπειτα κατασκευάζεται ένα αρχικό μοντέλο ταξινόμησης. Εμπνεόμενοι από πιο πρόσφατες μελέτες, ο κατακερματισμός των σημάτων γίνεται σε τμήματα διαφορετικής διάρκειας με σκοπό την ανάλυση της επίδρασης της διάρκειας των τμημάτων στις επιδόσεις των υπολογιστικών μοντέλων. Παράλληλα εξετάζονται δίαφορα μοντέλα μίας ή πολλαπλών διεργασιών. Τα αποτελέσματα της εργασίας δείχνουν την υπεροχή της μεθόδου κατακερματισμού των σημάτων σε σχέση με τις κλασικές μεθόδους ταξινόμησης, όπως επίσης και συγκεκριμένων πολυδιεργασικών μοντέλων ως προς τα κλασικά μονοδιεργασικά.

Στη συνέχεια γίνεται η προσπάθεια εξαγωγής αναπαραστάσεων των σημάτων. Ξεκινώντας από ένα σύνολο ακουστικών χαρακτηριστικών προερχόμενα από τον ευρύτερο κλάδο Αναγώρισης Συναισθήματος συγκρίνονται δίκτυα εξαγωγής αναπαραστάσεων με κλάσικους αλγόριθμους επιλογής χαρακτηριστικών. Οι εξαγόμενες αναπαραστάσεις δείχνουν να υπερτερούν των ακουστικών χαρακτηριστικών. Μάλιστα, τα δίκτυα εξαγωγής αναπαραστάσεων σημειώνουν καλύτερες επιδόσεις από τους αλγόριθμους επιλογής χαρακτηριστικών. Τέλος δίνεται περισσότερη εκφραστικότητα στα δίκτυα, εξετάζοντας την ικανότητά εξαγωγής χρήσιμων αναπαραστάσεων όχι από τον χώρο των ακουστικών χαρακτηριστικών αλλά από μια πίο αυτούσια μορφή του σήματος φωνής. Η παρούσα έρευνα δείχνει πως αυτή η μέθοδος δεν υστερεί σε τίποτα από τις υπόλοιπες μεθόδους εξαγωγής αναπαραστάσεων από τον χώρο των ακουστικών χαρακτηριστικών, ενώ μάλιστα σε συγκεκριμένες περιπτώσεις παρουσιάζει βελτίωση.

Abstract

Emotion is considered as a major factor in human communications, since it defines, to a great degree, the concept of propagated information from a transmitter to a receiver. Until now there has been considerable progress in the field of human - computer interactions. Thus, given the importance of emotion in human communications, it is essential to investigate computational Emotion Recognition. This work focuses on Stress Detection via Speech Signals, aiming to construct decent speech representations and develop an automated system for emotion classification.

In the broader field of Speech Emotion Recognition, a speech signal is described by extracting acoustic features, which are used for classification. Over the past years, different models have been suggested to perform speech signal emotion classification. In previous work, a speech signal is thought to contain emotional information throughout its duration. However, in human communications, emotion is contained to a few number of parts of the speech signal. Consequently, it may be beneficial for Speech Emotion Recognition to divide the signal into parts. Moreover, the acoustic features extracted from the signal reflect the empirical knowledge of humans modeling useful properties of the signal for classification. Since there is no mathematical formulation of human's knowledge in terms of emotion, there is an imminent doubt in the performance of a computational machine.

The first part of this work, examines the classification of speech signals by dividing them to smaller parts. Primarily a baseline model is implemented by extracting acoustic features from the whole signal. Inspired by more recent studies, signals are divided to a variable length illustrating the impact of the duration of each part in the performances of single or multi - tasking models. The results of this work support predominance of speech fragmentation, over traditional classification methods, and multi tasking over single tasking models.

The next step is to construct signal representations. Starting from a set of acoustic features originated from Emotion Recognition, the traditional feature selection algorithms are compared to representation learning networks. These representations are used for classification outperforming the original acoustic feature set. In particular, the representation learning networks achieved a higher score than the feature selection algorithms. Finally, the networks are allowed more expressiveness, by examining their ability in the extraction of useful representations, not from the space of acoustic features, but from the raw speech signal. This work illustrates that the raw speech signal approach is equal to the traditional extracting representation approaches while, in some cases, it shows improvement.

Ευχαριστίες

Η εργασία αυτή ολοκληρώνει τον κύκλο σπουδών μου στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών στο Εθνικό Μετσόβιο Πολυτεχνείο. Ύστερα από όλα αυτά τα χρόνια φτάνει το πιο σημαντικό κομμάτι, οι ευχαριστιές στους ανθρώπους που με περιβάλλουν.

Αρχικά θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της εργασίας κ. Αλέξανδρο Ποταμιάνο για την εμπιστοσύνη που μου έδειξε, τον χρόνο που μου αφιέρωσε και την καθοδήγηση που μου πρόσφερε. Ωστόσο πάνω απ' όλα τον ευχαριστώ για το ενδιαφέρον και τη δίψα για γνώση που μου προκάλεσαν οι διαλέξεις του.

Έπειτα θα ήθελα να ευχαριστήσω την οικογένειά μου, για τις θυσίες που εκάνε ώστε εγώ να μπορέσω να σπουδάσω πάνω σε ένα αντικείμενο που επέλεξα. Ήταν πάντα εκεί υποστηρίζοντας τα θέλω μου χωρίς διακρίσεις και πιέσεις. Επιπλέον, ένα μεγάλο ευχαριστώ στη Μαριαλένα για την υποστήριξη και την συντροφία μέσα και εξώ από τη σχολή.

Τέλος ευχαριστώ τους φίλους μου για όλες τις θετικές είτε αρνητικές αναμνήσεις δίοτι διαμόρφωσαν τον άνθρωπο που είμαι σήμερα. Νιώθω τυχερός και ευγνώμων που βρέθηκα στη ζωή τους.

Γεώργιος Μιχαήλ Πανταζόπουλος,

Αθήνα, 25η Ιουνίου 2018

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Συναίσθημα στην Ανθρώπινη Επικοινωνία

Στο παρελθόν αναπτύχθηκαν θεωρίες σχετικά με την εξέλιξη του ανθρώπινου συναισθήματος. Σύμφωνα με την προσέγγιση του Δαρβίνου [10], τα συναισθήματα αποτελούν εξελικτικά φαινόμενα που αποσκοπούν στην ενίσχυση των ικανοτήτων επιβίωσης. Με την πάροδο του χρόνου ο άνθρωπος έμαθε να ερμηνεύει τα ερεθίσματα χαράς και θυμού, καλλιεργώντας την συναισθηματική του νοημοσύνη. Σε γενικές γραμμές τα ίδια εξελικτικά φαινόμενα εντοπίζονται σε ολόκληρο το ανθρώπινο είδος και άρα εμφανίζονται παρόμοια συναισθήματα μεταξύ των ανθρώπων. Επιπλέον δεδομένου ότι ο άνθρωπος μοιράζεται κάποια εξελικτικά στοιχεία με συγκεκριμένα θηλαστικά τότε πρέπει να παρατηρούνται ομοιότητες ως προς τη νοημοσύνη μεταξύ εξελικτικά κοντά οργανισμών.

Η επικοινωνία αποτελεί αναπόσπαστο στοιχείο των ανθρώπινων σχέσεων παρέχοντας τόσο λεκτική όσο και μη λεκτική πληροφορία. Κατά την διεξαγωγή ενός διαλόγου, οι λέξεις συνδυάζονται με τις εκφράσεις του προσώπου, τις χειρονομίες και την στάση του σώματος του ομιλητή. Τα παραπάνω προσφέρουν επιπλέον πληροφορία στον συνομιλητή. Μάλιστα ο τρόπους με τον οποίο επικοινωνεί ο άνθρωπος μπορεί να κατηγοριοποιηθεί σε δύο κανάλια, το άμεσο (*explicit*) και το έμμεσο (*implicit*) τα οποία αλληλεπιδρούν μεταξύ τους [11]. Το *explicit* μεταδίδει τα σαφή μηνύματα ενώ το *implicit* προσδιορίζει τον τρόπο διαχείρισης των σαφών μηνυμάτων που μεταδίδονται. Ιδιαίτερο ενδιαφέρον παρουσιάζει το συναίσθημα το οποίο διαδίδεται μέσω του *implicit* καναλιού. Το συναίσθημα που βιώνει ένας ομιλητής καθορίζει σε μεγαλο βαθμό τον τρόπο ομιλίας του. Αντίστοιχα μέσω του συναισθήματος καθορίζεται και η αντίληψη των μηνυμάτων που λαμβάνει ο συνομιλητής. Τέλος αν και η φύση του συναισθήματος είναι παρόμοια μεταξύ των ανθρώπων η έκφραση του διαμορφώνεται από διάφορους παράγοντες όπως το κοινωνικό του περιβάλλον.

1.2 Αναγνώριση Συναισθήματος και Τεχνολογία

Τα τελευταία χρόνια έχει γεννηθεί η ανάγκη της αλληλεπίδρασης ανθρώπου και μηχανής. Η αλληλεπίδραση αυτή βρίσκει άμεση εφαρμογή σε τηλεφωνικά κέντρα εξυπηρέτησης πελατών, στη διάγνωση ψυχικών νοσών, στην ανθρώπινη διασκέδαση και εκπαίδευση [12]. Στις παραπάνω εφαρμογές ο υπολογιστής καλείται συχνά να προσδιορίσει την συναισθηματική κατάσταση του ανθρώπου και στη συνέχεια να αντιδράσει αναλόγως. Οι αντιδράσεις αυτές συνήθως αποτελούν προσέγγιση των ανθρωπίνων αντιδράσεων. Ένα τέτοιο παράδειγμα αποτελούν τα *chatbots*. Τα *chatbots* αποτελούν υπολογιστικές μηχανές οι οποίες αλληλεπιδρούν με τον άνθρωπο μέσω ηχητικών σημάτων ή κειμένου. Ο άνθρωπος μπορεί να συζητήσει για οποιοδήποτε θέμα την μηχανή. Στη συνέχεια το *chatbots* διατυπώνει μια απάντηση στον άνθρωπο διαμορφωμένη στο ύφος του διαλόγου. Η απάντηση ωστόσο δεν πρέπει να διαχωρίζεται από τις αντίστοιχες πιθανές απαντήσεις του ανθρώπου. Συνεπώς σε αυτό το παράδειγμα η μηχανή καλείται να μιμηθεί τον άνθρωπο. Εφόσον οι ανθρώπινες αντιδράσεις επηρεάζονται έντονα από την εκάστοτε συναισθηματική κατάσταση, κρίνεται αναγκαία η έρευνα για την

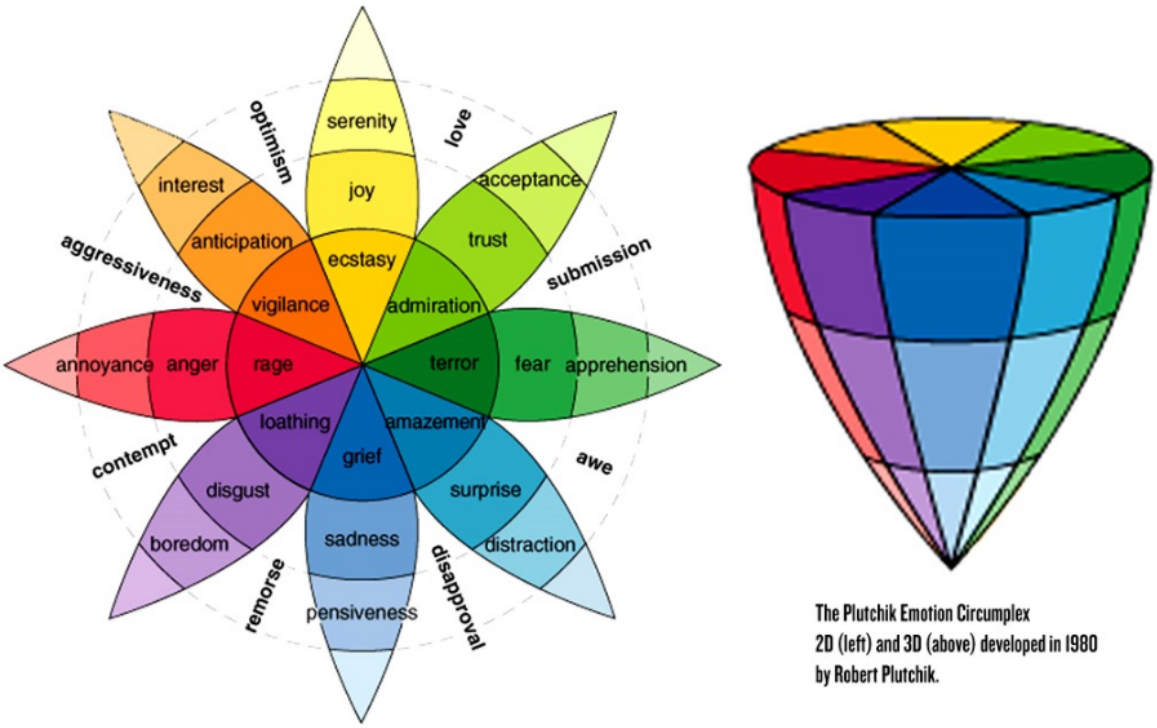
υπολογιστική αναγνώριση συναισθήματος.

Πρόσφατα έχει σημειωθεί σημαντική πρόοδος στην υπολογιστική Αναγνώριση Συναισθήματος. Συνήθως γίνεται η εξαγωγή κάποιων χαρακτηριστικών από δείγματα φωνής, βίντεο ή κειμένου, με σκοπό την κατηγοριοποίηση των δειγμάτων στο χώρο των συναισθημάτων. Τα προβλήματα αναγνώρισης συναισθήματος εντάσσονται στον κλάδο των προβλημάτων Αναγνώρισης Προτύπων. Το σύνολο δεδομένων συνοδεύεται συχνά από τις αντίστοιχες ετικέτες - κλάσεις. Στόχος είναι η κατασκευή ενός υπολογιστικού μοντέλου για την ταξινόμησή τους. Ειδικότερα κατά την Αναγνώριση Συναισθήματος τα δεδομένα προέρχονται είτε από φωνή ή βίντεο είτε απο κείμενο και οι κλάσεις αντιστοιχούν σε συνεχείς ή διακριτές τιμές στον χώρο των συναισθημάτων.

1.3 Αναπαράσταση Συναισθημάτων

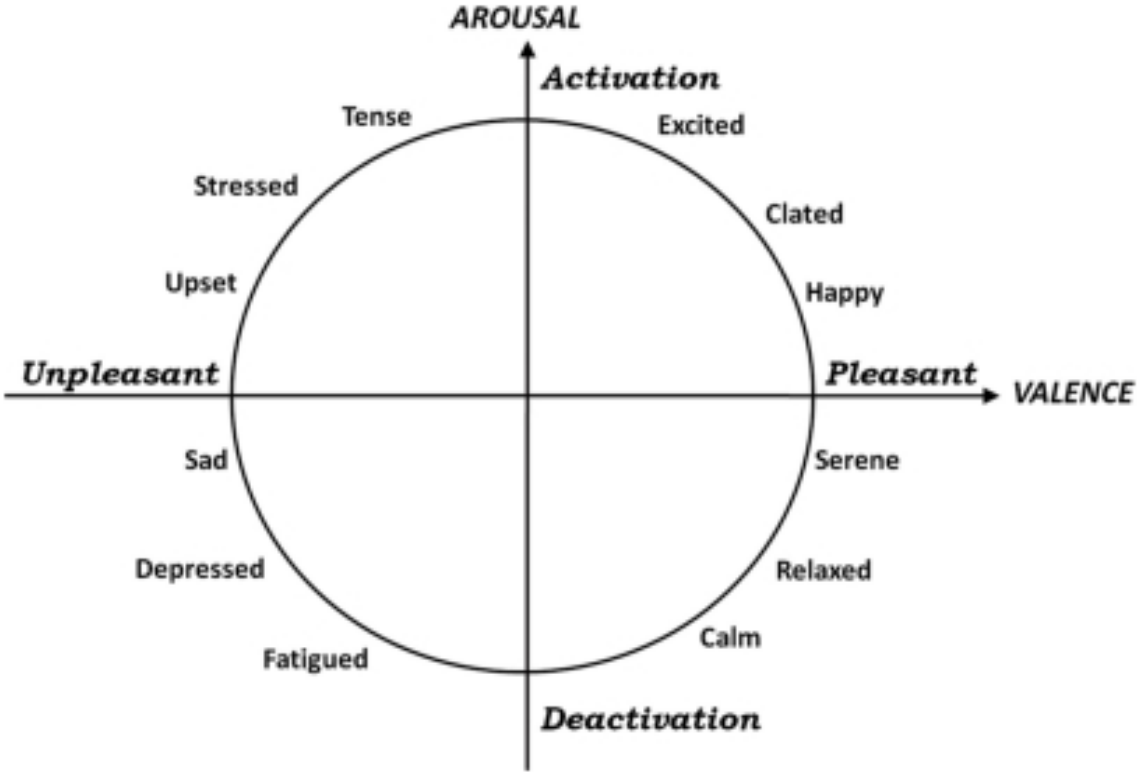
Μέχρι σήμερα έχουν προταθεί ποικίλλες αναπαραστάσεις των συναισθημάτων, οι οποίες μπορούν να κατακερματιστούν σε δύο κλαδους. Ο πρώτος κλάδος βασίζεται στην διακριτή κατηγοριοποίηση των συναισθημάτων. Σε αυτήν την αναπαράσταση ορίζονται κάποια πρωτεύοντα συναισθήματα, η μίξη των οποίων σε διαφορετικές αναλογίες δημιουργεί τα υπόλοιπα δευτερεύοντα. Ο δεύτερος κλάδος βασίζεται στην αναπαράσταση των συναισθημάτων σε συστήματα αξόνων. Κάθε άξονας του συστήματος περιγράφει ένα χαρακτηριστικό που προσδίδει το εκάστοτε συναίσθημα και η τιμή στον αξόνα αυτό εκφράζει τον βαθμό του χαρακτηριστικού αυτού.

Μια ιδιαίτερα γνωστή διακριτή κατηγοριοποίηση είναι αυτή του *Plutchik* [1]. Ο ίδιος ισχυρίστηκε πως υπάρχουν 8 πρωτεύονται συναισθήματα : θυμός (*anger*), φόβος (*fear*), λύπη (*sadness*), απέχθεια (*disgust*), έκπληξη (*surprise*), προσμονή (*anticipation*), εμπιστοσύνη (*trust*), και χαρά (*joy*). Ο *Plutchik* χρησιμοποίησε ένα διδιάστατο και ένα τρισδιάστατο έχρωμο μοντέλο για να περιγράψει τον τρόπο με τον οποίο σχετίζονται τα συναισθήματα μεταξύ τους όπως φαίνεται στο Σχήμα 1.1.



Σχήμα 1.1: Αναπαράσταση Συναισθημάτων του Plutchik [1].

Στη θεωρία του *Plutchik* τα 8 πρωτεύοντα συναισθήματα εντοπίζονται σε 4 αντίθετα ζεύγη : *joy* και *sadness*, *anger* και *fear*, *trust* και *disgust*, *surprise* και *anticipation*. Τα συναισθήματα αυτά βρίσκονται στο κέντρο του διδιάστατου μοντέλου. Τα συναισθήματα στις εξωτερικές περιοχές προκύπτουν από μίξεις 2 πρωτευόντων συναισθημάτων. Στο τρισδιάστατο μοντέλο η κατακόρυφη διάσταση αναπαριστά την ένταση του συναισθήματος ενώ ο κύκλος που βρίσκεται στην οριζόντια διάσταση αναπαριστά τον βαθμό ομοιότητας των συναισθημάτων.



Σχήμα 1.2: Αναπαράσταση Συναισθημάτων σε άξονες με συνεχείς τιμές. [2].

Από την άλλη πλευρά, η κατηγοριοποίηση των συναισθημάτων μπορεί να γίνει σε συστήματα αξόνων με συνεχείς τιμές [13]. Σε αυτήν την προσέγγιση κάθε συναίσθημα βρίσκεται σε μια συγκεκριμένη περιοχή του διδιάστατου επιπέδου που ορίζουν οι δύο κάθετοι άξονες. Στο Σχήμα 1.2 δίνεται μια διδιάστατη συνεχή κατηγοριοποίηση των συναισθημάτων. Η ιδέα αυτή επεκτείνεται και σε συστήματα 3 αξόνων *Valence*, *Arousal*, *Dominance* [14]. Ο άξονας *Valence* ποσοτικοποεί την ευχαρίστηση που προσδίδει κάποιο συνάισθημα. Ο άξονας Arousal αφορά την ένταση του συναισθήματος ενώ ο άξονας *Dominance* την κυρίαρχη φύση του συναισθήματος. Ενδεικτικά τόσο το *anger*, *fear* και *boredom* είναι δυσάρεστα συναισθήματα έχοντας μικρές τιμές στον άξονα *Valence*. Ωστόσο το *anger* είναι κυριάρχο συνάισθημα σε σχέση με το *fear* με αποτέλεσμα να έχει μεγαλύτερη τιμή στον άξονα *Dominance*. Από την άλλη, το *anger* προσδίδει μεγαλύτερη ένταση από το *boredom* έχοντας υψηλότερη τιμή στον άξονα *Arousal*.

1.4 Συνεισφορά Εργασίας

Η παρούσα εργασία πραγματεύεται την Αναγνώριση Άγχους σε Σήματα Φωνής. Με βάση την κατηγοριοποίηση του *Plutchik* το άγχος αποτελεί δευτερεύον συναίσθημα το οποίο προέρχεται από την μιξή προσμονής και φόβου. Στόχος είναι η εξαγωγή αναπαραστάσεων του άγχους στο χώρο των

χαρακτηριστικών των δειγμάτων καθώς και η κατασκευή ενός υπολογιστικού μοντέλου για την αναγνώρισή του.

Σύμφωνα με τον *Selye* [15], το άγχος (*stress*) διαχωρίζεται σε 2 είδη το ευεργετικό (*eustress*), και το επιβλαβές (*distress*). Το *eustress* αντιστοιχεί σε καταστάσεις στις οποίες το άτομο επωφελείται από αυτό. Τέτοιες καταστάσεις συνήθως σχετίζονται με το ένστικτο επιβίωσης. Σε περιπτώσεις ζωής ή θανάτου το *eustress* ενισχύει τις ψυχικές και σωματικές ικανότητες του ατόμου. Από την άλλη το distress εντοπίζεται σε καταστάσεις που το άτομο αδυνατεί να ανταποκριθεί στην καθημερινή του ζωή. Το *distress* οδηγεί σε εξασθένηση του ατόμου και πολλές φορές σε κατάθλιψη. Στην παρούσα εργασία το άγχος ταυτίζεται με το *distress*.

Μέχρι πρόσφατα έχει γίνει μικρή προσπάθεια για την κατασκεύη ενός υπολογιστικού συστήματος για Αναγνώριση Άγχους από Φωνή. Τα περισσότερα συστήματα επεξεργάζονται βιολογικά σήματα ως δείγματα [16, 17]. Σε άλλες περιπτώσεις λαμβάνονται δείγματα μέσω σκηνοθετικού και φυσικού άγχους [18]. Τα δείγματα που χρησιμοποιούνται στην παρούσα εργασία προέρχονται από φυσικούς διαλόγους. Περισσότερες πληροφορίες για τα δεδομένα που χρησιμοποιούνται δίνονται σε επόμενο κεφάλαιο.

1.5 Περίγραμμα Εργασίας

Η εργασία αυτή οργανώνεται ως ακολούθως. Στο Κεφάλαιο 2 αναλύεται ο μηχανισμός παραγωγής φωνής του ανθρώπου, περιγράφονται τα στάδια κωδικοποίησης και αποκωδικοποίησης των μηνυμάτων κατά την διεξαγωγή ενός διαλόγου καθώς και τα εξαγόμενα ακουστικά χαρακτηριστικά. Το Κεφάλαι 3 εστιάζει στον θεωρητικό και πρακτικό φορμαλισμό των μοντέλων που χρησιμοποιούνται. Σε πρώτο στάδιο εξηγούνται βασικές έννοιες απαραίτητες για την κατανόηση της προσέγγισης της εργασίας ενώ στη συνέχεια παρουσιάζονται κλασικά και ακολουθιακά μοντέλα. Στο Κεφάλαιο 4 αναλύεται το θεωρητικό υπόβαθρο της εργασίας σχετικά με την κατασκεύη αναπαραστάσεων και την μετάδοση γνώσης μεταξύ μοντέλων ταξινόμησης. Παράλληλα παρουσιάζονται οι αλγόριθμοι κατασκευής αναπαραστάσεων. Το Κεφάλαιο 5 αφιερώνεται στις πειραματικές διατάξεις και αποτελέσματα της εργασίας όπου συγκρίνονται διαφορετικά μοντέλα και διαφορετικοί αλγόριθμοι κατασκευής αναπαράστασης. Τέλος, το Κεφάλαιο 6 συμπυκνώνει τα αποτελέσματα της εργασίας καθώς και πιθανές μελλοντικές προεκτάσεις.

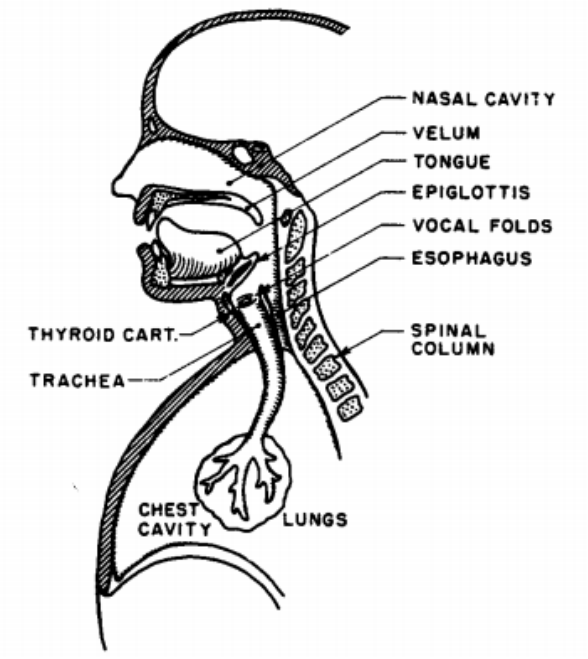
Κεφάλαιο 2

Χαρακτηριστικά Φωνής

2.1 Εισαγωγή

Θεμελιώδης σκοπός της φωνής είναι η ανθρώπινη επικοινωνία, δηλαδή η μεταδόση ενός μηνύματος από τον ομιλητή στον συνομιλητή του. Η φωνή είναι μια επίκτητη ικανότητα του ανθρώπου. Αναπτύσσεται, ελέγχεται και συντηρείται από την συνεχή ανατροφοδότηση του μηχανισμού ακόης, καθώς και των μυών υπεύθυνων για την παραγωγή της. Η πληροφορία που συγκροτείται από αυτά τα τμήματα του ανθρώπινου σώματος διαχειρίζεται από συγκεκρίμενα τμήματα του εγκεφάλου τα οποία συντονίζουν ολόκληρη την λειτουργία της φωνής [19]. Η φωνή παρουσιάζει ιδιαίτερο ενδιαφέρον, τόσο ως προς τον μηχανισμό παραγωγής της όσο και στα χαρακτηριστικά της. Ο μηχανισμος παραγωγής της φωνής περιλαμβάνει την συνδυασμένη χρήση του μυαλού, της μύτης του στόματος, των πνευμόνων καθως και κοιλοτήτων του ανθρώπινου σώματος. Τα χαρακτηριστικά της φωνής διαφέρουν από ανθρωπό σε άνθρωπο γεγονός που δυσχεραίνει την κατασκεύη υπολογιστικών συστημάτων.

Στο κεφάλαιο αυτό αρχικά αναλύεται ο τρόπος παραγωγής φωνής του ανθρώπου. Εν συνεχεία εξηγείται η αλυσίδα ομιλίας (*speech chain*), δηλαδή τα σταδία κωδικοποίησης και αποκωδικοποίησης μηνύματων κατά την διεξαγωγή ενός διαλόγου, μεταξύ δύο ομιλητών. Τέλος παρουσιάζονται τα χαρακτηριστικά φωνής που χρησιμοποιούνται στην παρούσα εργασία.



Σχήμα 2.1: Σχηματική αναπαράσταση της ανθρώπινης φωνητικής οδού. [3].

2.2 Παραγωγή Φωνής

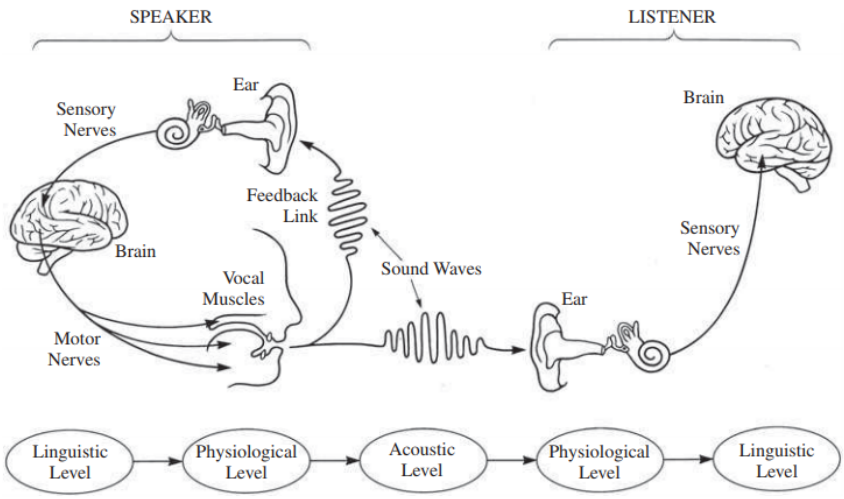
Στο Σχήμα 2.1 αναπαριστάται η ανθρώπινη φωνητική οδός καθώς και τα μέλη του ανθρώπινου σώματος υπεύθυνα για την παραγωγή της φωνής. Η φωνητική οδός αποτείλεται από τον φάρυγγα και την στοματική κοιλότητα. Ξεκινά από το άνοιγμα ανάμεσα στις φωνητικές χορδές, την γλωττίδα, και καταλήγει στα χείλη. Σε συγκεκριμένες περιπτώσεις παραγωγής φωνής χρησιμοποιείται και η ρινική οδός η οποία ξεκινά από τον ουρανίσκο και καταλήγει στα ρουθούνια.

Η διαδικασία παραγωγή φωνής διαφέρει στην περίπτωση έμφωνων και άφωνων ήχων. Τα μέλη της φωνητικής ή της ρινικής οδού αναγκάζονται να μετατοπιστούν προκειμένου να παραχθεί ο ζητούμενος ήχος. Στην γενική περίπτωση παραγωγής φωνής εισέρχεται αέρας μέσω της αναπνευστικής οδού. Καθώς ο αέρας εξέρχεται από τους πνεύμονες μέσω της τραχείας, οι φωνητικές χορδές πάλλονται περιοδικά. Στη συνέχεια οι περιοδικοί παλμοί διαμορφώνονται ως προς την συχνότητα τους και τέλος διοχετεύονται μέσω του φάρυγγα στην στοματική και πιθανώς στην ρινική κοιλότητα. Στην περίπτωση έμφωνων ήχων, όπως φωνηέντων, ο ήχος που παράγεται εξαρτάται άμεσα από την θέση της γλώσσας, του σαγονιού, και των χειλών. Οι άφωνοι ήχοι παράγονται μέσω ροής εξερχόμενου αέρα με σταθερό ρυθμό και στένωσης της φωνητικής οδού σε κάποιο σημείο της. Ανάλογα με το σημείο σύσφιξης της φωνητικής οδού παράγεται ο αντίστοιχος άφωνος ήχος. Ενδεικτικά για την παραγωγή του /φ/ η σύσφιξη γίνεται κόντα στα χείλη, ενώ για την παραγωγή του /θ/ η συσφίξη γίνεται στα δόντια.

2.3 Αλυσίδα Ομιλίας - *Speech Chain*

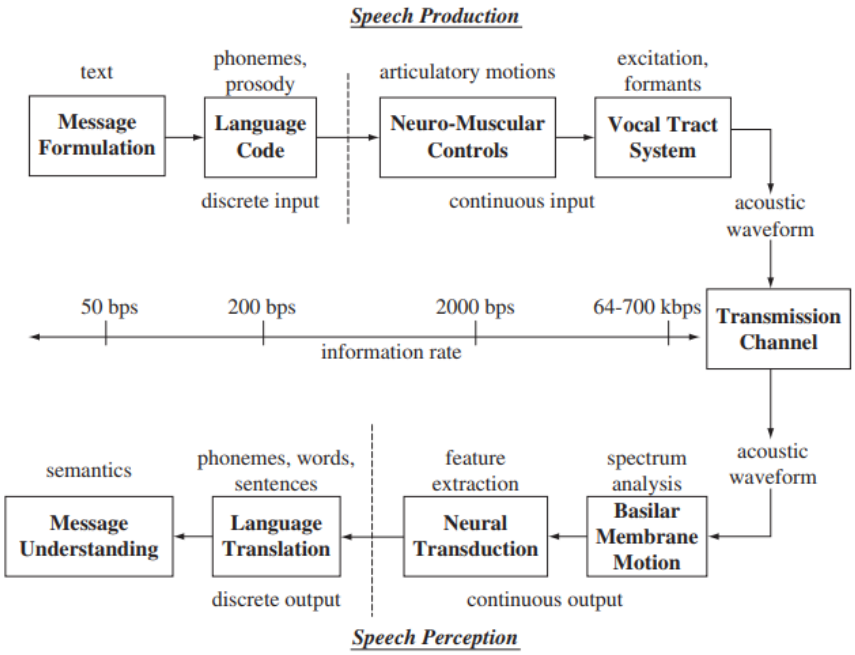
Η διαδικασία παραγωγής φωνής που αναφέρθηκε παραπάνω προϋποθέτει πως ο ομιλητής έχει κωδικοποιήσει το μήνυμα που επιθυμεί να μεταδόσει στον συνομιλητή του. Η κωδικοποίηση αυτή αντιστοιχεί σε μια ακολουθία συμβόλων στο λεξιλόγιο της γλώσσας. Από την πλευρά του ο συνομιλητής αποκωδικοποιεί την ακουλουθία ερμηνεύοντας το περιέχομενο της. Ύστερα κωδικοποιεί το δικό του μήνυμα σε μια νέα ακολουθία και η διαδικασία επαναλαμβάνεται. Μια πλήρης αναπαράσταση της παραγωγής και λήψης μηνυμάτων προτάθηκε από τους *Denes* και *Pinson* [4].

Η αναπαράσταση της *speech chain* παρουσιάζεται στο Σχήμα 2.2. Σύμφωνα με την αναπαράσταση της αυτή τα μεταδιδόμενα μηνύματα ακολουθούν συγκεκριμένη διαδρομή ανάμεσα σε 3 διακριτά επίπεδα. Η διαδρομή είναι η εξής: γλωσσικό επίπεδο (*linguistic level*) στο οποίο επιλέγονται οι βασικοί ήχοι επικοινωνίας με σκοπό την έκφραση ενός μηνύματος, φυσικό επίπεδο (*physiological level*) στο οποίο οι συνιστώσες της φωνητικής οδού παράγουν τους αντίστοιχους ήχους, ακουστικό επίπεδο (*acoustic level*) στο οποίο το κωδικοποιημένο μήνυμα εξέρχεται μέσω της φωνής και λαμβάνεται τόσο από τον ομιλητή όσο και από τον συνομιλητή. Από την πλευρά του συνομιλητή, υπάρχει ομοίως το physiological level στο οποίο η φωνή αναλύεται μέσω του ακουστικού συστήματος του συνομιλητή και *linguistic level* στο οποίο τελικά η φωνή αποκωδικοποιείται στο λεξιλόγιο της γλώσσας [5].



Σχήμα 2.2: Αναπαράσταση speech chain. [4].

Μια πιο λεπτομερής αναπαράσταση της *speech chain* παρουσιάζεται στο Σχήμα 2.3. Εδώ η *speech chain* χωρίζεται στα στάδια παραγωγής (*speech production*) και αντίληψης (*speech perception*). Το *speech production* ξεκινά πάνω αριστερά στο διάγραμμα. Αρχικά το μήνυμα αναπαριστάται σε μια πτυχή του μυαλού του ομιλητή. Στη συνέχεια κωδικοποιείται σε σύμβολα της γλώσσας και καθορίζεται ο τρόπος έκφρασής τους. Έπειτα το νευρομυϊκό σύστημα του ομιλητή καθορίζει τις κινήσεις των μελών υπεύθυνα για την παραγωγή ομιλίας. Τέλος το μήνυμα εκφράζεται μέσω της φωνητικής οδού με τον τρόπο που καθορίστηκε στο στάδιο της κωδικοποίησής του. Το *speech perception* εκτελεί με αντίστροφη διαδικασία την αποκωδικοποίηση και ερμηνεία του μηνύματος. Πρώτα η φωνή αναπαριστάται στο πεδίο της συχνότητας. Ύστερα ακολουθεί μια μορφή εξαγωγής χαρακτηριστικών και τέλος μεταφράζονται τα στοιχεία της γλώσσας σε φωνήματα, λέξεις και προτάσεις.



Σχήμα 2.3: Αναπαράσταση *speech production* και *speech perception*. [5].

2.4 Ακουστικά Χαρακτηριστικά

Για την μοντελοποίηση οποιουδήποτε δεδομένου είναι απαραίτητη η εξαγωγή χαρακτηριστικών που το περιγράφουν. Όσο αναφορά τα δεδομένα φωνής, τα χαρακτηριστικά εξάγονται στο *acoustic level*. Στον κλάδο της Αναγνώρισης Συναισθήματος μέσω Σημάτων Φωνής έχουν χρησιμοποιηθεί διάφορα ακουστικά χαρακτηρισικά, πολλά από τα οποίο προέρχονται από τον κλάδο της Αναγνώρισης Φωνής. Δεδομένου ότι η συναισθηματική κατάσταση του ανθρώπου επηρεάζει την φωνή του, τότε πρέπει να παρατηρούνται ομοιότητες μεταξύ ακουστικών χαρακτηριστικών σε παρόμοιες συναισθηματικές καταστάσεις. Οι ομοιότητες αυτές οδηγούν στην κατηγοριοποίηση των συναισθημάτων στον χωρό των χαρακτηριστικών και συμβάλλουν στην κατασκεύη της υπολογιστικής μηχανής για την Αναγνώριση Συναισθήματος.

Μια ηχητική εκφώνηση (*utterance*) αντιστοιχεί σε ένα δεδομένο εισόδου, και η εξαγωγή χαρακτηριστικών γίνεται σε επίπεδο πλαισίου (*frame - level*), ομάδας πλαισίων (*segment - level*) ή εκφώνησης (*utterance - level*). Συχνά τα χαρακτηριστικά αυτά αναφέρονται ως περιγραφητές της εκφώνησης, και διαχωρίζονται σε περιγραφητές χαμηλού επιπέδου (*low level descriptors - LLDs*) ή συναρτησιακά (*functionals*). Οι *LLDs* περιλαμβάνουν φασματικά (*spectral*), προσωδικά (*prosodic*), και ποιότητας φωνής (*voice quality*) χαρακτηριστικά, συμπεριλαμβανομένου και των παραγώγων τους (*delta*). Η εξαγωγή τους γίνεται σε *frame level* μετασχηματίζοντας το σήμα φωνής του πλαισίου από το πεδίο χρόνου στο πεδίο συχνότητας. Τα *functionals* εξάγονται σε *segment level* ή *utterance level* με εφαρμογή στατιστικών πάνω στους *LLDs* στο σύνολο των πλαισίων που ενδιαφέρει. Συνήθως το κάθε πλαίσιο διαρκεί 20 - $50msec$ ενώ η ομάδα πλαισίων μερικά δευτερόλεπτα.

Στην παρούσα εργασία χρησιμοποιούνται *utterance*, *segment* καθώς και *frame - level* ακουστικά χαρακτηριστικά. Τα *frame - level* χαρακτηριστικά περιλαμβάνουν ένα σύνολο *LLDs* οι οποίοι εξάγονται μέσω του ανοιτού λογισμικού *librosa* και περιγράφονται παρακάτω. Τα *utterance* και *segment - level* εφαρμόζονται σε προβλήματα Αναγνώρισης Φωνής, Συναισθήματος Φύλου και Ηλικίας του ομιλητή [20, 21] εστιάζοντας στην παραγλωσσική πληροφορία *(paralinguistic information)* της εκφώνησης [22]. Τα χαρακτηριστικά αυτά εξάγονται μέσω του ανοιχτού λογισμικού *openSmile* [23] για εξαγωγή *LLDs* όσο και *functionals*. Στον Πίνακα 2.1 και 2.2 δίνονται ονομαστικά τα *LLDs* και τα *functionals*. Ακολουθεί σύντομη περιγραφή των ακουστικών χαρακτηριστικών.

	LLDs
Spectral	Mel-Frequency Cepstral Coefficients (MFCCs)
Prosodic	Θεμελιώδης Συχνότητα ($F0$), Ένταση (Loudness) Πιθανότητα έμφωνου ήχου (Voice Probability)
Voice Quality	Jitter, Shimmer

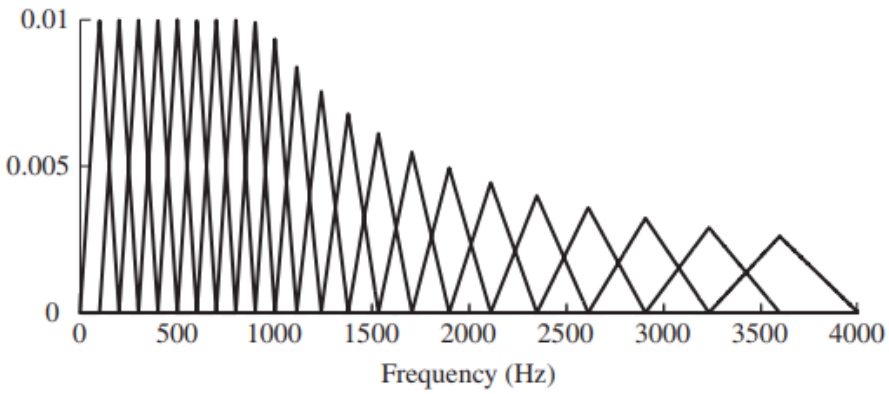
Πίνακας 2.1: Ακουστικά χαρακτηριστικά χαμηλού επιπέδου.

Functionals
Θέση Μεγίστου και Ελαχίστου (Maximum and Minimum Position)
Μέση Αριθμητική Τιμή, Τυπική Απόκλιση (Arithmetic Mean, Standard Deviation)
Ασυμμετρία, Κύρτωση (Skewness, Kurtosis)
Συντελεστές Γραμμικής Παρεμβολής (Linear Regression Coefficients)
Διαστήματα Τεταρτημορίου Εκατοστημορίου (Quartile, Percentile)
Up - Level Time

Πίνακας 2.2: *Functionals*.

Mel - Frequency Cepstral Coefficients (MFCCs)

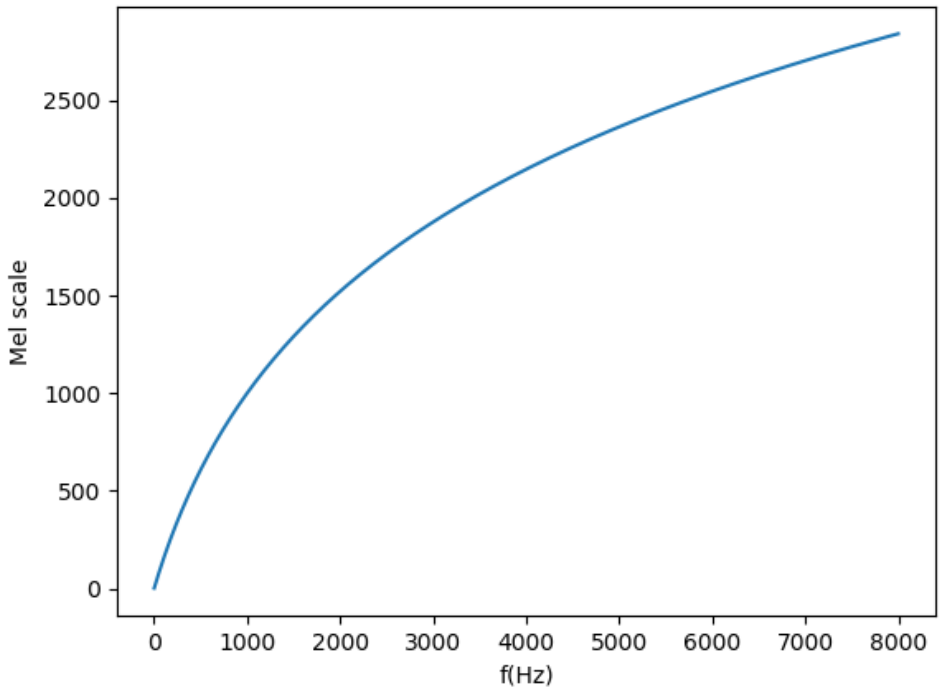
Τα *MFCCs* παρέχουν μια αναπαράσταση του σήματος προσομοιώνοντας τα εύρη ακοής του ανθρώπου [24]. Οι συχνότητες του σήματος αναλύονται με βάση μια τριγωνική συστοιχία φίλτρων στην κλίμακα mel. Στην κλίμακα *Hz* τα φίλτρα διαθέτουν σταθερό εύρος ζώνης για χαμηλές συχνότητες (συνήθως 1*Khz*) το οποίο αυξάνεται εκθετικά μέχρι την συχνότητα δειγματοληψίας F_s του σήματος. Μια τριγωνική συτοιχία φίλτρων δίνεται στο Σχήμα 2.4.



Σχήμα 2.4: Τριγωνική συστοιχία φίλτρων. [5].

Η κλίμακα *mel* (βλ. Σχήμα 2.5) αποσκοπεί στον μετασχηματισμό των συχνοτήτων από την κλίμακα *Hz* έτσι ώστε οι ίδιες να βρίσκονται ισοκατανεμημένες μεταξύ τους στο εύρος ακοής του ανθρώπου [25]. Επειδή ο άνθρωπος διακρίνει ευκολότερα μεταβολές σε μικρότερες συχνότητες από ότι σε μεγαλύτερες η κλίμακα *mel* είναι λογαριθμική. Παρ' όλα αυτά δεν υπάρχει μονοσήμαντος ορισμός της κλίμακας, διότι η αντίληψη συχνοτήτων του ανθρώπου είναι υποκειμενική [5]. Μια φόρμουλα για την κλίμακα *mel* δίνεται ως:

$$m(f) = 1127 log_{10}(1 + \frac{f}{700}) \tag{2.1}$$



Σχήμα 2.5: Αναπαράσταση της λογαριθμικής κλίμακας *mel*.

Ένα σήμα φωνής δίνεται στο Σχήμα 2.6. Αρχικά το σήμα χωρίζεται σε πλαίσια και παραθυρώνεται. Δημιουργούνται l επικαλυπτόμενα πλαίσια μήκους N δειγμάτων το καθένα. Το πρώτο βήμα για την εξαγωγή των *MFCCs* είναι ο Διακριτός Μετασχηματισμός Fourier (*Discrete Fourier Transform - DFT*) των πλαισίων από το πεδίο του χρόνου στο πεδίο της συχνότητας. Ο μετασχηματισμός ενός πλαισίου $x_m(n)$ για ένα τυχαίο πλαίσιο ορίζεται ως :

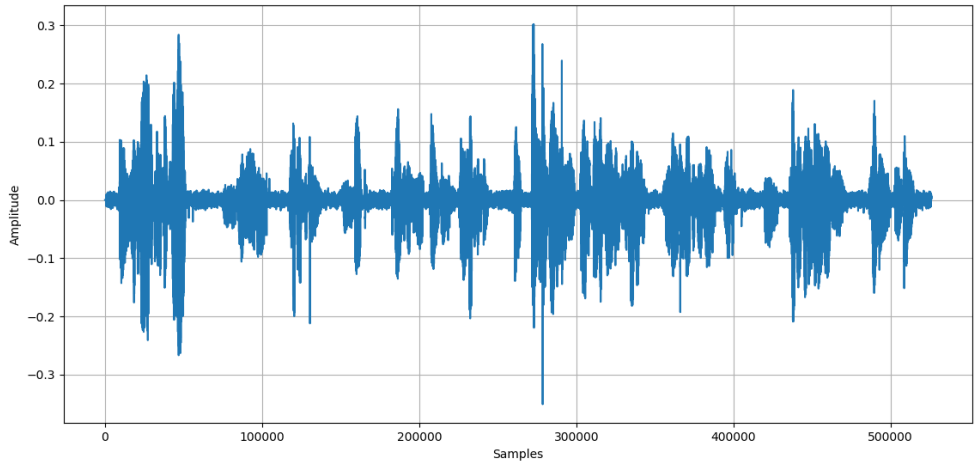
$$X_m[k] = \sum_{n=0}^{N-1} x_m[n] e^{(-j2\pi k/N)n} \tag{2.2}$$

Στη συνέχεια υπολογίζεται η ενέργεια εξόδου $P_m[r]$ κάθε φίλτρου $r = 1, 2 \dots R$ με συνάρτηση $V_r[k]$ και εύρος $[L_r, R_r]$ της συστοιχίας στην κλίμακα *mel* :

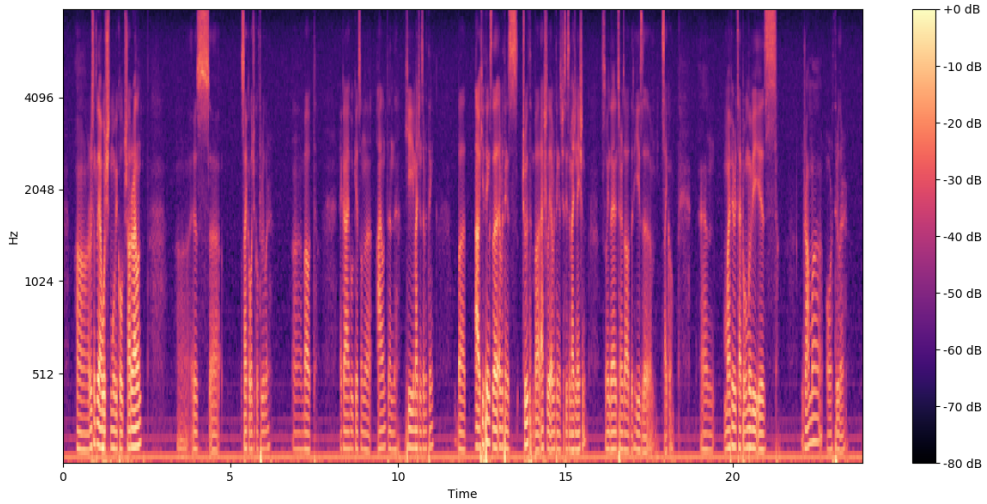
$$P_m[r] = \frac{\sum_{k=L_r}^{R_r} |V_r[k] X_m[k]|^2}{\sum_{k=L_r}^{R_r} |V_r[k]|^2} \tag{2.3}$$

Σε αυτό το σημείο το σήμα μπορεί να αναπαρασταθεί στο διδιάστατο επίπεδο που ορίζουν οι άξονες χρόνου και συχνότητας. Ο άξονας χρόνο περιλαμβάνει όλα τα παράθυρα στα οποία εφαρμόζεται η τριγωνική συστοιχία ενώ ο άξονας συχνότητας αντιστιστοιχεί στις επιμέρους ενέργειες των εξόδων των φίλτρων σε κάθε εύρος συχνότητας ξεχωριστά. Η αναπαράσταση αυτή αναφέρεται ως *spectgram* ή *melspectogram* αν ο άξονας της συχνότητας είναι στην κλίμακα *mel*, και φαίνεται στο Σχήμα 2.7. Στη συνέχεια για την εξαγωγή των *MFCC* συνιστώσων υπολογίζεται ο διακριτός μετασχηματισμός συνημιτόνου του λογαρίθμου της ενέργειας της εξόδου των φίλτρων :

$$MFCC_m[n] = \frac{1}{R} \sum_{r=1}^{R} log(P_m[r]) cos\left(\frac{2n\pi}{R}(r + \frac{1}{2})\right) \tag{2.4}$$



Σχήμα 2.6: Αναπαράσταση σήματος φωνής στο πεδίο του χρόνου.



Σχήμα 2.7: Αναπαράσταση σήματος φωνής μέσω *spectogram*.

Θεμελιώδης Συχνότητα ($F0$)

Η θεμελιώδης συχνότητα ενός σήματος φωνής αντιστοιχεί στην συχνότητα με την οποία ανοίγουν και κλείνουν οι φωνητικές χορδές [26]. Συμβολίζεται με $F0$ διότι αποτελεί την χαμηλότερη συχνότητα του σήματος, ακολουθούμενη από τις συχνότητες $F1, F2, \ldots$ οι οποίες ονομάζονται συχνότητες φωτονισμού. Για τον υπολογισμό της έχουν προταθεί ποικίλλοι αλγόριθμοι, από τους οποίους ξεχωρίζει ο αλγόριθμος με την χρήση αυτοσυσχέτισης [27].

Αρχικά το σήμα φωνής $s(n)$ φιλτράρεται μέσω ενός βαθυπερατού φίλτρου και στη συνέχεια παραθυρώνεται σε επικαλυπτόμενα πλαίσια $f_s(n, m)$, σύμφωνα με την εξίσωση 2.5a όπου $w(m-n)$ αντιστοιχεί στο παράθυρο μήκους N_w. Στα πλαίσια εφαρμόζεται μια συνθήκη κατωφλίου, 2.5b τιμής C_{thr}. Στη συνέχεια υπολογίζεται η αυτοσυσχέτιση σύμφωνα με την εξίσωση 2.5c όπου η αντιστοιχεί στην καθυστέρηση κατά τον υπολογισμό της αυτοσυσχέτισης. Τέλος η F_0 προσεγγίζεται σύμφωνα με την εξίσωση 2.5d, όπου F_s είναι η συχνότητα δειγματοληψίας, και F_h, F_l είναι η μέγιστη και ελάχιστη αντιληπτή συχνότητα του ανθρώπου. Με τον υπολογισμό της F_0 έμμεσα υπολογίζεται και η

Πιθανότητα Έμφωνου Ήχου (*Voice Probability*).

$$f_s(n,m) = s(n)w(m-n) \tag{2.5a}$$

$$\hat{f}_s(n,m) = \left\{ \begin{array}{cc} f_s(n,m) - C_{thr} & |f_s(n,m)| \geq C_{thr}, \\ 0 & |f_s(n,m)| < C_{thr} \end{array} \right\} \tag{2.5b}$$

$$r_s(\eta,m) = \frac{1}{N} \sum_{n=m-N_w+1}^{m} \hat{f}_s(n,m) f_s(n-\eta,m) \tag{2.5c}$$

$$\hat{F}_0 = \frac{F_s}{N_w} \arg\max_{\eta} \{|r_s(\eta,m)|\} \begin{array}{c} n = N_w(F_h/F_s), \\ n = N_w(F_l/F_s) \end{array} \tag{2.5d}$$

Jitter Shimmer

Τα χαρακτηριστικά *Jitter* και *Shimmer* ανήκουν στην οικογένεια των χαρακτηριστικών ποιότητας φωνής. Αποσκοπούν στην εκτίμηση της μεταβολής της θεμελιώδους συχνότητας και πλάτους του σήματος φωνής [28]. Βρίσκουν εφαρμογή σε περιπτώσεις αναγνώρισης φωνητικών παθήσεων [29]. Στην πράξη υπάρχουν διάφορα *Jitter* και *Shimmer*. Στο παραπάνω σύνολο χαρακτηριστικών χρησιμοποιούνται τα *frame-to-frame Jitter* και *Shimmer*, όπως και το *differential frame-to-frame Jitter*.

Κεφάλαιο 3

Ταξινομητές

3.1 Εισαγωγή

Ένα σύστημα αναγνώρισης προτύπων αποσκοπεί στην εύρεση επαναλαμβανόμενων μοτίβων στα δεδομένα εισόδου. Τέτοια συστήματα υπάρχουν σε αφθονία στην φύση. Για παράδειγμα ο άνθρωπος διαθέτει την ικανότητα αναγνώρισης αντικειμένων με βάση κάποια συγκεκριμένα χαρακτηριστικά όπως το σχήμα ή το χρώμα. Ο ίδιος μπορεί να αναγνωρίσει την συναισθηματική κατάσταση του συνομιλητή του και στη συνέχεια να προσαρμόσει το ύφος του στις απαιτήσεις το διαλόγου. Σε γενικές γραμμές αυτές οι ικανότητες σχετίζονται με την λήψη μιας απόφασης βασιζόμενη στα ερεθίσματα, δηλαδή στα δεδομένα εισόδου. Η λήψη αυτής της απόφασης πηγάζει από το νευρωνικό σύστημα του ανθρώπου, το οποίο εκπαιδεύεται συνεχώς με βάση τα δεδομένα εισόδου. Απαραίτητη προϋπόθεση για την λήψη της απόφασης είναι η έννοια της κατηγοριοποίησης. Το νευρωνικό σύστημα του ανθρώπου κατηγοριοποιεί τα δεδομένα εισόδου και στη συνέχεια μέσω αλγορίθμων καλείται να λάβει την βέλτιστη απόφαση για την εκάστοτε περίπτωση. Ένας ταξινομητής καλείται να προσομοιώσει την διαδικασία απόφασης του νευρωνικού συστήματος του ανθρώπου.

Σήμερα η τεχνολογία των ταξινομητών βρίσκει άμεση εφαρμογή σε προβλήματα διάγνωσης κάποιας ασθένειας, στον διαχωρισμό μεταξύ ελαττωματικών και μη αντικειμένων, ακόμα και στην ψυχαγωγία του ανθρώπου [12]. Ειδικότερα στον τομέα της Αναγνώρισης Προτύπων έχουν προταθεί ποίκιλλοι αλγόριθμοι για την επίλυση προβλημάτων που σχετίζονται με την ταξινόμηση. Τα προβλήματα αυτά προέρχονται από διάφορους κλάδους όπως Όραση Υπολογιστών, η Επεξεργασία Φωνής και Φυσικής Γλώσσας, καθώς και η Αναγνώριση Φωνής και Συναισθήματος.

Πάρα την διαφορετική προέλευση των παραπάνω προβλημάτων ταξινόμησης συνήθως διαθέτουν το εξής μοτίβο. Αρχικά τα δεδομένα εισόδου χρησιμοποιούνται για την εξαγωγή κάποιων χαρακτηριστικών που βοηθούν στην επίλυση του προβλήματος. Στη συνέχεια δίνεται ένα σύνολο κλάσεων $C_1, C_2, \ldots, Cn$ στις οποίες ταξινομείται οποιοδήποτε δεδομένο εισόδου με βάση τα χαρακτηριστικά του.

Η ανάπτυξη ενός ταξινομητή περιλαμβάνει δύο στάδια : το στάδιο εκπαίδευσης (*training*) και το στάδιο αξιολόγησης (*testing*). Στο στάδιο της εκπαίδευσης ο ταξινομητής μαθαίνει να λαμβάνει αποφάσεις με βάση τα δεδομένα εκπαίδευσης (training data). Υπάρχουν 2 διαφορετικές μέθοδοι εκπαίδευσης : με επίβλεψη (*supervised*) και χωρίς επίβλεψη (*unsupervised*). Στην *supervised* εκπαίδευση τα δεδομένα συνοδεύονται από τις αντίστοιχες κλάσεις - ετικέτες (*labels*). Η απόφαση του ταξινομητή αξιολογείται με βάση τα *labels* εκπαίδευσης. Αντίθετα, στη *unsupervised* εκπαίδευση τα δεδομένα εισόδου δεν διαθέτουν ετικέτες. Ο ταξινομητής καλείται να διαχωρίσει μόνος του τα δεδομένα εισόδου σε διαφορετικες, συνήθως συγκεκριμένες ως προς τον αριθμό, κλάσεις. Στο στάδιο της αξιολόγησης, ο ταξινομητής δοκιμάζεται στα δεδομένα αξιολόγησης (*test data*), σημειώνοντας την επίδοσή του στην μετρική αξιολόγησης.

Στον κλάδο της Αναγνώρισης Συναισθήματος από Φωνή χρησιμοποιούνται διάφοροι ταξινομητές. Οι κυρίοτεροι από αυτούς είναι : Μηχανές Υποστήριξης Διανυσμάτων (*Support Vector Machines*

- *SVMs*), Βαθιά Νευρωνικά Δίκτυα (*Deep Neural Networks - DNNs*), Αναδρομικά Νευρωνικά Δίκτυα (*Recurrent Neural Networks - RNNs*). Στο κεφάλαιο αυτό αρχικά παρουσιάζονται η θεωρία αποφάσεων κατά *Bayes* και τα γραφικά μοντέλα καθώς χρησιμοποιούνται σε επόμενο κεφάλαιο. Στη συνέχεια αναλύονται οι παραπάνω ταξινομητές που θα χρησιμοποιηθούν για την εξαγωγή συμπερασμάτων. Τέλος ακολουθεί αναφορά στην ικανότητα των ταξινομητών να λύνουν τόσο ένα πρόβλημα (*single - task*), όσο και πολλά προβλήματα ταυτόχρονα (*multi - task*).

3.2 Θεωρία Πιθανοτήτων

3.2.1 Θεωρία Αποφάσεων κατά *Bayes*

Η θεωρία αποφάσεων κατά *Bayes* αποτελεί βασική προσέγγιση στα προβλήματα Αναγνώρισης Προτύπων, ποσοτικοποιώντας τις αποφάσεις ταξινόμησης μέσω πιθανοτήτων [9]. Αν το πρόβλημα ταξινόμησης περιλαμβάνει $C_1, C_2, \ldots, Cn$ κλάσεις και κάθε πρότυπο εισόδου διαθέτει ένα διάνυσμα χαρακτηριστικών $\mathbf{x}$ τότε ο κανόνας του *Bayes* εκφράζει την εκ των υστέρων (*posterior*) πιθανότητα της κλάσης C_i δεδομένου του προτύπου ως το γινόμενο της συνάρτησης πιθανοφάνειας (*likelihood*) $p(\mathbf{x}|C_i)$, επί την εκ των προτέρων (*prior*) πιθανότητα της κλάσης C_i $p(C_i)$, προς την συνάρτηση πυκνότητας πιθανότητας (*evidence*) $p(\mathbf{x})$:

$$p(C_j|\mathbf{x}) = \frac{p(x|C_j)p(C_j)}{p(\mathbf{x})} \tag{3.1a}$$

$$p(\mathbf{x}) = \sum_{j=1}^{N} p(\mathbf{x}|C_j)p(C_j) \tag{3.1b}$$

Σύμφωνα με τον κανόνα του *Bayes* το πρότυπο ταξινομείται στην κλάση C_j^* με την μέγιστη *posterior* :

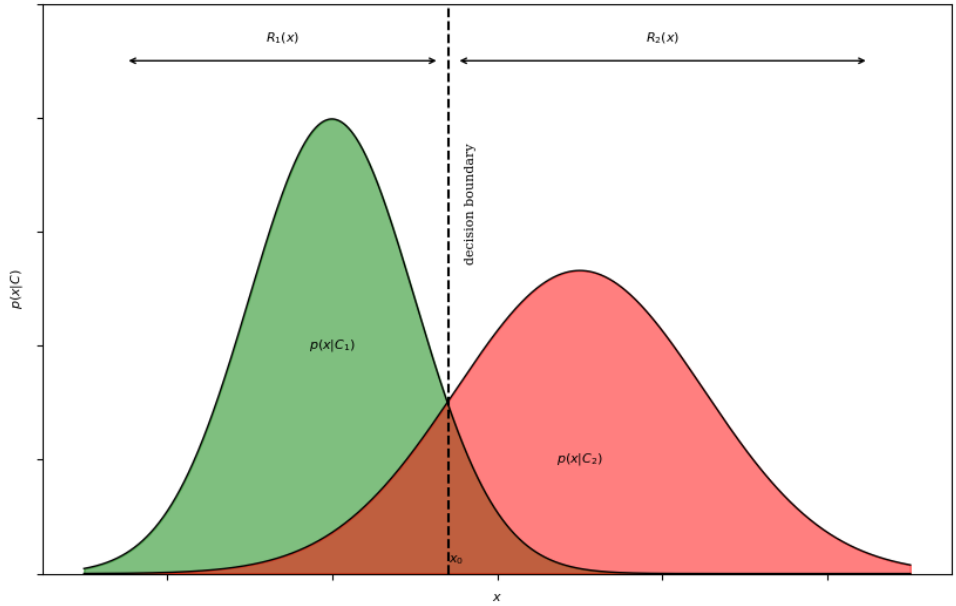
$$C_j^* = \arg\max_j p(C_j|\mathbf{x}) \tag{3.2}$$

Όπως φαίνεται και στο Σχήμα 3.1. στην περίπτωση δύο κλάσεων και ενός χαρακτηριστικού, δημιουργείται ένα σημείο απόφασης πάνω στον άξονα x, x_0. Το σημείο αυτό χωρίζει τον άξονα σε δύο περιοχές R_1 και R_2. Θεωρώντας ένα άγνωστο δείγμα, αν το χαρακτηριστικό του x βρίσκεται μέσα στην περιοχή R_1 τότε σύμφωνα με τον κανόνα του *Bayes* επιλέγεται η κλάση C_1 αλλίως επιλέγεται η κλάση C_2. Ενδέχεται όμως το δείγμα να μην ανήκει στην κλάση η οποία προήλθε από τον κανόνα. Σε αυτήν την περίπτωση πρόκειται για ένα δείγμα που δεν ταξινομήθηκε σωστά. Η περιοχή στην οποία ο κανόντας του *Bayes* ταξινομεί λάθος το πρότυπο αντιστοιχεί στο εμβάδο της καμπύλης που δημιουργείται από την τομή των $P(x|C_1), P(x|C_2)$ και του άξονα x. Η τιμή του σφάλματος $p(error|x)$ μπορεί να υπολογιστεί μέσω της εξίσωσης 3.3. Αποδεικνύεται ότι η επιλογή του ταξινομητή κατά *Bayes* ελαχιστοποιεί το σφάλμα ταξινόμησης [30]. Στη γενική περίπτωση περισσότερων χαρακτηριστικών ο διαχωρισμός των κλάσεων γίνεται μέσω της εύρεσης υπερεπιπέδων ή επιφανείων διαχωρισμού (*decision boundary*).

$$p(error|x) = min\{p(C_1|x), p(C_2|x)\} \tag{3.3}$$

3.2.2 Γραφικά Μοντέλα

Η χρήση πιθανοτήτων αποτελεί αναπόσπαστο στοιχεί στον κλάδο της Αναγνώρισης Προτύπων. Ανεξαρτήτως της πολυπλοκότητας ενός πιθανοκρατικού μοντέλου είναι εφικτή η επίλυση του μέσω



Σχήμα 3.1: Παράδειγμα σημείου απόφασης x_0 σύμφωνα με τον κανόνα ταξινόμησης κατά *Bayes* στην περίπτωση των δύο κλάσεων C_1, C_2 και ενός χαρακτηριστικού x

των θεμελιώδων κανόνων γινομένου και αθροίσματος των πιθανοτήτων. Ωστόσο στην πράξη η απεικόνιση ενός τέτοιου μοντέλου γίνεται με την βοηθεία γραφικών μοντέλων (*graphical models*) αφενός δίοτι η οπτικοποίηση είναι σχετικά απλή και αφετέρου επειδή οι ιδιότητες του μοντέλου προκύπτουν μέσω παρατήρησης του γράφου. Ένας γράφος $G = (V, E)$ αποτελείται από ένα σύνολο κόμβων V συνδεδεμένων με ακμές E. Στα πιθανοκρατικά γραφικά μοντέλα κάθε κόμβος v_i αντιπροσωπεύει μια τυχαία μεταβλητή, ενώ η ακμή μεταξύ δύο κόμβων v_i και v_j εκφράζει την σχέση που τις συνδέει. Έστω 3 τυχαίες μεταβλητές x_1, x_2, x_3. Με διαδοχική εφαρμογή του κανόνα γινομένου, η από κοινού πιθανότητα των μεταβλήτων ισούται :

$$p(x_1, x_2, x_3) = p(x_3|x_2, x_1)p(x_2|x_1)p(x_1) \tag{3.4}$$

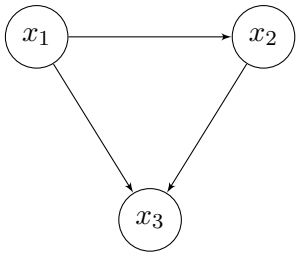
Το δεξί μέλος της εξίσωσης 3.4 μπορεί να αναπαρασταθεί με την βοήθεια ενός γραφικού μοντέλου (βλ. Σχήμα 3.2). Πιο συγκεκριμένα κάθε μεταβλητή x_i αντιστοιχεί σε ένα κόμβο του γράφου. Στη συνέχεια κάθε δεσμευμένη κατανομή που εμφανίζεται στην εξίσωση εκφράζεται ως μια κατευθυνόμενη ακμή από τους αντίστοιχους κόμβους που δεσμεύεται η κατανομή αυτή.

Για ένα σύνολο τυχαίων μεταβλητών $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ η από κοινού πιθανότητα μπορεί να γραφτεί ως εξής :

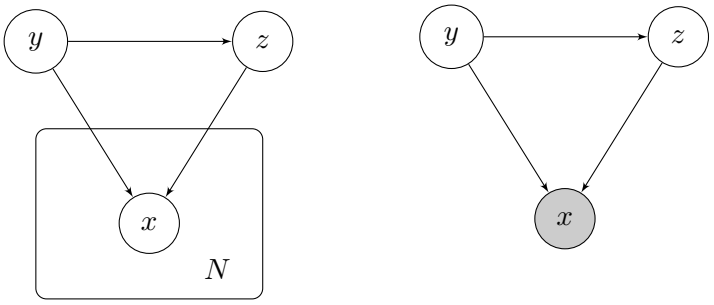
$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|px_i) \tag{3.5}$$

όπου px_i είναι το σύνολο των μεταβλητών όπου δεσμεύεται η x_i.

Σε περιπτώσεις όπου μια μεταβλητή εμφανίζεται πολλές φορές χρησιμοποιείται ένας μόνο κόμβος εκπροσωπώντας όλους τους υπόλοιπους όπως φαίνεται στο Σχήμα 3.3a. Επιπλέον, σε περιπτώσεις όπου η μεταβλητή είναι παρατηρήσιμη (*observed*) συμβολίζεται με γραμμοσκιασμένο κόμβο, ενώ με λευκό συμβολίζονται οι κρυφές (*latent*) μεταβλητές (βλ. Σχήμα 3.3b).



Σχήμα 3.2: Παράδειγμα γραφικού μοντέλου για την από κοινού πιθανότητα τριών τυχαίων μεταβλητών.



(a) Ενός κόμβου εκπροσώπησης.

(b) *Observed* και *latent* μεταβλητών.

Σχήμα 3.3: Παράδειγματα γραφικών μοντέλων.

3.3 Συναρτήσεις Κόστους

Οι περισσότεροι αλγόριθμοι μάθησης αποσκοπούν στην βελτιστοποίηση μιας μετρικής $J(\theta)$ ως προς ένα σύνολο μεταβλητών θ. Με τον όρο βελτιστοποίηση ορίζεται η διαδικασία μεγιστοποιήσης ή ελαχιστοποίησης της συνάρτησης $J(.)$, γνωστή ως συνάρτηση απώλειας (*loss function*) ή συνάρτηση σκοπού (*objective function*). Μολονότι οι περισσότεροι αλγόριθμοι εξετάζουν την ελαχιστοποίηση μιας *loss function*, εξίσου ικανή είναι η μεγιστοποίηση της, μέσω των ίδιων αλγορίθμων θέτοντας $J^{'}(\theta) = -J(\theta)$.

Το σύνολο μεταβλητών στο οποίο ενδιαφέρει η βελτιστοποίηση της *loss function* είναι το σύνολο των παραμέτρων του ταξινομητή. Κατά την *supervised* εκπαίδευση, το σύνολο εκπαίδευσης $\mathbf{X_{train}} = \{\mathbf{x_i}\}_{i=1}^{N}$ συνοδεύεται από τις αντίστοιχες ετικέτες $\mathbf{y_{train}} = \{y_i\}_{i=1}^{N}$. Στην περίπτωση των δύο κλάσεων ισχύει $y_i \in \{0, 1\}$. Αντίθετα αν υπάρχουν m δυνατές κλάσεις τότε κάθε ετικέτα αντιστοιχεί σε ενα *one hot vector* $\mathbf{y_i}$, όπου για κάθε συντεταγμένη του διανύσματος $y_{ik} = 1$ αν το πρότυπο x_i ανήκει στην κλάση C_k, αλλιώς $y_{ik} = 0$. Κάθε πρότυπο εκπαίδευσης συμβάλλει στην *loss function* κατά $J_i(\theta)$, ενώ η τιμή της συνήθως ισούται με την μέση τιμή του αθροίσματος των επιμέρους J_i :

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} J_i(\theta) \tag{3.6}$$

Η επιλογή της *loss function* διαφέρει από εφαρμογή σε εφαρμογή. Στην περίπτωση ενός προβλήματος ταξινόμησης προτύπων δύο κλάσεων, μια αρκετά συνηθισμένη *loss function* είναι η συνάρτηση δυαδικής εντροπίας (*binary cross entropy*). Αν $p_i \in [0, 1]$ είναι η πρόβλεψη του ταξινομητή τότε η *binary cross entropy* παίρνει την μορφή :

$$J(\theta) = -\frac{1}{N} \sum_{i=1}^{N} y_i log(p_i) + (1 - y_i) log(1 - p_i) \tag{3.7}$$

Οι λογαριθμικοί όροι έχουν ως στόχο να επιβάλλουν ένα είδος ποινής σε πρότυπα που δεν ταξινομούνται σωστά. Έτσι αν $y_i = 0$ και $p_i = 0.8$ τότε ο πρώτος όρος του αθροίσματος μηδενίζεται, ενώ ο δεύτερος όρος συμβάλλει κατά μια μεγάλη τιμή. Παρόμοια ο δεύτερος όρος μηδενίζεται και ο πρώτος όρος αυξάνεται αν $y_i = 1$ και $p_i = 0.2$. Ο ταξινομητής θεωρείται ιδανικός αν $J(\theta) = 0$ στο στάδιο της αξιολόγησης. Στην περίπτωση περισσότερων από δύο κλάσεων, ο ταξινομητής προβλέπει m πιθανές κλάσεις και η *binary cross entropy* γενικεύεται στην συνάρτηση κατηγορικής εντροπίας (categorical cross entropy) :

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m} y_{ij}log(p_{ij}) + (1 - y_{ij})log(1 - p_{ij}) \tag{3.8}$$

Ανεξάρτητα από την επιλογή της *loss function* η ελαχιστοποίησή της γίνεται μέσω αλγορίθμων βελτιστοποίησης. Βασική έννοια των αλγορίθμων είναι ο υπολογισμός της κλίσης (*gradient*) $\nabla_\theta J(\theta)$. Η κλίση προσδιορίζει την κατεύθυνση στον χώρο των παραμέτρων, πάνω στην οποία ελαχιστοποιείται η τιμή της *loss function*. Οι περισσότεροι αλγόριθμοι ακολουθούν την ίδια διαδικασία για τον υπολογισμό της κλίσης. Ενδεικτικά στην περίπτωση των νευρωνικών δικτύων χρησιμοποιείται ο αλγόριθμος *Back Propagation* [31]. Μετά τον υπολογισμό της κλίσης συνήθως ακολουθεί ένας κανόνας ενημέρωσης των παραμέτρων και επανάληψη του αλγορίθμου. Στο παράδειγμα των νευρωνικών δικτύων ενας ιδιαίτερα διαδεδομένος αλγόριθμος βελτιστοποίησης είναι ο αλγόριθμος *Adam* [32].

Τέλος σε αρκετές εφαρμογές εισάγονται επιπλέον όροι στην *loss function* παίρνοντας την μορφή της εξίσωσης 3.9. Οι όροι αυτοί αναφέρονται ως συνθήκες κανονικοποίησης των παραμέτρων του μοντέλου. Τυπικές τέτοιες συνθήκες είναι η νόρμα L_1 ή L_2, αναγκάζοντας τις τιμές των παραμέτρων να παραμείνουν σε σχετικά μικρό εύρος, γενικεύοντας καλύτερα το δίκτυο. Η παράμετρος λ αντιστοιχεί σε μια *hyperparameter* του μοντέλου η οποία προσδιορίζεται κατά την εκπαίδευσή του. Υψηλές τιμές της παραμέτρου δίνουν περαιτέρω ισχύ στον περιορισμό μειώνοντας τις τιμές των παραμέτρων.

$$J^{'}(\theta) = J(\theta) + \lambda R(\theta) \tag{3.9}$$

3.4 Μηχανές Υποστήριξης Διανυσμάτων (*Support Vector Machines - SVMs*)

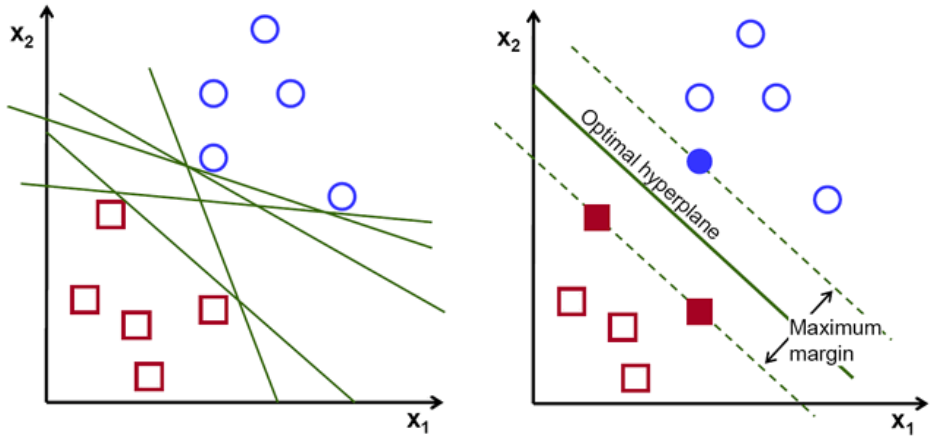
3.4.1 Γραμμικά Διαχωρίσιμα Πρότυπα.

Έστω ένα πρόβλημα ταξινόμησης προτύπων με δύο δυνατές ετικέτες $C_1 = -1, C_2 = 1$. Κάθε πρότυπο εισόδου διαθέτει ένα δισδιάστατο διάνυσμα χαρακτηριστικών $\mathbf{x_k}$ το οποίο συνοδεύεται από την αντίστοιχη ετικέτα $t_k \in \{-1, 1\}$. Μια απεικόνιση των προτύπων εκπαίδευσης δίνεται στο Σχήμα 3.4a, όπου με κόκκινο και τετράγωνο σχήμα απεικονίζονται τα πρότυπα της πρώτης κλάσης, ενώ με μπλε και κυκλικό της δεύτερης.

Στη γενική περίπτωση το υπερεπίπεδο διαχωρισμού περιγράφεται από την εξίσωση 3.10, όπου η συνάρτηση $\phi(.)$ αναπαριστά έναν μετασχηματισμό του διανύματος χαρακτηριστικών :

$$y(\mathbf{x}) = \mathbf{W}^T\phi(\mathbf{x}) + \mathbf{b} \tag{3.10}$$

Τα πρότυπα του σχήματος 3.4a ονομάζονται γραμμικά διαχωρίσιμα, επειδή υπάρχει τουλάχιστον μια ευθεία που τα διαχωρίζει, δηλαδή υπάρχει τουλάχιστον ένας συνδυασμός των $\mathbf{W}$ και $\mathbf{b}$ έτσι ώστε να ισχύει $sign(y(\mathbf{x_k})) = t_k$. Στο ίδιο σχήμα φαίνεται πως η επιλογή της ευθείας δεν είναι μοναδική. Δεδομένου ότι είναι επιθυμητό ο ταξινομητής να ανταποκρίνεται σε πρότυπα στα οποία δεν έχει



(a) Δυνατά υπερεπίπεδα διαχωρισμού. **(b)** Βέλτιστο υπερεπίπεδο διαχωρισμού.

Σχήμα 3.4: Παράδειγμα προβλήματος ταξινόμησης δύο γραμμικά διαχωρίσιμων κλάσεων [6].

συναντήσει προηγουμένως, η ευθεία οφείλει να διαχωρίζει τα πρότυπα όσο το δυνατό καλύτερα. Η ευθεία αυτή φαίνεται στο Σχήμα 3.4b.

Τα *Support Vector Machines* εντοπίζουν την ευθεία με την παραπάνω ιδιότητα, δηλαδή ασχολούνται με το πρόβλημα του περιθωρίου (*margin*) [33]. Ως *margin* ορίζεται η ελάχιστη απόσταση της επιφάνειας διαχωρισμού από οποιοδήποτε πρότυπο. Η ευθεία που διαχωρίζει βέλτιστα τα πρότυπα έχει το μέγιστο περιθώριο.

Γνωρίζοντας πως η απόσταση ενός προτύπου $\mathbf{x_k}$ από ένα υπερεπίπεδο $y(\mathbf{x}) = 0$ ισούται με $\frac{|y(\mathbf{x_k})|}{||\mathbf{W}||}$ καθώς και ότι οι ζητούμενες λύσεις οφείλουν να ταξινομούν σωστά τα πρότυπα, η απόσταση ενός προτύπου από την επιφάνεια διαχωρισμού παίρνει την μορφή :

$$\frac{t_k y(\mathbf{x})}{||\mathbf{W}||} = \frac{t_k(\mathbf{W}^T\phi(\mathbf{x}) + \mathbf{b})}{||\mathbf{W}||} \tag{3.11}$$

Χρησιμοποιώντας την εξίσωση 3.11 το πρόβλημα μεγιστοποίησης γράφεται ως:

$$J\big(\mathbf{W}, \mathbf{b}\big) = \arg\max_{\mathbf{W},\mathbf{b}} \left\{ \frac{1}{||\mathbf{W}||} \min_k \left\{ t_k(\mathbf{W}^\mathbf{T}\phi(\mathbf{x}) + \mathbf{b}) \right\} \right\} \tag{3.12}$$

Γράφοντας σε κανονική μορφή την $y(\mathbf{x})$ προκύπτει πως η τιμή της στα πλησιέστερα πρότυπα των κλάσεων C_1, C_2 είναι -1 και 1 αντίστοιχα η εξίσωση 3.12 απλοποιείται στην 3.13a υπό τον περιορισμό της 3.13b, η λύση της οποίας προκύπτει με πολλαπλασιαστές *Lagrange* [9].

$$J\big(\mathbf{W}\big) = \arg\min_{\mathbf{W}} \left\{ \frac{1}{2}||\mathbf{W}||^2 \right\} \tag{3.13a}$$

$$t_k\big(\mathbf{W}^T\phi(\mathbf{x}) + \mathbf{b}\big) \geq 1 \tag{3.13b}$$

Παρόμοια λογική εφαρμόζεται και σε προβλήματα ταξινόμησης περισσότερων των 2 κλάσεων. Σε αυτήν την περίπτωση έχουν αναπτυχθεί διάφορες τεχνικές όπως μίας - εναντίων - υπολοίπων και μίας - εναντίων - μίας, οι οποίες βασίζονται στην υλοποίηση περισσότερων ταξινομητών δύο κλάσεων και στον συνδυασμό τους για την λήψη της τελικής απόφασης [34, 35, 36].

3.4.2 Μη Γραμμικά Διαχωρίσιμα Πρότυπα

Στις περισσότερες περιπτώσεις τα πρότυπα δεν είναι γραμμικά διαχωρίσιμα. Σε αυτές τις περιπτώσεις η χρήση οποιουδήποτε υπερεπιπέδου στο χώρο χαρακτηριστικών οδηγεί σε εσφαλμενά ταξι-

νομημένα πρότυπα. Ωστόσο είναι δυνατή η μεγιστοποίηση του περιθωρίου όπως και προηγουμένως με χρήση συγκεκριμένων μεταβλητών ξ_k οι οποίες αναφέρονται ως μεταβλητές χαλάρωσης [37, 38]. Οι ίδιες ορίζονται ως εξής : $\xi_k = 0$ για οποιοδήποτε πρότυπο εκπαίδευσης k το όποιο ταξινομείται σωστά, και $\xi_k = |t_k - y(\mathbf{x_k})|$ διαφορετικά. Συνεπώς το πρόβλημα ελαχιστοποίησης της εξίσωσης 3.13a και του περιορισμού 3.13b γίνεται :

$$\arg\min_{\mathbf{W}} \left\{ C \sum_{i=1}^{n} \xi_i + \frac{1}{2}||\mathbf{W}||^2 \right\} \tag{3.14a}$$

$$t_k(\mathbf{W}^T \phi(\mathbf{x}) + \mathbf{b}) \geq 1 - \xi_k \tag{3.14b}$$

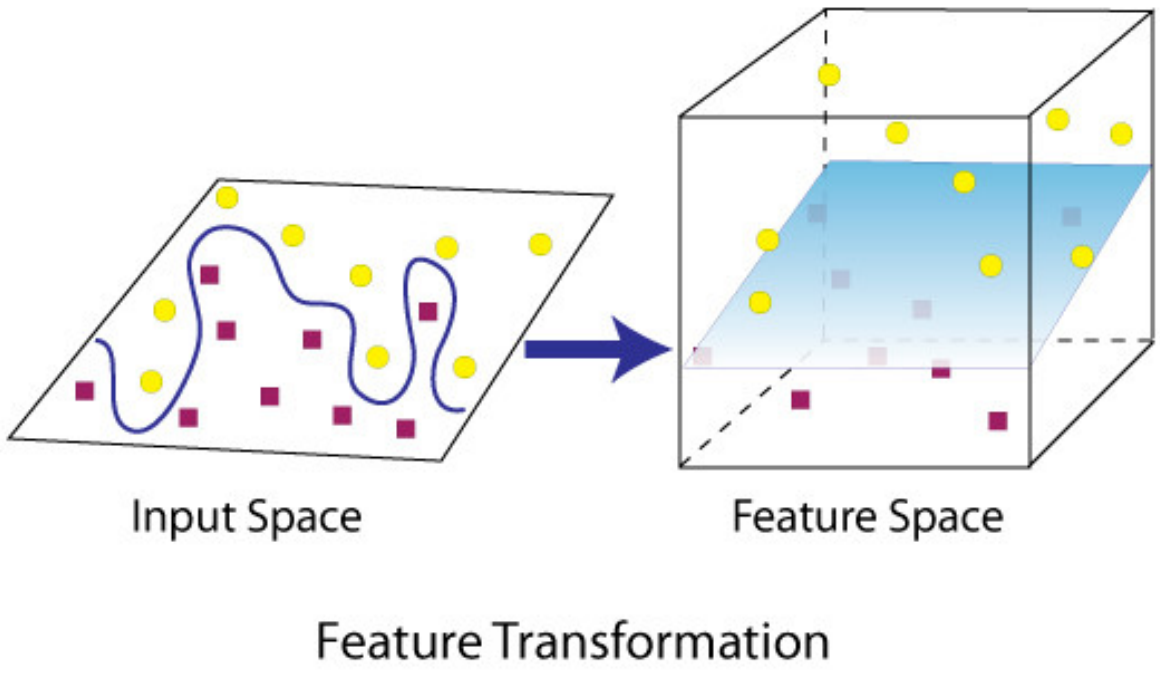
Όπου C είναι μια σταθερά αντιστάθμισης μεταξύ της ποινής των μεταβλητών χαλάρωσης και του *margin*. Όπως και στην περίπτωση των γραμμικά διαχωρίσιμων προτύπων, η λύση προκύπτει με χρήση πολλαπλασιαστών *Lagrange* [9].

3.4.3 Συναρτήσεις Πυρήνα

Όπως φαίνεται και στο Σχήμα 3.5, τα *SVMs* λαμβάνουν τα δεδομένα είσοδου και τα προβάλλουν σε έναν χώρο συνήθως μεγαλύτερης διάστασης. Η διαδικασιά αυτή γίνεται μέσω της μη γραμμικής απεικόνισης $\phi(.)$. Στη συνέχεια τα *SVMs* ταξινομούν τα πρότυπα στο νέο χώρο (*feature space*).

Ωστόσο, στην πράξη δεν χρειάζεται η απευθείας προβολή των σημειών στο χώρο προβολής, αλλά το εσωτερικό γινόμενο μεταξύ κάθε ζεύγους προτύπων. Επειδή η διαδικασία υπολογισμού του εσωτερικού γινομένου για κάθε ζεύγος είναι αρκετά ακριβή, χρησιμοποιείται το τέχνασμα του πυρήνα. Οι συναρτήσεις πυρήνα δέχονται ως είσοδο δύο διανύσματα $\mathbf{x_i}, \mathbf{x_j}$ και επιστρέφουν το εσωτερικό γινόμενο τους, με μικρότερο υπολογιστικό κόστος. Οι ίδιες ορίζονται ως $K(\mathbf{x_i}, \mathbf{x_j}) = \phi(\mathbf{x_i})^T \phi(\mathbf{x_j})$.

Συνήθως ως συναρτήση πυρήνα χρησιμοποιείται η γραμμική, η πολυωνυμική και η συνάρτηση ακτινικής βάσης (*RBF*). Αποδεικνύεται πως με την κατάλληλη απεικόνιση ένα πρόβλημα ταξινόμησης δύο κλάσεων μπορεί να γίνει γραμμικά διαχωρίσιμο [39].

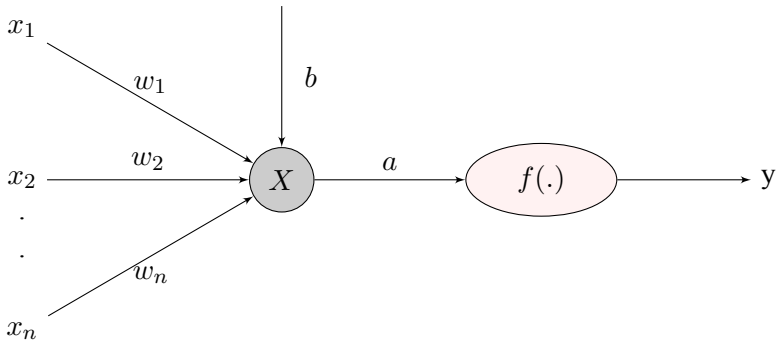


Σχήμα 3.5: Παράδειγμα προβληματος ταξινόμησης δύο μη γραμμικά διαχωρίσιμων κλάσεων [7].

3.5 Νευρωνικά Δίκτυα

Ένα νευρωνικό δίκτυο είναι ένας τεράστιος επεξεργαστής με κατανεμημένη αρχιτεκτονική, ο οποίος αποτελείται από απλές μονάδες επεξεργασίας και έχει από την φύση του τη δυνατότητα να αποθηκεύει εμπειρική γνώση και να την καθιστά διαθέσιμη για χρήση[33].

Η θεωρία των νευρωνικών δικτύων ως υπολογιστικές μηχανές εισάχθηκε από [40], στη συνέχεια διατυπώθηκε ο πρώτος κανόνας αυτό - οργανοούμενης μάθησης [41] και τελικά προτάθηκε το perceptron ως πρώτο μοντέλο μάθησης [42].



Σχήμα 3.6: Σχηματική αναπαράσταση ενός *Perceptron*.

3.5.1 Το *Pereptron*

Το *perceptron* αποτελεί την απλούστερή δυνατή μορφή νευρωνικού δικτύου, δέχεται ένα διάνυσμα χαρακτηριστικών $\mathbf{x} \in \mathbb{R}^n$. το οποίο πολλαπλασιάζεται με ένα διάνυσμα βαρών w. Οι πολλαπλασιασμένες συντεταγμένες αθροίζονται μαζί με μία σταθερά πόλωσης b. Η έξοδος της άθροισης, a, χρησιμοποιείται ως είσοδο στην συνάρτηση ενεργοποίησης (*activation function*) του perceptron, $f(.)$ παράγοντας την έξοδο του, y. Το *perceptron* δίνεται στο Σχήμα 3.6 και περιγράφεται από τις εξισώσεις 3.15a και 3.15b. Όπως και στην περίπτωση των *SVMs*, η εξίσωση 3.15a περιγράφει ένα υπερεπίπεδο διαχωρισμού. Ωστόσο για την εύρεση του υπερεπιπέδου δεν χρησιμοποιείται το περιθώριο. Αντίθετα, το υπερεπίπεδο προκύπτει από τις τελικές τιμές των παραμέτρων του *perceptron*, δηλαδή τα βάρη και την πόλωσή του, ύστερα από εφαρμογή ενός αλγορίθμου βελτιστοποίησης της *loss function*. Συνεπώς το perceptron λύνει το πρόβλημα ταξινόμησης προτύπων δύο κλάσεων, όταν τα πρότυπα είναι γραμμικά διαχωρίσιμα. Όσο αναφορά την *activation function* υπάρχουν πολλές δυνατές επιλογές. Στην πράξη οι *activation functions* που χρησιμοποιούνται συνήθως δίνονται στον Πίνακα 3.1.

$$a = \sum_{i=1}^{n} x_i w_i + b = \mathbf{w}^T \mathbf{x} + b \tag{3.15a}$$

$$y = f(a) \tag{3.15b}$$

3.5.2 Βαθίες Αρχιτεκτονικές.

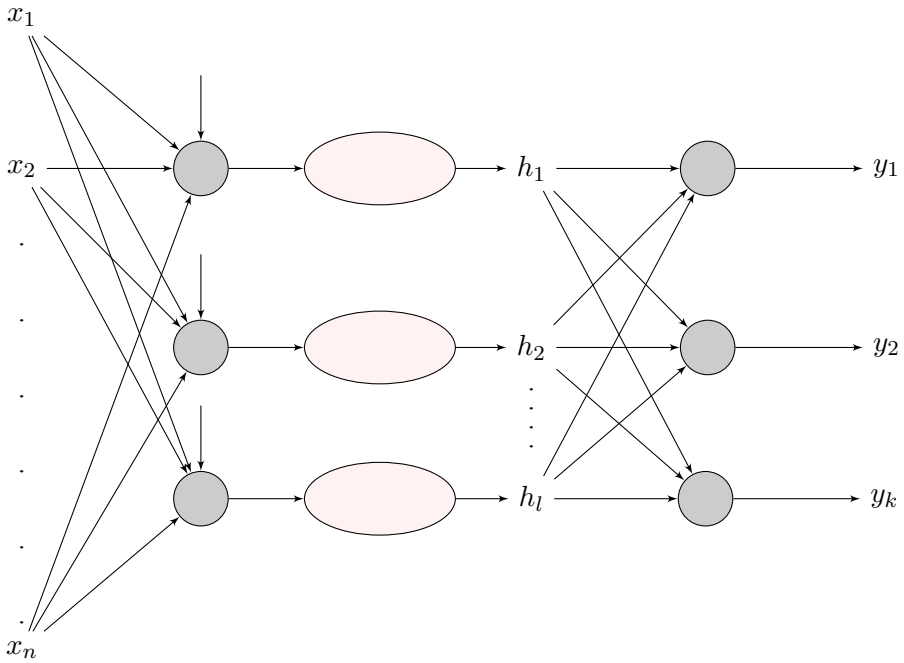
Όπως αναφέρθηκε, τα περισσότερα πραγματικά προβλήματα ταξινόμησης δεν έχουν γραμμικά διαχωρίσιμα πρότυπα ή διαθέτουν περισσότερες από δύο δυνατές κλάσεις ταξινόμησης. Συνεπώς το *perceptron* από μόνο του δεν βρίσκει ιδιαίτερη εφαρμογή. Τέτοια προβλήματα μπορούν να προσεγγιστούν μέσω βαθιών αρχιτεκτονικών.

Activation Function	Τύπος
sigmoid	$\frac{e^x}{1+e^x}$
tanh	$\frac{e^x-e^{-x}}{e^x+e^{-x}}$
relu	$max(0,x)$
softmax	$\frac{e_i^x}{\sum_i^k e_i^x}$
hard limit	$\begin{cases} 1 & x \geq 0 \\ 0 & x \leq 0 \end{cases}$

Πίνακας 3.1: Συνήθεις συναρτήσεις ενεργοποίησης.

Οι βαθιές αρχιτεκτονικές χρησιμοποιούν ως στοιχειώδη μονάδα το *perceptron*. Συνήθως διαθέτουν ένα επίπεδο εισόδου το διανύσματος **x**, κάποια κρυφά επίπεδα νευρώνων, και ένα επιπέδο εξόδου. Τα κρυφά επίπεδα αποτελούνται από perceptrons που επικοινωνούν μεταξύ τους μέσω συνάψεων - βάρη. Τόσο ο αριθμός των κρυφών επιπέδων όσο και οι συνάψεις μεταξύ νευρώνων μπορεί να διαφέρουν απο αρχιτεκτονική σε αρχιτεκτονική. Το επίπεδο εξόδου επικοινωνεί με το τελευταίο κρυφό επίπεδο. Στη γενική περίπτωση αν ο αριθμός δυνατών κλάσεων του προβλήματος, $n > 2$ η αρχιτεκτονική διαθέτει n εξόδους (στην περίπτωση των δύο κλάσεων αρκεί μία δυαδική έξοδος).

Ένα παράδειγμα βαθίας αρχιτεκτονικής φαίνεται στο Σχήμα 3.7. Το δίκτυο λαμβάνει ένα διάνυσμα εισόδου $\mathbf{x} \in \mathbb{R}^n$, το κρυφό επίπεδο μετασχηματίζει το διάνυσμα εισόδου στο διάνυσμα $\mathbf{h} \in \mathbb{R}^l$. Το επίπεδο εξόδου έχει k εξόδους δηλαδή $\mathbf{y} \in \mathbb{R}^k$. Στο παράδειγμα αυτό το δίκτυο είναι πλήρως συνδεδεμένο, δηλαδή υπάρχει σύναψη μεταξύ κάθε νευρώνα του προηγούμενου και του επόμενου επιπέδου.



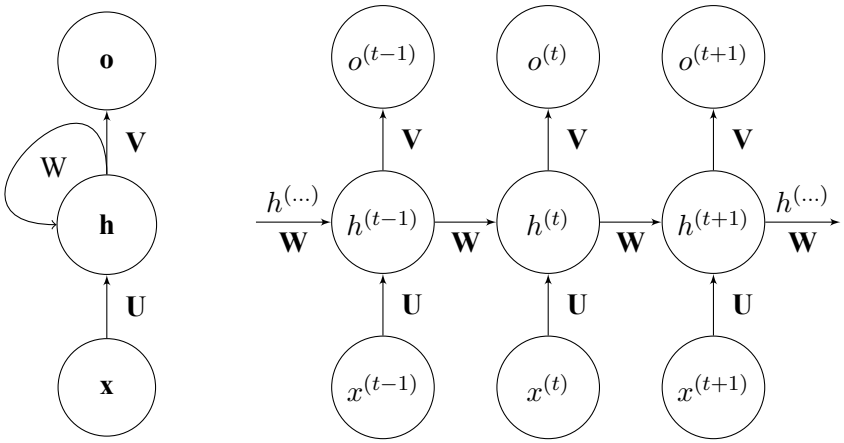
Σχήμα 3.7: Παράδειγμα αρχιτεκτονικής ενός κρυφού επιπέδου.

3.6 Αναδρομικά Νευρωνικά Δίκτυα (*Reccurent Neural Networks - RNNs*)

Τα *RNNs* ανήκουν σε ένα κλάδο νευρωνικών δικτύων ικανά να διαχειρίζονται ακολουθιακά δεδομένα της μορφής $x^{(1)}, x^{(2)}, \ldots x^{(\tau)}$ [31]. Στα κλασικά νευρωνικά δίκτυα γίνεται η υπόθεση πως τα δεδομένα εισόδου είναι ανεξάρτητα μεταξύ τους. Ωστόσο σε πολλά προβλήματα ταξινόμησης αυτή η υπόθεση είναι λάθος. Σε προβλήματα μετάφρασης προτάσεων, η μετάφραση μιας συγκεκριμένης λέξης είναι ευκολότερη δεδομένου των προηγούμενων μεταφράσεων. Υπό αυτό το πρίσμα τα RNNs διαθέτουν ένα είδος "μνήμης", μπορούν να διατηρήσουν την πληροφορία πού έχει επεξεργαστεί μέσα στο δίκτυο μέχρι την τωρινή χρονική στιγμή και έτσι να διαχειρίστούν δεδομένα εισόδου που διαφέρουν ως προς το μήκος τους. Με τον όρο χρονική στιγμή (*time step*) συνήθως αναφέρεται το δείγμα της ακολουθίας που εξετάζεται από το δίκτυο. Το δείγμα αυτό διαφέρει ανάλογα με το πρόβλημα ταξινόμησης. Σε προβλήματα Επεξεργασίας Φωνής το δείγμα μπορεί να περίεχει μια λέξη από μια πρόταση ενώ σε προβλήματα Όρασης κάποιο *pixel* της εικόνας. Στην Αναγνώριση Φωνής και Συναισθήματος το δείγμα μπορεί να αφόρα κάποιο τμήμα της ομιλίας, όπως ένα πλαίσιο ή μια ομάδα πλαισίων. Ένα τυπικό RNN δίνεται στο 3.8. Η αριστερή αναπαράσταση αντιστοιχεί στο τυλιγμένο (*folded*) *RNN*. Στα δεξία δίνεται το ξετυλιγμένο (*unfolded*) *RNN* ως προς το μήκος όλης της ακολουθίας.

Μία από τις σημαντικότερες ιδιότητες των *RNNs* είναι οι κοινές παράμετροι (*parameter sharing*), μεταξύ όλων των *time steps* της ακολουθίας [43]. Στα κλασικά νευρωνικά δίκτυα η παραπάνω ιδιότητα δεν είναι υποχρεωτική. Μάλιστα εκ πρώτης όψεως περιορίζει το δίκτυο. Εξαιτίας των κοινών παραμέτρων του δικτύου, πραγματοποιούνται οι ίδιοι υπολογισμοί σε διαφορετικά *time steps* της εισόδου. Συνεπώς το δίκτυο μπορεί να εφαρμοστεί σε ακολουθίες οι οποίες διαφέρουν ως προς το μήκος τους. Η δυνατότητα του δικτύου να διαχειρίζεται ακολουθίες διαφορετικού εξασφαλίζει την γενίκευσή του σε μήκη ακολουθιών τα οποία δεν εντοπίστηκαν κατά το στάδιο της εκπαίδευσης. Επιπλέον ο αριθμός των παραμέτρων προς εκπαίδευση μειώνεται σημαντικά.

Στη γενική περίπτωση οι παράμετροι των *RNNs* αντιστοιχούν σε 3 μήτρες βαρών $\mathbf{U}, \mathbf{W}, \mathbf{V}$. Η



Σχήμα 3.8: Σχηματική Αναπαράσταση ενός *RNN*.

πρώτη επίδρα στα *time steps* $x^{(t)}$ της ακολουθίας. Η δεύτερη αφορά την σύνδεση των κρυφών καταστάσεων (*hidden state*) δύο διαδοχικών time steps. Τέλος η τρίτη επιδρά στην κρυφή κατάσταση του δικτύου την χρονική στιγμή t παράγοντας την έξοδο $o^{(t)}$. Οι μήτρες αυτές συνοδεύονται από την συνάρτηση ενεργοποίησης κρυφής κατάστασης, f_h, και εξόδου, f_o. Το δίκτυο μπορεί να περιγραφεί από τις ακόλουθες εξισώσεις :

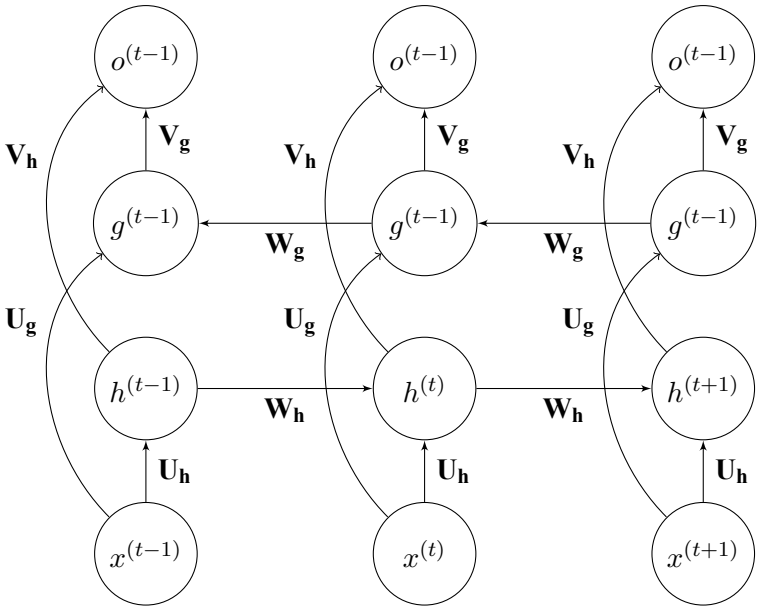
$$h_t = f_h(Ux_t + Wh_{t-1}) \tag{3.16a}$$

$$o_t = f_o(Vh_t) \tag{3.16b}$$

Η *hidden state*, h_t, του *RNN* λειτουργεί ως "μνήμη" του δικτύου. Περιέχει όλη την πληροφορία του δικτύου μέχρι την χρονική στιγμή t. Για τον υπολογισμό της τωρινής κρυφής κατάστασης χρειάζεται ο υπολογισμός όλων των προηγούμενων. Αντίθετα η έξοδος του δικτύου $o^{(t)}$ υπολογίζεται δεδομένου μόνο της τωρινής κρυφής κατάστασης του δικτύου.

Στη διάταξη του Σχήματος 3.8, το δίκτυο παράγει σε κάθε χρονική στιγμή έξοδο. Η υπόθεση αυτή δεν είναι υποχρεωτική. Σε αρκετές εφαρμογές το δίκτυο παράγει μια έξοδο μετά την παρατήρηση όλης της ακολουθίας. Η έξοδος αυτή $o(\tau)$ περιγράφει όλους τους ενδιάμεσους υπολογισμούς του δικτύου. Επιπλέον σε ορισμένες εφαρμογές προϋπόθέτουν πως η έξοδος την χρονική στιγμή t, διαμορφώνεται τόσο από προηγούμενα όσο και από επόμενα time steps της ακολουθίας. Για τον σκοπό αυτό χρησιμοποιούνται τα Αμφίδρομα Αναδρομικά Δίκτυα (*Bidirectional Recurrent Neural Networks - BRNNs*) [44]. Τα *BRNNs* αποτελούνται από δύο απλά *RNNs* τα οποία παράγουν τις ακολουθίες $h()$ και $g()$ αντίστοιχα (βλ. Σχήμα 3.9). Κάθε επιμέρους *RNN* διαθέτει τις δικές του παραμέτρους. Η ακολουθία $h(.)$ παράγεται παρατηρώντας την ακολουθία από την χρονική στιγμή $t = 0$ έως $t = \tau$, ενώ η $g(.)$ παρατηρώντας την ακολουθία με αντίστροφη σειρά. Η έξοδος σε κάθε χρονική στιγμή αποτελείται από την εφαρμογή της *activation function* στην συνδυασμένη είσοδο $[h(t), g(t)]$.

Τέλος, όπως και στην περίπτωση των κλασικών νευρωνικών δικτύων είναι η δυνατή η στοίβαξη *RNNs* για την δημιουργία Βαθέων Αναδρομικών Νευρωνικά Δικτύων (*Deep Reccurent Neural Networks - DRNNs*). Σε αυτήν την περίπτωση η έξοδος του επιπέδου i την χρονική στιγμή t τροφοδοτείται ως είσοδος στο επίπεδο $i + 1$. Στο Σχήμα 3.10 δίνεται ένα *RNN* 2 επιπέδων. Επεκτείνοντας την ιδέα των DRNNs μπορούν να χρησιμοποιηθούν και Βαθιά Αμφίδρομα Αναδρομικά Δίκτυα (*Deep Bidirectional Recurrent Neural Networks - DBRNNs*).



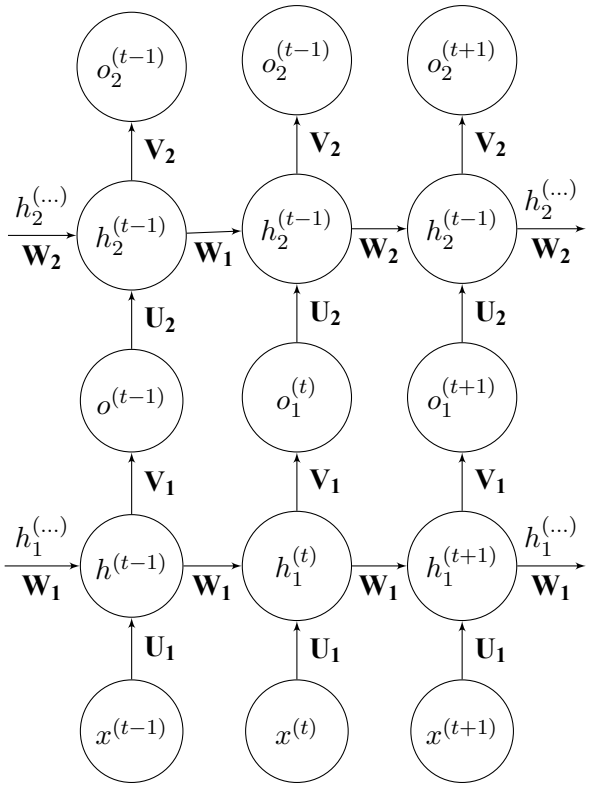
Σχήμα 3.9: Σχηματική Αναπαράσταση ενός *BRNN*.

3.6.1 Δίκτυα Μακράς - Βραχείας Μνήμης (*Long - Short Term Memory Networks - LSTM Networks*)

Στην θεωρία τα *RNNs* είναι ικανά να συγκρατήσουν την πληροφορία από προηγούμενα time steps. Ωστόσο ενδέχεται το δίκτυο να χρειαστεί κάποια πληροφορία η οποία βρίσκεται στα πρώτα *time steps* της ακολουθίας. Για μεγάλου μήκους ακολουθίες η απόσταση μεταξύ των τωρινών και αρχικών *time steps* καθιστά αδύνατη την πρόσβαση στην παραπάνω πληφορορία. Συνεπώς το δίκτυο αποτυγχάνει. Το πρόβλημα αυτό αναφέρεται ως *Challenge of Long Term Dependencies* [45].

Για την επίλυση του παραπάνω προβλήματος χρησιμοποιούνται τα Δίκτυα Μακράς - Βραχείας Μνήμης (*Long - Short Term Memory Networks - LSTM Networks*) [46]. Μολονότι τα *LSTMs* δεν διαφέρουν δομικά από τα *RNNs* υπολογίζουν το *hidden state* με διαφορετικό τρόπο. Η ”μνήμη” των *LSTMs* αποτελείται από στοιχειώδεις κυψελίδες (*cell*). Σε κάθε τέτοια κυψελίδα αποφασίζεται ποία τμήματα της μνήμης θεωρούνται χρήσιμα και στη συνέχεια αποθηκεύονται στη μνήμη. Συνδυάζοντας την προηγούμενη *hidden state*, την τωρινή μνήμη και είσοδο, υπολογίζεται η επόμενη κατάσταση.

Στα RNNs η σύναψη μεταξύ προηγούμενης και τωρινής κρυφής κατάστασης περιγράφεται μέσω της εξίσωσης 3.16a. Αντίθετα στα *LSTMs*, η σύναψη αυτή περιγράφεται μέσω 4 επιμέρους σχέσεων. Για τον σκοπό αυτό κάθε cell διαθέτει 3 πύλες : απόρριψης (*forget gate*), εισόδου (*input gate*), και εξόδου (*output gate*). Το διάνυσμα κατάστασης του cell $\mathbf{C_t}$, επηρεάζεται από τις *forget* και *input gate*, ενώ το διάνυσμα εξόδου του *cell* $\mathbf{o_t}$ διαμορφώνεται από την *output gate*. Στο Σχήμα 3.11 δίνεται η αναπαράσταση ενός *cell*. Παρακάτω περιγράφεται η λειτουργία των πυλών για την διάταξη του σχήματος.



Σχήμα 3.10: Σχηματική Αναπαράσταση ενός *RNN* 2 επιπέδων.

Forget Gate

Αρχικά καθορίζονται τα χρήσιμα τμήματα της μνήμης. Η *forget gate* λαμβάνει την συνδυασμένη είσοδο της προηγούμενης εξόδου και της τωρινής εισόδου. Οι παράμετροί της αντιστοιχούν σε μια μήτρα βαρών $\mathbf{W}_f$ καθώς και ένα διάνυσμα πόλωσης $\mathbf{b}_f$. Το διάνυσμα εξόδου $\mathbf{f}_t$ διάστασης ίδιας με το $\mathbf{C}_t$, περιέχει μονάδες και μηδενικά στις χρήσιμες και άχρηστες θέσεις μνήμης αντίστοιχα:

$$\mathbf{f}_t = sigmoid\Big(\mathbf{W}_f[\mathbf{o}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f\Big) \tag{3.17}$$

Input Gate

Στη συνέχεια υπολογίζονται τα υποψήφια νέα τμήματα προς αποθήκευση στη μνήμη. Σε πρώτο στάδιο παράγεται το διάνυσμα που προσδιορίζει τις τιμές οι οποίες ενημερώνονται $\mathbf{i_t}$. Όπως και προηγουμένως η πύλη λαμβάνει την συνδυασμένη είσοδο της προηγούμενης εξόδου και της τωρινής εισόδου. Οι αντίστοιχες παράμετροι είναι η μήτρα βαρών $\mathbf{W}_i$ και το διάνυσμα πολώσεων $\mathbf{b}_i$. Οι υποψήφιες μεταβολές στην μνήμη $\bar{\mathbf{C}}_t$ υπολογίζονται μέσω μιας ξεχωριστής μήτρας $\mathbf{W}_c$, και πόλωσης $\mathbf{b_c}$. Το διάνυσμα κατάστασης του *cell* ανανεώνεται μέσω των $\mathbf{f}_t$, $\mathbf{i}_t$ και $\bar{\mathbf{C}}_t$:

$$\mathbf{i}_t = sigmoid\Big(\mathbf{W}_i[\mathbf{o}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i\Big) \tag{3.18a}$$

$$\bar{\mathbf{C}}_t = tanh\Big(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c\Big) \tag{3.18b}$$

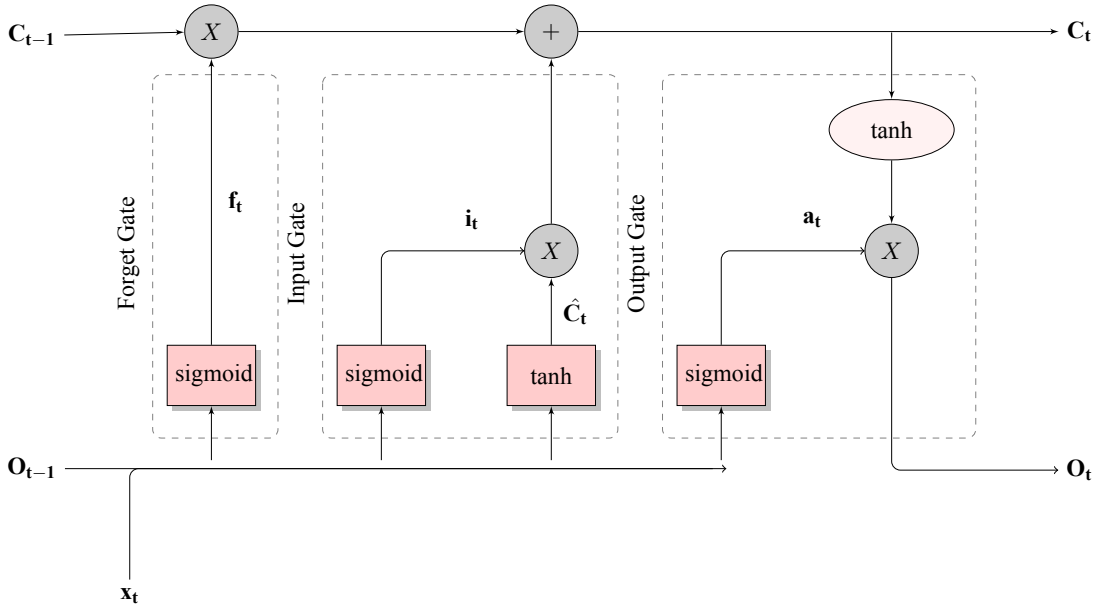
$$\mathbf{C}_t = \mathbf{f}_t * \mathbf{C}_{t-1} + \mathbf{i}_t * \bar{\mathbf{C}}_t \tag{3.18c}$$

Output Gate

Τέλος υπολογίζεται η έξοδος του cell για την χρονική στιγμή t :

$$\mathbf{a}_t = sigmoid\Big(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o\Big) \tag{3.19a}$$

$$\mathbf{o}_t = \mathbf{a}_t * tanh\Big(\mathbf{C}_t\Big) \tag{3.19b}$$



Σχήμα 3.11: Σχηματική αναπαράσταση ενός *LSTM* cell.

3.6.2 Ο μηχανισμός *Attention*

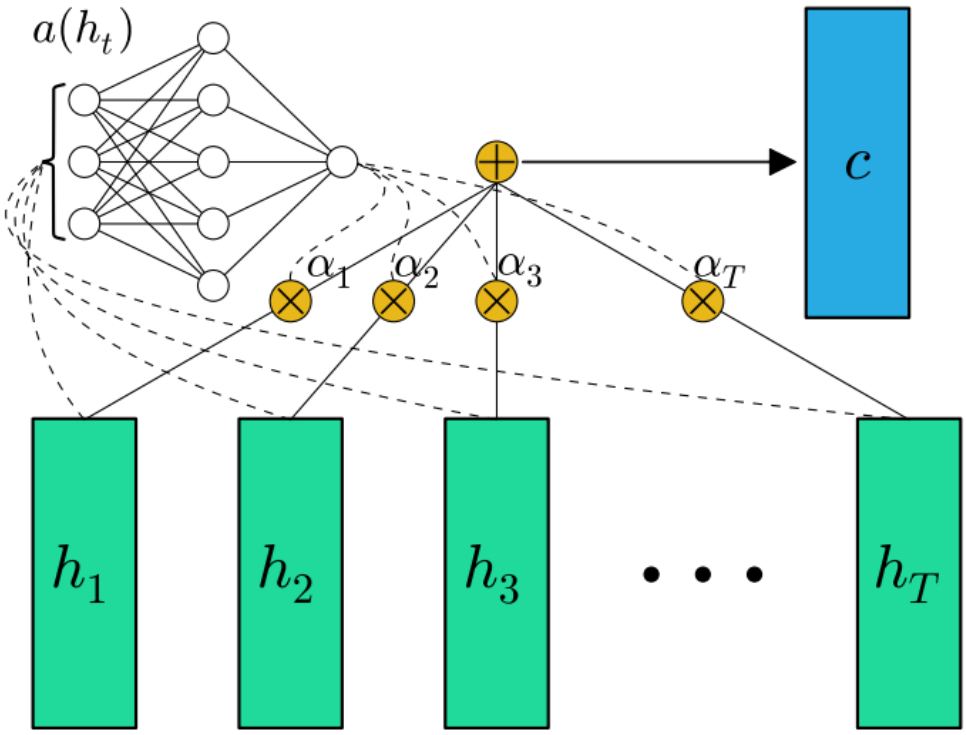
Ο άνθρωπος συνήθως σε προβλήματα επεξεργασίας ακολουθιακών δεδομένων δίνει περαιτέρω σημασία στο παρόν γεγονός, παρά σε προηγούμενα ή μελλοντικά. Ενδεικτικά κατά την μετάφραση ενός κειμένου δίνεται κυρίως έμφαση στην επί του παρόντος μεταφραζόμενη λέξη. Στα νευρωνικά δίκτυα η ικανότητα αυτή επιτυγχάνεται μέσω του μηχανισμού *Attention* [47]. Στην πράξη ο μηχανισμός αυτός αντιστοιχεί ως ένα επιπλέον επίπεδο νευρώνων στο δίκτυο και αποδίδει συντελεστές - βάρη στις εξόδους που παράγονται από ένα ακολουθιακό μοντέλο, εστιάζοντας σε συγκεκριμένες από αυτές.

Στο Σχήμα 3.12 δίνεται η σχηματική αναπαράσταση του μηχανισμού. Αν h_t είναι η έξοδος του μοντέλου σε κάθε *time step*, τότε υπολογίζεται ένα διάνυσμα συμφραζόμενων (*"context" vector"*), $\mathbf{c}$ [48]. Το διάνυσμα αυτό αντιστοιχεί στον σταθμισμένο μέσο όρο των εξόδων h_t, $t = 1, 2, \ldots, T$:

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t h_t \tag{3.20}$$

Τα βάρη a_t υπολογίζονται μέσω των :

$$e_t = f(h_t) \tag{3.21a}$$



Σχήμα 3.12: Σχηματική αναπαράσταση του μηχανισμού Attention [8].

$$a_t = softmax(e_t) = \frac{exp(e_t)}{\sum_{t=1}^{T} exp(e_k)} \tag{3.21b}$$

Όπως φάινεται και στην εξίσωση 3.20 ο μηχανισμός *Attention* επιστρέφει ένα διάνυσμα διάστασης ίσης με το μήκος της ακολουθίας. Τα βάρη $a_t \in [0, 1]$ προσδιορίζονται με εφαρμογή της *softmax* σε κάθε συντεταγμένη e_t της εξόδου της συνάρτησης ενεργοποίησης του επιπέδου f. Υψηλές τιμές στα βάρη a_t αντιστοιχούν σε time steps μεγάλης σημασίας. Ουσιαστικά μέσω το μηχανισμού το δίκτυο δίνει έμφαση σε συγκεκριμένες περιοχές στον χώρο των εξόδων.

3.7 *Single and Multi - tasking*

Μέχρι στιγμής, για ένα πρόβλημα Αναγνώρισης Προτύπων εκπαιδεύεται ένα μοντέλο μέσω αλγορίθμων βελτιστοποίησης κάποιας μετρικής. Σε γενικές με την κλασική αυτή τεχνική (*single - tasking*) επιτυγχάνεται μια αξιοπρεπής επίδοση. Ωστόσο εξετάζοντας τον τρόπο λειτουργίας του ανθρώπινου μυαλού, παρατηρείται πως εφαρμόζοντας *single - tasking* συνήθως αγνοείται ένα τμήμα πληροφορίας ικανό να προσφέρει επιπλέον βελτίωση στο μοντέλο. Ο ανθρώπινος εγκέφαλος λύνει πολλά προβλήματα ταυτοχρόνα και μεταδίδει την γνώση που αποκτά από ένα πρόβλημα για να λύσει ένα άλλο. Ως *multi - tasking* αναφέρεται η ικανότητα ενός μοντέλου να λύνει πολλά προβλήματα ταυτόχρονα. Θεωρώντας ότι χρησιμοποιείται ένα μοντέλο για κάθε πρόβλημα, τότε για την επίλυση n προβλημάτων θα χρειαζόταν η εκπαίδευση n μοντέλων. Από την άλλη πλευρά, επιτρέποντας ένα μοντέλο να λύσει τα n προβλήματα ταυτόχρονα τότε εκπαιδεύεται μόνο ένα μοντέλο ενώ πιθανώς επιτυγχάνει καλύτερη επίδοση από τα n ξεχωριστά μοντέλα.

Το *multi - tasking* ενισχύει την ικανότητα γενίκευσης ενός μοντέλου επεκτείνοντας την πληροφορία που περιέχεται στα δεδομένα εκπαίδευσης μετάξυ παρόμοιων προβλημάτων [49]. Η έννοια του multitask έχει χρησιμοποιηθεί σε προβλήματα Όρασης Υπολογιστών [50], Επεξεργασίας Φωνής [51] και Φυσικής Γλώσσας καθώς και Αναγνώρισης Φωνής [52]. Για ένα νευρωνικό δίκτυο, οι δύο πιο γνωστές μέθοδοι *multi - tasking* είναι η ολική (*hard*) και η μερική (*soft*) *parameter sharing*. Κατά *hard parameter sharing* το δίκτυο μοιράζεται όλα τα κρυφά του επίπεδα με όλα τα προβλήματα, ενώ διαθέτει μερικά επιπέδα κοντά στην έξοδο συγκεκριμένα για κάθε προβλήμα [53]. Σε γενικές γραμμές, η *hard parameter sharing* δεν επιτρέπει στο δίκτυο να γενικευτεί καθώς για να λύσει όλα τα προβλήματα, το μοντέλο αναγκάζεται να μάθει γενικού τύπου χαρακτηριστικά [54]. Από την άλλη στη *soft parameter sharing* κάθε πρόβλημα διαθέτει το δικό του μοντέλο με τις δικές του παραμέτρους. Με σκοπό να υπάρξει σχετική ομοιομορφία μεταξύ των παραμέτρων, εισάγονται συνθήκες κανονικοποίησης όπως η νόρμα l_2 στην *loss function* δικτύου [55]. Τέλος και στις δύο μεθόδους κάθε πρόβλημα διαθέτει την δικία του *loss function*. Το συνολικό δίκτυο εκπαιδεύεται μέσω συνήθων αλγορίθμων βελτιστοποίησης της ολικής *loss function* του δικτύου, η οποία αντιστοιχεί στο άθροισμα των επιμέρους *loss function* του κάθε προβλήματος.

Κεφάλαιο 4

Υπόβαθρο Εργασίας

4.1 Εισαγωγή

Σκοπός των μοντέλων της μηχανικής μάθησης είναι ικανότητα να λαμβάνουν καλές αποφάσεις. Για την λήψη της απόφασης μια μηχανή χρειάζεται γνώση η οποία προέρχεται από τα παραδείγματα εκπαίδευσης. Με την πάροδο του χρόνου έγινε μεγάλη προσπάθεια από ερευνητές έτσι ώστε να μεταδώσουν την ανθρώπινη γνώση στην μηχανή. Ωστόσο δεν είναι ακόμα σαφές πώς ο άνθρωπος μαθαίνει και κατά συνέπεια η προσπάθεια μετάδοσης δεν επέφερε σημαντικά αποτελέσματα. Σε γενικές γραμμές ο άνθρωπος μαθαίνει παρατηρώντας το περιβάλλον γύρω του, γενικεύοντας από παραδείγματα τα οποία έχει ήδη παρατηρήσει σε νεά. Η διαδικασία γενίκευσης αναφέρεται ως μάθηση (*learning*). Συνήθως τα παραδείγματα γενίκευσης περιέχουν ένα είδος πληροφορίας η οποία εκφράζεται μέσω μιας αναπαράστασης (*representation*). Για παράδειγμα ο άνθρωπος καταλαβαίνει την υφή των αντικειμένων μέσω της δύναμης αντίδρασης κατά την επαφή με το αντικείμενο. Αντίστοιχα λαμβάνει τα ακουστικά μηνύματα μέσω κυμάτων διαμορφωμένα σε διαφορετικά μήκη και πλάτη.

Στην πράξη ο άνθρωπος εφαρμόζει κάποια επεξεργασία της πληροφορίας που εκφράζεται με την αντίστοιχη *representation* με σκοπό να λύσει ένα πρόβλημα. Πολλά από αυτά τα προβλήματα εξαρτώνται από τον τρόπο με τον οποίο αναπαρίσταται η πληροφορία. Είναι σχετικά απλό για τον άνθρωπο να υπολογίσει το άθροισμα 2 αριθμών εκφρασμένων στο αραβικό αριθμητικό σύστημα. Από την άλλη πλευρά, ο υπολογισμός γίνεται λιγότερο απλός αν οι αριθμοί είναι εκφρασμένοι στο δυαδικό αριθμητικό σύστημα. Με άλλα λόγια η επίλυση ενός προβλήματος εξαρτάται άμεσα από τον τρόπο απεικόνισης της πληροφορίας.

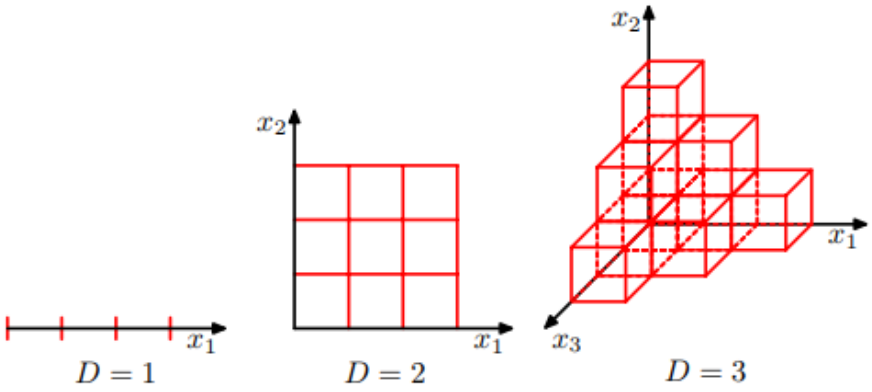
Η ίδια ιδέα επεκτείνεται και στις μηχανές. Όπως και στην περίπτωση των ανθρώπων η απόδοση μιας μηχανής εξαρτάται άμεσα από την *representation* των δεδομένων. Για τον λόγο αυτό έχουν αναπτυχθεί τεχνικές μετασχηματισμού των δεδομένων έτσι ώστε οι αλγόριθμοι μάθησης να μπορούν να εφαρμοστούν ευκολότερα στα νέα αναπαριστώμενα δεδομένα. Ο κλάδος που ασχολείται με τους μετασχηματισμούς αυτούς αναφέρεται ως *representation learning* και εξετάζει την εύρεση *representations* των δεδομένων έτσι ώστε να είναι ευκολότερη η εξαγωγή της χρησιμης πληροφορίας [56].

Στο προηγούμενο κεφάλαιο δόθηκε έμφαση στον φορμαλισμό των μοντέλων καθώς και των ικανοτήτων τους, με την ελπίδα ότι τα μοντέλα επιλύουν το εκάστοτε πρόβλημα ανεξάρτητα από την *representation* των δεδομένων. Το παρόν κεφάλαιο εστιάζει στις *representations* των δεδομένων. Αρχικά παρουσιάζεται η πρόκληση που προκύπτει προβάλλοντας τα δεδομένα σε χώρους χαρακτηριστικών μεγάλης διαστασιμότητας. Στη συνέχεια αναλύονται οι βασικές έννοιες του *representation learning*. Περιγράφονται οι επιθυμητές ιδιότητες μιας *representation* και αναλύονται δύο αλγόριθμοι εξαγωγής *representation* των δεδομένων. Τέλος εξετάζεται η μεταφορά γνώσης μεταξύ δύο μοντέλων, η ικανότητα δηλαδη μεταφοράς την γνώσης ενός μοντέλου το οποίο χρησιμοποιείται για την επίλυση ενός αρχικού προβλήματος σε ένα δεύτερο μοντέλο για την επίλυση ενός διαφορετικού προβλήματος.

4.2 *Curse of Dimensionality*

Στις περισσότερες εφαρμογές Αναγνώρισης Προτύπων τα δεδομένα προβάλλονται σε χώρους χαρακτηριστικών μεγάλης διαστασιμότητας. Εκ πρώτης όψεως οι χώροι αυτοί διευκολύνουν την ταξινόμηση των προτύπων, καθώς μεγάλη διαστασιμότητα του χώρου συνεπάγεται πληθώρα διαθέσιμων τιμών των μεταβλητών των χαρακτηριστικών. Έτσι υπάρχει μεγάλη ελευθερία στην αναπαράσταση των προτύπων και είναι ευκολότερη η εύρεση περιοχών στις οποίες ομαδοποιούνται τα πρότυπα της ίδιας κλάσης. Ωστόσο η παραπάνω υπόθεση θέτει σημαντικές προκλήσεις. Ο όρος *Curse of Dimensionality* [57] χρησιμοποιείται για να περιγράψει τις προκλήσεις που προκύπτουν κατά την επεξεργασία δεδομένων σε μεγάλες διαστάσεις.

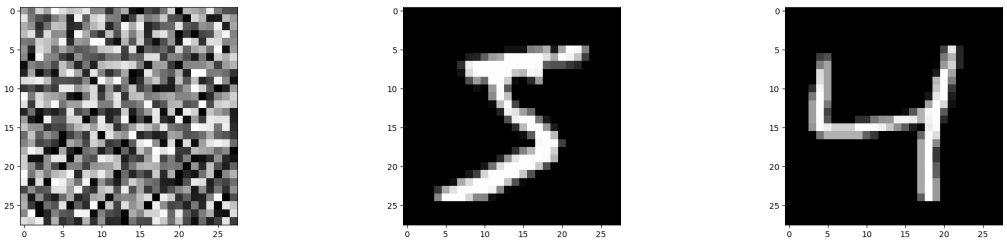
Για την κατανόηση των προκλήσεων θεωρείται ένα απλό πρόβλημα ταξινόμησης. Μια απλή προσέγγιση είναι ο κατακερματισμός του χώρου χαρακτηριστικών σε περιοχές. Τα δείγματα που ανήκουν στην ίδια κλάση οφείλουν να βρίσκονται κοντά μεταξύ τους. Έτσι, για ένα νέο δείγμα x προς ταξινόμηση αρχικά προσδιορίζεται η περιοχή στην οποία ανήκει και στην συνέχεια κατατάσσεται στην κλάση η οποία διαθέτει τα περισσότερα δείγματα μέσα σε αυτήν την περιοχή. Αν κάθε δείγμα περιγράφεται από ένα χαρακτηριστικό η παραπάνω προσέγγιση είναι εφικτή. Ωστόσο, όπως φαίνεται και στο Σχήμα 4.1 καθώς αυξάνεται η διαστασιμότητα του χώρου, τότε ο αριθμός των περιοχών αυξάνεται εκθετικά. Προκειμένου να υπάρχουν μη κενές περιοχές χρειάζεται τουλάχιστον ένα δείγμα για κάθε περιοχή του χώρου. Σε πρακτικές εφαρμογές, τα διαθέσιμα δείγματα είναι περιορισμένα. Έτσι σε κάθε περιοχή τοποθετούνται ελάχιστα εώς και κανένα δείγματα, συνεπώς οι χώροι μεγάλων διαστάσεων οδηγούν σε αραιή αναπαράσταση των δεδομένων. Μια αραιή αναπαράσταση συνήθως δυσκολεύει την ικανότητα γενίκευσης καθώς η απόσταση των δεδομένων ίδιων κλάσεων μπορεί να γίνει μεγάλη, καθιστώντας την παραπάνω προσέγγιση απαγορευτική.



Σχήμα 4.1: Παράδειγμα της εκθετικής αύξησης των περιοχών του χώρου καθώς αυξάνεται η διαστασιμότητά του [9].

Η παραπάνω προσέγγιση προϋπόθετει ότι όλες οι περιοχές του χώρου είναι χρήσιμες για το πρόβλημα της ταξινόμησης. Ευτυχώς, σε ρεαλιστικές περιπτώσεις σπανίως χρησιμοποιούνται όλες οι περιοχές του χώρου. Ενδεικτικά στο πρόβλημα αναγνώρισης χειρόγραφων ψηφίων, μια δυαδική εικόνα 28×28 *pixels* αναπαρίσταται σε χώρο διάστασης 784. Τα συνολικά διανύσματα που περιέχονται στον χώρο είναι 2^{784} και αντιστοιχούν σε όλες τις δυνατές 28×28 ψηφιακές εικόνες. Θεωρώντας πως τα διανύσματα αυτά προέρχονται από μια κατανομή *Bernoulli*, τότε δειγματοληπτώντας από την κατανομή λαμβάνονται τυχαία διανύσματα του χώρου, δηλαδή τυχαίες ψηφιακές 28×28 εικονές. Στο Σχήμα 4.2 δίνονται 3 ψηφιακές εικόνες 28×28 *pixels*. Η αριστερή αντιστοιχεί σε ένα τυχαίο διάνυσμα, ένω οι υπόλοιπες προέρχονται από το σύνολο ψηφιακών χειρόγραφων ψηφίων *MNIST* [58].

Όπως είναι φανερό το τυχαίο διάνυσμα του χώρου δεν μπορεί να θεωρηθεί ως κάποιο ρεαλιστικό δείγμα ψηφίου. Τα ρεαλιστικά δείγματα ψηφίων διαγράφουν συγκεκριμένες καμπύλες στον χώρο, δηλαδή μέσα στον 784-διάστατο χώρο εμπεριέχονται καμπύλες χαμηλότερης διάστασης (*manifolds*) ικανές να περιγράψουν τα εκάστοτε ψηφία. Με άλλα λόγια η πραγματική διαστασιμότητα ενός προβλήματος ενδέχεται να είναι μικρότερη από εκεινή στην οποία αναπαρίστανται τα δεδομένα.



Σχήμα 4.2: Παράδειγμα ψηφιακών εικόνων.

4.3 *Representation Learning*

Όπως φάνηκε στην προηγούμενη παράγραφο η επιλογή της διαστασιμότητας του χώρου χαρακτηριστικών παίζει καθοριστικό ρόλο στην επίδοση μιας υπολογιστικής μηχανής. Χώροι χαμηλής διαστασιμότητας δίνουν συνήθως φτωχή *representation* των δεδομένων και έτσι ενδέχεται τα πρότυπα διαφορετικών κλάσεων να βρίσκονται αρκετά κοντά μεταξύ τους. Από την άλλη πλευρά χώροι υψηλής διαστασιμότητας τοποθετούν τα πρότυπα αρκετά μακριά μεταξύ τους στερώντας την ικανότητα γενίκευσης του μοντέλου. Ο κλάδος *representation learning* πραγματεύεται τις ιδιότητες που οφείλει να έχει μια καλή *representation*. Με βάση τις ιδιότητες αυτές αναπτύσσονται αλγόριθμοι μετασχηματισμού των δεδομένων και εφαρμόζονται σε διάφορα προβλήματα όπως Αναγνώριση Φωνής [59, 60], Όρασης Υπολογιστών, [61, 62] αλλά και Επεξεργασίας Κειμένου [63, 64].

Σε γενικές γραμμές, μια καλή *representation* είναι εκεινή στην οποία το πρόβλημα λύνεται ευκολότερα μέσω των μετασχηματισμένων δεδομένων. Πιο συγκεκριμένα μια καλή *representation* συνήθως διαθέτει μια συνθήκη ομαλότητας. Αν f είναι η συνάρτηση προς μάθηση και ισχύει $x \approx y$ τότε $f(x) \approx f(y)$. Ένα άλλο στοιχείο που ξεχωρίζει σε μια καλή *representation* είναι η υπάρξη πολλών περιγραφών οργανωμένων σε μια ιεραρχική δομή, ξεκινώντας από τις πιο ειδικές και καταλήγοντας στις πιο γενικές, όπως στην περίπτωση των βαθέων νευρωνικών δικτύων. Σε άλλες περιπτώσεις μια καλή *representation* περιέχει κάποιο *manifold*, κάποιο φυσικό κατακερματισμό ή ικανότητα αραιών περιγραφών [56].

Ο κλάδος *representation learning* φέρει ιδιαίτερο ενδιαφέρον καθώς περιλαμβάνει *supervised*, *semi - supervised* και *unsupervised* αλγορίθμους Στα προβλήματα Αναγνώρισης Προτύπων, λαμβάνονται χειροτεχνικά χαρακτηριστικά από τα αυτούσια δεδομένα. Τα χαρακτηριστικά αυτά συνήθως προέρχονται από εμπειρική γνώση ή κρίση του ανθρώπου για το εκάστοτε πρόβλημα. Ωστόσο για πολλά προβλήματα είτε η εμπειρική γνώση και κρίση είναι ελλιπείς είτε η κατασκεύη χειροτεχνικών χαρακτηριστικών είναι αρκετά χρονοβόρα. Αντίθετα οι *representation learning* παρέχουν τρόπους μάθησης χρήσιμων χαρακτηριστικών (*feature learning*) για το πρόβλημα.

Τα νευρωνικά δίκτυα που χρησιμοποιούν *supervised* αλγορίθμους εκπαίδευσης μπορούν να θεωρήθούν επίσης ως μια τεχνική *representation learning*. Συχνά το τελευταίο επίπεδο του δικτύου περιλαμβάνει ένα γραμμικό ταξινομητή, όπως μια συνάρτηση *softmax*. Το υπόλοιπο δίκτυο αποσκοπεί στην εύρεση μιας *representation* για τον ταξινομητή. Η εκπαίδευση του δικτύου οδηγεί σε μια κλιμάκωση της *representation* του δικτύου όπου τα πρώτα επίπεδα του δικτύου αντιστοιχούν σε χαμηλού

επιπέδου περιγράφες ενώ τα τελευταία επίπεδα παρέχουν μια πιο υψηλού επιπέδου περιγραφή.

Από την άλλη πλευρά, σε αρκετές εφαρμογές η πλειονότητα του συνόλου δεδομένων δεν περιέχει ετικέτες. Τέτοιες περιπτώσεις αναφέρονται ως *semi supervised learning*. Αν η εκπαίδευση ενός μοντέλου γίνει μόνο μέσω του υποσυνόλου των δεδομένων εκπαίδευσης με ετικέτες τότε ενδέχεται το μοντέλο να υπερειδικευτεί *overfit* σε αυτά, αδυνατόντας να γενικευτεί σε νέα δεδομένα. Μια λύση είναι η εξαγωγή *representations* μέσω *unsupervised* αλγορίθμων πάνω στο σύνολο εκπαίδευσης χωρίς ετικέτες και η χρήση τους για την εκπαίδευση ενός μοντέλου με τα δεδομένα που διαθέτουν ετικέτες.

Τέλος υπάρχουν αποκλειστικά *unsupervised representation learning* τεχνικές. Για οποιοδήποτε πρόβλημα στόχος είναι η εκμάθηση αναλλοίωτων χαρακτηριστικών (*invariant features*), δηλαδή χαρακτηριστικά τα οποία για μικρές μεταβολές επηρεάζουν την έξοδο του ταξινομητή. Προφανώς υπάρχουν διαφορετικά *invariant features* για κάθε πρόβλημα. Αν τα δεδομένα αφορούν σήματα φωνής τότε για το πρόβλημα Αναγνώρισης Φωνής τα *invariant features* δεν περιλαμβάνουν την ένταση του μικροφώνου ηχογράφησης. Αντίστοιχα αν το πρόβλημα αφορά την ταυτοποίηση του ομιλητή, δεν θα πρέπει να δίνεται έμφαση στα λόγια του ομιλητή. Οι *unsupervised representation learning* τεχνικές επεκτείνουν την ιδέα μαθαίνοντας γενικού τύπου χαρακτηριστικά (*disentangling features*). Ακριβώς επειδή το πρόβλημα σε αυτές τις περιπτώσεις δεν είναι γνωστό, οι *unsupervised* τεχνικές βρίσκουν χρήσιμη πληροφορία ανεξαρτήτο προβλήματος.

Στην παρούσα εργασία χρησιμοποιούνται οι *unsupervised representation learning* τεχνικές. Πιο συγκεκριμένα εφαρμόζεται η κλάσικη μέθοδος *Principal Compoment Analysis - PCA*, μια από τις κλασικές μεθόδους *feature extraction*, η οποία αποσκοπεί στην εξαγωγή μιας καλύτερης *representation* σε ένα χώρο μικρότερης διαστασιμότητας. Επιπλέον χρησιμοποιούνται οι *autoencoders*, δηλάδη νευρωνικά δίκτυα τα οποία παρέχουν χρησίμες *representations* μέσω των ενδιάμεσων κρυφών επιπέδων.

4.4 *Principal Component Analysis - PCA*

Με τον όρο *feature extraction* αναφέρεται η εξαγωγή νέων χαρακτηριστικών έτσι ώστε η νέα *representation* των δεδομένων να έχει αφενός λιγότερος μεταβλητές και αφετέρου να διατηρεί όσο το δυνατό περισσότερη πληροφορία για τα δεδομένα. Ως πληροφορία συνήθως αναφέρεται η δομή του συνόλου δεδομένων, δηλαδή οι αποστάσεις μεταξύ τους. Πιο συγκεκριμένα έστω ότι η αρχική διάσταση των χαρακτηριστικών των δεδομένω ισούται με D. Η βασική ιδέα του *feature extraction* είναι η κατασκευή M νέων διαστάσεων συνδιάζοντας τα αρχικά δεδομένα μέσω ενός μετασχηματισμού $f(.)$:

$$N_i = f(X_1, X_2, \ldots, X_d), \quad i = 1, 2, \ldots M \tag{4.1}$$

Τόσο οι *autoencoders* όσο και η μέθοδος *PCA* κατασκευάζουν τον μετασχηματισμό $f(.)$ και στη συνέχεια προσδιορίζουν το νέο σύνολο διαστάσεων $\mathbf{N} = \{N_1, N_2, \ldots, N_m\}$. Ωστόσο στους *autoencoders* ο μετασχηματισμός $f(.)$ είναι αυθάιρετος, ενώ στη *PCA* ο μετασχηματισμός είναι γραμμικός.

Στην πράξη η *PCA* ορίζει ένα σύστημα M κυρίων συνιστωσών. Οι συνιστώσες αυτές αντιστοιχούν στους κάθετους άξονες πάνω στους οποίους το σύνολο δεδομένων $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{R}^D$, παρουσιάζει την μεγαλύτερη διασπορά. Μέτα την εύρεση του συστήματος συντεταγμένων, το σύνολο δεδομένων μετασχηματίζεται προβάλλοντας τα $\mathbf{x}_i$ στο νέο σύστημα αξόνων. Η επιλογή των αξόνων με την μεγαλύτερη διασπορά διατηρεί τις αποστάσεις μεταξύ ενός ζεύγους δειγμάτων. Έτσι δύο δείγματα που βρίσκονται αρχικά μακρία στον χώρο των χαρακτηριστικών, συνεχίζουν να βρίσκονται μακριά στον χώρο που δημιουργούν οι νέοι άξονες.Στην απλή περίπτωση όπου $M = 1$ αναζητείται το διάνυσμα $\mathbf{e}_1$ το οποίο αντιστοιχεί στην κατεύθυνση της μεγαλύτερης διασποράς. Κάθε δεδομένο $\mathbf{x}_k$, προβάλλεται σε μία σταθερά που ισούται με ${\mathbf{e}_1}^\intercal \mathbf{x}_k$.

Η μέση τιμή των προβαλλομένων διανυσμάτων είναι $\mathbf{e_1}^\intercal \bar{\mathbf{x}}$ όπου $\bar{\mathbf{x}}$ είναι η μέση τιμή των δεδομένων:

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i} \tag{4.2}$$

Η αντίστοιχη διασπορά ισούται με:

$$V = \frac{1}{N}\sum_{i=1}^{N} \{\mathbf{e}_1{}^\intercal \mathbf{x}_i - \mathbf{e}_1{}^\intercal \bar{\mathbf{x}}\} = \mathbf{e}_1{}^\intercal \mathbf{S} \mathbf{e}_1 \tag{4.3}$$

Όπου $\mathbf{S}$ ο πίνακας συνδιακύμανσης (*Covariance Matrix*):

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\intercal \tag{4.4}$$

Για την εύρεση του άξονα $\mathbf{e}_1$ χρειάζεται η μεγιστοποίηση της εξίσωσης 4.3, υπό τον περιορισμό $||\mathbf{e}_1|| = 1$, έτσι ώστε να βρεθεί η διεύθυνση του διανύσματος $\mathbf{e}_1$. Εισάγοντας έναν πολλαπλιαστή *Lagrange* η μεγιστοποιήση της εξίσωσης 4.3 ανάγεται στην μεγιστοποίηση της:

$$J = \frac{1}{N}\sum_{i=1}^{N} \{\mathbf{e}_1{}^\intercal \mathbf{x_i} - \mathbf{e}_1{}^\intercal \bar{\mathbf{x}}\} = \mathbf{e}_1{}^\intercal \mathbf{S} \mathbf{e}_1 + \lambda_1(1 - \mathbf{e}_1{}^\intercal \mathbf{e}_1) \tag{4.5}$$

Θέτοντας την παράγωγό της εξίσωσης 4.5 ίση με 0 προκύπτει:

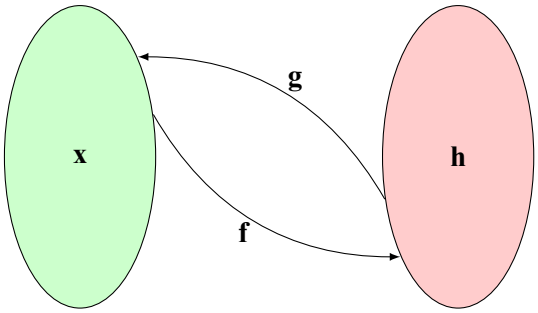
$$\mathbf{e}_1{}^\intercal \mathbf{S} \mathbf{e}_1 = \lambda_1 \tag{4.6}$$

Η παραπάνω συνθήκη δείχνει ότι το διάνυσμα $\mathbf{e_1}$ αντιστοιχεί στο ιδιοδιάνυσμα της *covariance matrix* με την μέγιστη ιδιοτιμή λ_1 [9]. Το διάνυσμα $\mathbf{e}_1$ αντιστοιχεί στην πρώτη κύρια συνιστώσα της *PCA*. Αντίστοιχα μπορούν να προκύψουν και επόμενες συνιστώσες μεγιστοποιώντας την διασπορά των προβαλλόμενων διανυσμάτων ως προς όλες τις κατευθύνσεις οι οποίες είναι ορθογώνιες σε όλες τις προηγούμενες συνιστώσες.

4.5 *Autoencoders*

Ένας *autoencoder* είναι ένα νευρωνικό δίκτυο το οποίο εκπαιδεύεται με σκοπό να αντιγράψει την είσοδο στην έξοδο. Το δίκτυο αποτελείται από 2 τμηματα : τον *encoder* ο οποίος κωδικοποιεί την είσοδο $\mathbf{x}$ σε μια κρυφή αναπαράσταση $\mathbf{h} = f(\mathbf{x})$ και τον *decoder* ο οποίος αποκωδικοποίει την αναπαράσταση $\mathbf{r} = g(\mathbf{h})$. Στο Σχήμα 4.3 δίνεται το *mapping* μεταξύ του χώρου χαρακτηριστικών και χώρου της αναπαράστασης. Ένα δείγμα $\mathbf{x} \in \mathbb{R}^N$ απεικονίζεται μέσω της συνάρτησης $f : \mathbb{R}^N \to \mathbb{R}^D$ σε μια κρυφή αναπαράσταση. Αντίστροφα η κρυφή αναπαράσταση $\mathbf{h} \in \mathbb{R}^D$ απεικονίζεται στον χώρο τον χαρακτηριστικών $g : \mathbb{R}^D \to \mathbb{R}^N$. Για λόγους που φαίνονται παρακάτω συνήθως ισχύει $D < N$.

Η εκπαίδευση ενός *autoencoder* δεν διαφέρει από την εκπαίδευση ενός νευρωνικού δικτύου. Οι ίδιοι αλγόριθμοι εκμάθησης νευρωνικών δικτύων μπορούν να εφαρμστούν και στην περίπτωση των *autoencoders*. Στην περίπτωση των *autoencoders* η ετικέτα καθενός δείγματος $\mathbf{x}_i$ ταυτίζεται με το ίδιο το δείγμα $\mathbf{y}_i = \mathbf{x}_i$. Αν και η εκμάθηση της συνάρτησης $\mathbf{x} = g(f(\mathbf{x}))$ δεν παρουσιάζει ιδιαίτερο ενδιαφέρον, θέτοντας περιορισμούς στο δίκτυο μπορούν να βρεθούν ιδιαίτερες δομές των δεδομένων. Οι περιορισμοί συνήθως αφορούν την αρχιτεκτονική του δικτύου ή τιμές των βαρών οι οποίες εμφανίζονται ως επιπλέον όροι στην *loss function*.



Σχήμα 4.3: Σχηματική αναπαράσταση του *autoencoder mapping*.

Παραδοσιακά οι *autoencoders* χρησιμοποιούνται σε προβλήματα επεξεργασίας των χαρακτηστικών όπως *dimensionality reduction* ή *feature learing* [65, 66, 67]. Το πλεονέκτημα των *autoencoders* σε σχέση με τις κλασικές μεθόδους επιλογής χαρακτηστικών είναι η δυνατότητα χρήσης μη γραμμικού μετασχηματισμού μέσω των ντετερμινιστικών συναρτήσεων $f(.)$ και $g(.)$. Όπως αναφέρθηκε η μέθοδος *PCA* χρησιμοποιεί ένα γραμμικό μετασχηματισμό των χαρακτηριστικών και επιλέγει εκείνα που διαχωρίζουν τα δεδομένα καλύτερα. Ωστόσο αυτό περιορίζει την ισχύ του μετασχηματισμού καθώς η προβολή σε ένα μη γραμμικό χώρο ενδέχεται να διαχωρίζει τα πρότυπα ακόμα περισσότερο.

Πιο σύγχρονοι *autoencoders* έχουν γενικεύσει την ιδέα των ντετερμινιστικών συναρτήσεων χρησιμοποιώντας ένα είδος *stochastic mapping* μεταξύ 2 κατανομών $p_{encoder}(\mathbf{h}|\mathbf{x})$ και της $p_{decoder}(\mathbf{x}|\mathbf{h})$ αντίστοιχα. Τα μοντέλα που χρησιμοποιούν *stochastic mapping* αναφέρονται ως *generative models* επειδή μολις εκπαιδευτούν μπορούν να δημιουργήσουν νέα δεδομένα $\tilde{\mathbf{x}}$ δειγματοληπτώντας τις εκπαιδευμένες κατανομές.

Στη συνέχεια αναλύεται ο κλασικός *autoencoder* (*Vanilla Autoencoder*) και συγκρίνεται με την μέθοδο *PCA*. Τέλος παρουσιάζονται δύο *generative autoencoders*, ο *Variational Autoencoder* και ο *Conditional Variational Autoencoder*. Σε πειραματικό επίπεδο, όλοι οι παραπάνω *autoencoders* χρησιμοποιούνται για *feature selection*, ενώ εξετάζεται παράλληλα η ικανότητα δειγματοληψίας των *generative autoencoders*.

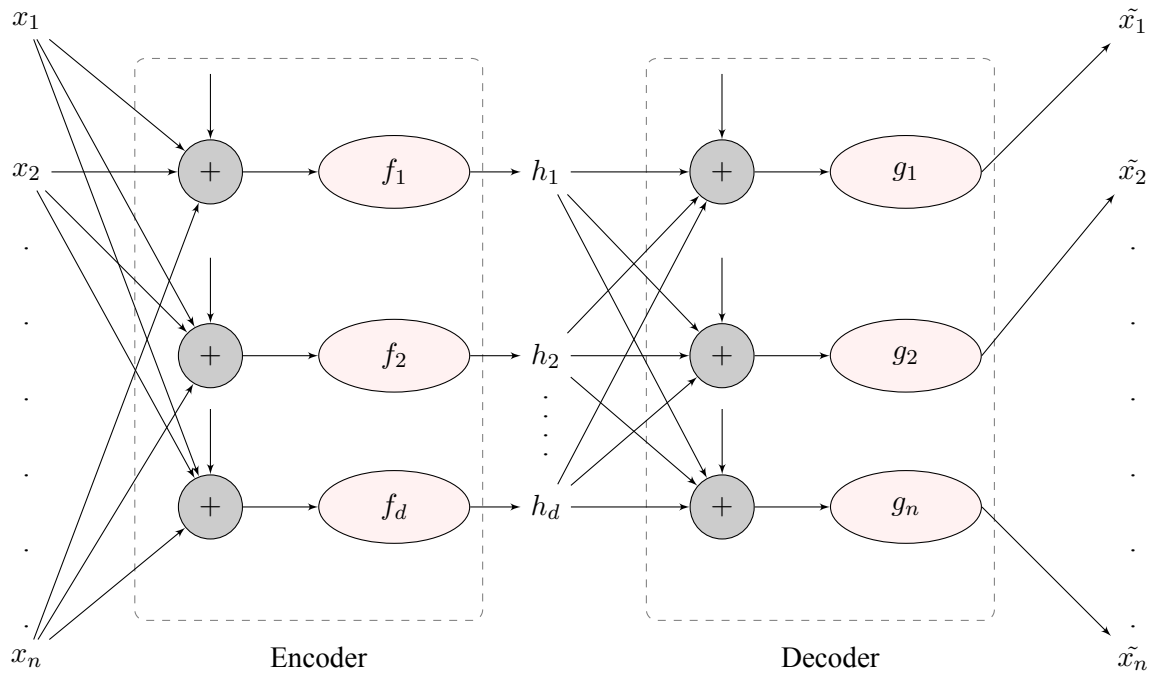
4.5.1 *Vanilla Autoencoder - VAN*

Ο κλασικός (*vanilla*) *autoencoder* δίνεται στο Σχήμα 4.4. Η διάταξη περιέχει ένα κρυφό επιπέδο αποτελούμενο από D νευρώνες. Ο *encoder* κωδικοποιεί το διάνυσμα εισόδου $\mathbf{x}$ στο διάνυσμα $\mathbf{h}$. Κάθε συντεταγμένη h_i αντιστοιχεί στην έξοδο ενός νευρώνα του κρυφού επιπέδου : $h_i = f_i(\mathbf{w}_{ei}^T\mathbf{x} + b_{ei})$ όπου f_i η *activation function*, $\mathbf{w}_{ei}$ και b_{ei} οι παράμετροι του i-οστού νευρώνα του *encoder*. Στη συνέχεια ο *decoder* αποκωδικοποιεί την αναπαράσταση παράγοντας το $\tilde{x}_i = g_i(\mathbf{w}_{di}^T\mathbf{h} + b_{di})$ στην έξοδο i του *VAN*, όπου g_i, η *activation function*, $\mathbf{w}_{di}$ και b_{di} οι παράμετροι του i-οστού νευρώνα του *decoder*.

Η εκπαίδευση ενός *VAN* γίνεται μέσω ελαχιστοποίησης της *loss function*:

$$J(\mathbf{x}, g(f(\mathbf{x}))) \tag{4.7}$$

Η επιλογή της *loss function* διαφέρει από εφαρμογή σε εφαρμογή. Στην ειδική περίπτωση όπου οι συναρτήσεις $f(.)$ και $g(.)$ είναι γραμμικές και επιπλέον η *loss function* είναι η συνάρτηση *mean squared error*, ο *VAN* και η *PCA* παράγουν τον ίδιο χώρο κρυφής αναπαράστασης [68]. Από την άλλη πλεύρα, με χρήση μη γραμμικών *activation functions* ο *VAN* μπορεί να μάθει μια πιο ισχυρή αναπαράσταση από την *PCA*.



Σχήμα 4.4: Σχηματική αναπαράσταση *Vanilla Autoencoder* 1 κρυφού επιπέδου.

Ένας εύκολος τρόπος για να μάθει ο *VAN* χρήσιμα χαρακτηριστικά προκύπτει μέσω του περιορισμού $D < N$, δηλαδή η διαστασιμότητα της κρυφής αναπαράστασης να είναι μικρότερη από την διαστασιμότητα των δεδομένων. Έστω ότι ο *VAN* καλείται να ανακατασκευάσει τις ψηφιακές εικόνες του συνόλου δεδομένων *MNIST*. Κάθε ψηφιακή εικόνα αναπαριστάται με ένα δυαδικό διάνυσμα $\mathbf{x} \in \{0,1\}^{784}$. Αν ισχύει $D < 784$ τότε ο $encoder$ κωδικοποιεί κάθε $\mathbf{x}$ σε ένα $\mathbf{h} \in \mathbb{R}^D$. Στη συνέχεια ο *decoder* προσπαθεί να ανακατασκευάσει την είσοδο. Επειδή ισχύει $D < N$, χάνεται ένα τμήμα της πληροφορίας που περιέχεται στον χώρο χαρακτηριστικών. Ο *decoder* προσπαθεί να ανακτήσει την χαμένη πληροφορία μέσω του h. Συνεπώς το δίκτυο προσπαθεί να εγκλωβίσει όσο το δυνατό περισσότερη πληροφορία στο διάνυσμα $\mathbf{h}$, αμελώντας πιθανώς άχρηστη πληροφορία που βρίσκεται μέσα στον χώρο χαρακτηριστικών. Αν κάθε x_i προέρχεται από μια *independent and identically - iid* κατανομή ανεξάρτητο από τα άλλα, τότε το $\mathbf{h}$ σπάνια περιέχει κάποια χρήσιμη πληροφορία. Ωστόσο αν υπάρχει κάποια δομή μεταξύ των δεδομένων ο *VAN* μπορεί να την ανακαλύψει. Έστω τώρα ότι ισχύει $D \geq N$. Σε αυτήν την περίπτωση ο *VAN* μπορεί να αντιγράψει την είσοδο στην έξοδο επεκτείνοντας το $\mathbf{x}$ έτσι ώστε $\mathbf{x} \in \mathbb{R}^D$ εισάγοντας περαιτέρω μηδενικά. Το διάνυσμα $\mathbf{h}$ που προκύπτει δεν διαφέρει από το αρχικό $\mathbf{x}$, συνεπως δεν περιέχει περαιτέρω πληφορορία για το δεδομένο και άρα ο *VAN* δεν παρέχει κάποια ουσιαστική πληροφορία.

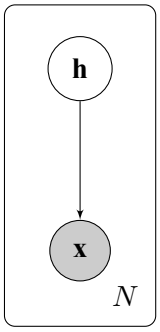
Στην πράξη ο παραπάνω περιορισμός αφορά την αρχιτεκτονική του δικτύου. Ένας άλλος τρόπος εξαγωγής χρήσιμων χαρακτηριστικών προκύπτει με εφαρμογή *sparse* περιορισμών στο δίκτυο (*sparse autoencoders*). Για τον σκοπό αυτό εισάγονται επιπλέον όροι στην *loss function* παίρνοντας την μορφή της εξίσωσης 4.8. Οι περιορισμοί αυτοί αναγκάζουν τους νευρώνες του δικτύου να ενεργοποιούνται πιό σπάνια έτσι ώστε τα $\mathbf{h}_i$ να είναι κατά το δυνατό διαχωρίσιμα. Τυπικοί περιορισμοί $\Omega(\mathbf{h})$ αφορούν την μήτρα των βαρών του δικτύου, όπως η νόρμα L_1 ή L_2. Η παράμετρος λ αντιστοιχεί σε μια *hyperparameter* του δικτύου η οποία προσδιορίζεται κατά την εκπαίδευσή του. Υψηλές τιμές της παραμέτρου δίνουν περαιτέρω ισχύ στον περιορισμό μειώνοντας τις τιμές των βαρών του δικτύου.

$$J(\mathbf{x}, g(f(\mathbf{x}))) + \lambda\Omega(\mathbf{h}) \tag{4.8}$$

4.5.2 *Variational Autoencoder - VAE*

Σε αντίθεση με τον *VAN*, ο *VAE* υποθέτει κάποια άγνωστη κατανομή πάνω στα δεδομένα. Στόχος του είναι ο προσδιορισμός των παραμέτρων της κατανομής. Πιο συγκεκριμένα έστω το σύνολο δεδομένων $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ αποτελούμενο από N *iid* δείγματα. Κάθε δείγμα $\mathbf{x}^{(i)}$ προέρχεται από μια τυχαία διαδικασία μιας μη παρατηρίσημης τυχαίας μεταβλητής (*latent variable*) $\mathbf{h}$ η οποία προέρχεται από κάποια *prior* κατανομή $p_{\theta^*}(\mathbf{h})$:

1. Από την *prior* κατανομή $p_{\theta^*}(\mathbf{h})$ λαμβάνεται ένα δείγμα $\mathbf{h}^{(i)}$.
2. Από την δεσμευμένη κατανομή $p_{\theta^*}(\mathbf{x}|\mathbf{h})$ λαμβάνεται ένα δείγμα $\mathbf{x}^{(i)}$.



Σχήμα 4.5: Γραφικό μοντέλο της γέννησης του συνόλου δεδομένων.

Η διαδικασία γέννησης των δειγμάτων $\mathbf{x}$ δίνεται στο Σχήμα 4.5 όπου κάθε $\mathbf{x_i}$ προέρχεται από την δικία του ξεχωριστή *latent variable* $\mathbf{h}_i$ την οποία δεν μοιράζεται με κανένα άλλο δείγμα $\mathbf{x}_j$, δηλαδή δεν υπάρχουν *global latent variables*. Με βάση την παραπάνω υπόθεση στόχος του *VAE* είναι ο προσδιορισμός της $p_{\theta^*}(\mathbf{x}|\mathbf{h})$. Ωστόσο τόσο οι τυχαίες μεταβλητές $\mathbf{h}_i$ όσο και οι παράμετροι της κατανομής θ^* είναι άγνωστες. Σύμφωνα με το θεώρημα του *Bayes* η ζητούμενη πιθανότητα γράφεται:

$$p_{\theta^*}(\mathbf{h}|\mathbf{x}) = \frac{p_{\theta^*}(\mathbf{x}|\mathbf{h})p(\mathbf{h})}{p_{\theta^*}(\mathbf{x})} \tag{4.9a}$$

$$p_{\theta^*}(\mathbf{x}) = \int p_{\theta^*}(\mathbf{x}|\mathbf{h})p_{\theta^*}(\mathbf{h})dh \tag{4.9b}$$

Για τον υπολογισμό της *posterior* $p_{\theta^*}(\mathbf{x}|\mathbf{h})$ της εξίσωσης 4.9a χρειάζεται ο υπολογισμός της *evidence* $p_{\theta^*}(\mathbf{x})$. Ακόμα και στην περίπτωση όπου οι παράμετροι της κατανομής είναι γνωστές αυτό ειναι πρακτικά αδύνατο αφενός εξαιτίας της εκθετικής πολυπλοκότητας του υπολογισμού ολοκληρώματος της 4.9b (χρειάζεται κάθε δυνατή σύνθεση της $\mathbf{h}$) και αφετέρου ότι ενδέχεται να μην υπάρχει λύση κλειστής μορφής. Προκειμένου να λύσουν το πρόβλημα υπολογισμού της *evidence* τα *variational* μοντέλα χρησιμοποιούν την τεχνική *variational inference* [69]. Σύμφωνα με την αυτή η ζητούμενη *posterior* προσεγγίζεται μέσω μιας οικογένειας κατανομών $q_\lambda(\mathbf{h}|\mathbf{x})$, όπου λ αντιστοιχεί στο σύνολο παραμέτρων της οικογένειας της κατανομής. Ως μετρική εκτίμησης χρησιμοποιείται η απόσταση *Kullback - Leibler divergence* [70] η οποία ποσοτικοποιεί την ομοιότητα μεταξύ 2 κατανομών q και p:

$$KL\{q_\lambda(\mathbf{h}|\mathbf{x})||p(\mathbf{x}|\mathbf{h})\} = \mathbb{E}_{q_\lambda}\{log(q_\lambda(\mathbf{h}|\mathbf{x}))\} - \mathbb{E}_{q_\lambda}\{log(p(\mathbf{h},\mathbf{x}))\} + logp(\mathbf{x}) \tag{4.10}$$

Η κατανομή $q_\lambda(\mathbf{h}|\mathbf{x})$ που ελαχιστοποιεί την KL προσεγγίζει καλύτερα την άγνωστη $p(\mathbf{h}|\mathbf{x})$:

$$q^*_\lambda(\mathbf{h}|\mathbf{x}) = \arg\min_\lambda KL\big\{q_\lambda(\mathbf{h}|\mathbf{x})||p(\mathbf{x}|\mathbf{h})\big\} \tag{4.11}$$

Στην εξίσωση 4.10 εμφανίζεται πάλι ο όρος $p(\mathbf{x})$. Συνεπώς δεν είναι δυνατή η εύρεση της βέλτιστης κατανομής $q^*_\lambda(\mathbf{h}|\mathbf{x})$. Για την εξάλειψη του όρου θεωρείται η συνάρτηση $J(\lambda)$:

$$J(\lambda) = \mathbb{E}_{q_\lambda}\big\{log\big(p(\mathbf{h},\mathbf{x})\big)\big\} - \mathbb{E}_{q_\lambda}\big\{log\big(q_\lambda(\mathbf{h}|\mathbf{x})\big)\big\} \tag{4.12}$$

Λύνοντας την εξίσωση 4.10 χρησιμοποιώντας την σχέση 4.12 προκύπτει:

$$logp(\mathbf{x}) = J(\lambda) + KL\big\{q_\lambda(\mathbf{h}|\mathbf{x})||p(\mathbf{x}|\mathbf{h})\big\} \tag{4.13}$$

Τέλος με χρήση της ανισότητας *Jensen* παρατηρείται ότι $KL(.) \geq 0$. Συνεπώς σύμφωνα με την εξίσωση 4.13 η ελαχιστοποίηση της $KL\big\{q_\lambda(\mathbf{h}|\mathbf{x})||p(\mathbf{x}|\mathbf{h})\big\}$ ισοδυναμέι με την μεγιστοποίηση της $J(\lambda)$. Με άλλα λόγια, δεν χρειάζεται ο υπολογισμός της KL μεταξύ της προσέγγισης q και της ακριβούς τιμής της *posterior* p καθώς η ζητούμενη κατανομή q μπορεί να προκύψει ισοδύναμα μεγιστοποιώντας την $J(\lambda)$ η οποία συχνά εμφανίζεται συχνά ως *Evidence Lower Bound*. Επειδή στα *variational* μοντέλα δεν υπάρχουν *global latent variables*, η συνάρτηση $J(\lambda)$ μπορεί να εκφραστεί ως άθροισμα των επιμέρους συνεισφορών του κάθε δείγματος i. Χρησιμοποιώντας την ιδιότητα γινομένου του λογαρίθμου στην από κοινού πιθανότητα $p(\mathbf{h},\mathbf{x})$ της εξίσωσης 4.12 η $J(\lambda)$ παίρνει την παρακάτω μορφή:

$$J(\lambda) = \sum_{i=1}^{N} J_i(\lambda) = \sum_{i=1}^{N} \mathbb{E}_{q_\lambda(\mathbf{h}|\mathbf{x}_i)}\big\{log\big(p(\mathbf{h}_i|\mathbf{x}_i)\big)\big\} - KL\big\{q_\lambda(\mathbf{h}|\mathbf{x}_i)||p(\mathbf{h})\big\} \tag{4.14}$$

Το τελευταίο βήμα για την κατασκευή του *VAE* είναι η ταύτιση των κατανομών q και p με τα επιμέρους τμήματα ενός *autoencoder*. Η προσέγγιση της *posterior*, $q_\theta(\mathbf{h}|\mathbf{x})$, αντιστοιχεί στον *encoder*, ο οποίος δεδομένου ενός δείγματος εισόδου $\mathbf{x}$ παράγει μια *latent variable* της κατανομής. Από την άλλη πλευρά η *likelihood* $p_\phi(\mathbf{h}|\mathbf{x})$, αντιστοιχεί στον *decoder* ο οποίος δεδομένου της *latent variable* $\mathbf{h}$, παράγει ένα δείγμα $\tilde{\mathbf{x}}$. Οι παράμετροι των κατανομών θ και ϕ αντιστοιχούν στις παραμέτρους των αρχιτεκτονικών των επιμέρους δικτύων, δηλαδή στα βάρη και στις πολώσεις. Στο μοντέλο του *VAE* η προσέγγιση της *posterior* είναι μια *Gaussian* κατανομή, όχι απαραίτητα κανονική, ενώ η *prior* των *latent variables* $p(\mathbf{h})$ ακολουθεί την κανονική κατανομή. Εκφράζοντας το πρόβλημα μεγιστοποίησης ως πρόβλημα ελαχιστοποίησης η *loss function* του *VAE* γράφεται :

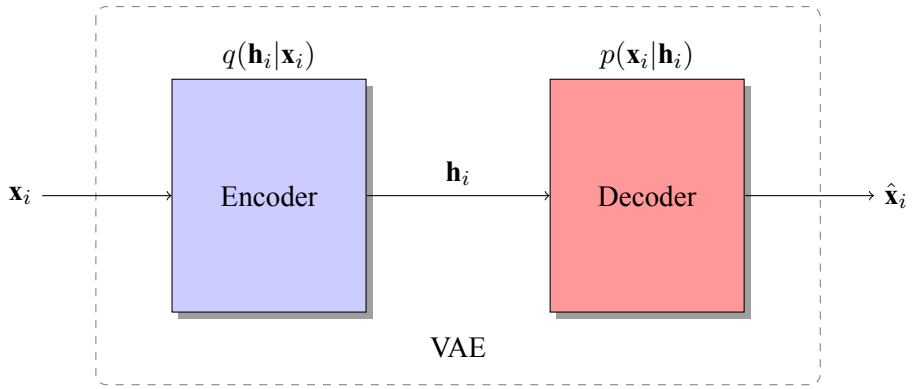
$$J(\mathbf{X},\theta,\phi) = \sum_{i=1}^{N} J_i(\mathbf{x}_i,\theta,\phi) \tag{4.15a}$$

$$J_i(\mathbf{x}_i,\theta,\phi) = -\mathbb{E}_{\mathbf{h}_i\sim q_\theta(\mathbf{h}_i|\mathbf{x}_i)}\big\{log\big(p_\phi(\mathbf{x_i}|\mathbf{h}_i)\big)\big\} + KL\big\{q_\theta(\mathbf{h}_i|\mathbf{x}_i)||p(\mathbf{h}_i)\big\} \tag{4.15b}$$

Καθένας από τους όρους του δεξιού μέλους της εξίσωσης 4.15b περιγράφει την ποινή της *loss function* που δέχεται ο *VAE* εξαιτίας ενός δείγματος. x_i. Αν ο *decoder* δεν ανακατασκευάζει ικανοποιητικά το δείγμα τότε δέχεται την ποινή του όρου $-\mathbb{E}_{\mathbf{h}_i\sim q_\theta(\mathbf{h}_i|\mathbf{x}_i)}\big\{log\big(p_\phi(\mathbf{x}_i|\mathbf{h}_i)\big)\big\}$. Αντίθετα αν ο *encoder* δεν παράγει *latent variables* $\sim \mathbb{N}(0,1)$ τότε δέχεται την ποινή του όρου $KL\big\{q_\theta(\mathbf{h}_i|\mathbf{x}_i)||p(\mathbf{h}_i)\big\}$.

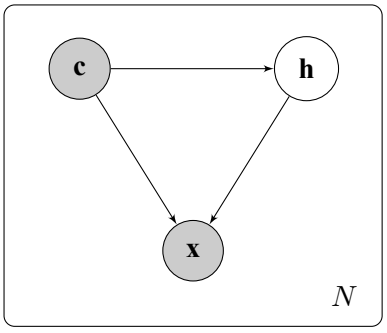
Ο όρος κανονικοποίησης $KL\big\{q_\theta(\mathbf{h}_i|\mathbf{x}_i)||p(\mathbf{h}_i)\big\}$ αποτελεί το κλειδί του *VAE*. Στην πράξη εξασφαλίζει ότι τα δεδομένα που προέρχονται από την ίδια κλάση βρίσκονται σχετικά κοντά και παράλληλα φροντίζει τα σύνολα των δεδομένων των κλάσεων να είναι κατά το δυνατό διαχωρίσιμα. Στο παράδειγμα των χειρόγραφων ψηφίων, δυό εικόνες του ίδιου αριθμού έχουν παρόμοια κωδικοποίηση. Αντίθετα δύο εικόνες των αριθμών 5 και 7 οφείλουν να έχουν αρκετά διαφορετική κωδικοποίηση. Αν

ο όρος αυτος απουσιάζε τότε ενδέχεται ο *encoder* να τοποθετούσε κάθε δείγμα σε μια διαφορετική περιοχή του χώρου [71], με αποτέλεσμα δύο εικόνες του ίδιου ψηφιού να βρεθούν αρκετά μακρία .



Σχήμα 4.6: Σχηματική αναπαράσταση ενός *VAE*.

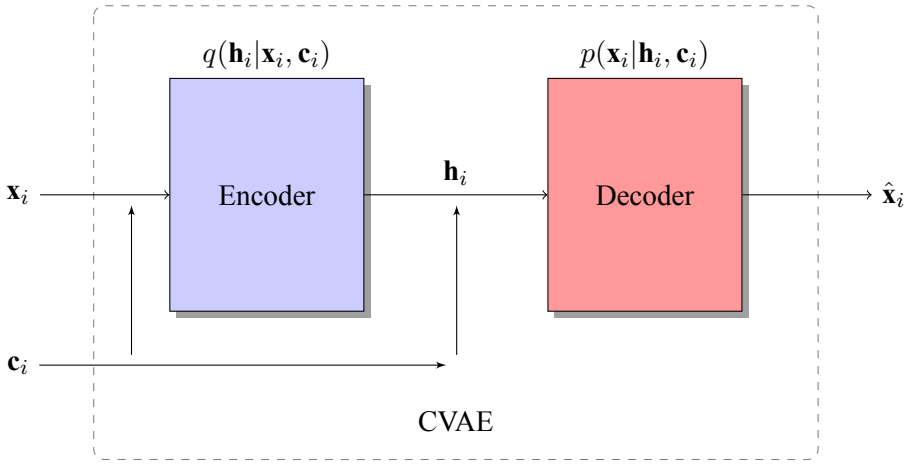
Η αναπαράσταση ενός *VAE* δίνεται στο Σχήμα 4.6. Συχνά ο *encoder* και ο *decoder* αναφέρονται ως *inference* και *generative networks* αντίστοια. Όπως δείχθηκε παραπάνω ο όρος *inference network* χρησιμοποιοιείται για ένα δίκτυο το οποίο καλείται να προβλέψει μια *latent representation* ενός καινούργιου δείγματος, ενώ ο όρος *generative network* χρησιμοποιείται για ένα δίκτυο που αποκωδικοποιεί την *latent representation* γεννώντας ένα νεό δείγμα της κατανομής του συνόλου δεδομένων. Για αυτόν τον λόγο οι *VAE* χαρακτηρίζονται ως *generative models*.



Σχήμα 4.7: Γραφικό μοντέλο της γέννησης του συνόλου δεδομένων.

4.5.3 *Conditional Variational Autoencoder - CVAE*

Εξετάζοντας το μοντέλο του *VAE* παρατηρείται πως δεν υπάρχουν περιορισμοί στα δείγματα που γεννούνται. Ο *encoder* κωδικοποιεί το διάνυσμα εισόδου αδιαφορώντας για την προέλευσή του, δηλαδή για την ετικέτα του. Παρόμοια λογική εφαρμόζεται και στην περίπτωση του *decoder*. Με άλλα λόγια είναι δύσκολο να ερμηνευτούν τα δεδομένα που παράγονται, καθώς δεν είναι εφικτή η ένταξή τους σε μια κλάση. Ένα παράδειγμα που περιγράφει το πρόβλημα ταυτοποίησης των παραγόμενων δεδομένων προκύπτει πάλι με χρήση των ψηφιακών εικόνων *MNIST*. Έστω ότι επιθυμείται η δημιουργία μιας εικόνας που αντιστοιχεί στον αριθμό *5*. Αρχικά λαμβάνεται μια *latent variable* δειγματοληπτώντας την κατανομή του *encoder* και στη συνέχεια αποκωδικοποιείται μέσω του *decoder* γεννώντας ένα νέο δείγμα. Δεδομένου ότι ο *VAE* αδιαφορεί για το ποιόν αριθμό θα παράξει, η εικόνα στην έξοδο ενδέχεται να αναπαριστά τον αριθμό *7*. Αντίστοιχα προβλήματα μπορούν να προκύψουν σε δείγματα φωνής τα οποία διαθέτουν ετικέτες συναισθήματος.



Σχήμα 4.8: Σχηματική αναπαράσταση ενός *CVAE*.

Για την επίλυση του προβλήματος τόσο οι *encoder* όσο και ο *decoder* χρειάζονται περαιτέρω πληροφορία για το δείγμα. Έστω ότι η πληροφορία αυτή βρίσκεται σε ένα διανύσμα $\mathbf{c}_i$ το οποίο αναφέρεται στην προέλευση του δείγματος $\mathbf{x}_i$. Προφανώς ενδέχεται να υπάρχουν δείγματα όμοιας προέλευσης δηλάδη $\exists\, i, j : \{\mathbf{c}_i = \mathbf{c}_j,\ i \neq j\}$. Επείδη το δείγμα $\mathbf{x}$ παράγεται μέσω της *latent variable* $\mathbf{h}$, έπεται ότι η $\mathbf{h}$ δεσμέυεται από το διάνυσμα $\mathbf{c}$ (βλ. Σχήμα 4.7). Για την ένταξη της πληροφορίας στο δίκτυο δεσμεύονται οι κατανομές q και p ως προς το διάνυσμα $\mathbf{c}$. Όσο αναφορά την εκπαίδευση του δικτύο χρησιμοποιείται η *loss function* της εξίσωσης 4.15b στην οποία πλέον οι όροι δεσμεύνται ως προς $\mathbf{c}_i$:

$$J(\mathbf{X}, \theta, \phi) = \sum_{i=1}^{N} J_i(\mathbf{x}_i, \theta, \phi) \tag{4.16a}$$

$$J_i(\mathbf{x}_i, \theta, \phi) = -\mathbb{E}_{\mathbf{h}_i \sim q_\theta(\mathbf{h}_i|\mathbf{x}_i, \mathbf{c}_i)}\big\{log\big(p_\phi(\mathbf{x}_i|\mathbf{h_i}, \mathbf{c}_i)\big)\big\} + KL\big\{q_\theta(\mathbf{h}_i|\mathbf{x}_i, \mathbf{c}_i)||p(\mathbf{h}_i|\mathbf{c}_i)\big\} \tag{4.16b}$$

Όσο αναφορά του διανύσμα $\mathbf{c}$ η πιο προφανής επιλογή είναι η χρήση των ετικετών του καθενός δείγματος, συνήθως σε μορφή *one hot vector*. Το μοντέλο δίνει σημασία στην προέλευεση των δειγμάτων και κατά συνέπεια μπορεί να παράξει στην έξοδο ερμηνέυσιμα συνθετικά δείγματα. Ο παραπάνω *autoencoder* ονομάζεται *Conditional Variational Autoencoder - CVAE* [72] (βλ. Σχήμα 4.8)

4.6 *Transfer Learning*

Ανεξάρτητα από την ποιότητα της *representation* των δεδομένων, οι περισότεροι αλγόριθμοι μηχανικής μάθησης υποθέτουν πως τα δεδομένα εκπαίδευσης και αξιολόγησης προέρχονται από τον ίδιο χώρο χαρακτηριστικών και από την ίδια κατανομή. Υπάρχουν περιπτώσεις στις οποίες η κατανομή μπορεί να αλλάξει, όπως για παραδείγμα μετά την συλλογή νέων δεδομένων και την υποσημείωση της αντίστοιχης ετικέτας τους. Σε τέτοιες περιπτώσεις τα περισσότερα μοντέλα χρειάζονται να επανεκπαιδευτούν, κάτι το οποίο μπορεί να γίνει αρκετα χρονοβόρο. Επίσης σε άλλες περιπτώσεις τα δεδομένα εκπαίδευσης δεν επαρκούν για την εκπαίδευση του μοντέλου. Στόχος του *transfer learning* είναι η μεταφόρα της γνώσης μεταξύ διαφόρων μοντέλων αντιμετωπίζοντας τις περιπτώσεις αλλαγής κατανομής και ανεπάρκειας δειγμάτων.

Σύμφωνα με την ορολογία [73], το *transfer learning* περιλαμβάνει τις έννοεις ενός προβλήματος (*task*) και ενός πεδίου (*domain*). Ένα *domain* $\mathbb{D} = \{\mathbb{X}, p(\mathbf{X})\}$ αποτελείται από τον χώρο χα-

ρακτηριστικών $\mathbb{X}$ καθώς και την από κοινού πιθανότητα του συνόλου δεδομένων $p(\mathbf{X})$, όπου $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{X}$. Δεδομένου ενός *domain*, ένα *task* $\mathbb{T} = \{\mathbb{Y}, p(y_i|\mathbf{x}_i)\}$ αποτελείται από τον χώρο των ετικετών $\mathbb{Y}$ και την δεσμευμένη πιθανότητα $p(y_i|\mathbf{x_i})$ η οποία μαθαίνεται από το μοντέλο. Έτσι για ένα *source domain* $\mathbb{D}_s$ και *source task* $\mathbb{T}_s$, ένα *target domain* $\mathbb{D}_t$ και ένα *target task* $\mathbb{T}_t$, το *transfer learning* στοχεύει στην εκμάθηση της δεσμευμένης κατανομής $p(y_t|\mathbf{x}_t)$ μέσω της πληροφορίας από τα $\mathbb{D}_s$ και $\mathbb{T}_s$ όπου $\mathbb{D}_s \neq \mathbb{D}_t$ ή $\mathbb{T}_t \neq \mathbb{T}_t$. Αν $\mathbb{D}_s = \mathbb{D}_t$ και $\mathbb{T}_t = \mathbb{T}_t$ τότε πρόκειται για την κλασική *supervised learning* εκπαίδευση ενός μοντέλου. Σε κάθε άλλη περίπτωση υπάρχουν διαφορετικές μεθοδοι *transfer learning* οι οποίες συνοψίζονται στον Πίνακα 4.1.

Στο *transfer learning* υπάρχουν 3 θέματα προς έρευνα:

1. *"What to transfer"*: ποίο κομμάτι της γνώσης μπορεί να μεταφερθεί μεταξύ διαφορετικών *domains* ή *tasks*.

2. *"When to transfer"*: κάτω από ποιές συνθήκες η μεταφορά της γνώσης επωφελεί το *target task*. Υπάρχουν περιπτώσεις στις οποίες η μεταφορά γνώσης μπορεί να δυσκολέψει το *target task*. Για παράδειγμα είναι φανερό πως η μεταφορά γνώσης μεταξύ ενός *source task* Επεξεργασίας Εικόνας και ενός *target task* Επεξεργασίας Φυσικής Γλώσσας δεν θα επιφέρει επιθυμητά αποτελέσματα, αφού τα δύο *tasks* διαφέρουν αισθητά. Αντίθετα, η μεταφορά της γνώσης μεταξύ ενός *source task* Αναγνώρισης Ομιλητή και ενός *target task* Αναγνώρισης Συναισθήματος από Φωνή μπορεί να επιφέρει σημαντική βελτίωση.

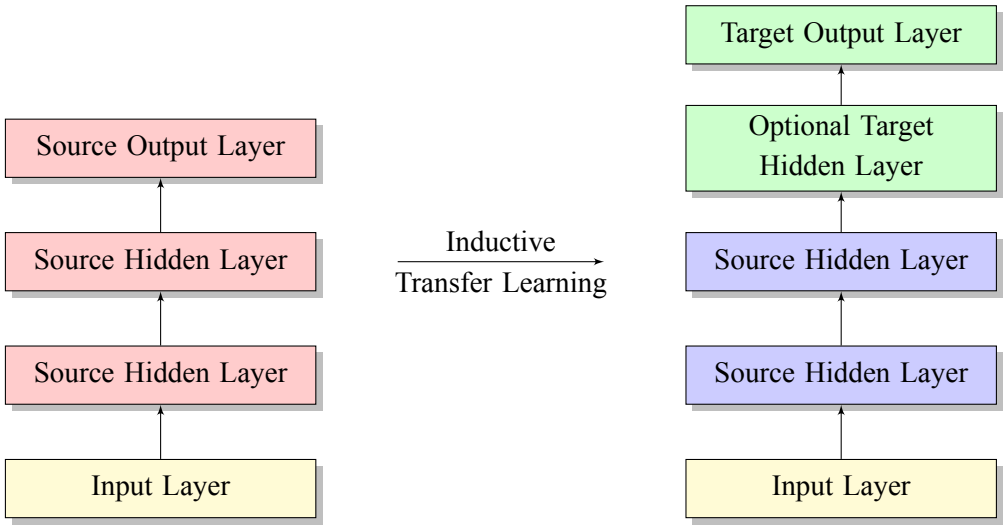
3. *"How to transfer"*: ποίοι αλγόριθμοι εφαρμόζονται για την μεταφόρα της γνώσης.

Learning Settings	Source and Target Domains	Source and Target Taks
Traditional Machine Learning	same	same
Inductive Transfer Learning	same	diff but related
Unsupervised Transfer Learning	diff but related	diff but related
Transductive Transfer Learning	diff but related	same

Πίνακας 4.1: Δυνατές περιπτώσεις *transfer learning*.

Στην παρούσα εργασία ενδιαφέρουν οι μέθοδοι *Inductive Transfer Learning*. Ο πιο απλός τρόπος *Inductive Transfer Learning* γίνεται μέσω βαθιών νευρωνικών δικτύων. Ένα δίκτυο προεκπαιδεύεται (*pretraining*) σε ένα *source task* και *domain*. Τα κρύμμενα επίπεδα του δικτύου συνιστούν μια αλληλουχία *representations* των δεδομένων. Αυτά που βρίσκονται κοντά στην είσοδο του δικτύου αναπαριστούν τα δεδομένα σε μια πιο χαμηλού επιπέδου *representation* ενώ εκείνα που βρίσκονται πιο κοντά στην εξόδο εκφράζουν μια πιο υψηλού επιπέδου *representation*. Το τελευτάιο επίπεδο του δικτύου ασχολείται με την επίλυση του *source task*. Μετά το *pretraining* του δικτύου, αφαιρούνται επίπεδα ξεκινώντας από την έξοδο του δικτύου. Τέλος τοποθετείται ένα νέο επιπέδο εξόδου για την επίλυση του *target task* και πιθανώς κάποια νέα κρυφά επίπεδα. Επειδή τα *domains* των δύο *tasks* είναι ίδια, το νέο δίκτου που πρόκυπτει γνωρίζει πως να εξάγει την χαμηλού επιπέδου πληροφορία των δεδομένων και στη συνέχεια να την χρησιμοποιήσει για να λύσει το *target task*. Συνήθως οι παράμετροι προς εκπαίδευση του νέου δικτύου περιλαμβάνουν μόνο τα επίπεδα που αντικαταστάθηκαν έτσι ώστε η γνώση που έλαβε το δίκτυο κατά το *pretrain* να παραμείνει αυτούσια (*fine - tunning*). Ένα παράδειγμα *Inductive Transfer Learning* δίνεται στο Σχήμα 4.9. Το αρχικό δίκτυο προεκπαιδεύεται για να λύσει το *source task*. Στην συνέχεια αφαιρείται το επίπεδο εξόδου και προστίθεται το αντίστοιχο επίπεδο εξόδου για το *target task*. Κατά το *fine - tunning* εκπαιδεύονται τα νέα επιπέδα του δικτού, ενώ τα ήδη προεκπαιδευμένα επίπεδα παγώνονται. Έτσι το νέο δίκτυο αποκτά την γνώση από

το *source task* και την χρησιμοποιεί για να λύσει το *target task*. Περισσότερα για το *Inductive Transfer Learning* καθώς και την επιλογή *source domain* και *task* δίνονται στο επόμενο κεφάλαιο.



Σχήμα 4.9: Παράδειγμα *Inductive Transfer Learning* ενός νευρωνικού δικτύου.

Κεφάλαιο 5

Αναγνώριση Άγχους

Προηγουμένως δόθηκε έμφαση στο θεωρητικό υπόβαθρο της εργασίας. Στο Κεφάλαιο 1 κατηγοριοποιήθηκαν τα συναισθήματα σε συνεχείς και διακριτούς χώρους. Στο Κεφάλαιο 2 παρουσιάστηκε η παραγωγή φωνής όπως και των χαρακτηριστικών που χρησιμοποιούνται, ενώ στα Κεφάλαια 3 και 4 αναλύθηκαν τα μοντέλα και οι τεχνικές μάθησης. Το κεφάλαιο αυτό αφιερώνεται στην σχετική και προϋπάρχουσα έρευνα, στις πειραματικές διατάξεις και στα αποτελέσματα της εργασίας.

5.1 Προηγούμενη και Σχετική Έρευνα

Μολονότι το *stress* αποτελεί σημαντικό πρόβλημα στη μοντέρνα κοινωνία, δεν υπάρχει σχετική έρευνα με την Αναγνώριση Άγχους από Σήματα Φωνής. Η επί το πλείστον μέχρι τώρα έρευνα εστιάζει στον θεωρητικό φορμαλισμό, στην κατηγοριοποίηση του *stress* μέσω μιας αύξουσας κλίμακας και στα χαρακτηριστικά φωνής που βοηθούν στην αναγνώριση *stress*. Σε άλλες μελέτες το *stress* εξετάζεται μέσω βιοσημάτων ένω σπάνια συμπεριλαμβάνονται στοιχεία φωνής ή βίντεο.

Μια από τις πρώτες προσπάθειες έρευνας σχετικά με το *stress* έγινε από τον *Hansen* [74]. Κατά τον *Hansen* ο ορισμός του *stress* είναι αμφιλεγόμενος, καθώς κάθε άνθρωπος αισθάνεται σε διαφορετικό βαθμό το *stress*. Σε γενικές γραμμές, τα χαρακτηριστικά της φωνής του ανθρώπου κάτω από *stressful* συνθήκες αλλοιώνονται εξαιτίας είτε κάποιου περιβαντολογικού παράγονται είτε της συναισθηματικής κατάστασης η οποία διαταράσσει την παραγωγή της φωνής από το φυσικό, καθομιλούμενο πλαίσιο. Για τον *Hansen* υπάρχουν πολλές τέτοιες συνθήκες στις οποίες οι ψυχικές και σωματικές καταστάσεις αλλάζουν εστιάζοντας σε παραδείγματα προσωπικού εκτάκτης ανάγκης όπως αστυνομίας, πυροσβεστικής και ασθενοφόρων, και ψυχιατρικής όπως της συναισθηματικής κατάστασης ενός ασθενή. Για την μελέτη του *stress* ο *Hansen* δημιούργησε το σύνολο δεδομένων *Speech under Simulated and Actual Stress* [18, 75] και χρησιμοποίησε προσωδικά και φασματικά χαρακτηριστικά για την μελέτη του *stress* σε διαφορετικούς ανθρώπους [74, 76, 77]. Ωστόσο το σύνολο δεδομένων εμπεριέχει δεδομένα τα οποία προέρχονται από σπάνιες μη καθημερινές περιπτώσεις και συνθήκες, όπως ομιλίες από ρυθμιστές εναέριας κυκλοφορίας ή κραυγές από ανθρώπους σε *roller coaster*.

Πιο πρόσφατες έρευνες εστιάζουν σε καθημερινές περιπτώσεις χρησιμοποιώντας δεδομένα που προέρχονται από βιοαισθητήρες, αφού θεωρείται πως παρέχουν καλύτερη πληροφορία σχετικά με τις ψυχοσωματικές αλλαγές του ανθρώπου υπό την επίδραση του *stress* [78]. Οι βιοαισθητήρες συνήθως λαμβάνουν ενδείξεις σχετικά με την υγεία του δέρματος, την πίεση του αίματος, τους καρδιακούς παλμούς και την διάμετρο της κόρης του ματιού. Για παράδειγμα στην μελέτη [79], συλλέχθηκαν δεδομένα σχετικά με το δέρμα και την καρδία από 24 διαφορετικούς οδηγούς σε 3 διαφορετικές συνθήκες οδήγησης, κάθε μία από αυτές τις συνθήκες προκαλούσε διαφορετικού βαθμού *stress* στον οδηγό. Μια άλλη μελέτη εστιάσε στις *stressed* και *unstressed* αντιδράσεις στο πλαίσιο ενός εργαστηρίου, όπου συλλέχθηκαν δεδομένα σχετικά με την πίεση του αίματος, την ηλεκτρική αγωγιμότητα και θερμοκρασία του δέρματος. Στο ίδιο πλαίσιο, η έρευνα [80] επεξεργάζεται δεδομένα αγωγιμότητας του δέρματος μεταξύ 9 διαφορετικών υπαλλήλων σε ένα κέντρο τηλεφωνικής εξυπηρέτησης, όπου

κάθε υπάλληλος φορούσε ένα αισθητήρα για μία εβδομάδα και σημείωνε τα επίπεδα *stress* ύστερα από κάθε κλήση. Μια άλλη μελέτη [81] εστίασε στις *stressed* και *unstressed* αντιδράσεων 22 ατόμων μεταξύ 4 *stressful* και 6 *relaxed* συνεδριών. Παρόμοιες έρευνες που εστιάζουν περισσότερο δεδομένα καρδιογραφημάτων αλλά και δέρματος εντοπίζονται στις περιπτώσεις [82, 83, 84, 85], ενώ λίγες έρευνες ενσωματώνουν στοιχεία φωνής και βίντεο [86, 87, 88].

Η εργασία αυτή διαφέρει από τις προϋπάρχουσες μελέτες σε δύο κυρίως τομείς. Αρχικά, μολονότι έχουν σημειωθεί ικανοποιητικές επιδόσεις με την χρήση αισθητήρων, τα δεδομένα που προέρχονται από αισθητήρες είναι, σε αρκετές περιπτώσεις, δύσκολο να αποκτηθούν και ενδέχεται η χρήση τους να επηρεάζει έντονα την έκβαση των πειραμάτων. Αντίθετα η απόκτηση των δεδομένων φωνής είναι πιο εύκολη. Επίσης όλες οι παραπάνω μελέτες εξετάζουν τα δεδομένα από μια στατική σκοπία, καθώς εξάγουν χαρακτηριστικά πάνω σε ολόκληρο το μήκος των δεδομένων, χωρίς να εκμεταλλεύονται την ακολουθιακή πληφορορία που βρίσκεται μέσα τους. Η ακολουθιακή πληροφορία μπορεί να αποδειχθεί αρκετά χρήσιμη, καθώς σε δείγματα μεγάλου μήκους η συναισθηματική κατάσταση συνήθως εντοπίζεται σε μικρότερα υποτμήματα του δείγματος.

Για αυτούς τους λόγους η εργασία εμπνέεται από τον ευρύτερο κλάδο Αναγνώρισης Συναισθήματος μέσω Φωνής όπου εξάγονται φασματικά και προσωδικά χαρακτηριστικα και στη συνέχεια εφαρμόζονται *functionals* κατά μήκος του δείγματος [89]. Προς εκμετάλλευση της ακολουθιακής πληροφορίας αναπτύχθηκαν απλά *segment - based* και *frame-based LSTMs* [90], υπερτερώντας των κλασικών ταξινομητών όπως *SVMs* [91] και *DNNs* [92]. Ώστοσο με την προσέγγιση αυτή κάθε *time step* της ακολουθίας συνεισφέρει το ίδιο στην απόφαση του ταξινομητή. Δεδομένου ότι κάποιο *time step* μπορεί να περιέχει περισσότερη πληροφορία από ένα άλλο, εφαρμόστηκαν *attention - based* μοντέλα [93, 94], σημειώνοντας καλύτερη επίδοση από τα *baseline no attention - based* μοντέλα. Στον τομέα των *autoencoders* αναπτύχθηκαν *sparse variational* μοντέλα, αφενός για την κωδικοποίηση διανυσμάτων μεγάλης διάστασης σε μικρότερα χωρίς να χάνεται κρίσιμη πληροφορία, και αφετέρου για την γέννηση συνθετικών δειγμάτων στον χώρο χαρακτηριστικών [95, 96, 97]. Σε άλλες περιπτώσεις οι *autoencoders* χρησιμοποιήθηκαν για τεχνικές *transfer learning* ενισχύοντας τις επιδόσεις στα αντίστοιχα *target tasks* [98, 99].

5.2 Περιγραφή Βάσης Δεδομένων

Τα δεδομένα που εξετάζονται στην παρούσα εργασία προέρχονται από ηχογραφήσεις συνεντεύξεων φοιτητών κατά την εξεταστική τους περίοδο [100]. Με σκοπό την αποφυγή σκηνοθετημένων συναισθημάτων οι ίδιοι γνώριζαν πως οι συνεντεύξεις αφορούν έρευνα του πανεπιστημίου. Επιπλέον για την εξάλειψη της νευρικότητας των ερωτηθέντων, υπήρξε μια ολιγόλεπτη συνομιλία μεταξύ κάθε ερωτώμενου και του συνεντευξιάζων. Τέλος η συχνότητα δειγματοληψίας των ηχογραφήσεων είναι $16KHz$ σε μονό κανάλι με 16 - bit κβάντωση.

Το πλήρες σύνολο δεδομένων περιλαμβάνει 61 Μανδαρινούς, 42 Άγγλους και 69 Καντονέζους φοιτητές-ριες, οι οποίοι διαθέτουν σχετική άνεση με την εκάστοτε γλώσσα. Κατά την διάρκεια της συνέντευξης κάθε ερωτώμενος καλούταν να απαντήσει σε 12 ερωτήσεις, εστιασμένες στην καθημερινότητα και στους μελλοντικούς στόχους των φοιτητών. Οι ερωτήσεις αφορούν την προσωπική ζωή, την ακαδημαϊκή πίεση καθώς και τις βλέψεις καριέρας. Προκειμένου να υπάρξει μια φυσική έκφραση του εκάστοτε συναισθήματος, οι ερωτήσεις αυτές διατάσσονται σε σειρά κλιμακώμενου συναισθήματος [101].

Στη συνέχεια κάθε ηχογράφηση επισημειώθηκε από 2 κριτές με δύο δυνατές επιλογές : αγχωμένη ή φυσικη (*stressed or unstressed*). Με σκοπό την ποσοτικοποίηση της συμφωνίας μεταξύ των 2 κριτών μετρήθηκε ο συντελεστής *Kappa inter - labeler agreement* [102]. Πιο συγκεκριμένα για το αγγλικό υποσύνολο δεδομένων οι επισημειώσεις των δύο κριτών δίνεται στον παρακάτω πίνακα:

	Stressed	Unstressed
Κριτής Α	254	250
Κριτής Β	232	272

Πίνακας 5.1: Επισημειώσεις αγγλικών εκφωνήσεων.

Κριτής Α / Κριτής Β	Stressed	Unstressed
Stressed	213	19
Unstressed	41	231

Πίνακας 5.2: Επισημειώσεις συμφωνίας μεταξύ των κριτών των αγγλικών εκφωνήσεων.

Ο συντελεστής κ υπολογίζεται ως :

$$\kappa = \frac{Pr(\alpha) - Pr(e)}{1 - Pr(e)} \tag{5.1}$$

Όπου $Pr(\alpha)$ η παρατηρούμενη συμφωνία και $Pr(e)$ η πιθανότητα τυχαίας συμφωνίας. Οι επισημειώσεις συμφωνίας μεταξύ των κριτών δίνεται στον Πίνακα 5.1. Συνεπώς η παρατηρούμενη συμφωνία ισούται :

$$Pr(\alpha) = \frac{213 + 231}{504} = 0.8810 \tag{5.2}$$

Δεδομένου ότι ο κριτής Α δήλωσε *stressed* 50.40% και ο κριτής Β 46.03% αντίστοιχα προκύπτει ότι :

$$Pr(e) = 0.5040 * 0.4603 + 0.496 * 0.5397 = 0.2320 + 0.2676 = 0.4996 \tag{5.3}$$

$$\kappa = \frac{0.8810 - 0.4996}{1 - 0.4996} = 0.7621 \tag{5.4}$$

Στην παρούσα εργασία χρησιμοποιείται ένα υποσύνολο των αγγλικών εκφωνήσεων. Συγκεκριμένα διατίθενται 300 εκφωνήσεις. Αυτές χωρίζονται σε 156 γυναικείες και σε 144 αντρικές. Επιπλέον 41 και 51 από τις γυναικείες και αντρικές εκφωνήσεις αντίστοιχα έχουν *stressed* ετικέτα. Τέλος ο συντελεστής κ θα χρησιομοποιηθεί ως μετρική σύγκρισης μεταξύ του αναπτυσσόμενου μοντέλου και της ικανότητας του ανθρώπου να διαχωρίζει τα δείγματα. Τα παραπάνω στοιχεία των δεδομένων συνοψίζονται στον ακόλουθο πίνακα :

	Stressed	Unstressed	Total
Male	51	93	144
Female	46	110	156
Total	97	203	300

Πίνακας 5.3: Ετικέτες εκφωνήσεων.

Εξαιτίας του μεγάλου αριθμού παραμέτρων των νευρωνικών δικτύων, το παραπάνω σύνολο δεδομένων δεν επαρκεί για την εκπαίδευση των μοντέλων που εξετάζονται (Σχήμα 5.2, 5.3 και 5.4). Για τον λόγο αυτό χρησιμοποιούνται τεχνικές *inductive transfer learning*, όπου τα μοντέλα των σχημάτων προεκπαιδεύονται σε ένα *source domain* και *task*, δηλαδή σε μια μεγαλύτερη βάση δεδομένων αποτελούμενο δείγματα φωνής. Ως βάση δεδομένων προεκπαίδευσης χρησιμοποιείται η *Interactive Emotional Dyadic Motion Capture database - IEMOCAP* [103]. Η βάση αποτελείται από περίπου

12 ώρες οπτικο - ακουστικών δεδομένων οργανωμένη σε 5 συνεδρίες. Κάθε συνεδρία περιλαμβάνει διαλόγους μεταξύ ενός άντρα και μιας γυναίκας στους οποίους επισημαίνεται η ετικέτα του συναισθήματος σε συνεχή και διακριτό χώρο. Έτσι για την προεκπαίδευση διατίθενται 7529 εκφωνήσεις όπου οι συνεχείς τιμές των ετικετών των εκφωνήσεων στους άξονες *Valence*, *Activation* και *Dominance* κυμαίνονται στο εύρη τιμών $[1.0, 5.5]$, $[1.0, 5.0]$ και $[0.5, 5.0]$ αντίστοιχα, ενώ υπάρχουν 9 διακριτά πιθανά συναισθήματα :

Emotion	Number of utterances
angry	1103
disgusted	2
excited	1041
fearful	40
frustrated	1849
happy	595
neutral	1708
sad	107

Πίνακας 5.4: Ετικέτες διακριτού συναισθήματος των εκφωνήσεων της *IEMOCAP*.

5.3 Περιγραφή Μοντέλων

5.3.1 Αξιολόγηση Μοντέλων

Στόχος της εργασίας είναι η εξαγωγή καλών *representations* για την Αναγνώριση Άγχους μέσω Σημάτων Φωνής. Για τον σκοπό αυτό χρησιμοποιούνται *utterance - based*, *frame - based*, και *segment - based* μοντέλα. Σε κάθε περίπτωση το σύνολο δεδομένων $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ κατακερματίζεται μέσω ενός *10-fold cross validation - 10CV* σε 10 ξένα μεταξύ τους αλλά ίσα ως προς το μέγεθος υποσύνολα έτσι ώστε $\mathbf{X} = \left\{ \bigcup_{i=1}^{N} \mathbf{X}_i : \mathbf{X}_j \neq \mathbf{X}_i \, \forall i \neq j \right\}$. Στη συνέχεια κάθε μοντέλο αξιολογείται για 10 επαναλήψεις όπου στην *i*-οστή επανάληψη εκπαιδεύεται με όλα τα υποσύνολα $\mathbf{X}_j : i \neq j$ και αξιολογείται μέσω κάποιας μετρικής στο υποσύνολο $\mathbf{X}_i$ σημειώνοντας επίδοση $score_i$. Η τελική επίδοση του μοντέλου ισοδυναμεί με την αριθμητική μέση τιμή των επιμέρους επιδόσεων. Ως μετρική επίδοσης χρησιμοποιείται το καθαρό ποσοστό επιτυχίας (*unweighted accuracy - UA*), δηλαδή ο αριθμός των σωστά ταξινομημένων δεδομένων προς το συνολικό αριθμό δεδομένων.

5.3.2 Ακουστικά Χαρακτηριστικά

Utterance - Level

Ως πρώτη προσέγγιση εξάγονται *LLDs* και εφαρμόζονται *functionals* σε όλο το μήκος του *utterance*. Όπως αναφέρθηκε και στο Κεφάλαιο 2, εξάγονται 1582 *utterance - level* χαρακτηριστικά ως εξής : Αρχικά κάθε *utterance* $u_i = \{f_{i1}, f_{i2}, \dots f_{im}\}$ χωρίζεται σε επικαλυπτόμενα πλαίσια όπου εξάγονται 38 *LLDs* μάζι με τις *first order delta coefficients* (βλ. Πίνακα 5.5). Επειδή συγκεκριμένοι *LLDs* χρειάζονται διαφορετικό μήκος *frame*, η διάρκεια των *frames* είναι 25 *msec* με 10 *msec* επικάλυψη για όλους τους *LLDs* εκτός από το *pitch* όπου χρησιμοποιείται 40 *msec* με 10 *msec* επικάλυψη. Στους 38 *LLDs* μαζί με τις *delta coefficients* εφαρμόζονται 21 *functionals* παράγοντας $2 * 38 * 21 = 1596$ χαρακτηριστικά 16 από τα οποία απορρίπτονται καθώς παρουσιάζουν μηδενική τιμή. Επιπλέον εφαρμόζονται 2 *functionals* στην F_0 παράγοντας τελικά 1582 χαρακτηριστικά για κάθε *utterance*.

LLDs	Delta coefficients	Functionals
PCM loudness	✓	A,B
MFCC [0-14]	✓	A,B
Log mel freq. band [0-7]	✓	A,B
LSP freq[0-7]	✓	A,B
F_0 by SHS	✓	A, C
F_0 Envelope	✓	A,B
Voicing probability	✓	A,B
Jitter local	✓	A
Jitter DDP	✓	A
Shimmer local	✓	A

Πίνακας 5.5: Οι επιλεγμένοι *LLDs* μαζί με τις *delta coefficients*και το σύνολο των *functionals* που εφαρμόζονται σε κάθε *LLD*. Συντομογραφίες: *DDP*: *difference of difference of periods*, *LSP*: *linear spectral pairs*, *SHS: sub-harmonic sum*, *Q/A*: *quadratic, absolute*.

Functionals	Set
position max / min arith. mean std. deviation skewness kurtosis lin. regression coeff 1/2 lin. regression error Q/A quartile 1/2/3 quartile range 2-1/3-2/3-1 percentile 99 up level time 75/90	A
percentile 1, percentile range 99-1	B
turn onSets number, turn duration	C

Πίνακας 5.6: Σύνολα functionals που εφαρμόζονται στους *LLDs* και τις *delta coefficients*.

Frame - Level

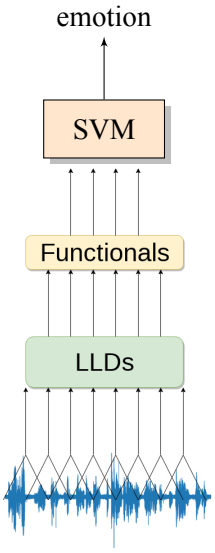
Για την *frame - level* χαρακτηριστικά κάθε *utterance* $u_i = \{f_{i1}, f_{i2}, \dots f_{im}\}$ χωρίζεται σε επικαλυπτόμενα *frames*, διάρκειας 25 *ms* και επικάλυψης 10 *ms*. Ακολουθώντας τις μελέτες [96, 104, 105, 106], σε κάθε *frame* εφαρμόζεται *DFT* 512 σημείων και στη συνέχεια εξάγονται από κάθε *frame* οι πρώτες 80 συνιστώσες του λογαρίθμου της ενέργειας.

Segment - Level

Τα *segment - level* χαρακτηριστικά ταυτίζονται με τα *utterance - level* χαρακτηριστικά με τη διαφορά ότι πλέον το σύνολο των 1582 χαρακτηριστικών εξάγεται σε επίπεδο *segment* και όχι σε όλο το μήκος του utterance [107].

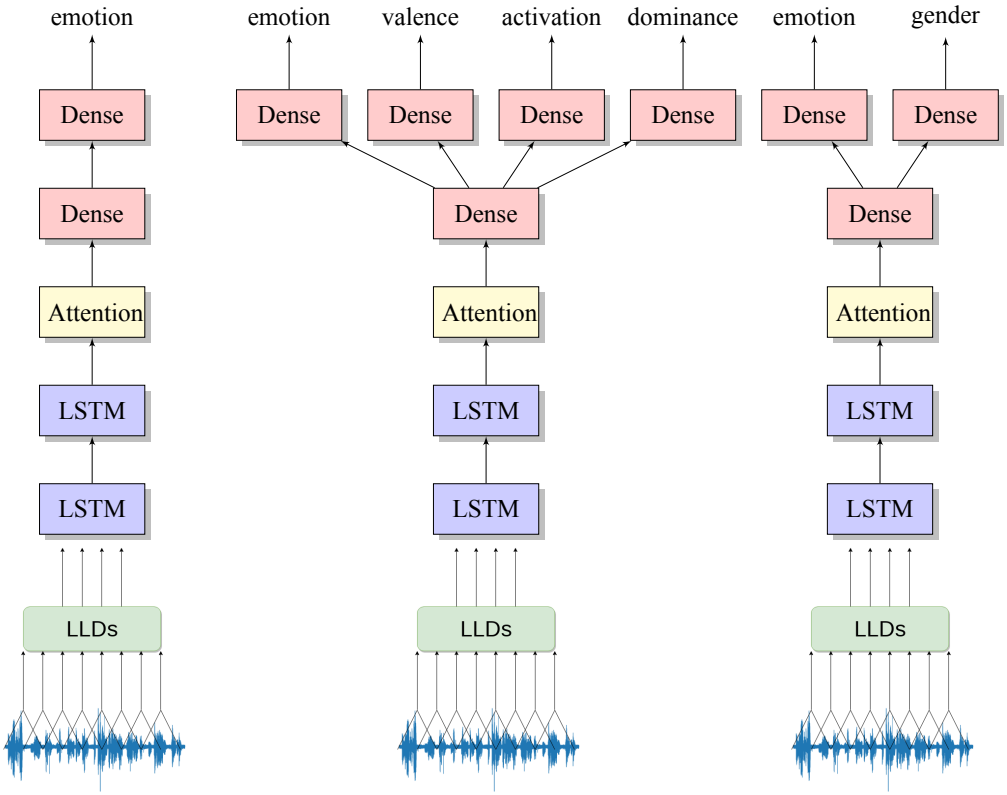
5.3.3 *Utterance - Based* Μοντέλο

Ως *utterance-based* μοντέλο χρησιμοποιείται ένα *SVM* όπου κάθε *utterance* αναπαριστάται από ένα σταθερού μήκους διάνυσμα χαρακτηριστικών. Για την εξαγωγή των χαρακτηριστικών αρχικά κάθε *utterance* u_i χωρίζεται σε επιμέρους επικαλυπτόμενα *frames* : $u_i = \{f_{i1}, f_{i2}, \dots f_{im}\}$. Σε κάθε *frame*



Σχήμα 5.1: *Utterance - based model*.

f_{ij} εξάγονται κάποιοι *LLDs* και στη συνέχεια εφαρμόζονται *functionals* κατά μήκος όλων των *frames* της ακολουθίας δημιουργώντας το σταθερού μήκους διάνυσμα χαρακτηριστικών του *utterance*.

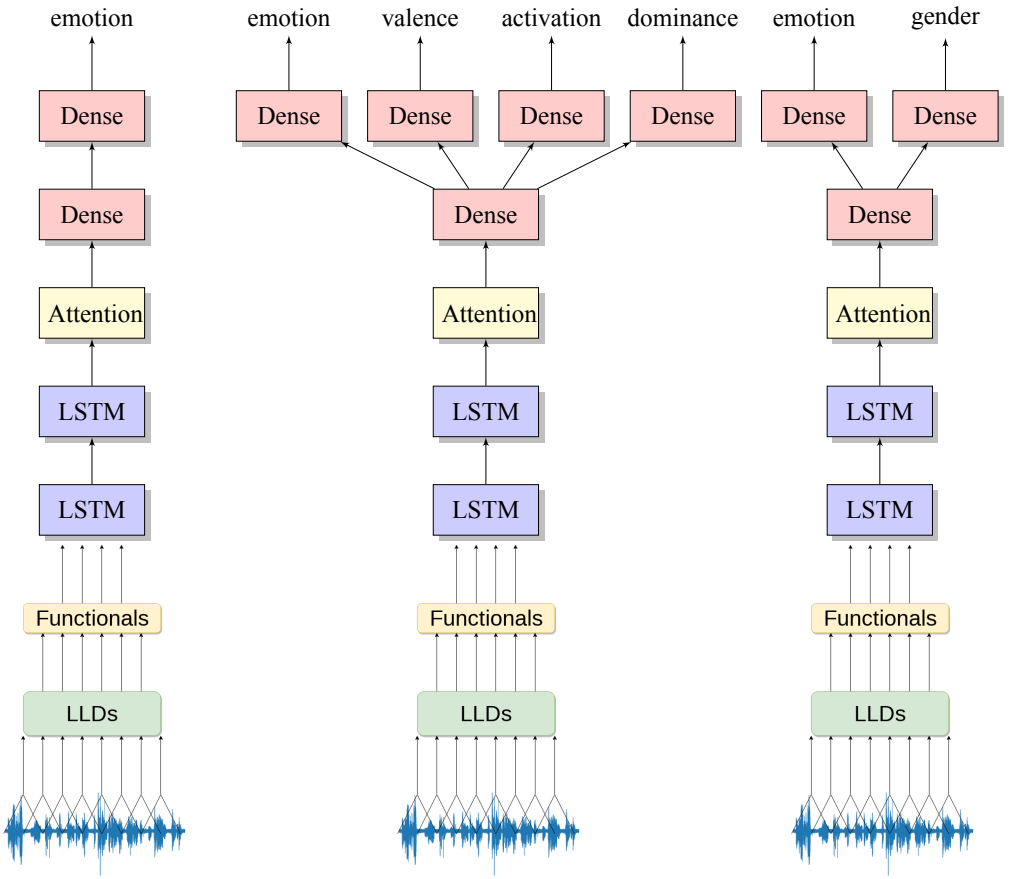


Σχήμα 5.2: *Pretrain frame - based models*.

5.3.4 *Frame - Based* Μοντέλα

Σε αντίθεση με το *utterance - based* μοντέλο, στα *frame - based* και *segment - based* μοντέλα κάθε *utterance* αναπαριστάται μέσω μιας ακολουθίας από *time steps* όπου τα χαρακτηριστικά της ακολου-

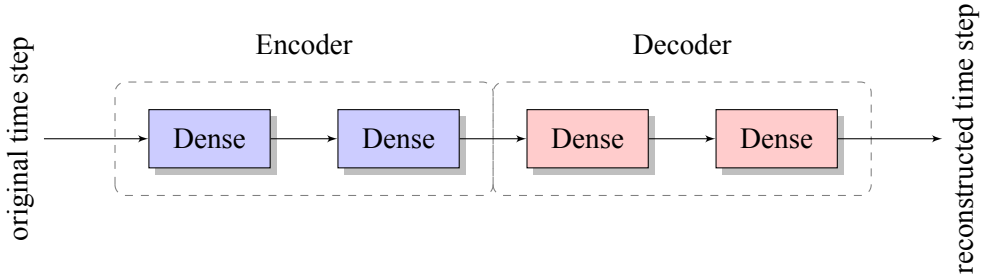
θίας αντιστοιχούν στην συνένωση των επιμέρους χαρακτηριστικών των *time steps*. Έτσι για τα *frame - based* μοντέλα κάθε *time step* αντιστοιχεί σε ένα ή μια μικρή ομάδα $(5 - 10)$ από *frames* στα οποία εξάγονται οι *LLDs* και η ακολουθία αναπαριστάται από την επιμέρους ένωση των διανυσμάτων των *frames*, το ένα δίπλα στο άλλο. Έδω χρησιμοποιούνται 3 διαφορετικά *frame - based* μοντέλα, που δίνονται στο Σχήμα 5.2 που διαφέρουν μόνο ως προς την έξοδό τους. Το αριστερά *frame - based* μοντέλο αντιστοιχεί σε ένα κλασικό *single - task* μοντέλο για την αναγνώριση συναισθήματος στο διακριτό χώρο συναισθημάτων. Το μεσαίο είναι ένα *multi - tasking* μοντέλο για την αναγνώριση συναισθήματος αναγώριση συναισθήματος στον διακριτό και στο συνεχή χώρο (*Valence - Activation - Dominance - VAD*). Τέλος το δεξία είναι επίσης ένα *multi - tasking* μοντέλο για την αναγνώριση συναισθήματος και φύλου του ομιλητή. Τα μοντέλα διαθέτουν 2 ενδιάμεσα επιπέδα *LSTM* νευρώνων με 1 έξοδο ανά *time step*. Ακολουθεί ο μηχανισμός *attention* μαζί με ένα τελευταίο επίπεδο πλήρως συνδεδεμένων νευρώνων (*Dense*). Οι επιμέρους λεπτομέρειες της αρχιτεκτονικής των μοντέλων δίνεται σε κάθε πείραμα ξεχωριστά.



Σχήμα 5.3: *Pretrain segment - based models*.

5.3.5 *Segment - Based* Μοντέλα

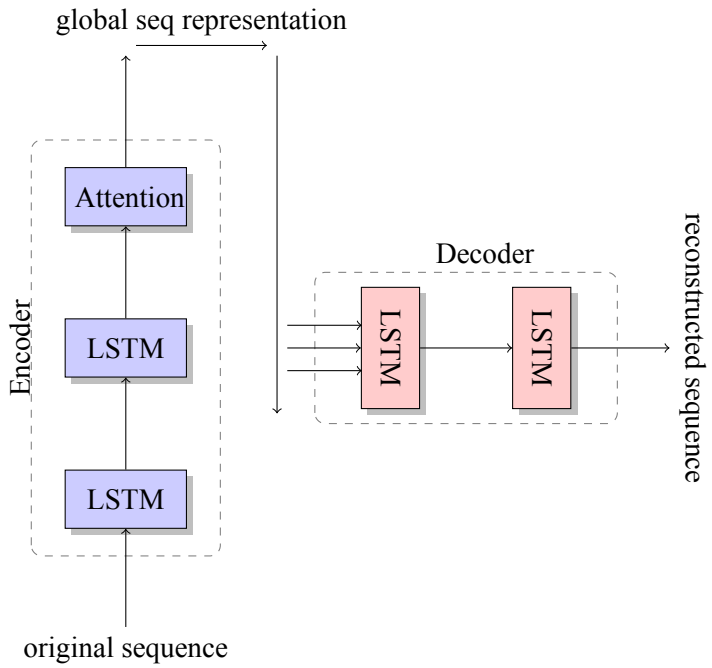
Στα *segment - based* μοντέλα ένα *time step* αντιστοιχεί σε ένα *segment*. Έτσι αν η ακολουθία $u_i = \{s_{i1}, s_{i2}, \ldots, s_{ij}\}$ αποτελείται από j επικαλυπτόμενα *segments*, τότε κάθε *segment* $s_{ij} = \{f_{ij1}, f_{ij2}, \ldots, f_{ijk}\}$ χωρίζεται σε επιμέρους *frames*. Έπειτα εξάγονται οι *LLDs* σε κάθε *frame* και εφαρμόζονται τα *functionals* σε όλο το μήκος του *segment*, δίνοντας ένα σταθερού μήκους διάνυσμα χαρακτηριστικών για κάθε *segment*. Τα διανύσματα τοποθετούνται το ένα διπλά στο άλλο αναπαριστώντας την ακολουθία σε ολόκληρο το μήκος της. Όπως και στην περίπτωση των *frame - based* (βλ. Σχήμα 5.3) εδώ χρησιμοποιούνται 3 διαφορετικά μοντέλα για τα ίδια *tasks* που αναφέρθηκαν, διατηρώντας την ίδια ακριβως αρχιτεκτονική. Ωστόσο, επειδή τα *frame - based* διαχειρίζονται διαφορετικά χαρακτηριστικά από τα *segment - based* μοντέλα, οι λεπτομέρειες των αρχιτεκτονικών διαφέρουν και αναφέρονται ξεχωριστά σε κάθε πείραμα.



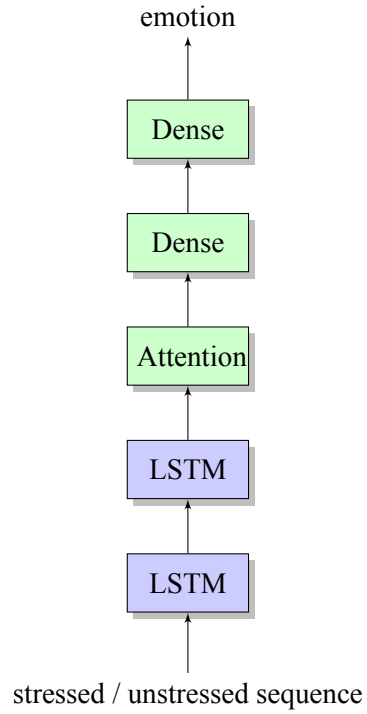
Σχήμα 5.4: *Autoencoders* επιπέδου *time step*.

5.3.6 *Autoencoders*

Εκτός από τα μοντέλα για ταξινόμηση χρησιμοποιούνται και *autoencoders* για *representation learning* και *feature extraction*. Εξετάζονται οι 3 διαφορετικοί *autoencoders* που περιγραφήκαν στο προηγούμενο κεφάλαιο, δηλαδή ο *VAN*, *VAE*, και *CVAE* σε *time step* και *sequence* επίπεδο. Σε επίπεδο *time step* κάθε *autoencoder* λαμβάνει 1 *time step* της ακολουθίας και προσπαθεί να το ανακατασκευάσει. Συνεπώς για *frame - based autoencoders* και *segment - based autoencoders* η είσοδος και η έξοδος αντιστοιχεί σε *frames*, και *segments* αντίστοιχα. Για την κατασκεύη ιεραρχικής *representation* χρησιμοποιούνται βαθιοί *autoencoders* 2 επιπέδων στον *encoder* και *decoder* όπως φαίνεται στο Σχήμα 5.4. Σε επίπεδο *sequence* οι *autoencoders* λαμβάνουν ως είσοδο ολόκληρη την ακολουθία με σκοπό την ανακατασκευή της. Έτσι οι *autoencoders* του Σχήματος 5.4 αντικαθίστανται από *sequence to sequence - seq2seq* μοντέλα [108, 109], δηλάδη αντί για πλήρως συνδεδεμένα επίπεδα νευρώνων χρησιμοποιούνται επίπεδα *LSTM* (βλ. Σχήμα 5.5). Όπως και στην περίπτωση των *time step autoencoders*, οι *seq2seq autoencoders* διαθέτουν 2 επιπέδα στον *encoder* και στον *decoder*. Το πρώτο επίπεδο του *encoder* διαθέτει 1 έξοδο ανά *time step* της ακολουθίας, ενώ το δεύτερο επίπεδο έχωντας παρατηρήσει ολόκληρη την ακολουθία, επιστρέφει ένα διάνυσμα σταθερού μήκους το οποίο περιγράφει την ακολουθία σε όλο το μήκος της *global sequence representation* [110, 111]. Στην προσέγγιση της εργασίας, το διάνυσμα υπεύθυνο για την καθολική αναπαράσταση της ακολουθίας δεν προέρχεται από την έξοδο του τελευταίου επιπέδου του *encoder*. Αντίθετα το τελευταίο επίπεδο του *encoder* διαθέτει 1 έξοδο ανά *time step* της ακολουθίας και στη συνέχεια προστίθεται το επίπεδο του μηχανισμού *Attention*. Το διάνυσμα που παράγεται από τον μηχανισμό προωθείται στο πρώτο επίπεδο του *decoder* αρχικοποιώντας τις καταστάσεις του. Από την πλευρά του *decoder*, τα επίπεδα *LSTM* του *decoder* διαθέτουν 1 έξοδο ανά *time step* της ακολουθίας όπου το τελευταίο επίπεδο, αντιστοιχεί και στην έξοδο του συστήματος, υπεύθυνη για την ανακατασκευή της ακολουθίας.



Σχήμα 5.5: *Autoencoders* επιπέδου *sequence to sequence*.



Σχήμα 5.6: *Fine - tuned Models*

5.3.7 *Transfer Learning*

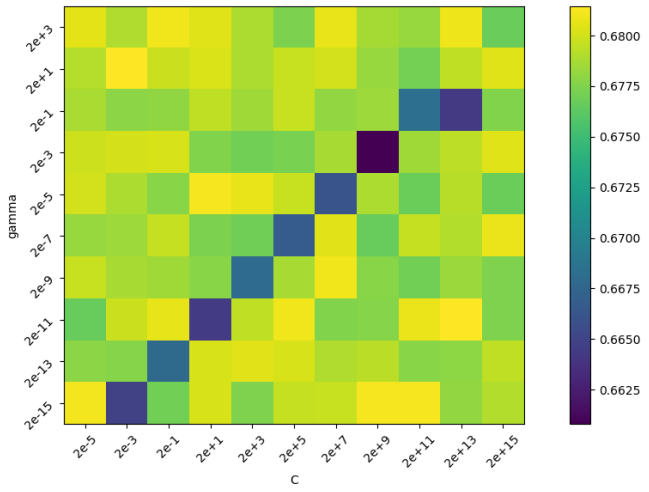
Όπως αναφέρθηκε παραπάνω τα μοντέλα ταξινόμησης και οι *autoencoders* των Σχημάτων 5.2 - 5.5 προεκπαιδεύονται μέσω της *IEMOCAP*. Αρχικά, για την μεταφορά της γνώσης των προεκπαιδευμένων μοντέλων ταξινόμησης από το *source task* και *domain* της *IEMOCAP* στο *target task* και *domain* του συνόλου *stressed* / *unstressed* δεδομένων αφαιρείται το επίπεδο εξόδου, αφού τα 2 *tasks* διαφέρουν. Επίσης αφαιρούνται τα 2 υψηλότερα κρυφά επίπεδα, δηλαδή ο μηχανισμός *Attention* και

το τελευταίο κρυφό επίπεδο πλήρως συνδεδεμένων νευρώνων. Στη συνέχεια τοποθετείται ένας νέος μηχανισμός *Attention* μαζί με ένα νέο επίπεδο πλήρως συνδεδεμένων νευρώνων. Οι παράμετροι των προηγούμενων *LSTM* επιπέδων διατηρούνται στις τιμές ύστερα από το στάδιο προεκπαίδευσης και δεν εκπαιδεύονται περαιτέρω κατά την επίλυση του *target task*. Έτσι όλη η γνώση που απέκτησαν τα χαμηλότερα κρυφά επίπεδα διατηρείται αναλλοίωτη, ενώ τα υψηλότερα κρυφά επίπεδα εστιάζουν στην χρήση της γνώσης των χαμηλότερων κρυφών επιπέδων για την επίλυση του *target task*. Στο Σχήμα 5.6 δίνονται τα μοντέλα που χρησιμοποιούνται για το *target task*, όπου μόνο τα επίπεδα με πράσινο χρώμα εκπαιδεύονται περαιτέρω. Από την άλλη πλευρά, οι *autoencoders* προεκπαιδεύονται και χρησιμοποιούνται στο ίδιο *domain* και *task* και κατά συνέπεια διατηρούν την αρχιτεκτονική τους.

5.4 Πειραματικές Διατάξεις και Αποτελέσματα απλής Ταξινόμησης

5.4.1 *Utterance - based* Ταξινόμηση - *baseline* μοντέλο

Η πρώτη μέθοδος ταξινόμησης αντιστοιχεί στην *baseline* περίπτωση, όπου χρησιμοποιείται ένα απλό *SVM* μαζί με τα *utterance - level* χαρακτηριστικά για κάθε *utterance*. Ως συνάρτηση πυρήνα του *SVM* ορίζεται η συνάρτηση *RBF* της εξίσωσης 5.5. Το μοντέλο εκπαιδεύεται και αξιολογείται για διαφορετικό συνδυασμό των παραμέτρων C της εξίσωσης 3.14a και γ της συνάρτησης πυρήνα, σημειώνοντας τις εξής επιδόσεις για κάθε συνδυασμό:



Σχήμα 5.7: Μετρική *UA*(%) αναγώρισης *stressed* / *unstressed utterances* για το *utterance - based SVM* μοντέλο.

$$K(\mathbf{x}, \tilde{\mathbf{x}}) = exp(-\gamma||\mathbf{x} - \tilde{\mathbf{x}}||^2) \tag{5.5}$$

Το απλό αυτό μοντέλο σημειώνει επίδοση ($\sim$ 68%). Δεδομένου ότι οι *priors* των δύο κλάσεων είναι $p_{stressed} = \frac{93}{300} = 0.31$ και $p_{unstressed} = \frac{207}{300} = 0.69$ το *baseline* μοντέλο δεν παρουσιάζει κάποια βελτίωση καθώς η τυχαία επιλοδή ισοδυναμεί με την *prior* της μεγαλύτερης κλάσης. Μια πιθανή εξήγηση δίνεται από την δυσκολία της αντικειμενικής επισήμανση των δεδομένων, καθώς δεν υπάρχει κάποιο καθολικό μοτίβο με το οποίο ο ανθρωπος κατηγοριοποιεί τα συναισθήματα. Ωστόσο εδώ εξετάζονται αποκλειστικά οι δυνατότητες των μοντέλων, και όχι ο τρόπος επισήμανσης των εκφωνήσεων. Φαίνεται πως τα *utterance - based* μοντέλα αποτυγχάνουν όταν τα δείγματα φωνής έχουν

μεγάλο μήκος και οι αποστάσεις μεταξύ των κλάσεων στις οποίες ταξινομούνται είναι αρκετά μικρές. Το γεγονός αυτό ενισχύει την υπόθεση ότι η συναισθηματική πληροφορία βρίσκεται σε μικρά υποτμήματα των εκφωνήσεων. Έτσι αν η συναισθηματική πληροφορία βρίσκεται σε ένα μικρό υποτμήμα διάρκειας λίγων δευτερολέπτων μιας εκφώνησης 1 λεπτού, τότε οι εξαγωγή καθολικών αναπαραστάσεων μέσω *functionals* σε όλο το μήκος της εκφώνησης αποκρύπτει αυτήν την πληροφορία. Για το καλύτερο διαχωρισμό των κλάσεων χρειάζεται περαιτέρω προσοχή στον τρόπο εξαγωγής χαρακτηριστικών έτσι ώστε η πληροφορία σε μικρότερα υποτμήματα να γίνει εμφανής στα μοντέλα ταξινόμησης. Κρίνεται αναγκαία λοιπόν η μελέτη των εκφωνήσεων μέσω μικρότερων υποτμημάτων, σε επίπεδο *frames* ή *segments*. Τόσο σε *frame* όσο και σε *segment - based* ταξινόμηση, μια ακολουθία $u_i = \{t_{i_1}, t_{i_2}, \dots t_{i_m}\}$ χωρίζεται σε επιμέρους *time steps* όπου σε κάθε t_{ij} εξάγονται *frame* ή *segment - level* χαρακτηριστικά. Η ακολουθία περιγράφεται ως η συνένωση των επιμέρους *frame* ή *segment - level* διανυσμάτων και στη συνέχεια χρησιμοποιείται ως είσοδος στα ακολουθιακά μοντέλα.

5.4.2 *Frame - based* Ταξινόμηση

Για την *frame - based* ταξινόμηση κάθε ακολουθία $u_i = \{f_{i1}, f_{i2}, \dots f_{im}\}$ χωρίζεται σε επικαλυπτόμενα *frames* καθένα από τα οποία περιγράφεται με 80 χαρακτηριστικά. Στη συνέχεια τα *frames* ομαδοποιούνται ανά 10, δημιουργώντας ενα διάνυσμα διάστασης 800 για κάθε *time step* της ακολουθίας. Οι αρχιτεκτονικές των μοντέλων περιέχουν 2 κρυμμένα επίπεδα *LSTM* με 512 και 256 νευρώνες αντίστοιχα και 1 κρυμμένο επίπεδο *Dense* με 128 νευρώνες. Για την αναγνώριση του συναισθήματος στο διακριτό χώρο, η έξοδος των *single - task* και *multi - task* μοντέλων αποτελείται από ένα επίπεδο *softmax* 9 νευρώνων. Επιπλέον για την αναγνωριση συναισθήματος στο συνεχή χώρο καθώς και την αναγώριση φύλου, τα *multi - task* μοντέλα διαθέτουν εξόδους που αποτελούνται από μια *sigmoid activation function* και 1 νευρώνα. Καθένα από αυτά τα μοντέλα προεκπαιδεύεται μέσω της *IEMOCAP*. Ως *loss function* για κάθε έξοδο ορίζεται η *categorical cross entropy*, για την αναγνώριση συναισθήματος στο συνεχή χώρο, η *mean squared error* για την αναγνώριση συναισθήματος στο διακριτό χώρο καθώς και η *binary cross entropy* για την αναγνώριση φύλου. Στη συνέχεια εκπαιδεύεται και αξιολογείται σε *stressed* / *unstressed utterances* όπου ως *loss function* ορίζεται η *binary cross entropy*. Παρακάτω αναγράφονται τα αποτελέσματα από κάθε προεκπαιδευμένο μοντέλο.

Pretrain Task	UA
Emotion	68%
Emotion - Gender	69%
Emotion - VAD	69%

Πίνακας 5.7: Μετρική *UA*(%) αναγώρισης *stressed* / *unstressed utterances* για κάθε *frame - based* προεκπαιδευμένο μοντέλο.

Η *frame - based* απλή ταξινόμηση δεν παρουσιάζει ιδιαίτερη βελτίωση σε σχέση με την *utterance - based* ταξινόμηση. Προηγουμένως θεωρήθηκε πως το συναίσθημα εκφράζεται σε όλο το μήκος του δείγματος. Εδώ από τα δείγματα κατακερματίστηκαν σε πολύ μικρά υποτμήματα. Συνεπώς πολλά από τα οποία περιέχουν άφωνους ήχους ενώ τα έμφωνα τμήματα διαρκούν τόσο λίγο που ο ομιλητής δεν προλαβαίνει να εκφράσει κάποια συναισθηματική κατάσταση. Σε επίπεδο σύγκρισης των επιδόσεων των *pretrain tasks* παρατηρείται πως βρίσκονται αρκετά κοντά, με μια μικρή υπεροχή των *pretrain multi - tasking* μοντέλων. Μέχρι στιγμής εξετάστηκαν 2 αντίθετες περιπτώσεις όπου εφαρμόστηκαν χαρακτηριστικά σε όλο το μήκος του δείγματος αλλά και σε πολύ μικρά υποτμήματά του. Σε αυτές τις περιπτώσεις φάνηκε η αδυναμία των μοντέλων τα οποία προσεγγίζουν το συναίσθημα σε πολύ μεγάλα ή μικρά τμήματα ομιλίας αντίστοιχα. Το επαγόμενο βήμα είναι η ταξινόμηση δειγμάτων σε

αρκετά μεγάλα τμήματα έτσι ώστε ο ομιλητής να προλαβαίνει να εκφράσει κάποιο συναίσθημα αλλά και σε αρκετά μικρά προκειμένου να μην αποσιωποίεται η συναισθηματική πληροφορία.

5.4.3 *Segment - based* Ταξινόμηση

Για την *segment - based* ταξινόμηση κάθε ακολουθία $u_i = \{s_{i1}, s_{i2}, \ldots, s_{ij}\}$ χωρίζεται σε επικαλυπτόμενα *segments*. Σε κάθε *segment* εξάγονται τα ίδια χαρακτηριστικά με την περίπτωση της *utterance - based* ταξινόμησης, παράγοντας ένα διάνυσμα διάστασης 1582. Στη συνέχεια τα *segments* τοποθετούνται το ένα διπλά στο άλλο αναπαριστώντας την ακολουθία. Εδώ θεωρείται πως κάθε *segment* διαρκεί 3, 4 ή 5 *seconds* με 50% επικάλυψη μεταξύ τους. Τέλος οι ίδιες αρχιτεκτονικές καθώς και οι *loss functions*, που χρησιμοποιούνται στην περίπτωση της *frame - based* ταξινόμησης εφαρμόζονται και εδώ. Παρακάτω αναγράφονται τα αντίστοιχα αποτελέσματα για κάθε προεκπαιδευμένο μοντέλο και για κάθε διάρκεια των *segments*.

	Segment Duration		
Pretrain Task	UA (3secs)	UA (4 secs)	UA (5secs)
Emotion	68%	68%	68%
Emotion - Gender	71%	70%	70%
Emotion - VAD	70%	70%	**72**%

Πίνακας 5.8: Μετρική *UA*(%) αναγώρισης *stressed* / *unstressed utterances* για κάθε *segment - based* προεκπαιδευμένο μοντέλο.

Στον Πίνακα 5.8 φαίνεται πως επηρεάζει την επίδοση των μοντέλων το μήκος των *segments* στα οποία κατακερματίζεται η ακολουθία καθώς και τα *pretrain tasks* ξεχωριστά. Όσο αναφορά την διάρκεια των *segments*, φαίνεται πως στην περίπτωση των 5 *sec* η συναισθηματική κατάσταση του ομιλητή εκφράζεται χωρις να αποσιωποιείται. Ωστόσο, ακόμα και στις περιπτώσεις των 3 και 4 *sec* οι επιδόσεις των *segment - based* μοντέλων υπερτερούν των *frame* και *utterance - based*, καθώς το συναίσθημα αποτυπώνεται καλύτερα στα αντίστοιχα χρονικά διαστήματα. Από την άλλη πλευρά φαίνεται στην περίπτωση των *segment - based* μοντέλων φαίνεται η υπεροχή των *multi - tasking* μοντέλων. Το κλασικό *emotion single - task* δεν συνεισφέρει στην επίδοση του μοντέλου για καμία από τις διαφορετικές διάρκειες των *segments*, ενώ η επίδοση του μοντέλου ταυτίζεται με την περίπτωση *utterance* και *frame - based* ταξινόμησης. Η εικόνα άλλαζει στην περίπτωση των *segment - based* μοντέλων όπου σημειώνεται η καλύτερη επίδοση μέσω ενός *multi - tasking emotion - valence, activation, dominance* μοντέλου στο οποίο κάθε *segment* διαρκεί 5 *sec*.Στο στάδιο της προεκπαίδευσης, η ταυτόχρονη αναγνώριση συναισθήματος στο συνεχή και διακριτό χώρο δίνει περισσότερη γνώση στο μοντέλο, με αποτέλεσμα να ανταποκρίνεται καλύτερα στο επόμενο *task*. Ακολουθεί με μικρή διαφορά η επίδοση του *multi - tasking emotion - gender* μοντέλου όπου κάθε *segment* διαρκεί 3 *sec*.

5.5 Πειραματικές Διατάξεις και Αποτελέσματα Ταξινόμησης μέσω *handcrafted features representation learning*

Στα παραπάνω πειράματα χρησιμοποιήθηκαν *handcrafted features* από τον ευρύτερο χώρο της Αναγνώρισης Συναισθήματος μέσω Φωνής, με την ελπίδα ότι ανταποκρίνονται ικανοποιητικά στην Αναγνώριση Άγχους μέσω Φωνής. Φάνηκε πως η εξέταση η εφαρμογή καθολικών χαρακτηριστικών σε όλο το μήκος των δειγμάτων αλλα και η εφαρμογή χαρακτηριστικών σε τμήματα μικρού μήκους δεν επιφέρει σημαντικά αποτελέσματα. Αντίθετα η εφαρμογή μιας μέσης λύσης όπως η *segment - based* ταξινόμηση δείχνει πιο υποσχόμενη.

Στόχος των επόμενων πειραμάτων είναι η σύγκριση αλγορίθμων εξαγωγής *representations* από τον χώρο των *handcrafted* σε ένα χώρο αυθαίρετο χώρο μικρότερης διαστασιμότητας. Για τον σκοπό αυτό χρησιμοποιείται ο αλγόριθμος ο αλγόριθμος *PCA* καθώς και οι *autoencoders*, ενώ για να υπάρξει δίκαιη σύγκριση μεταξύ των αλγορίθμων η διαστασιμότητα του χώρου αναπαράστασης θεωρείται ίδια και στις 2 περιπτώσεις. Έτσι αν η $\mathbf{x}_{ti}$ είναι το διάνυσμα χαρακτηριστικών για το *time step* t της ακολουθίας u_i, τότε οι *representation learning* αλγόριθμοι προβάλλουν το $\mathbf{x}_{ti}$ σε χώρους διάστασης 128 και 256. Από την πλευρά της *PCA* επιλέγονται οι πρώτες 128 ή 256 κύριες συνιστώσες του $\mathbf{x}_{ti}$, ενώ από τους *autoencoders* χρησιμοποιείται η ενδιάμεση αναπαράσταση μεταξύ *encoder* και *decoder*, όπως φαίνεται στο Σχήμα 5.4. Ειδικότερα για τους *autoencoders* αν $\mathbf{x}_t$, $\tilde{\mathbf{x}}_t$ η είσοδος και η έξοδος των δικτύων τότε για τις *loss function* προκύπτει :

VAN

Αν $\mathbf{W}$, $\mathbf{b}$ είναι οι παράμετροι του δικτύου τότε από την εξίσωση 4.8 η *loss function* ορίζεται ως το άθροισμα της *mean squared error* μαζί με ένα όρο κανονικοποιήσης των παραμέτρων:

$$J(\mathbf{x}_t, \tilde{\mathbf{x}}_t) = \frac{1}{N}\sum_{i=1}^{N} ||x_{ti} - \tilde{x}_{ti}||^2 + \lambda\Omega(\mathbf{W}, \mathbf{b}) \tag{5.6}$$

VAE

Αν θ, ϕ είναι οι παράμετροι του *encoder* και *decoder* αντίστοιχα τότε από την εξίσωση 4.15b η *loss function* ως το άθροισμα της αρνητικής αναμενόμενης τιμής του λογαριθμού της *likelihood* μαζί με την απόκλιση *KL* μεταξύ των κατανομών του *encoder* και *decoder*, και ενός όρου κανονικοποίησης των παραμέτρων:

$$J_i(\mathbf{x}_i, \theta, \phi) = -\mathbb{E}_{\mathbf{h}_i \sim q_\theta(\mathbf{h}_i|\mathbf{x}_i)}\big\{log\big(p_\phi(\mathbf{h}_i|\mathbf{x}_i)\big)\big\} + KL\big\{q_\theta(\mathbf{h}_i|\mathbf{x}_i)||p(\mathbf{h}_i)\big\} + \lambda\Omega(\theta, \phi) \tag{5.7}$$

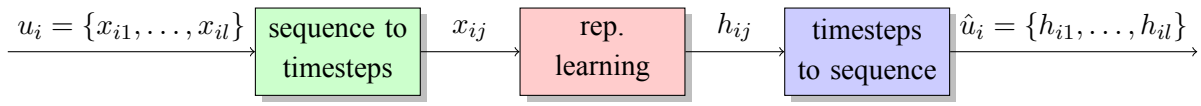
CVAE

Παρόμοια με την περίπτωση του *VAE*, η *loss function* του *CVAE* προκύπτει μέσω της εξίσωσης 4.16b :

$$J_i(\mathbf{x}_i, \theta, \phi) = -\mathbb{E}_{\mathbf{h}_i \sim q_\theta(\mathbf{h}_i|\mathbf{x}_i, \mathbf{c}_i)}\big\{log\big(p_\phi(\mathbf{h}_i|\mathbf{x}_i, \mathbf{c}_i)\big)\big\} + KL\big\{q_\theta(\mathbf{h}_i|\mathbf{x}_i, \mathbf{c}_i)||p(\mathbf{h}_i|\mathbf{c}_i)\big\} + \lambda\Omega(\theta, \phi) \tag{5.8}$$

Τέλος, ο όρος κανονικοποίησης αφορά τις L_2 νόρμες των πολώσεων και βαρών των δικτύων :

$$\Omega(\mathbf{W}, \mathbf{b}) = ||\mathbf{W}||^2 + ||\mathbf{b}||^2 \tag{5.9}$$



Σχήμα 5.8: Σχηματική αναπαράσταση *representation learning* μιας ακολουθίας.

Η διαδικασία κατασκευής *representations* μιας ακολουθίας περιγράφεται με το Σχήμα 5.8. Σε γενικές γραμμές κάθε ακολουθία κατακερματίζεται σε *time steps* όπου προβάλλονται στο χώρο αναπαράστασης. Η νέα αναπαράσταση της ακολουθίας θεωρείται η συνένωση των επιμέρους νέων αναπαραστάσεων των *time steps*. Οι αλγόριθμοι *representation learning* εφαρμόζονται αρχικά στην *IEMOCAP*

όπου μαθαίνεται ο μετασχηματισμός και στη συνέχεια ο μετασχηματισμός αυτός εφαρμόζεται σε *stressed* / *unstressed utterances*.

Από τις μεθόδους *representation learning* που εξετάζονται, ιδιαίτερο ενδιαφέρον παρουσιάζει η περίπτωση του *CVAE*. Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, κάθε είσοδος $\mathbf{x}_i$ του *CVAE* δεσμεύεται σε ένα διάνυσμα $\mathbf{c}_i$ οπού συνήθως αναφέρεται στην ετικέτα της εισόδου. Υπάρχουν 2 θέματα σχετκά με την επιλογή του $\mathbf{c}_i$. Αρχικά το $\mathbf{x}_i$ αντιστοιχεί σε ενα *time step* της ακολουθίας, επομένως δεν υπάρχει κάποια ετικέτα για το συγκεκριμένο *time step*, παρά μόνο η ετικέτα ολόκληρης της ακολουθίας. Μία απλή προσέγγιση είναι η αποδοχή της ετικέτας της ακολουθίας από κάθε *time step*. Έτσι αν η ακολουθία έχει κάποια ετικέτα c τότε όλα τα *time steps* της ακολουθίας έχουν την ίδια ετικέτα. Επιπλέον ο *CVAE* προεκπαιδεύεται μέσω της *IEMOCAP* και ύστερα εφαρμόζεται σε *stressed* / *unstressed utterances*. Ακόμα και στην παραπάνω απλή προσέγγιση, υπάρχουν διαφωνίες μεταξύ των ετικετών των 2 συνόλων. Μια λύση στο παραπάνω πρόβλημα είναι η εύρεση ενός μετασχηματισμού που συνδέει τα 2 σύνολα ετικετών. Ωστόσο η εύρεση τέτοιου μετασχηματισμού είναι αρκετά δύσκολη ή σε πολλές περιπτώσεις μη ρεαλιστική εξαιτίας της μη αντικειμενικής επισήμανσης των εκφωνήσεων, καθώς υπάρχουν διαφορετικοί κριτές υπεύθυνοι για τις ετικέτες μεταξύ των 2 συνόλων. Έτσι, θεωρείται πως ο μετασχηματισμός αυτός δεν υπάρχει, δηλαδή τα σύνολα ετικετών μεταξύ των συνόλων είναι ξένα. Με βάση αυτές τις προσεγγίσεις το διάνυσμα $\mathbf{c}_i$ αντιστοιχεί σε ένα *one - hot - vector* διάστασης ίσης με το άθροισμα των πιθανών ετικετών των συνόλων : $9 + 2 = 11$.

Autoencoders		
interm dim	256	512
latent dim	128	256
interm dim	256	512

Πίνακας 5.9: Αριθμός νευρώνων ανά επίπεδο των *autoencoders*.

5.5.1 *Frame - based* Ταξινόμηση

Όπως και στην περίπτωση της απλής *frame - based* ταξινόμησης κάθε *time step* αποτελείται από ένα διάνυσμα 800 χαρακτηριστικών. Από το *time step* εξάγονται *representations* διάστασης 128 ή 256 μέσω *PCA* ή *autoencoders*. Ειδικότερα η αρχιτεκτονική των *autoencoders* δίνεται από Πίνακα 5.9. Οι ακολουθίες χρησιμοποιούνται μέσω των νέων αναπαραστάσεων για την προεκπαίδευση των μοντέλων στα ίδια *tasks* της απλής *frame - based* ταξινόμησης με τις ίδιες *loss functions*. Στον Πίνακα 5.10 δίνονται οι επιδόσεις των μοντέλων για διαφορετική διαστασιμότητα της *representation* και διαφορετικούς αλγόριθμους εξαγωγής.

250 msec		128 latent dim				256 latent dim			
	Pretrain Task	PCA	VAN	VAE	CVAE	PCA	VAN	VAE	CVAE
	Emotion	69%	70%	69%	69%	69%	68%	69%	69%
	Emotion - Gender	69%	68%	68%	67%	70%	**71%**	68%	68%
	Emotion - VAD	70%	69%	68%	69%	69%	68%	**71%**	68%

Πίνακας 5.10: Μετρική *UA*(%) αναγώρισης *stressed* / *unstressed utterances* για κάθε *frame - based* προεκπαιδευμένο μοντέλο με *handcrafted features representation learning*.

Συγκρίνοντας τους Πίνακες 5.7 και 5.10 παρατηρείται μικρή αύξηση στην επίδοση της *frame - based* ταξινόμησης σε κάθε *task* ξεχωριστά. Το *single - task emotion* μοντέλο παρουσίαζει καλύτερο ποσοστό επιτύχίας 70% με χρήση αναπαράστασης του *VAN*. Τα *multi - tasking emotion - gender* και

emotion - valence, activation, dominance παρουσιάζουν ποσοστό επιτυχίας 71% με χρήση αναπαράστασης του *VAN* και *VAE* αντίστοιχα. Η επίδοση αυτή είναι και η καλύτερη για την περίπτωση της *frame - based* ταξινόμησης παρουσιάζοντας βελτίωση 4.22% σε σχέση με την *utterance - based baseline* ταξινόμηση. Εξετάζοντας κάθε *representation learning* αλγόριθμο ξεχωριστά παρατηρείται ότι οι γραμμικές αναπαράστασεις της *PCA* υστερούν από τις μη γραμμικές αναπαραστάσεις των *autoencoders* σε κάθε *pretrain task* και σε κάθε διαφορετική διάσταση της αναπαράστασης. Ωστόσο ακόμα και με *frame - based representation learning*, η απλή *segment - based* υπερτερεί της *frame - based* ταξινόμησης.

5.5.2 *Segment - based* Ταξινόμηση

Παρόμοια λογική εφαρμόζεται και στην *segment - based* ταξινόμηση με *representation learning*. Κάθε *time step* αντιστοιχεί σε ένα *segment* 1582 χαρακτηριστικών στο οποίο εξάγονται *segment - based representations* μέσω *PCA* ή *autoencoders*. Στη συνέχεια τα μοντέλα προεκπαιδεύονται όπως και στην περίπτωση της απλής *segment - based* ταξινόμησης όπου οι λεπτομέρειες των αρχιτεκτονικών δίνονται παλί από τον Πίνακα 5.9. Παρακάτω δίνονται οι επιδόσεις των μοντέλων για διαφορετικό μήκος *segment*.

		128 latent dim				256 latent dim			
3 secs	Pretrain Task	PCA	VAN	VAE	CVAE	PCA	VAN	VAE	CVAE
	Emotion	68%	69%	68%	68%	68%	70%	68%	69%
	Emotion - Gender	69%	68%	70%	69%	**72%**	70%	71%	69%
	Emotion - VAD	68%	68%	68%	68%	69%	68%	69%	70%

Πίνακας 5.11: Μετρική *UA*(%) αναγώρισης *stressed* / *unstressed utterances* για κάθε *segment - based* προεκπαιδευμένο μοντέλο με *handcrafted features representation learning* (κάθε *segment* διαρκεί 3 δευτερόλεπτα).

		128 latent dim				256 latent dim			
4 secs	Pretrain Task	PCA	VAN	VAE	CVAE	PCA	VAN	VAE	CVAE
	Emotion	70%	71%	70%	72%	72%	72%	72%	70%
	Emotion - Gender	71%	73%	74%	**75**%	69%	74%	74%	71%
	Emotion - VAD	71%	73%	72%	73%	70%	71%	74%	72%

Πίνακας 5.12: Μετρική *UA*(%) αναγώρισης *stressed* / *unstressed utterances* για κάθε *segment - based* προεκπαιδευμένο μοντέλο με *handcrafted features representation learning* (κάθε *segment* διαρκεί 4 δευτερόλεπτα).

		128 latent dim				256 latent dim			
5 secs	Pretrain Task	PCA	VAN	VAE	CVAE	PCA	VAN	VAE	CVAE
	Emotion	70%	71%	71%	70%	69%	71%	73%	71%
	Emotion - Gender	70%	73%	73%	73%	68%	73%	73%	71%
	Emotion - VAD	72%	73%	**74**%	70%	71%	73%	71%	**74**%

Πίνακας 5.13: Μετρική *UA*(%) αναγώρισης *stressed* / *unstressed utterances* για κάθε *segment - based* προεκπαιδευμένο μοντέλο με *handcrafted features representation learning* (κάθε *segment* διαρκεί 5 δευτερόλεπτα).

Στην *frame - based* ταξινόμηση υπήρξε μικρή βελτίωση με χρήση *representation learning* αλγορίθμων. Η επίδραση των αναπαραστάσεων γίνεται πιο αισθητή συγκρίνοντας τις *segment - based* ταξινομήσεις των Πινάκων 5.8 και 5.11 - 5.13. Πιο συγκεκριμένα το *single task - emotion* μοντέλο παρουσιάζει ποσοστό επιτυχίας 73% για 5 *sec segment* μέσω μιας *VAE* αναπαράστασης η οποία παρουσιάζει αύξηση κατά 6.84% σε σχέση με την περίπτωση της απλής ταξινόμησης, το *multi tasking emotion - gender* μοντέλο 75% για 4 *sec segment* μέσω μιας *CVAE* αναπαράστασης παρουσιάζοντας αύξηση κατά 5.33%, ενώ το *multi tasking emotion - valence, activation, dominance* μοντέλο 74% για 4 και 5 *sec segment* με αναπαραστάσεις *VAE* και *CVAE* με αύξηση 2.70%. Με άλλα λόγια όλα τα *pretrain tasks* μοντέλα επωφελούνται από το *representation learning*. Από την σκοπία των αλγορίθμων φαίνεται πως *autoencoders* υπερτερούν της κλασικής *PCA* στα περισσότερα μοντέλα, ενώ όσο αναφορά τους *autoencoders*, τα *variational* μοντέλα, *VAE* και *CVAE*, δείχνουν να αναπαραστούν καλύτερα τα τα δείγματα σε χώρους μικρότερης διάστασης από τον *VAN*. Επιπλέον ενισχύεται ο ισχυρισμός του κατακερματισμού των δεδομένων σε υποτμήματα. Όπως φαίνεται και στους παραπάνω πίνακες, η *segment - based* ταξινόμηση υπερτερεί των υπολοίπων. Η επίδοση του *multi tasking emotion - gender* μοντέλου με *representation learning* παρουσιάζει βελτίωση κατά 9.33% σε σχέση με την *baseline* περίπτωση και υπερτερεί όλων των υπολοίπων μοντέλων. Μπορεί να μην υπάρχει καθολική βέλτιστη επιλογή για την διάρκεια των *segments*, φαίνεται ωστόσο πως ο κατακερματισμός των δειγμάτων σε μικρά ή μεγάλα τμήμα δεν αποτελεί την καλύτερη επιλογή.

5.6 Πειραματικές Διατάξεις και Αποτελέσματα Ταξινόμησης μέσω συνθετικών δειγμάτων

Στην προηγούμενη παράγραφο έγινε εκτεταμένη χρήση της *representations* των *autoencoders*, όπου ύτερα από την εκπαίδευσή τους αποκόπτεται το δίκτυο του *decoder* και στη συνέχεια χρησιμοποιείται μόνο ο *encoder* για την εξαγωγή της *representations*. Μολονότι αυτή είναι η κύρια λειτουργία του *VAN*, δεν ισχύει το ίδιο για τα *variational* μοντέλα. Όπως αναφέρθηκε και σε προηγούμενο οι *VAE* και *CVAE* μπορούν να εκτιμήσουν συνθετικά μοντέλα μέσω των κατανομών των επιμέρους δικτύων. Εδώ εξετάζεται η ικανότητα γέννησης των *variational* μοντέλων. Σε αρκετές περιπτώσεις τα δεδομένα εκπαίδευσης δεν επαρκούν για την εκπαίδευση μοντέλων με αποτέλεσμα την χαμηλή τους επίδοση. Η ιδέα αυτή πρόερχεται από μοντέλα σύνθεσης μουσικής [72] ή εικόνας [112] όπου σε αυτές τις περιπτώσεις είναι εφικτή η χρήση του *VAE* αλλά και του *CVAE*. Ωστόσο με σκοπό την αύξηση του συνόλου εκπαίδευσης κατά την *supervised* εκπαίδευση χρησιμοποιείται μόνο ο *CVAE*. Έστω ότι χρησιμοποιούνται τα *variational* μοντέλα για αύξηση του συνόλου εκπαίδευσης σε ένα πρόβλημα αναγνώρισης ζώων από εικόνες. Σε αυτήν την περίπτωση τα συνθετικά δείγματα που παράγει ο *VAE* δεν μπορούν να ταυτοποιηθούν. Με άλλα λόγια δεν μπορεί να προσδιοριστεί αυτόματα αν το παραγόμενο δείγμα, δηλαδή η εικόνα αφορά ένα σκύλο ή ενα άλογο. Αντίθετα, στην περίπτωση του *CVAE*, επειδή τα δείγματα δεσμεύονται ως προς τις αντίστοιχες ετικέτες, μπορούν να γεννηθούν δείγματα τα οποία ανήκουν σε συγκεκριμένες κλάσεις. Εντελώς αντίστοιχα με την περίπτωση των εικόνων, ένα συνθετικό δείγμα φωνής το οποίο προέρχεται από έναν *VAE* δεν μπορεί να κατηγοριοποιηθεί σε κάποια κλάση συναισθήματος. Για τον λόγο αυτό, στα επόμενα πειράματα χρησιμοποιείται μόνο ο *CVAE*.

Έστω λοιπόν η αξιολόγηση ενός μοντέλου μέσω ενός 10 - *CV*, όπου σε κάθε *fold* το μοντέλο εκπαιδεύεται με X_{train} και αξιολογείται με X_{test} δείγματα. Επιθυμείται ο διπλασιασμός του συνόλου εκπαίδευσης, δηλαδή αν N_i είναι τα δείγματα της κλάσης *i* τότε επιθυμείται η κατασκευή N_i συνθετικών δειγμάτων. Έστω επιπλέον ότι τα δείγματα αντιστοιχούν σε ακολουθίες $u_i = \{t_{i1}, t_{i2}, \ldots, t_{ij}\}$ αποτελούμενες από επικαλυπτόμενα *time steps*. Η συνθετική ακολουθία $\hat{u}_i$ προκύπτει ως εξής :

1. Κατασκευή του ιστογράμματος κάθε κλάσης ως προς τον αριθμό των *time steps* που περιέχουν οι ακολουθίες που ανήκουν στην κλάση, *hist(number of time steps)*.

2. Ταίριασμα μιας κατανομής πάνω στο ιστόγραμμα p_i. Η κατανομή που χρησιμοποιείται αντιστοιχεί σε μια κανονική κατανομή που ταιρίαζει πάνω στο ιστόγραμμα μέσω υπολογισμού της μέσης τιμής και της τυπικής απόκλισης του αριθμού των *time steps* των ακολουθιών.

3. Από την κατανομή του αριθμού των *time steps* p_i της κλάσης *i* λαμβάνεται ένα δείγμα n. Η τιμή n αντιστοιχεί στον αριθμό των *time steps* που περιέχει η συνθετική ακολουθία.

4. Κατασκευή n συνθετικών *time steps* από τον *CVAE* τα οποία στη συνέχεια τοποθετούνται το ένα διπλά στο άλλο παράγοντας της συνθετική ακολουθία $\hat{u}_i$.

Επιστρέφοντας στο σύνολο *stressed / unstressed* δειγμάτων, σε κάθε *fold* τα μοντέλα εκπαιδεύονται με συνολικα 270 δείγματα και αξιολογούνται στα υπόλοιπα 30. Με την παραπάνω διαδικασία παράγονται επιπλέον 270 συνθετικά δείγματα με τις ίδιες αναλογίες κλάσεων. Τα μοντέλα πλέον εκπαιδεύονται με 540 δείγματα και αξιολογούνται στα αρχικά 30. Σημειώνεται ότι η συνθετική ακολουθία $\hat{u}_i$ έχει την ιδια ετικέτα με την γνήσια u_i, ωστόσο ενδέχεται να διαφέρουν ως προς το μήκος τους.

Η παραπάνω τεχνική κατασκευής συνθετικών δειγμάτων εφαρμόζεται σε *single - task emotion* μοντέλα, με *segment - based* ταξινόμηση με 2 τρόπους. Ο πρώτος τρόπος αντιστοιχεί στον διπλασιασμό των δειγμάτων εκπαίδευσης κατά την διάρκεια του 10 - *CV* όπως περιγράφηκε (*augmenting train*). Ο δεύτερος τρόπος αντιστοιχεί στην ταξινόμηση των γνήσιων ακολουθιών μέσω μοντέλων που έχουν εκπαιδευτεί αποκλειστικά από συνθετικές ακολουθίες. Σε αυτόν τον τρόπο παράγονται 300 συνθετικές ακολουθίες με τις οποίες εκπαιδεύονται τα μοντέλα και αξιολογούνται στις 300 αρχικές γνήσιες ακολουθίες (*synthesized vs real - svr*). Παρακάτω δίνονται οι επιδόσεις των μοντέλων για διαφορετικό μήκος *segment*, 3, 4 και 5 *sec* και για διαφορετικές διαστάσεις της κατανομής του *CVAE*, 128 και 256.

	3 secs		4 secs		5 secs	
Latent dim	128	256	128	256	128	256
Augmenting Train	69%	68%	68%	68%	69%	69%
Svr	67%	66%	69%	68%	69%	69%

Πίνακας 5.14: Μετρική *UA*(%) αναγώρισης *stressed / unstressed utterances* για κάθε *segment - based* προεκπαιδευμένο μοντέλο με *synthesized* δείγματα.

Παρά την αναγνωρισμένη χρήση των *variational* μοντέλων σε διαφορετικούς κλάδους για την γεννηση συνθετικών δειγμάτων, παρατηρείται πως τα συνθετικά δείγματα για την αναγνώριση συναισθήματος δυσχαιρένουν τα μοντέλα. Οι επιδόσεις του πίνακα 5.14 παρουσιάζουν πτώση σε σχέση με την αρχική *segment - based*. Τα *variational* αποδείχτηκαν αρκετά ικάνα για εξαγωγή κάποιας χρήσιμης αναπαράστασης. Δεν ισχύει όμως το ίδιο για την κατασκεύη εξ ολοκλήρου κατασκεύη συνθετικών δειγμάτων.

5.7 Πειραματικές Διατάξεις και Αποτελέσματα Ταξινόμησης μέσω *raw signals representation learning*.

Στα προηγούμενα πειράματα εξάχθηκαν *utterance*, *segment* και *frame - level representations* από *handcrafted features* προβάλλοντας *time steps* ακολουθιών σε ένα υψηλότερου επιπέδου αφηρημένο χώρο αναπαράστασης χαρακτηριστικών. Αυτές οι αναπαραστάσεις εξάγονται από μοντέλα που δέχονται σταθερού μήκους είσοδο και στη συνέχεια χρησιμοποιούνται σε *LSTM* μοντέλα για περαιτέρω *pretrain* και *fine tunning*. Παρά τις υποσχόμενες επιδόσεις των μοντέλων η τεχνική αυτή περιορίζει την ικανότητα των *representation learning* αλγορίθμων με 2 τρόπους :

1. Παραδοσιακά τα *handcrafted features* αποτυπώνουν την εμπειρική γνώση του ανθρώπου ως προς το "τι είναι χρήσιμο" για να περιγραφεί ένα σήμα φωνής. Δεν υπάρχει κάποιος καθολικά αποδεκτός φορμαλισμός που υποστηρίζει πως αυτά τα *handcrafted features* είναι πάντα μια καλή επιλογή αναπαράστασης των δεδομένων.

2. Συνήθως ένα *time step* της ακολουθίας θεωρείται ως η μικρότερη χρονική μονάδα που περιέχει κάποια χρήσιμη πληροφορία. Υπάρχουν 2 προβλήματα σχετικά με αυτήν την προσέγγιση. Αρχικά, όπως αναφέρθηκε και στα προηγούμενα πειράματα, δεν υπάρχει περαιτέρω φορμαλισμός που υποστηρίζει την καθολική βέλτιστη χρονική διάρκεια ενός *time step*. Για παράδειγμα ενδέχεται η βέλτιστη διάρκεια ενός *time step* για την αναγνώριση χαράς στην ομιλία μιας ομάδας ανθρώπων να ισούται με κάποια τιμή, δεν ισχύει απαραίτητα το ίδιο για την αναγνώριση διαφορετικού συναισθήματος στην ίδια ομάδα ανθρώπων είτε ίδιο συναισθήματος σε διαφορετική ομάδα ανθρώπων. Επιπλέον επειδή κάθε *time step* δεν κατακεριματίζεται περαιτέρω αγνοείται πιθανή ακολουθιακή πληροφορία που περιέχεται μέσα στο *time step*.

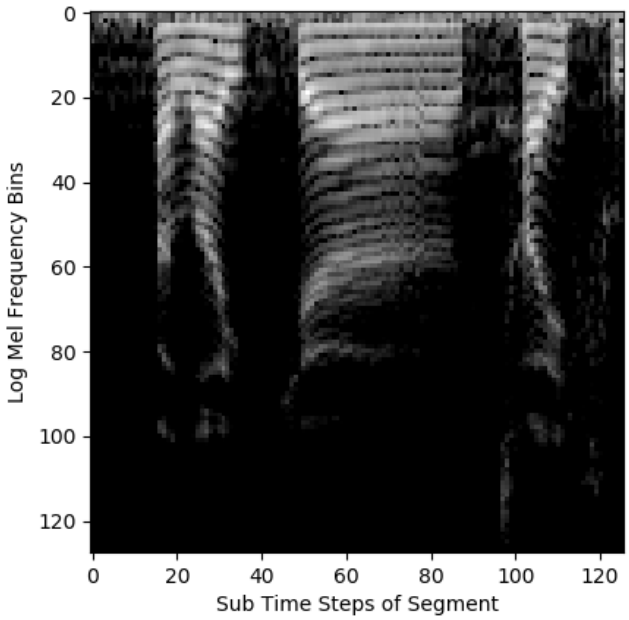
Με σκοπό την αντιμετώπιση των παραπάνω προβλημάτων προτείνεται ένα *seq2seq representation learning* μοντέλο όπου τα δεδομένα φωνής απεικονίζονται σε μια "πρωιμη" (*raw*) μορφή. Η *raw* μορφή των δεδομένων επιτρέπει στο μοντέλο να μάθει αφηρημένα χρήσιμα χαρακτηριστικά από την είσοδο. Έτσι οι *representation learning* αλγορίθμους διαθέτουν όσο το δυνατό περισσότερο ελευθερία ως προς την εκφραστικότητα τους, παρακαπτοντας των περιορισμό γνώσης που επιβάλλουν τα *handcrafted features*. Επιπλέον αφού κάθε *time step* απεικονίζεται στην *raw* μορφή του ενδέχεται να υπάρχει κάποια ακολουθιακή πληροφορία μέσα στο ίδιο το *time step*.

Πιο συγκεκριμένα έστω $u = \{s_0, s_1, \ldots, s_l\}$ μια ακολουθία ενός σήματος φωνής χωρισμένη σε *l time steps* σε επίπεδο *segment*. Επιθυμείται η εξαγωγή μιας *utterance representation* η οποία προέρχεται από την συνένωση των επιμέρους *segment representation*. Επιπλέον επιθυμείται μια αρκετά βολική *raw representation* για κάθε *segment*, έτσι ώστε να χρησιμοποιηθούν στους *representation learning* αλγορίθμους. Με βάση την απαίτηση και εμπνέομενοι από τις μελέτες [113, 114, 115] τα *segments* απεικονίζονται πριν την εξαγωγή των *MFCCs* δηλαδή σε μορφή *spectogram* όπως φαίνεται και στο Σχήμα 5.9. Έτσι, κάθε *segment* απεικονίζεται σε μορφή *spectogram* $P(F_{bins}, N)$ όπου F_{bins} ο αριθμός της συστοιχίας των τριγωνικών φίλτρων και N ο αριθμός των χρονικών υπότμημάτων που χωρίζεται το *segment* κατά τον μετασχηματισμό DFT. Με αυτή την προσέγγιση τα δεδομένα μοντελοποιούνται ως εξής :

1. *Segment - level representation* : Αρχικά υπενθυμίζεται πως ένα *segment* αντιστοιχεί σε σταθερού μήκους υποτμήμα της ακολουθίας. Έτσι αν $P_i(F_{bins}, N)$ είναι το *melspectogram* του i - οστού *segment* τότε η διαστασιμότητα της αναπαράστασης του κάθε *segment* είναι σταθερή, δηλαδή η τιμή N είναι ίδια για όλα τα *segments*. Επιπλέον ενσωματώνεται η ακολουθιακή πληροφορία που πιθανώς να βρίσκεται μέσα στο *segment*. Στο Σχήμα 5.9 ο οριζόντιος άξονας

αντιστοιχεί στις χρονικές υπομονάδες που χωρίζεται ένα *segment* (*sub - time steps*). Κάθε τέτοιο *sub - time step* αναπαριστάται από F_{bins} χαρακτηριστικά (κάθετος άξονας) στο αντίστοιχο εύρος συχνοτήτων. Αυτό επιτρέπτει *seq2seq representation learning* σε επίπεδο *segment*. Σημειώνεται πως η ικανότητα αυτή δεν είναι δεδομένη στην περίπτωση των *handcrafted features* οπού εξάγονται *functionals* για να περιγράψουν το *segment* σε όλο το μήκος του αγνοώντας την χρονική σειρά.

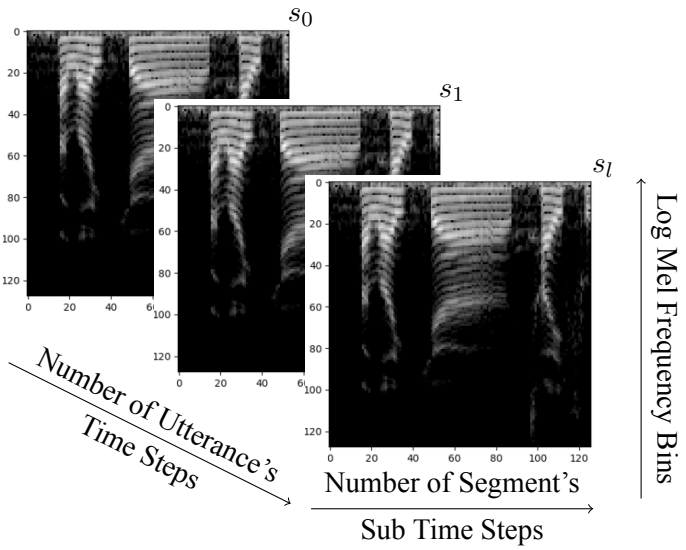
2. *Utterance - level representation* : με βάση την *segment - level representation* αν $P_i(F_{bins}, N)$ είναι το *melspectogram* του i - οστού *segment* της ακολουθίας u, τότε η u αναπαριστάται από την επιμέρους συνένωση των *melspectograms*, όπως φαίνεται και στο Σχήμα 5.10.



Σχήμα 5.9: *Spectogram* ενός τυχαίου *segment*. Στον οριζόντιο και κάθετο άξονα δίνονται τα υποτμήματα του *segment* και τα *bins* της συχνότητας στην κλίμακα *mel* αντίστοιχα.

Τα πλεονεκτήματα της αναπαράστασης της ακολουθίας με αυτόν τον τρόπο μπορούν αν εκφραστούν ως εξής. Όπως αναφέρθηκε παραπάνω δεν χρησιμοποιούνται *handcrafted features*. Κάθε *segment* αναπαριστάται αυτούσιο στο διδιάστατο επίπεδο που ορίζουν οι χρονοσυχνοτικοί άξονες. Επιπλέον κάθε σταθερού μήκους *segment* εμπεριέχει την δικία του ακολουθιακή πληροφορία και έτσι μπορούν να εφαρμοστούν ακολουθιακά μοντέλα για την κατασκεύη *representations*.

Παρά το γεγονός ότι εδώ χρησιμοποιούνται ακολουθιακά μοντέλα, οι ίδιοι *autoencoders* εφαρμόζονται για την εύρεση *representations*. Ωστόσο εδώ οι *autoencoders* αναφέρονται στο Σχήμα 5.5 όπου τα πλήρως συνδεδεμένα επίπεδα νευρώνων του *encoder* και *decoder* του Σχήματος 5.4 έχουν αντικαθισταθεί από επίπεδα *LSTMs*. Ως είσοδος των μοντέλων θεωρείται το *spectogram* $P_i(F_{bins}, N)$ ενός *segment*. Για την κατασκευή των *spectograms* χρησιμοποιούνται $F_{bins} = 128$ εύρη συχνοτήτων στην κλίμακα *mel* και N χρονικές μονάδες οι οποίες προέρχονται από μετασχηματισμό *DFT* 512 σημείων με επικάλυψη 256 σημειών. Τα *spectograms* εισάγονται στους *autoencoders* όπου θεωρείται η έξοδος του *encoder* g_i ως η αναπαράστασή τους στον αφηρημένο χώρο. Τέλος η ακολουθία αναπαριστάται ως η επιμέρους συνένωση των αναπαραστάσεων g_i δηλαδή $u = \{g_1, g_2, \dots g_l\}$.



Σχήμα 5.10: Αναπαράσταση ακολουθίας μέσω *spectograms* σε επίπεδο *segment*.

Για την δίκαιη σύγκριση μεταξύ *handcrafted features* και *raw signals representation learning* αλγορίθμους, πραγματοποιούνται ακρίβως τα ίδια πείραματα, δηλαδή εξετάζονται οι ίδιες διαστασιμότητες της κρυφής αναπαράστασης, τα ίδια μήκη των *segments* καθώς και τα ίδια προεκπαιδευόμενα μοντέλα. Παρακάτω δίνονται οι αντίστοιχες επιδόσεις.

		128 latent dim			256 latent dim		
3 secs	Pretrain Task	VAN	VAE	CVAE	VAN	VAE	CVAE
	Emotion	73%	72%	73%	73%	74%	74%
	Emotion - Gender	75%	73%	74%	75%	74%	73%
	Emotion - VAD	73%	74%	73%	72%	72%	71%

Πίνακας 5.15: Μετρική *UA*(%) αναγώρισης *stressed* / *unstressed utterances* για κάθε *segment - based* προεκπαιδευμένο μοντέλο με *raw signals representation learning* (κάθε *segment* διαρκεί 3 δευτερόλεπτα)

		128 latent dim			256 latent dim		
4 secs	Pretrain Task	VAN	VAE	CVAE	VAN	VAE	CVAE
	Emotion	72%	72%	71%	71%	73%	72%
	Emotion - Gender	71%	73%	73%	71%	72%	71%
	Emotion - VAD	72%	72%	72%	71%	72%	72%

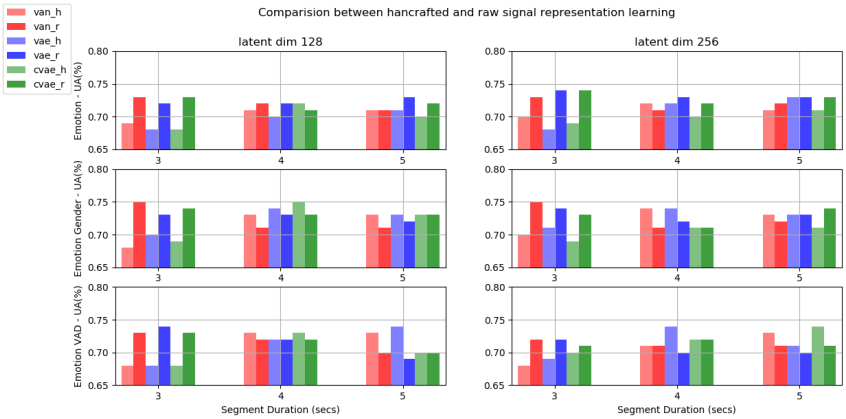
Πίνακας 5.16: Μετρική *UA*(%) αναγώρισης *stressed* / *unstressed utterances* για κάθε *segment - based* προεκπαιδευμένο μοντέλο με *raw signals representation learning* (κάθε *segment* διαρκεί 4 δευτερόλεπτα)

		128 latent dim			256 latent dim		
5 secs	Pretrain Task	VAN	VAE	CVAE	VAN	VAE	CVAE
	Emotion	71%	73%	72%	72%	73%	73%
	Emotion - Gender	71%	72%	73%	72%	73%	74%
	Emotion - VAD	70%	69%	70%	71%	70%	71%

Πίνακας 5.17: Μετρική *UA*(%) αναγώρισης *stressed* / *unstressed utterances* για κάθε *segment - based* προεκπαιδευμένο μοντέλο με *raw signals representation learning* (κάθε *segment* διαρκεί 5 δευτερόλεπτα)

Εξετάζοντας αποκλειστικά την τεχνική *raw signals representation learning*, αρχικά παρατηρείται πως υπερτερεί της *baseline* μεθόδου παρουσιάζοντας καλύτερο ποσοστό επιτυχίας 75%, δηλαδή αύξηση κατά 10.29% σε σχέση με την *baseline* επίδοση. Επιπλέον, όπως αναμενόταν με βάση και τα προηγούμενα πειράματα, φαίνεται πως το κλασικό *single task - emotion* υστερεί από τα *multi - tasking* μοντέλα, για διαφορετικές παραμέτρους των πειραμάτων. Η καλύτερη επίδοση εντοπίζεται στην περίπτωση ενός *multi - tasking emotion - gender* με χρήση της *VAN* αναπαράστασης για την περίπτωση 3 *sec segment*. Μολονότι η υπεροχή των *variational* μοντέλων σε σχέση με τον κλασικό *autoencoder* ηταν εμφανής στην περίπτωση της *handcrafted features representation learning* δεν ισχύει το ίδιο για την περίπτωση της *raw signals representation learning*. Εδώ ο *VAN* παρουσιάζει, με μικρή διαφορά, την καλύτερη επίδοση πιθανώς δίοτι η εκτίμηση κατανομών των *variational* είναι πιο δύσκολη στην περίπτωση όπου τα *time steps* αναπαρίστανται μέσω *spectograms*.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η σύγκριση των μοντέλων για διαφορετική διάρκεια των *time steps*. Τα μοντέλα των πειραμάτων *handcrafted features representation learning* παρουσιάσαν μικρές επιδόσεις για *time steps* μικρής διάρκειας. Μάλιστα η *segment - based* ταξινόμηση φάνηκε ανώτερη *frame - based* ταξινόμησης. Στην *segment - based* ταξινόμηση τα μοντέλα σημείωσαν μικρές επιδόσεις για την περίπτωση των 3 *sec*, μέγιστη στην περίπτωση των 4 *sec* ακολουθούμενη από μια μικρή πτώση για 5 *sec*. Εδώ η εικόνα αλλάζει καθώς οι καλύτερες επιδόσεις των μοντέλων εντοπίζονται επί το πλείστον στην περίπτωση των *segments* διάρκειας 3 *sec*. Στα *seq2seq* μοντέλα το ισοζύγιο μεταξύ της συναισθηματικής πληροφορίας και της διάρκειας των *time steps* αποτυπωνεται με διαφορετικό τρόπο σε σχέση με τα κλασικά, σταθερού εισόδου μοντέλα. Όσο μειώνεται το μήκος των *segments* τόσο μειώνεται και ο αριθμός N των χρονικών υπομονάδων μέσα σε κάθε *segment*. Επιπλέον η κάθε ακολουθία κατακερματίζεται σε περισσότερα τμήματα μικρότερου μήκους, παρέχοντας περισσότερα παραδείγματα για τα μοντέλα. Έτσι το πρόβλημα ανακατασκευής των *segments* για τα *seq2seq* γίνεται ευκολότερο, αφού διαθέτουν περισσότερα παραδείγματα με μικρότερη διάσταση το καθένα. Συνεπώς μαθαίνουν καλύτερες κρυφές αναπαραστάσεις οι οποίες χρησιμοποιούνται για την αναγνώριση του συναισθήματος. Ωστόσο μικρότερα *segments* συνεπάγεται μικρότερη συναισθηματική πληροφορία μέσα σε κάθε *segment*, δηλαδή το μετέπειτα προβλημα αναγνώρισης συναισθήματος δυσκολέυει. Αντίθετα όσο αυξάνεται το μήκος των *segments* τόσο αυξάνεται ο αριθμός των χρονικών υπομονάδων. Τα μοντέλα εκπαιδεύονται με λιγότερα παραδείγματα μεγαλύτερης διάστασης, δυσκολέυοντας το πρόβλημα της ανακατασκεύης ενώ παράλληλα διευκολύνεται το πρόβλημα της αναγνώρισης του συναισθήματος αφού υπάρχει περισσότερη συναισθηματική πληροφορία μέσα στο *segment*.



Σχήμα 5.11: Σύγκριση μεταξύ *handcrafted features representation learning* και *raw signals representation learning*. Με έντονο και ανοιχτό χρώμα απεικονίζονται οι επίδοσεις των *handcrafted features representation learning* και *raw signals representation learning* αντίστοιχα.

Μια αναλυτική σύγκριση των *representation learning* αλγορίθμων σε επίπεδο *segment* δίνεται στο Σχήμα 5.11 για όλα τα *pretrain tasks* και για όλες τις *latent dims*. Αρχικά παρατηρείται πως και οι 2 αλγόριθμοι επωφελούν περισσότερο τα *multi task* μοντέλα κυρίως το μοντέλο *emotion - gender*. Όσο αναφορά το μήκος των μήκος των *segment* οι *handcrafted features* αναπαραστάσεις δείχνουν να λειτουργούν καλύτερα για *segments* μεσαίας και μεγάλης διάρκειας δηλαδή για 4 και 5 *sec*. Αντίθετα οι *raw signals* αναπαραστάσεις δείχνουν να λειτουργούν καλύτερα για μικρά *segments* 3 *sec*. Ωστόσο και οι 2 μέθοδοι σημειώνουν παρόμοιες μέγιστες επιδόσεις οι οποίες εντοπίζονται είτε σε διαφορετικά *pretrain tasks*, *latent dims* ή *autoencoders*. Συμπερασματικά, οι *raw signals* αναπαράστασεις δεν υστερούν σε τίποτα από τις παραδοσιακές *handcrafted features* αναπαράστασεις. Τα μοντέλα είναι ικανά από μόνα τους να ανακαλύψουν χρήσιμα χαρακτηριστικά των σημάτων φωνής χωρίς την επίβλεψη και την εμπειρική γνώση των ανθρώπων που διοχεύεται στα μοντέλα μέσω των *handcrafted features*.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντικές Προεκτάσεις της Εργασίας

6.1 Συμπεράσματα

Σε αυτήν την εργασία εξετάσθηκε η ανάλυση του άγχους εμπεριεχόμενου σε σήματα φωνής, σε ένα πρόβλημα δυαδικής ταξινόμησης. Στόχος ήταν η κατασκευή ικανών αναπαραστάσεων έτσι ώστε να περιγραφούν τα σήματα φωνής και στη συνέχεια η κατασκεύη υπολογιστικών μηχανών με σκοπό την αναγνώρισή του άγχους. Ως βασική προσέγγιση θεωρήθηκε πως το συναίσθημα εκφράζεται σε όλοκληρο το σήμα φωνής. Κατασκευάστηκαν αναπαραστάσεις σε επίπεδο *utterance* μέσω εμπειρικών γνώσεων και χρησιμοποιήθηκαν *utterance - based* μοντέλα ως μια βασική πρώτη επίδοση. Στη συνέχεια θεωρήθηκε πως το συναίσθημα δεν εκφράζεται σε ολόκληρο το σήμα φωνής. Αντίθετα, το σήμα κατακερματίζεται σε τμήματα από τα οποία ένας μικρός αριθμός τους περιέχει κάποια συναισθηματική πληροφορία, ενώ τα υπόλοιπα συνιστούν τις μεταβάσεις μεταξύ των συναισθηματικών καταστάσεων του ομιλητή. Χρησιμοποιώντας παρόμοια λογική με τις αναπαραστάσεις σε επίπεδο *utterance* εξάχθηκαν αναπαραστάσεις τμημάτων όπου η συνένωσή των επιμέρους αναπαραστάσεων συνιστά την αναπαράσταση του σήματος σε ολόκληρο το μήκος του. Μεταβάλλοντας την χρονική διάρκεια των τμημάτων από *frames* σε *segments* εξετάστηκαν διάφορα μοντέλα δείχνοντας την επίδραση της χρονικής διάρκειας των τμημάτων στην επίδοση των μοντέλων. Το επόμενο βήμα είναι η κατασκεύη αυθαίρετων αναπαραστάσεων από τον χώρο των εμπειρικών χαρακτηριστικών. Συγκρίθηκε ο παραδοσιακός αλγόριθμος *representation learning PCA* με τα δίκτυα *autoencoders* όπου φάνηκε η υπεροχή των δικτύων. Δεδομένου ότι κάποια από αυτά τα δίκτυα είναι *generative*, χρησιμοποιήθηκαν για την γέννηση συνθετικών δειγμάτων σε επίπεδο *frames* και *segments*. Τέλος εξετάστηκε η ικανότητα των *autoencoders* να μαθαίνουν από μόνοι τους χρήσιμες αναπαραστάσεις, αφαιρώντας τον περιορισμό στη γνώση που επέβαλλαν τα εμπειρικά χαρακτηριστικά.

Αναλυτικότερα ως *utterance - based* μοντέλο χρησιμοποιήθηκε ο *SVM* με διαφορετικές παραμέτρους συνάρτησης πυρήνα. Η χαμηλή του επίδοση όδηγησε στην υπόθεση του κατακερματισμού των σημάτων φωνής, δηλαδή στην χρήση *frame* και *segment - based LSTM* μοντέλα. Λόγω της μεγάλης εκφραστικότητας των *LSTM* μοντέλων ο αριθμός των παραμέτρων προς εκπαίδευση είναι αρκετά μεγάλος. Έτσι τα *LSTM* μοντέλα προεκπαιδεύτηκαν σε ένα πρόβλημα μέσω ενός συνόλου προεκπαίδευσης. Εξετάστηκαν διαφορετικά προβλήματα προεκπαίδευσης όπως η αναγνώριση συναισθήματος στο συνεχή χώρο, η ταυτόχρονη αναγνώριση συναισθηματος στο συνεχή χώρο και η αναγνώριση φύλου και η ταυτόχρονη αναγνώριση συναισθήματος στο συνεχή και διακριτό χώρο. Τα *frame - based LSTM* αποδείχθηκαν εξίσου αδύναμα με το *utterance - based* παρουσιάζοντας μικρή βελτίωση στην περίπτωση του *multi tasking emotion - gender*. Από την άλλη πλευρά υπήρξε σαφής βελτίωση στην περίπτωση των *segment - based* μοντέλων όπου όπως και στην περίπτωση των *frame - based* παρουσιάσαν την μέγιστη τιμή τους για το *multi tasking emotion - gender* πρόβλημα προεκπαίδευσης. Τα παραπάνω πειράματα έδειξαν πως η αναγνώριση συναισθήματος λειτουργεί καλύτερα στην περίπτωση όπου τα σήματα κατακερματίζονται σε *segments*. Σε αντίθεση με τα *frames*, η χρονική τους διάρκεια επιτρέπει στον ομιλητή να εκφράσει κάποια συναισθηματική κατάσταση. Από την άλλη πλεύρα, η αναγνώριση

σε επίπεδο *utterance* επισκιάζει εκείνα τα τμήματα που περιέχουν συναισθηματική πληροφορία. Για σήματα μεγάλου μήκους τα τμήματα που δεν περιέχουν κάποια συναισθηματική πληροφορία υπερτερούν αριθμητικά των συναισθηματικών τμημάτων. Έτσι η εφαρμογή καθολικών χαρακτηριστικών αποκρύπτει από τα μοντέλα τα τμήματα του σήματος που περίεχουν την σημαντικότερη πληροφορία.

Για την καλύτερη αναπαράσταση των σημάτων χρησιμοποιήθηκαν *representation learning* αλγόριθμοι. Εδώ τα *LSTM* μοντέλα των προηγούμενων πειραματων δέχθηκαν ως είσοδο σήματα φωνής κατακερματισμένα σε τμήματα και αναπαριστώμενα σε ενά αυθαίρετο και μικρότερης διαστασιμότητας χώρο. Συγκρίθηκε ο αλγόριθμος *PCA* με τον *VAN*, *VAE*, και *CVAE* σε *frame* και *segment - based* μοντέλα. Τα αποτελέσματα των πειραμάτων υποστηρίζουν το γεγονός ότι η *segment - based* αποτελεί την καλύτερη μέθοδο αναγνώρισης συναισθήματος. Επιπλέον οι γραμμικές απεικονίσεις του αλγορίθμου *PCA* υστερούν από τους *autoencoders*. Ειδικότερα τα *variational* μοντέλα *VAE* και *CVAE* σημείωσαν υποσχόμενες. Για τον λόγο αυτό εξετάσθηκε και η *generative* ικανότητα τους. Πιο συγκεκριμένα χρησιμοποιήθηκε ο *CVAE* για την γέννηση συνθετικών δειγμάτων με σκοπό την αύξηση του συνόλου δεδομένων εκπαίδευσης. Ωστόσο τα συνθετικά δείγματα έδειξαν πως δυσχεραίνουν την αναγνώριση σημειώνοντας χαμήλοτερη επίδοση από τους *representation learning* αλγορίθμους.

Τέλος δόθηκε περαιτέρω ελευθερία στους *representation learning* αλγορίθμους. Τα παραδοσιακά *handcrafted features* εκφράζουν την εμπειρική γνώση του ανθρώπου χωρίς ωστόσο να αποτελεί πάντα μια καλή αναπαράσταση των δεδομένων. Προτάθηκαν *seq2seq autoencoders* οι οποίοι σε επίπεδο *segment*. Οι ίδιοι δέχθηκαν ως είσοδο *spectograms* από *segments* με σκοπό την ανακατασκεύη τους. Η αναπαράσταση των *segments* σε μορφή *spectograms* είναι ένας βολικός τρόπος περιγραφής του σήματος σε χωροσυγχνοτικούς άξονες όπου απουσιάζουν τα *handcrafted features*. Έτσι τα μοντέλα έμαθαν από μόνα τους χρήσιμα χαρακτηριστικά για τα *segments* και στη συνέχεια τα χαρακτηριστικά αυτά χρησιμοποιήθηκαν για ταξινόμηση με τα μοντέλα των προηγούμενων πειραμάτων. Παρατηρήθηκε πως αυτή η μέθοδος δεν έχει κάτι να ζηλέψει από την παραδοσιακοί *handcrafted features representation learning*. Μάλιστα για κάποια προβλήματα προεκπαίδευσης, χρονική διάρκεια των *segments* και *autoencoders*, οι εξαγόμενες από τα *spectograms* αναπαραστάσεις σημείωσαν καλύτερες επιδόσεις από τις αναπαραστάσεις των *handcrafted features*.

Τα συμπεράσματα της εργασίας διαμορφώνονται γύρω από 3 άξονες. Αρχικά, όσο αναφορά τον τρόπο διαχείρισης των δεδομένων παρατηρείται πως τα *segment - based* μοντέλα υπερτερούν από τα *utterance - based* και *frame - based*. Το συναίσθημα εντοπίζεται σε μικρότερα τμήματα από ολόκληρο το σήμα φωνής, όχι τόσο μικρά όμως όσο τα *frames*. Σε επίπεδο *segment* ο ομιλητής προλαβαίνει να εκφράσει κάποια συναισθηματική κατάσταση η οποία γίνεται εμφανής στα μοντέλα. Δεν ισχύει το ίδιο στην περίπτωση των *utterance - based* όπου αφενός ο ομιλητής έχει όλο το χρόνο στη διάθεσή του, ωστόσο η συναισθηματική του κατάσταση σπάνια γίνεται εμφανής στα μοντέλα. Από την άλλη πλευρά σε επίπεδο *frame* τα μοντέλα μπορούν να διακρίνουν συναισθηματικές μεταβάσεις καλύτερα, ωστόσο ο ομιλητής δεν προλαβαίνει να εκφράσει κάποιο συναίσθημα οδηγώντας σε χαμήλη επίδοση των μοντέλων. Όσο αναφόρα το πρόβλημα προεκπαίδευσης παρατηρήθηκε πως οι καλύτερες επιδόσεις εντοπιστήκαν για *multi tasking* μοντέλα, κυρίως για την περίπτωση του *emotion - gender*. Η συναισθηματική αντίληψη του συναισθήματος διαφέρει από άτομο σε άτομο. Έτσι, κατά την επισήμανση των δεδομένων δεν υπάρχει κάποια καθολική και αντικειμενική σκοπία. Τα μοντέλα που προεκπαιδεύονται αποκλειστικά για αναγνώριση συναισθηματος στο συνεχή ή (και) στο διακριτό χώρο αποκτούν γνώση για ένα παρόμοιο πρόβλημα που ωστόσο έχει μοντελοποιηθεί με διαφορετικό τρόπο. Αντίθετα στην περίπτωση του *multi tasking emotion - gender* μοντέλου εισάγεται επιπλέον ο αντικειμενικός παράγοντας αναγνώρισης φύλου. Με άλλα λόγια η αναγνώριση φύλου προσφέρει επιπλέον πληροφορία στο μοντέλο σε σχέση με το *single task emotion* μοντέλο, ενώ είναι εξασφαλισμένο πως αυτή η πληροφορία θα αποδειχθεί χρήσιμη στο μετέπειτα πρόβλημα της αναγώρισης άγχους. Τέλος όσο αναφορά τους αλγορίθμους *representation learning*, παρατηρήθηκε ότι στις περισσότερες περι-

πτώσεις οι εξαγόμενες αναπαραστάσεις των *autoencoders* υπερτερούν από εκείνες του αλγορίθμου *PCA*. Μάλιστα τα *variational* μοντέλα σημείωσαν τις καλύτερες επιδόσεις ακολουθούμενες από τον κλασικό *autoencoder*. Τα *variational* έδειξαν πως είναι ικάνα για την εξαγωγή καλών αναπαραστάσεων είτε από *handcrafted features* είτε από *spectograms*. Δεν ισχύει το ίδιο ωστόσο για την ικανότητά τους να συνθέτουν δείγματα, καθώς δυσκολεύουν το πρόβλημα της αναγνώρισης.

6.2 Προεκτάσεις Εργασίας

Σε γενικές γραμμές ο κλάδος της αναγνώρισης συναισθήματος είναι ακόμα ανοιχτος καθώς διαρκώς εμφανίζονται διάφορες υποσχόμενες ιδέες. Εδώ αναφέρεται μια μελλοντικλες προέκτασεις της εργασίας.

Αρχικά, στην παρούσα εργασία εξετάστηκαν *single* και *multi - tasking* μοντέλα, για την αναγνώριση φύλου καθώς και συναισθήματος σε διακριτό και συνεχή χώρο. Σε γενικές γραμμές το *multi - tasking emotion - gender* μοντέλο φάνηκε να ανταποκρίνεται καλύτερα στη φύση του προβλήματος. Το επαγομένο βήμα είναι η κατασκευή ενός μοντέλου για αναγνώριση φύλου, διακριτού και συνεχούς συναισθήματος δηλαδή *emotion - gender - valence, activation, dominance*. Χρησιμοποιώντας τους ίδιους *autoencoders* μπορούν να συγκριθούν οι επιδόσεις αυτού του μοντέλου με τις επιδόσεις των μοντέλων που σημειώθηκαν προηγουμένως.

Σε επίπεδο *representation learning* αλγορίθμων, μολονότι ο *CVAE* σημείωσε αξιόλογες επιδόσεις, έγινε μια αναγκαία υπόθεση ως προς το διάνυσμα **c** που δεσμεύτηκαν τα δεδομένα εισόδου. Όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο συνήθως το διάνυσμα αυτό αντιστοιχεί στην κλάση στην οποία ανήκουν τα δείγματα. Στην εργασία τα δεδομένα εισόδου του *CVAE* αντιστοιχούν σε υποτμήματα των *utterances*, δηλαδή το διάνυσμα **c** των τμημάτων κληρονομήθηκε από την κλάση στην οποία ανήκει το αρχικό σήμα φωνής. Ωστόσο επείδη ο *CVAE* εκπαιδεύτηκε τόσο στην *IEMOCAP* όσο και σε *stressed* / *unstressed* υπάρχουν διαφωνείες μεταξύ των ετικετών των 2 συνόλων. Εδώ έγινε η υπόθεση πως τα σύνολα των κλάσεων των 2 συνόλων είναι ξένα μεταξύ τους. Μια διαφορετική και ενδεχομένως καλύτερη προσέγγιση είναι η εύρεση ενός μετασχηματισμού ο οποίος προβάλλει τα σύνολα των κλάσεων σε ένα ενιάιο αφηρημένο σύνολο συναισθημάτων. Έτσι όλα τα δεδομένα δεσμεύονται σε ένα διάνυσμα **c** το οποίο ωστόσο προέρχεται από τον ίδιο χώρο. Στη συνέχεια οι αναπαραστάσεις του *CVAE* μπορούν να χρησιμοποιηθούν με τον ίδιο τρόπο που χρησιμοποιήθηκαν και στην εργασία. Επιπλέον με την χρήση του μετασχηματισμού είναι εφικτή η μετάβαση από τον χώρο *stressed* / *unstressed* δειγμάτων και πιθανώς η γέννηση καλύτερων συνθετικών δειγμάτων.

Μια άλλη ενδιαφέρουσα προσέγγιση είναι η κατασκευή αναπαραστάσεων από οπτικοακουστικά δεδομένα. Σε αυτήν την περίπτωση τα παραπάνω μοντέλα μπορούν να χρησιμοποιηθούν με 2 τρόπους. Πρώτον οι αναπαραστάσεις μπορούν να κατασκευαστούν από μίξη των ακουστικών αλλά και των οπτικών χαρακτηριστικών. Έτσι ένα *time step* αποτελείται από την συνένωση των οπτικοακουστικών χαρακτηριστικών. Όπως και στην εργασία, με τους ίδιους *representation learning* αλγορίθμους η ακολουθία αναπαριστάται από την ένωση των επιμέρους *time step*, τα μοντέλα προεκπαιδεύονται στο *source domain* και *task* και χρησιμοποιούνται για την επίλυση του *target domain* και *task*. Από την άλλη πλευρά μπορούν να κατασκευαστούν αναπαραστάσεις για τα οπτικά και ακουστικά χαρακτηριστικά ξεχωριστά. Μια τέτοια προσέγγιση μπορεί δείξει την σύγκριση μεταξύ του είδους των δεδομένων χρησιμοποιώντας τον μηχανισμό *Attention*δίνοντας ένα συντελεστή σε κάθε αναπαράσταση ανάλογα με την χρησιμότητα της.

Βιβλιογραφία

[1] R. Plutchik, "Emotion: Theory, research, and experience: Vol. 1. theories of emotion," 1980.

[2] G. Valenza, P. Allegrini, A. Lanatà, and E. P. Scilingo, "Dominant lyapunov exponent and approximate entropy in heart rate variability during emotional visual elicitation," *Frontiers in neuroengineering*, vol. 5, p. 3, 2012.

[3] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[4] P. B. Denes and E. Pinson, *The speech chain*. Macmillan, 1993.

[5] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*. Pearson, 2011.

[6] https://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html.

[7] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC medical informatics and decision making*, vol. 10, no. 1, p. 16, 2010.

[8] C. Raffel and D. P. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," *arXiv preprint arXiv:1512.08756*, 2015.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2011.

[10] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[12] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and language—state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[13] T. C. Schneirla, "An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal." 1959.

[14] A. Mehrabian, "Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies," 1980.

[15] H. Selye, "Confusion and controversy in the stress field," *Journal of human stress*, vol. 1, no. 2, pp. 37–44, 1975.

[16] V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, and O. M. Mozos, "Stress detection using wearable physiological sensors," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 2015, pp. 526–532.

[17] R. Sioni and L. Chittaro, "Stress detection using physiological sensors," *Computer*, vol. 48, no. 10, pp. 26–33, 2015.

[18] J. H. Hansen and S. E. Bou-Ghazale, "Getting started with susas: A speech under simulated and actual stress database," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[19] J. L. Flanagan, *Speech analysis synthesis and perception*. Springer Science & Business Media, 2013, vol. 3.

[20] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[21] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2794–2797.

[22] C. Ishi, H. Ishiguro, and N. Hagita, "Using prosodic and voice quality features for paralinguistic information extraction," in *Proc. of Speech Prosody*. Citeseer, 2006, pp. 883–886.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Openear—introducing the munich open-source emotion and affect recognition toolkit," in *Affective computing and intelligent interaction and workshops, 2009. ACII 2009. 3rd international conference on*. IEEE, 2009, pp. 1–6.

[24] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Readings in speech recognition*. Elsevier, 1990, pp. 65–74.

[25] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[26] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[27] M. Sondhi, "New methods of pitch extraction," *IEEE Transactions on audio and electroacoustics*, vol. 16, no. 2, pp. 262–266, 1968.

[28] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[29] J. Kreiman and B. R. Gerratt, "Perception of aperiodicity in pathological voice," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2201–2211, 2005.

[30] R. O. Duda, P. E. Hart, D. G. Stork *et al.*, *Pattern classification*. Wiley New York, 1973, vol. 2.

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, p. 533, 1986.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neural networks and learning machines*. Pearson Upper Saddle River, NJ, USA:, 2009, vol. 3.

[34] U.-G. Krebel, "Pairwise classification and support vector machines," *Advances in Kernel Methods*, 1999.

[35] O. L. Mangasarian and D. R. Musicant, "Lagrangian support vector machines," *Journal of Machine Learning Research*, vol. 1, no. Mar, pp. 161–177, 2001.

[36] J. Weston, C. Watkins *et al.*, "Support vector machines for multi-class pattern recognition." in *Esann*, vol. 99, 1999, pp. 219–224.

[37] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[38] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization methods and software*, vol. 1, no. 1, pp. 23–34, 1992.

[39] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE transactions on electronic computers*, no. 3, pp. 326–334, 1965.

[40] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[41] D. O. Hebb, *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[42] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[43] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[44] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[45] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[47] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[48] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4945–4949.

[49] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[50] R. Girshick, "Fast r-cnn," *arXiv preprint arXiv:1504.08083*, 2015.

[51] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[52] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8599–8603.

[53] R. Caruna, "Multitask learning: A knowledge-based source of inductive bias," in *Machine Learning: Proceedings of the Tenth International Conference*, 1993, pp. 41–48.

[54] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[55] L. Duong, T. Cohn, S. Bird, and P. Cook, "Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, vol. 2, 2015, pp. 845–850.

[56] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[57] R. Bellman, *Adaptive control process: a guided tour*, 1961.

[58] http://yann.lecun.com/exdb/mnist/.

[59] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[60] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[61] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[62] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on*. IEEE, 2012, pp. 3642–3649.

[63] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the seventh international conference on Information and knowledge management*. ACM, 1998, pp. 148–155.

[64] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[65] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[66] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[67] G. Alain and Y. Bengio, "What regularized auto-encoders learn from the data-generating distribution," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3563–3593, 2014.

[68] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural networks*, vol. 2, no. 1, pp. 53–58, 1989.

[69] M. J. Wainwright, M. I. Jordan *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[70] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[71] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[72] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015, pp. 3483–3491.

[73] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[74] J. H. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," in *Speaker classification I*. Springer, 2007, pp. 108–137.

[75] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," 1988.

[76] J. H. Hansen, "Evaluation of acoustic correlates of speech under stress for robust speech recognition," in *Bioengineering Conference, 1989., Proceedings of the 1989 Fifteenth Annual Northeast*. IEEE, 1989, pp. 31–32.

[77] ——, "Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2. IEEE, 1993, pp. 95–98.

[78] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[79] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.

[80] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 125–134.

[81] Y. Shi, M. H. Nguyen, P. Blitz, B. French, S. Fisk, F. De la Torre, A. Smailagic, D. P. Siewiorek, M. al'Absi, E. Ertin *et al.*, "Personalized stress detection from physiological measurements," in *International symposium on quality of life technology*, 2010, pp. 28–29.

[82] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable eda device," *IEEE Transactions on information technology in biomedicine*, vol. 14, no. 2, pp. 410–417, 2010.

[83] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: in laboratory and real life," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 1185–1193.

[84] S. S. Rajagopalan, O. R. Murthy, R. Goecke, and A. Rozga, "Play with me—measuring a child's engagement in a social interaction," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8.

[85] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Transactions on Affective Computing*, 2016.

[86] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy, "Stress detection from speech and galvanic skin response signals," in *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*. IEEE, 2013, pp. 209–214.

[87] H. Gao, A. Yüce, and J.-P. Thiran, "Detecting emotional stress from facial expressions for driving safety," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5961–5965.

[88] Y. Nakashima, J. Kim, S. Flutura, A. Seiderer, and E. André, "Stress recognition in daily work," in *International Symposium on Pervasive Computing Paradigms for Mental Health*. Springer, 2015, pp. 23–33.

[89] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[90] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," *arXiv preprint arXiv:1701.08071*, 2017.

[91] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 698–704.

[92] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Signal and information processing association annual summit and conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016, pp. 1–4.

[93] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2227–2231.

[94] C.-W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition." in *INTERSPEECH*, 2016, pp. 1387–1391.

[95] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," *Proc. Interspeech 2017*, pp. 1243–1247, 2017.

[96] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion," *arXiv preprint arXiv:1712.08708*, 2017.

[97] R. Xia, J. Deng, B. Schuller, and Y. Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 990–994.

[98] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 511–516.

[99] P. Song, Y. Jin, L. Zhao, and M. Xin, "Speech emotion recognition using transfer learning," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 9, pp. 2530–2532, 2014.

[100] X. Zuo, L. Lin, and P. Fung, "A multilingual database of natural stress emotion," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 1174–1178.

[101] R. S. Lazarus, *Stress and emotion: A new synthesis*. Springer Publishing Company, 2006.

[102] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[103] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[104] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[105] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2741–2745.

[106] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[107] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 190–195.

[108] Y.-A. Chung, C.-C. Wu, C.-H. Shen, H.-Y. Lee, and L.-S. Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," *arXiv preprint arXiv:1603.00982*, 2016.

[109] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International conference on machine learning*, 2015, pp. 843–852.

[110] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[111] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[112] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, “Cvae-gan: fine-grained image generation through asymmetric training,” *arXiv preprint arXiv:1703.10155*, 2017.

[113] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. Hinton, “Binary coding of speech spectrograms using a deep auto-encoder,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[114] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, “Neural audio synthesis of musical notes with wavenet autoencoders,” *arXiv preprint arXiv:1704.01279*, 2017.

[115] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, “Sequence to sequence autoencoders for unsupervised representation learning from audio,” in *Proc. of the DCASE 2017 Workshop*, 2017.