ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΠΤΥΞΗ ΣΥΣΤΗΜΑΤΟΣ ΑΝΑΓΝΩΡΙΣΗΣ ΦΩΝΗΣ ΜΕ ΧΡΗΣΗ ΒΑΘΕΩΝ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΟΥΖΟΥΝΟΓΛΟΥ ΑΝΑΡΓΥΡΟΥ

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΠΤΥΞΗ ΣΥΣΤΗΜΑΤΟΣ ΑΝΑΓΝΩΡΙΣΗΣ ΦΩΝΗΣ ΜΕ ΧΡΗΣΗ ΒΑΘΕΩΝ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΟΥΖΟΥΝΟΓΛΟΥ ΑΝΑΡΓΥΡΟΥ

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 11 Ιουλίου 2023.

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
…….	……..	…….
Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Καθηγητής Ε.Μ.Π.	Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

Copyright © – All rights reserved. Με την επιφύλαξη παντός δικαιώματος.
Ουζούνογλου Ανάργυρος, 2023.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα.

Το περιεχόμενο αυτής της εργασίας δεν απηχεί απαραίτητα τις απόψεις του Τμήματος, του Επιβλέποντα, ή της επιτροπής που την ενέκρινε.

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

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

(Υπογραφή)

……...
Ουζούνογλου Ανάργυρος
3 Ιουλίου 2023

Περίληψη

Ένας από τους σημαντικότερους λόγους που ο άνθρωπος κατάφερε να επιβιώσει, να χτίσει πολύπλοκες κοινωνίες και εν τέλει να δαμάσει την φύση είναι η γλώσσα. Ο άνθρωπος από την φύση του κοινωνικό ων, έχει δημιουργήσει διαφόρους τρόπους επικοινωνίας ανά τους αιώνες, ο πιο σύνθετος όμως είναι αυτός της ομιλίας, μέσω της φωνής μπορεί και εκφράζει τα πολύπλοκα συναισθήματα του και τις ιδέες του. Από την κατασκευή των πρώτων υπολογιστών δημιουργήθηκε η ανάγκη για ανάπτυξη ενός τρόπου επικοινωνίας ανθρώπου μηχανής. Αρχικά αυτού του τύπου η επικοινωνία στηριζόταν περισσότερο στις “ανάγκες” της μηχανής, δηλαδή οι χρήστες έπρεπε να εκπαιδευτούν στην γλώσσα μηχανής. Με την πάροδο του χρόνου δημιουργήθηκαν τρόποι επικοινωνίας πιο κοντά σε αυτήν που χρησιμοποιούν οι άνθρωποι για να επικοινωνήσουν μεταξύ τους.

Η εξέλιξη των υπολογιστών τόσο σε επίπεδο λογισμικού όσο και σε επίπεδο υλικού, οδήγησε στην ανάπτυξη των τομέων της μηχανικής μάθησης και της επεξεργασίας φυσικής γλώσσας. Σήμερα, λόγο των παραπάνω, ο άνθρωπος μπορεί να επικοινωνήσει με την μηχανή χρησιμοποιώντας τον πιο εκφραστικό και συνηθισμένο από τον ίδιο τρόπο επικοινωνίας, την ομιλία. Τα τελευταία χρόνια η έρευνα στην ανάπτυξη συστημάτων αναγνώρισης ομιλίας είναι αξιοσημείωτη. Όμως, η εξέλιξη των γλωσσών και η ολοένα αυξανόμενη ανάγκη του ανθρώπου για επικοινωνία με την μηχανή (με σκοπό φυσικά την διευκόλυνση της ζωής του) δημιουργούν προκλήσεις.

Ένα σύστημα αναγνώρισης φωνής δέχεται ως είσοδο μία έκφραση δοσμένη ως ήχο και έχει στόχο την μετάφραση αυτής σε μορφή κειμένου. Η κατασκευή ενός τέτοιου συστήματος προϋποθέτει την ανάπτυξη δύο επιμέρους μοντέλων του ακουστικού που ανταποκρίνεται στο πώς ηχεί μία λέξη/έκφραση και του γλωσσικού που ανταποκρίνεται στο συντακτικό και στην γραμματική μίας γλώσσας, δηλαδή στην δομή, που επιτρέπει η γλώσσα, να έχει μία έκφραση. Για την εκπόνηση της παρούσας διπλωματικής εργασίας, μελετήθηκαν και θα παρουσιαστούν τόσο τεχνικές στατιστικής μάθησης όσο και μηχανικής μάθησης, και για τα δύο μοντέλα. Ενώ για την βελτίωση του συστήματος χρησιμοποιήθηκαν τεχνικές που εξάγουν τα χαρακτηριστικά του ομιλητή. Τέλος, χρησιμοποιήθηκε σύνολο δεδομένων από διαλόγους στην αγγλική γλώσσα σε πραγματικό περιβάλλον (δηλαδή, όχι σε χώρο κατάλληλο για ηχογραφήσεις).

Λέξεις Κλειδιά

Σύστημα αναγνώρισης φωνής, Ακουστικό μοντέλο, Γλωσσικό μοντέλο, Επεξεργασία φυσικής γλώσσας, Μηχανική μάθηση, Στατιστική μάθηση, Βαθιά νευρωνικά δίκτυα, Ομιλία.

Abstract

One of the most important reasons that human kind managed to survive, build complex societies and even tame the nature is language. Humans by their nature are social creatures, over the centuries they have created various ways of communication, the most complex is that of speech, through voice they could express their complex feelings and ideas. Since the construction of the first computers, there has been a need to develop a way of human-machine communication. Initially this type of communication was based more on the "needs" of the machine, users should be trained in machine language. Over time, have been developed ways of communication closer to those which people use to communicate with each other.

The evolution of computers at both software and hardware levels has led to the development of the fields of machine learning and natural language processing. Today, because of those fields, human kind could communicate with the machine using the most expressive and common way of communication, speech. In recent years, research into the development of speech recognition systems is remarkable. However, the evolution of languages and the ever increasing need for man to communicate with the machine (with the aim of course to make his life easier) create challenges.

A speech recognition system, gets as input an utterance as sound and aims to translate it into text. The construction of a system like this, requires the development of two separate models, the acoustic model which corresponds to how a word/phrase sounds and the language model which corresponds to the syntax and the grammar of a language, that is the structure of an utterance allowed by the language. For the preparation of this thesis, both statistical learning and machine learning techniques were studied and will be presented, for both models. In order to improve the system, speaker adaptive training techniques, were used. Finally, a dataset of English language dialogues in a real environment (not in a recording studio) was used.

Keywords:
Speech recognition system, Acoustic model, Language model, Natural language processing, Machine learning, Statistical learning, Deep neural networks, Speech.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα μου κ. Ανδρέα Σταφυλοπάτη για την ευκαιρία που μου έδωσε, να εκπονήσω την διπλωματική μου εργασία στον τομέα Τεχνολογίας Πληροφορικής και Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου δίνοντας μου την ελευθερία να επιλέξω ένα θέμα που το θεωρώ αρκετά ενδιαφέρον.

Επίσης θα ήθελα να πω ένα μεγάλο ευχαριστώ στον υποψήφιο διδάκτορα Γεώργιο Ιωάννου για την άψογη συνεργασία και την διαρκή διάθεση για παροχή βοήθειας, υπήρξε καταλυτής για την ολοκλήρωση της παρούσας διπλωματικής εργασίας.

Ακόμα θα ήθελα να ευχαριστήσω το Εθνικό Μετσόβιο Πολυτεχνείο και πιο συγκεκριμένα την Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, διότι πέρα από τις γνώσεις που μου πρόσφερε, μου δίδαξε ένα ιδιαίτερο και πολυδιάστατο τρόπο σκέψης που είμαι βέβαιος ότι θα με συντροφεύει για την υπόλοιπη και όχι μόνο επαγγελματική, ζωή μου.

Έπειτα, θα ήθελα να ευχαριστήσω τους φίλους μου για την ψυχολογική υποστήριξη και κυρίως τον φίλο και συνάδελφο Χρήστο Πέτρου διότι πέρα από την ψυχολογική υποστήριξη στάθηκε παρόν και επιστημονικά όταν υπήρξε ανάγκη.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου, εύχομαι όταν έρθει η στιγμή να βρεθώ στην θέση τους να πράξω αντίστοιχα και αντάξια.

Περιεχόμενα

Κατάλογος Διαγραμμάτων

Κατάλογος Πινάκων

1. Εισαγωγή

1.1 Ιστορική Αναδρομή

Ο άνθρωπος στα πρώτα του βήματα παρατηρεί την φύση γοητεύεται, και θεοποιεί το ανεξερεύνητο. Τους επόμενους αιώνες μελετά και αντιλαμβάνεται τους φυσικούς μηχανισμούς, ενσωματώνοντας τις παρατηρήσεις του στις κατασκευές του. Τα παραδείγματα αυτών των κατασκευών είναι πολλά, κάποια από τα πιο σημαντικά είναι το σόναρ, το αεροπλάνο και η μπαταρία.

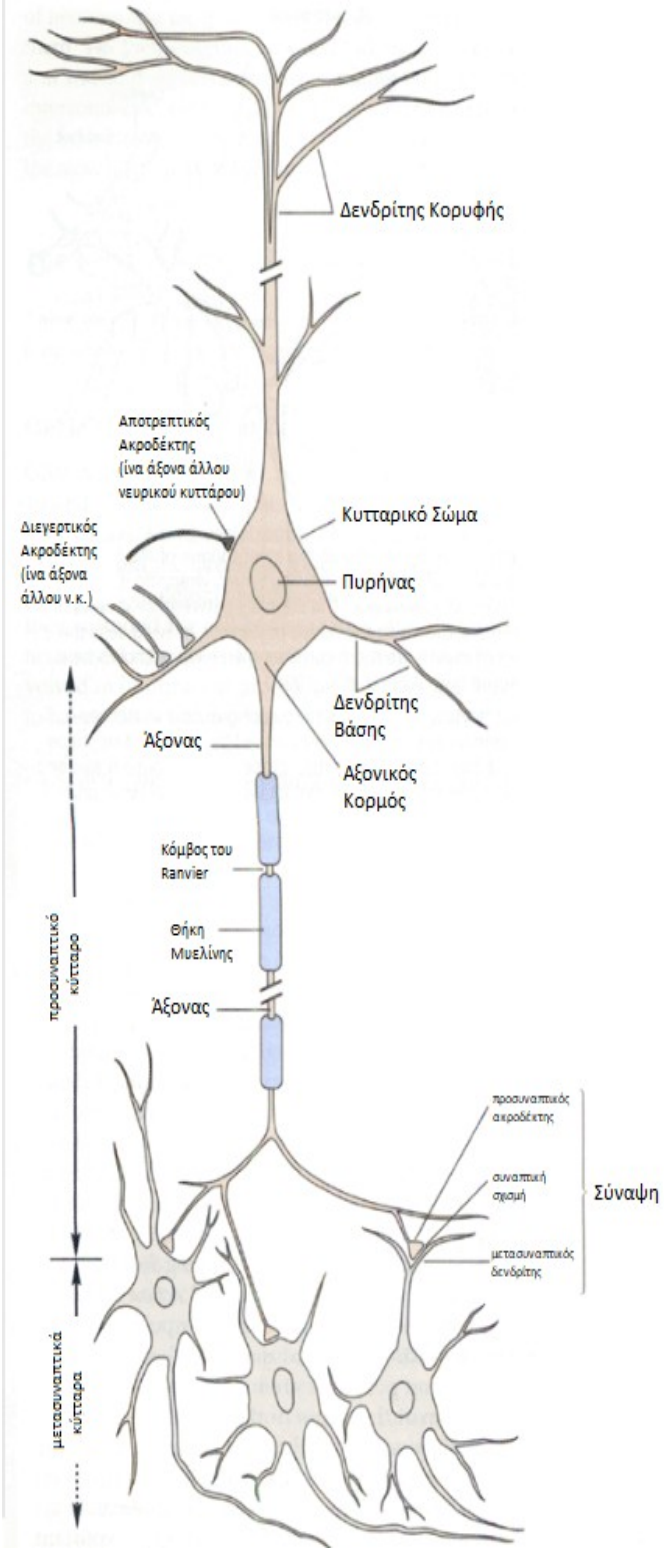
Στην αρχαιότητα ο Γαληνός προτείνει ότι τα νεύρα μεταφέρουν πληροφορία διαμέσου ενός υγρού από τον εγκέφαλο και την σπονδυλική στήλη στο υπόλοιπο σώμα. Στα τέλη του 18ου αιώνα ο Luigi Galvani παρατηρεί την ύπαρξη ηλεκτρισμού σε έμβια όντα. Τον επόμενο αιώνα Γερμανοί φυσιολόγοι παρατηρούν ότι η ηλεκτρική δραστηριότητα στα έμβια όντα είναι προβλέψιμη. Τον ίδιο αιώνα επιβεβαιώνεται η εικασία του Γαληνού καθώς παρατηρείτε στο μικροσκόπιο η δομή των νευρικών κυττάρων από τους Golgi και Gayal. Η παρατήρηση των νευρικών κυττάρων οδηγεί στην ολοένα και αυξανόμενη ενασχόληση των επιστημόνων με τον εγκέφαλο, στις αρχές του 20ου αιώνα παρατηρώντας τις χαρακτηριστικές δομές νευρικών κυττάρων και τις χαρακτηριστικές διατάξεις των κυτταρικών στοιβάδων ο Korbinian Brodmann χωρίζει τον εγκεφαλικό φλοιό σε 52 διακριτές περιοχές που η κάθε μια σχετίζεται και με μία λειτουργία του ανθρώπινου σώματος. Την δεκαετία του 1940 ο Donald Hebb εισάγει την θεωρία της πλαστικότητας των νεύρων δηλαδή της δυνατότητας του νευρικού συστήματος να προσαρμόζεται στο περιβάλλον του (συνάψεις που χρησιμοποιούνται περισσότερο δημιουργούν πιο ισχυρούς δεσμούς καθώς πεθαίνουν τα γηραιότερα κύτταρα και δημιουργούνται νέα). [1]

Το 1943 οι McCulloch και Pitts μοντελοποιούν μαθηματικά την λειτουργία των βιολογικών νευρικών δικτύων (θεώρηση τους ως υπολογιστικές μηχανές) [2]. Το 1950 ο A.M. Turing εισάγει στην επιστήμη των υπολογιστών την έννοια της ευφυίας και της μηχανικής μάθησης [3]. Τα επόμενα χρόνια ξεπηδάνε τα πρώτα προγράμματα τεχνητής νοημοσύνης όπως το πρόγραμμα του C. S. Strachey που έπαιζε Ντάμα [4] και το στηριζόμενο σε αυτό πρόγραμμα του A.L. Samuel που μάθαινε μέσω εκπαίδευσης [5].Το 1958, ένας ψυχολόγος, ο Rosenblatt εισάγει την έννοια του perceptron του πρώτου μοντέλου επιβλεπομένης μάθησης [6]. Το perceptron είναι η απλούστερη δυνατή μορφή νευρωνικού δικτύου αποτελείται από έναν μεμονωμένο νευρώνα και χρησιμοποιείται για γραμμικό διαχωρισμό προτύπων. Η επέκταση στις περισσότερες διαστάσεις είναι απλή όμως η έρευνα των Minsky και Papert το 1969 “παγώνει” την έρευνα στην τεχνητή νοημοσύνη καθώς οι perceptron (χωρίς κρυφά επίπεδα) αποτυγχάνουν στο πρόβλημα της αποκλειστικής διάζευξης (XOR). Το 1986 οι Rumelhart και McClelland αναπτύσσουν τον αλγόριθμο back propagation που αποτέλεσε ορόσημο στην εξέλιξη του κλάδου των τεχνιτών νευρωνικών δικτύων διότι παρείχε μια υπολογιστικώς αποτελεσματική μέθοδο για την εκπαίδευση perceptron πολλών επιπέδων (επιλύοντας και το πρόβλημα της XOR) [7]. Τα επόμενα χρόνια η πρόοδος στην επιστήμη των υπολογιστών τόσο στο τομέα του λογισμικού όσο και στον τομέα του υλικού καθώς και η ανάπτυξη του κυβερνοχώρου (internet) οδήγησε το 2006 στην ανάπτυξη πολυεπίπεδων δικτύων μεγάλου βάθους (βαθιά νευρωνικά δίκτυα) που έχουν την δυνατότητα να ανιχνεύσουν διάφορα χαρακτηριστικά για την ίδια είσοδο [8] (π.χ σε μία φωτογραφία ενός γραφείου εύρεση των διαφόρων αντικειμένων που υπάρχουν σε αυτό).

Τα τελευταία χρόνια η τεχνητή νοημοσύνη έχει μπει για τα καλά στις ζωές μας, ολοένα και περισσότερες εμπορικές εφαρμογές χρησιμοποιούν τεχνητή νοημοσύνη όπως οι ψηφιακοί βοηθοί Cortana και Siri, προγράμματα διαλόγου όπως το ChatGPT, προγράμματα δημιουργίας εικόνων όπως το Gencraft κ.α.

1.2. Βιολογικά Νευρωνικά Δίκτυα

Η βασική δομική μονάδα του εγκεφάλου είναι τα ανεξάρτητα νευρικά κύτταρα. Παρ' όλο που ο αριθμός των νευρώνων, στον άνθρωπο, είναι αρκετά μεγάλος (περίπου 10^{11}) και μπορούν να κατηγοριοποιηθούν σε τουλάχιστον χίλιους διαφορετικούς τύπους, η βασική δομή (Εικόνα 1.1) αυτών είναι η ίδια. Η περιπλοκότητα της ανθρώπινης συμπεριφοράς δεν εξαρτάται τόσο από τους διαφορετικούς τύπους (εξειδικεύσεις) των νευρώνων αλλά στις συνδέσεις μεταξύ αυτών (βιολογικά νευρωνικά δίκτυα). Τα βασικά χαρακτηριστικά του νευρικού συστήματος είναι :

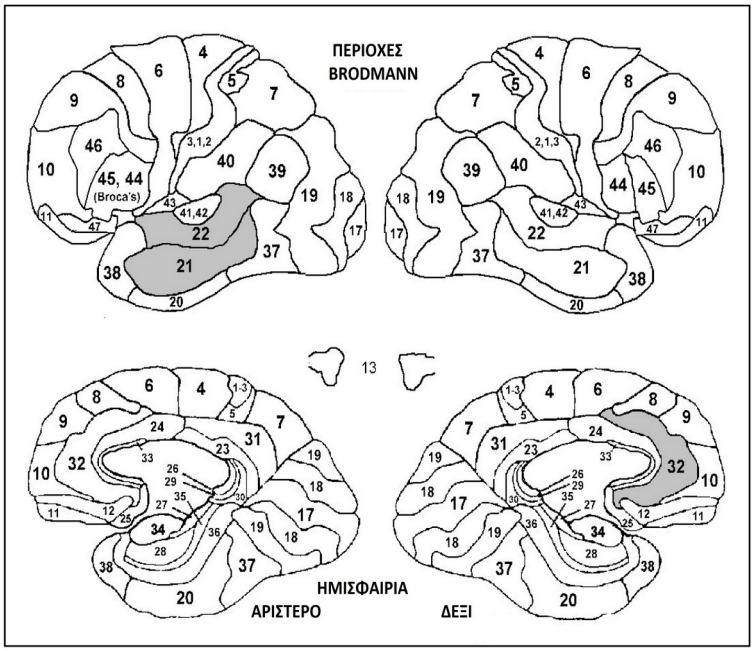


(Εικόνα 1.1):Βασική δομή νευρώνα. [1]

α) Οι μηχανισμοί που οι νευρώνες παράγουν σήματα.
β) Οι τρόποι με τους οποίους γίνεται η σύνδεση των νευρικών κυττάρων.
γ) Η σχέση μεταξύ των διαφορετικών τρόπων σύνδεσης και των διαφόρων συμπεριφορών του ανθρώπου.
δ) Ο τρόπος που οι νευρώνες και οι συνδέσεις τους μεταβάλλονται από την εμπειρία.
Η συστηματική μελέτη των παραπάνω χαρακτηριστικών αποτελεί το θέμα της επιστήμης της ιατρικής νευρολογίας.

Οι περισσότεροι νευρώνες ενός σπονδυλωτού νευρικού συστήματος (όπως το ανθρώπινο) έχουν αρκετά ίδια κύρια χαρακτηριστικά. Αυτοί αποτελούνται από το κυτταρικό σώμα, τον άξονα και τους δενδρίτες. Το κυτταρικό σώμα έχει διάμετρο τουλάχιστον 50μm και εμπεριέχει τον πυρήνα που είναι το σημείο αποθήκευσης της γενετικής πληροφορίας. Ο άξονας είναι πολύ πιο λεπτός σε σχέση με τον πυρήνα (διάμετρος από 0.2 έως 20μm) αλλά το μήκος του μπορεί να φτάσει τα 3m εντός του σώματος, συχνά περιβάλετε από θήκες μυελίνης μία λιποειδής ουσία που βοηθά στη διάδοση του δυναμικού κατά μήκος αυτού. Το δυναμικό ξεκινά είτε στον αξονικό κορμό, είτε στην περιοχή του άξονα κοντά σε αυτόν είτε στον πρώτο κόμβο του Ranvier ανάλογα την εξειδίκευση του νευρώνα. Ο άξονας καταλήγει στους δενδρίτες, άλλων νευρικών κυττάρων,ένας άξονας μπορεί να μεταφέρει πληροφορία σε έως και σε χίλιους νευρώνες. Οι δενδρίτες αποτελούν την είσοδο του νευρώνα, το σημείο ένωσης τους με τους άξονες άλλων νευρικών κυττάρων ονομάζεται σύναψη. Κάθε σύναψη αποτελείται από τρία στοιχεία τον προσυναπτικό ακροδέκτη, την συναπτική σχισμή και τον μετασυναπτικό δενδρίτη. Η συναπτική σχισμή ουσιαστικά είναι κενός χώρος,μέσω του οποίου μεταφέρονται οι νευροδιαβιβαστές δηλαδή οι χημικές ενώσεις που χρησιμοποιούνται για την μεταφορά του δυναμικού (πληροφορία) από τον προσυναπτικό ακροδέκτη στον μετασυναπτικό δενδρίτη. Τέλος κάποια από νευρικά κύτταρα

δημιουργούν συνάψεις με άλλα, για την ενεργοποίηση ή αποτροπή της λειτουργίας τους.



(Εικόνα 1.2): Οι διάφορες Περιοχές του εγκεφάλου χωρισμένες σε περιοχές ανάλογα την λειτουργία τους από τον Brodmann (ακουστικός φλοιός 41,42). [9]

Βιολογικά ακουστικά νευρωνικά δίκτυα

Τα τελευταία χρόνια οι επιστήμονες έχουν αρχίσει και βγάζουν κάποια συμπεράσματα σχετικά με το ανθρώπινο ακουστικό σύστημα. Έχει παρατηρηθεί ότι το όργανο Corti (όργανο υπεύθυνο για την μετατροπή του ακουστικού σήματος σε ηλεκτρικό σήμα) συνδέεται κατά μήκος του με αρκετά νεύρα, κάθε νεύρο είναι υπεύθυνο για την μεταφορά ενός σήματος με συγκεκριμένη συχνότητα στον ακουστικό φλοιό (Εικόνα 1.2). Τα νεύρα συγκεντρώνονται στο σπειροειδές γάγγλιο εντός του κοχλία και από εκεί η πληροφορία διανέμεται στις διάφορες περιοχές του εγκεφάλου που είναι υπεύθυνες για την ακουστική επεξεργασία. Οκτώ νευρικές ίνες διανέμουν την πληροφορία εντός του εγκεφάλου (τέσσερεις για κάθε αφτί). Το τί συμβαίνει στα κέντρα ακουστικής επεξεργασίας του εγκεφάλου μας είναι ακόμα άγνωστο. Έχουμε παρατηρήσει όμως τα νευρωνικά δίκτυα από το πρώτο κέντρο επεξεργασίας του εγκεφάλου, τον κοχλιακό πυρήνα. Εκεί υπάρχουν τέσσερεις τύποι νευρωνικών δικτύων. Από τις αποκρίσεις τους έχουμε σχηματίσει άποψη σχετικά με την λειτουργία τους. Τα δίκτυα έχουν πάρει το όνομα τους από το σχήμα τους και είναι τα εξής:

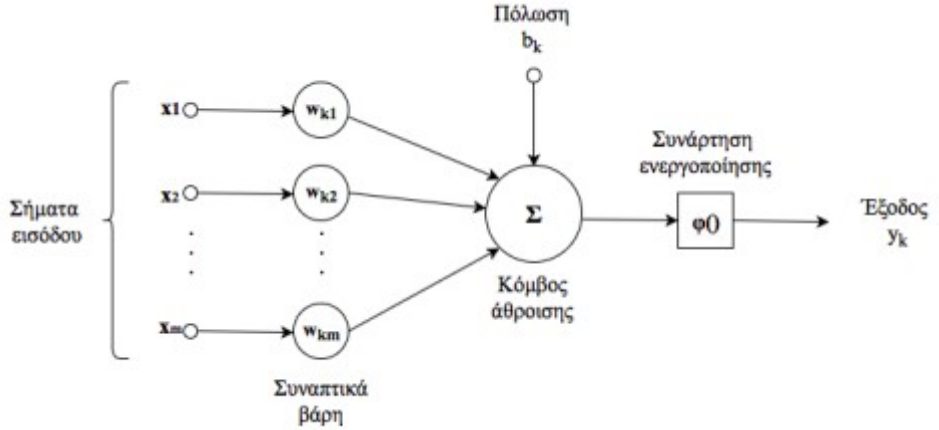
α) Αστρικό Δίκτυο: αποτελείται από συμμετρικούς δενδρίτες . Λειτουργεί ως φίλτρο απορρίπτοντας τις κρουστικές αποκρίσεις της εισόδου ενώ μπορούν και κωδικοποιούν (παρά τον θόρυβο) την συχνότητα του ήχου εισόδου.

β) Θαμνοειδή Δίκτυο: Ανταποκρίνονται σε συγκεκριμένη διαφορά δυναμικού στην αρχή του ηχητικού σήματος σχετίζονται πιθανότατα με την χρονομέτρηση του σήματος, πληροφορία που μας βοηθά στην τοποθέτηση της ακουστικής πηγής στον οριζόντιο άξονα.

γ) Δίκτυο Χταπόδι: Όμοια λειτουργία με το θαμνοειδή δίκτυο πιθανός βοηθά στην τοποθέτηση της ακουστικής πηγής στον κάθετο άξονα.

δ) Ασαφούς σχήματος Δίκτυο: πιθανών συμπιέζει το ακουστικό σήμα παράγοντας μία καθυστερημένη έξοδο. [1]

1.3 Μηχανική Μάθηση και Τεχνητά Νευρωνικά Δίκτυα



(Διάγραμμα 1.1): Βασικό μοντέλο τεχνητού νευρώνα. [7]

Όπως και στα βιολογικά έτσι και στα τεχνητά νευρωνικά δίκτυα ο νευρώνας είναι η στοιχειώδεις μονάδα επεξεργασίας. Στο διάγραμμα 1.1 φαίνεται το μοντέλο νευρώνα που αποτελεί βάση για την σχεδίαση των περισσότερων νευρωνικών δικτύων. Σε αυτό το διάγραμμα μπορούμε να διακρίνουμε τα τρία βασικά χαρακτηριστικά των τεχνιτών νευρώνων:

α) Ένα σύνολο συνάψεων, που η κάθε μία χαρακτηρίζεται από το δικό της βάρος. Δηλαδή το σήμα εισόδου x_j , της σύναψης j που συνδέεται με έναν νευρώνα k, θα πολλαπλασιαστεί με το συναπτικό βάρος w_{jk} . Σε αντίθεση με τα βιολογικά νευρωνικά δίκτυα, τα βάρη μπορούν να πάρουν θετικές και αρνητικές τιμές.

β) Τον αθροιστή, που χρησιμοποιείται για την άθροιση των σημάτων εισόδου σταθμισμένων από τα αντίστοιχα συναπτικά βάρη (στο συγκεκριμένο μοντέλο δρα ως γραμμικός συνδυαστής). Επίσης λαμβάνει μία επιπλέον είσοδο, εξωτερικά εφαρμοζόμενη, την πόλωση (b_k), η οποία έχει ως αποτέλεσμα την αύξηση ($b_k > 0$) ή την μείωση ($b_k < 0$) της δικτυακής διέγερσης της συνάρτησης ενεργοποίησης.

γ) Μία συνάρτηση ενεργοποίησης, που χρησιμοποιείται για τον περιορισμό του πλάτους του σήματος εξόδου. Η συγκεκριμένη συνάρτηση αναφέρεται συχνά και ως συνάρτηση περιορισμού επειδή ουσιαστικά περιορίζει το επιτρεπτό εύρος πλάτους του της εξόδου σε κάποια πεπερασμένη τιμή.

Από τα παραπάνω μπορούμε να ορίσουμε μαθηματικά την σχέση που διέπει την είσοδο και την έξοδο του παραπάνω νευρώνα: $y_k = \varphi\left(\sum_{j=i}^{m} [w_{jk} x_j] + b_k\right)$.

Συνδέοντας πολλούς νευρώνες μεταξύ τους μπορούμε και σχηματίζουμε ένα νευρωνικό δίκτυο που στην ουσία πρόκειται για έναν τεράστιο παράλληλο επεξεργαστή με κατανεμημένη αρχιτεκτονική, ο οποίος αποτελείται από απλές μονάδες επεξεργασίας (νευρώνες) και έχει από την φύση του την δυνατότητα να αποθηκεύει εμπειρική γνώση και να την καθιστά διαθέσιμη για χρήση. Το δίκτυο προσλαμβάνει γνώση από το περιβάλλον (μέσω μιας διαδικασίας μάθησης) και η αποθήκευση της γνώσης γίνεται στα συναπτικά βάρη (ισχύς συνδέσεων μεταξύ νευρώνων).

Ο τρόπος με τον οποίο δομούνται οι νευρώνες εντός του δικτύου σχετίζεται στενά με τον αλγόριθμο μάθησης που χρησιμοποιείται για την εκπαίδευση. Εστιάζοντας στην δομή μπορούμε να κάνουμε μία απλή κατηγοριοποίηση βάσει στην ύπαρξη ή όχι ανάδρασης:

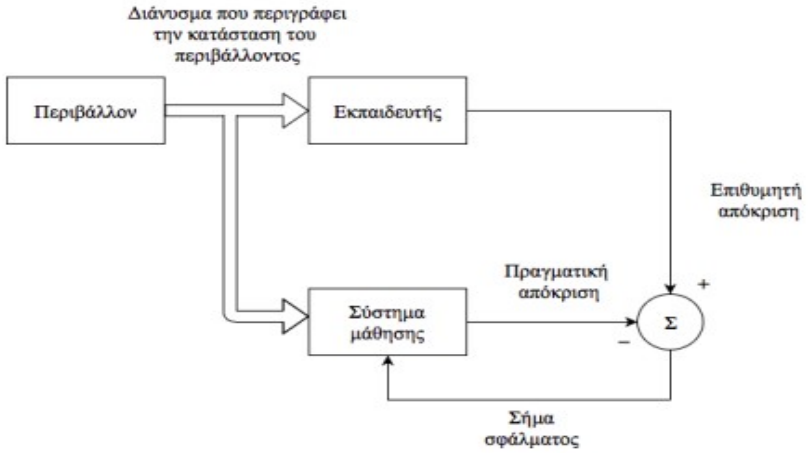
α) Δίκτυα πρόσθιας τροφοδότησης (feedforward): Η πορεία της πληροφορίας γίνεται από την είσοδο προς την έξοδο , χωρίς ανάδραση.

β) Αναδρομικά Δίκτυα (Recurent Neural Networks): Είναι δίκτυα με τουλάχιστον ένα βρόχο ανάδρασης δηλαδή κάθε νευρώνας εξόδου τροφοδοτεί την εξόδου του πίσω στις εισόδους των άλλων νευρώνων. Η παρουσία των βρόγχων ανάδρασης επιδρά σημαντικά στη δυνατότητα μάθησης του δικτύου και στην απόδοση του. Αυτοί προϋποθέτουν την χρήση συγκεκριμένων κλάδων, αποτελουμένων από στοιχεία μοναδιαίας χρονικής καθυστέρησης τα οποία έχουν ως αποτέλεσμα μη γραμμική δυναμική συμπεριφορά, εάν φυσικά το δίκτυο περιέχει μη γραμμικές μονάδες.

Όπως και στους ανθρώπους, έτσι και στα νευρωνικά δίκτυα υπάρχουν πολλοί τρόποι με τους οποίους αποκτάται η γνώση. Αυτοί οι τρόποι ονομάζονται διαδικασίες ή αλγόριθμοι μάθησης, που κατηγοριοποιούνται ως εξής :

α) Μάθηση με εκπαιδευτή (ή μη επιβλεπόμενη μάθηση). Ο εκπαιδευτής έχει γνώση του περιβάλλοντος, αυτή η γνώση αντιπροσωπεύεται από ένα σύνολο παραδειγμάτων εισόδου-εξόδου (σύνολο ετικετοποιημένων δεδομένων). Ωστόσο, το περιβάλλον είναι άγνωστο στο νευρωνικό δίκτυο. Χάρη στην εγγενή του γνώση, ο εκπαιδευτής είναι σε θέση να παρέχει στο δίκτυο μία επιθυμητή απόκριση για το κάθε διάνυσμα εκπαίδευσης (είσοδο). Η επιθυμητή απόκριση αντιπροσωπεύει τη βέλτιστη ενέργεια που πρέπει να εκτελείται από το νευρωνικό δίκτυο. Οι παράμετροι του δικτύου προσαρμόζονται υπό τη συνδυασμένη επιρροή του διανύσματος εκπαίδευσης και του σήματος σφάλματος. Το σήμα σφάλματος ορίζεται ως η διαφορά μεταξύ της επιθυμητής απόκρισης και της πραγματικής απόκρισης του δικτύου. Αυτή η προσαρμογή εκτελείται με επαναληπτικό τρόπο, βήμα προς βήμα, με στόχο να φέρει τελικά το νευρωνικό δίκτυο σε μία κατάσταση όπου θα προσομοιώνει τη συμπεριφορά του εκπαιδευτή.

Κατ’ αυτό τον τρόπο, η γνώση του περιβάλλοντος μεταφέρεται στο νευρωνικό δίκτυο και αποθηκεύεται με τη μορφή σταθερών συναπτικών βαρών, τα οποία αντιπροσωπεύουν την μακροπρόθεσμη μνήμη. Η διαδικασία επιβλεπομένης μάθησης συνιστά ένα σύστημα ανάδρασης κλειστού βρόχου, αλλά το άγνωστο περιβάλλον βρίσκεται εκτός του βρόχου (Διάγραμμα 1.2).



(Διάγραμμα 1.2): Σχηματική αναπαράσταση μάθησης με εκπαιδευτή. [7]

β) Μάθηση χωρίς εκπαιδευτή, που με την σειρά της χωρίζεται στις εξής περιπτώσεις:

1) Ενισχυτική Μάθηση. Η εκμάθηση μίας αντιστοίχισης εισόδου-εξόδου εκτελείται μέσω μίας συνεχούς αλληλεπίδρασης με το περιβάλλον με στόχο την ελαχιστοποίηση ενός βαθμωτού δείκτη απόδοσης. Στηρίζεται σε έναν μηχανισμό που δρα ως κριτής, ο οποίος μετατρέπει ένα κύριο σήμα ενίσχυσης λαμβανόμενο από το περιβάλλον σε ένα υψηλότερης ποιότητας σήμα που αποκαλείται ευρετικό σήμα ενίσχυσης (αμφότερα τα σήματα είναι βαθμωτές είσοδοι). Το σύστημα σχεδιάζεται ώστε να μαθαίνει βάσει καθυστερούμενης ενίσχυσης, δηλαδή παρατηρεί μία χρονική ακολουθία ερεθισμάτων που λαμβάνει από το περιβάλλον που καταλήγουν στην παραγωγή του ευρετικού σήματος ενίσχυσης. Στόχος της είναι η ελαχιστοποίηση μίας συνάρτησης τρέχοντος κόστους, η οποία ορίζεται ως πρόβλεψη του αθροιστικού ενεργειών που εκτελούνται σε μία σειρά βημάτων αντί απλώς του άμεσου κόστους μίας ενέργειας. Ενδεχομένως ορισμένες από τις ενέργειες που εκτελούνται σε αυτή την σειρά να είναι οι καλύτερες ορίζουσες της συνολικής συμπεριφοράς του συστήματος. Στην πράξη δεν χρησιμοποιείται καθώς η μηχανή πρέπει να είναι σε θέση να καθορίζει τον βαθμό επιτυχίας για κάθε ενέργεια στην αλληλουχία των βημάτων που οδήγησαν στο τελικό αποτέλεσμα, ενώ ο μηχανισμός ενίσχυσης αποτιμά μόνο το τελικό αποτέλεσμα.

2) Μη επιβλεπόμενη μάθηση (ή αυτο-οργανούμενη μάθηση): Όπως φανερώνει το όνομα της δεν υπάρχει εξωτερικός κριτής ή παρατηρητής που να επιβλέπει την διαδικασία μάθησης, αντ’ αυτού, υπάρχει ένα ανεξάρτητο από την εργασία μέτρο της ποιότητας της αναπαράστασης που καλείται να μάθει το δίκτυο και τα συναπτικά βάρη ρυθμίζονται βάσει αυτού. Στόχος του δικτύου ο συντονισμός στις στατιστικές κανονικότητες των δεδομένων εισόδου , αναπτύσσει τη δυνατότητα να σχηματίζει εσωτερικές αναπαραστάσεις για την κωδικοποίηση των χαρακτηριστικών της εισόδου και απ’ αυτές σχηματίζει νέες κλάσεις (αυτόματα). Σε ένα απλό δίκτυο μη επιβλεπομένης μάθησης, για παράδειγμα, δύο επιπέδων- ένα επίπεδο εισόδου και ένα ανταγωνισμού. Στο πρώτο επίπεδο εισέρχονται τα δεδομένα, ενώ οι νευρώνες του δευτέρου ανταγωνίζονται ο ένας τον άλλο για την “ευκαιρία” να αποκριθούν σε χαρακτηριστικά που περιέχονται στα δεδομένα εισόδου. [7]

<u>Βαθιά μάθηση (Deep Learning)</u>

Η Βαθιά μάθηση αποτελεί υποκατηγορία της μηχανικής μάθησης. Πρόκειται για μια αλληλουχία πολλαπλών επιπέδων από μη γραμμικές επεξεργαστικές μονάδες (νευρώνες) για την εξαγωγή και τον μετασχηματισμό χαρακτηριστικών. Κάθε επίπεδο δέχεται ως είσοδο την έξοδο του προηγούμενου επιπέδου. Η γνώση των πολλαπλών επιπέδων των αναπαραστάσεων ανταποκρίνεται

σε διαφορετικά επίπεδα αφαιρετικότητας. Τα επίπεδα σχηματίζουν μία ιεραρχία εννοιών. Τέλος, για την δημιουργία ενός δικτύου με βαθιά δομή μπορούν να χρησιμοποιηθούν τόσο τεχνικές επιβλεπομένης μάθησης όσο και μη. [10]

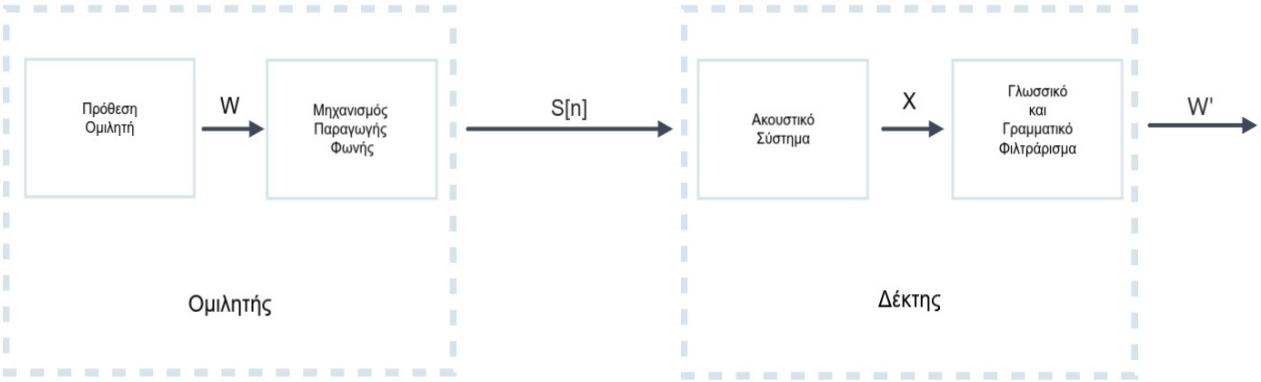
Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing – NLP)

Το αντικείμενο της παρούσας διπλωματικής εργασίας ανήκει στην ευρύτερη υποκατηγορία της μηχανικής μάθησης, επεξεργασία φυσικής γλώσσας. Με τον όρο αυτό εννοείται, η συστηματική μελέτη και έρευνα με στόχο την ανάπτυξη μηχανών που θα επιτελούν χρήσιμες εργασίες που συμπεριλαμβάνουν την ανθρώπινη γλώσσα (φωνή και γραφή). Τέτοιου τύπου εργασίες σηματοδοτούν την ανάπτυξη επικοινωνίας ανθρώπου μηχανής, βοηθούν στην επικοινωνία ανθρώπου με άνθρωπο και προσφέρουν την δυνατότητα ανάπτυξης νέων τεχνικών στην επεξεργασία κειμένου και φωνής. [11]

2.Περιγραφή του προβλήματος

2.1 Σύστημα αναγνώρισης φωνής

Για να υλοποιήσουμε ένα αυτόματο σύστημα αναγνώρισης φωνής θα πρέπει αρχικά να αναλογισθούμε πώς αναγνωρίζουμε εμείς ως άνθρωποι την φωνή. Έστω ότι έχουμε έναν ομιλητή ο οποίος έχει την πρόθεση να εκφράσει μία άποψη W ως μία πρόταση διαδοχικών λέξεων. Αφού σχηματίσει την πρόταση, ο εγκέφαλος του δίνει εντολή στο σύστημα παραγωγής φωνής που με την σειρά του μετατρέπει την σκέψη σε μορφή ακουστικού σήματος ήχου s[n]. Το σήμα φτάνει στο ακουστικό σύστημα του δέκτη εκεί το μετατρέπεται σε ακουστικά χαρακτηριστικά X. Τέλος τα χαρακτηριστικά φιλτράρονται από την γλωσσική εμπειρία/γνώση του δέκτη δημιουργώντας μια εκτίμηση της πρότασης του ομιλητή W', που αποτελεί και την αναγνωριζόμενη πρόταση (Διάγραμμα 2.1).



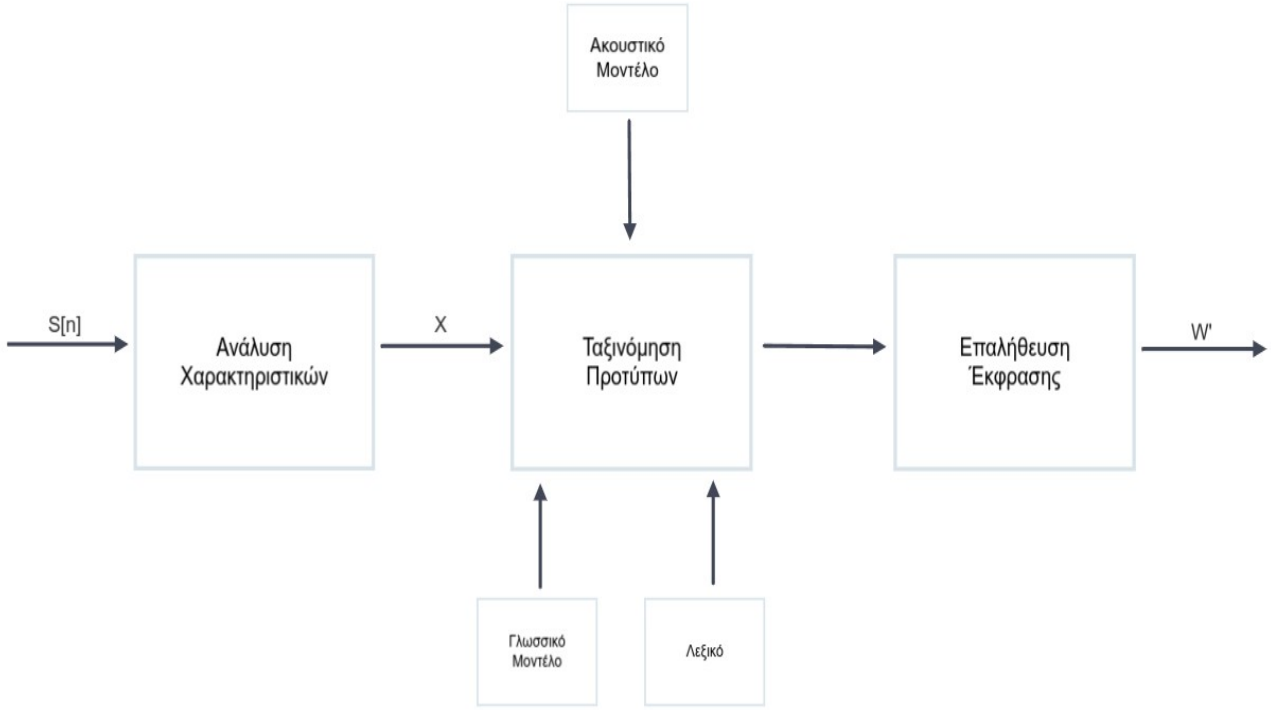
(Διάγραμμα2.1): Διαδικασία παραγωγής και αναγνώρισης φωνής.[12]

Από τα παραπάνω μπορούμε να χωρίσουμε σε στάδια το μοντέλο αναγνώρισης φωνής (Διάγραμμα 2.2). Τα στάδια αυτά είναι τα παρακάτω:

Ανάλυση χαρακτηριστικών (ανάλυση φάσματος): Δεχόμαστε ως είσοδο ένα ακουστικό σήμα φωνής,το πρώτο πράγμα που θα πρέπει να γίνει είναι να εξαχθεί η πληροφορία του σήματος και να εκφραστεί με τρόπο κατανοητό στο σύστημα επεξεργασίας, αναδεικνύοντας τα ακουστικά χαρακτηριστικά που περιέχει το σήμα και να απορριφθεί η πληροφορία που δυσχεραίνει την αναγνώριση (θόρυβος) .

Ταξινόμηση προτύπων (αποκωδικοποίηση και αναζήτηση): Έπειτα θα πρέπει να αποκωδικοποιήσουμε τα ακουστικά χαρακτηριστικά σε μια συμβολική απεικόνιση εφαρμόζοντας ακουστικούς και γραμματικούς περιορισμούς και βάσει των αποτελεσμάτων να αναζητήσουμε την πιθανότερη από όλες της πιθανές εκφράσεις που υπάρχουν στην μνήμη. Για να εφαρμόσουμε τους περιορισμούς θα χρειαστούμε, ένα λεξικό που εμπεριέχει την αντιστοίχηση των λέξεων με την προφορά τους, ένα ακουστικό μοντέλο που αντιστοιχεί τα χαρακτηριστικά με την συμβολική έκφραση της προφοράς της λέξης και ένα γραμματικό μοντέλο που εμπεριέχει κανόνες γραμματικής και συντακτικού.

Επαλήθευση έκφρασης: Τέλος θα πρέπει να μετρικοποιήσουμε την εμπιστοσύνη (μέγιστη πιθανοφάνεια) για την ύπαρξη της εκάστοτε λέξης στην πρόταση. Το στάδιο αυτό έχει ως στόχο την βελτίωση του προηγούμενου σταδίου πχ. προσθήκη νέων λέξεων στο λεξικό. [12]
Τα παραπάνω στάδια θα αναλυθούν διεξοδικά τόσο στη το θεωρία όσο και στην πράξη στα κεφάλαια 3 και 4.



(Διάγραμμα 2.2): Στάδια συστήματος αναγνώρισης φωνής.[12]

2.2 Παραγωγή φωνής και η συμβολική της έκφραση

Ένα σύστημα αναγνώρισης φωνής αποτελεί ουσιαστικά τον δέκτη μίας επικοινωνίας. Είναι σημαντικό πριν την ανάλυση του συστήματος να γίνει κατανοητή η απόκριση του πομπού,που αποτελεί την είσοδο σε ένα τέτοιο σύστημα, δηλαδή την φωνή. Η φωνή παράγεται από την διέγερση του ακουστικού σωλήνα, δηλαδή της φωνητικής οδού, η οποία τερματίζεται από τα χείλη στην μία πλευρά και στην άλλη, στην γλωττίδα. Υπάρχουν τρεις βασικές κατηγορίες ήχων φωνής:

α) Έμφωνοι ήχοι παράγονται από τη διέγερση της φωνητικής οδού με ημιπεριοδικούς παλμούς ροής αέρα που προκαλείται από το άνοιγμα και το κλείσιμο της γλωττίδας.

β) Τυρβώδεις ήχοι παράγονται σχηματίζοντας ένα στένωμα κάπου στην φωνητική οδό και εξαναγκάζοντας τον αέρα να περάσει μέσα από το στένωμα έτσι ώστε να δημιουργείται αναταραχή, παράγοντας μία διέγερση που μοιάζει με ήχο.

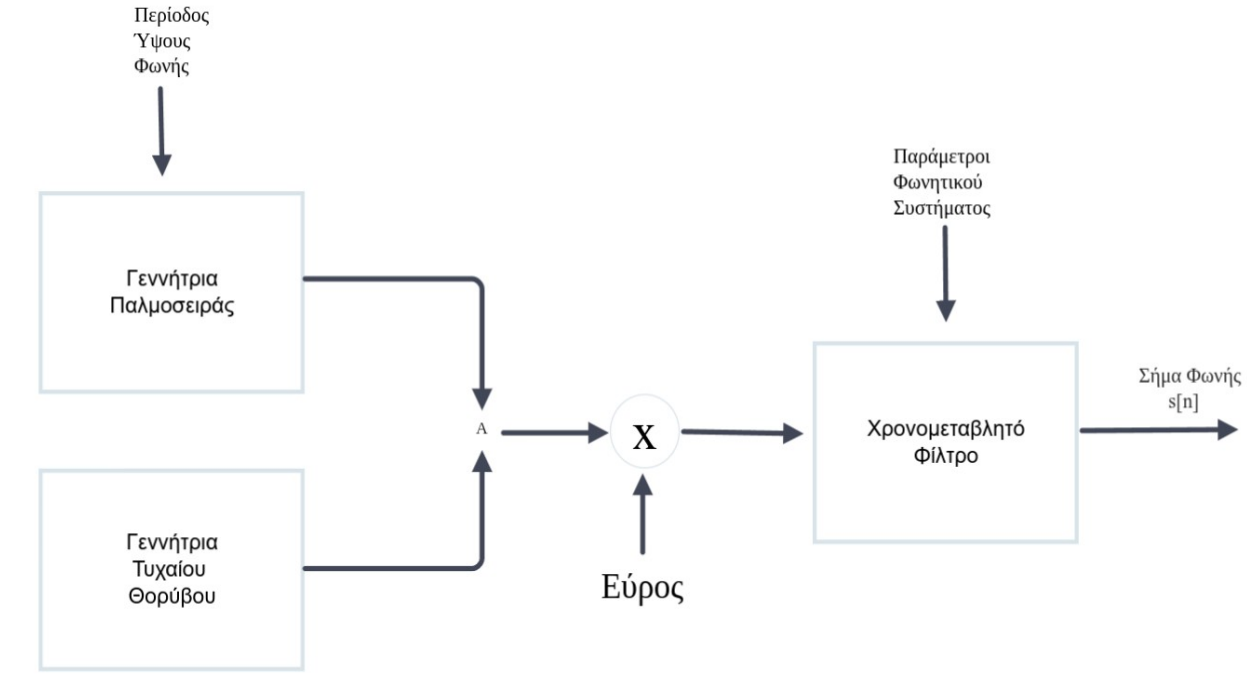
γ) Αποφρακτικοί ήχοι παράγονται από το ερμητικό κλείσιμο της φωνητικής οδού, δημιουργώντας πίεση πίσω από το κλείσιμο, και στη συνέχεια απελευθερώνοντας την πίεση.

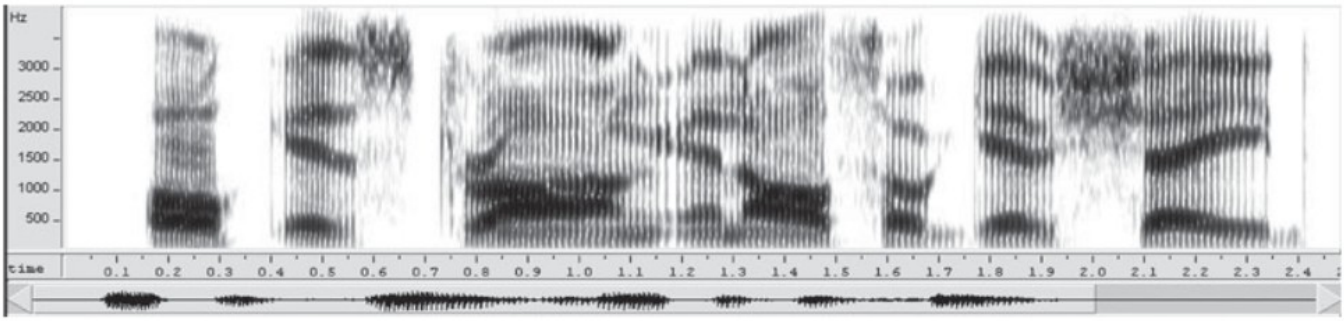
Για σταθερό σχήμα της φωνητικής οδού, η φωνή μπορεί να θεωρηθεί ως απόκριση ενός γραμμικά χρονικού ανεξάρτητου συστήματος, πιο συγκεκριμένα μία ημιπεριοδική παλμοσειρά για έμφωνους ήχους ή ευρυζωνικός θόρυβος για άφωνους . Έτσι λοιπόν η φωνητική οδός αποτελεί ένα ακουστικό σύστημα μετάδοσης που χαρακτηρίζεται από τις φυσικές του συχνότητες (ιδιοσυχνότητες) που αντιστοιχούν σε συντονισμούς στην απόκριση συχνότητας (φωνοσυντονισμοί), οι ιδιότητες της απόκρισης συχνότητας εξαρτώνται από το φάσμα διέγερσης. Στην κανονική ομιλία η φωνητική οδός αλλάζει σχήμα και λειτουργεί ως χρονομεταβλητό φίλτρο, γεγονός που σημαίνει ότι το σήμα φωνής δεν είναι στατικό. Όμως μπορεί να θεωρηθεί για διαστήματα 30-40 ms.

Στο διάγραμμα 2.3 φαίνεται το μοντέλο παραγωγής φωνής όπως το περιγράφηκε παραπάνω, το συγκεκριμένο μοντέλο αποτελεί το απλούστερο αλλά και το πιο ευρέως χρησιμοποιούμενο, το σημείο Α του Διαγράμματος συμβολίζει την επιλογή έμφωνου ή άφωνου ήχου, κατ’ επέκταση επιλογή γεννήτριας. [13]

Όπως είναι ευρέως γνωστό η γραμματική απεικόνιση των λέξεων δεν ταυτίζεται με την προφορά τους. Οι γλωσσολόγοι παρατήρησαν ότι υπάρχουν διακριτά και ελάχιστα φωνητικά στοιχεία που απαρτίζουν την κάθε λέξη, τα φωνήματα. Κάθε λέξη αποτελείται από μία σειρά

φωνημάτων και κάθε πρόταση από μια σειρά λέξεων. Κάθε φώνημα ανταποκρίνεται σε συγκεκριμένη διάρκεια και συχνότητα, τα μεγέθη αυτά εξαρτώνται από διάφορους παράγοντες. Ένας από τους πιο σημαντικούς παράγοντες εξάρτησης είναι αυτός που αποτελεί και το κριτήριο κατηγοριοποίησης των φωνημάτων από τους γλωσσολόγους δηλαδή ο τρόπος παραγωγής της φωνής. Πιο συγκεκριμένα η θέση ή η κίνηση της γλώσσας και των χειλιών και η χρήση ή μη της μύτης. Λόγω αυτού μπορούμε εύκολα να συμπεράνουμε ότι άλλος ένας παράγοντας εξάρτησης είναι η ανατομία του ομιλητή, κατά μέσο όρο παρατηρούμε αλλαγή της συχνότητας της φωνής σχετικά με το φίλο ή την ηλικία. Τέλος, σημαντικό ρόλο παίζει και η συναισθηματική κατάσταση του ομιλητή.[12]





(Διάγραμμα 2.3) πάνω: Απλό μοντέλο παραγωγής φωνής. [13]
(Εικόνα 2.1) κάτω: Ευρυζωνικό φασματογράφημα της έκφρασης “Oak is strong and also gives shade”. [12]

Ο αριθμός των διαφορετικών φωνημάτων διαφέρει ανάλογα την γλώσσα και την διάλεκτο. Στην νέα ελληνική γλώσσα, αν και υπάρχουν διαφωνίες μεταξύ των φιλολόγων,τα φωνήματα είναι 25 [14]. Στην αγγλική είναι 48, αν και τα 39 αρκούν για να περιγράψουν όλους τους πιθανούς ήχους (τα υπόλοιπα 9 μπορούν να εκφραστούν ως συνδυασμός)(Πίνακας 2.1.).Στο σημείο αυτό θα πρέπει να αναφερθεί ότι για την ανάπτυξη ενός συστήματος αναγνώρισης φωνής θα πρέπει να συμπεριληφθεί και το φώνημα της σιωπής ενώ μπορούν προαιρετικά (ανάλογα την φύση του προβλήματος) να χρησιμοποιηθούν και άλλα π.χ. φώνημα θορύβου,γέλιου κ.α.

Η συμβολική έκφραση της προφοράς αποτελεί ένα σημαντικό εργαλείο για ένα σύστημα αναγνώρισης φωνής,αφού αποτελεί τον συνδετικό κρίκο μεταξύ ήχου και κειμένου. Παίρνοντας μία φωνητική πρόταση εκφρασμένη στην συχνότητα μπορούμε να εντοπίσουμε τις μεταβολές και να

διαχωρίσουμε το σήμα στα ελάχιστα φωνητικά στοιχεία, έπειτα ετικετοποιώντας τα στοιχεία με ένα μοναδικό φώνημα παίρνουμε την πρόταση εκφρασμένη ως σειρά φωνημάτων, τέλος αν αντιστοιχίσουμε την σειρά αυτή με την γραμματική απεικόνιση των λέξεων έχουμε ως αποτέλεσμα την αρχική φωνητική πρόταση γραμμένη σε μορφή κειμένου.(Εικόνα 2.1) [12]

Φώνημα	Παράδειγμα	Φώνημα	Παράδειγμα
AA	Odd (AA D)	L	Lee (L IY)
AE	At (AE T)	M	Me (M IY)
AH	Hut (HH AH T)	N	Knee (N IY)
AO	Ought (AO T)	NG	Ping (P IH NG)
AW	Cow (K AW)	OW	Oat (OW T)
AY	Hide (HH AY D)	OY	Toy (T OY)
B	Be (B IY)	P	Pee (P IY)
CH	Cheese (CH IY Z)	R	Read (R IY D)
D	Dee (D IY)	S	Sea (S IY)
DH	Thee (DH IY)	SH	She (SH IY)
EH	Ed (EH D)	T	Tea (T IY)
ER	Hurt (HH ER T)	TH	Theta (TH EY T AH)
EY	Ate (EY T)	UH	Hood (HH UH D)
F	Fee (F IY)	UW	Two (T UW)
G	Green (G R IY N)	V	Vee (V IY)
HH	He (HH IY)	W	We (W IY)
IH	It (IH T)	Y	Yield (Y IY L D)
IY	EAT (IY T)	Z	Zee (Z IY)
JH	Gee (JH IY)	ZH	Seizure (S IY ZH ER)
K	Key (K IY)	SIL (silence)	“ “

(Πίνακας 2.1) : Φωνήματα της αγγλικής γλώσσας και παραδείγματα χρήσης.[15]

2.3 Στόχος

Από τα παραπάνω μπορούμε να ορίσουμε τον στόχο της παρούσας διπλωματικής εργασίας που δεν είναι άλλος από την ανάπτυξη ενός συστήματος αναγνώρισης φωνής και πιο συγκεκριμένα διαλόγου στην αγγλική γλώσσα, σε πραγματικό περιβάλλον. Το σύστημα θα είναι καθολικό δηλαδή δεν θα λαμβάνει υπόψη το φύλο ή την ηλικία των ομιλητών. Με τον όρο πραγματικό περιβάλλον εννοούμε ότι οι διάλογοι δεν θα λαμβάνουν χώρα σε κάποιο δωμάτιο ηχογράφησης αλλά σε μέρος που εγκυμονείτε θόρυβος.

Για την ανάπτυξη θα χρησιμοποιηθούν όλα τα σύγχρονα μέσα και τεχνικές καθώς θα υπάρξει σύγκριση μεταξύ στατιστικών μεθόδων και μεθόδων βαθιάς μηχανικής μάθησης.

3. Θεωρητικό Μέρος

Σε αυτό το κεφάλαιο θα αναλυθεί διεξοδικά η θεωρητική διάσταση όλων των βημάτων που χρειάζονται για την ανάπτυξη ενός συστήματος αναγνώρισης φωνής. Η δομή του κεφαλαίου ακολουθεί την πορεία της πληροφορίας εντός του συστήματος, από την ανίχνευση ενός φυσικού (αναλογικού) σήματος φωνής έως την έξοδο του συστήματος (πληροφορία εκφρασμένη σε κείμενο).

3.1 Ανάλυση Χαρακτηριστικών (εξαγωγή mel frequency cepstrum coefficients)

Όπως αναφέρθηκε στο κεφάλαιο 2.1 ένα αναλογικό σήμα φωνής S[t] δεν είναι “κατανοητό” στο σύστημα αναγνώρισης έτσι λοιπόν θα πρέπει να επεξεργαστεί αρχικά το σήμα, ώστε να εξαχθεί μία σειρά από διανύσματα χαρακτηριστικών X_T.

3.1.1 Μετατροπή αναλογικού σήματος σε ψηφιακό

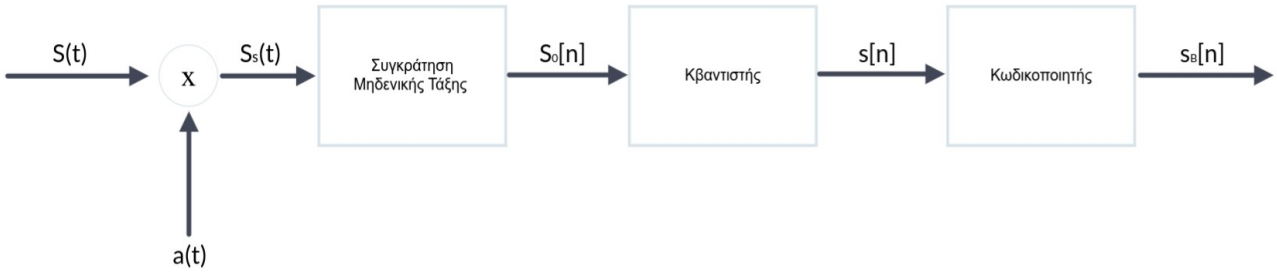
Δεχόμαστε λοιπόν ως είσοδο ένα σήμα συνεχούς χρόνου S[t], αρχικά θα πρέπει αυτό να μετατραπεί σε σήμα διακριτού χρόνου s[n] με σκοπό να γίνει εφικτή η ψηφιακή επεξεργασία του. Για να επιτευχθεί αυτό θα πρέπει να εφαρμοστούν στο σήμα δύο θεμελιώδεις τεχνικές, η δειγματοληψία και η κβαντοποίηση.

Μία χρονοδιακριτή αναπαράσταση ενός συνεχούς σήματος προκύπτει μέσω περιοδικής δειγματοληψίας,πολλαπλασιάζοντας το σήμα με μία περιοδική κρουστική παλμοσειρά της μορφής: $a(t)=\sum_{n\geq -\infty}^{\infty}\delta(t-nT)$, όπου T η περίοδος δειγματοληψίας (το αντίστροφο $f_a=1/T$ ονομάζεται συχνότητα δειγματοληψίας), n τα δείγματα και δ(t) η κρουστική συνάρτηση (δέλτα Dirac).

Άρα $S_s(t)=S(t)a(t)=S(t)\sum_{n\geq -\infty}^{\infty}\delta(t-nT)=\sum_{n\geq -\infty}^{\infty}[\delta(t-nT)S(t)]$ (3.1) εφαρμόζοντας τώρα την ιδιότητα της χρονοσυνεχούς κρουστικής συνάρτησης (ιδιότητα μετατόπισης) S(t)δ(t)=S(0)δ(t) η σχέση (3.1) εκφράζεται ως: $S_s(t)=\sum_{n\geq -\infty}^{\infty}S(nT)\delta(nT)$. Το μέγεθος (εμβαδό) της κρούσης στα χρονικά τμήματα nT είναι ίσο με την τιμή του σήματος εισόδου την ίδια χρονική στιγμή. Το σήμα που δίνεται ως έξοδο δεν είναι ίσο με s[n] αλλά είναι, κατά μία έννοια ένα χρονοσυνεχές σήμα (κρουστική παλμοσειρά) το οποίο είναι μηδενικό, εκτός από της θέσεις των ακέραιων πολλαπλάσιων της περιόδου T. Στο σημείο αυτό θα πρέπει να αναφερθεί ο θεμελιώδης κανόνας δειγματοληψίας ή θεώρημα Nyquist-Shannon δηλαδή για να μην υπάρχει επικάλυψη κατά την δειγματοληψία ενός σήματος θα πρέπει η συχνότητα δειγματοληψίας να είναι τουλάχιστον δύο φορές μεγαλύτερη από την συχνότητα του σήματος [13]. Για σήματα φωνής χρησιμοποιούνται ρυθμοί δειγματοληψίας από 8000 samples/sec έως 20000 samples/sec [12].

Αμέσως μετά την δειγματοληψία θα πρέπει το σήμα να περάσει από ένα σύστημα συγκράτησης μηδενικής τάξης (zero order hold), δηλαδή: $h_0(t)=\begin{cases}1, & 0<t<T\\ 0, & αλλού\end{cases}$ και άρα θα πάρει την μορφή $S_0(n)=h_0(t)*\sum_{n\geq -\infty}^{\infty}S(nT)\delta(nT)$. Αφού λοιπόν έχει συγκρατηθεί η δειγματική τιμή, το δείγμα S_0[n] μετασχηματίζεται σε πεπερασμένο σύνολο προκαθορισμένων τιμών μέσω ενός μη γραμμικού συστήματος του κβαντιστή, s[n]=Q(S_0[n]). Ο κβαντιστής είναι δυνατό να ορισθεί είτε ομοιόμορφα, είτε με ανομοιόμορφα απέχοντα μεταξύ τους επίπεδα κβαντισμού. Στον ομοιόμορφο κβαντιστή οι δειγματικές τιμές στρογγυλοποιούνται στο πλησιέστερο επίπεδο κβαντισμού. Τα ελάχιστα επίπεδα κβαντισμού, δηλαδή το ελάχιστο σημαντικό δυφίο (bit) του δυαδικού σήματος, προέρχονται από την σχέση $\Delta=\frac{X_m}{2^B}$, όπου B ο αριθμός των bit που χρησιμοποιούνται για την αναπαράσταση και X_m η παράμετρος που καθορίζει την πλήρη κλιμάκωση του μετατροπέα.

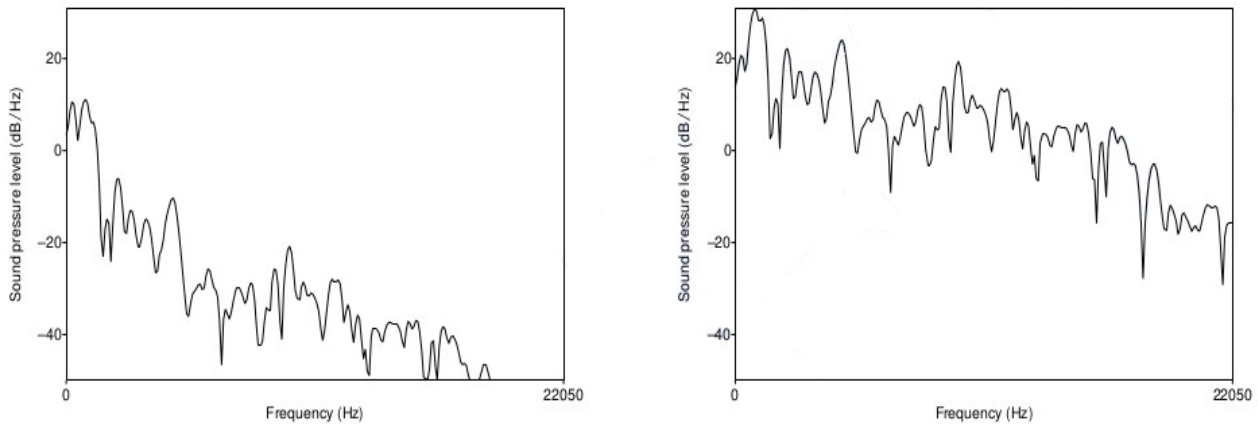
Υποθέτοντας τώρα ότι $s_B[n]$ δυαδικός αριθμός τέτοιος ώστε $-1 \le s_b[n] < 1$ (στο συμπλήρωμα ως προς 2), ισχύει $s[n]=X_m s_B[n]$. Τα δυαδικά κωδικοποιημένα δείγματα $s_B[n]$ είναι, λοιπόν, ευθέως ανάλογα των κβαντισμένων δειγμάτων s[n]. Συνεπώς είναι δυνατό να χρησιμοποιηθούν ως μία αριθμητική αναπαράσταση του εύρους των δειγμάτων. Γενικά είναι κατάλληλη η υπόθεση ότι το σήμα είναι κανονικοποιημένο και άρα τα s[n] και $s_B[n]$ είναι ταυτόσημα. [13]



(Διάγραμμα 3.1): Μοντέλο μετατροπής αναλογικού/ψηφιακού. [13]

3.1.2 Προ-έμφαση

Μετά την ψηφιοποίηση του σήματος θα πρέπει να ενισχυθεί η ενέργεια των υψηλών συχνοτήτων. Η διαδικασία αυτή είναι απαραίτητη διότι σε αρκετούς ήχους, όπως οι έμφωνοι, υπάρχει περισσότερη ενέργεια στις χαμηλές συχνότητες απ' ότι στις υψηλές. Έχει παρατηρηθεί ότι ενίσχυση αυτή, κάνει την πληροφορία των υψηλών ιδιοσυχνοτήτων (formants) του ακουστικού συστήματος πιο "διαθέσιμη" για το ακουστικό μοντέλο και βελτιώνει την ανίχνευση φωνημάτων. Η προ-έμφαση επιτυγχάνεται με την χρήση ενός υψιπερατού φίλτρου πρώτης τάξης [11] με συνάρτηση μεταφοράς: $h_1(z)=(1-\gamma z^{-1})$ [12], όπου $0.9 \le \gamma \le 1$. Έτσι λοιπόν για είσοδο s[n] το σήμα θα πάρει την μορφή $s_p[n]=s[n]-\gamma s[n-1]$ [11].



(Εικόνα 3.1) : Φασματική απεικόνιση του φωνήματος [aa] πριν την προ-έμφαση (αριστερά) και μετά (δεξιά) .[11]

3.1.3 Παραθυροποίηση

Αναφέρθηκε στο κεφάλαιο 2.2 ότι το σήμα φωνής δεν είναι στατικό σήμα, δηλαδή τα στατιστικά του χαρακτηριστικά δεν είναι σταθερά κατά την διάρκεια του χρόνου, αλλά μπορεί να θεωρηθεί στατικό για ένα μικρό χρονικό διάστημα, μερικών ms. Λόγω αυτού, για να εξαχθούν τα φασματικά χαρακτηριστικά χρησιμοποιείται ένα παράθυρο. Το παράθυρο είναι μη μηδενικό για ένα χρονικό διάστημα και μηδενικό εκτός αυτού και "κινείται" δια μήκος (άξονας χρόνου) του σήματος εξάγοντας τις στατικές κυματομορφές (frames) εντός του. Αυτό χαρακτηρίζεται από τρία βασικά χαρακτηριστικά το εύρος δηλαδή το χρονικό διάστημα που είναι μη μηδενικό, το σχήμα που αποτελεί την τιμή του στο μη μηδενικό χρονικό διάστημα και την επικάλυψη των παραθύρων.

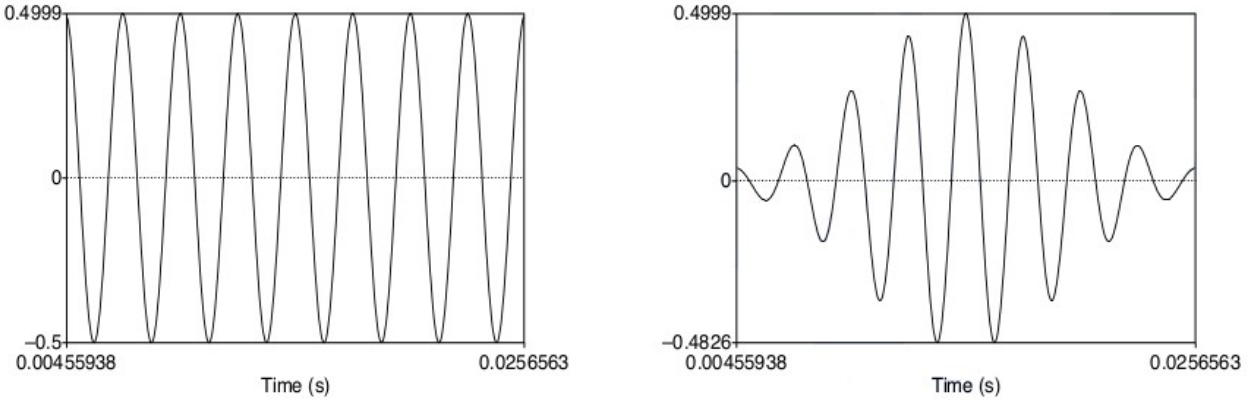
Το κάθε frame διέπεται από την σχέση $s'_p[n]=s_p[n]w[n]$, όπου w[n] η τιμή του παραθύρου. Υπάρχουν διάφοροι τύποι παραθύρων που μπορούν να χρησιμοποιηθούν για την διαδικασία αυτή το πιο απλό είναι το τετραγωνικό παράθυρο με εξίσωση :

$$w_R(n)=\begin{cases}1, & 0<n<L-1\\ 0, & \alpha\lambda\lambda o\acute{u}\end{cases}$$, όπου L το μέγεθος του frame.

Το τετραγωνικό παράθυρο δεν χρησιμοποιείται συχνά καθώς δημιουργεί προβλήματα στην ανάλυση φάσματος. Για σήματα φωνής συνήθως χρησιμοποιείται παράθυρο Hamming που διέπεται από την εξίσωση:

$$w_H(n)=\begin{cases}0.54-0.46\cos\left(\frac{2\pi n}{L}\right), & 0<n<L-1\\ 0, & \alpha\lambda\lambda o\acute{u}\end{cases}$$.

Ο λόγος που είναι πιο ευρέως χρησιμοποιούμενο για τέτοιες εφαρμογές είναι ότι συρρικνώνει τις τιμές στα άκρα του, αποφεύγοντας τις ασυνέχειες [11].



(Εικόνα 3.2): Παραθυροποίηση σήματος ημίτονου με χρήση (αριστερά) ορθογώνιου παραθύρου και (δεξιά) παραθύρου Hamming. [11]

3.1.4 Φασματική ανάλυση

Το επόμενο βήμα είναι η εξαγωγή της φασματικής πληροφορίας από το παραθυροποιημένο σήμα, κατά αυτών των τρόπο θα παραχθεί η γνώση για το πόση ενέργεια εμπεριέχει το σήμα στις διάφορες συχνότητες. Ως φάσμα ορίζεται η αντιστοίχηση των συχνοτήτων του σήματος με τα πλάτη (τιμή σήματος) τους. Για να παρθεί η φασματική πληροφορία, από ένα σήμα όπως το $s'_p[n]$, θα πρέπει να εφαρμοσθεί σε αυτό ένας διακριτού χρόνου μετασχηματισμός Fourier (Discrete Fourier Transform – DFT) [11]. Ο DFT ορίζεται ως :

$$S[k]=\sum_{n=0}^{N-1} s'_p[n]e^{-j(2\pi/N)kn}$$, όπου N η περίοδος του σήματος εισόδου.

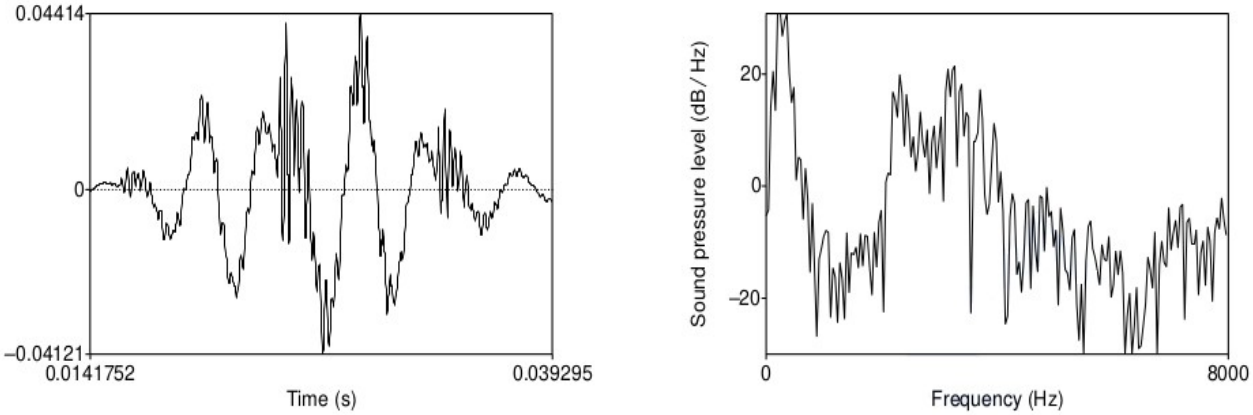
Φυσικά υπάρχει και ο αντίστροφος DFT (Inverse DFT - IDFT) που εφαρμόζεται σε σήμα εκφρασμένο στην συχνότητα και παρέχει την αντιστοίχιση του πλάτους με τον χρόνο, δίνεται από την σχέση: $s'_p[n]=\frac{1}{N}\sum_{n=0}^{N-1} S[k]e^{j(2\pi/N)kn}$ [13].

Ο πιο κοινός αλγόριθμος για τον υπολογισμό του DFT είναι ο γρήγορος μετασχηματισμός Fourier (Fast Fourier Transform – FFT) που υπολογίζει με μεγάλη ακρίβεια την τιμή του DFT αλλά μπορεί να εφαρμοστεί μόνο για τιμές του N πολλαπλάσιες του δύο [11].

3.1.5 Φίλτρα mel και λογάριθμος

Το ανθρώπινο αυτί δεν διακρίνει με την ίδια αποτελεσματικότητα τον ήχο σε όλες τις συχνότητες. Πιο συγκεκριμένα στις υψηλές συχνότητες είναι λιγότερο ευαίσθητο. Έχει παρατηρηθεί ότι βελτιώνεται η αποτελεσματικότητα ενός συστήματος αναγνώρισης φωνής αν

εφαρμοσθεί αυτό το χαρακτηριστικό του αυτιού στην εξαγωγή χαρακτηριστικών. Υπάρχουν διάφοροι τρόποι για να επιτευχθεί η αυτή η εφαρμογή, αυτοί εξαρτώνται από τα διάφορα



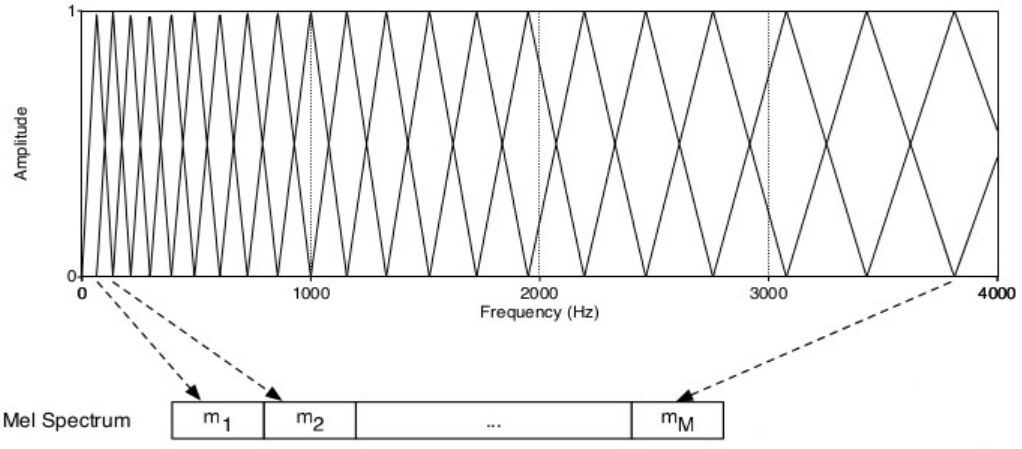
(Εικόνα 3.3): Παραθυροποιημένο με Hamming παράθυρο σήμα του φωνήματος [iy] πριν (αριστερά) και μετά (δεξιά) το DFT [11].

ψυχοακουστικά μοντέλα που υπάρχουν, το πιο συχνά χρησιμοποιούμενο μοντέλο είναι η κλίμακα mel. Τα ψυχοακουστικά αυτά μοντέλα σχετίζονται με την αντίληψη της τονικότητας, ως τονικότητα (pitch) ενός ήχου ορίζεται η διανοητική αίσθηση ή αντιληπτική συσχέτιση της θεμελιώδους συχνότητας του ήχου.

Ένα mel είναι μία μονάδα τονικότητας ορισμένη ώστε ήχοι με αντιληπτικά ισαπέχον τονικότητα να διαχωρίζονται με ίδιο αριθμό από mels και υπολογίζεται από τον παρακάτω τύπο:

$$mel(f) = 1127 \ln\left(1 + \frac{f}{700}\right) \quad .$$

Για να εφαρμοσθεί η κλίμακα αυτή σε σήμα S[k], θα πρέπει το σήμα να περάσει από μία συστοιχία φίλτρων (mel filter banks) ώστε να συλλεχθεί η ενέργεια από κάθε συχνότητα. Η συστοιχία αυτή αποτελείται από δέκα φίλτρα (τριγωνικού σχήματος) γραμμικά τοποθετημένα κάτω από τα 1000Hz και τα υπόλοιπα φίλτρα (επίσης τριγωνικού σχήματος) απλώνονται λογαριθμικά στις υπόλοιπες συχνότητες, χωρίς ο αριθμός τους να είναι συγκεκριμένος αλλά εξαρτάται από την εκάστοτε εφαρμογή. Το φάσμα που προκύπτει μετά το φιλτράρισμα ονομάζεται φάσμα mel.



(Εικόνα 3.4): Συστοιχία φίλτρων mel και εξαγωγή φάσματος mel. [11]

Ένα ακόμα χαρακτηριστικό του ανθρώπινου αυτιού που πρέπει να εισαχθεί είναι ότι έχει λιγότερη ευαισθησία στις μικρές διαφοροποιήσεις πλάτους για υψηλά απ' ότι για χαμηλά ηχητικά πλάτη. Για να εφαρμοστεί αυτό το χαρακτηριστικό θα πρέπει να εφαρμοσθεί λογάριθμος στο φάσμα mel. [11] Άρα λοιπόν από όλη την παραπάνω διαδικασία το σήμα S[k] έχει πάρει την μορφή: $S'[k]=\log(S[k]*H_{MEL}(k))$, όπου $H_{MEL}(k)$ η συνάρτηση μεταφοράς τις συστοιχίας φίλτρων mel [11].

3.1.6 Ανάφασμα

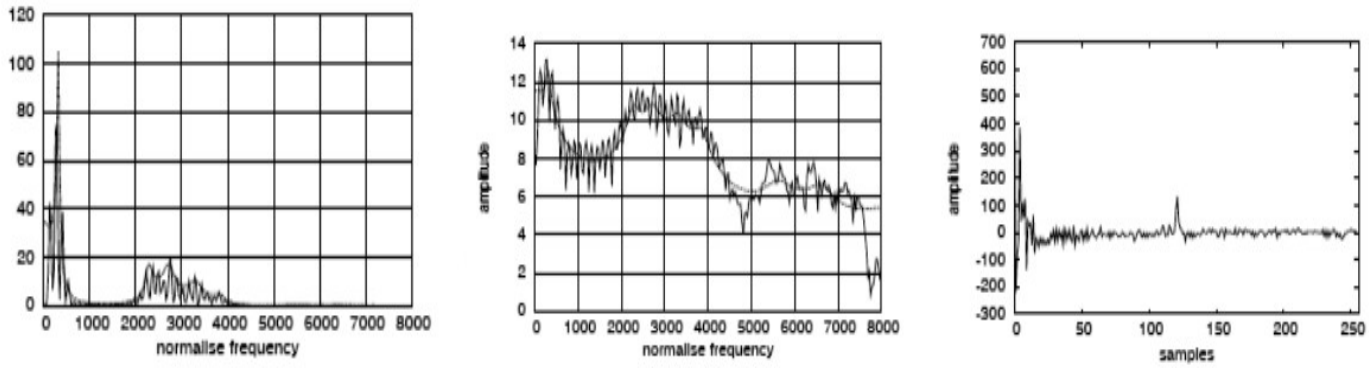
Το φάσμα mel θα μπορούσε να χρησιμοποιηθεί άμεσα για την αναγνώριση φωνημάτων, αν συμβεί κάτι τέτοιο όμως δεν θα παρθούν ικανοποιητικά αποτελέσματα. Ο λόγος που συμβαίνει αυτό έχει να κάνει την παραγωγή φωνής (Κεφάλαιο 2.2), όπως αναφέρθηκε ένα σήμα φωνής δημιουργείται όταν διεγείρεται η φωνητική οδός από μία θεμελιώδη συχνότητα καθώς ο αέρας ταξιδεύει εντός της οδού φιλτράρεται (ανάλογα το σχήμα της οδού) και σχηματίζεται ο τελικός ήχος. Άρα, αυτό που μπορεί να γίνει κατανοητό είναι ότι για την αναγνώριση του φωνήματος δεν έχει τόσο μεγάλη σημασία η θεμελιώδη συχνότητα αλλά το σχήμα της φωνητικής οδού δηλαδή η λειτουργία της ως φίλτρο. Ουσιαστικά θα πρέπει να γίνει μια αποσυνέληξη του σήματος για να διαχωριστεί η πηγή από το φίλτρο. Αυτό μπορεί να επιτευχθεί με την κατασκευή του αναφάσματος.

Ως ανάφασμα ορίζεται το φάσμα του λογαριθμικού φάσματος. Για την επεξήγηση αυτού του όρου θα χρειαστεί να παρατηρήσουμε την εικόνα 3.5 . Στην αριστερή γραφική παράσταση φαίνεται το φάσμα ενός φωνήματος, αν εφαρμοστεί λογάριθμος στο πλάτος του φάσματος θα παρθεί το γράφημα που απεικονίζεται στο κέντρο (λογαριθμικό φάσμα). Το μεσαίο γράφημα αν παρατηρηθεί χωρίς τις ετικέτες στους άξονες, μπορεί να θεωρηθεί ως σήμα φωνής εκφρασμένο στον χρόνο. Αυτό το ψευδοσήμα αξίζει να παρατηρηθεί επιπλέον καθώς εμπεριέχει πληροφορία που αξίζει να αναφερθεί. Υπάρχει μία συνιστώσα υψηλής συχνότητας που δημιουργεί επαναλήψεις (μικρά κύματα 8 ανά 1000 τιμές στα 120Hz) και αναπαριστάται από μικρές κορυφές (μία για κάθε αρμονική), αυτή σχετίζεται άμεσα με την θεμελιώδη συχνότητα του σήματος. Στο δεξιά γράφημα φαίνεται το ανάφασμα, δηλαδή το φάσμα του ψευδοσήματος, μετριέται σε δείγματα (samples) διότι πλέον το σήμα έχει φύγει από το πεδίο της συχνότητας και βρίσκεται στο πεδίο του χρόνου. Στο γράφημα του αναφάσματος φαίνεται μία κορυφή στα 120 samples που αναπαριστά την θεμελιώδη συχνότητα και διάφορες διακυμάνσεις στην αρχή του άξονα x που αναπαριστούν το φίλτρο. Ανάλογα την εφαρμογή μπορούν να χρησιμοποιηθεί είτε η θεμελιώδη συχνότητα (αναγνώριση τονικότητας) είτε το φίλτρο (αναγνώριση φωνής).

Από τα παραπάνω μπορεί να οριστεί μαθηματικά ο τύπος του αναφάσματος ως ο αντίστροφος DFT του λαγαριθμημένου φάσματος ενός σήματος, πιο συγκεκριμένα για ένα παραθυροποιημένο σήμα x[n] ισχύει:

$$c[n]=\sum_{n=0}^{N-1}\log\left(\sum_{n=0}^{N-1}x[n]e^{-j(2\pi/N)kn}\right)e^{j(2\pi/N)kn} \quad .$$

Από τον υπολογισμό του αναφάσματος μπορεί να εξαχθούν 12 συντελεστές αναφάσματος mel (mel frequency cepstral coefficients -MFCC) που αντιπροσωπεύουν την πληροφορία της φωνητικής οδού απαλλαγμένη από την τονικότητα. Οι συντελεστές αυτοί χρησιμοποιούνται συχνότερα για την αναγνώριση φωνής, αυτό οφείλεται στο γεγονός ότι έχουν ασύνδετη μεταξύ τους μέση τιμή και δεν χρειάζεται ο υπολογισμός της συνδιακύμανσης από το Gaussian Mixture Model (GMM) που θα δούμε στην συνέχεια, κάτι που σημαίνει δραματική μείωση των υπολογισμών [11].



(Εικόνα 3.5): (αριστερά) φάσμα παραθυροποιημένου φωνήματος, (μέση) λογάριθμος του φάσματος (ίδιου σήματος) (δεξιά) ανάφασμα (ίδιου σήματος) [11].

3.1.7 Παράγωγοι και Ενέργεια

Οι 12 συντελεστές που έχουν εξαχθεί παρ' ότι δίνουν σημαντική πληροφορία δεν αρκούν να περιγράψουν ένα σήμα φωνής καθώς όπως ειπώθηκε το σήμα φωνής δεν είναι στατικό άρα θα πρέπει να εισαχθούν συντελεστές που σχετίζονται με τις εναλλαγές του σήματος αλλά και πληροφορία σχετικά με την ενέργεια αυτού καθώς αυτή είναι διαφορετική για φωνήεντα και σύμφωνα. Ο δέκατος τρίτος συντελεστής είναι η ενέργεια που είναι το άθροισμα της δύναμης δειγμάτων (ενός παραθυροποιημένου σήματος x[t]) στο πέρασμα του χρόνου και δίνεται από τον τύπο:

$$Ενέργεια = \sum_{t=t_1}^{t_2} x^2[t] \quad .$$

Για να εισαχθεί η πληροφορία των εναλλαγών θα πρέπει να υπολογιστούν άλλοι 26 συντελεστές, 13 (12 συντελεστές mel και ένας συντελεστής ενέργειας) για την ταχύτητα (δέλτα) και άλλοι 13 (όμοιοι) για την επιτάχυνση (δέλτα τετράγωνο) του σήματος. Ο υπολογισμός τους είναι απλός και δίνεται από τον τύπο:

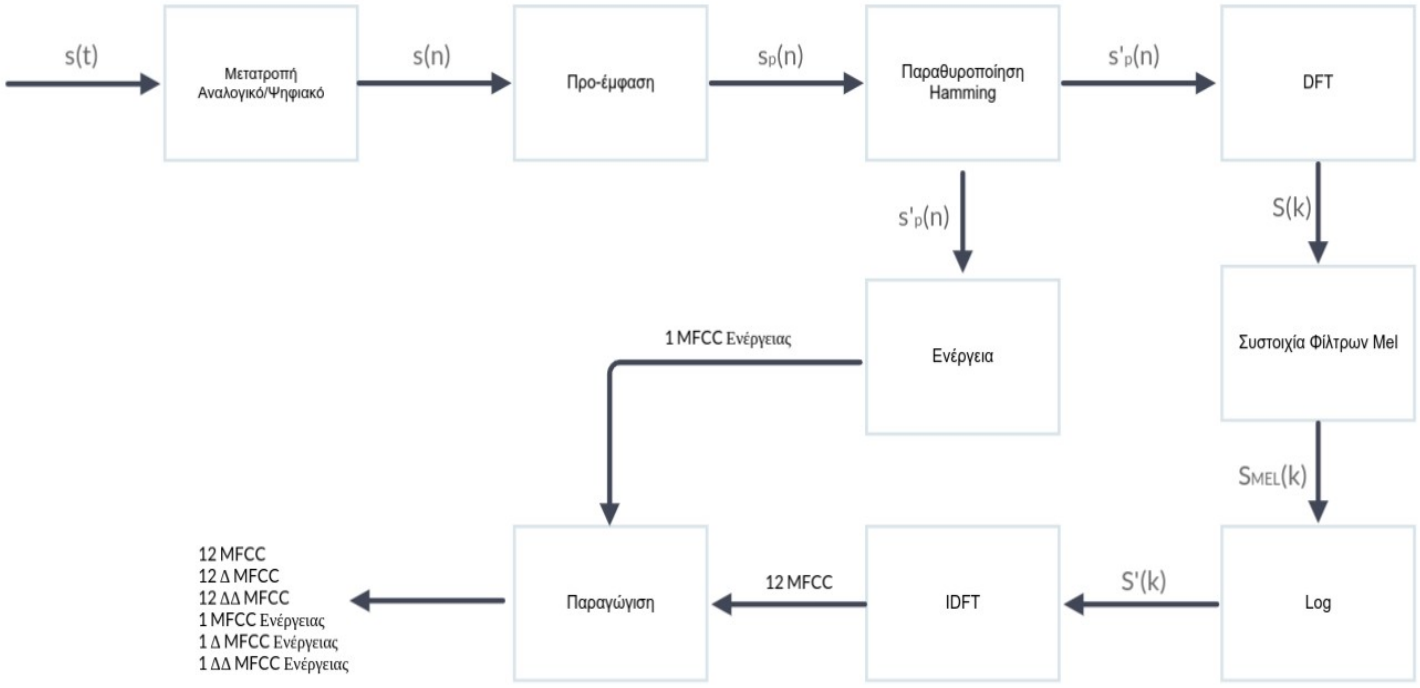
$$d(t) = \frac{c(t+1) - c(t-1)}{2} \quad .$$

Για τον υπολογισμό της ταχύτητας το d(t) συμβολίζει την ταχύτητα του frame το c(t+1) το ανάφασμα του επόμενου frame ενώ το c(t-1) το ανάφασμα του προηγούμενου. Όμοια για την επιτάχυνση με την διαφορά ότι αντί των αναφασμάτων χρησιμοποιείται η ταχύτητα αυτών [11].

3.1 Ανάλυση Χαρακτηριστικών (συνέχεια)

Ακολουθώντας την παραπάνω διαδικασία εξάγονται 39 MFCC που αποτελούν το διάνυσμα των ακουστικών χαρακτηριστικών ενός συστήματος αναγνώρισης φωνής. Στο Διάγραμμα 3.2 φαίνεται το συνολικό μοντέλο εξαγωγής χαρακτηριστικών (MFCCs).

Η προ-επεξεργασία των κειμενικών δεδομένων δεν εμπεριέχει κάποιο ιδιαίτερο θεωρητικό υπόβαθρο καθώς όπως αναφέρθηκε και στο κεφάλαιο 2 το μόνο που χρειάζεται είναι η μετάφραση τους σε μία συμβολική γλώσσα προφοράς.



(Διάγραμμα 3.2): Συνολικό μοντέλο εξαγωγής διανυσματικών χαρακτηριστικών (MFCC).

3.2 Αναγνώριση Προτύπων

3.2.1 Μοντελοποίηση

Στόχος ενός συστήματος αναγνώρισης φωνής είναι η εύρεση της έκφρασης W' που αντιστοιχεί στην πληροφορία των δοσμένων διανυσματικών χαρακτηριστικών X. Η παραπάνω διαδικασία ακόμα και για ένα βιολογικό ακουστικό σύστημα δεν έχει βέβαιο αποτέλεσμα, στην πραγματικότητα ο άνθρωπος (ως δέκτης) πιθανολογεί για το τι ακριβώς είπε ο συνομιλητής του. Από τα παραπάνω συμπεραίνεται ότι το πρόβλημα θα πρέπει να αντιμετωπισθεί ως πρόβλημα στατιστικών αποφάσεων. Πιο συγκεκριμένα μπορεί να τυποποιηθεί σαν μια διαδικασία προόδου, με ζητούμενο την έκφραση W', που μεγιστοποιεί την a posteriori πιθανότητα μια έκφραση W να προέρχεται από τα X, δηλαδή:

$$W' = \arg\max_{W} P(W|X) \quad .$$

Αν χρησιμοποιηθεί ο κανόνας Bayes στην παραπάνω έκφραση ισχύει:

$$W' = \arg\max_{W} P(W|X) = \arg\max_{W} \frac{P(X|W)P(W)}{P(X)} \quad .$$

Σε αυτό το σημείο θα πρέπει να αναλυθεί τόσο η φυσική όσο και η μαθηματική σημασία των όρων της παραπάνω εξίσωσης.

1) Η a priori πιθανότητα P(W) πρόκειται για την πιθανότητα της σειράς των λέξεων της έκφρασης W, ουσιαστικά αντιπροσωπεύει αυτό που αναφέρθηκε στο κεφάλαιο 2 ως γλωσσική (γραμματική και συντακτικό) γνώση του δέκτη. Αυτή υλοποιείται μέσω ενός γλωσσικού μοντέλου (language model). Τέλος είναι φανερό ότι αν M ο αριθμός των λέξεων σε μία έκφραση τότε $W=\{W_1,W_2,\ldots,W_M\}$.
2) Η a posteriori πιθανότητα P(X|W), δηλαδή η πιθανότητα η έκφραση W να παράγει (ηχητικά) τα διανύσματα X. Αυτή σχετίζεται άμεσα με την ακουστική φύση μίας έκφρασης, υλοποιείται από ένα ακουστικό μοντέλο. Επίσης αν T ο αριθμός των frames μίας έκφρασης (με κάθε frame να αποτελείται από 39 MFCC) τότε $X=\{X_1,X_2,\ldots,X_T\}$.
3) Η a priori πιθανότητα P(X) δεν έχει κάποια ιδιαίτερη φυσική σημασία, είναι ανεξάρτητη της W και παραλείπεται. Άρα η εξίσωση της W' γίνεται:
$$W' = \arg\max_{W} P(X|W)P(W) \quad .$$
4) Ο όρος $\arg\max_{W}$ αντιπροσωπεύει την αναζήτηση της πιθανότερης έκφρασης, πρόκειται για τον υπολογισμό της μέγιστης πιθανοφάνειας μίας πρότασης σε όλες τις πιθανές εκφράσεις [12].

3.2.2 Γλωσσικό Μοντέλο (Language Model - LM)

Το μοντέλο αυτό σχετίζει την πιθανότητα μία έκφραση να είναι αληθείς βάσει το περιεχόμενό της. Κατά αυτόν τον τρόπο το σύστημα μπορεί να διαχωρίσει ομόηχες λέξεις (π.χ. τύχη, τοίχοι, τείχη) και να διορθώσει τυχόν ακουστικές αστοχίες. Όπως αναφέρθηκε στην προηγούμενη παράγραφο στόχος του μοντέλου αυτού είναι η εύρεση της a priori πιθανότητας P(W) μίας έκφρασης, αυτός επιτυγχάνεται με την χρήση μίας τεχνικής (στατιστικής ή μηχανικής) μάθησης που συγκρίνει όλες τις έγκυρες εκφράσεις σε μία γλώσσα και εξάγει τις πιθανότητες αυτών. Ως είσοδο το μοντέλο χρησιμοποιεί ένα μεγάλο σύνολο δεδομένων από εκφράσεις γραμμένες σε κείμενο, που μπορεί να προέρχονται από τα δεδομένα που χρησιμοποιούνται και για την ακουστική εκπαίδευση ή/και από άλλες πηγές π.χ. βιβλία, εφημερίδες κ.α. Από το σύνολο δεδομένων υπολογίζεται μια N-Gram γραμματική που στηρίζεται στην εξής υπόθεση ότι η πιθανότητα ύπαρξης μίας λέξης σε μία πρόταση εξαρτάται μόνο από τις προηγούμενες N-1 λέξεις της πρότασης. Αν εφαρμοσθεί η υπόθεση αυτή στην P(W) θα ισχύει:

$$P(W) = P(W_1, W_2, \ldots, W_M) = \prod_{n=1}^{M} P(W_n | W_{n-1}, W_{n-2}, \ldots, W_{n-N+1}) \quad . [12]$$

Η παραπάνω πιθανότητα για να υπολογιστεί θα χρειαστεί να υπολογισθούν οι σχετικές συχνότητες των σειρών, μετρώντας δηλαδή τις φορές που εμφανίζεται μία σειρά Ν λέξεων και διαιρώντας το αποτέλεσμα με τον αριθμό που εμφανίζεται η ίδια σειρά με Ν-1 λέξεις. Άρα

$$P(W)=\frac{C(W_{n-N+1}...W_{n-2}W_{n-1}W_n)}{C(W_{n-N+1}...W_{n-2}W_{n-1})}$$, όπου C ο αριθμός των σειρών.

Αυτή η διαδικασία ονομάζεται εκτίμηση μέγιστης πιθανοφάνειας (maximum likehood estimation – MLE).

Από τα παραπάνω είναι εύκολο να συμπεραθεί ένα σημαντικό πρόβλημα που σχετίζεται με το γεγονός ότι, οι προτάσεις υπάρχουν σε ένα πεπερασμένο σύνολο δεδομένων που όσο μεγάλο κι αν είναι δεν μπορεί να καλύψει όλες τις λέξεις και τις εκφράσεις μίας γλώσσας καθώς οι γλώσσες εξελίσσονται αφού εμπεριέχουν το στοιχείο της δημιουργικότητας. Αν μία λέξη η πρόταση δεν υπάρχει στο σύνολο τότε η πιθανότητα P(W) θα είναι 0 κάτι που επί της ουσίας μηδενίζει την αξιοπιστία του μοντέλου. Το ίδιο ισχύει και για προτάσεις που εμφανίζονται σπάνια στο σύνολο δεδομένων καθώς η χρήση της MLE θα τους δώσει πιθανότητα περίπου 0. Η λύση για τις λέξεις που δεν υπάρχουν στο σύνολο είναι απλή καθώς μπορεί να προστεθεί μία λέξη στο λεξιλόγιο που συμβολίζει όλες της λέξεις εκτός αυτού (out of vocabulary - <oov>). Για της εκφράσεις όμως που εμφανίζονται σπάνια έως και καθόλου θα πρέπει να χρησιμοποιηθούν τεχνικές που “βελτιώνουν” την MLE με τέτοιο τρόπο ώστε να αυξάνει την πιθανότητα αυτών των εκφράσεων και να μειώνει την πιθανότητα αυτών που εμφανίζονται αρκετά συχνά στο σύνολο, οι τεχνικές αυτές ονομάζονται τεχνικές ομαλοποίησης (smoothing) ή έκπτωσης (discounting). Αρκετές τεχνικές αυτού του τύπου στηρίζονται σε θεωρήσεις εκτίμησης πληθυσμών. Οι τεχνικές που χρησιμοποιήθηκαν στην εκπόνηση της παρούσας διπλωματικής εργασίας θα αναλυθούν παρακάτω.

3.2.2.1 Τεχνική έκπτωσης Good-Turing

Η ιδέα που στηρίζεται αυτή η τεχνική είναι η επανεκτίμηση της τιμής της μάζας πιθανότητας των N-grams εκφράσεων που δεν εμφανίζονται, μέσω των εκφράσεων που εμφανίζονται μόνο μία φορά (sigletons). Βασίζεται στον υπολογισμό των N_c δηλαδή στον αριθμό των N-grams που συμβαίνουν c φορές, πρόκειται ουσιαστικά για την συχνότητα της συχνότητας c. Άρα η τιμή N_0 αντιπροσωπεύει τον αριθμό των εκφράσεων που δεν εμφανίζονται στο σύνολο (εκπαίδευσης) αλλά εμφανίζονται στο σύνολο ανάπτυξης, η N_1 τον αριθμό που των εκφράσεων που εμφανίζονται μία φορά κ.ο.κ. Δηλαδή :

$$N_c=\sum_{x:\,count(x)=c} 1 \quad .$$

Η MLE δίδει c για κάθε N_c, η Good Turing αντικαθιστά την c με μία ομαλοποιημένη c’ ως συναρτήσει της N_{c+1} ως εξής :

$$c'=(c+1)\frac{N_{c+1}}{N_c} \quad .$$

Η εξίσωση θα εφαρμοστεί για όλα τα N_c με c>0 και για c=0 θα χρησιμοποιηθεί η εξίσωση πιθανότητας μηδενικής μάζας (missing mass) που φαίνεται παρακάτω:

$$P_{GT}=\frac{N_1}{N}$$, όπου N ο αριθμός όλων των εκφράσεων που υπάρχουν στο σύνολο.

Από τα παραπάνω επιλύεται το πρόβλημα των εκφράσεων που δεν εμφανίζονται καθόλου στο σύνολο αλλά προκύπτει ένα νέο πρόβλημα, Αφού η c’ εξαρτάται από το N_{c+1}, εγκυμονεί ο κίνδυνος κάποιο N_{c+1}= 0. Μία εύκολη λύση είναι αφότου υπολογισθούν τα N_c και πριν υπολογισθούν τα c’ να ομαλοποιηθούν τα N_c , αυτή η τεχνική ονομάζεται απλός (simple) Good-Turing. Η πιο απλή μορφή ομαλοποίησης είναι αντικατασταθεί το N_c από τον λογάριθμο

$$\log(N_c)=a+b\log(c)$$, όπου a,b σταθερές.

Παρ’ όλα αυτά στην πράξη εφαρμόζεται ένα κατώφλι k, για τον υπολογισμό της c’ για μεγάλο αριθμό του c, συνήθως k>5. Πιο συγκεκριμένα :

$$c'=c \quad \text{για} \quad c>k.$$

Ενώ για τιμές από 1<c<k ισχύει:

$$c' = \frac{(c+1)\frac{N_{c+1}}{N_c} - c\frac{(k+1)N_{k+1}}{N_1}}{1-\frac{(k+1)N_{k+1}}{N_1}} \quad .$$

3.2.2.2 Τεχνική ομαλοποίησης Kneiser-Nay

Πριν την ανάλυση της επόμενης τεχνικής θα πρέπει να επεξηγηθούν δύο σημαντικές τεχνικές που χρησιμοποιούνται για τον υπολογισμό της πιθανότητας P(W), την παρεμβολή (interpolation) και το πίσω βήμα (backoff).

Στην παρεμβολή για τον υπολογισμό της πιθανότητας μίας N-gram σειράς χρησιμοποιούνται όλα τα προηγούμενα grams και βάρη λ_i τέτοια ώστε:

$$\sum_i \lambda_i = 1$$

Ουσιαστικά κατ' αυτόν τον τρόπο ο κατασκευαστής επιλέγει την σημαντικότητα του κάθε gram. Υπάρχουν δύο τύποι παρεμβολής η γραμμική (linear interpolation) και η υποθετική (conditional). Για την επεξήγηση της παρεμβολής θα υποτεθεί μία 3-gram έκφραση, για τον υπολογισμό της πιθανότητας $P'(W_n|W_{n-1}W_{n-2})$, ισχύει:

$$P'(W_n|W_{n-1}W_{n-2}) = \lambda_1 P(W_n|W_{n-1}W_{n-2}) + \lambda_2 P(W_n|W_{n-1}) + \lambda_3 P(W_n)$$

τα βάρη στην γραμμική παρεμβολή τα ορίζει ο κατασκευαστής ενώ στην υποθετική ορίζονται βάσει του περιεχομένου ενός συνόλου ανάπτυξης το οποίο δεν χρησιμοποιείται για τον υπολογισμό των πιθανοτήτων αλλά άλλων συντελεστών όπως τα βάρη, σε αυτήν την περίπτωση χρησιμοποιούνται ως βάρη οι πιθανότητες τον N-grams στο σύνολο ανάπτυξης.

Η τεχνική backoff χρησιμοποιείται για τις περιπτώσεις που το μοντέλο καλείται να υπολογίσει την πιθανότητα μίας συγκεκριμένης N-gram σειράς που δεν εμφανίζεται στο σύνολο, αυτό επιτυγχάνεται χρησιμοποιώντας την πιθανότητα της (N-1)-gram σειράς . Για παράδειγμα έστω ένα 3-gram καλείται να υπολογίσει την $P(W_n|W_{n-1}W_{n-2})$ αλλά δεν υπάρχουν στοιχεία για $W_n W_{n-1} W_{n-2}$ με την χρήση της backoff μπορεί να υπολογισθεί η P μέσω της $P(W_n|W_{n-1})$ αντίστοιχα αν δεν υπάρχει και αυτή μπορεί να χρησιμοποιηθεί η $P(W_n)$. Για το παράδειγμα ισχύει:

$$P_B(W_n|W_{n-1}W_{n-2}) = \begin{cases} P'(W_n|W_{n-1}W_{n-2}) & Αν\, C(W_n|W_{n-1}W_{n-2})>0 \\ \lambda_i P_B(W_n|W_{n-1}) & αλλιώς\, αν\, C(W_n|W_{n-1})>0 \\ P'(W_n) & σε\, κάθε\, άλλη\, περίπτωση \end{cases} \quad με$$

$$P_B(W_n|W_{n-1}) = \begin{cases} P'(W_n|W_{n-1}) & Αν\, C(W_n|W_{n-1})>0 \\ \lambda_i P'(W_n) & σε\, κάθε\, άλλη\, περίπτωση \end{cases} \quad .$$

Γενικεύοντας το παράδειγμα ο γενικός τύπος πιθανότητας για την τεχνική backoff είναι:

$$P_B(W_n|W_{n-1},W_{n-2},...,W_{n-N+1}) = \begin{cases} P'(W_n|W_{n-1},W_{n-2},...,W_{n-N+1}) & Αν\, C(W_n|W_{n-1},W_{n-2},...,W_{n-N+1})>0 \\ \lambda_i P_B(W_n|W_{n-1},W_{n-2},...,W_{n-N+2}) & σε\, κάθε\, άλλη\, περίπτωση \end{cases}$$

Στις παραπάνω πιθανότητες μπορούν να γίνουν οι εξής παρατηρήσεις, για τον υπολογισμό της P_B δεν χρησιμοποιούνται οι πιθανότητες εκτιμημένες με MLE αλλά οι ομαλοποιημένες πιθανότητες P' αυτό οφείλεται στο γεγονός του ότι για μία πιθανότητα εκτιμημένη 0 με MLE, χρησιμοποιηθεί η τεχνική backoff, τότε αν υπολογισθούν και οι πιθανότητες (N-1)-grams με MLE ουσιαστικά προστίθεται επιπλέον μάζα πιθανότητας στην εξίσωση έτσι υπάρχει περίπτωση να υπάρξει πιθανότητα μεγαλύτερη από ένα. Για αντίστοιχο λόγο χρησιμοποιούνται και τα βάρη λ_i ώστε να υπάρχει η βεβαιότητα το άθροισμα της μάζας πιθανότητας των μικρότερης τάξης N-grams να αντιστοιχεί στην τιμή έκπτωσης των N-grams μεγαλύτερης τάξης αφαιρεμένη από την μονάδα.

Η τεχνική Kneiser-Nay βασίζεται στον συνδυασμό μίας μεθόδου έκπτωσης που ονομάζεται απόλυτη έκπτωση (absolute discounting) και μίας τεχνικής υπολογισμού πιθανότητας από αυτές που περιγράφηκαν παραπάνω. Με την μέθοδο αυτή, για τον υπολογισμό της ομαλοποιημένης συχνότητας c', μία σταθερά d (με τιμή από 0 έως 1) αφαιρείται από την συχνότητα c. Αυτό

προέκυψε από την παρατήρηση ότι η διαφορά, των ομαλοποιημένων συχνοτήτων c' που προκύπτουν από την τεχνική Good-Turing με τις συχνότητες c, είναι περίπου σταθερή για κάθε c. Από την μέθοδο αυτή προκύπτει:

$$P_A(W_n|W_{n-1},W_{n-2},\dots,W_{n-N+1})=\begin{cases}\dfrac{C(W_nW_{n-1},W_{n-2},\dots,W_{n-N+1})-d}{C(W_{n-1},W_{n-2},\dots,W_{n-N+1})} & Αν\, C(W_nW_{n-1},W_{n-2},\dots,W_{n-N+1})>0\\ \lambda_i P_A(W_n|W_{n-1},W_{n-2},\dots,W_{n-N+2}) & \textit{σε κάθε άλλη περίπτωση}\end{cases}$$

Τα βάρη λ_i αποτελούν βάρη μίας εκ των τεχνικών που περιγράφηκαν προηγουμένως (η παραπάνω εξίσωση έχει εκφραστεί με την τεχνική backoff). Η τεχνική Kneiser-Nay βασίζει την εκτίμηση της στον αριθμό των διαφορετικών περιεχομένων που έχει εμφανιστεί η λέξη w (λέξεις που εμφανίζονται συχνότερα είναι πιθανότερο να ξαναεμφανιστούν). Αντιμετωπίζει δηλαδή την πιθανότητα ως συνεχόμενη (continuation propability). Δηλαδή ως:

$$P_C(Wn)=\frac{\|[W_n:C(W_nW_{n-1},W_{n-2},\dots,W_{n-N+1})>0]\|}{\sum_{Wn}\|[W_n:]C(W_nW_{n-1},W_{n-2},\dots,W_{n-N+1})>0|}\quad .$$

Εφαρμόζοντας σε όλα τα παραπάνω την τεχνική backoff προκύπτει ο τύπος πιθανότητας της τεχνικής Kneiser-Nay backoff:

$$P_{KN}(W_n|W_{n-1},W_{n-2},\dots,W_{n-N+1})=\begin{cases}\dfrac{C(W_nW_{n-1},W_{n-2},\dots,W_{n-N+1})-d}{C(W_{n-1},W_{n-2},\dots,W_{n-N+1})} & Αν\, C(W_nW_{n-1},W_{n-2},\dots,W_{n-N+1})>0\\ \lambda_i\dfrac{\|[W_n:C(W_nW_{n-1},W_{n-2},\dots,W_{n-N+1})>0]\|}{\sum_{Wn}\|[W_n:]C(W_nW_{n-1},W_{n-2},\dots,W_{n-N+1})>0|} & \textit{σε κάθε άλλη περίπτωση}\end{cases}$$

Ενώ αν εφαρμοστεί η τεχνική της παρεμβολής προκύπτει ο τύπος πιθανότητας της τεχνικής Kneiser-Nay interpolation:

$$P_{KN}(W_n|W_{n-1},W_{n-2},\dots,W_{n-N+1})=\frac{C(W_nW_{n-1},W_{n-2},\dots,W_{n-N+1})-d}{C(W_{n-1},W_{n-2},\dots,W_{n-N+1})}+\lambda_i\frac{\|[W_n:C(W_nW_{n-1},W_{n-2},\dots,W_{n-N+1})>0]\|}{\sum_{Wn}\|[W_n:]C(W_nW_{n-1},W_{n-2},\dots,W_{n-N+1})>0|}\quad .$$

Έχει παρατηρηθεί ότι η τεχνική Kneiser-Nay interpolation δίνει πιο αξιόπιστα αποτελέσματα απ' ότι η τεχνική Kneiser-Nay backoff. [11]

3.2.2.3 Τεχνική μέγιστης εντροπίας (Maximum Entropy – MaxEnt)

Η τεχνική αυτή στηρίζεται στην αρχή της μέγιστης εντροπίας, δηλαδή: τα συμπεράσματα που προέρχονται από ένα σύνολο με ελλειπή πληροφορία θα πρέπει να αντληθούν από την κατανομή πιθανοτήτων που έχει την μέγιστη εντροπία που επιτρέπει το σύνολο [16]. Πιο συγκεκριμένα σε ένα ακουστικό μοντέλο, γίνεται η προσπάθεια περιγραφής μίας φυσικής γλώσσας μέσω ενός σύνολου εκπαίδευσης (δείγμα με ελλειπή πληροφορία). Η τεχνική αυτή εξισώνει την κατανομή που περιγράφει καλύτερα την γλώσσα με την κατανομή μέγιστης εντροπίας που ταιριάζει (fit) στο σύνολο. Γενικά ως εντροπία (N διακριτών δοκιμών με n^N διαφορετικών αποτελεσμάτων) ορίζεται από:

$$H(f_1\dots f_n)=-\sum_{i=1}^{n} f_i\, logf_i \quad ,$$ όπου οι συχνότητες $f_i=N_i/N$ [16].

Με βάση τα παραπάνω θα πρέπει να ξανά ορισθεί το πρόβλημα του γλωσσικού μοντέλου. Στόχος του είναι η εύρεση της κατανομής πιθανοτήτων p(h,w) όπου $w=W_n$ και $h=W_{n-1},W_{n-2},\dots,W_{n-N+1}$. Από την θεωρία της από κοινού κατανομής πιθανότητας ισχύει $p(h,w) = p(w|h)p_E(h)$ όπου $p_E(h)$ η κατανομή της h που προκύπτει από το σύνολο εκπαίδευσης (εμπειρική κατανομή) και P(w|h) η κατανομή της πιθανότητας να ακολουθεί η λέξη w αν έχει προηγηθεί η φράση h.

Βάσει αυτών μπορεί να ορισθεί η εντροπία της p(w|h) ως:

$$H(p) = -\sum_{h,w} p(h)\, p(w|h)\, log p(w|h) \quad .$$

Έστω f(h,w) η συνάρτηση χαρακτηριστικών που αντιστοιχεί στον δείκτη πραγματοποίησης του ζευγαριού (h,w) και άρα παίρνει τιμή 1 όταν ακολουθεί η λέξη w και έχει προηγηθεί η φράση h και 0 σε κάθε άλλη περίπτωση. Για λόγους συντομίας θα αναπαριστάται ως f και f_i οι συναρτήσεις n χαρακτηριστικών. Έχει αποδειχθεί ότι υπάρχει για κάθε σύνολο που ικανοποιεί την εξίσωση:

$C = [\, p \in P \,|\, p(f_i) = p_E(f_i) \; για \; i \in (1,..,n)]$ όπου P σύνολο όλων των πιθανών κατανομών, μοναδική κατανομή p' με μέγιστη εντροπία. Άρα το πρόβλημα εύρεσης είναι:

$$p' = \underset{p \in C}{argmax}\, H(p) \quad .$$

Στο σημείο αυτό θα εισαχθούν τα βάρη λ_i (πολλαπλασιαστές Lagrange) και θα ορισθεί η συνάρτηση Lagrange Λ(p,λ) ως:

$$\Lambda(p,\lambda) = H(p) + \sum_i \lambda_i (p(f_i) - p_E(f_i)) \quad .$$

Για σταθερά λ μπορεί να υπολογισθεί το μέγιστο της συνάρτησης Lagrange Λ(p,λ) για κάθε p που ανήκει στο P. Έστω p_λ η p που μεγιστοποιεί την Λ(p,λ) και Ψ(λ) η μέγιστη τιμή, τότε:

$$p_\lambda = \underset{p \in P}{argmax} \quad \text{και} \quad \Psi(\lambda) = \Lambda(p_\lambda, \lambda)$$

Οι παραπάνω συναρτήσεις μπορούν να υπολογιστούν εύκολα με απλούς υπολογισμούς και τα αποτελέσματα τους είναι:

$p_\lambda(w|h) = \dfrac{\exp(\sum_i \lambda_i f_i(h,w))}{Z_\lambda(h)}$ και $\Psi(\lambda) = -\sum_h p_E(h)\, log Z_\lambda(h) + \sum_i \lambda_i\, p_E(f_i)$, όπου Z_λ(h) κανονικοποιημένη σταθερά που προκύπτει από τον περιορισμό της πιθανότητας $\sum_w p_\lambda(w|h) = 1$ για όλα τα h και ισούται: $Z_\lambda(h) = \sum_w \exp(\sum_i \lambda_i f_i(h,w))$. Άρα από τα παραπάνω, για λ' τον πολλαπλασιαστή Lagrange που μεγιστοποιεί την Ψ(λ), το πρόβλημα εύρεσης μετασχηματίζεται σε :

$$\lambda' = \underset{\lambda}{argmax}\, \Psi(\lambda) \quad .$$

Ο υπολογισμός της παραμέτρου λ' μπορεί να επιτευχθεί με διάφορες μεθόδους υπολογισμού, η πιο απλή είναι ο Scaling αλγόριθμος:

$\lambda_i = 0$ για όλα τα i σε {1,..,n}
for i from 1 to n:

Βρες το $\Delta\lambda_i$ που ικανοποιεί την $p_E(f_i) = \sum_{h,w} p_E(h)\, p(w|h)\, f_i(h,w) \exp(\Delta\lambda_i \sum_{i=1}^{n} f_i(h,w))$

$\lambda_i = \lambda_i + \Delta\lambda_i$

Αν $\sum_{i=1}^{n} f_i(h,w)$ σταθερό τα $\Delta\lambda_i$ υπολογίζονται εύκολα (με την μέθοδο του Newton) ως :

$$\Delta\lambda_i = \frac{1}{\sum_{i=1}^{n} f_i(h,w)} \log \frac{p_E(f_i)}{p_\lambda(f_i)} \quad .\ [17]$$

Στην πράξη αντί της συνάρτησης Lagrange Λ(p,λ) χρησιμοποιείται η ομαλοποιημένη αυτής:

$$\Lambda'(p,\lambda) = \Lambda(p,\lambda) - \sum_{i=1}^{n} \frac{\lambda_i^2}{2\sigma_i^2} \quad .$$

Συνήθως χρησιμοποιείται $\sigma_i = \sigma$(σταθερό) που εξάγεται από το σύνολο ανάπτυξης. [18]

3.2.2 Γλωσσικό Μοντέλο (συνέχεια)

Τέλος, αφού αναλύθηκαν οι τεχνικές θα πρέπει να εισαχθεί μία μετρική που θα βοηθήσει τον κατασκευαστή να επιλέξει ποιο από τα ακουστικά μοντέλα που αναπτύχθηκαν είναι κατάλληλο για χρήση. Η μετρική αυτή είναι η γλωσσική πολυπλοκότητα (language perplexity) ορίζεται ως ο μέσος αριθμός των λέξεων που μπορεί να ακολουθήσουν ως συνέχεια μίας φράσης και υπολογίζεται για Μ αριθμό λέξεων από τον τύπο:

$$Perplexity = P(W_1, W_2, \ldots, W_M)^{-1/M}$$, καθώς το Μ τείνει το άπειρο.[12]

Στο σημείο αυτό θα πρέπει να αναφερθεί ότι και στις φυσικές γλώσσες υπάρχει πολυπλοκότητα καθώς γλώσσες όπως τα Ελληνικά και τα Γερμανικά που είναι πιο εύκολο να σχηματιστούν σύνθετες λέξεις είναι πιο πολύπλοκες από άλλες γλώσσες π.χ. Αγγλικά.

3.2.3 Ακουστικό Μοντέλο

Όπως αναφέρθηκε παραπάνω το μοντέλο αυτό υπολογίζει την πιθανότητα του ακουστικού διανύσματος $X=\{X_1,X_2,..,X_T\}$, όπου Τ τα frames, να προέρχεται από την σειρά φωνημάτων/λέξεων $W=\{W_1,W_2,\ldots,W_M\}$ όπου Μ τα φωνήματα/λέξεις. Στόχος του είναι αντιστοίχηση της ακουστικής πραγματικότητας των παρατηρούμενων διανυσματικών χαρακτηριστικών,με την γραφική απεικόνιση των εκφράσεων.

$$P(X|W) = P(X_1, X_2, \ldots, X_T|W_1, W_2, \ldots, W_M)$$.

Αν ληφθεί υπόψη ότι κάθε frame του μοντέλου αντιστοιχεί σε ένα φώνημα τότε μπορεί η παραπάνω πιθανότητα να εκφραστεί ως σύνολο πεπερασμένων καταστάσεων [12]. Πριν την περαιτέρω ανάλυση του μοντέλου θα πρέπει να αναφερθεί πώς υπολογίζονται οι πιθανότητες σειρών από τυχαίες καταστάσεις. Αυτό γίνεται με την χρήση αλυσίδων Markov (Markov chain).

3.2.3.1 Κρυφά μοντέλα Markov (Hidden Markov Models-HMM)

Οι αλυσίδες Markov στηρίζονται στην υπόθεση ότι μόνο η γνώση της παρούσας κατάστασης χρειάζεται για την πρόβλεψη της επόμενης, δηλαδή οι καταστάσεις πριν την παρούσα δεν έχουν απολύτως καμία σημασία. Αυτή η μέθοδος πρόβλεψης όμως μπορεί και υπολογίζει το “μέλλον” μόνο για καταστάσεις που είναι άμεσα παρατηρήσιμες, όταν υπάρχουν κρυφές καταστάσεις στα δεδομένα χρησιμοποιούνται τα κρυφά μοντέλα Markov (Hidden Markov Models-HMM). Στην περίπτωση του ακουστικού μοντέλου, έστω μία έκφραση εκφρασμένη σε φωνήματα, όταν τελειώνει μία λέξη, με την χρήση αλυσίδων Markov το μοντέλο θα ψάξει την επόμενη κατάσταση σε κάποιο φώνημα ήχου καθώς δεν παρατηρεί την πληροφορία των λέξεων, τα κρυφά μοντέλα Markov όμως που παρατηρούν την πληροφορία των λέξεων θα ψάξουν την επόμενη κατάσταση στο φώνημα της σιωπής. Για τον λόγο αυτό για την ανάπτυξη του ακουστικού μοντέλου χρησιμοποιούνται HMM.

Για την επεξήγηση των HMM θα ξεχάσουμε για λίγο το ακουστικό σύστημα. Ένα HMM χαρακτηρίζεται από τα εξής στοιχεία:

$Q=\{q_1,q_2,\ldots,q_N\}$	Το σύνολο Ν καταστάσεων.
$A=\{a_{11},..,a_{ij}\ldots,a_{NN}\}$	Ο πίνακας πιθανοτήτων Α, κάθε a_{ij} αντιπροσωπεύει την πιθανότητα μετάβασης από μία κατάσταση i σε μία κατάσταση j με $\sum_i^N a_{ij}=1\ \forall i$.
$O=\{o_1,o_2,\ldots,o_T\}$	Μία σειρά παρατηρήσεων Τ προερχόμενη από ένα λεξικό $V=\{v_1,v_2,..,v_N\}$.
$B=b_i(o_t)$	Η σειρά της πιαθανότητας των παρατηρίσεων o_t, αντιπροσωπεύουν την πιθανότητα μία παρατήρηση να προέρχεται από την κατάσταση i.
$\pi=\{\pi_1,\pi_2,\ldots,\pi_N\}$	Η αρχική κατανομή πιθανότητας των καταστάσεων ουσιαστικά πρόκειται για την πιθανότητα το μοντέλο να ξεκινήσει από μια κατάσταση i. $\sum_i^N \pi_i=1$.

Κάθε κρυφή κατάσταση παράγει μόνο μία παρατήρηση άρα ο αριθμός της σειράς των κρυφών καταστάσεων έχει τον ίδιο αριθμό με τις παρατηρήσεις. Λόγω αυτού μπορεί να συμπεραθεί ότι:

$$P(O|Q)=\prod_{i=1}^{T} P(o_i|q_i)$$

όμως το μοντέλο δεν μπορεί να γνωρίζει σε ποια κρυφή κατάσταση μίας σειράς κρυφών καταστάσεων αντιστοιχεί η κάθε παρατήρηση άρα η παραπάνω πιθανότητα μπορεί να υπολογισθεί με την χρήση της από κοινού κατανομής πιθανοτήτων ως:

$$P(O,Q)=P(O|Q)P(Q)=\prod_{i=1}^{T} P(o_i|q_i)\prod_{i=1}^{T} P(q_i|q_{i-1}) \quad .$$

Αν αθροιστούν, όλων των σειρών, οι από κοινού πιθανότητες μπορεί να βρεθεί η πιθανότητα των παρατηρήσεων P(O):

$$P(O)=\sum_{Q} P(O,Q)=\sum_{Q} P(O|Q)P(Q) \quad .$$

Για ένα HMM με N κρυφές καταστάσεις και σειρές παρατηρήσεων αποτελούμενες από T παρατηρήσεις μπορούν να δημιουργηθούν N^T πιθανές κρυφές σειρές. Στην πράξη ο αριθμός N και ο αριθμός T είναι πολύ μεγάλος και καθιστά αδύνατο τον υπολογισμό της P(O). Για τον λόγο αυτό έχει αναπτυχθεί ένας δυναμικά προγραμματισμένος αλγόριθμος, ο εμπρόσθιος αλγόριθμος (forward algorithm), με πολυπλοκότητα $O(N^2T)$. Ο αλγόριθμος αυτός υπολογίζει την P(O) μέσω ενός διανύσματος (διάνυσμα trellis) $a_t(j)$ που αντιπροσωπεύει την πιθανότητα της κατάστασης j αν έχουν προηγηθεί t παρατηρήσεις για ένα αυτόματο λ και άρα:

$$a_t(j)=P(o_1,o_2,\ldots,o_t,q_t=j|\lambda)$$

κάθε “κελί” του διανύσματος αυτού υπολογίζεται από τον τύπο:

$$a_t(j)=\sum_{i=1}^{N} a_{t-1}(i)a_{ij}b_j(o_t)$$

Ο υπολογισμός του της P(O) με την χρήση του εμπρόσθιου αλγορίθμου γίνεται ως:

1)Βήμα αρχικοποίησης:

$$a_1(j)=\pi_j b_j(o_1) \quad \text{για} \quad 1\le j\le N \quad .$$

2)Βήμα ανάδρασης:

$$a_t(j)=\sum_{i=1}^{N} a_{t-1}(i)a_{ij}b_j(o_t) \quad \text{για} \quad 1\le j\le N, 1<t\le T \quad .$$ Ως ότου υπολογισθεί το διάνυσμα trellis.

3)Βήμα Τερματισμού:

$$P(O|\lambda)=\sum_{i=1}^{N} a_T(i) \quad .$$

Παρ’ ότι ο παραπάνω αλγόριθμος επιλύει το πρόβλημα των HMM δεν είναι και ο πιο αποτελεσματικός. Λόγω αυτού, έχει αναπτυχθεί και είναι ο πιο ευρέως χρησιμοποιούμενος, αλγόριθμος για τον υπολογισμό των HMM, ο αλγόριθμος Baum-Welch ή forward-backward ή Expectation Maximization, που παρέχει δυνατότητα μάθησης αφού εκτιμά αρχικά τις πιθανότητες και έπειτα χρησιμοποιεί τα αποτελέσματα αυτά με σκοπό την ολοένα και καλύτερη εκτίμηση τους. Αυτό επιτυγχάνεται με την χρήση του εμπρόσθιου αλγόριθμου και μίας πιθανότητας $\beta_t(i)$ που ονομάζεται backward propability, που υπολογίζεται από τον οπίσθιο αλγόριθμο (backward algorithm). Το $\beta_t(i)$ με το διάνυσμα $a_t(j)$ είναι όμοια, με την διαφορά ότι το $\beta_t(i)$ “βλέπει” παρατηρήσεις από την χρονική στιγμή t+1 έως το τέλος, δεδομένου ότι το μοντέλο βρίσκεται στην κατάσταση i την χρονική στιγμή t για ένα αυτόματο λ, όμοια λοιπόν με το $a_t(j)$:

$$\beta_t(i)=P(o_{t+1},o_{t+2},\ldots,o_T|q_t=i,\lambda)$$

1)Βήμα αρχικοποίησης:

$$\beta_T(i)=1 \quad \text{για} \quad 1\le j\le N \quad .$$

2)Βήμα ανάδρασης:

$$\beta_t(i)=\sum_{j=1}^{N} \beta_{t+1}(j)a_{ij}b_j(o_{t+1}) \quad \text{για} \quad 1\le j\le N, 1<t\le T \quad .$$

3)Βήμα Τερματισμού:

$$P(O|\lambda)=\sum_{j=1}^{N}\beta_1(j)b_j(o_1)\pi_j \quad .$$

Οι αλγόριθμοι forward και backward βοηθούν στον υπολογισμό του πίνακα καταστάσεων μετάβασης a_{ij} και της πιθανότητας των παρατηρήσεων $b_i(o_t)$. Για τον πίνακα καταστάσεων a_{ij} εξ' ορισμού ισχύει:

$$a_{ij}=\frac{\textit{προσδοκώμενος αριθμός μεταβάσεων από την κατάσταση i στην κατάσταση j}}{\textit{προσδοκόμενος αριθμός μεταβάσεων από την κατάσταση i}}$$

Έστω ξ_t η πιθανότητα το μοντέλο να βρίσκεται στην κατάσταση i την χρονική στιγμή t και στην κατάσταση j την χρονική στιγμή t+1, και ξ_t' η πιθανότητα το μοντέλο να βρίσκεται στην κατάσταση i την χρονική στιγμή t και στην κατάσταση j την χρονική στιγμή t+1 με δεδομένη όμως την παρατήρηση O. Τότε:

$$\xi_t(i,j)=P(q_t=i,q_{t+1}=j|O,\lambda) \quad \text{και} \quad \xi'_t(i,j)=P(q_t=i,q_{t+1}=j,O|\lambda)=a_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j) \quad .$$

Η πιθανότητα των παρατηρήσεων που δίνονται στο μοντέλο είναι :

$$P(O|\lambda)=\sum_{j=1}^{N}a_t(j)\beta_t(j) \quad .$$

Άρα από την κοινού κατανομή πιθανότητας ισχύει:

$$P(q_t=i,q_{t+1}=j,O|\lambda)=P(q_t=i,q_{t+1}=j|O,\lambda)P(O|\lambda) \quad \text{και άρα:}$$

$$\xi_t(i,j)=\frac{a_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{j=1}^{N}a_t(j)\beta_t(j)} \quad .$$

Αθροίζοντας όλα τα t για κάθε ξ εξάγεται ο προσδοκώμενος αριθμός μεταβάσεων από την κατάσταση i στην κατάσταση j και αθροίζοντας όλες τις πιθανές μεταβάσεις από την κατάσταση i εξάγεται ο πίνακας καταστάσεων μετάβασης a_{ij} ,ως:

$$a_{ij}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\sum_{k=1}^{N}\xi_t(i,k)} \quad .$$

Για την πιθανότητα των παρατηρήσεων $b_i(o_t)$, έστω v_k η παρατήρηση από ένα σύνολο V της κατάστασης j. Εξ' ορισμού ισχύει:

$$b_j(v_k)=\frac{\textit{προσδοκώμενος αριθμός στην κατάσταση j να παρατηρηθεί το } v_k}{\textit{προσδοκόμενος αριθμός στην κατάσταση j}} \quad .$$

Έστω $\gamma_t(j)$ η πιθανότητα το μοντέλο να βρίσκεται στην κατάσταση j την χρονική στιγμή t. Τότε:

$$\gamma_t(j)=P(q_t=j|O,\lambda)=\frac{P(q_t=j,O|\lambda)}{P(O|\lambda)}=\frac{a_t(j)\beta_t(j)}{\sum_{j=1}^{N}a_t(j)\beta_t(j)}\text{¿} \quad .$$

Αθροίζοντας όλα τα t για κάθε γ εξάγεται ο προσδοκώμενος αριθμός που το μοντέλο θα βρεθεί στην κατάσταση j και αθροίζοντας όλα τα t που ισχύει $o_t=v_k$ εξάγεται ο προσδοκώμενος αριθμός που θα βρεθεί το μοντέλο στην κατάσταση j και θα παρατηρήσει v_k :

$$b_j(v_k)=\frac{\sum_{t=1,\ which\ o_t=v_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)} \quad .$$

Έχοντας υπολογίσει όλες τις μεταβλητές του αλγορίθμου μπορεί να παρουσιαστεί ο αλγόριθμος Expectation-Maximization:

1)Αρχικοποίηση :

Χρήση forward και backward αλγορίθμου για τον υπολογισμό των παραμέτρων .

2)Βήμα Εκτίμησης (Expectation):

Υπολογισμός των ξ_t και γ_t.

3)Βήμα Μεγιστοποίησης (Maximization):

Υπολογισμός των a_{ij} και $b_i(o_t)$.

4)Βήμα ανάδρασης ή Τερματισμού:

Επιστροφή στο βήμα 2 έως η μεταβολή των παραμέτρων να είναι μηδαμινή.

Ο παραπάνω αλγόριθμος είναι καλός για τον υπολογισμό των πιθανοτήτων αλλά δεν μπορεί να εφαρμοσθεί σε προβλήματα αποκωδικοποίησης, δηλαδή σε προβλήματα που δίνεται ένα HMM και ένα σύνολο παρατηρήσεων, και ζητείται η πιθανότερη σειρά καταστάσεων. Για τον λόγο αυτό έχει αναπτυχθεί ένας αντίστοιχος αλγόριθμος, ο αλγόριθμος Viterbi. Ο Viterbi αντικαθιστά το διάνυσμα $a_t(j)$ του εμπρόσθιου αλγορίθμου με ένα διάνυσμα $v_t(j)$ που αντιπροσωπεύει την πιθανότερη διαδρομή που καταλήγει στην κατάσταση j αν έχουν προηγηθεί t παρατηρήσεις για ένα αυτόματο λ και άρα:

$$v_t(j) = \max_{q_1, q_2, \ldots, q_{t-1}} P(o_1, o_2, \ldots, o_t, q_t = j | \lambda)$$

ομοίως υπολογίζεται το κάθε “κελί” του διανύσματος από τον τύπο:

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, a_{ij}\, b_j(o_t) \quad .$$

Στον αλγόριθμο αυτόν εισάγεται το στοιχείο του οπίσθιου εντοπισμού άλλωστε δεν θα είχε νόημα η εύρεση της πιθανότητας χωρίς την εύρεση της διαδρομής που ακολουθήθηκε. Ο αλγόριθμος Viterbi φαίνεται παρακάτω:

1)Βήμα αρχικοποίησης:

$$v_1(j) = \pi_j b_j(o_1) \quad \text{για} \quad 1 \le j \le N \quad \text{και} \quad bt_1(j) = 0 \quad \text{για} \quad 1 \le j \le N \quad .$$

2)Βήμα ανάδρασης:

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, a_{ij}\, b_j(o_t) \quad \text{για} \quad 1 \le j \le N, 1 < t \le T$$

$$bt(j) = \operatorname*{argmax}_{i=1}^{N} v_{t-1}(i)\, a_{ij}\, b_j(o_t) \quad \text{για} \quad 1 \le j \le N, 1 < t \le T$$

Ως ότου υπολογισθεί το διανύσματα.

3)Βήμα Τερματισμού:

μέγιστη πιθανότητα $P' = \max_{i=1}^{N} v_T(i)$ και αρχική κατάσταση $q_T = \operatorname*{argmax}_{i=1}^{N} v_T(i)$. [19]

Συχνά στην πράξη δεν χρησιμοποιείται ο Viterbi αλλά ο log-Viterbi διότι είναι πιο γρήγορος αφού αντικαθιστά τους παραπάνω πολλαπλασιασμούς με προσθέσεις. [12]

Στο σημείο αυτό μπορούμε να συνοψίσουμε για τα HMM, τα βήματα επίλυσης προβλημάτων με την χρήση αυτών είναι: αρχικά η εκπαίδευση HMM με τον αλγόριθμο Expectation-Maximization για ένα σύνολο εκπαίδευσης και έπειτα η αντιστοίχηση της πιθανότερης σειράς καταστάσεων για μία είσοδο από παρατηρήσεις (δεδομένα συνόλου εκπαίδευσης) με τον αλγόριθμο Viterbi.

3.2.3 Ακουστικό Μοντέλο (συνέχεια)

Αφού, λοιπόν αναλύθηκε ο τρόπος χρήσης των HMM τώρα θα πρέπει να ενσωματωθούν στο πρόβλημα του ακουστικού μοντέλου. Μπορεί να γίνει η υπόθεση ότι κάθε frame t αντιστοιχεί σε ένα μοντέλο φωνήματος i και σε μία κατάσταση j του HMM μέσω μίας συνάρτησης w_j^i επίσης μπορεί να υποτεθεί ότι κάθε frame είναι ανεξάρτητο από τα υπόλοιπα και άρα:

$$P(X|W)=\prod_{t=1}^{T} P(X_t|w_j^i) \quad .$$

Συνδυάζοντας την παραπάνω εξίσωση με την ανάλυση των HMM μπορεί κάποιος να παρατηρήσει ότι τα frames X_t αντιστοιχούν στις παρατηρήσεις o_t και τα μοντέλα φωνημάτων w_j^i στις καταστάσεις q_i [13]. Για την κατανομή των παρατηρήσεων $b_j(X_t)$ μπορεί να θεωρηθεί ότι ακολουθεί κατανομή Gauss, επειδή τα διανύσματα των ακουστικών χαρακτηριστικών αποτελούνται από 39 mfcc χρησιμοποιείται κατανομή Gauss πολλών μεταβλητών. Επίσης η αναφασματική φύση των διανυσμάτων δεν ταιριάζει στην κατανομή Gauss πλήρως, για τον λόγο αυτό χρησιμοποιείται σταθμισμένη κατανομή Gauss πολλών μεταβλητών. Τα μοντέλα αυτού του τύπου ονομάζονται Γκαουσιανά μοντέλα μίξης (Gaussian Mixture Models – GMM). Από τα παραπάνω προκύπτει:

$$b_j(X_T)=\sum_{k=1}^{K} c_{jk}\mathbb{N}[X_T,\mu_{jk},\Sigma_{jk}]=\sum_{k=1}^{k} c_{jk}\frac{1}{\sqrt{2\pi|\Sigma_{jk}|}}\exp[(X-\mu_{jk})^T\Sigma_{jk}^{-1}(X_t-\mu_{jk})] \quad .[11]$$

Η παράμετρος c_{jk} είναι το βάρος του k-οστού στοιχείου μίξης (mixture component) της κατάστασης j και υπόκεινται στους περιορισμούς:

$$c_{jk}\geq 0 \quad \text{και} \quad \sum_{k=1}^{K} c_{jk}=1 \;\; \gamma\iota\alpha\; 1\leq j\leq N \quad .[13]$$

Οι παράμετροι μ_{jk} και Σ_{jk}, είναι η μέση τιμή και ο πίνακας συνδιακύμανσης αντίστοιχα. Έστω ξ_{tk} η πιθανότητα το μοντέλο να βρίσκεται στην κατάσταση j την χρονική στιγμή t με το k-οστό στοιχείο μίξης να υπολογίζεται για την έξοδο της παρατήρησης X_t . Τότε :

$$\xi_{tk}(j)=\frac{\sum_{i=1}^{} Na_{t-1}(j)a_{ij}c_{jk}b_{jk}(X_t)\beta_t(j)}{\sum_{j=1}^{N} a_t(j)b_t(j)} \quad .$$

Μέσω αυτής της πιθανότητας μπορούν να υπολογιστούν οι παράμετροι c_{jk}, μ_{jk} και Σ_{jk} ως:

$$c_{jk}=\frac{\sum_{t=1}^{T}\xi_{tk}(j)}{\sum_{t=1}^{T}\sum_{k=1}^{K}\xi_{tk}(j)} \quad , \quad \mu_{jk}=\frac{\sum_{t=1}^{T}\xi_{tk}(j)X_t}{\sum_{t=1}^{T}\sum_{k=1}^{K}\xi_{tk}(j)} \quad \text{και} \quad \Sigma_{jk}=\frac{\sum_{t=1}^{T}\xi_{tk}(j)(X_t-\mu_{jk})(X_t-\mu_{jk})^T}{\sum_{t=1}^{T}\sum_{k=1}^{K}\xi_{tk}(j)} \quad .[11]$$

Αντικαθιστώντας τους παραπάνω τύπους στους αλγορίθμους Expectation-Maximization και Viterbi επιτυγχάνεται η (στατιστική) εκπαίδευση και η αποκωδικοποίηση, αντίστοιχα, του ακουστικού μοντέλου. Στην συνέχεια θα αναλυθούν κάποιες τεχνικές που χρησιμοποιούνται ευρέως για την βελτίωση των αποτελεσμάτων του ακουστικού μοντέλου.

3.2.3.2 Τριφωνικά Γκαουσιανά Μοντέλα Μίξης (GMM)

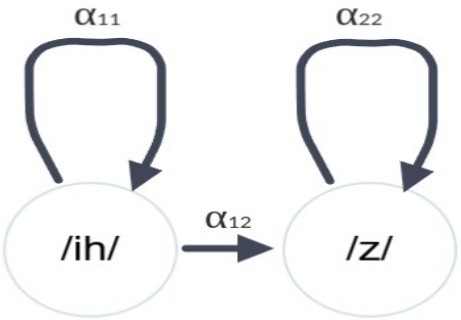
Σε ένα Γκαουσιανό Μοντέλο Μίξης μπορεί η κάθε μία κατάσταση να αντιπροσωπεύει και ένα φώνημα τότε το μοντέλο ονομάζεται μονοφωνικό. Έχει παρατηρηθεί ότι αν κάθε φώνημα αντιστοιχεί σε τρεις καταστάσεις αντί μίας, τα αποτελέσματα του μοντέλου είναι καλύτερα. Αυτό συμβαίνει διότι τα στατιστικά χαρακτηριστικά αντιπροσωπεύονται καλύτερα κατά αυτόν τον τρόπο δηλαδή διαφορετικά χαρακτηριστικά υπάρχουν στην αρχή, στην μέση και στο τέλος του ήχου. Είναι φανερά τα αποτελέσματα αυτής της επέκτασης στα “διπλά” φωνήματα (όπως /iy/, / ay/ , /ae/ κ.α.) καθώς το φώνημα είναι “αδιάσπαστο” αλλά τα ηχητικά χαρακτηριστικά του αλλάζουν κατά την διάρκεια του.

3.2.3.3 Linear Discriminant Analysis - LDA

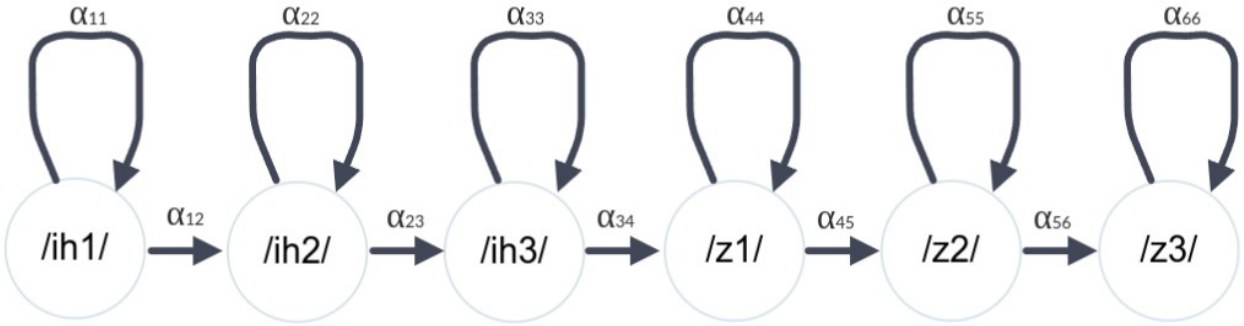
Πρόκειται για μία κατηγορία τεχνικών που χρησιμοποιείται ευρέως με στόχο την μείωση των υπολογισμών, κατ’ επέκταση του χρόνου για την κατασκευή των μοντέλων. Στόχος των τεχνικών αυτών είναι η εύρεση της σχέσης των κατηγορηματικών μεταβλητών (κλάσεις - στην περίπτωση του ακουστικού μοντέλου φωνήματα ή λέξεις ή ακόμα και ολόκληρες εκφράσεις) με ένα

σύνολο αλληλένδετων μεταβλητών (σύνολο δεδομένων) [20]. Η τεχνική LDA επιτυγχάνει μείωση των υπολογισμών μέσω της μείωσης της διαστατικότητας των δεδομένων. Στηρίζεται στην ιδέα εύρεσης ενός γραμμικού μετασχηματισμού τέτοιου ώστε τα διανύσματα χαρακτηριστικών X n-διαστάσεων να μετατραπούν σε διανύσματα Y m-διαστάσεων με m<n με τρόπο που προβλέπει τον βέλτιστο διαχωρισμό μεταξύ των κλάσεων. Χρησιμοποιούνται Scatter πίνακες (πίνακες που χρησιμοποιούνται για τον υπολογισμό της συνδιακύμανσης) για την μοντελοποίηση του προβλήματος βελτιστοποίησης. Έστω ο συνολικός πίνακας Scatter S_1=T, πίνακας S_2=W ο πίνακας Scatter εσωτερικής κλάσης (within-class) και B η ενδιάμεση κλάση (between class). Τα πιο ευρέως χρησιμοποιούμενα κριτήρια μετασχηματισμού είναι η μεγιστοποίηση του ίχνους και της ορίζουσας:

$$J_1(m)=tr(S_{2y}^{-1}S_{1y}) \quad \text{και} \quad J_2(m)=det(S_{2y}^{-1}S_{1y}) \quad . \text{ [21]}$$



(Διάγραμμα 3.3): Διάγραμμα καταστάσεων μονοφωνικού GMM, για την λέξη “is”.



(Διάγραμμα 3.4): Διάγραμμα καταστάσεων τριφωνικού GMM, για την λέξη “is”. [13]

3.2.3.4 Γραμμικός μετασχηματισμός μέγιστης πιθανοφάνειας

Ο Γραμμικός μετασχηματισμός μέγιστης πιθανοφάνειας (maximum likehood linear transformation – MLLT) αντί της χρήσης ενός πίνακα συνδιακύμανσης και μέσης τιμής για κάθε στοιχείο (mixture component), χρησιμοποιεί κοινούς πίνακες για κάποια από αυτά. Αυτό βοηθάει στην αποτελεσματικότητα της ταξινόμησης αφού η θεώρηση του διακριτού πίνακα συνδιακύμανσης και μέσης τιμής για κάθε στοιχείο ανταποκρίνεται καλά όταν τα στοιχεία είναι αμοιβαίως ανεξάρτητα. Ο πίνακας συνδιακύμανσης κάθε στοιχείου, βάσει του μετασχηματισμού αυτού, αποτελείται από δύο πίνακες, έναν πίνακα γραμμικού μετασχηματισμού H^T που μοιράζονται τα στοιχεία μεταξύ τους και τον διαγώνιο $\Lambda_j = diag(\lambda_j)=(\lambda_j^1,\lambda_j^2,\ldots,\lambda_j^n)$.Για τον μετασχηματισμό της μέσης τιμής χρησιμοποιείται η μέση τιμή, ένας πίνακας μετασχηματισμού A^T αλλά και ένα διάνυσμα προκατάληψης (bias vector) b. Οι μετασχηματισμοί ορίζονται ως :

$$\Sigma_j^{-1} \approx H\Lambda_j H^T = \sum_{k=1}^{n} \lambda_j^{\kappa} h_k h_k^T \quad \text{και} \quad \hat{\mu} = A^T \mu_{jk} + b \quad .$$

Κατ’ αυτόν τον τρόπο επιτυγχάνεται εκτίμηση των παραμέτρων του GMM μέσω της μέγιστης πιθανοφάνειας.

Για σύνολο παραμέτρων Θ ισχύει:

$$\hat{\Theta}=\underset{\Theta}{argmax}\sum_{t=1}^{T} p(X_t|\Theta) \quad .[22][23]$$

Όμοια ορίζεται και ο περιορισμένος γραμμικός μετασχηματισμός μέγιστης πιθανοφάνειας (constrained maximum likehood linear transformation – CMLLT) πιο συγκεκριμένα οι εξισώσεις μετασχηματισμού είναι:

$$\Sigma_j^{-1}\approx W\Lambda_j W^T=\sum_{k=1}^{n}\lambda_j^{\kappa} w_k w_k^T \quad \text{και} \quad \hat{\mu}=W^T\mu_{jk}-b \quad .$$

Η διαφορά τους είναι ο κοινός πίνακας μετασχηματισμού μέσης τιμής (W^T) και συνδιακύμανσης και το διάνυσμα προκατάληψης δρα αρνητικά στον υπολογισμό της νέας μέσης τιμής.

Για τον υπολογισμό αυτών στην πράξη χρησιμοποιείται η γραμμική παλινδρόμηση μέγιστης πιθανοφάνειας (maximum likehood linear regression – MLLR) και η επέκταση αυτής η περιορισμένη γραμμική παλινδρόμηση μέγιστης πιθανοφάνειας (constrained maximum likehood linear regression – CMLLR) [23]. Ο αναλυτικός υπολογισμός των παραπάνω εξισώσεων δεν μπορεί να γίνει στην ανάλυση της παρούσας διπλωματικής καθώς οι υπολογισμοί θα είναι αρκετά μακροσκελείς και η ανάλυση τους θα παρέκλινε την ροή, αν ο αναγνώστης ενδιαφέρεται παρ' όλα αυτά για την ανάπτυξη των παραπάνω εξισώσεων μπορεί να την βρει στην έρευνα του M.J.F. Gales [23].

3.2.3.5 Speaker adaptive training (SAT)

Πρόκειται για μία κατηγορία τεχνικών που χρησιμοποιείται ευρέως με στόχο την ενσωμάτωση της πληροφορίας της προφοράς του ομιλητή στην αναζήτηση. Η ιδέα στηρίζεται στο ότι ένα σύστημα αναζήτησης μπορεί και είναι πιο αποδοτικό όταν προσαρμόζεται στον ομιλητή. Όπως αναφέρθηκε στο κεφάλαιο 2 στόχος του συστήματος που αναπτύσσετε στην παρούσα διπλωματική είναι να μπορεί να αναγνωρίζει φωνή ανεξάρτητα του των χαρακτηριστικών του ομιλητή (φύλο, ηλικία κτλπ). Συνυπολογίζοντας την απόδοση και τον στόχο όμως, μπορούν να ενσωματωθούν τεχνικές που μπορούν και προσαρμόζουν την προφορά νέων ομιλητών αρκετά γρήγορα. Οι τεχνικές αυτές στηρίζονται στους γραμμικούς μετασχηματισμός μέγιστης πιθανοφάνειας και πιο συγκεκριμένα στους αλγορίθμους MLLR και CMLLR εισάγοντας τα χαρακτηριστικά του ομιλητή στους πίνακες μετασχηματισμού, προσθέτοντας μία διάσταση στους πίνακες των παραμέτρων και των χαρακτηριστικών.

Αξίζει να σημειωθεί ότι στην περίπτωση του SAT με CMLLR επειδή ο πίνακας μετασχηματισμού είναι κοινός για συνδιακύμανση και μέση τιμή, αυτός μπορεί να εφαρμοσθεί άμεσα στα διανύσματα χαρακτηριστικών για τον υπολογισμό των παραμέτρων του GMM,

$$\hat{X}_t=W_s X_t+b_s \quad .$$

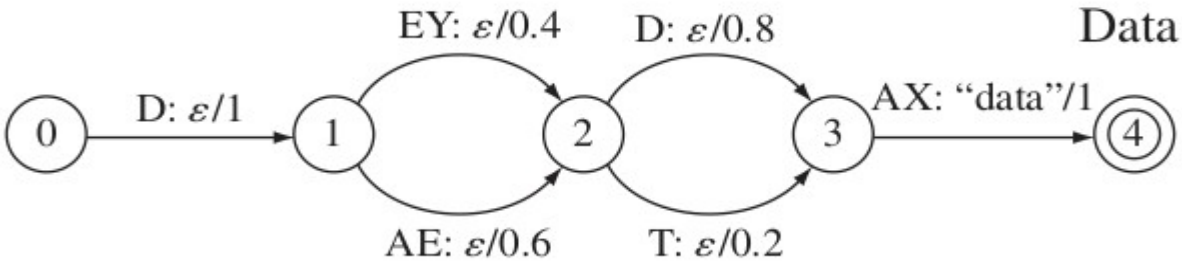
Για τον λόγο αυτό συχνά η τεχνική SAT με CMLLR αναφέρεται ως γραμμική παλινδρόμηση μέγιστης πιθανοφάνειας στον χώρο των χαρακτηριστικών (feature space maximum likehood linear regression – fMLLR). Η τεχνική αυτή μειώνει αρκετά τον αριθμό των υπολογισμών σε σχέση με την απλή SAT που για τον υπολογισμό της χρειάζονται σημαντικοί υπολογιστικοί πόροι.[23][24]

3.2.4 Αναζήτηση πιθανότερης έκφρασης

Για την ολοκλήρωση της ανάλυσης της μοντελοποίησης του συστήματος αναγνώρισης προτύπων θα πρέπει να επεξηγηθεί το πώς γίνεται η αναζήτηση της πιθανότερης έκφρασης σε όλες τις πιθανές εκφράσεις του γλωσσικού μοντέλου. Από μία πρώτη ματιά η παραπάνω αναζήτηση φαντάζει αδύνατη από άποψη υπολογιστικών πόρων, λόγω του όγκου των πιθανών εκφράσεων και της πολυπλοκότητας του γλωσσικού μοντέλου. Για την αντιμετώπιση του παραπάνω ζητήματος χρησιμοποιούνται μέθοδοι και τεχνικές της θεωρίας των αυτομάτων πεπερασμένων καταστάσεων (finite state automata). Οι μέθοδοι αυτοί δίνουν την δυνατότητα λύσεων μέγιστης πιθανοφάνειας με

χρήση εφικτών υπολογιστικών πόρων. Αυτό επιτυγχάνεται κατασκευάζοντας δίκτυα πεπερασμένων καταστάσεων (finite state networks – FSN) με χρήση σταθμισμένων μετατροπέων πεπερασμένων καταστάσεων (weighted finite state transducers – WFST).

Στο διάγραμμα 3.5 φαίνεται το δίκτυο που αντιστοιχεί στην προφορά της λέξης /Data/. Είναι φανερό ότι υπάρχουν διάφοροι τρόποι έκφρασης (τέσσερεις για την λέξη /Data/) και κάθε ακμή του δικτύου εμπεριέχει ένα φώνημα και ένα βάρος που αντιστοιχεί στην πιθανότητα του φωνήματος αυτού. Συνδυάζοντας τις διάφορες προφορές για την εκάστοτε λέξη, μειώνεται ο όγκος των



(Διάγραμμα 3.5.): Απεικόνιση της λέξης Data ως FSN. [12]

αναζητήσεων αφού σε αντίθετη περίπτωση θα έπρεπε να μετρηθούν (ή να ετικετοποιηθούν) ξεχωριστά οι όποιες διαφορετικές εκφράσεις της ίδιας λέξης, το γεγονός αυτό θα μείωνε και την αποτελεσματικότητα του συστήματος αφού η πιθανότητα, των λέξεων με παραπάνω από μία προφορές, θα μειωνόταν. Άρα, αφότου κατασκευαστούν τα δίκτυα των λέξεων (του λεξικού) βάσει της προφοράς τους μπορούν να συνδυαστούν βάσει του γλωσσικού μοντέλου σχηματίζοντας το ένα FSN. Όμοια μπορεί να εκφραστεί και το ακουστικό μοντέλο όπως έχει προαναφερθεί.

Έτσι λοιπόν χρησιμοποιώντας τις καταστάσεις των HMM μπορούν να εκφραστούν οι πιθανότητες μετάβασης των FSN των φωνημάτων, από τα φωνήματα μπορούν να εκφραστούν τα FSN των λέξεων και ομοίως τα FSN των εκφράσεων. Ο συνδυασμός των παραπάνω δικτύων οδηγεί στην κατασκευή ενός μεγάλου γράφου. Οι WFST χρησιμοποιούν μία ενοποιημένη μαθηματική δομή ώστε να μεταγλωττίσουν (compile), τον προερχόμενο από ένα μεγάλο σύνολο δεδομένων, γράφο σε μικρές αναπαραστάσεις που μπορούν να αποκωδικοποιηθούν εύκολα από τον αλγόριθμο Viterbi.[12]

3.3 Επαλήθευση επίδοσης συστημάτων αναγνώρισης φωνής

Στο σημείο αυτό θα πρέπει να εισαχθεί ο τρόπος επιλογής του πιο αποτελεσματικού συστήματος αναγνώρισης προτύπων μεταξύ αυτών που κατασκευάσθηκαν. Για την επιλογή αυτή υπάρχουν διάφορες μετρικές αποτελεσματικότητας ενός συστήματος αναγνώρισης φωνής, η πιο συχνή στη χρήση μετρική και αυτή που χρησιμοποιείται στην παρούσα εργασία, είναι αυτή του ρυθμού των λάθος λέξεων (Word Error Rate – WER). Η μετρική αυτή αντιπροσωπεύει το ποσοστό των λανθασμένων λέξεων που υπάρχουν σε μία έκφραση. Στηρίζεται στις τρεις κατηγορίες λάθους που μπορεί να κάνει ένα σύστημα αναγνώρισης φωνής, που είναι:

1) Η εισαγωγή λέξεων, πρόκειται για λέξεις που δεν υπάρχουν στην έκφραση αλλά το σύστημα τις αναγνωρίζει. Το λάθος αυτό παρατηρείται συνήθως όταν ο ομιλητής κομπιάζει ή σταματά για λίγο την πρόταση του, στις περιπτώσεις αυτές αναγνωρίζονται λέξεις μικρής διάρκειας όπως άρθρα.
2) Αντικατάσταση λέξεων, δηλαδή στην θέση μίας λέξης να εμφανίζεται μία άλλη, συνήθως με παρόμοια προφορά.
3) Η διαγραφή λέξεων, όταν μία λέξη που ειπώθηκε στην έκφραση δεν εμφανίζεται καθόλου αλλά ούτε δίνεται κάποια άλλη ως εναλλακτική. Το λάθος αυτό παρατηρείται συνήθως όταν δύο λέξεις στη σειρά μπορούν και σχηματίζουν μία σύνθετη, ουσιαστικά σε αυτή την περίπτωση πρόκειται για τον συνδυασμό των δύο παραπάνω λαθών.

Από τα παραπάνω μπορεί να οριστεί ο WER ως:

$$WER = \frac{Αριθμός\ Εισαγωγών + Αριθμός\ Αντικατάστασης + Αριθμός\ Διαγραφών}{Αριθμός\ λέξεων\ έκφρασης\ που\ ειπώθηκε}$$

Συνήθως εκφράζεται ως WER επί της εκατό (WER%). [12] Λόγω των λαθών που προέρχονται από εισαγωγή λέξεων το WER% μπορεί να ξεπεράσει το 100%. [11]

3.4 Βελτιστοποίηση συστημάτων αναγνώρισης φωνής με χρήση μηχανικής μάθησης

Από τα παραπάνω μπορεί να κατασκευαστεί ένα σύστημα αναγνώρισης φωνής με την χρήση στατιστικής μάθησης, τις μέρες που γράφεται η παρούσα διπλωματική τέτοιου τύπου συστήματα θεωρούνται ξεπερασμένα από συστήματα που εμπεριέχουν νευρωνικά δίκτυα καθώς τα τελευταία παρουσιάζουν καλύτερη αποδοτικότητα. Τα νέα συστήματα ακολουθούν την παραπάνω διαδικασία αντικαθιστώντας τόσο τις τεχνικές που σχετίζονται με την εύρεση της μεγαλύτερης πιθανοφάνειας όσο και τις τεχνικές εκείνες που σχετίζονται με την βελτιστοποίηση των μοντέλων. Φυσικά για να μπορέσει κάποιος να κατασκευάσει ένα σύγχρονο σύστημα αναγνώρισης φωνής είναι αναγκαία η κατασκευή του στατιστικού συστήματος. Παρακάτω θα αναλυθούν τόσο οι τεχνικές που βοηθούν στον σχεδιασμό των μοντέλων με χρήση νευρωνικών δικτύων όσο και τα ίδια τα νευρωνικά δίκτυα καθώς και ότι είναι απαραίτητο για την εκπαίδευση τους.

3.4.1 I-vectors

Στην προσπάθεια ανάπτυξης σύγχρονων συστημάτων αναγνώρισης ομιλητή αναπτύχθηκε μία τεχνική που διαχωρίζει την πληροφορία του ομιλητή τόσο από το περιβάλλον (διαύλος) αυτού όσο και από τα λεγόμενα του, η τεχνική αυτή στηρίζεται στην εξαγωγή των διανυσμάτων ταυτοποίησης (identity vectors ή i-vectors). Πριν την ανάπτυξη των i-vectors θα γίνει μία προσπάθεια επεξήγησης της παραπάνω πρότασης. Ως περιβάλλον θα πρέπει να θεωρηθεί το οτιδήποτε δεν σχετίζεται με τον ομιλητή, από τον χώρο που βρίσκεται έως και τα μέσα που χρησιμοποιούνται για την ηχογράφηση αυτού. Η πληροφορία του περιβάλλοντος μπορεί να χρησιμοποιηθεί αντίστοιχα σε εφαρμογές αναγνώρισης μέσου π.χ. τηλέφωνο, συστοιχία μικροφώνων στο χώρο (πολυκαναλικό μικρόφωνο) κ.α. αλλά και σε εφαρμογές αναγνώρισης θορύβου. Για τον ομιλητή εξάγονται τα χαρακτηριστικά της φωνής αυτού ανεξαρτήτως γλώσσας ή διαλέκτου που χρησιμοποιεί, φύλου και ηλικίας. Σε εφαρμογές αναγνώρισης φωνής χρησιμοποιούνται ως τεχνική SAT, δηλαδή ο ρόλος τους είναι να κατηγοριοποιούν μία έκφραση/ηχογράφηση στον χώρο των χαρακτηριστικών του ομιλητή. Είναι σημαντικό στο σημείο αυτό να αναφερθεί ότι η τεχνική αυτή είναι πιο εύκολο να εισαχθεί σε συστήματα αναγνώρισης φωνής απ’ ότι ομιλητή μιας και ένα σύστημα αναγνώρισης φωνής “ενδιαφέρεται” μόνο για την αλλαγή ομιλητή, γεγονός που αντανακλά στον όγκο τον δεδομένων που χρειάζεται την εξαγωγή των i-vectors.

Η μετατροπή ενός τμήματος φωνής σε αναπαράσταση ενός μικρής διαστατικότητας i-vector στηρίζεται στην παραγοντική ανάλυση (Factor Analysis) [25]. Έστω M ένα υπερδιάνυσμα που αναπαριστά μία έκφραση ομιλίας δηλαδή εμπεριέχει τόσο την πληροφορία του ομιλητή (χαρακτηριστικά ομιλητή και λεγόμενα) ως διάνυσμα s όσο και την πληροφορία του περιβάλλοντος ως διάνυσμα c, μπορεί να γίνει η θεώρηση ότι M=s+c. Η παραπάνω θεώρηση οφείλεται στο γεγονός του ότι η πληροφορία του ομιλητή είναι ανεξάρτητη από το περιβάλλον. Έστω τώρα, ένα στατιστικό μοντέλο (καθολικό μοντέλο παρασκηνίου (universal background montel – UBM)) όπως το GMM που περιγράφηκε παραπάνω. Έστω C ο αριθμός των στοιχείων (μίξης για GMM) και F η διάσταση των ακουστικών χαρακτηριστικών (αριθμός των MFCC). Αρχικά θα πρέπει να ενωθούν τα στοιχεία C με το διάνυσμα μέσης τιμής του GMM (με διάσταση F) σχηματίζοντας ένα υπερδυάνισμα διάστασης CF.

Από κοινού οι τα δύο διανύσματα (s και c) αποσυντίθενται σε ένα σύνολο μικρότερης διαστατικότητας παραγόντων, κάθε παράγον δρα στην διάσταση που ανταποκρίνεται στα στοιχεία.

Έστω η κατανομή του διανύσματος του ομιλητή s έχει κρυφή παραγοντική απεικόνιση της μορφής:

$$s = m + Vy + Dz \quad .$$

Όπου, m=CFx1 το υπερδιάνυσμα του UBM(ανεξάρτητο από τον ομιλητή και το περιβάλλον ουσιαστικά ανταποκρίνεται στα λεγόμενα του ομιλητή), V τετραγωνικός πίνακας μικρής διάστασης που ονομάζεται ιδιοφωνικός πίνακας (eigenvoice matrix) τα στοιχεία του αντιστοιχούν στις αποκρίσεις των παραγόντων y, D είναι ο CFxCF ο υπολειπόμενος διαγώνιος πίνακας (residual diagonial matrix), y είναι το διάνυσμα που απεικονίζει τους παράγοντες του ομιλητή και z είναι κανονικά κατανεμημένο τυχαίο διάνυσμα διάστασης CF που απεικονίζει συγκεκριμένους και υπολειπόμενους (residual) παράγοντες που σχετίζονται με τον ομιλητή. Για το διάνυσμα c ισχύει:

$$c = Ux \quad .$$

Όπου U ο ιδιοφωνικός πίνακας του περιβάλλοντος και x είναι το διάνυσμα που απεικονίζει τους παράγοντες του περιβάλλοντος.[26]

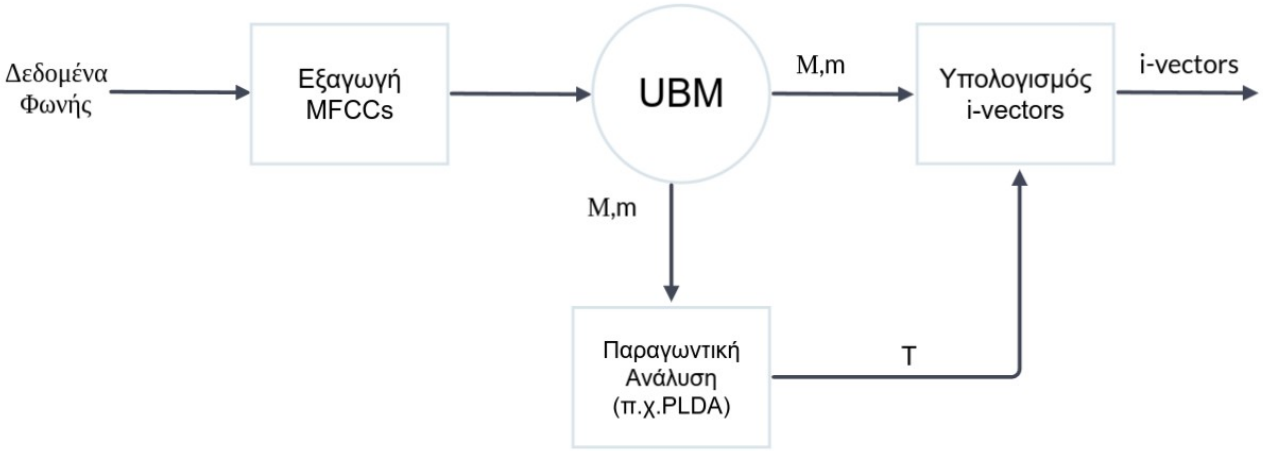
Έχει παρατηρηθεί [27] ότι δεν μπορεί να απεξαρτηθούν στην πράξη τα διανύσματα s και c καθώς θα εμπεριέχεται πάντα πληροφορία του ομιλητή στο c αυτό συμβαίνει λόγω του διαύλου (τηλέφωνο,μικρόφωνο κ.α.), λόγω αυτής της παρατήρησης η παραπάνω θεώρηση χαρακτηρίζεται ως ιδανική,και επαναπροσδιορίζεται εξ αρχής ως εξής:

$$M = m + Tx \quad .$$

Όπου m το υπερδιάνυσμα του UBM , ο T μικρής διαστατικότητας τετραγωνικός πίνακας που συνδέει τον υποχώρο που καλύπτει της μεταβολές των παραγόντων τόσο του s όσο και του c με τον χώρο του υπερδιανύσματος, και x είναι το κανονικά κατανεμημένο τυχαίο διάνυσμα που εμπεριέχει της μεταβολές των παραγόντων του ομιλητή και του περιβάλλοντος και ονομάζεται i-vector. Σε αυτή την θεώρηση το M ακολουθεί κανονική κατανομή με m διάνυσμα μέσων τιμών και TT’ πίνακας συνδιακύμανσης. [25][26]

Στο Διάγραμμα 3.6 φαίνεται το μοντέλο που χρησιμοποιείται για την εξαγωγή των i-vectors η διαδικασία που ακολουθείται είναι η παρακάτω :

1) Αρχικά εξάγονται τα διανύσματα χαρακτηριστικών (MFCC) και εκπαιδεύεται το UBM .
2) Για κάθε έκφραση εξάγονται οι μέγιστες a posteriori (maximum a posteriori -MAP) πιθανότητες των φωνητικών χαρακτηριστικών και υπολογίζεται τα υπερδιανύσματα M και m.
3) Έπειτα εφαρμόζεται παραγοντική ανάλυση στο M, συνήθως χρησιμοποιείται Probabilistic LDA, που πρόκειται για επέκταση της LDA με την διαφορά του ότι οι γραμμικά διαχωρίσιμες κλάσεις που ανήκουν τα στοιχεία του T ακολουθάνε κανονική κατανομή.
4) Τέλος, έχοντας όλα τα στοιχεία μπορεί να υπολογιστούν τα i-vectors.[26][28]



(Διάγραμμα 3.6): Μοντέλο εξαγωγής i-vector.

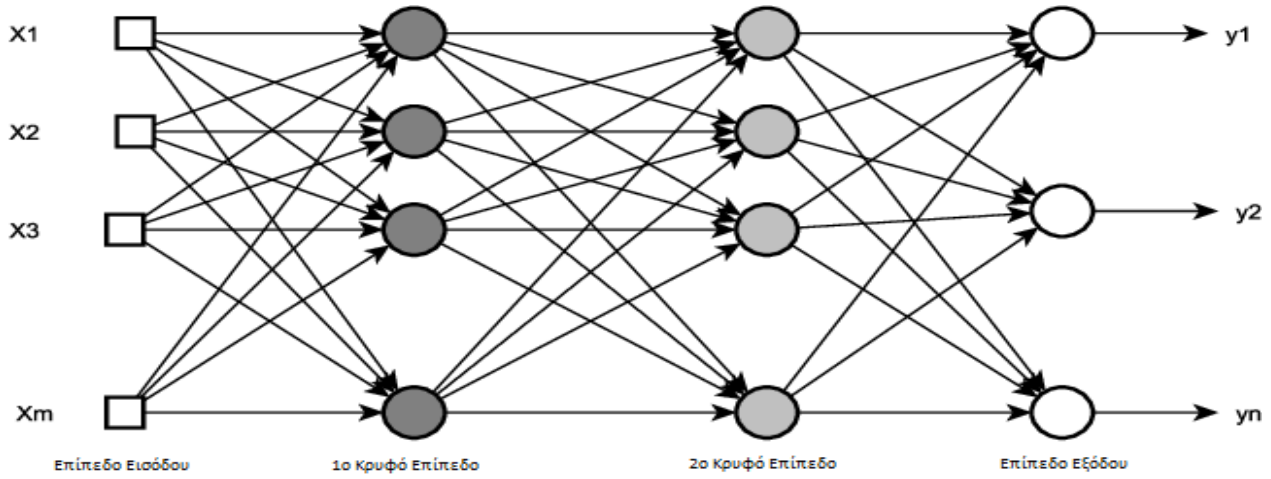
Τέλος αξίζει να σημειωθεί ότι αν παρθεί ως UBM μοντέλο κατασκευασμένο από βαθιά νευρωνικά δίκτυα και ακολουθηθεί η παραπάνω διαδικασία τα διανύσματα που δημιουργούνται ονομάζονται x-vectors ή διανύσματα ενσωμάτωσης (embedding vectors). Τα x-vectors συγκριτικά με τα i-vectors παρουσιάζουν καλύτερη αποδοτικότητα στην απομόνωση της πληροφορίας του ομιλητή, παρ’ όλα αυτά για την εξαγωγή τους χρειάζεται επιπλέον ώρες δεδομένων. Έχει

παρατηρηθεί ότι για την εξαγωγή αποδοτικών x-vectors χρειάζεται ως και δύο τάξεις μεγέθους επιπλέον δεδομένα κάτι που τα κάνει δυσπρόσιτα προς το παρόν σε εφαρμογές αναγνώρισης φωνής ή ομιλητή.

3.4.2. Νευρωνικά Δίκτυα

Η βασική δομή του απλού νευρώνα περιγράφηκε στο κεφάλαιο 1, σε αυτό το κεφάλαιο θα γίνει εμβάθυνση στον τομέα της μηχανικής μάθησης καθώς θα εισαχθούν πολυπλοκότερες δομές των νευρωνικών δικτύων.

3.4.2.1 Perceptron πολλών επιπέδων (multi layer perceptron – MLP)



(Διάγραμμα 3.7): Αρχιτεκτονικός γράφος πλήρως συνδεδεμένου perceptron πολλών επιπέδων με δύο κρυφά επίπεδα.

Όπως γίνεται κατανοητό από το όνομα του ένα MLP αποτελείται από πολλά επίπεδα νευρώνων. Κάθε επίπεδο αποτελείται από πολλούς απλούς νευρώνες. Τα επίπεδα χαρακτηρίζονται είτε ως φανερά (επίπεδα εισόδου και εξόδου) είτε ως κρυφά (επίπεδα ενδιάμεσα από την είσοδο και την έξοδο). Η λειτουργία των κρυφών επιπέδων είναι να δρουν ως ανιχνευτές των χαρακτηριστικών, κατά την διαδικασία μάθησης αρχίζουν σταδιακά να ανακαλύπτουν την σημαντική πληροφορία των διανυσμάτων εισόδου. Αυτό συμβαίνει λόγω ενός μη γραμμικού μετασχηματισμού των διανυσμάτων εισόδου σε έναν νέο χώρο (χώρος χαρακτηριστικών) που ενδιαφέρει την εργασία ταξινόμησης προτύπων. Ανάλογα την σύνδεση μεταξύ των επιπέδων ένα MLP δίκτυο χωρίζεται είτε ως πλήρως συνδεόμενο (fully conectected) είτε ως μερικώς συνδεμένο (partially conected), ενώ ανάλογα την πορεία της πληροφορίας χωρίζεται είτε ως εμπρόσθιο (feedforward) είτε ως αναδρομικό (recurrent).

Από τα παραπάνω μπορούν να εξαχθούν τα βασικά χαρακτηριστικά ενός MLP είναι:

1) Το μοντέλο κάθε νευρώνα στο δίκτυο περιλαμβάνει μια μη γραμμική συνάρτηση ενεργοποίησης, η οποία είναι διαφορίσιμη.
2) Το δίκτυο περιέχει ένα ή περισσότερα επίπεδα τα οποία παραμένουν κρυφά. Συχνά όταν τα κρυφά επίπεδα είναι περισσότερα από δύο το δίκτυο αποκαλείται βαθύ νευρωνικό δίκτυο (Deep Neural Network – DNN).
3) Το δίκτυο επιδεικνύει μεγάλη διασυνδεσιμότητα, ο βαθμός της οποίας καθορίζεται από τα συναπτικά βάρη του δικτύου. [7][29]

Έστω τώρα ένα MLP με L+1 επιπέδων, για το διάνυσμα ενεργοποίησης θα ισχύει:

$$v^l = \varphi(z^l) = \varphi(W^l v^{l-1} + b^l) \quad \text{για } 0<l<L.$$

Όπου l το επίπεδο του νευρωνικού δικτύου για l=0 το επίπεδο εισόδου και l=L το επίπεδο εξόδου. Όπου $v^l \in \mathbb{R}^{N_l x 1}$ το διάνυσμα ενεργοποίησης, για l=0 το διάνυσμα ενεργοποίησης ισούται με την παρατήρηση προς επεξεργασία (διάνυσμα χαρακτηριστικών) $v^0 = o \in \mathbb{R}^{N_0 x 1}$. $z^l \in \mathbb{R}^{N_l x 1}$ είναι

το διάνυσμα διέγερσης, φ μία από τις συναρτήσεις ενεργοποίησης, οι νευρώνες κάθε επίπεδου συχνά έχουν διαφορετική συνάρτηση ενεργοποίησης ανάλογα τον σκοπό που επιτελούν.
$W^l \in \mathbb{R}^{N_l x N_{l-1}}$ ο πίνακας των συναπτικών βαρών και $b^l \in \mathbb{R}^{N_l x 1}$ το διάνυσμα προκατάληψης (bias). Τέλος ο $N_l \in \mathbb{R}$ είναι ο αριθμός της διάστασης, δηλαδή ο αριθμός των νευρώνων που υπάρχουν σε κάθε επίπεδο, για l=0 N_0 ισούται με τον αριθμό των χαρακτηριστικών και για l=L με τον αριθμό των κλάσεων ταξινόμησης. [29]

3.4.2.2 Συναρτήσεις ενεργοποίησης

Πριν την ανάλυση της εκπαίδευσης των νευρωνικών δικτύων θα πρέπει να αναφερθούν κάποια σημαντικά εργαλεία που χρησιμοποιούνται στην εκπαίδευση. Αναφέρθηκε στο κεφάλαιο 1 ότι ένας νευρώνας χαρακτηρίζεται από μία συνάρτηση ενεργοποίησης, σε αυτή την παράγραφο θα παρουσιαστούν κάποιες από τις πιο σημαντικές και ευρέως χρησιμοποιούμενες συναρτήσεις.

α) Συνάρτηση κατωφλίου (Threshold Function): $\varphi(\upsilon)=\begin{cases}1 \ για \ \upsilon \geq 0 \\ 0 \ για \ \upsilon < 0\end{cases}$.

Στον τομέα της μηχανικής μάθησης το όνομα αυτής της είναι “όλα ή τίποτα” καθώς παίρνει την τιμή 1 για θετικό ή μηδέν δυναμικό ενεργοποίησης και 0 για αρνητικό. Επίσης συναντάται και ως McCulloch & Pits, εις αναγνώριση του πρωτοποριακού έργου τους, άλλωστε κατά αυτόν τον τρόπο συμβαίνει και η ενεργοποίηση των βιολογικών νευρωνικών δικτύων.

β) Σιγμοειδής συνάρτηση (Sigmoid Function): $\varphi(\upsilon)=\frac{1}{1+\exp(-\alpha\upsilon)}$.

Πήρε το όνομα της από την γραφική της παράσταση που έχει σχήμα “S”. Όπου α η κλίση της συνάρτησης. Είναι αυστηρώς αύξουσα και επιδεικνύει ισορροπία μεταξύ γραμμικής και μη γραμμικής συμπεριφοράς.

γ) Συνάρτηση προσήμου (Sign Fuction): $\varphi(\upsilon)=\begin{cases}1 \ για \ \upsilon > 0 \\ 0 \ για \ \upsilon = 0 \\ -1 \ για \ \upsilon < 0\end{cases}$.

Πρόκειται για επέκταση της συνάρτησης κατωφλίου χρησιμοποιείται στις περιπτώσεις που επιθυμείται ενεργοποίηση σε πεδίο τιμών με αρνητικές τιμές.

δ) Υπερβολική εφαπτομένη: $\varphi(\upsilon)=\tanh(\upsilon)$.

Πρόκειται για επέκταση της σιγμοειδής συνάρτησης χρησιμοποιείται στις περιπτώσεις που επιθυμείται ενεργοποίηση σε πεδίο τιμών με αρνητικές τιμές. [7]

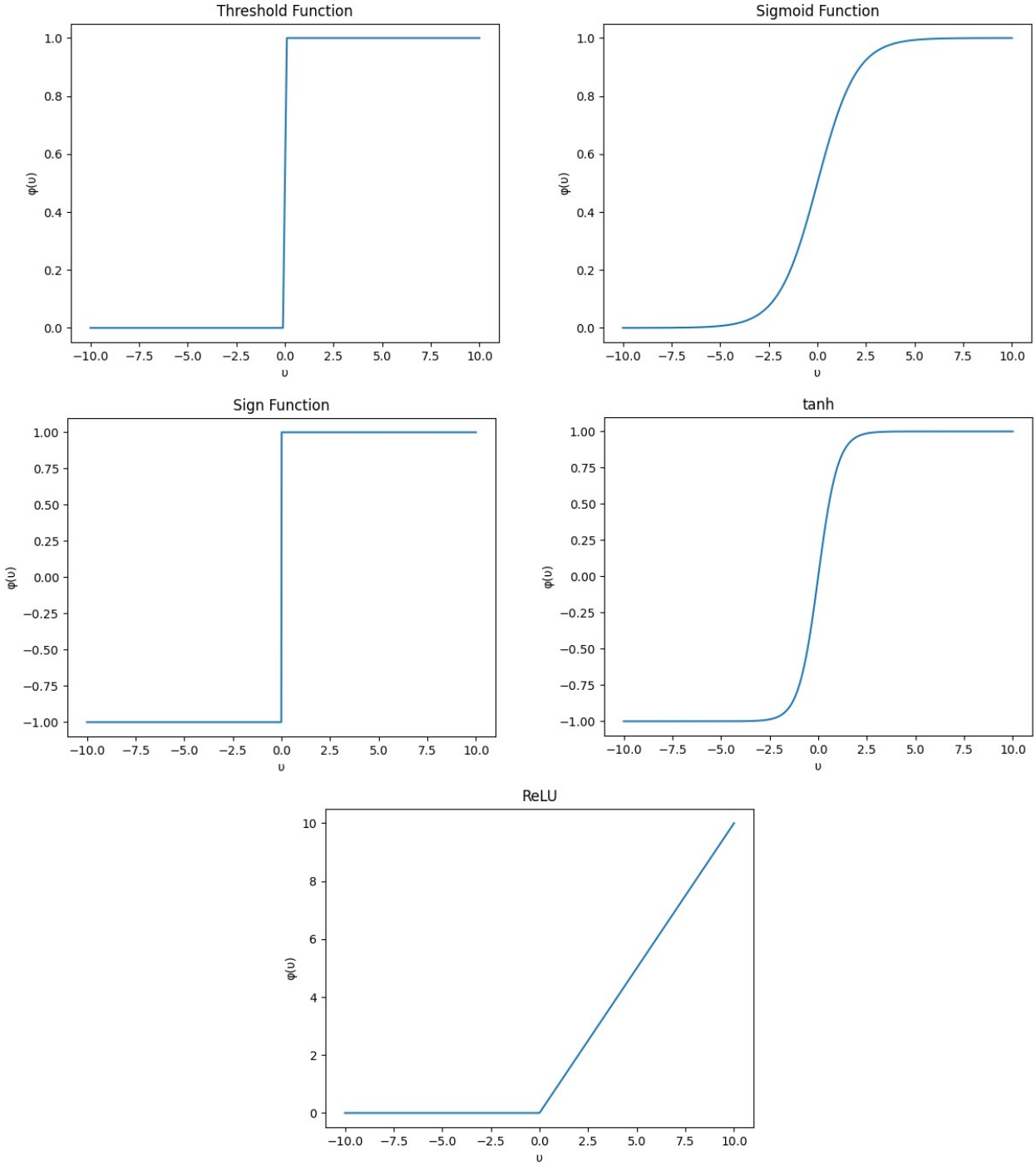
ε) Συνάρτηση διορθωμένης γραμμικής μονάδας (ReLU): $\varphi(\upsilon)=max(0,\upsilon)$.

Η συνάρτηση διορθωμένης γραμμικής μονάδας (Rectified Linear Unit Function- ReLU) πρόκειται για μία συνάρτηση που εξαναγκάζει σποραδικές ενεργοποιήσεις (η σιγμοειδής συνάρτηση δίνει τιμές κοντά στο 0, η ReLU μπορεί να δώσει ακριβώς τιμή 0) και έχει πολύ απλή παράγωγο. Υπάρχουν αρκετές προεκτάσεις της συγκεκριμένης συνάρτησης που μπορούν να χρησιμοποιηθούν.

στ) Κανονικοποιημένη εκθετική συνάρτηση (Softmax): $\varphi(\upsilon_i)=\frac{e^{\upsilon_i}}{\sum_{k=1}^{K} e^{\upsilon_k}}$.

Είναι η πιο συχνά χρησιμοποιούμενη συνάρτηση ενεργοποίησης του επιπέδου εξόδου. Αυτό συμβαίνει διότι αν θεωρηθεί ότι κάθε νευρώνας εξόδου αντιπροσωπεύει μία κλάση ταξινόμησης τότε η τιμή της κάθε εξόδου αντιπροσωπεύει την πιθανότητα η παρατήρηση να ανήκει στην κλάση αυτή. [29]

Στην Εικόνα 3.6 φαίνονται οι γραφικές αναπαραστάσεις των παραπάνω συναρτήσεων πλην της Softmax, μιας αναπαριστά πιθανότητα η γραφική απεικόνιση της δεν έχει κάποιο ιδιαίτερο ενδιαφέρον.



(Εικόνα 3.6) Γραφικές παραστάσεις των συναρτήσεων ενεργοποίησης (στην σιγμοειδής συνάρτηση α=1),

3.4.2.3 Εμπρόσθιος (forward) αλγόριθμος εκπαίδευσης

Από τα παραπάνω μπορεί να εισαχθεί εμπρόσθιος αλγόριθμος για την εκπαίδευση των MLP. Έστω O το διάνυσμα χαρακτηριστικών.

$v^0 \leftarrow O$

Για l από 1 εώς L με βήμα $l = l + 1$

$z^l = W^l v^{l-1} + b^l$

$v^l = \varphi(z^l)$

Αν η χρήση γίνεται για παλινδρόμηση (regression).

$v^L = Z^L$

Αλλιώς $v^L = softmax(z^L)$

Τέλος αλγορίθμου. [29]

3.4.2.4 Κριτήρια εκπαίδευσης και συναρτήσεις κόστους

Τα κριτήρια εκπαίδευσης πρέπει να είναι εύκολα στην επαλήθευση και άμεσα συσχετιζόμενα με τον στόχο και άρα η βελτιστοποίηση τους θα πρέπει να σημαίνει βελτιστοποίηση του συνολικού βαθμού επίδοσης. Ιδανικά οι παράμετροι των μοντέλων θα πρέπει να εκπαιδεύονται ώστε να ελαχιστοποιούν το αναμενόμενο κόστος (expected loss).

$$J_{CE}=E(J(W,b;o,y))=\int_{o} J(W,b;o,y)\,p(o)\,d(o) \quad .$$

Όπου $J(W,b;o,y)$ είναι η συνάρτηση κόστους, οι παράμετροι του μοντέλου W,b (βάρη και προκατάληψη αντίστοιχα), ο οι παρατηρήσεις εισόδου και y το διάνυσμα εξόδου. Η συνάρτηση πυκνότητας πιθανότητας p(o) των παρατηρήσεων ο είναι η άγνωστη συνάρτηση προς εκτίμηση από το σύνολο δεδομένων. Από τα παραπάνω μπορεί να συμπεραθεί ότι η συνάρτηση κόστους δεν μπορεί να ορισθεί για τον λόγο αυτό χρησιμοποιούνται εμπειρικά κριτήρια. Παρακάτω θα αναλυθούν κάποια από τα πιο σημαντικά.

1) Το πιο συχνά χρησιμοποιούμενο κριτήριο για παλινδρόμηση (regression) είναι αυτό του μέσου τετραγωνικού σφάλματος (mean square error – MSE).

$$J_{MSE}(W,b;S)=\frac{1}{M}\sum_{m=1}^{M} J_{MSE}(W,b;o^m,y^m)=\frac{1}{2}\|v^L-y\|^2=\frac{1}{2}(v^L-y)^T(v^L-y) \quad .$$

2) Για εφαρμογές ταξινόμησης (y η κατανομή πιθανότητας) χρησιμοποιείται συχνότερα το κριτήριο cross-entropy (CE).

$$J_{CE}(W,b;S)=\frac{1}{M}\sum_{m=1}^{M} J_{CE}(W,b;o^m,y^m)=-\sum_{i=1}^{C} y_i \log(v_i^L) \quad .$$

Το πρόβλημα αναγνώρισης φωνής είναι πρόβλημα ταξινόμησης αλληλουχιών, για τέτοιου τύπου θέματα έχουν αναπτυχθεί κριτήρια εκπαίδευσης (sequence descriminative training criteria) που λαμβάνουν υπόψη τους περισσότερους παράγοντες, για παράδειγμα στην αναγνώριση φωνής θα πρέπει να ληφθεί υπόψη το λεξικό, το γλωσσικό μοντέλο καθώς και οι περιορισμοί των HMM (λόγο της προσέγγισης ενσωμάτωσης των νευρωνικών δικτύων, θα αναλυθεί σε επόμενη παράγραφο του παρόντος κεφαλαίου). Τέτοιου τύπου κριτήρια είναι το κριτήριο ελαχίστου ρίσκου κατά Bayes (minimum Bayes error – MBE), ελαχίστου φωνηματικού λάθους (minimum phone error -MPE) κ.α.. Στην παρούσα διπλωματική θα αναλυθεί το κριτήριο μέγιστης αμοιβαίας πληροφορίας και η προέκταση αυτού, το κριτήριο ελευθέρου πλέγματος - μέγιστης αμοιβαίας πληροφορίας (lattice free – maximum mutual information - LF-MMI) που χρησιμοποιείται και ως κριτήριο εκπαίδευσης στην κατασκευή του συστήματος.

Το κριτήριο MMI είναι στενά συνδεδεμένο με την μείωση του αναμενόμενου ακολουθιακού λάθους, στόχος του είναι η αύξηση της μέγιστης αμοιβαίας πληροφορίας μεταξύ των παρατηρήσεων και των κατανομών (των φωνημάτων). Έστω $o^m=o_1^m,\dots,o_t^m,\dots o_{Tm}^m$ η αλληλουχία των παρατηρήσεων με Tm τον συνολικό αριθμό των frames στην έκφραση m και $w^m=w_1^m,\dots,w_t^m,\dots,w_{Nm}^m$ η σωστή αναπαράσταση σε κείμενο της m-οστής έκφρασης με Nm τον συνολικό αριθμό των λέξεων. Έστω ένα σύνολο εκπαίδευσης S={(o^m,w^m)|0<m<M}τότε για το κριτήριο MMI θα ισχύει:

$$J_{MMI}(\theta;S)=\sum_{m=1}^{M} J_{MMI}(\theta;o^m,w^m)=\sum_{m=1}^{M} logP(w^m|o^m;\theta)=\sum_{m=1}^{M}\log\frac{p(o^m|s^m;\theta)^{\kappa}P(w^m)}{\sum_{w} p(o^m|s^m;\theta)^{\kappa}P(w)} \quad .$$

Όπου θ οι παράμετροι των DNN όπως τα βάρη και οι προκαταλήψεις, $s^m=s_1^m,\dots,s_t^m,\dots s_{Tm}^m$ η αλληλουχία των καταστάσεων που ανταποκρίνονται στο w^m και κ ο παράγοντας ακουστικής κλίμακας. Θεωρητικά το άθροισμα στον παρονομαστή θα έπρεπε να παρθεί από όλες τις πιθανές εκφράσεις που υπάρχουν στο σύνολο δεδομένων. Στην πράξη για την μείωση των υπολογισμών χρησιμοποιείται ο γράφος (πλέγμα - lattice) αποκωδικοποίησης για την m-οστή έκφραση. Η παράγωγος του κριτηρίου που σχετίζεται με τις παραμέτρους του μοντέλου υπολογίζεται ως:

$$\nabla_{\theta} J_{MMI}(\theta;o^m,w^m)=\sum_{m}\sum_{t}\nabla_{z_{mt}^L} J_{MMI}(\theta;o^m,w^m)\frac{\partial z_{mt}^L}{\partial\theta}=\sum_{m}\sum_{t} e_{mt}^L\frac{\partial z_{mt}^L}{\partial\theta}$$

Όπου z_{mt}^{L} το αποτέλεσμα εφαρμογής της softmax στο frame t της έκφρασης m, e_{mt}^{L} το σήμα λάθους (error signal). Το σήμα λάθους υπολογίζεται ως:

$$e_{mt}^{L}(i)=\nabla_{z_{mt}^{L}(i)}J_{MMI}(\theta;o^{m},w^{m})=\sum_{r}\frac{\partial J_{MMI}(\theta;o^{m},w^{m})}{\partial\log p(o_{t}^{m}|r)}\frac{\partial p(o_{t}^{m}|r)}{\partial z_{mt}^{L}(i)}$$

$$=\sum_{r}\kappa(\delta(r=s^{m})-\frac{\sum_{w:s_{t=r}}p(o_{t}^{m}|s)^{\kappa}P(w)}{\sum_{w}p(o_{t}^{m}|s^{w})P(w)})x(\frac{\partial(logP(r|o_{t}^{m})-logP(r)+p(o_{t}^{m}))}{\partial z_{mt}^{L}(i)})$$

$$=\kappa(\delta(r=s^{m})-\frac{\sum_{w:s_{t=r}}p(o_{t}^{m}|s)^{\kappa}P(w)}{\sum_{w}p(o_{t}^{m}|s^{w})P(w)})\frac{\partial\log v_{mt}^{L}(r)}{\partial z_{mt}^{L}(i)}=\kappa(\delta(i=s^{m})-\frac{\sum_{w:s_{t=i}}p(o_{t}^{m}|s)^{\kappa}P(w)}{\sum_{w}p(o_{t}^{m}|s^{w})P(w)})\quad .$$

Όπου $e_{mt}^{L}(i)$ είναι το i-οστό στοιχείο του σήματος λάθους, $v_{mt}^{L}(r)=P(r|o_{t}^{m})$ r-οστή έξοδος του DNN, δ η κρουστική συνάρτηση, ο αφαιρέτης δηλαδή το κλάσμα της παραπάνω συνάρτησης αντιπροσωπεύει την a posteriori πιθανότητα το σύστημα να βρίσκεται στην κατάσταση r την χρονική στιγμή t και υπολογίζεται από τον παρονομαστή του J_{MMI} μέσω του γράφου (πλέγμα - lattice) αποκωδικοποίησης για την m-οστή έκφραση.

Η προέκταση του κριτηρίου MMI, LF-MMI χρησιμοποιεί και τον αριθμητή της εξίσωσης του J_{MMI}. Αυτό συμβαίνει διότι στο κριτήριο του απλού MMI γίνεται η θεώρηση ότι τα ακουστικά δεδομένα συνδέονται αυστηρά με τα κειμενικά δεδομένα, ενώ στην περίπτωση του LF-MMI γίνεται η παραδοχή του ότι όλες οι πιθανές καταστάσεις μίας αλληλουχίας μπορούν να οδηγήσουν στην w^{m}, [29] λόγο αυτής της παραδοχής τα αποτελέσματα βελτιώνονται αφού είναι πιθανό να χρησιμοποιείται διαφορετική προφορά για κάποια λέξη για παράδειγμα η λέξη at μπορεί να προφερθεί και ως /ae/ /t/ και ως /aa/ /t/. Άρα για την παράγωγο θα ισχύει:

$$\text{για}\quad \gamma(r)=\frac{\sum_{w:s_{t=r}}p(o_{t}^{m}|s)^{\kappa}P(w)}{\sum_{w}p(o_{t}^{m}|s^{w})P(w)}\quad\text{και}\quad \gamma(i)=\frac{\sum_{w:s_{t=i}}p(o_{t}^{m}|s)^{\kappa}P(w)}{\sum_{w}p(o_{t}^{m}|s^{w})P(w)}\quad\text{τότε:}$$

$$e_{mt}^{L}=\sum_{r}[\gamma(r)-\gamma(i)]\quad .$$

Στην πράξη κατασκευάζονται δύο γράφοι ένας (ακυκλικός) για τον αριθμητή που κωδικοποιεί την πληροφορία των παρατηρήσεων εισόδου και ένας για τον παρονομαστή που κωδικοποιεί την πληροφορία όλων των πιθανών εκφράσεων και είναι ο ίδιος ανεξαρτήτως εισόδου. Παρά το όνομα του το κριτήριο LF-MMI είναι ένα κριτήριο MMI που βασίζεται σε πλέγματα (lattice based). Οι γράφοι κατασκευάζονται ως FSA και οι τιμές τους παίρνονται από δυο περάσματα (ένα για κάθε γράφο) του forward-backword αλγορίθμου. [30][31]

Τέλος, θα πρέπει να σημειωθεί στο σημείο αυτό ότι αρκετές φορές στην προσπάθεια να το ελαχιστοποιηθεί το αναμενόμενο κόστος (loss function) μειώνεται το εμπειρικό κόστος που ορίζεται από το σύνολο δεδομένων το φαινόμενο αυτό ονομάζεται overfit. Ένας τρόπος για τον έλεγχο του overfit είναι η κανονικοποίηση (regularization) του κριτηρίου εκπαίδευσης ώστε οι παράμετροι του μοντέλου να μην ταιριάζουν (fit) τόσο καλά στα δεδομένα εκπαίδευσης. Ένας άλλος είναι η χρήση επιπέδων Dropout τα επίπεδα αυτά τυχαία παραλείπουν (από το επόμενο κρυφό επίπεδο) ένα ποσοστό από νευρώνες, κατά αυτόν τον τρόπο οι νευρώνες εξαρτώνται λιγότερο ο ένας από τον άλλο στην αναγνώριση προτύπων [29]. Φυσικά υπάρχουν και άλλοι τρόποι όπως, η αύξηση των δεδομένων εκπαίδευσης είτε άμεσα να εισαχθούν νέα δεδομένα με νέα πληροφορία είτε έμμεσα να εισαχθούν δεδομένα που προέρχονται από την επεξεργασία των ήδη υπαρχόντων,κ.α.

3.4.2.5 Αλγόριθμος Backpropagation

Αφού έχουν επεξηγηθεί τα κριτήρια εκπαίδευσης, μπορεί να γίνει η ανάλυση του πιο πολυχρησιμοποιούμενου αλγορίθμου εκπαίδευσης του Backpropagation (Bp), ο οποίος πηγάζει στον κανόνα αλυσίδας για τον υπολογισμό της παραγώγου. Οι παράγοντες του μοντέλου μπορούν

να βελτιωθούν με την χρήση της πρώτης παραγώγου ως:

$$w_{t+1}^l = w_t^l + \varepsilon \nabla w_t^l \quad \text{και} \quad b_{t+1}^l = b_t^l + \varepsilon \nabla b_t^l \quad \text{με}$$

$$\nabla w_t^l = \frac{1}{M_b} \sum_{m=1}^{M_b} \nabla_{w_t^l} J(W, b; o^m, y^m) \quad \text{και} \quad \nabla b_t^l = \frac{1}{M_b} \sum_{m=1}^{M_b} \nabla_{b_t^l} J(W, b; o^m, y^m) \quad .$$

Όπου t η επανάληψη (iteration), ε ο ρυθμός μάθησης (learning rate – lr), M_b ο αριθμός των δειγμάτων. Για S σύνολο δεδομένων με M δείγματα και L ο συνολικός αριθμός επιπέδων, O παρατηρήσεις, Y διανύσματα εξόδου, G η παράγωγος και φ η συνάρτηση ενεργοποίησης, ο αλγόριθμος Bp ορίζεται ως:

Τυχαία αρχικοποίηση για w_0^l, b_0^l ,0<l<L

Όσο δεν επιβεβαιώνεται το κριτήριο σταματημού επανέλαβε:

Επέλεξε ένα μικρό σύνολο (minibatch) από O,Y και M δείγματα.

Κάλεσε τον εμπρόσθιο (forward) αλγόριθμο (O)

$e_t^l = V_t^l - Y$

$G_t^l \leftarrow e_t^l$

Για l=L και όσο l>0 με βήμα l=l-1:

$\nabla w_t^l \leftarrow G_t^l (v_t^{l-1})^T$

$\nabla b_t^l \leftarrow G_t^l$

$w_{t+1}^l = w_t^l - \frac{\varepsilon}{M} \nabla w_t^l$

$b_{t+1}^l = b_t^l - \frac{\varepsilon}{M} \nabla b_t^l$

$e_t^{l-1} \leftarrow (w_t^l)^T G_t^l$

Αν l>1 τότε:

$G_t^{l-1} \leftarrow \varphi'(z_t^{l-1}) \circ e_t^{l-1}$ (Πολλαπλασιασμός πινάκων στοιχείο με στοιχείο)

Επέστρεψε DNN={ w^l, b^l },0<l<L

Το κριτήριο τερματισμού επιβεβαιώνεται είτε όταν τελειώσουν οι ορισμένες από τον κατασκευαστή επαναλήψεις (iterations), είτε όταν η βελτίωση του κριτηρίου εκπαίδευσης είναι μηδαμινή. Τέλος, ο ρυθμός μάθησης είναι είναι μία πολύ σημαντική παράμετρος στην κατασκευή DNN αφού επηρεάζει σημαντικά τις τιμές των παραμέτρων του μοντέλου. Η πιο σύνηθες στρατηγική που χρησιμοποιείται σχετίζει τον ρυθμό με το μέγεθος του συνόλου (batch), πιο συγκεκριμένα στις πρώτες επαναλήψεις επιλέγεται ένα μικρό batch ενώ ορίζεται από τον κατασκευαστή μία τιμή για lr, όσο προχωράν οι επαναλήψεις το μέγεθος του batch αυξάνει ενώ το lr μειώνεται αναλογικά, δηλαδή $lr_t = lr_0/\text{bach size}$. Αυτό συμβαίνει εμπειρικά αφού όσο περισσότερο μειώνεται η βελτίωση του κριτηρίου εκπαίδευσης τόσο λιγότερη πρέπει να είναι και η μεταβολή των παραμέτρων του μοντέλου. [29]

3.4.2.6 Τύποι επιπέδων και δικτύων DNN

Έχουν αναλυθεί ως τώρα όλα τα στοιχεία που είναι απαραίτητα για την κατασκευή ενός DNN καθώς και αλγόριθμοι εκπαίδευσης. Το τελευταίο πράγμα που μένει να αναφερθεί είναι οι τύποι των επιπέδων ενός DNN, που χωρίζονται βάσει της εργασίας που αποτελούν και χαρακτηρίζουν το DNN είτε ως υβριδικό όταν αυτό αποτελείται από διάφορους τύπους επιπέδων είτε με το όνομα του επιπέδου όταν αποτελούνται μόνο από έναν τύπο. Παρακάτω θα αναλυθούν οι τύποι των επιπέδων που χρησιμοποιούνται ευρέως στην αναγνώριση φωνής.

Επίπεδα συνέλιξης (convolutional layers): πρόκειται για έναν από τα πιο ευρέως χρησιμοποιούμενους τύπους επιπέδων, χρησιμοποιείται κυρίως σε εφαρμογές αναγνώρισης εικόνας, παρ' όλα αυτά εμφανίζεται στην βιβλιογραφία σε αρκετές εφαρμογές αναγνώρισης φωνής είτε σε συνδυασμό με άλλους τύπους επιπέδων είτε σε DNNs αποτελούμενα με μόνο συνελικτικά

επίπεδα (convolutional neural networks – CNN) [29]. Ένα συνελικτικό δίκτυο παρουσιάζει υψηλό βαθμό μη ευαισθησίας (ιδιότητα ως αμετάβλητο) κατά την διάρκεια του χώρου και του χρόνου και έχει τα εξής χαρακτηριστικά:

1) Εξαγωγή χαρακτηριστικών: Κάθε νευρώνας λαμβάνει συναπτικές εισόδους από ένα τοπικό δεκτικό πεδίο του προηγούμενου επιπέδου, υποχρεώνοντας το να εξάγει τοπικά χαρακτηριστικά. Αφού εξαχθεί ένα χαρακτηριστικό, η ακριβής θέση του γίνεται λιγότερο σημαντική, εφόσον διατηρείται η σχετική του θέση ως προς άλλα χαρακτηριστικά.
2) Αντιστοίχηση χαρακτηριστικών: Κάθε υπολογιστικό επίπεδο του δικτύου απαρτίζεται από πολλούς χάρτες χαρακτηριστικών (feature maps), με κάθε χάρτη χαρακτηριστικών να είναι στην μορφή ενός επιπέδου μέσα στο οποίο οι μεμονωμένοι νευρώνες περιορίζονται στο να μοιράζονται το ίδιο σύνολο συναπτικών βαρών. Το γεγονός αυτό έχει τα εξής επακόλουθα:
   α) Μείωση του αριθμού των ελεύθερων παραμέτρων, η οποία επιτυγχάνεται μέσω του διαμοιρασμού των βαρών.
   β) Μη ευαισθησία ως προς την μετατόπιση (shift invariance), η οποία επιβάλλεται στη λειτουργία ενός χάρτη χαρακτηριστικών μέσω της χρήσης μίας συνέλιξης με έναν πυρήνα (kernel) μικρού μεγέθους.
3) Υποδειγματοληψία: κάθε συνελικτικό επίπεδο ακολουθείται από ένα υπολογιστικό επίπεδο το οποίο εκτελεί τοπικό υπολογισμό και υποδειγματοληψία, δια των οποίων η ανάλυση του χάρτη χαρακτηριστικών μειώνεται. Η λειτουργία αυτή έχει ως αποτέλεσμα τη μείωση της ευαισθησίας της εξόδου του χάρτη χαρακτηριστικών στις μετατοπίσεις και άλλες μορφές παραμόρφωσης. [7]

Από τα παραπάνω για ένα συνελικτικό επίπεδο μπορεί να εξαχθεί ότι, έστω k τα κανάλια εισόδου και l τα κανάλια εξόδου, κάθε πυρήνας K_{kl} είναι ένας πίνακας που κινείται (stride) με βήμα S διαμήκος των οριζόντιων γραμμών και στηλών της εισόδου X, φιλτράροντας "τεμάχια" εισόδου μεγέθους που ισούται με το μέγεθος (διαστάσεις) του πυρήνα. Η πράξη φιλτραρίσματος μεταξύ του πυρήνα και των "τεμαχίων" δεν είναι η συνέλιξη αλλά μια πράξη που ονομάζεται cross-correlation η διαφορά της με την συνέλιξη φαίνεται στον παρακάτω τύπο,

$$X * K = X\,(cross-correlation)\,rotate_{180^\circ}[K] \quad ,$$

η πράξη αυτή ορίζεται ως το άθροισμα του πολλαπλασιασμού πινάκων στοιχείο με στοιχείο. Από τα παραπάνω μπορεί να ορισθεί η έξοδος ενός συνελικτικού νευρώνα ως:

$$Y_i = b_i + \sum_{j=1}^{k} X_j \circ K_{ij} \quad \text{, όπου b η προκατάληψη.[29][32]}$$

Επίσης,αν λάβουμε υπόψη το βήμα κίνησης S, δημιουργείται θέμα στις περιπτώσεις που ο πίνακας εισόδου έχει διαστάσεις που δεν είναι ακέραια πολλαπλάσια των διαστάσεων του πυρήνα. Για την επίλυση αυτού του θέματος υπάρχουν διάφορες προσεγγίσεις με τις πιο σύνηθες είναι no padding και η zero padding.

1) Στην no padding προσέγγιση, ο πυρήνας ξεκινάει από την μία γωνία τις εισόδου και σταματά όταν φτάσει στην άκρη του πίνακα,όπως στο παρακάτω παράδειγμα για $S_c=S_r=1$.

$$\begin{vmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{vmatrix} \circ \begin{vmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{vmatrix} = \begin{vmatrix} X_{11}K_{11}+X_{12}K_{12}+X_{21}K_{21}+X_{22}K_{22} & X_{12}K_{11}+X_{13}K_{12}+X_{22}K_{21}+X_{23}K_{22} \\ X_{21}K_{11}+X_{22}K_{12}+X_{31}K_{21}+X_{32}K_{22} & X_{22}K_{11}+X_{23}K_{12}+X_{32}K_{21}+X_{33}K_{22} \end{vmatrix}$$

Οι παρατηρήσεις που μπορούν να εξαχθούν είναι ότι κάποια στοιχεία του πίνακα εισόδου χρησιμοποιούνται περισσότερο από άλλα αυτό έχει σαν αποτέλεσμα η πληροφορία που περιέχουν να "τονίζεται" περισσότερο στην έξοδο, και μία ακόμα παρατήρηση έχει να κάνει με την διάσταση της εξόδου η οποία χαρακτηρίζεται από τον εξής τύπο:

$$Y[r,c] = \left[\frac{X_r - K_r}{S_r} + 1, \frac{X_c - K_c}{S_c} + 1\right] \quad ,$$

(οι δείκτες r και c αντιστοιχούν στις γραμμές και τις στήλες αντίστοιχα).

2) Στην zero padding προσέγγιση, προστίθενται διαστάσεις στις διαστάσεις του πίνακα εισόδου, αυτές οι διαστάσεις γεμίζουν με μηδενικά, όπως στο παρακάτω παράδειγμα για $S_c=S_r=1$.

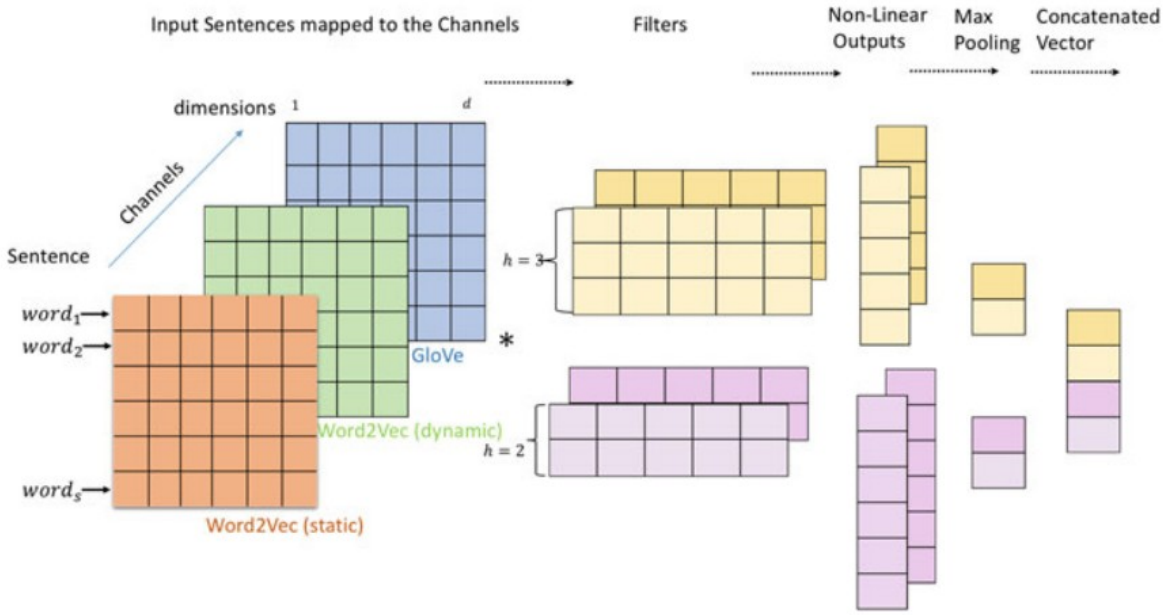
$$\begin{vmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & X_{11} & X_{12} & X_{13} & 0 \\ 0 & X_{21} & X_{22} & X_{23} & 0 \\ 0 & X_{31} & X_{32} & X_{33} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{vmatrix} \circ \begin{vmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{vmatrix} =$$

$$\begin{vmatrix} X_{11}K_{22} & X_{11}K_{21}+X_{12}K_{22} & X_{12}K_{21}+X_{13}K_{22} & X_{13}K_{22} \\ X_{11}K_{11}+X_{21}K_{22} & X_{11}K_{11}+X_{12}K_{12}+X_{21}K_{21}+X_{22}K_{22} & X_{12}K_{11}+X_{13}K_{12}+X_{22}K_{21}+X_{23}K_{22} & X_{13}K_{11}+X_{23}K_{21} \\ X_{21}K_{12}+X_{31}K_{22} & X_{21}K_{11}+X_{22}K_{12}+X_{31}K_{21}+X_{32}K_{22} & X_{22}K_{11}+X_{23}K_{12}+X_{32}K_{21}+X_{33}K_{22} & X_{23}K_{11}+X_{33}K_{21} \\ X_{31}K_{12} & X_{31}K_{11}+X_{32}K_{12} & X_{32}K_{11}+X_{33}K_{12} & X_{33}K_{11} \end{vmatrix}$$

Σε αυτή την περίπτωση, όλα τα δεδομένα του πίνακα εισόδου “επεξεργάζονται” από τον πυρήνα εξίσου. Οι διαστάσεις του πίνακα εξόδου αυξάνονται κατά τον εξής τύπο:

$$Y[r,c]=[\frac{X_r-mod(K_r,2)}{S_r}+1, \frac{X_c-mod(K_c,2)}{S_c}+1] \quad .$$

Τέλος, για τα επίπεδα υπολογισμού και δειγματοληψίας (pooling layers) και σε αυτά υπάρχει ένας πυρήνας που “κινείται” στα δεδομένα εισόδου του επιπέδου με ένα βήμα S. Σε αυτήν την περίπτωση όμως ο πυρήνας επιλέγει ή υπολογίζει ένα στοιχείο από αυτά που “βλέπει” (ένα για κάθε μετατόπιση). Υπάρχουν διάφοροι τρόποι επιλογής που αντίστοιχα με τον υπολογισμό χαρακτηρίζονται. Οι πιο συχνά χρησιμοποιούμενους είναι, ο max pooling που επιλέγει το στοιχείο με την μεγαλύτερη τιμή και ο avarage pooling που υπολογίζει τον μέσο όρο από τις τιμές που “βλέπει” και προωθεί αυτή την τιμή στην έξοδο. [29]



(Εικόνα 3.7): Παράδειγμα χαρτογράφησης με CNN, κειμενικών δεδομένων .[32]

Επίπεδα χρονικής καθυστέρησης (time delay layers): πρόκειται για έναν από τα πιο ευρέως χρησιμοποιούμενους τύπους επιπέδων στον τομέα της αναγνώρισης φωνής αλλά και γενικότερα σε εφαρμογές χρονικών σειρών (time series). Συναντώνται είτε σε συνδυασμό με άλλους τύπους επιπέδων είτε σε DNNs αποτελούμενα με μόνο τέτοιου τύπου επίπεδα (time delay neural networks – TDNN). Τα TDNNs αποτελούν ειδική κατηγορία CNNs ως μονοδιάστατα CNNs [33]. Ένα TD επίπεδο δέχεται ως είσοδο διανύσματα διάστασης D (frames), κάθε νευρώνας επεξεργάζεται από ένα frame, ενώ μαζί με τους διπλανούς του επεξεργάζονται ένα συνεχές παράθυρο από frames. Ουσιαστικά η πληροφορία που επεξεργάζονται οι νευρώνες συνδέεται. Η επιθυμητή πληροφορία προς προώθηση σε έναν νευρώνα του επόμενου επιπέδου είναι η πληροφορία του συνεχούς παραθύρου, εδώ εισάγεται και το στοιχείο καθυστέρησης δηλαδή το delay offset [d_1,d_2] (συνήθως [-d,d] με $d \in \mathbb{Z}$), ονομάζεται έτσι καθώς για να γίνει η προώθηση θα πρέπει να τελειώσουν την

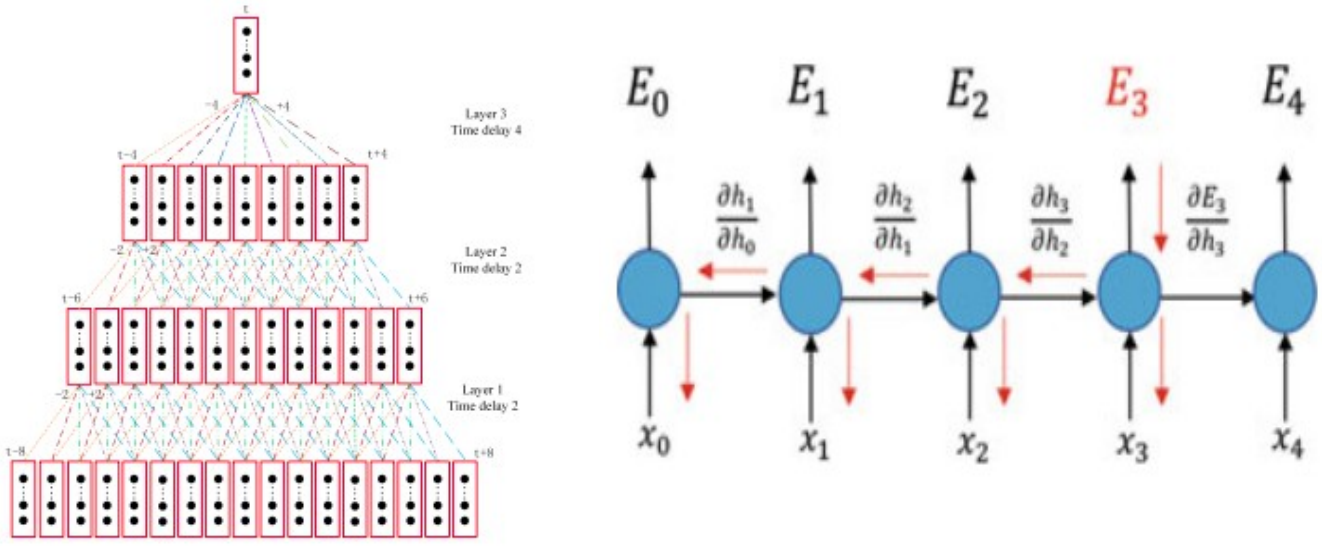
επεξεργασία τους d_2-d_1+1 νευρώνες (μέγεθος παραθύρου). Μπορεί να συμπεραθεί ότι το delay offset δρα σε ένα σύνολο διανυσμάτων (π.χ. μία έκφραση που κάθε λέξη αποτελεί ένα παράθυρο και κάθε φώνημα ένα frame) όπως το stride στα CNN σε έναν πίνακα. Σε αυτό το σημείο μπορεί να ορισθεί το παράθυρο εισόδου V_t και η σειρά εισόδου (σύνολο διανυσμάτων) s, για N αριθμό διανυσμάτων και $W=d_2-d_1$, ως:

$$V_t=[V_{t-W}, V_{t-W+1}\dots, V_t, \dots, V_{t+W-1}, V_{t+W}] \quad , \quad s=[V_1,..,V_N] \quad \text{με} \quad s\in \mathbb{R}^{Nx(WxD)} \quad .$$

Όμοια με τα CNNs όταν δεν υπάρχουν τιμές για πριν ή μετά (π.χ. πρώτη ή τελευταία τιμή του διανύσματος αντίστοιχα) εφαρμόζεται padding. Σε αυτό το σημείο αξίζει να αναφερθεί ότι το επίπεδο εισόδου σε εφαρμογές NLP συχνά αναφέρεται και ως επίπεδο ενσωμάτωσης (embedding layer). Τα TD επίπεδα δρουν ως φίλτρα (πράξη συνέλιξης) λόγο του ότι είναι μονοδιάστατα δεν υπάρχει ανάγκη για pooling.[34]

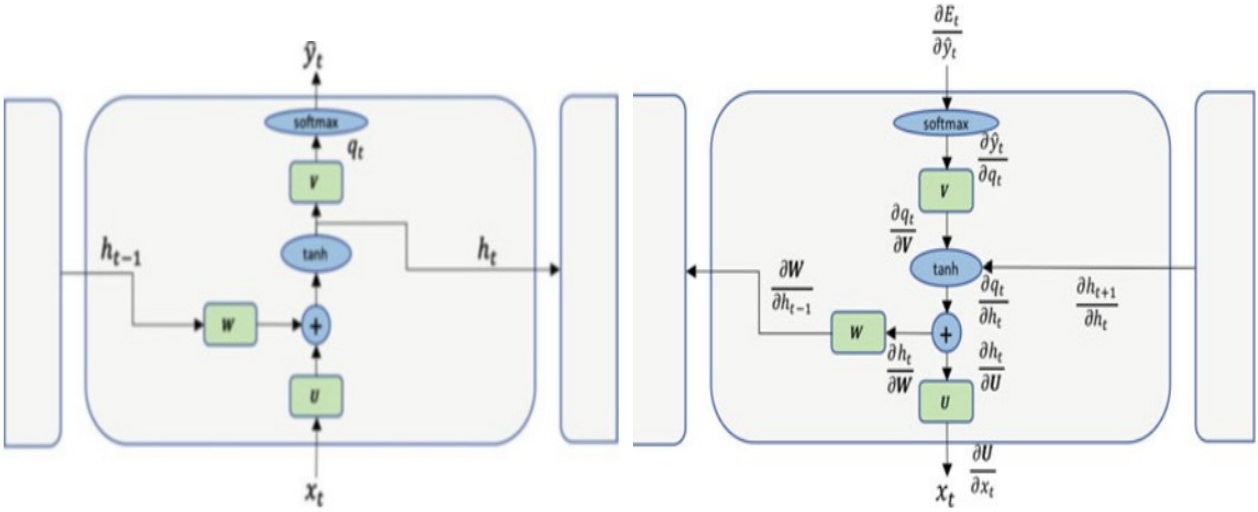
Αναδρομικά νευρωνικά δίκτυα (recurrent neural networks – RNN): ένα αναδρομικό δίκτυο χαρακτηρίζεται από τουλάχιστον μία ανάδραση, δηλαδή από μία σύνδεση μεταξύ των νευρώνων του αναδρομικού επιπέδου, αυτό σημαίνει ότι η τιμή ενός νευρώνα του επιπέδου την χρονική στιγμή t επηρεάζει την τιμή του διπλανού του την χρονική στιγμή t+1. Τέτοιου τύπου δίκτυα έχουν αποδειχθεί πολύ αποτελεσματικά, κυρίως σε εφαρμογές που σχετίζονται με την πρόβλεψη σειρών (sequence prediction). Ο σύνδεσμος ανάδρασης παρέχει μία μορφή μνήμης ή περιεχομένου που κωδικοποιεί την προηγούμενη επεξεργασία και πληροφορεί για τις αποφάσεις που έχουν παρθεί στους διπλανούς νευρώνες (hidden states -hs). Αυτή η πρόσθεση των δεδομένων με τα διανύσματα εισόδου φαντάζει αύξηση της πολυπλοκότητας στην πράξη όμως το μόνη σημαντική αλλαγή είναι η προσθήκη νέων βαρών που αντιστοιχούν στον σύνδεσμο ανάδρασης και θα συμβολίζονται ως U και V που αντιστοιχούν στην έξοδο. Στην εκπαίδευση του δικτύου τα παραπάνω αντανακλώνται ως εξής:

1) Για τον εμπρόσθιο αλγόριθμο, έστω h_t η τιμή του hs που καταλήγει ο κόμβος της ανάδρασης και h_{t-1} η τιμή του hs που ξεκινά ο κόμβος της ανάδρασης, θα ισχύει:
$$z_t=Uh_{t-1}+Wx_t+b_t \quad , \quad y_t=\varphi(Vh_t) \quad .$$
Το V συμβολίζει βάρη που μαθαίνουν να μετασχηματίζουν το μέγεθος της διάστασης εξόδου του hs.
2) Για τον αλγόριθμο backpropagation, θα πρέπει να υπολογισθούν οι παράγωγοι της συνάρτησης κόστους για όλα τα χρονικά βήματα, αυτό σημαίνει ότι θα πρέπει να αποθηκευτούν οι τιμές των κρυφών επιπέδων ώστε να επαναχρησιμοποιηθούν στα επόμενα βήματα. Αυτός ο νέος



(Εικόνα 3.8):(Αριστερά) σχηματική απεικόνιση TDNN [34],(Δεξιά) σχηματική απεικόνιση βημάτων Αλγορίθμου backpropagation through time [32].

αλγόριθμος ονομάζεται backpropagation through time (BPTT). Στην δεξιά εικόνα 3.8 φαίνονται με κόκκινα βέλη τα βήματα που θα πρέπει να ακολουθηθούν για τον υπολογισμό του λάθους στην τρίτη επανάληψη E_3.



(Εικόνα 3.9):(Αριστερά) υπολογισμός των παραμέτρων εμπρόσθιου αλγορίθμου για ένα hs. (Δεξιά) υπολογισμός των παραγώγων Backpropagation για ένα hs. (έχουν θεωρηθεί U τα βάρη εισόδου και W τα βάρη ανάδρασης).[32]

Υπάρχουν αρχιτεκτονικές RNN επιπέδων πιο σύνθετες όπως η bidirectional RNN (bRNN), στην συγκεκριμένη αρχιτεκτονική η ανάδραση μεταδίδεται και στις δύο κατευθύνσεις. Υπάρχουν δύο επίπεδα από hs, κάθε είσοδος x_i τροφοδοτείται στις αντίστοιχες καταστάσεις των δύο επιπέδων $h^i_{t->t+1}$ και $h^i_{t->t-1}$ οι δύο έξοδοι των hidden states δημιουργούν ένα τελικό διάνυσμα βάση κάποιας μεθόδου (π.χ. άθροιση, μέσος όρος, συνένωση κ.α.). Στο σημείο αυτό θα πρέπει να αναφερθεί το πρόβλημα που δημιουργείται στα RNN δίκτυα, δηλαδή ότι μετά από μερικές επαναλήψεις παρατηρείται είτε η εκθετική εκτόξευση της τιμής της παραγώγου είτε η εκθετική συρρίκνωση αυτής. Αυτό συμβαίνει επειδή στον υπολογισμό της παραγώγου συνεισφέρουν αντιστρόφως ανάλογα τα βάρη των προηγούμενων βημάτων. Υπάρχουν διάφοροι τρόποι για την επίλυση αυτού του προβλήματος όπως η προσεκτική επιλογή των παραμέτρων ή η χρήση κατάλληλου optimizer (π.χ. Adam), ο πιο βέβαιος όμως τρόπος επίλυσης είναι η χρήση κρυφών μονάδων μακράς βραχυπρόθεσμης μνήμης (Long short-term memory LSTM).[19][32]

Μονάδες μακράς βραχυπρόθεσμης μνήμης (Long short-term memory LSTM): διαιρούν το πρόβλημα της παραγώγου σε δύο υποπροβλήματα, σε ένα αφαίρεσης περιεχομένου που δεν χρειάζεται πλέον και σε ένα προσθήκης περιεχομένου που θα χρειαστεί για το μελλοντικό πάρσιμο αποφάσεων. Το κλειδί για την επίλυση των παραπάνω υποπροβλημάτων είναι η χρήση νευρωνικών μονάδων που δρουν ως πύλες ελέγχοντας την ροή της πληροφορίας εντός και εκτός των hs. Οι πύλες έχουν κοινό σχεδιασμό, κάθε μία αποτελείται από ένα εμπρόσθιο επίπεδο, που ακολουθείται από μία σιγμοειδής συνάρτηση ενεργοποίησης (δράση ως δυική μάσκα) το αποτέλεσμα πολλαπλασιάζεται με την τιμή του επιπέδου που ελέγχεται. Οι πύλες των LSTMs είναι:

1) Forget gate, στόχος της είναι η διαγραφή του περιεχομένου που δεν χρειάζεται πλέον. Οι πράξεις που επιτελεί είναι η εξής, όπου f_t το εμπρόσθιο επίπεδο της πύλης και c το διάνυσμα του περιεχομένου που μοιράζονται τα επίπεδα:

$$f_t = sigmoid(U_f h_{t-1} + W_f x_t) \quad \text{και} \quad k_t = c_{t-1} \circ f_t \quad .$$

Έπειτα, υπολογίζεται η πληροφορία που πρέπει να εξαχθεί από το προηγούμενο hs και την τωρινή είσοδο:

$$g_t = \tanh(U_g h_{t-1} + W_g x_t) \quad .$$

2) Add gate, στόχος της είναι η επιλογή νέου περιεχομένου για μελλοντικούς υπολογισμούς. Οι πράξεις που επιτελεί είναι η εξής, όπου i_t το εμπρόσθιο επίπεδο της πύλης:

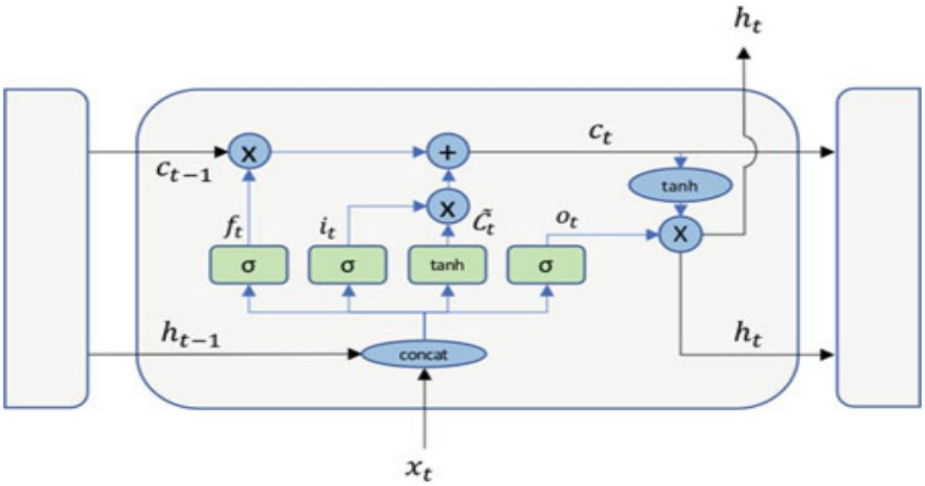
$$i_t = sigmoid(U_i h_{t-1} + W_i x_t) \quad \text{και} \quad j_t = g_t \circ i_t \quad .$$

Έπειτα, υπολογίζεται το νέο διάνυσμα περιεχομένου:

$$c_t = j_t \circ k_t \quad .$$

3) Πύλη εξόδου (output gate), στόχος της είναι η επιλογή της πληροφορίας που θα πρέπει να εξαχθεί από αυτό το hs (σε αντίθεση με την προηγούμενη πύλη που επιλέγει την πληροφορία που θα πρέπει να διατηρηθεί για μελλοντικούς υπολογισμούς):

$$o_t = sigmoid(U_o h_{t-1} + W_o x_t) \quad \text{και} \quad h_t = o_t \circ \tanh(c_t) \quad . [19]$$



(Εικόνα 3.10) Διαγραμματική απεικόνιση LSTM.[32]

3.4.3 Ενσωμάτωση DNN στο πρόβλημα αναγνώρισης φωνής

Στην προηγούμενη παράγραφο παρουσιάστηκαν τα βαθειά νευρωνικά δίκτυα (DNN) που χρησιμοποιούνται σε αρκετές εφαρμογές όπως η αναγνώριση φωνή. Στην παρούσα παράγραφο θα αναλυθεί η ενσωμάτωση των δικτύων στην αναγνώριση φωνής, τόσο για το γλωσσικό όσο και για το ακουστικό μοντέλο.

Ξεκινώντας από το γλωσσικό μοντέλο, οι λέξεις χρησιμοποιούνται ως μία χρονοσειρά. Εισάγονται σε ένα επίπεδο ενσωμάτωσης (embedding layer), που αποτελείται από έναν πίνακα, ως διανύσματα που περιέχουν 1 στον αύξοντα αριθμό της λέξης στο λεξικό και 0 αλλού (ετικετοποίηση).Στο επίπεδο αυτό κωδικοποιούνται ως ένα διάνυσμα μικρότερης διάστασης (που απεικονίζει μία έκφραση ή ένα τμήμα αυτής) και αυτό δίνεται ως είσοδο στο νευρωνικό δίκτυο [19]. Χρησιμοποιούνται κυρίως RNN και TDNN για την κατασκευή τέτοιων μοντέλων, γιατί όπως αναφέρθηκε και παραπάνω παρουσιάζουν υψηλή αποδοτικότητα σε εφαρμογές χρονοσειρών, αφού λαμβάνουν υπόψη για κάθε υπολογισμό και τις τιμές των διπλανών hs.

Για το ακουστικό μοντέλο υπάρχουν δύο τρόποι ενσωμάτωσης των DNN και οι δύο σχετίζονται με τον τρόπο που συνδυάζεται ένα DNN με ένα HMM, αυτοί είναι:

1) Υβριδική DNN-HMM προσέγγιση. Το σήμα φωνής μοντελοποιείται με HMM, ενώ οι πιθανότητες των παρατηρήσεων προσεγγίζονται με χρήση DNN. Κάθε νευρωνική έξοδος εκπαιδεύεται στον υπολογισμό της a posteriori πιθανότητας της συνεχούς κατανομής των καταστάσεων των HMM για είσοδο τις δοσμένες ακουστικές παρατηρήσεις. Κατά αυτόν τον τρόπο η νέα εξίσωση του ακουστικού μοντέλου είναι:

$$p(x|w) = \sum_q p(x|q,w)\, p(q|w) \approx max\ \pi(q_0) \prod_{t=1}^{T} a_{q_{t-1}q_t} \prod_{t=0}^{T} p(q_t|x_t) / p(q_t) \quad .$$

Όπου η $p(q_t|x_t)$ υπολογίζεται από DNN, $p(q_t)$ η a priori πιθανότητα της κατάστασης q_t ($q_t = T_q/T$ όπου T ο συνολικός αριθμός των frames και T_q ο συνολικός αριθμός των frames που εμφανίζεται η κατάσταση q_t), $\pi(q_0)$ η αρχική κατάσταση και $a_{q_{t-1}q_t}$ ο πίνακας μετάβασης οι παράγοντες αυτοί υπολογίζονται από το HMM. Η παραπάνω εξίσωση προκύπτει λόγω της θεώρησης εξάρτησης περιεχομένου (condex dependent - CD) φωνημάτων, δηλαδή για c_j κλάση ταξινόμησης περιεχομένου και s_i κατάσταση ανεξάρτητου φωνήματος (από το περιεχόμενο) θα ισχύει:

$$p(s_i, c_j|x_t) = p(s_i|x_t)\, p(c_i|s_j, x_t) \quad \text{ή} \quad p(s_i, c_j|x_t) = p(c_i|x_t)\, p(s_i|c_j, x_t) \quad .$$

2) Προσέγγιση σύμπτυξης DNN-HMM. Κατασκευάζεται ένα DNN με στόχο την επαύξηση (augmentation) των δεδομένων εισόδου του GMM. Πιο συγκεκριμένα εξάγονται τα διανυσματικά δεδομένα με την μικρότερη διαστατικότητα από το κρυφό επίπεδο bottleneck (ως bottleneck επίπεδο χαρακτηρίζεται το επίπεδο με τους λιγότερους νευρώνες, συνήθως είναι το επίπεδο πριν το επίπεδο εξόδου) του DNN και ενσωματώνονται στα διανύσματα χαρακτηριστικών, δημιουργώντας ένα νέο διάνυσμα χαρακτηριστικών με διάσταση ίση με το άθροισμα των διαστάσεων των δεδομένων (bottleneck και MFCC) έπειτα εφαρμόζεται LDA στα νέα διανύσματα και δίνονται ως είσοδο στο GMM. Η προσέγγιση αυτή προέκυψε από την παρατήρηση ότι το συγκεκριμένο επίπεδο του δικτύου επιβάλει περιορισμούς και εξαναγκάζει την πληροφορία που σχετίζεται με την ταξινόμηση σε αναπαραστάσεις μικρότερης διαστατικότητας. Το συγκεκριμένο επίπεδο συχνά χρησιμοποιείται ως αυτοκωδικοποιητής (autoencoder) δηλαδή εκπαιδεύεται με σκοπό την πρόβλεψη των δεδομένων εισόδου. [29]

Η επιλογή της προσέγγισης εξαρτάται από τον κατασκευαστή και την εφαρμογή, καμία απ' της δύο δεν υπερτερεί της άλλης πριν την κατασκευή και την αξιολόγηση του μοντέλου. Στην βιβλιογραφία εμφανίζεται συχνότερα η υβριδική προσέγγιση πιθανώς για λόγους διευκόλυνσης του κατασκευαστή όπως η επιλογή διαστάσεων (οι εκφράσεις στο σύνολο εκπαίδευσης δεν έχουν ίδιο αριθμό λέξεων), οι τεχνικές βελτιστοποίησης (έχουν αναπτυχθεί αρκετές τεχνικές που μπορούν να βελτιώσουν την αποτελεσματικότητα), το επίπεδο bottleneck ανήκει στα κρυφά επίπεδα του DNN γεγονός που καθιστά μη ορατό από την διαδικασία εκπαίδευσης (λόγω αυτού συχνά εξαναγκάζεται από τον κατασκευαστή το επίπεδο πριν την έξοδο να μετατραπεί σε επίπεδο bottleneck) και κυρίως για λόγους κατανόησης καθώς η υβριδική προσέγγιση αποτελείται από στάδια που ελέγχονται πιο εύκολα για τυχόν αστοχίες.

4 Πειραματικό Μέρος

4.1 Σύνολο Δεδομένων (Dataset)

Για την εκπόνηση της παρούσας διπλωματικής εργασίας επιλέχθηκε το σύνολο δεδομένων ChiME-5 καθώς ανταποκρίνεται σε μεγάλο βαθμό στους περιορισμούς που θέσαμε στον στόχο (Κεφ.2.3.) . Το συγκεκριμένο σύνολο δημιουργήθηκε για τις ανάγκες της πέμπτης πρόκλησης διαχωρισμού και αναγνώρισης ομιλίας της σειράς CHiME, που έχει ως στόχο την ανάπτυξη εύρωστων αυτόματων συστημάτων αναγνώρισης ομιλίας με μικρόφωνα διαποστατικής συστοιχίας σε περιβάλλον πραγματικού σπιτιού , προωθώντας έτσι την έρευνα στη διεπαφή επεξεργασίας λόγου και γλώσσας, στη ψηφιακή επεξεργασία σήματος και στη μηχανική μάθηση.

Για την συλλογή των δεδομένων πραγματοποιήθηκαν είκοσι ξεχωριστά δείπνα σε πραγματικά σπίτια. Στο κάθε δείπνο έλαβαν μέρος τέσσερα άτομα, δύο που είχαν τον ρόλο του οικοδεσπότη και δύο που είχαν τον ρόλο του καλεσμένου. Οι συμμετέχοντες είναι φίλοι μεταξύ τους και η οδηγία που τους δόθηκε είναι να συμπεριφέρονται όσο πιο φυσικά μπορούν. Το κάθε δείπνο αποτελούνταν από τρία διαφορετικά στάδια:
1) Την προετοιμασία του φαγητού στην κουζίνα.
2) Το δείπνο στην τραπεζαρία.
3) Το μετά το δείπνο στο σαλόνι.
Στους συμμετέχοντες επιτρέπονταν οι μετακινήσεις ανάμεσα στους χώρους καθώς και να συζητήσουν για όποιο θέμα επιθυμούν . Οι περιορισμοί που τους επιβλήθηκαν ήταν οι εξής: το δείπνο δεν θα έπρεπε να κρατήσει το λιγότερο δύο ώρες, εκ των οποίων θα περνούσαν τουλάχιστον τριάντα λεπτά στον κάθε χώρο, ενώ απαγορεύεται η χρήση τηλεόρασης και μουσικής με στόχο την αποφυγή θεμάτων που αφορούν πνευματικά δικαιώματα.

Για την ηχογράφηση του κάθε δείπνου χρησιμοποιήθηκαν συνολικά δέκα μικρόφωνα. Έξι μικρόφωνα τοποθετήθηκαν βέλτιστα ανά δύο σε κάθε χώρο, το κάθε ένα από αυτά έχει ένα γραμμικό διάνυσμα τεσσάρων σύγχρονων δειγματοληπτικών καναλιών , ενώ τέσσερα αμφιωτικά μικρόφωνα τοποθετήθηκαν στους συμμετέχοντες ένα στον καθένα. Τα αμφιωτικά μικρόφωνα είναι μικρόφωνα που τοποθετούνται στα αυτιά με στόχο να αποτυπώνουν των ήχο βάση της θέσης του ατόμου στον χώρο αλλά και της γεωμετρίας του έξω αυτιού (πτερύγιο και ακουστικό κανάλι), κάθε ένα τέτοιου τύπου μικρόφωνο αποτελείται από δύο κανάλια. Τα Μικρόφωνα με το γραμμικό διάνυσμα δεν είναι τέλεια συγχρονισμένα μεταξύ τους. Οι ηχογραφήσεις από τα αμφιωτικά μικρόφωνα χρησιμοποιήθηκαν για την μετατροπή ήχου σε κείμενο ανά χρονική στιγμή. Στο σημείο αυτό θα πρέπει να αναφερθεί ότι σε καμία ηχογράφηση δεν υπάρχει επικάλυψη στον λόγο κάποιου από τους ομιλητές . Για την παρούσα διπλωματική εργασία χρησιμοποιήθηκε μόνο η πληροφορία που πάρθηκε από τα αμφιωτικά μικρόφωνα μιας και ο όγκος των δεδομένων ήταν αρκετά μεγάλος.

Ελήφθησαν είκοσι ηχογραφήσεις δείπνων μεγέθους 114.3 GB εκ των οποίων χρησιμοποιήθηκαν 46.2 GB. Δεκαέξι δείπνα χρησιμοποιήθηκαν ως σύνολο εκπαίδευσης μεγέθους 37.3 GB,οι ηχογραφήσεις εκπαίδευσης έχουν με συνολική διάρκεια 40 ώρες και 33 λεπτά. Ενώ τέσσερα δείπνα ως σύνολο ελέγχου μεγέθους 8.9GB. Τα δύο από αυτά χρησιμοποιήθηκαν ως σύνολο ανάπτυξης με συνολική διάρκεια 4 ώρες και 27 λεπτά, τα υπόλοιπα δύο χρησιμοποιήθηκαν ως σύνολο επαλήθευσης με διάρκεια 5 ώρες και 12 λεπτά. [35]

4.1.1 Επιλογή δεδομένων

Από τις εκφράσεις που ελήφθησαν αφαιρέθηκαν οι εκφράσεις που αποτελούνταν καθαρά από θόρυβο, γέλιο, εκφράσεις που υπάρχουν ψίθυροι ή ήχοι που δεν μπορεί να τους αναγνωρίσει άνθρωπος και εκφράσεις που έχουν παραληφθεί από τους δημιουργούς του συνόλου για άγνωστους προς τους χρήστες λόγους. Ο λόγος που συνέβη αυτό είναι διότι όλες οι παραπάνω περιπτώσεις μπορούν να θεωρηθούν ως τυχαίος θόρυβος που θα μπορούσε να μοιάζει με φωνή είτε να μην έχει καμία απολύτως σχέση. Επίσης οι εκφράσεις που έχουν ετικετοποιηθεί από τον δημιουργό ως μία

από τις παραπάνω κατηγορίες, η μεταξύ τους ηχητική υπόσταση να μην έχει καμία σχέση (π.χ. στην κατηγορία θόρυβος να υπάρχει το σπάσιμο κάποιου πιάτου και ο ήχος ενός αεροπλάνου).

4.1.2 Επαύξηση δεδομένων (Data augmentation)

Δεν έχει υπάρξει κάποια επαύξηση δεδομένων από πλευράς εκπόνησης της εργασίας καθώς έχει γίνει η θεώρηση του ότι τα δεδομένα έχουν παρθεί επαυξημένα αφού για κάθε έκφραση υπάρχουν τέσσερεις ηχογραφήσεις που διαφέρουν ανάλογα την θέση του κάθε συμμετέχοντα στον χώρο αλλά και της γεωμετρίας του αυτιού του.

Δεδομένα	Δείπνα	Ομιλητές	Ώρες:Λεπτά	Εκφράσεις που χρησιμοποιήθηκαν
Εκπαίδευσης (train)	16	32	40:33	119966
Ανάπτυξης (development-dev)	2	8	4:27	12210
Επαλήθευσης (Evaluation-eval)	2	8	5:12	17488

(Πίνακας 4.1): Όγκος δεδομένων που χρησιμοποιήθηκαν (ο χρόνος των ηχογραφήσεων για κάθε σύνολο είναι η τιμή του πίνακα επί 4).

4.2 Εργαλεία

Σε αυτή την παράγραφο θα περιγραφούν τα εργαλεία που χρησιμοποιήθηκαν για την εκπόνηση της παρούσας διπλωματικής εργασίας, που είναι ανοιχτού κώδικα (open source).

4.2.1 Kaldi

Το Kaldi είναι μία εργαλειοθήκη (toolkit) γραμμένη σε C++ και python. Έχει ως στόχο την κατανόηση, επεξεργασία και ανάπτυξη ευέλικτου κώδικα για εφαρμογές αναγνώρισης φωνής. Χρησιμοποιεί την βιβλιοθήκη OpenFST για την κατασκευή και αναπαράσταση Finite State Transducers [36]. Οι βιβλιοθήκες που περιέχει μπορούν να χρησιμοποιηθούν είτε άμεσα σε shell scripts είτε έμμεσα ως βιβλιοθήκες κάποιας άλλης γλώσσας (π.χ. python). Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν για όλη την κατασκευή του συστήματος shell scripts, σε αρκετές εφαρμογές προτιμάτε η χρήση του Kaldi ως βιβλιοθήκη σε κώδικα python καθώς κατά αυτόν τον τρόπο επιτυγχάνετε ο συνδυασμός ευελιξίας και ευκολίας ανάπτυξης κώδικα, που παρέχει η python, με την εύρωστη εξαγωγή και αναπαράσταση διαφόρων τύπων δεδομένων που παρέχει η εργαλειοθήκη.

4.2.2 CMU pronouncing dictionary

Πρόκειται για μία βάση δεδομένων (database) που αναπτύχθηκε από το Carnegie Mellon University και χρησιμοποιείται με στόχο την αυτόματη μετάφραση λέξεων της αγγλικής γλώσσας από μορφή κειμένου σε μορφή φωνημάτων. [15]

4.2.3 Phonetisaurus G2P

Πρόκειται για ένα μοντέλο που χρησιμοποιεί την βιβλιοθήκη OpenFST και έχει στόχο την αυτόματη μετάφραση λέξεων της αγγλικής γλώσσας από μορφή κειμένου σε μορφή φωνημάτων. Σε αντίθεση με το προηγούμενο εργαλείο δεν έχει κάποια βάση δεδομένων αλλά έχει εκπαιδευτεί να επιτυγχάνει την μετάφραση με χρήση τεχνικών αναγνώρισης προτύπων.

4.3 Στατιστική Ανάπτυξη

4.3.1 Λεξικό

Η κατασκευή του συστήματος ξεκινά με την δημιουργία του λεξικού. Το λεξικό χρησιμοποιείται και στα δύο μοντέλα του συστήματος, χρησιμοποιείται δηλαδή και για την μετάφραση λέξεων σε σειρές φωνημάτων αλλά είναι και οντότητα που περιέχει την πληροφορία για την a priori πιθανότητα εμφάνισης των λέξεων σε όλο το σύνολο δεδομένων.

Για την δημιουργία του λεξικού, αρχικά απομονώνονται οι λέξεις, μετριέται η συχνότητα τους και ο συνολικός αριθμός τους και τροφοδοτούνται στο CMU dictionary με στόχο την μετάφραση τους. Μαζί με την μετάφραση του εξάγονται και λέξεις που δεν μπορούν να μεταφραστούν από την βάση δεδομένων, οι πιο συχνές φαίνονται στον πίνακα 4.2.

Συχνότερες Λέξεις OOV (συμπεριλαμβανομένων και των “κομμένων”)		Συχνότερες Λέξεις OOV	
Λέξεις	Συχνότητα	Λέξεις	Συχνότητα
s-	5588	woah	1980
w-	4488	mkay	1936
y-	4422	arjan	990
i-	3586	wenny	924
th-	3366	ibs	660
d-	2376	shandryn	594
woah	1980	ped	462
t-	1980	woulda	396
mkay	1936	pesto	396
a-	1892	thrones	374

(Πίνακας 4.2): Συχνότερες λέξεις <oov>.

Μπορεί να γίνει η παρατήρηση ότι οι περισσότερες άγνωστες λέξεις είναι λέξεις κομμένες δηλαδή κομπιάσματα στην ροή της ομιλίας, οι λέξεις αυτές πρέπει να ετικετοποιηθούν και να συμπεριληφθούν στο σύστημα διότι κατά αυτόν τον τρόπο συμπεριλαμβάνονται και οι ιδιαιτερότητες του προφορικού λόγου.

Τέλος, όλες οι <oov> λέξεις εισάγονται στο Phonetisaurus G2P με στόχο την μετάφραση τους σε σειρά φωνημάτων και εξάγεται το λεξικό ολοκληρωμένο. Κατά αυτόν τον τρόπο επιτυγχάνεται η προεπεξεργασία των κειμενικών δεδομένων.

4.3.2 Γλωσσικό Μοντέλο

Το επόμενο βήμα είναι η κατασκευή του γλωσσικού μοντέλου. Ο λόγος είναι ότι έχοντας αυτό μπορεί ο κατασκευαστής και ελέγχει την απόδοση του συνολικού συστήματος ανά πάσα στιγμή, άλλωστε όπως μπορεί να παρατηρηθεί από το προηγούμενο κεφάλαιο οι τεχνικές που αφορούν το ακουστικό μοντέλο είναι αρκετά περισσότερες και είναι σημαντικό να υπάρχει μία εποπτεία για το πώς αυτές επηρεάζουν την απόδοση του συστήματος.

Για την εκπόνηση της παρούσας διπλωματικής εργασίας έγινε η κατασκευή έξι ακουστικών μοντέλων, δύο με την τεχνική έκπτωσης Good turing, δύο με την τεχνική ομαλοποίησης Kneiser-Nay και δύο με την τεχνική μέγιστης εντροπίας. Για κάθε μία τεχνική εκπαιδευτικέ ένα 3-gram και ένα 4-gram μοντέλο. Τέλος με την χρήση του συνόλου ανάπτυξης υπολογίσθηκαν οι

πολυπλοκότητες και επιλέχθηκε αυτή με την μικρότερη, στον πίνακα 4.3 φαίνονται οι πολυπλοκότητες για κάθε τεχνική και κάθε N-gram.

Τεχνική	N-gram	Πολυπλοκότητα
Good Turing	3-gram	428,16
	4-gram	428,34
Kneiser Nay	3-gram	307,80
	4-gram	307,44
Μέγιστη Εντροπία	3-gram	300,32
	4-gram	300,67

(Πίνακας 4.3): Πολυπλοκότητες γλωσσικών μοντέλων ανά τεχνική και N-gram.

Μπορεί να παρατηρηθεί από τον παραπάνω πίνακα ότι οι τεχνικές Kneiser Nay και μέγιστης εντροπίας παρουσιάζουν μικρή διαφορά στην πολυπλοκότητα για το συγκεκριμένο σύνολο δεδομένων. Ενώ, όσων αφορά την πολυπλοκότητα, ο αριθμός των N-grams δείχνει να μην την επηρεάζει. Για την συνέχεια της ανάπτυξης του συστήματος χρησιμοποιήθηκε το 3-gram μοντέλο μέγιστης εντροπίας.

4.3.3 Εξαγωγή και επεξεργασία ακουστικών χαρακτηριστικών

Για την εξαγωγή των MFCC ακολουθήθηκε η διαδικασία που περιγράφηκε στην παράγραφο 3.1 και φαίνεται στο διάγραμμα 3.2. Η μόνη διαφορά, στην πράξη, είναι για την κατασκευή του αναφάσματος αντί της χρήσης του αντιστρόφου DFT χρησιμοποιήθηκε αντίστροφος μετασχηματισμός συνημιτόνου (inverse discrete cosine transform – IDCT). Οι δύο μετασχηματισμοί είναι παρόμοιοι η κύρια διαφορά τους είναι ότι ο DCT χρησιμοποιεί τις πραγματικές τιμές του σήματος, ενώ ο DFT χρησιμοποιεί τις μιγαδικές. Για πραγματικά σήματα (ή σχεδόν πραγματικά όπως το σήμα φωνής) οι δύο μετασχηματισμοί παρουσιάζουν το ίδιο μέσο τετραγωνικό σφάλμα. Η ιδιότητα όμως που κάνει τον DCT προτιμητέο για τέτοιου τύπου εφαρμογές απορρέει από το θεώρημα του Parseval (θεώρημα κατανομής της ενέργειας στο πεδίο της συχνότητας για πεπερασμένου μήκους σήματα), είναι η ιδιότητα της ενεργειακής συμπύκνωσης. Η ακολουθία του DCT παρουσιάζει συγκέντρωση στις χαμηλές συχνότητες ενώ στις υψηλές μηδενίζεται. Λόγω αυτού το ανάφασμα μπορεί να κοπεί και να κρατηθούν μόνο οι χαμηλές συχνότητες, αυτό αντανακλά στον όγκο αναπαράστασης των MFCC και άρα στην μνήμη που καλύπτουν τα δεδομένα και κατ’ επέκταση στον χρόνο εκπαίδευσης μοντέλου. [13][37]

Μετά την εξαγωγή των MFCC εφαρμόσθηκε σε αυτές κανονικοποίηση μέσης τιμής και διασποράς αναφάσματος (cepstral mean and variance normalization – CMVN) δηλαδή για κάθε έκφραση με χρήση σταθερού παραθύρου, υπολογίσθηκε η μέση τιμή και η διασπορά των MFCC της, εφαρμόζοντας τον παρακάτω τύπο:

$$\frac{MFCC_t - \mu_{MFCC}(i)}{\sigma_{MFCC}(i)} \quad .$$

Η κανονικοποίηση αυτή έχει στόχο να κάνει τα δεδομένα πιο εύρωστα, δηλαδή αμετάβλητα όσων αφορά τον τυχαίο θόρυβο αφού κατ’ αυτόν τον τρόπο ο στατικός θόρυβος που μπορεί να υπάρχει σε μία ηχογράφηση μίας έκφρασης εξαλείφεται. [38]

4.3.4 Ακουστικό Μοντέλο

Αφότου έχουν εξαχθεί τα διανύσματα χαρακτηριστικών και έχει δημιουργηθεί το λεξικό μπορεί να γίνει η κατασκευή του ακουστικού μοντέλου. Πριν την ανάπτυξη της διαδικασίας που ακολουθήθηκε, θα πρέπει να γίνει αναφορά σχετικά με ένα τεχνικό θέμα που προκύπτει. Αυτό έχει

να κάνει με ότι ένα GMM δέχεται ως είσοδο ένα πεπερασμένο αριθμό διανυσμάτων πράγμα που σημαίνει ότι σπάνια σε αυτόν τον αριθμό θα υπάρχει ακριβώς μία ολόκληρη έκφραση, συνήθως θα υπάρχει είτε ένα τμήμα αν η έκφραση είναι μεγάλη είτε μία έκφραση ολόκληρη (αν αυτή είναι μικρή) και ένα τμήμα μίας άλλης. Ως επίλυση αυτού του θέματος εισάγεται το φώνημα της προαιρετικής σιωπής </s> (optional silence) που αντικαθιστά το φώνημα της σιωπής στην αρχή και στο τέλος μίας πρότασης. Για τον λόγο αυτό ο δημιουργός του συνόλου δεδομένων θα πρέπει να είναι προσεκτικός αφήνοντας (ή εισάγοντας) κάποια milisecond σιωπής πριν και μετά από κάθε έκφραση. Το ίδιο ισχύει και για τον κατασκευαστή καθώς η αυθαίρετη εισαγωγή του φωνήματος σιωπής στην αρχή και στο τέλος της έκφρασης εγκυμονεί τον κίνδυνο το φώνημα της προαιρετικής σιωπής να θεωρηθεί ως κάποιο άλλο και αλυσιδωτά όλα τα φωνήματα μίας έκφρασης να θεωρηθούν ως το διπλανό τους φώνημα. Για αυτές περιπτώσεις συχνά η λύση είναι η αύξηση της βαρύτητας του φωνήματος της σιωπής έως 1.5 φορές.

Για την διαδικασία κατασκευής του ακουστικού μοντέλου ακολουθήθηκαν τα εξής βήματα:

1) Κατασκευάστηκε ένα μονοφωνικό GMM και έγινε ευθυγράμμιση των καταστάσεων του (φωνήματα) με τα κειμενικά δεδομένα. (mono)
2) Έπειτα κατασκευάστηκε ένα τριφωνικό GMM και έγινε ευθυγράμμιση των καταστάσεων του (φωνήματα) με τα κειμενικά δεδομένα.(tri1)
3) Επέκταση του τριφωνικού μοντέλου του προηγούμενου βήματος χρησιμοποιώντας τις τεχνικές LDA και MLT. (tri2)
4) Τέλος έγινε μία επιπλέον επέκταση χρησιμοποιώντας την τεχνική SAT. (tri3)

Στον πίνακα 4.4 φαίνονται τα χαρακτηριστικά των GMM. Σε όλα τα μοντέλα χρησιμοποιήθηκε αύξηση της βαρύτητας του φωνήματος της σιωπής κατά 1.25 φορές, δοκιμάστηκε και επιπλέον αύξηση χωρίς κάποια ιδιαίτερη συνεισφορά στην εμφάνιση του φωνήματος προαιρετικής σιωπής.

Μοντέλο	Επαναλήψεις	Αριθμός Καταστάσεων	Αριθμός Γκαουσιανών
mono	39	142	985
tri1	34	1936	30079
tri2	34	3224	50097
tri3	34	3968	100108

(Πίνακας 4.4): Χαρακτηριστικά κατασκευής GMM.

4.3.5 Αναζήτηση πιθανότερης έκφρασης

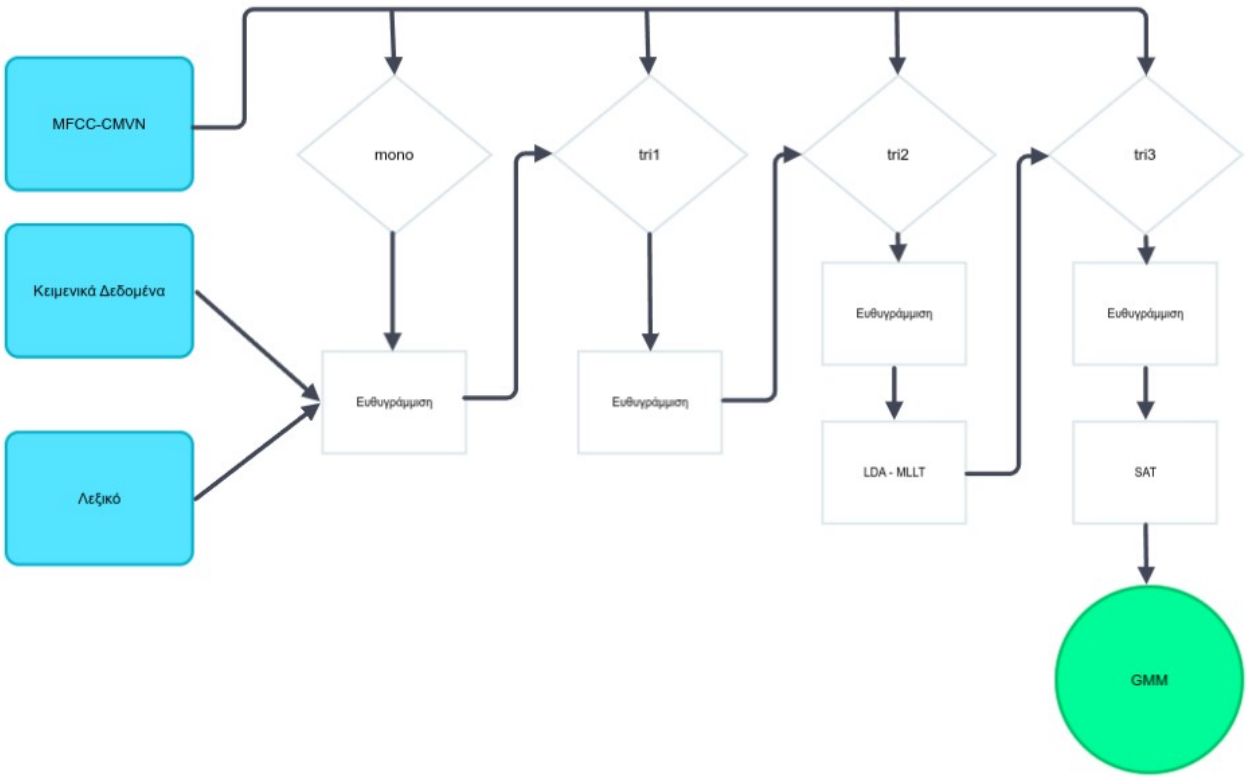
Στην συνέχεια εξήχθησαν τα lattices για κάθε μοντέλο και με την χρήση του γλωσσικού μοντέλου μέγιστης εντροπίας κατασκευάστηκαν οι γράφοι αποκωδικοποίησης. Τέλος εξήχθησαν τα WER% για το σύνολο ανάπτυξης και αναζήτησης. Αναλυτικά τα αποτελέσματα θα παρουσιαστούν στο επόμενο κεφάλαιο. Το σύστημα που προέκυψε από το tri3 χρησιμοποιήθηκε για την συνέχεια της ανάπτυξης.

4.4 i-vecrors

Χρησιμοποιώντας το tri3 GMM ως UBM εξάγονται τα i-vectors ακολουθώντας την διαδικασία που περιγράφηκε στην παράγραφο 3.4.1.

4.5 Ανάπτυξη με μηχανική μάθηση

Για την εκπόνηση της παρούσας διπλωματικής εργασίας κατασκευάστηκε ακουστικό και γλωσσικό μοντέλο με χρήση DNN. Πιο συγκεκριμένα δοκιμάστηκαν τρεις αρχιτεκτονικές για ακουστικό μοντέλο και δύο αρχιτεκτονικές για γλωσσικό μοντέλο.



(Διάγραμμα 4.1): Βήματα κατασκευής GMM.

4.5.1 Ακουστικό μοντέλο

Η προσέγγιση που χρησιμοποιήθηκε είναι η υβριδική προσέγγιση DNN-HMM και για όλες τις αρχιτεκτονικές χρησιμοποιήθηκε ως κριτήριο εκπαίδευσης το LF-MMI. Αρχικά από τα lattices κατασκευάζεται μια τοπολογία HMM, αριθμώντας όλες τις πιθανές σειρές λέξεων και όλες πιθανές προφορές για κάθε λέξη. Ουσιαστικά η τοπολογία αυτή είναι ο γράφος όλων των πιθανών σειρών λέξεων ή όπως αναφέρθηκε στο προηγούμενο κεφάλαιο ο γράφος του παρονομαστή της συνάρτησης κόστους MMI. Στην συνέχεια για κάθε τμήμα εισόδου (chunk) του DNN κατασκευάζεται ένας γράφος που συνδέει την αληθή σειρά των λέξεων με τα ακουστικό chunk, πρόκειται δηλαδή για τον γράφο του παρονομαστή της συνάρτησης κόστους MMI. Κατά αυτόν τον τρόπο παίρνοντας την έξοδο από το DNN για κάθε chunk και εισάγοντας την στους finite state transducters, που ανταποκρίνονται στους παραπάνω γράφους, με την χρήση του εμπρόσθιου και οπίσθιου αλγορίθμου μπορεί να βρεθεί η παράγωγος της συνάρτησης MMI.

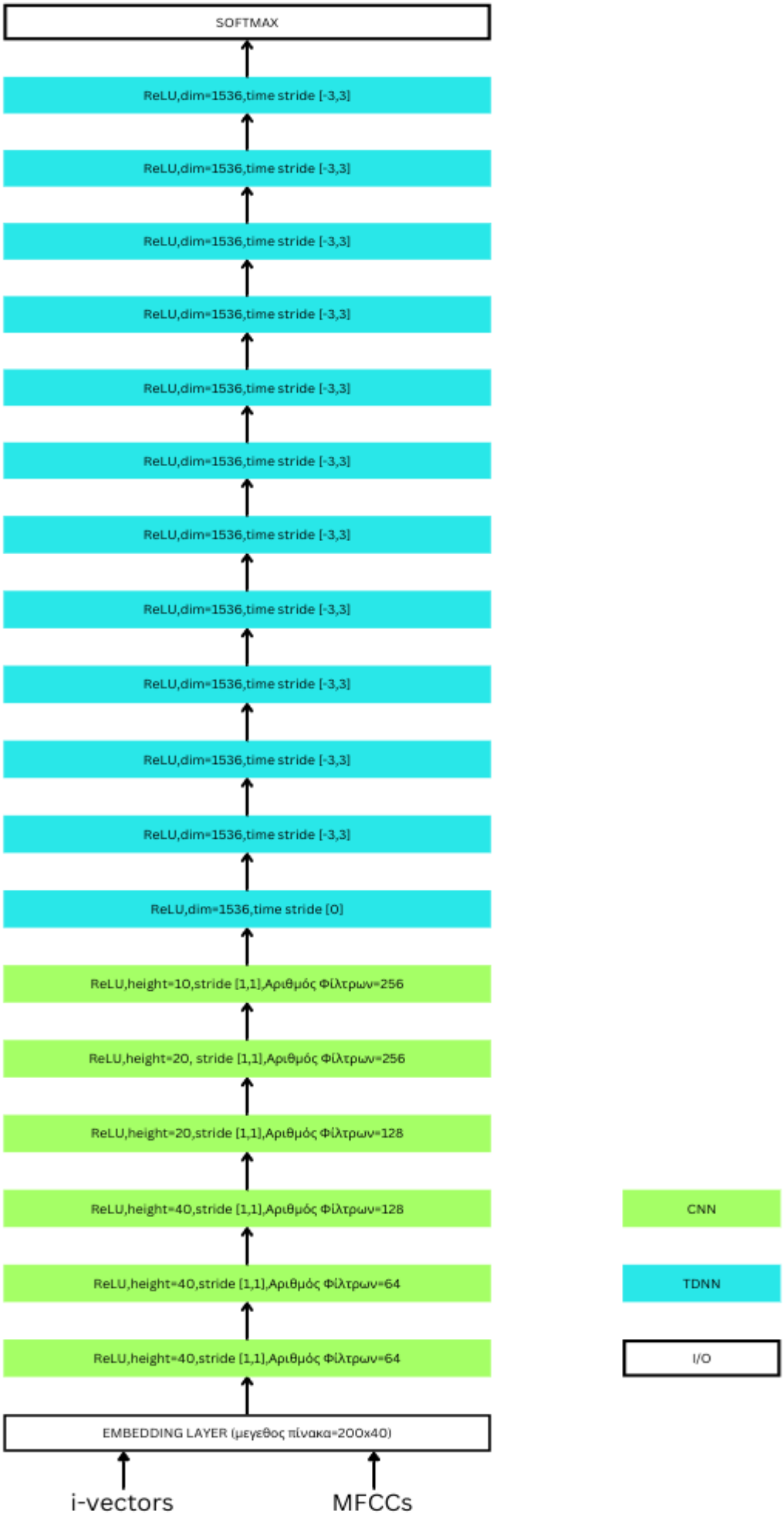
Για την κατασκευή του ακουστικού μοντέλου δοκιμάστηκαν τρεις αρχιτεκτονικές και συγκρίθηκαν διάφορες τιμές ρυθμού μάθησης. Οι αρχιτεκτονικές προς εκπαίδευση είναι οι παρακάτω:

1) TDNN αρχιτεκτονική που αποτελείται από 17 επίπεδα (15 κρυφά TDNN).
2) Υβριδική CNN-TDNN αρχιτεκτονική που αποτελείται από 20 επίπεδα (6 κρυφά CNN και 12 κρυφά TDNN).
3) Υβριδική CNN-TDNN-LSTM αρχιτεκτονική που αποτελείται από 16 επίπεδα (2 κρυφά CNN και 3 κρυφά LSTM, 9 κρυφά TDNN).

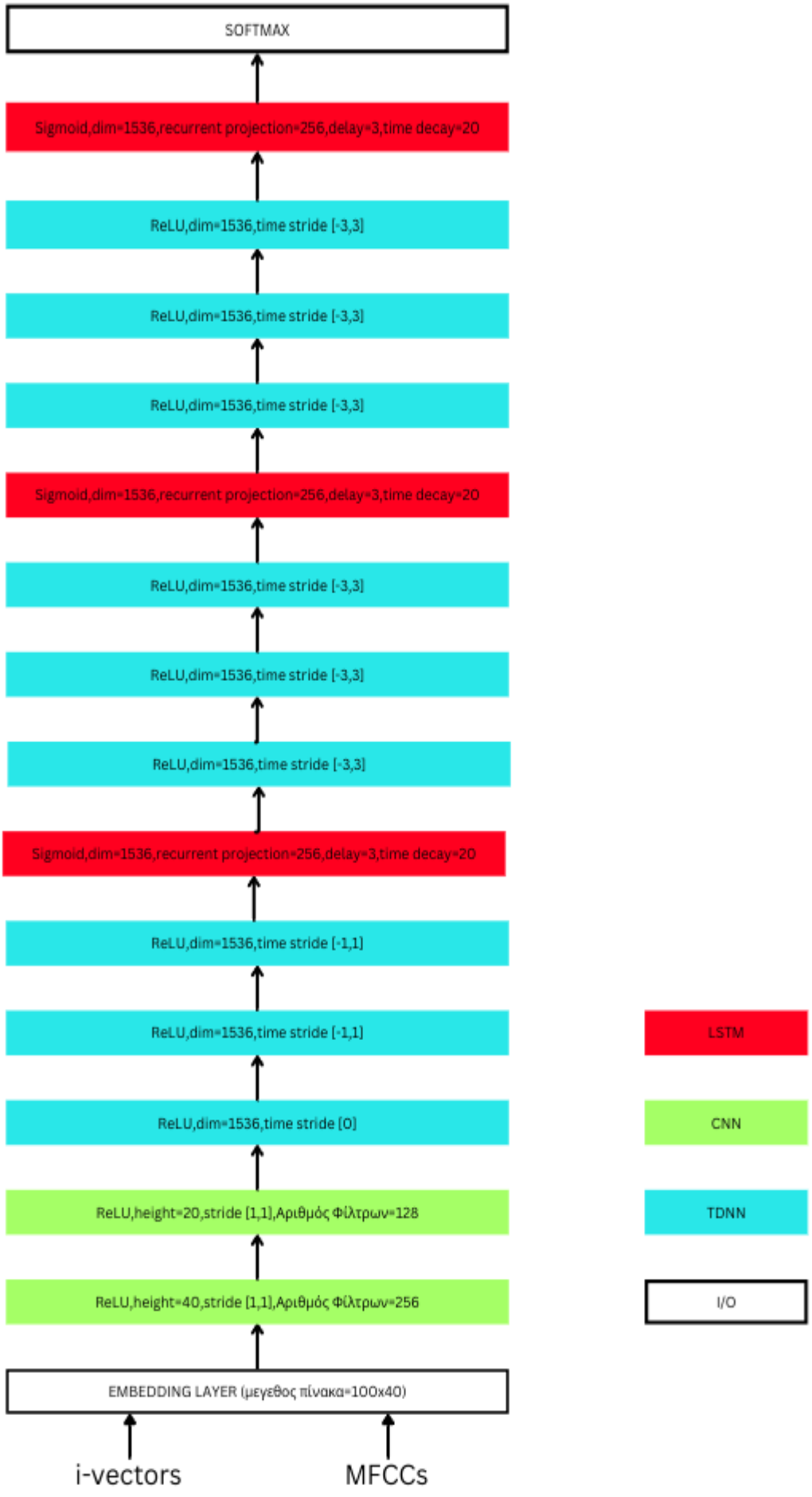
Στα παρακάτω διαγράμματα (4.2,4.3,4.4) φαίνονται οι αρχιτεκτονικές, για κάθε επίπεδο φαίνεται η συνάρτηση ενεργοποίησης και το μέγεθος του επιπέδου (στα LSTM η Sigmoid είναι η συνάρτηση ενεργοποίησης εξόδου, για τον κόμβο ανάδρασης χρησιμοποιείται η tanh). Όλες οι αρχιτεκτονικές εκπαιδευτήκαν για 4 εποχές που σημαίνει 492 iterations για την TDNN αρχιτεκτονική,300 iterations για την CNN-TDNN αρχιτεκτονική και 492 iterations για την CNN-TDNN-LSTM αρχιτεκτονική.

Στην συνέχεια εξήχθησαν τα lattices για κάθε αρχιτεκτονική και με την χρήση του γλωσσικού μοντέλου μέγιστης εντροπίας κατασκευάστηκαν οι γράφοι αποκωδικοποίησης. Τέλος εξήχθησαν τα WER% για το σύνολο ανάπτυξης και αναζήτησης. Αναλυτικά τα αποτελέσματα θα παρουσιαστούν στο επόμενο κεφάλαιο.

(Διάγραμμα 4.2): Αρχιτεκτονική TDNN.



(Διάγραμμα 4.3): Αρχιτεκτονική CNN-TDNN.

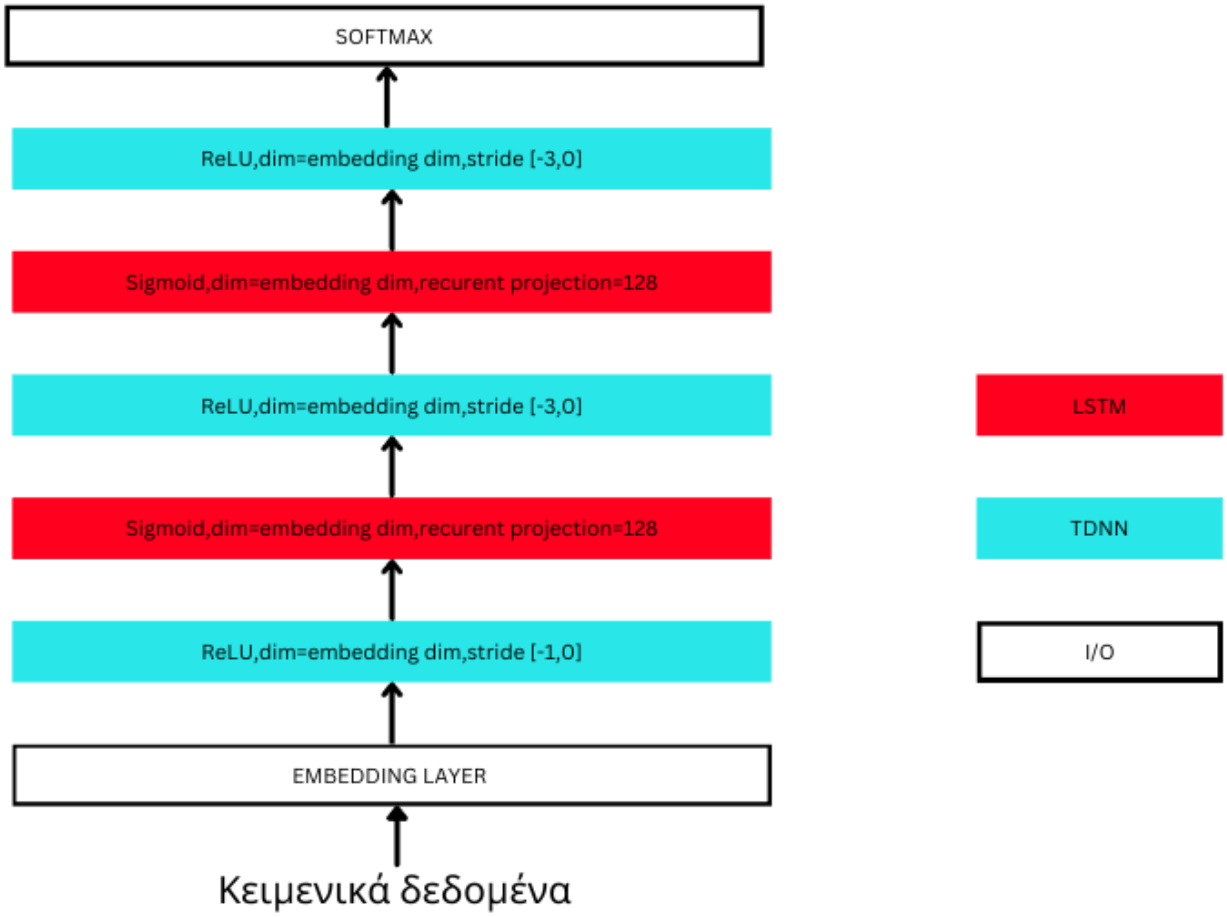


(Διάγραμμα 4.4):Αρχιτεκτονική CNN-TDNN-LSTM

4.5.2 Γλωσσικό μοντέλο

Για την κατασκευή του γλωσσικού μοντέλου δοκιμάστηκαν δύο αρχιτεκτονικές. Οι αρχιτεκτονικές είναι όμοιες δομούνται από 7 επίπεδα (3 κρυφά TDNN και 2 κρυφά LSTM), η διαφορά τους είναι ότι στη πρώτη χρησιμοποιήθηκαν απλά LSTM ενώ στην δεύτερη b-LSTM (bidirectional- LSTM). Για τις αρχιτεκτονικές δοκιμάστηκαν διάφορες τιμές μεγέθους των επιπέδων, η τιμή που κατασκευάζει το μοντέλο με την μικρότερη πολυπλοκότητα χρησιμοποιήθηκε για την αποκωδικοποίηση των ακουστικών μοντέλων.

Τέλος για την κατασκευή του γράφου αποκωδικοποίησης δεν χρειάστηκε η επανεκπαίδευση του ακουστικού μοντέλου. ουσιαστικά χρησιμοποιήθηκαν τα lattices που είχαν εξαχθεί για κάθε μοντέλο, σε αυτά αφαιρέθηκε η συνεισφορά του γλωσσικού μοντέλου (μέγιστης εντροπίας) και αθροίστηκε η συνεισφορά του νέου μοντέλου. Στην περίπτωση της παρούσας διπλωματικής εργασίας χρησιμοποιήθηκε μια πιο πολύπλοκη τεχνική, η τεχνική N-best. Δηλαδή, πριν την διαγραφή της συνεισφοράς της μέγιστης εντροπίας εξήχθησαν (για κάθε lattce) μία λίστα των N=10 πιο πιθανών προτάσεων και κρίθηκαν από το νέο μοντέλο για ποια από τις προτάσεις είναι η πιθανότερη. Τέλος εξήχθησαν τα WER% για το σύνολο ανάπτυξης και αναζήτησης. Αναλυτικά τα αποτελέσματα θα παρουσιαστούν στο επόμενο κεφάλαιο.



(Διάγραμμα 4.5): Αρχιτεκτονική γλωσσικού μοντέλου με DNN.

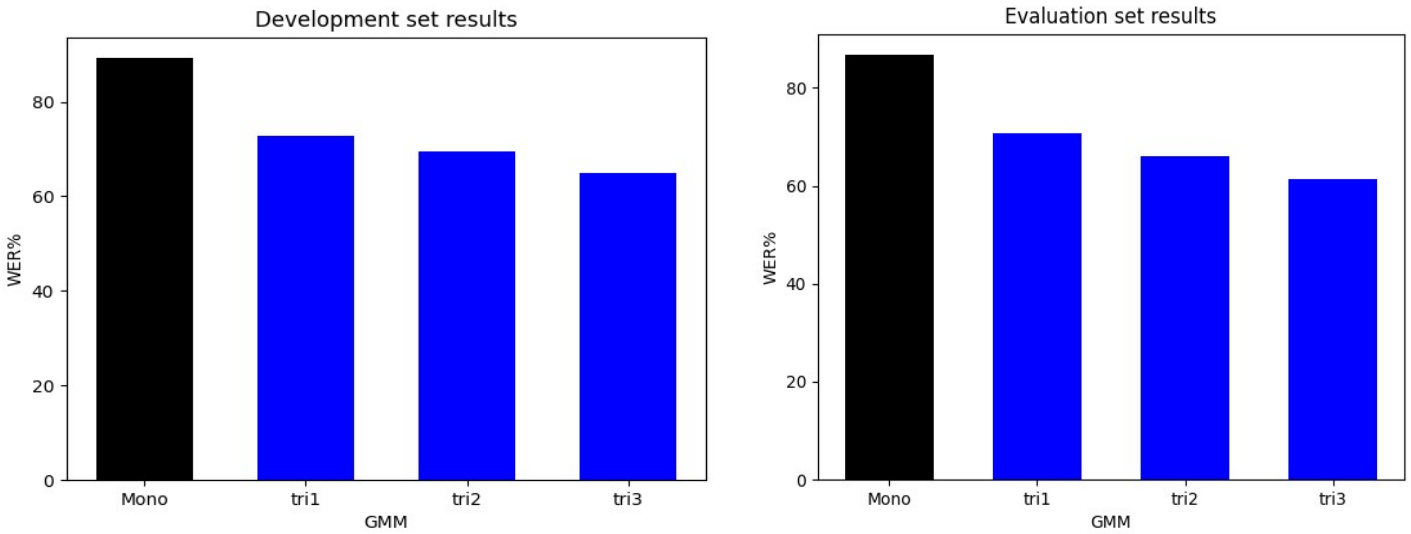
5. Αποτελέσματα

5.1 Αποτελέσματα στατιστικής ανάπτυξης

Στον πίνακα 5.1 φαίνονται τα αποτελέσματα των GMM που κατασκευάστηκαν. Αρχικά η πρώτη παρατήρηση είναι ότι τα τριφωνικά μοντέλα παρουσιάζουν κατά μέσο όρο περίπου 20% καλύτερη απόδοση σε σχέση με το μονοφωνικό μοντέλο. Επίσης συγκρίνοντας τα τριφωνικά μοντέλα μεταξύ τους παρατηρείται ότι οι τεχνικές που χρησιμοποιήθηκαν παρουσιάζουν σημαντική βελτίωση της τάξεως του 10%.

GMM	WER% dev	WER% eval
mono	89.22	86.68
tri1	72.67	70.84
tri2	69.61	65.94
tri3	64.94	61.42

(Πίνακας 5.1): Αποτελέσματα GMM.



(Εικόνα 5.1):Γράφημα αναπαράστασης WER% ανά GMM.

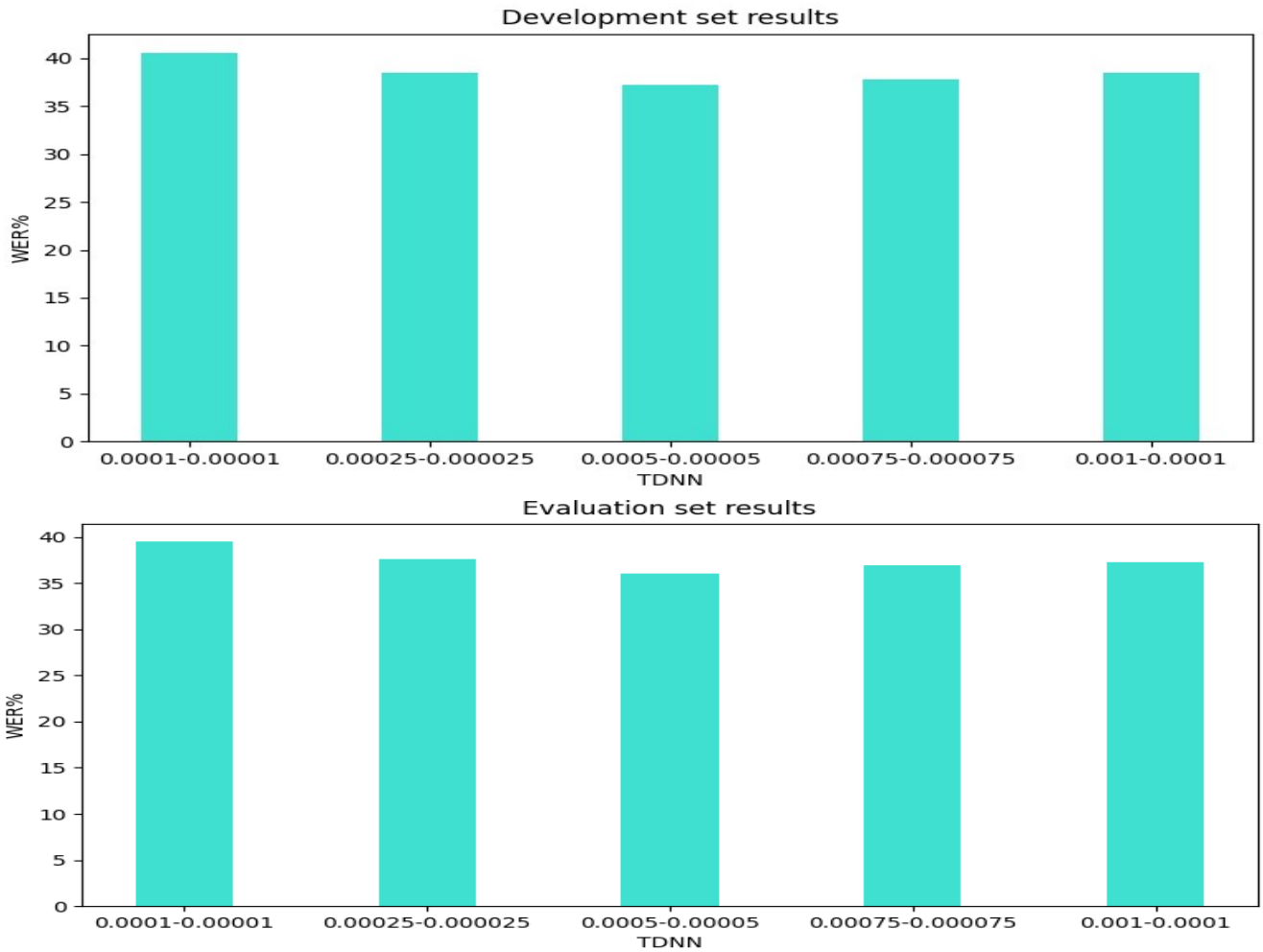
5.2 Αποτελέσματα αρχιτεκτονικών DNN

5.2.1 Ακουστικό Μοντέλο

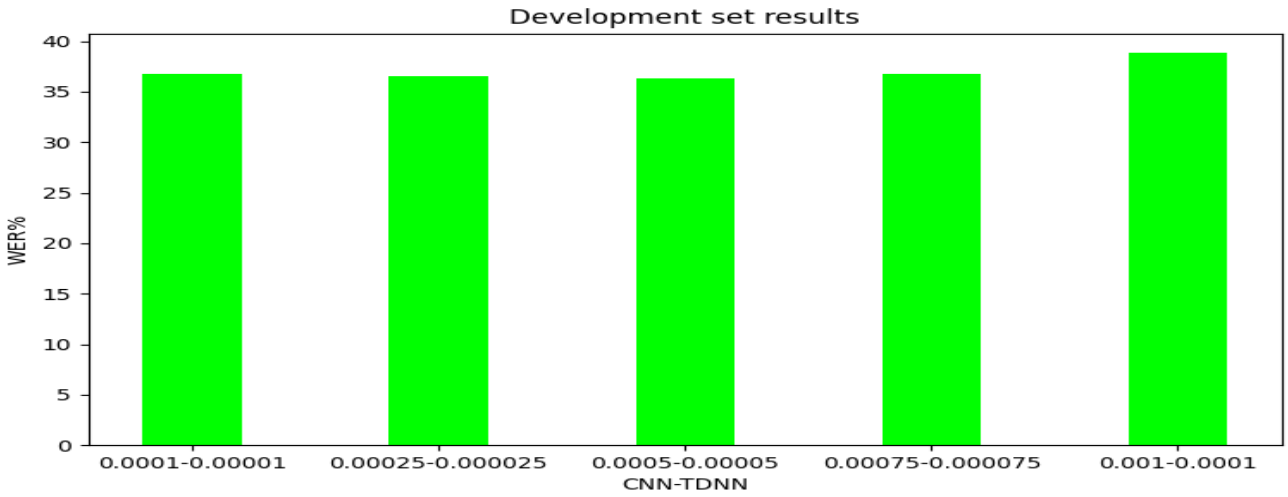
Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, με την χρήση του γλωσσικού μοντέλου μέγιστης εντροπίας αποκωδικοποιήθηκαν οι αρχιτεκτονικές που παρουσιάστηκαν, για διάφορες τιμές του ρυθμού μάθησης. Στον πίνακα 5.2 φαίνονται οι τιμές WER% για κάθε αρχιτεκτονική, ρυθμό και σύνολο δεδομένων. Είναι φανερός ο λόγος που τα GMM θεωρούνται ξεπερασμένα αφού τα συστήματα με DNN παρουσιάζουν περίπου τη διπλάσια απόδοση. Οι αρχιτεκτονικές μεταξύ τους παρουσιάζουν μικρές διαφορές, κάτι που κοιτάζοντας την σχετική βιβλιογραφία για το συγκεκριμένο σύνολο δεδομένων (πρόκληση CHiME 5) ήταν αναμενόμενο. Η αρχιτεκτονική CNN-TDNN παρουσιάζει συνολικά την καλύτερη απόδοση.

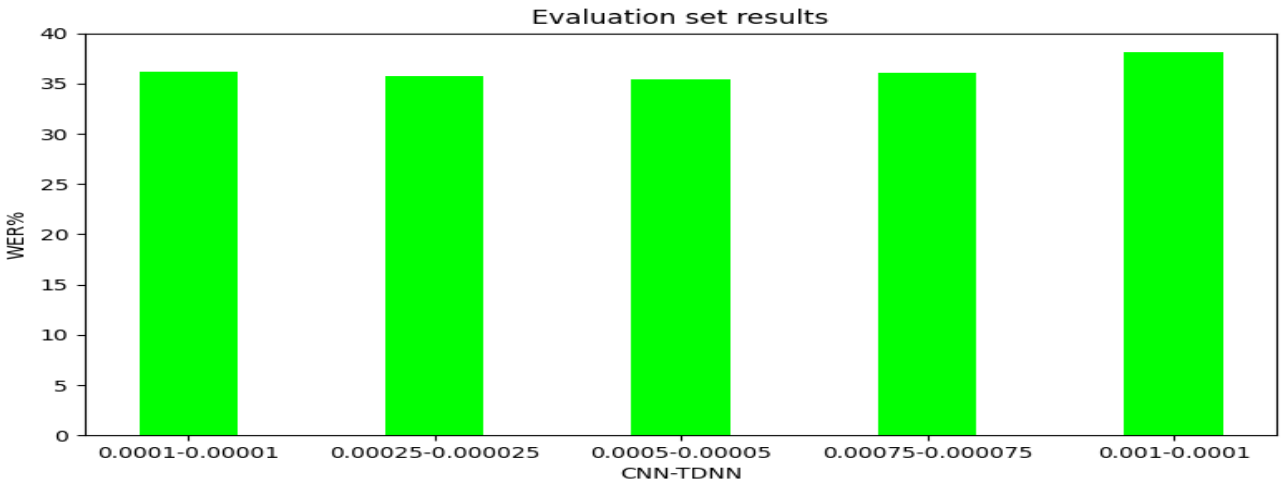
DNN	Ρυθμός Μάθησης	WER% dev	WER% eval
TDNN	0.001-0.0001	38.45	37.30
	0.00075-0.000075	37.77	36.91
	0.0005-0.00005	37.17	36.01
	0.00025-0.000025	38.54	37.55
	0.0001-0.00001	40.54	39.48
CNN-TDNN	0.001-0.0001	36.73	36.14
	0.00075-0.000075	36.56	35.73
	0.0005-0.00005	36.28	35.44
	0.00025-0.000025	36.70	36.07
	0.0001-0.00001	38.82	38.11
CNN-TDNN-LSTM	0.002-0.0002	41.34	40.08
	0.0015-0.00015	41.14	40.65
	0.001-0.0001	41.77	40.91
	0.0005-0.00005	42.65	41.96
	0.0001-0.00001	45.65	44.74

(Πίνακας 5.2):Αποτελέσματα DNN.

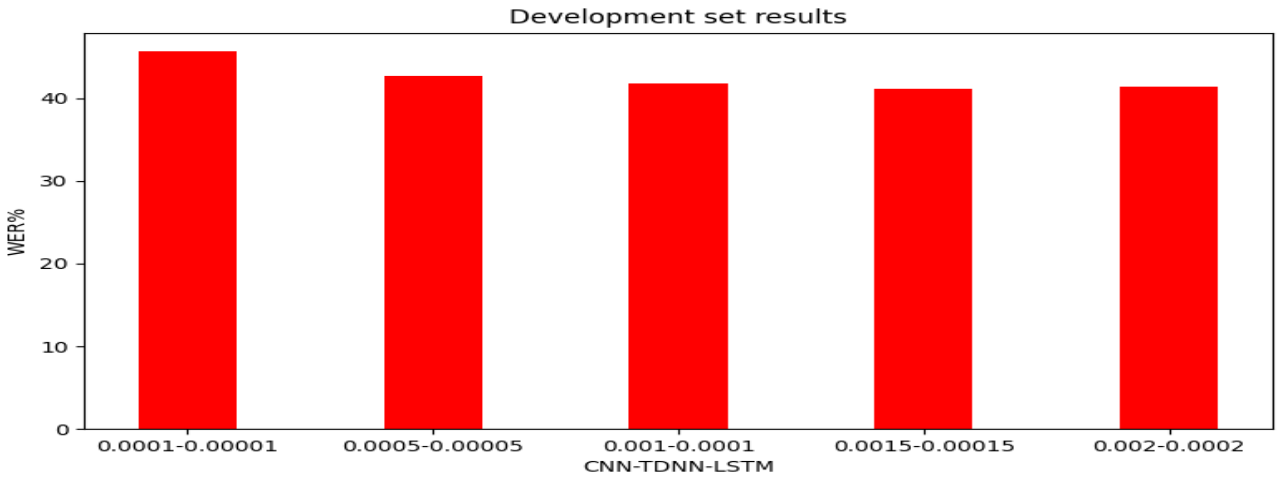


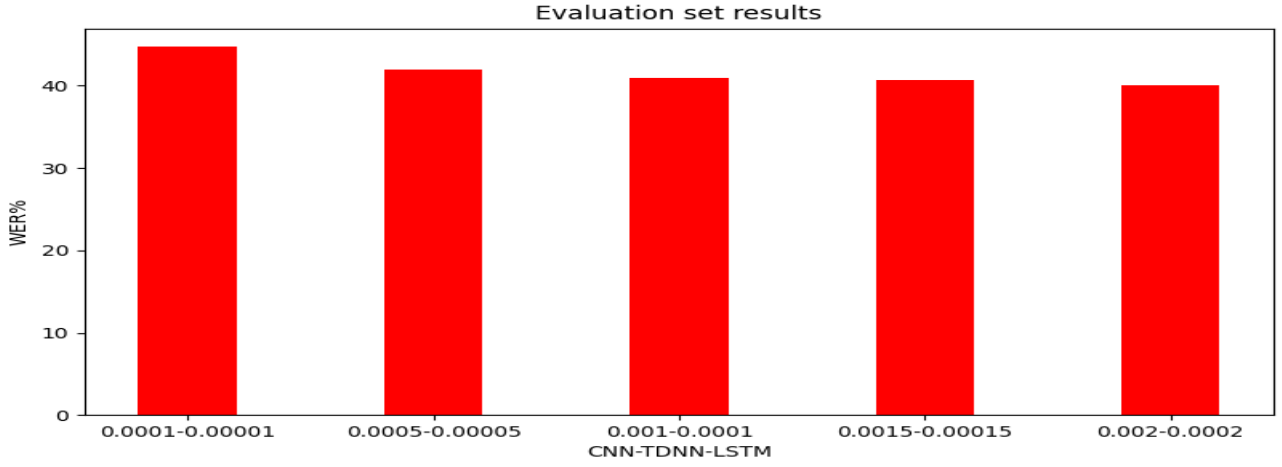
(Εικόνα 5.2):Γράφημα αναπαράστασης WER% ανά ρυθμό εκπαίδευσης για την αρχιτεκτονική TDNN..





(Εικόνα 5.3):Γράφημα αναπαράστασης WER% ανά ρυθμό εκπαίδευσης για την αρχιτεκτονική CNN-TDNN..





(Εικόνα 5.4):Γράφημα αναπαράστασης WER% ανά ρυθμό εκπαίδευσης για την αρχιτεκτονική CNN-TDNN-LSTM.

5.2.2 Γλωσσικό Μοντέλο

Έπειτα εκπαιδεύτηκαν τα γλωσσικά μοντέλα και χρησιμοποιήθηκαν για την αποκωδικοποίηση των ακουστικών μοντέλων με την καλύτερη απόδοση για κάθε αρχιτεκτονική. Στον πίνακα 5.3 φαίνονται οι πολυπλοκότητες όλων των γλωσσικών μοντέλων (συμπεριλαμβανομένων και αυτών που σχεδιαστήκαν με στατιστικές μεθόδους), οι πολυπλοκότητες για τις αρχιτεκτονικές που κατασκευάστηκαν με DNN είναι οι ίδιες και γι’ αυτό παρουσιάζονται μόνο μια φορά. Το μοντέλο που κατασκευάστηκε με LSTM και διάσταση επιπέδων 512 παρουσιάζει την μικρότερη πολυπλοκότητα από όλα τα μοντέλα που δοκιμάστηκαν. Στον πίνακα 5.4 φαίνονται τα αποτελέσματα για κάθε σύστημα. Τα συστήματα που χρησιμοποιήθηκε γλωσσικό μοντέλο με απλά LSTM παρουσιάζουν την καλύτερη απόδοση. Η βελτίωση που πρόσφερε το γλωσσικό μοντέλο ήταν η αναμενόμενη, τόσο στη βιβλιογραφία που χρησιμοποιείται το ίδιο σύνολο δεδομένων (πρόκληση CHiME 5) όσο και στη γενική βιβλιογραφία που σχετίζεται με την αναγνώριση φωνής, η βελτίωση είναι αντίστοιχη.

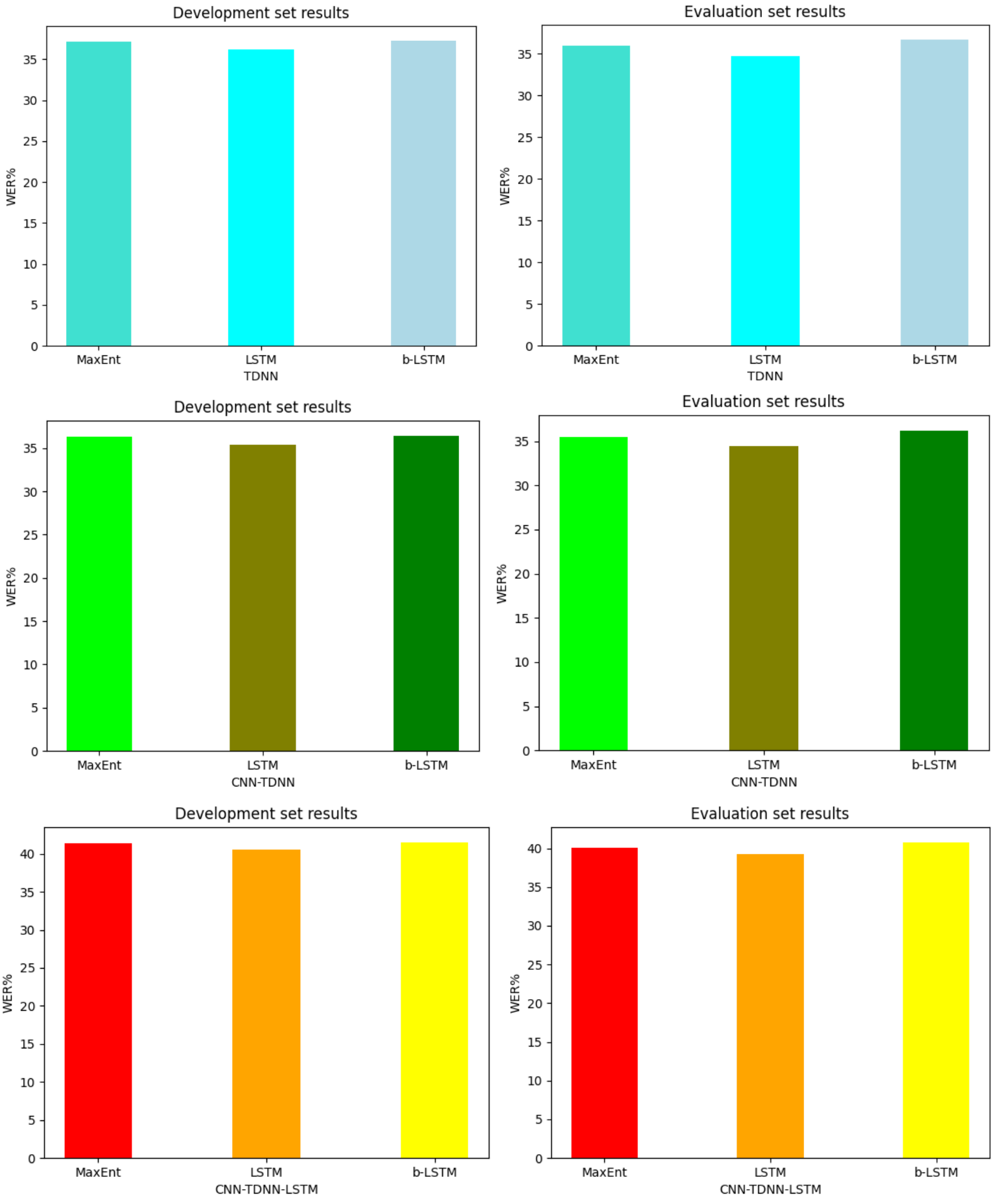
Το καλύτερο σύστημα αναγνώρισης φωνής που κατασκευάστηκε είναι αυτό που χρησιμοποιήθηκε CNN-TDNN για το ακουστικό μοντέλο και απλά LSTM για το γλωσσικό με 35.35 WER% για το σύνολο ανάπτυξης και 34.41 WER% για το σύνολο επαλήθευσης.

Μοντέλο		Πολυπλοκότητα
Good Turing	3-gram	428,16
	4-gram	428,34
Kneiser Nay	3-gram	307,80
	4-gram	307,44
Μέγιστη Εντροπία	3-gram	300,32
	4-gram	300,67
LSTM	Embedding dim=512	281
	Embedding dim=1024	299
	Embedding dim=2048	540

(Πίνακας 5.3):Πολυπλοκότητες γλωσσικών μοντέλων.

DNN	Ρυθμός Μάθησης	L.M.	WER % dev	WER % eval
TDNN	0.0005	MaxEnt	37,17	36,01
		LSTM	36,2	34,72
		b-LSTM	37,23	36,68
CNN-TDNN	0.0005	MaxEnt	36,28	35,44
		LSTM	35.35	34.41
		b-LSTM	36,39	36,17
CNN-TDNN-LSTM	0.002	MaxEnt	41,34	40,08
		LSTM	40,58	39,28
		b-LSTM	41,47	40,74

(Πίνακας 5.4):Αποτελέσματα συστημάτων αναγνώρισης φωνής.



(Εικόνα 5.4):Γράφημα αναπαράστασης WER% ανά σύστημα.

6 Συμπεράσματα και επεκτάσεις

Στην παρούσα διπλωματική εργασία αναλύθηκε διεξοδικά η κατασκευή ενός σύγχρονου συστήματος αναγνώρισης φωνής. Μελετήθηκαν και εφαρμόστηκαν αρκετές τεχνικές βελτιστοποίησης τόσο για το ακουστικό όσο και για το γλωσσικό μοντέλο. Παρόλο που το σύστημα με την καλύτερη απόδοση είναι πολύ καλύτερο από τα συστήματα που κατασκευάστηκαν για την πρόκληση CHiME 5 (αντίστοιχη βιβλιογραφία για το αντίστοιχο σύνολο δεδομένων), δεν μπορεί να εφαρμοστεί για την κατασκευή εφαρμογών, καθώς το ποσοστό λάθους (περίπου μία στις τρεις λέξεις) το καθιστά αδύνατο.

Το σύνολο δεδομένων που χρησιμοποιήθηκε συναντά πλήρως τον στόχο που είχε τεθεί (Κεφάλαιο 2.3). Το περιεχόμενο της πληροφορίας που εμπεριέχει, όμως, δημιουργεί μία σύγχυση, αυτό οφείλετε στο γεγονός του ότι οι συμμετέχοντες (φοιτητές) χρησιμοποιούν την αγγλική αργκό. Στη σύγχρονη αγγλική αργκό, παρατηρείται η τάση να κόβονται ή να προστίθενται καταλήξεις, να δημιουργούνται σύνθετες λέξεις με την ίδια σημασία με ένα από τα συνθετικά τους (π.χ. mkay = I am OK), να χρησιμοποιούνται αρκτικόλεξα και να επαναλαμβάνονται λέξεις με χαμηλή σημασία στο περιεχόμενο. Όλα τα παραπάνω χαρακτηριστικά δυσκολεύουν το έργο ενός συστήματος αναγνώρισης φωνής. Αν ο αναγνώστης ενδιαφέρεται να ασχοληθεί με το συγκεκριμένο σύνολο θα μπορούσε να συμπεριλάβει και τα δεδομένα από τα μικρόφωνα με το γραμμικό διάνυσμα, με χρήση τεχνικών ψηφιακής επεξεργασίας σήματος και επιπλέον υπολογιστικούς πόρους. Επίσης θα μπορούσε να χρησιμοποιήσει μεταφορά μάθησης με χρήση ενός μοντέλου που θα έχει εκπαιδευτεί σε πιο γενική πληροφορία πάνω στην αγγλική γλώσσα και να ενσωματώσει έπειτα τα δεδομένα του συνόλου αυτού, τόσο για το γλωσσικό όσο και για το ακουστικό μοντέλο. Τέλος για να συμπεριληφθούν οι εκφράσεις που αποτελούνταν μόνο από θόρυβο,γέλια κ.α. θα μπορούσε να γίνει η κατασκευή ενός πιο γενικού συστήματος που θα αναγνωρίσει ένα πιο ευρύ φάσμα ήχων και θα εμπεριείχε κάποιο γενικό σύστημα κατηγοριοποίησης ήχων που θα συνεργαζόταν με πιο εξειδικευμένα συστήματα όπως το σύστημα αναγνώρισης ομιλίας.

Η διαδικασία που ακολουθήθηκε είναι προϊόν συστηματικής μελέτης, η βελτίωση των αποτελεσμάτων καθ’ όλη την διάρκεια της είναι σημαντική και παρουσιάζει ιδιαίτερο ενδιαφέρον. Η βελτίωση επιτυγχάνεται με την χρήση DNN είναι αξιοσημείωτη. Αν ο αναγνώστης ενδιαφέρεται να ασχοληθεί με τα συστήματα αναγνώρισης φωνής συνιστάται να ακολουθήσει αντίστοιχη διαδικασία. Με μικρές αλλαγές στις παραμέτρους των αρχιτεκτονικών μπορεί να ενσωματωθεί και να εκπαιδευτεί οποιοδήποτε σύνολο δεδομένων που θα μπορούσε να κατασκευάσει ένα σύστημα με ελάχιστο ρυθμό λάθους. Ένα τέτοιο σύστημα μπορεί εύκολα (με την χρήση βιβλιοθηκών του Kaldi) να ενσωματωθεί σε εφαρμογές αλληλεπίδρασης ανθρώπου μηχανής.

Βιβλιογραφία

[1]Eric R. Kandel, James Harris Schwartz, Thomas M. Jessell - Principles of neural science-McGraw-Hill, Health Professions Division (2000).
[2]McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity.
[3] A. M. Turing (1950), Computing Machinery and Intelligence.
[4] C. S. Strachey (1952), LOGICAL OR NON-MATHEMATICAL PROGRAMMES.
[5]A.L. Samuel (1959), Some Studies in Machine Learning Using the Game of Checkers.
[6]ROSENBLATT F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958 Nov;65(6):386-408. doi: 10.1037/h0042519. PMID: 13602029.
[7]Simon S.Haykin Neural Networks and Learning Machines,Third Edition (2009).
[8]Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. Neural Computation,
[9]Valotassiou, V.; Sifakis, N.; Tzavara, C.; Lykou, E.; Tsinia, N.; Kamtsadeli, V.; Sali, D.; Angelidis, G.; Psimadas, D.; Theodorou, E.; Tsougos, I.; Papageorgiou, S.G.; Georgoulias, P.; Papatriantafyllou, J. Anosognosia in Dementia: Evaluation of Perfusion Correlates Using 99mTc-HMPAO SPECT and Automated Brodmann Areas Analysis.*Diagnostics* 2022, *12*, 1136. https://doi.org/10.3390/diagnostics12051136.
[10]W. J. Zhang, G. Yang, Y. Lin, C. Ji and M. M. Gupta, "On Definition of Deep Learning," *2018 World Automation Congress (WAC)*, Stevenson, WA, USA, 2018, pp. 1-5, doi: 10.23919/WAC.2018.8430387.
[11]Daniel Jurafsky, James H. Martin - Speech and Language Processing_ An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition-Prentice Hall (2008).
[12]L.R. Rabiner, R.W. Schafer (1975), Theory and Applications of Digital Speech Processing, First Edition.
[13]Oppenheim, Alan V., Schafer, Ronald W. and Buck, John R.. *Discrete-Time Signal Processing*. Third : Prentice-hall, 2009.
[14]Νεοελληνική γραμματική Μανώλης Τριανταφιλίδης.
[15]The CMU Pronouncing Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.
[16]Jaynes, Edwin T.. “On the rationale of maximum-entropy methods.” *Proceedings of the. IEEE* 70 (1982): 939-952.
[17]Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. Comput. Linguist. 22, 1 (March 1996), 39–71.
[18]Alumäe, Tanel and Mikko Kurimo. “Domain Adaptation of Maximum Entropy Language Models.” *Annual Meeting of the Association for Computational Linguistics* (2010).
[19]Daniel Jurafsky, James H. Martin - Speech and Language Processing_ An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition-Third Edition draft (2023).
[20]Geoffrey J. McLachlan - Discriminant Analysis and Statistical Pattern Recognition-Wiley-Interscience (2004)
[21]R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition,"*[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, CA, USA, 1992, pp. 13-16 vol.1, doi: 10.1109/ICASSP.1992.225984
[22]Psutka, J.V. (2007). Benefit of Maximum Likelihood Linear Transform (MLLT) Used at Different Levels of Covariance Matrices Clustering in ASR Systems. In: Matoušek, V., Mautner, P. (eds) Text, Speech and Dialogue. TSD 2007. Lecture Notes in Computer Science(), vol 4629. Springer, Berlin, Heidelberg
[23]Maximum likelihood linear transformations for HMM-based speech recognition M.J.F. Gales (1998).

[24]Stolcke, Andreas, Luciana Ferrer, Sachin S. Kajarekar, Elizabeth Shriberg and Anand Venkataraman. “MLLR transforms as features in speaker recognition.”*Interspeech*(2005).
[25]I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction. Noor Salwani Ibrahim, Dzati Athiar Ramli (2018).
[26]i-Vectors in speech processing applications: a survey.Pulkit Verma, Pradip K. Das (2015).
[27]Dehak, N. (2009). Discriminative and generative approaches for longand short-term speaker characteristics modeling: Application to speaker verification. PhD thesis, Ecole de Technologie Superieure (Canada), aAINR50490.
[28]O. Glembek, L. Burget, P. Matějka, M. Karafiát and P. Kenny, "Simplification and optimization of i-vector extraction,"*2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 4516-4519, doi: 10.1109/ICASSP.2011.5947358.
[29]Automatic Speech Recognition A Deep Learning Approach, Dong Yu, Li Deng(2015)
[30]Purely sequence-trained neural networks for ASR based on lattice-free MMI.Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, Sanjeev Khudanpur (2016)
[31]H. Hadian, D. Povey, H. Sameti, J. Trmal and S. Khudanpur, "Improving LF-MMI Using Unconstrained Supervisions for ASR,"*2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, 2018, pp. 43-47, doi: 10.1109/SLT.2018.8639684.
[32]Deep Learning for NLP and Speech Recognition. Uday Kamath, John Liu, James Whitaker (2019).
[33]Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., Khudanpur, S. (2018) Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. Proc. Interspeech 2018, 3743-3747, doi: 10.21437/Interspeech.2018-1417
[34]Zhang, Zhen, Hao Huang, and Kai Wang. 2020. "Using Deep Time Delay Neural Network for Slot Filling in Spoken Language Understanding"*Symmetry* 12, no. 6: 993. https://doi.org/10.3390/sym12060993
[35]Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal.The fifth `CHiME’ Speech Separation and Recognition Challenge: Dataset, task and baselines Submitted to Interspeech, 2018.
[36]D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2011.
[37]D. Prabakaran and S. Sriuppili 2021”Speech Processing: MFCC Based Feature Extraction Techniques- An Investigation”
[38]Ole Morten Strand, Andreas Egeberg.”Cepstral mean and variance normalization in the model domain”(2004)