Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
Τομέας Επικοινωνιών, Ηλεκτρονικής & Συστημάτων Πληροφορικής

Σχεδίαση & Υλοποίηση Συστήματος Αναγνώρισης Ομιλητή

Διπλωματική Εργασία

Αθανάσιος Δ. Ανδριόπουλος

Επιβλέπων Καθηγητής
Παύλος-Πέτρος Σωτηριάδης
Αναπλ. Καθηγητής Ε.Μ.Π.

Εργαστήριο Ηλεκτρονικής
Αθήνα, Απρίλιος 2018

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
Τομέας Επικοινωνιών, Ηλεκτρονικής & Συστημάτων Πληροφορικής

Σχεδίαση & Υλοποίηση Συστήματος Αναγνώρισης Ομιλητή

Διπλωματική Εργασία

Αθανάσιος Δ. Ανδριόπουλος

Επιβλέπων Καθηγητής
Παύλος-Πέτρος Σωτηριάδης
Αναπλ. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 20η Απριλίου 2018

.....	.....	.....
Παύλος-Πέτρος Σωτηριάδης	Κιαμάλ Πεκμεστζή	Ανδρέας-Γεώργιος Σταφυλοπάτης
Αναπλ. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Εργαστήριο Ηλεκτρονικής
Αθήνα, Απρίλιος 2018

.....
Αθανάσιος Δ. Ανδριόπουλος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στη σύγχρονη πραγματικότητα των μέσων δικτύωσης, των έξυπνων συσκευών και της αξιοποίησης κάθε είδους πληροφορίας, η ασφάλεια συστημάτων κάθε φύσεως είναι ένα ζήτημα που πρέπει να μας απασχολεί ιδιαίτερα. Μια προσέγγιση ως προς την βελτίωση των συστημάτων ασφαλείας είναι η αξιοποίηση βιομετρικών χαρακτηριστικών όπως η εμφάνιση, η φωνή αλλά και ο συνδυασμός χαρακτηριστικών που αυτές οι κατηγορίες μας παρέχουν. Τέτοιου είδους χαρακτηριστικά παρουσιάζουν ιδιαίτερο ενδιαφέρων επειδή αντιγράφονται δύσκολα.
Στην παρούσα εργασία γίνεται σχεδίαση και υλοποίηση ενός συστήματος Αναγνώρισης Ομιλητή, ως χαρακτηριστικά αναγνώρισης βασισμένο σε βραχυπρόθεσμα φασματικά χαρακτηριστικά του ομιλητή, συγκεκριμένα έγινε χρήση των Mel-Frequency Cepstral Coefficients (MFCC). Για την λήψη της απόφασης στο σύστημα αυτό χρησιμοποιήθηκε ένα Αυτόνομο Νευρωνικό Δίκτυο (ANN) σε Feed-Forward μορφή ελαχιστοποιώντας την επεξεργαστική ισχύ που απαιτείται από το τελικό σύστημα.
Στην έκταση της διπλωματικής θα αναλυθούν αρχικά τα μέρη του συστήματος, έπειτα οι αλγόριθμοι ανάλυσης σήματος και εξόρυξης χαρακτηριστικών, ο αλγόριθμος εκπαίδευσης του νευρωνικού συστήματος, η γενικότερη μορφή του καθώς και η μεθοδολογία πάνω στην οποία δοκιμάστηκε το σύστημα. Σαν αποτέλεσμα θα δούμε αν ένα τέτοιο σύστημα αξιοποιώντας μόνο τα φωνητικά χαρακτηριστικά του χρήστη μπορεί να προσφέρει αξιόπιστα αποτελέσματα.

Λέξεις-Κλειδιά

Αναγνώριση Ομιλητή, Ανάλυση Σήματος, Mel-Frequency Cepstral Coefficiets(MFCC), Νευρωνικά Δίκτυα,Λεκτικά Ανεξάρτητη ΑΟ, Ολοκληρωμένα Συστήματα, Σχεδίαση PCB, Μικροεπεξεργαστής τύπου ARM

Abstract

Design and Implementation of Autonomous Text-Independent Speaker Recognition

Nowadays, our life is fixated around social media and smart devices which lead to the birth of the big data. As a result, the concept of security is threatened to a point there is a need to evolve our security measures. A widely accepted idea is to use biometric characteristics of the user as an authentication method. As biometric characteristics we define human traits such as voice, appearance, fingerprint or ear shape that differ from one human to an other. The reason biometric characteristics are a viable solution lies in the effort needed in order to copy all those traits.

In this diploma thesis, a text-independent speaker recognition system is going to be designed and implemented. Utilizing a short term spectral characteristic extraction method, the Mel-Frequency Cepstral Coefficients (MFCC) which is considered a state-of-the-art method for speech processing. As a classifier an Artificial Neural Network (ANN) is going to be used in order to classify the extracted characteristics. Furthermore, we used a metric utilizing the number of identified inputs for each user in order make a decision.

In Conclusion, in this diploma thesis there are several aspects to examine the system such as the parts used, the desgning process, the algorithm for signal processing to extract the voice features, the training algorithm of our neural network and the evaluation process for our system. So, our results will define if a voice characteristics alone can be used as a security system.

Keywords

Speaker Recongition System(ASR), Mel-Frequency Cepstral Coefficiets(MFCC), Artificial Neural Network(ANN), Signal Processing, Text-Independent ASR, Embedded Systems, PCB Design, ARM Processor

Στοιχεία Συγγραφέα

Ο Ανδριόπουλος Αθανάσιος είναι προπτυχιακός φοιτητής στο τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Ηλεκτρονικών Υπολογιστών του Εθνικού Μετσοβίου Πολυτεχνείου.

e-mail: andriopoulosthanassis@gmail.com

Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω όλους εκείνους που ήταν δίπλα μου, καθένας με τον δικό του τρόπο, κατά την εκπόνηση της διπλωματικής μου εργασίας. Συγκεκριμένα, θα ήθελα να ευχαριστήσω:

Τον καθηγητή μου και επιβλέπων στη διπλωματική μου εργασία, κο Πέτρο Παύλο Σωτηριάδη, για την εμπιστοσύνη που μου έδειξε, την ευκαιρία που μου έδωσε να ασχοληθώ με αυτό το θέμα διπλωματικής εργασίας και την γνώση που έλαβα από τα μαθήματά του.

Τον Διδακτορικό φοιτητή κ. Κωσταντίνο Παπαφώτη για την πολύτιμη καθοδήγηση και στήριξη κατά τη διάρκεια αυτής της διπλωματικής εργασίας καθώς και τους υπόλοιπους διδακτορικούς φοιτητές του Circuits & Systems Group Κ. Ούστογλου, Κ. Ασημακόπουλο, Χ. Δήμα, Ν. Χατζιγεωργίου, Δ. Μπαξεβανάκη και Ν. Τέμενο για το υπέροχο κλίμα συνεργασίας που έχουν δημιουργήσει στο εργαστήριο αλλά και για τις πολύτιμες συμβουλές τους.

Τους φίλους μου, για την κατανόηση και την στήριξη τους κατά την εκπόνηση αυτής της διπλωματικής εργασίας.

Την οικογένεια μου, για την πλήρη στήριξη τους σε όλες μου τις επιλογές όλα αυτά τα χρόνια, την κατανόηση τους, και την δυνατότητα που μου προσέφεραν ώστε να βρίσκομαι σήμερα εδώ.

Ακρωνύμια

ASR	Automatic Speaker Recognition
ANN	Artifitial Neural Network
CNN	Convolutional Neural Network
PCB	Printed Circuit Board
MFCC	Mel Frequency Cepstral Coefficients
LPC	Linear Predictive Coefficients
FFT	Fast Fourier Transformation
ADC	Analog to Digital Converter
SDIO	Secure Digital Input Output
MFCC	Mel Frequency Cepstral Coefficients
FatFs	File Allocation Table File System
GMM	Gaussian Mixture Models
HMM	Hidden Markov Model
SPI	Serial Periferal Interface
MPU	Mircoprocessor Unit
UART	Universal Asynchronous Receive-Transmit
PCM	Pulse-Code Modulation
DCT	Discrete Cosine Transformation
I^2S	Inter-IC Sound
DMA	Direct Memory Access
SWD	Serial Wire Debugging

Πίνακας 1: Πίνακας Ακρωνύμιων

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1 Εισαγωγή

1.1 Γενικά

Η ασφάλεια τόσο σε προσωπικό όσο και σε επαγγελματικό επίπεδο αποτελεί ένα ζήτημα με ιδιαίτερο ενδιαφέρον. Από τις κλειδαριές και τους φύλακες εισόδου μέχρι το σύστημα ασφαλείας του Fort Knox, οι άνθρωποι πάντα ένιωθαν την ανάγκη για ασφάλεια. Ο Abraham Maslow ερευνώντας την ιεραρχία των ανθρωπίνων αναγκών καταλήγει ότι η ασφάλεια αποτελεί την δεύτερη πιο σημαντική κατηγορία αναγκών μετά τις φυσιολογικές ανάγκες.
Ειδικότερα, σε επαγγελματικούς χώρους η ασφάλεια τόσο του εξοπλισμού όσο και των εργαζομένων είναι θέμα μέγιστης σημασίας για τη βιωσιμότητα της επιχείρησης. Αποτέλεσμα αυτού, είναι η επένδυση μεγάλων κεφαλαίων, στην κατασκευή συστημάτων παρακολούθησης των χώρων της, στη στελέχωση των κτηρίων με ανθρώπινο δυναμικό με σκοπό την προστασία του κτηρίου και στη δημιουργία συστήματος δικαιωμάτων πρόσβασης για τους χώρους της.

1.2 Ορισμός του προβλήματος

Σήμερα, όντας σε μια εποχή όπου παρέχουμε υπερβολικά πολλά προσωπικά δεδομένα, στον οποιοδήποτε, με τη χρήση των μέσων κοινωνικής δικτύωσης(social media) τη στιγμή που, ο μέσος όρος των χρηστών του διαδικτύου δεν έχει πλήρη επίγνωση των κινδύνων που υπάρχουν, η υποκλοπή προσωπικών δεδομένων και κάθε λογής κωδικών πρόσβασης δεν πρέπει να μας κάνει εντύπωση.
Όμως, η χρήση ενός αλφαριθμητικού κωδικού δεν είναι ο μόνος τρόπος ταυτοποίησης κάποιου χρήστη. Στο σημείο αυτό χρήσιμο θα ήταν να δούμε τι χαρακτηριστικά μπορούμε να αντλήσουμε από τον χρήστη. Αρχικά, την εμφάνιση του και τον τρόπο που μιλάει. Έπειτα, τις κινήσεις του και τέλος το τι λέει. Άρα, με χρήση των παραπάνω παρατηρήσεων για τον χρήστη πρέπει να κρίνουμε αν έχει δικαιώματα πρόσβασης σε αυτό τον χώρο. Αυτή τη στιγμή έχουμε στη διάθεση μας αρκετά ικανούς αισθητήρες που σε συνδυασμό με εξίσου ικανούς αλγορίθμους ανάλυσης δεδομένων είμαστε σε θέση να πετύχουμε ένα ιδιαίτερα αξιόπιστο σύστημα ασφαλείας.
Επομένως, ο τρόπος να προστατέψουμε τον χώρο που μας ενδιαφέρει είναι η χρήση ενός πολυπαρα-

γοντικού(Multifactor) συστήματος ασφαλείας. Ένα τέτοιο σύστημα πρέπει να αξιοποιεί κατάλληλα τα βιομετρικά στοιχεία του ατόμου εστιάζοντας σε χαρακτηριστικά που δεν μπορεί κάποιος να "αντιγράψει". Κάποια τέτοια χαρακτηριστικά είναι οι απόσταση μεταξύ των ματιών, της ίριδας του ματιού και ακουστικών χαρακτηριστικών της φωνής.

1.3 Στόχος της Διπλωματικής

Στόχος της παρούσας διπλωματικής εργασίας είναι ο σχεδιασμός και η υλοποίηση ενός συστήματος αναγνώρισης εστιάζοντας μόνο σε ηχητικές πληροφορίες για τον χρήστη. Συγκεκριμένα, στοχεύουμε το σύστημα μας να μπορεί να αναγνωρίσει ποιος, από τους γνωστούς χρήστες για αυτό, μόλις ηχογραφήθηκε.
Φυσικά, κάτι τέτοιο προϋποθέτει την εξόρυξη ορισμένων χαρακτηριστικών της φωνής ικανών να διαχωρίσουν τους χρήστες του συστήματος. Προσφέροντας στο σύστημα μια επιπλέον αξιοπιστία και καθιστώντας σχεδόν αδύνατη την πλαστογράφηση αυτών των χαρακτηριστικών από άλλους χρήστες του συστήματος.
Ακόμη, στην παρούσα εργασία χρειάστηκε πέρα από τη σχεδίαση του συστήματος και η κατασκευή ενός προτύπου συστήματος για την δοκιμή της εν λόγω εφαρμογής σε πραγματικές συνθήκες.

1.4 Οργάνωση Κεφαλαίων

Η συνέχεια της διπλωματικής οργανώνεται ως εξής:

- ***Κεφάλαιο 2 - Θεωρητικό Υπόβαθρο:*** Στο κεφάλαιο 2 περιγράφονται αναλυτικά όλες οι έννοιες που σχετίζονται με την Ανάλυση Σήματος καθώς και τη μορφή γενικότερη μορφή ενός Αυτόνομου Νευρωνικού Δικτύου.Πιο συγκεκριμένα, το κεφάλαιο ξεκινάει με τον ορισμό ενός σήματος φωνής και πως αυτό παράγεται από τον ανθρώπινο παράγοντα(Κεφ. 2.1). Στη συνέχεια, αναπτύσσονται τα χαρακτηριστικά που μπορούμε να εξορύξουμε από ένα φωνητικό σήμα και παρουσιάζεται εν συντομία η μέθοδος εξόρυξης χαρακτηριστικών που χρησιμοποιήθηκε σε αυτή τη διπλωματική(Κεφ 2.2).Έπειτα, ορίζονται οι κατηγορίες προβλημάτων που μπορούν να λυθούν με χρήση των παραπάνω χαρακτηριστικών και γίνεται ορισμός και διαχωρισμός κατηγοριών αναγνώρισης Ομιλητή(Κεφ 2.3). Ακολουθεί μια εισαγωγή στη μορφή και τον τρόπο λειτουργίας ενός Αυτόνομου Νευρωνικού Δικτύου(Κεφ 2.4). Τέλος γίνεται μαι βιβλιογραφική αναφορά στον τρόπο που έχει αντιμετωπιστεί το πρόβλημα Αναγνώρισης Ομιλητή μέχρι σήμερα(Κεφ. 2.5).

- ***Κεφάλαιο 3 - Τεχνολογίες:*** Στο κεφάλαιο 3 περιγράφεται η διαδικασία σχεδίασης πλακέτας(PCB) αναφέρονται τα εξαρτήματα που χρησιμοποιήθηκαν και γίνεται περιγραφή των πρωτοκόλλων επικοινωνίας μεταξύ του επεξεργαστή και τον περιφερειακών(Peripherals). Εκτενέστερα, το κεφάλαιο ξεκινάει με τον ορισμό των εξαρτημάτων που χρησιμοποιήθηκαν(Κεφ. 3.1). Ακολουθούν τα σχηματικά της διάταξης τα οποία έχουν κατηγοριοποιηθεί ανάλογα με

την χρήση τους στο σύστημα αυτό(Κεφ 3.2).Στη συνέχεια, παρουσιάζεται τα 2D και 3D Μοντέλα της πλακέτας(PCB) δημιουργώντας έτσι μια εικόνα στον αναγνώστη για το πώς θα μοιάζει το τελικό σύστημα(Κεφ 3.3). Τέλος, αναλύονται τα πρωτόκολλα επικοινωνίας ως προς το πως χρησιμοποιήθηκαν σε αυτό το σύστημα αλλά και ως προς τις γενικότερες δυνατότητες τους(Κεφ. 3.4).

- ***Κεφάλαιο 4 - Προσέγγιση Συστήματος:*** Στο κεφάλαιο 4 γίνεται αναλυτική περιγραφή του συστήματος, του αλγορίθμου που χρησιμοποιήθηκε για την ανάλυση σήματος, του σταδίου εκπαίδευσης του νευρωνικού δικτύου και των κανόνων ορθής λειτουργίας του συστήματος. Αναλυτικότερα, το κεφάλαιο ξεκινάει με τη γενική περιγραφή του συστήματος που αναπτύχθηκε(Κεφ 4.1). Ακολουθεί, η εκτενέστερη περιγραφή του αλγόριθμου ανάλυσης σήματος που χρησιμοποιήθηκε εστιάζοντας στις βελτιστοποιήσεις που έγιναν ώστε να επιταχύνουμε την διαδικασία(Κεφ 4.2). Στη συνέχεια, περιγράφεται η μορφή του νευρωνικού δικτύου καθώς και η διαδικασία εκπαίδευσης του(κεφ 4.3). Τέλος, διατυπώνονται οι κανόνες λειτουργίας που διέπουν το σύστημα αυτό εξασφαλίζοντας τον ομαλότερη του λειτουργία(Κεφ. 4.4).

- ***Κεφάλαιο 5 - Πειράματα και Αποτελέσματα:*** Στο Κεφάλαιο 5 παρατίθενται τα πειράματα που εκτελέστηκαν στα πλαίσια της διπλωματικής. Τα πειράματα αφορούν τις διάφορες δοκιμές που έγιναν για την εύρεση των καλύτερων παραμέτρων εισόδου με στόχο την βέλτιστη εκπαίδευση του νευρωνικού δικτύου. Τα αποτελέσματα αφορούν την παρουσίαση της ομαδοποίησης που έκανε το νευρωνικό δίκτυο, ανάλογα με τις παραμέτρους εισόδου, να παρουσιάζει καλύτερα αποτελέσματα αλλά και τον ορισμό κατάλληλου ποσοστού επιτυχίας για να φτάσουμε στην αναγνώριση.

- ***Κεφάλαιο 6 - Συμπεράσματα & Μελλοντικές Επεκτάσεις:*** Η διπλωματική εργασία ολοκληρώνεται με το Κεφάλαιο 6, όπου παρατίθενται τα συμπεράσματα που εξήχθησαν και προτείνονται κάποιες ιδέες για μελλοντική εργασία πάνω στο αντικείμενο που πραγματεύεται.

2 Θεωρητικό Υπόβαθρο

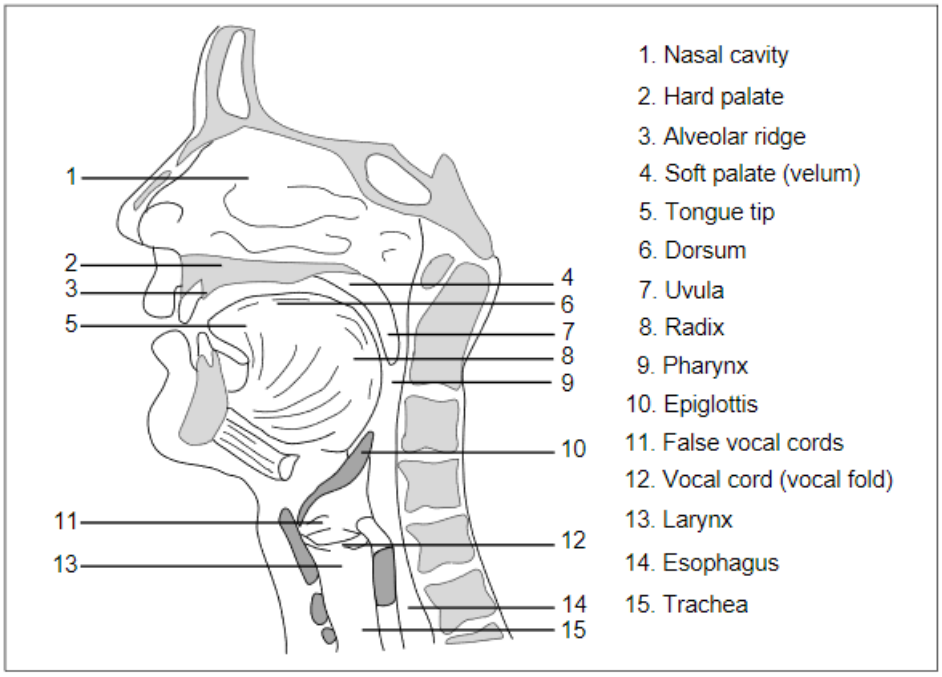
Στο σημείο αυτό κρίνεται αναγκαίο, από τον συγγραφέα, με γνώμονα την καλύτερη κατανόηση του συστήματος μια εισαγωγή στον τρόπο που παράγεται η ανθρώπινη ομιλία. Αναλυτικότερα, παρακάτω αναφέρονται τα όργανα του ανθρώπινου σώματος που παράγουν τη φωνή και εξετάζονται σε μορφή συστήματος, με σκοπό την παραμετροποίηση του.

2.1 Ανρθώπινη Φωνή

Η παραγωγή της ανθρώπινης φωνής μπορεί χονδροειδώς να χωριστεί σε τρία μέρη: τους πνεύμονες, τις φωνητικές χορδές και τη φωνητική οδό. Οι πνεύμονες λειτουργούν ως η πηγή της ροής αέρα και πίεσης. Τη στιγμή που παράγεται η ομιλία, οι φωνητικές χορδές ανοίγουν και κλείνουν περιοδικά με αποτέλεσμα να τη μετατροπή της ροής αέρα, από τους πνεύμονες σε μια ακολουθία παλμών, η οποία λειτουργεί ως η ακουστική διέγερση και πηγή της ομιλίας. Η φωνητική οδός είναι ένα σύνολο από κοιλότητες πάνω από τις φωνητικές χορδές μέχρι τα χείλη και τη μύτη. Λειτουργεί σαν ένα ακουστικό φίλτρο το οποίο μορφοποιεί το φάσμα του ήχου. Τελικά, η φωνή διαδίδεται μέσω του αέρα στον γύρω χώρο μέσω των χειλιών και της μύτης. Ο παραπάνω μηχανισμός απεικονίζεται στο (Σχήμα 2.1).

2.1.1 Φωνητικές Χορδές

Οι φωνητικές χορδές αποτελούνται από μαλακό, ελαστικό ιστό που βρίσκονται οριζόντια μέσα στον λάριγκα(Larynx). Ο χώρος μεταξύ των φωνητικών χορδών ονομάζεται γλωττίδα(glottis). Ο μηχανισμός υποστηρίζεται και ελέγχεται από πολλούς χόνδρους και μύες.



Σχήμα 2.1: Μηχανισμός Παραγωγής Ανθρώπινης Φωνής

Οι φωνητικές χορδές είναι συνδεδεμένες με τον θυρεοειδή χόνδρο στα επάνω του άκρα και στα κάτω άκρα του με τους αριτενοειδείς χόνδρους. Οι αριτενοειδείς χόνδροι μπορούν να κινηθούν διαφορετικά από τον μυ του λάρυγγα, επιτρέποντας έτσι το πλάτος των φωνητικών χορδών να μεταβάλλεται. Κατά τη διαδικασία της αναπνοής οι φωνητικές χορδές είναι τελείως ανοικτές σε αντίθεση με τη διαδικασία της ομιλίας κατά την οποία πλησιάζουν μεταξύ τους.

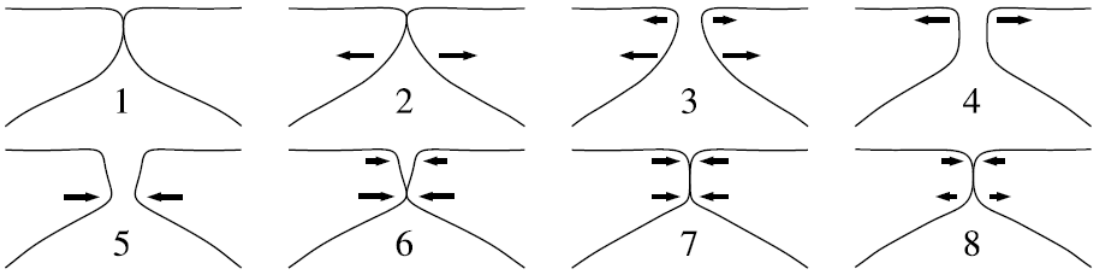
Όταν περνάει αέρας από τους πνεύμονες τη στιγμή που οι φωνητικές χορδές είναι κοντά μεταξύ τους τότε αυτές αρχίζουν να ταλαντεύονται. Αυτή η δόνηση μετατρέπει τη ροή αέρα σε μια ακολουθία παλμών που αποτελούν πηγή της φωνής.

Το μήκος και η τάση των φωνητικών χορδών μπορεί επίσης να ελεγχθεί από μυικές κινήσεις με στόχο να ρυθμίσουν την συχνότητα της ταλάντωσης και την ποιότητα της φωνής. Οι δονούμενες φωνητικές χορδές έχουν μήκος περίπου 16 mm για τον μέσο ενήλικο άνδρα και 10 mm για τη μέση ενήλικη γυναίκα. Ακόμη, οι φωνητικές χορδές μπορούν να τεντωθούν μερικά χιλιοστά(mm) κατά τις κινήσεις των μυών του λάρυγγα.

Η συχνότητα της γλωττιδικής(glottal)ταλάντωσης καθορίζει την θεμελιώδη συχνότητα της ομιλίας, η οποία στο εξής θα αναφέρεται ως f_0. Η μέση συχνότητα f_0 για έναν άνδρα εκτιμάται κοντά στα 120 Hz, 200 Hz για μια γυναίκα και 300 για παιδιά. Το εύρος της θεμελιώδους συχνότητας είναι αρκετά μεγάλο: f_0 μικρότερη από 100 Hz δεν θεωρούνται ασυνήθιστες για άνδρες, όμως ένας τενόρος μπορεί να φτάνει συχνότητες πάνω από 600 Hz. Στις γυναίκες η χαμηλότερη f_0 είναι κάτω από τα 150 Hz ενώ μια σοπράνο μπορεί να ξεπεράσει ακόμη και τα 1300 Hz. [1]

Οι φωνητικές χορδές αποτελούνται από αρκετά επίπεδα με διαφορετική σκληρότητα το κάθε ένα. Το υψηλότερο επίπεδο(επιθήλιο) αποτελεί την επιφάνεια βλεννογόνου ιστού, ο οποίος μπορεί να χωριστεί σε τρία επίπεδα. Η σκληρότητα του βλεννογόνου ιστού αυξάνεται καθώς διεισδύουμε στο

εσωτερικό της. Στο εσωτερικό στρώμα των φωνητικών χορδών συναντάμε έναν ιδιαίτερα ελαστικό μυ(musculus thyroarytenoids). Η δόνηση αυτή δημιουργείται κυρίως μέσα στον βλεννογόνο ιστό. Η ταλάντωση αυτή καθ' αυτή δεν χρειάζεται κάποια μυική κίνηση. Διατηρείται από τις διακυμάνσεις της πίεσης του αέρα και την ελαστικότητα των ιστών. Ωστόσο, οι μύες λειτουργούν ως ρυθμιστές της σκληρότητας(άρα και της δόνησης)και της απόστασης μεταξύ των φωνητικών χορδών.[2]
Παρατηρήσεις δονήσεων των φωνητικών χορδών με διάφορες μεθόδους έχουν δείξει πως τα πάνω και κάτω μέρη των φωνητικών χορδών δεν ταλαντώνονται συμφασικά. Συγκεκριμένα, ένα κύμα μεταδίδεται από το χαμηλότερο κομμάτι τους προς το υψηλότερο, δημιουργώντας έτσι μια κυματόμορφη κίνηση η οποία διασχίζει προς τα πάνω το εξωτερικό στρώμα των φωνητικών χορδών. Αυτό το φαινόμενο ονομάζεται βλεννογόνο κύμα(mucosal wave)και αναπαριστάται στο Σχήμα 2.2 [3].



Σχήμα 2.2: Ιδεατός κύκλος δονήσεων των φωνητικών χορδών

Τέλος, η οριζόντια κίνηση των φωνητικών χορδών συχνά δεν συμβαίνει ταυτόχρονα σε ολόκληρο το μήκος των φωνητικών χορδών προς το οριζόντιο επίπεδο. Αντίθετα, αυτή η οριζόντια κίνηση πολλές φορές θυμίζει την κίνηση ενός φερμουάρ [4].Το παραπάνω σχήμα απεικονίζει το βλεννογόνο κύμα(mucosal wave)εστιάζοντας στη διαφορά φάσης του άνω και κάτω μέρους των φωνητικών χορδών.[3]

2.1.2 Φωνητική Οδός

Οι δονήσεις των φωνητικών χορδών παρέχουν μια πλούσια σε φασματικό περιεχόμενο διέγερση το οποίο έπειτα διαμορφώνεται κατάλληλα από τις κοιλότητες του υπολοίπου συστήματος. Η σωληνοειδής μορφή του λάρυγγα μαζί με τον φάρυγγα και τη στοματική κοιλότητα σχηματίζουν τη φωνητική οδό[5]. Το μήκος αυτής είναι περίπου 17 cm για τους άνδρες, 15 cm για τις γυναίκες και 14 cm για παιδιά [6]. Σύμφωνα με άλλους ορισμούς σε αυτή ανήκει και η ρινική κοιλότητα [2].
Ο ρόλος της φωνητικής οδού είναι ένα μεταβλητό ακουστικό φίλτρο που μεταβάλει το φάσμα της διέγερσης του σήματος. Κάθε συλλαβή παράγει ένα χαρακτηριστικό φασματικό προφίλ. Αυτό το φασματικό προφίλ εξαρτάται από το σχήμα της φωνητικής οδού, το οποίο καθορίζεται από τη θέση της γλώσσας, τον χειλιών και τον ουρανίσκο.

2.2 Χαρακτηριστικά Φωνής

Στοιχειώδης κρίνεται η ανάλυση των ειδών χαρακτηριστικών που μπορούμε να εξορύξουμε κάθε σήμα ομιλίας. Τα χαρακτηριστικά αυτά μπορούν να χωριστούν σε:
(1) ***Βραχυπρόθεσμα Φασματικά Χαρακτηριστικά***
(2) ***Χαρακτηριστικά Πηγής***
(3) ***Φασματοχρονικά και Προσωδιακά Χαρακτηριστικά***
(4) ***Υψηλού Επιπέδο Χαρακτηριστικά***.
Ξεκινώντας με τα βραχυπρόθεσμα χαρακτηριστικά όπως υποδεικνύει και το όνομα αναφέρονται σε χαρακτηριστικά που μπορούμε να εξορύξουμε από μικρά σε χρόνο διαστήματα (20-30 ms). Συνήθως αναπαριστούν το βραχυπρόθεσμο φασματικό περιεχόμενο και θα μπορούσαμε να τα συσχετίσουμε με τον τόνο, την ποιότητα, τους "χρωματισμούς" της φωνής, καθώς και τις ιδιότητες συντονισμού της φωνητικής οδού. Τα χαρακτηριστικά Πηγής με τη σειρά τους προσδιορίζουν τη ροή της ομιλίας. Τα προσωδιακά και φασματοχρονικά χαρακτηριστικά ασχολούνται με χαρακτηριστικά που μπορούμε να εξορύξουμε από εκατοντάδες ms, για παράδειγμα προσωδιακά χαρακτηριστικά θεωρούνται ο ρυθμός και ο τονισμός της ομιλίας. Κλείνοντας, τα χαρακτηριστικά υψηλού επιπέδου που προσπαθούν να κρατήσουν πληροφορίες σε επίπεδο συζήτησης από τη φωνή κάποιου. Παράδειγμά αυτών αποτελούν η επαναλαμβανόμενη χρήση λέξεων και εκφράσεων από τον ομιλητή.

Σχήμα 2.3: Χαρακτηριστικά Ανάλογα με τη Φυσική τους Σημασία

Όπως γίνεται φανερό από το παραπάνω διάγραμμα η επιλογή κατάλληλων χαρακτηριστικών ανάλογα με την εφαρμογή εξαρτάται από αρκετούς παράγοντες.
Μια μέθοδος εξόρυξης βραχυπρόθεσμων φασματικών χαρακτηριστικών είναι η χρήση των Mel Frequency Cepstral Coefficients (MFCC). Για την εξόρυξη αυτών πρέπει:
(1) Να μετατρέψουμε το σήμα φωνής, το οποίο εκ φύσεως αλλάζει δυναμικά με την πάροδο του χρόνου, σε στατικό. Αυτό επιτυγχάνεται "τεμαχίζοντας" το σήμα σε επικαλυπτόμενα σήματα των

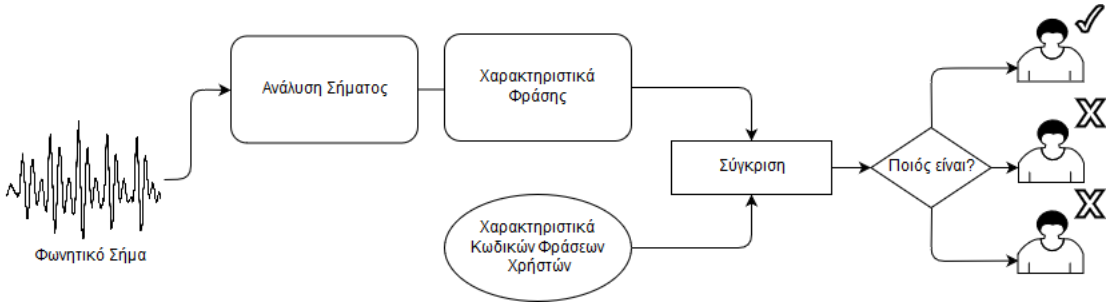
20-30 ms και εφαρμόζοντας κάποιο παράθυρο σε αυτό, συνήθως του Hamming.
(2) Την μετατροπή αυτών στο πεδίο της συχνότητας ώστε να μπορούμε να βρούμε το φασματικό τους περιεχόμενο το οποίο επιτυγχάνεται εφαρμόζοντας FFT στα frames.
(3) Περνώντας το φασματικό περιεχόμενο του κάθε frame μέσα από ένα ψυχοακουστικό φίλτρο στην κλίμακα Mel (Η κλίμακα Mel είναι γραμμική μέχρι τα 1000 Hz και λογαριθμική έπειτα προσπαθώντας να προσομοιώσει την ανθρώπινη ακοή)και στη συνέχεια υπολογίζουμε τον λογάριθμο των αποτελεσμάτων.
(4) Εφαρμόζοντας Inverse Discrete Cosine Transformation (IDCT) καταλήγουμε στους εν λόγω συντελεστές. Χρήσιμες είναι επίσης οι μεταβολές των συντελεστών αυτών, ως προς τα προηγούμενα δύο από αυτό χαρακτηριστικά, οι οποίες παρουσιάζουν την δυναμική των συντελεστών αυτών.

2.3 Κατηγορίες Αναγνώρησης Ομιλητή

Έχοντας αναλύσει το σύστημα παραγωγής της ανθρώπινης φωνής και τα χαρακτηριστικά που μπορούμε εξορύξουμε από αυτή μπορούμε να εξετάσουμε τα είδη αναγνώρισης ομιλητή που υπάρχουν. Ιδιαίτερη σημασία έχει να τονίσουμε ότι εξετάζουμε ένα σύστημα με αποστολή να συγκρίνει τα χαρακτηριστικά της φωνής του άγνωστου ομιλητή με τα χαρακτηριστικά των γνωστών για το σύστημα ομιλητές και να κάνει μια εκτίμηση σε ποιόν από τους γνωστούς σε αυτό χρήστες μοιάζει περισσότερο. Οι δύο μεγαλύτερες κατηγορίες τέτοιων συστημάτων είναι τα Λεκτικά Εξαρτώμενα συστήματα και τα Λεκτικά Ανεξάρτητα συστήματα. Το καθένα με πλεονεκτήματα και μειονεκτήματα που θα αναλυθούν στις δύο επόμενες υποενότητες.

2.3.1 Λεκτικά Εξαρτώμενο

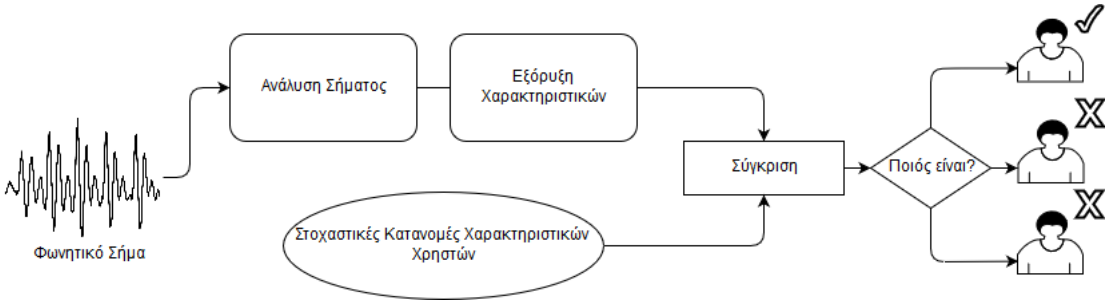
Ένα λεκτικά εξαρτώμενο σύστημα θα μπορούσε να παρομοιαστεί και με ένα σύστημα αναγνώρισης ομιλίας ή αναγνώρισης λέξεων. Ουσιαστικά, έχουμε ένα σύστημα το οποίο εξετάζει αποκλειστικά τι ειπώθηκε από τον ομιλητή ώστε να αναγνωρίσει την ταυτότητα του. Τέτοια συστήματα για να λειτουργήσουν χρειάζονται τη χρήση μιας συγκεκριμένης φράσης από κάθε χρήστη ώστε να μπορεί με αναφορά σε αυτή να συγκρίνει το περιεχόμενο του ηχογραφημένου σήματος ομιλίας που εξετάζει. Συνοψίζοντας, ένα τέτοιου είδους σύστημα βασίζεται σε έναν προσωπικό κωδικό για τον κάθε χρήστη τον οποίο πρέπει το σύστημα να βρίσκεται σε θέση να αναγνωρίσει. Το Σχήμα 2.4 μπορεί να δώσει μια πιο εύπεπτη οπτική αναπαράσταση ενός τέτοιου συστήματος.



Σχήμα 2.4: Αναπαράσταση ενός Λεκτικά Εξαρτώμενου Συστήματος Αναγνώρισης Ομιλητή

2.3.2 Λεκτικά Ανεξάρτητο

Ένα λεκτικά ανεξάρτητο σύστημα βασίζει τη λειτουργικότητα του σε στοχαστικά μοντέλα για τον κάθε χρήστη, τα οποία δημιουργούνται με τη χρήση κάποιου είδους ταξινομητή(Classifier), τα οποία συγκρίνει με τα χαρακτηριστικά που εξόρυξε, από το ηχογραφημένο σήμα ομιλίας, και σύμφωνα με το κριτήριο μέγιστης ομοιότητας διαλέγει έναν από τους χρήστες του. Μεγάλο πλεονέκτημα και ταυτόχρονα μειονέκτημα ενός τέτοιου συστήματος είναι ότι εξετάζει "το πως λέγεται κάτι", ο λόγος που αυτό θεωρείται πλεονέκτημα είναι ότι δίνει σημασία σε οργανικά χαρακτηριστικά του ομιλητή και μειονέκτημα επειδή είναι ευαίσθητο σε μια οποιαδήποτε ηχογράφηση του χρήστη αν αυτός είναι ο μοναδικός τρόπος ταυτοποίησης. Ακόμη, ένα τέτοιο σύστημα παρόλο που χρειάζεται μεγαλύτερο δείγμα φωνής του κάθε χρήστη ώστε να εκπαιδευτεί σωστά τελικά απαιτεί την ελάχιστη δυνατή προσπάθεια από τους χρήστες για την ορθή του λειτουργία. Τέλος, ένα τέτοιο σύστημα ασχολείται με την κατανομή των των χαρακτηριστικών στον χώρο σε αντίθεση με τα λεκτικά εξαρτώμενα συστήματα που ασχολούνται κυρίως με την σχέση μεταξύ των χαρακτηριστικών και με τη σειρά που αυτά εμφανίζονται. Το Σχήμα 2.5 αναπαριστά ένα τέτοιο σύστημα.



Σχήμα 2.5: Αναπαράσταση ενός Λεκτικά Ανεξάρτητου Συστήματος Αναγνώρισης Ομιλητή

2.4 Αυτόνομο Νευρωνικό Δίκτυο

Η προηγούμενη ενότητα ασχέτως του είδους συστήματος κάνει προφανή την ανάγκη για τη χρήση κάποιου ταξινομητή ώστε να κατασκευάσουμε τα μοντέλα που χρησιμοποιούνται ως αναφορά για την λήψη της τελικής απόφασης του συστήματος. Για αυτό τον σκοπό έχουν δοκιμαστεί, με εξίσου αξιόλογα αποτελέσματα, αρκετές μέθοδοι. Μερικά παραδείγματα αποτελούν Συνδυαστικά Γκαουσιανά Μοντέλα (GMM's),[7] Κρυφά Μοντέλα Μάρκοβ (HMM's), Κβαντοποίηση Φορέα(VQ) [8] και τέλος Αυτόνομα Νευρωνικά Δίκτυα (ANN's) με τα οποία θα ασχοληθούμε σε αυτό το κεφάλαιο.

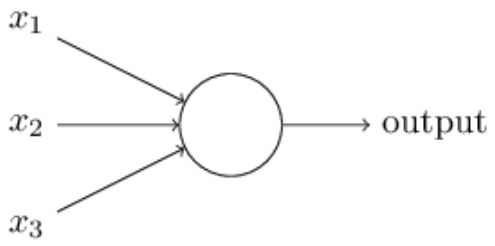
2.4.1 Εισαγωγή

Ένα νευρωνικό δίκτυο έχει ως αποστολή με δεδομένο έναν σύνολο εισόδων, που εισέρχονται σε αυτό κατά ομάδες, για τις οποίες ξέρει τα σωστά αποτελέσματα να προσαρμόσει τις τιμές των βαρών, που έχει σε κάθε μία από τις εισόδους του κάθε νευρώνα, ώστε μετά από ορισμένο αριθμό επαναλήψεων αυτού του συνόλου εισόδων να είναι σε θέση να αναγνωρίσει το σωστό αποτέλεσμα

για εισόδους που δεν έχει συναντήσει μέχρι στιγμής κατά την εκπαίδευση του.
Η εκπαίδευση του βασίζεται στην σύγκριση των αποτελεσμάτων του νευρωνικού δικτύου με τα γνωστά αποτελέσματα και την μεταφορά αυτού του σφάλματος μέσα από μια συνάρτηση κόστους στα πιο πίσω επίπεδα του δικτύου μεταβάλλοντας τα βάρη του κατάλληλα αποσκοπώντας στην μείωση του σφάλματος.
Αποτελείται από το επίπεδο εισόδου στο οποίο εισέρχονται στο δίκτυο οι ομάδες του συνόλου εισόδων, το κρυφό επίπεδο το οποίο μπορεί να είναι και περισσότερα από ένα και το επίπεδο εξόδου στο οποίο μπορούμε να δούμε το αποτέλεσμα. Σε αυτό το σημείο θα γίνει παρουσίαση των παραπάνω εννοιών ενός πλήρως συνδεδεμένου Αυτόνομου Νευρωνικού Δικτύου.

2.4.2 Μαθηματικό Μοντέλο Δυαδικού Νευρώνα

Η παρουσίαση αυτή θα ξεκινήσει από το πιο βασικό κομμάτι τον απλό δυαδικό νευρώνα με σιγμοειδείς εισόδους όπως τον όρισε ο Frank Rosenblatt.



Σχήμα 2.6: Δυαδικός Νευρώνας

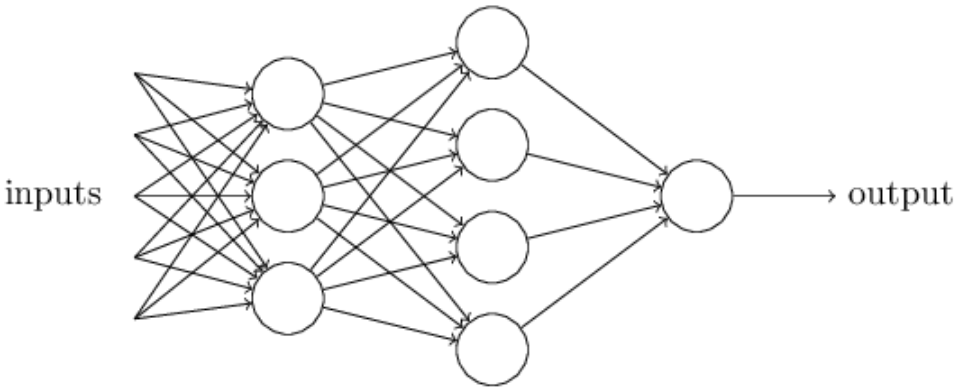
Ο παραπάνω νευρώνας χάρη απλότητας έχει μόνο τρεις εισόδους, x_1,x_2,x_3. Ο Rosenblatt πρότεινε έναν απλό κανόνα ώστε να υπολογιστεί το αποτέλεσμα(output). Ορίζοντας τα βάρη,w_1,w_2,w_3,τα οποία είναι πραγματικοί αριθμοί και εκφράζουν την σημαντικότητα της εισόδου ως προς το αποτέλεσμα της εισόδου. Στην πραγματικότητα η δουλειά που κάνει ο νευρώνας είναι να υπολογίζει το άθροισμα των εισόδων πολλαπλασιασμένο με το αντίστοιχο βάρος και να το συγκρίνει με κάποιο όριο κάτω από το οποίο ο νευρώνας ως έξοδο έχει την τιμή μηδέν, αντίθετα όταν αυτό το άθροισμα είναι πάνω από αυτό το όριο τότε η έξοδος είναι η μονάδα.

$$Output = \begin{cases} 0 & \text{if } \sum_{i=1}^{3} w_i x_i \leq \text{ Threshold} \\ 1 & \text{if } \sum_{i=1}^{3} w_i x_i > \text{Threshold} \end{cases}$$

Το παραπάνω αποτελεί το μαθηματικό μοντέλο στο οποίο βασίζονται οι νευρώνες. Επομένως, ένα νευρωνικό δίκτυο αποτελείται από ένα σύνολο νευρώνων κατά μια συγκεκριμένη τοπολογία.

2.4.3 Βασική Τοπολογία

Η τοπολογία ενός βασικού Αυτόνομου Νευρωνικού Δικτύου αποτελείται από 3 επίπεδα: (1) ***Το Επίπεδο Εισόδου:*** το οποίο όπως γίνεται κατανοητό και από το όνομα του αποτελεί την είσοδο στο δίκτυο των χαρακτηριστικών που θέλουμε να περάσουμε μέσα από αυτό το δίκτυο νευρώνων. (2) ***Το Κρυφό Επίπεδο:*** το οποίο μπορεί να είναι και περισσότερα από ένα και (3) ***Το Επίπεδο Εξόδου:*** από το οποίο λαμβάνουμε την εκτίμηση του νευρωνικού δικτύου για την δεδομένη ομάδα εισόδου.
Ως πλήρως συνδεδεμένο νοείται πως κάθε νευρώνας του πρώτου κρυφού επιπέδου συνδέεται με όλους τους νευρώνες εισόδου και όλους τους νευρώνες εξόδου, αν έχει μόνο ένα κρυφό επίπεδο, ή με όλους τους νευρώνες του επόμενου κρυφού επιπέδου αν έχει παραπάνω από ένα κρυφό επίπεδο. Ακολουθεί μια αναπαράσταση ενός πλήρους συνδεδεμένου νευρωνικού δικτύου με δύο κρυφά επίπεδα πέντε εισόδους και μία έξοδο (Σχήμα 2.7).



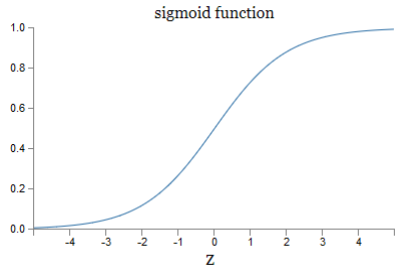
Σχήμα 2.7: Νευρωνικό Δίκτυο με Δύο Κρυφά Επίπεδα

2.4.4 Συνάρτηση Ενεργοποίησης και Συνάρτηση Κόστους

Συνάρτηση Ενεργοποίησης: Στον δυαδικό νευρώνα οι μόνες τιμές που μπορεί η έξοδος του να πάρει είναι μηδέν και ένα όπως είδαμε παραπάνω. Όμως, σε πολλά προβλήματα αυτό δεν είναι αρκετό γι αυτόν ακριβώς τον λόγο χρησιμοποιούνται και οι συναρτήσεις ενεργοποίησης. Ένα χαρακτηριστικό παράδειγμα συνάρτησης ενεργοποίησης αποτελεί η σιγμοειδής συνάρτηση ενεργοποίησης με μαθηματικό τύπο:

$$\sigma(z) = \frac{1}{1+\exp(-z)} \text{ με } z = \sum_{i=1}^{n} w_i x_i - b$$

και γραφική παράσταση:



Σχήμα 2.8: Γράφημα Σιγμοειδής Συνάρτησης Ενεργοποίησης

Όπου, x_i είναι οι είσοδοι του νευρώνα, w_i είναι τα βάρη που αντιστοιχούν σε κάθε είσοδο του νευρώνα και b είναι μια πραγματική τιμή για κάθε νευρώνα.

Συνάρτηση Κόστους: Αυτή η συνάρτηση παίζει τον ρόλο του κριτή των αποτελεσμάτων του νευρωνικού δικτύου. Δηλαδή συγκρίνοντας το αποτέλεσμα του νευρωνικού δικτύου με το πραγματικό αποτέλεσμα καθορίζει κατά πόσο και προς ποια κατεύθυνση πρέπει να αλλάξουν τα βάρη του συστήματος. Χαρακτηριστικά παραδείγματα τέτοιων συναρτήσεων αποτελούν η συνάρτηση μέσου τετραγωνικού σφάλματος με μαθηματικό τύπο:

$$C(w,b) = \frac{1}{2n}\sum_{x}^{n}\|y(x) - a\|^2$$

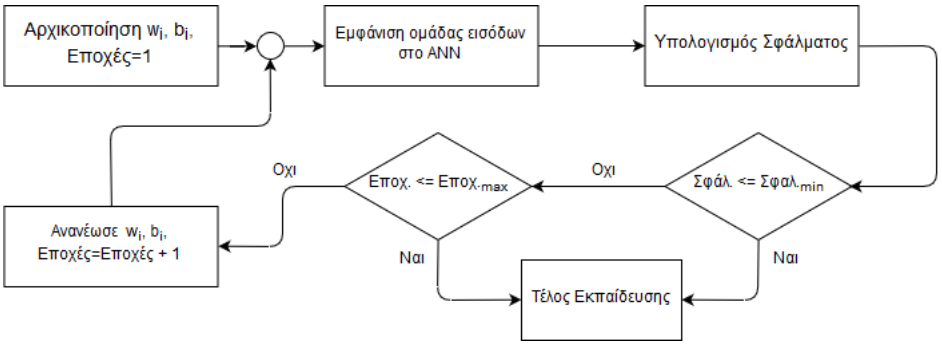
με n τον συνολικό αριθμό ομάδων εισόδου, α είναι η έξοδος του ANN για είσοδο x και y(x) είναι η επιθυμητή έξοδος του ANN για είσοδο x και η Cross-Entropy συνάρτηση κόστους με μαθηματικό τύπο:

$$C = -\frac{1}{n}\sum_{x}^{n}[yln(a) + (1-y)ln(1-a)]$$

Τα σύμβολα συνεχίζουν να έχουν την ίδια σημασία όπως και στην MSE συνάρτηση κόστους με μόνη διαφορά ότι το y(x) γράφετε εδώ ως y.

2.4.5 Στάδιο Εκπαίδευσης

Η εκπαίδευση του νευρωνικού δικτύου βασίζεται στα αποτελέσματα της συνάρτησης κόστους και στον στον αλγόριθμο της πίσω-μετάδοσης(back-propagation algorithm) για τη μεταφορά του σφάλματος για πιο πίσω επίπεδα αλλάζοντας έτσι τις τιμές των βαρών και των biases του κάθε νευρώνα. Όμως πρώτα ας δούμε τις διαδικασίες που γίνονται σε ένα βασικό νευρωνικό δίκτυο κατά τη διαδικασία της εκπαίδευσης(Σχήμα 2.9).



Σχήμα 2.9: Γράφημα του Σταδίου Εκπαίδευσης ενός ANN

Ο αλγόριθμος της πίσω-μετάδοσης αναπαριστάται στο παραπάνω γράφημα με την ανανέωση των w_i, b_i. Η διαδικασία αυτή δεν είναι κάτι παραπάνω από τον υπολογισμό των μερικών παραγώγων της συνάρτησης κόστους ως προς w_i, b_i. Όμως για να υπολογίσουμε αυτές θα χρησιμοποιήσουμε μια ενδιάμεση συνάρτηση δ_j^l που είναι η συνάρτηση σφάλματος στον j^{th} νευρώνα του l^{th} επιπέδου.Ό αλγόριθμος αυτός θα μας δώσει μια διαδικασία για τον υπολογισμό τους σφάλματος δ_j^l και τη συσχέτιση του με τις ζητούμενες μερικές παραγώγους. Αρχικά παρουσιάζονται οι τέσσερις βασικές συναρτήσεις του αλγορίθμου αυτού(Σχήμα 2.10).

$$\delta^L = \nabla_a C \odot \sigma'(z^L)$$

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l$$

$$\frac{\partial C}{\partial w_{j\,k}^l} = a_k^{l-1} \delta_j^l$$

Σχήμα 2.10: Μαθηματική Έκφραση Αλγόριθμου Πίσω Μετάδοσης

Όπου, w_{jk}^l ορίζουμε ως το βάρος της σύνδεσης μεταξύ των του k^{th} νευρώνα στο $(l-1)^{th}$ επίπεδο με τον j^{th} νευρώνα στο l^{th} επίπεδο. Η πρώτη σχέση αποτελεί την συνάρτηση υπολογισμού του σφάλματος για το στάδιο εξόδου. Η δεύτερη είναι για τον υπολογισμό του σφάλματος στο επίπεδο l^{th} με δεδομένο το σφάλμα του $(l+1)^{th}$ επιπέδου. Ακολουθεί η συνάρτηση που εκφράζει την μεταβολή της συνάρτησης κόστους ως προς b_i^l. Με την τελευταία συνάρτηση εκφράζεται η μεταβολή της συνάρτησης κόστους ως προς κάθε w_{jk}^l του δικτύου.[9]

2.5 Παλιότερες Εργασίες

Σύμφωνα με την εργασία [8], έγινε χρήση βραχυπρόθεσμων φασματικών χαρακτηριστικών και το ρόλο του classifier έχει ένας VQ. Πρόκειται για ένα σύστημα που αξιοποιεί τη μέθοδο Linear Predictive Coefficients (LPC) για την εξόρυξη βραχυπρόθεσμων φασματικών χαρακτηριστικών από το σήμα. Η μέθοδος αυτή αποτελείται από 2 βασικά μέρη το πρώτο είναι η κωδικοποίηση και η ανακατασκευή του σήματος. Κατά την κωδικοποίηση το σήμα χωρίζεται σε μικρότερα σε διάρκεια σήματα τα οποία καθορίζουν τις τιμές ενός φίλτρου ικανό να αποκωδικοποιήσει το συγκεκριμένο σήμα ομιλίας. Στην ανακατασκευή γίνεται ανακατασκευή του φίλτρου βασιζόμενη στα αποτελέσματα της κωδικοποίησης.Ακόμη, υποστηρίζει ότι μπορεί να εφαρμοστεί τόσο σε λεκτικά εξαρτώμενα όσο και σε λεκτικά ανεξάρτητα συστήματα δοκιμάζοντας τα αποτελέσματα αναγνώρισης σε διάφορες ομαδοποιήσεις χαρακτηριστικών. Πριν εμφανιστεί η MFCC η LPC ήταν η καλύτερη επιλογή για εφαρμογές φωνής.

Στην εργασία [7] αξιοποιούνται πάλι τα βραχυπρόθεσμα φασματικά χαρακτηριστικά της φωνής με τη χρήση των MFCC όμως εδώ γίνεται χρήση κι άλλων φωνητικών χαρακτηριστικών δεδομένου ότι πρόκειται για ένα σύστημα αναγνώρισης ομιλητή και ταυτόχρονα αναγνώρισης ομιλίας. Ακόμη, χρησιμοποιείται μια GMM για την διαδικασία του classification. Αναλυτικότερα, με σκοπό να βελτιωθεί η αναγνώριση ομιλίας σε θορυβώδεις συνθήκες που περισσότερες από μια φωνές ηχογραφούνται στο ίδιο σήμα το σύστημα ξεχωρίζει την ύπαρξη δεύτερης φωνής και επεξεργάζεται μόνο το σήμα που προέρχεται από τον γνωστό σε εκείνη χρήστη.

3 Τεχνολογίες

Στα πλαίσια αυτής της διπλωματικής εργασίας χρειάστηκε πέρα από το θεωρητικό υπόβαθρο, πάνω στο οποίο βασίστηκε η ανάπτυξη του εν λόγω συστήματος, η σχεδίαση και η κατασκευή αυτού σε μορφή πλακέτας. Αυτό το κομμάτι της εργασίας θα αναλύσουμε παρακάτω.

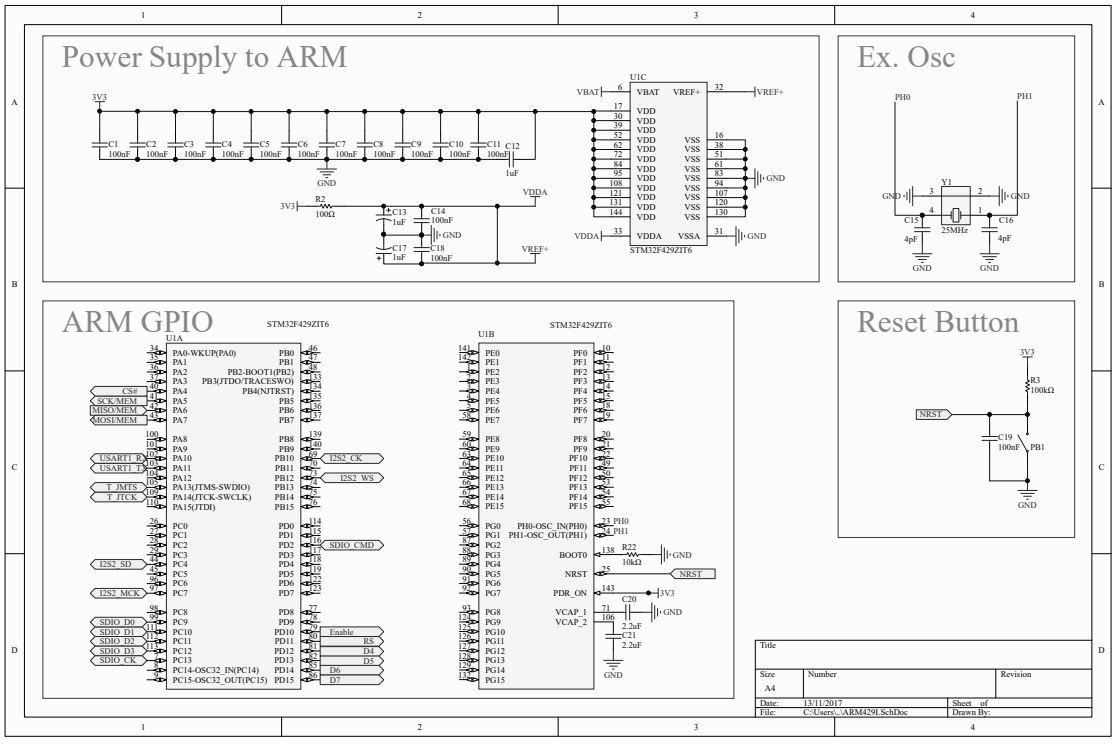
3.1 Εξαρτήματα και Μέθοδος

Εξαρτήματα:Αρχικά, θα δούμε τα εξαρτήματα που χρησιμοποιήθηκαν στο σύστημα αυτό.
(1) Για τον ρόλο του μικροεπεξεργαστή επιλέχθηκε ένας STM23F4 ARM λαμβάνοντας υπόψιν τις υψηλές του δυνατότητες σε συχνότητες που φτάνουν τα 180 Mhz. Τα χαρακτηριστικά του αυτά μας επιτρέπουν να έχουμε ένα σχεδόν real-time σύστημα με πολύ μικρή κατανάλωση σε κατάσταση αδράνειας δίνοντας έτσι τη δυνατότητα για χρήση ακόμη και με μπαταρία.
(2) Φυσικά, δεν μπορούμε να μιλάμε για ένα σύστημα αναγνώρισης ομιλητή χωρίς τη δυνατότητα ηχογράφησης. Για μικρόφωνο αρχικά επιλέχθηκε ένα ψηφιακό μικρόφωνο με 18-bit ποιότητα ήχου με σκοπό να προσφέρει πολύ καλή ποιότητα ήχου αξιοποιώντας το πρωτόκολλο I2S και ταυτόχρονα να καταλαμβάνει μικρό χώρο στην πλακέτα. Δυστυχώς, δεν είχε τα επιθυμητά αποτελέσματα και αντικαταστάθηκε με ένα απλό αναλογικό μικρόφωνο με 8-bit ποιότητα ήχου το οποίο διαβάζεται από τον εσωτερικό ADC του μικροεπεξεργαστή.(Τα σχηματικά που ακολουθούν αναφέρονται μόνο στο ψηφιακό μικρόφωνο δεδομένου ότι το αναλογικό προστέθηκε στο σύστημα πολύ αργότερα από τη σχεδίαση της πλακέτας)
(3) Για τις ανάγκες αποθήκευσης δεδομένων χρησιμοποιήθηκαν μια 64 MB NOR Flash μνήμη, που με τη χρήση του SPI πρωτοκόλλου πετυχαίνει αρκετά γρήγορη μεταφορά δεδομένων κατά την ηχογράφηση, και μια SD κάρτα, που με τη χρήση του πρωτοκόλλου SDIO σε FatFs είναι και ο τελικός προορισμός των δεδομένων της ηχογράφησης.
(4) Για την καλύτερη αλληλεπίδραση των χρηστών με το σύστημα προστέθηκε και μια 16x2 LED οθόνη, την οποία χειριζόμαστε σε 4-bit mode.
(5) Τέλος, για την τροφοδοσία του συστήματος, δεδομένου ότι λειτουργεί με επαναφορτιζόμενη μπαταρία, χρειαστήκαμε ένα γραμμικό φορτιστή και δύο μετατροπείς (Buck-Boost Converters) στα 3.3 Volt για τον επεξεργαστή, το μικρόφωνο και τις μνήμες και στα 5 Volt για την οθόνη.

Μέθοδος: Για την σχεδίαση του PCB εργάστηκα στο Altium Designer από τη δημιουργία των εξαρτημάτων σε μορφή βιβλιοθηκών τόσο σε μορφή σχηματικού όσο και σε μορφή footprtint και 3D μοντέλου, στη σύνθεση των εξαρτημάτων μεταξύ τους καταλήγοντας στα τελικά σχηματικά του PCB, μέχρι τον προσδιορισμό της τοπολογίας και των συνδέσεων μεταξύ των εξαρτημάτων που είχαν ως αποτέλεσμα το τελικό σύστημα. Στην επόμενη ενότητα παρουσιάζεται ακριβώς αυτή η διαδικασία. (παραλείπεται η παρουσίαση των βιβλιοθηκών)

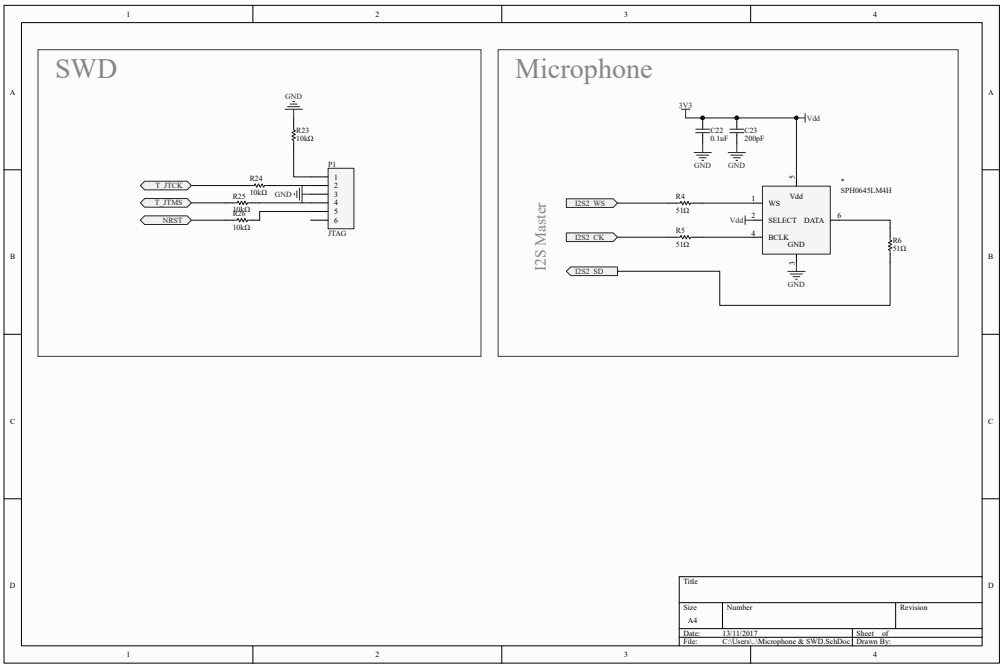
3.2 Σχηματικά, 2D και 3D Μοντέλα

3.2.1 Σχηματικά

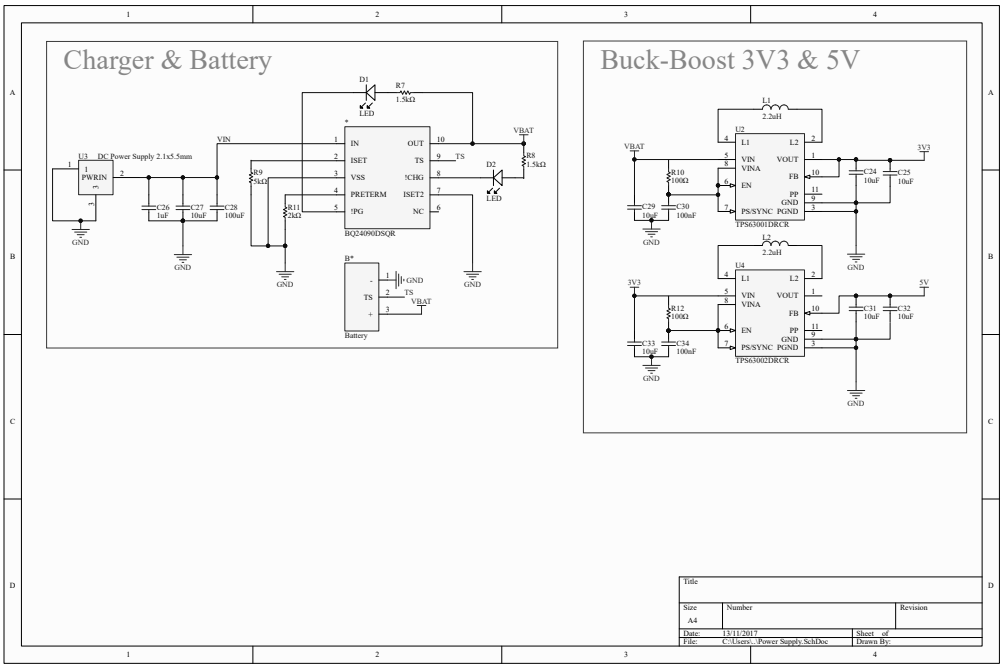


Σχήμα 3.1: Σχηματικό Μικροεπεξεργαστή

Στο παραπάνω σχηματικό αναπαριστάται ο μικροεπεξεργαστής του συστήματος ο οποίος έχει "τεμαχισθεί" σε τρία αντικείμενα το πρώτο είναι αποκλειστικά για την τροφοδοσία του ενώ τα επόμενα δύο έχουν τις υπόλοιπες του συνδέσεις και έχουν διαχωριστεί για να έχουμε ένα ευανάγνωστο σχηματικό. Ακόμη, στο παραπάνω σχηματικό συμπεριλαμβάνεται ένας εξωτερικός κρύσταλλος που αποτελεί την πηγή ρολογιού για το MPU μας και ένα κουμπί για ώστε να κάνει επαναφορά το σύστημα στην αρχική του κατάσταση.



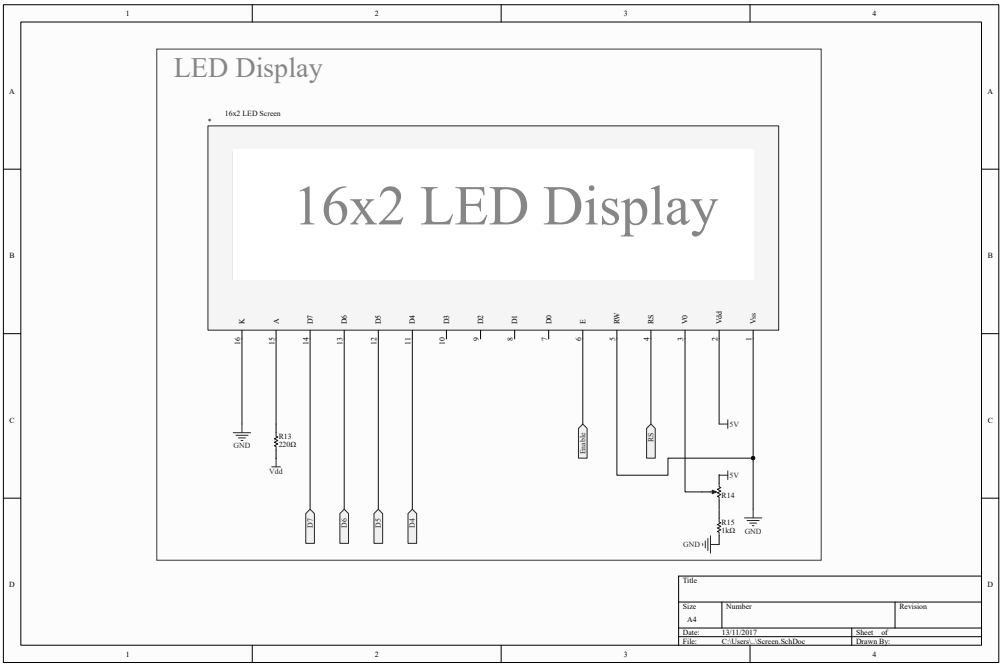
Σχήμα 3.2: Σχηματικό Μικροφώνου και SWD



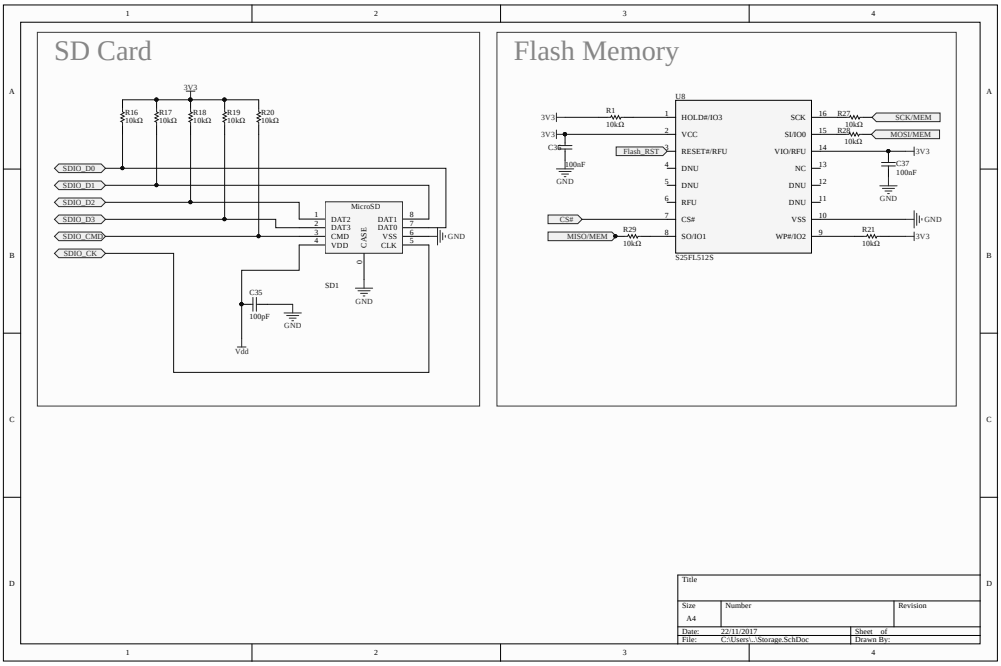
Σχήμα 3.3: Σχηματικό Τροφοδοσίας του Συστήματος

Στα παραπάνω δύο σχηματικά (Σχήμα 3.2, Σχήμα 3.3) συναντάμε τη συνδεσμολογία για το ψηφιακό μικρόφωνο, το pinout για το SWD πρωτόκολλο και τα σχηματικά για την τροφοδοσία του

συστήματος . Αξίζει να σημειωθεί πως οι τιμές των αντιστάσεων και πυκνωτών ίσως να διαφέρουν από τις τελικές, όπως της αντίστασης του ISET για το γραμμικό φορτιστή η που όπως λέει και το όνομα της ρυθμίζει το ρεύμα που θα φορτίζει την μπαταρία.



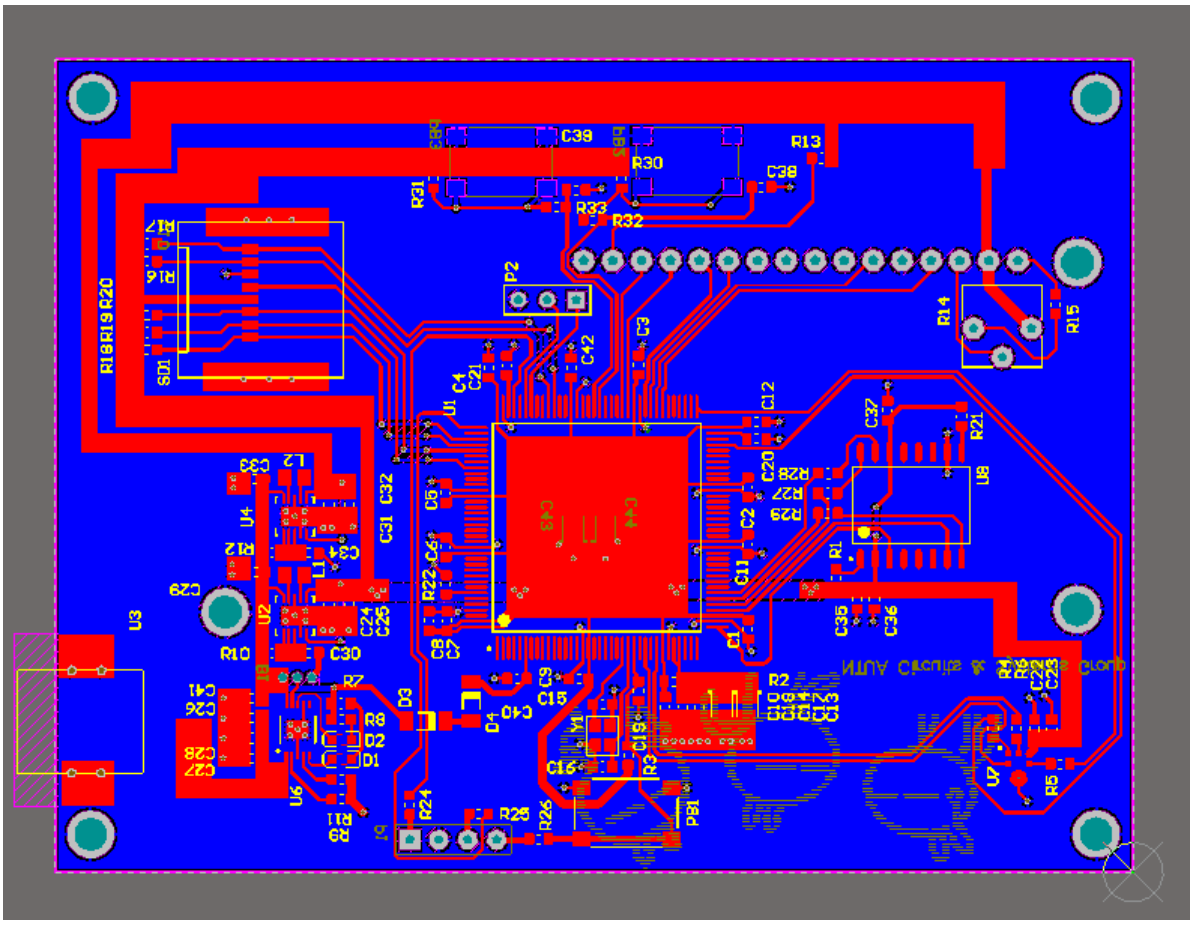
Σχήμα 3.4: Σχηματικό Οθόνης



Σχήμα 3.5: Σχηματικό Αποθηκευτικών Μέσων

Τέλος, έχουμε τα σχηματικά της οθόνης και των αποθηκευτικών μέσων του συστήματος. Έχοντας πλέον μια εικόνα για το σύστημα αυτό μπορούμε να προχωρήσουμε στην παρουσίαση των 2D & 3D μοντέλων του.

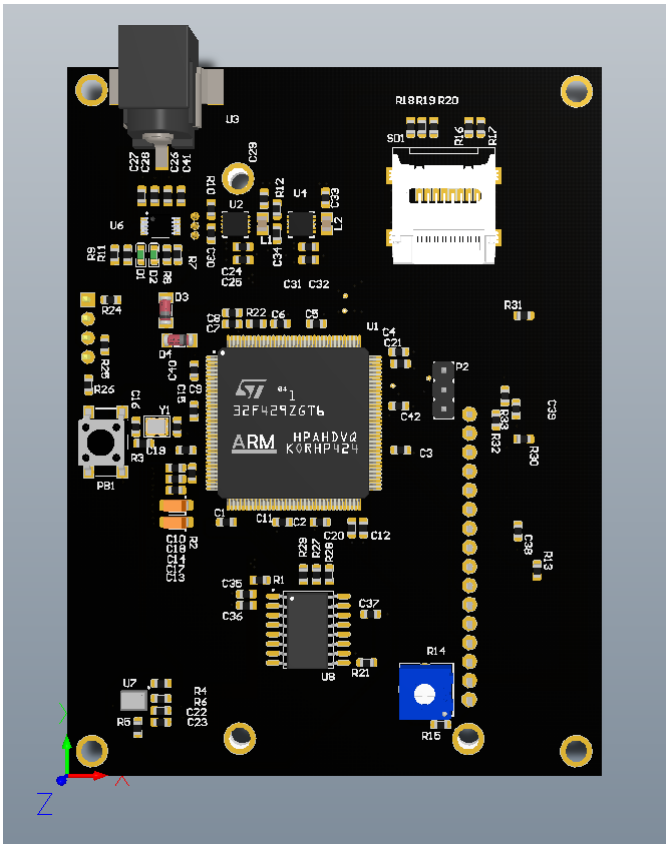
3.2.2 2D Μοντέλο



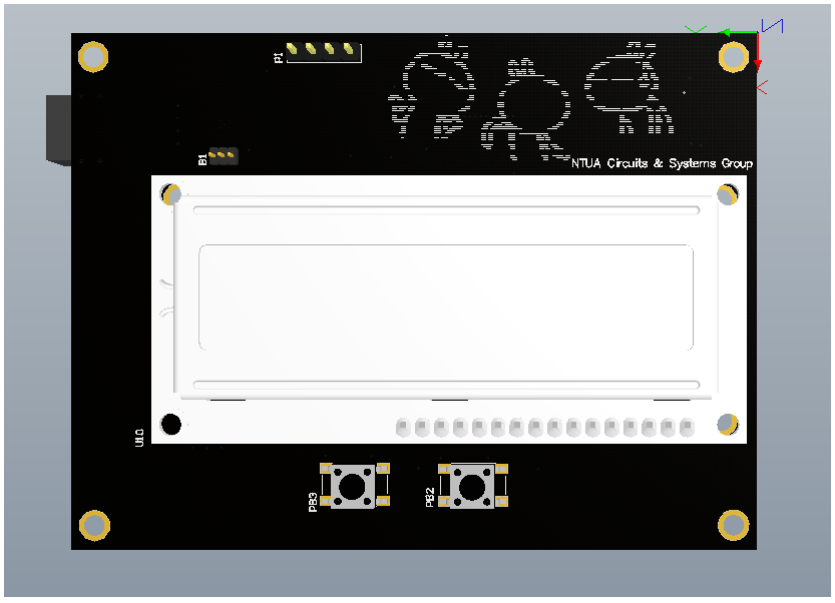
Σχήμα 3.6: 2D Μοντέλο του Συστήματος

Στο παραπάνω σχήμα βλέπουμε το 2D μοντέλο του συστήματος, στο οποίο έχουν χρησιμοποιηθεί μόνο δύο επίπεδα για τη μεταφορά των σημάτων και για την τροφοδοσία. Ελαχιστοποιώντας έτσι το κόστος κατασκευής του PCB. Ακόμη, όσο ήταν δυνατόν περισσότερο χρησιμοποιήθηκε το ένα επίπεδο(κόκκινο) για τη μεταφορά σημάτων και τη διανομή των 3.3 και 5 Volt αντίστοιχα σε όλα τα εξαρτήματα κρατώντας έτσι το δεύτερο επίπεδο (μπλε) για να ένα μεγάλο ενιαίο GND. Πέραν των εξαρτημάτων που αναφέρθηκαν παραπάνω υπάρχει και ένα pinout δεξιά της MCU που στο στάδιο προγραμματισμού της πλακέτας είχε το ρόλο της UART ενώ τελικά μπορεί να χρησιμοποιηθεί ως έλεγχος για την εκάστοτε εφαρμογή που μπορεί να χρησιμοποιηθεί.

3.2.3 3D Μοντέλο



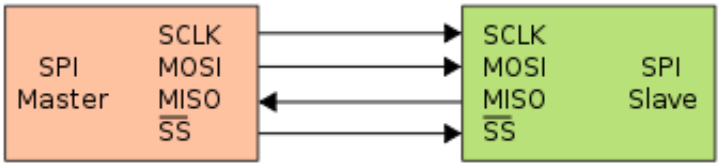
Σχήμα 3.7: 3D Μοντέλο του Συστήματος(α)



Σχήμα 3.8: 3D Μοντέλο του Συστήματος(β)

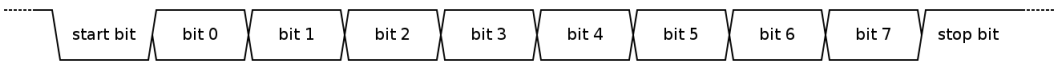
3.3 Πρωτόκολλα Επικοινωνίας

SPI: Το ακρωνύμιο σημαίνει Serial Peripheral Interface και πρόκειται για ένα σύγχρονο σειριακό πρωτόκολλο επικοινωνίας σχεδιασμένο για μικρές σε απόσταση μεταφορές δεδομένων με ιδιαίτερα υψηλές ταχύτητες. Το SPI έχει βρει εφαρμογή σε embedded συστήματα. Αναλυτικότερα, το λειτουργικό διάγραμμα του συστήματος (Σχήμα 3.9) απεικονίζει τη βασική λειτουργία του πρωτοκόλλου αυτού. Για τη χρήση αυτού χρειαζόμαστε δύο οντότητες έναν οδηγό (Master) και έναν ακόλουθο (Slave), οι οποίοι σχηματίζουν την μεταξύ τους ζεύξη με τουλάχιστον τέσσερεις συνδέσεις. Οι συνδέσεις αυτές παίζουν τον ρόλο του συγχρονισμού μεταξύ των δύο με την σύνδεση SCLK, την μεταφορά δεδομένων από τον οδηγό στον ακόλουθο MOSI(Master Out Slave In), την μεταφορά δεδομένων από τον ακόλουθο στον οδηγό MISO(Master In Slave Out) και τέλος την σύνδεση SS(Slave Select) όπου γίνεται φανερή η δυνατότητα του οδηγού να έχει περισσότερους από έναν ακόλουθο και αυτή η σύνδεση ορίζει με ποιόν από τους ακόλουθους επικοινωνεί ο οδηγός.



Σχήμα 3.9: Βασική Χρήση SPI

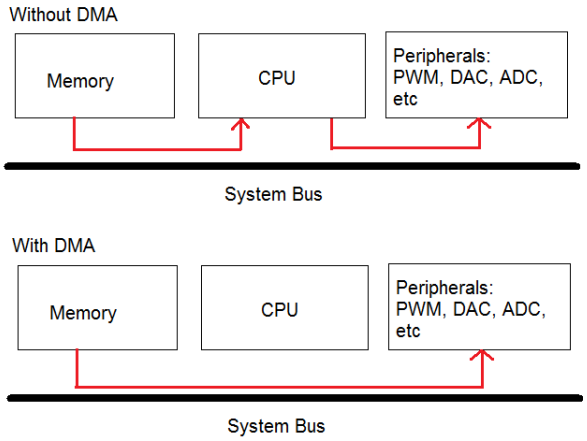
UART: Σημαίνει Universal Asynchronous Receiver-Transmitter, χρησιμοποιείται όπως δηλώνει και το όνομα του για την αμφίδρομη μεταφορά δεδομένων από το ένα άκρο στο άλλο της σύνδεσης. Σε αυτή την εργασία κυρίως χρησιμοποιήθηκε για την αποσφαλμάτωση του κώδικα. Για τη μεταφορά δεδομένων είτε από είτε προς τον MPU παίρνει τα bytes δεδομένων και τα μεταφέρει με τη σειρά ανά ένα bit. Με σκοπό να γνωρίζει ο παραλήπτης πριν τη μεταφορά δεδομένων στέλνει ένα start bit και στο τέλος για να αναγνωρίσει το τέλος ένα stop bit. Ακόμη μια σημαντική παράμετρος της UART είναι ο αποστολέας και ο παραλήπτης να έχουν την ίδια ταχύτητα μεταφοράς των bit ώστε να γίνεται έχουμε σωστή αποκωδικοποίηση των δεδομένων.



Σχήμα 3.10: Διάγραμμα Χρονισμού UART

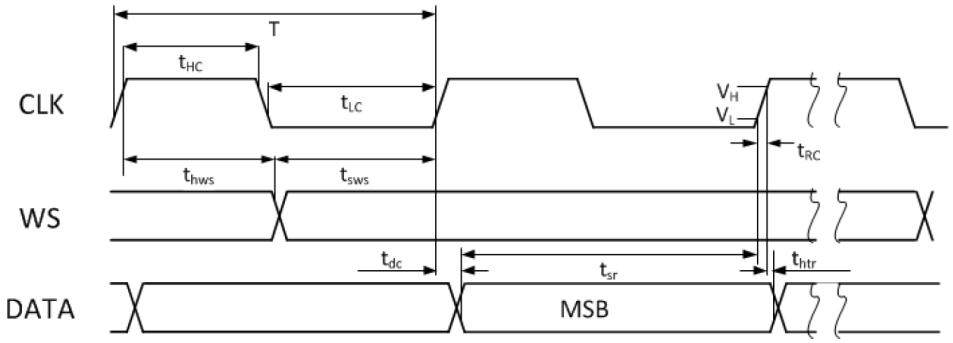
DMA: Ως Direct Memory Access ορίζεται η δυνατότητα της μεταφοράς δεδομένων από τα περιφερειακά του συστήματος σε θέσεις μνήμης αλλά και από θέσεις μνήμης σε θέσεις μνήμης χωρίς την χρήση του MPU. Επιτρέποντας έτσι στο σύστημα να εκτελεί παράλληλα πολλές διεργασίες

επιταχύνοντας έτσι την εκτέλεση μη εξαρτώμενων λειτουργιών του συστήματος. Ακολουθεί μια γραφική με το άμεσο κέρδος χρήσης του DMA.



Σχήμα 3.11: Σύγκριση Διεργασίας Με Χρήση DMA και Χωρίς DMA

I^2***S:*** Inter-IC Sound είναι ένα σειριακό πρωτόκολλο επικοινωνίας μεταξύ συσκευές ψηφιακού ήχου. Συγκεκριμένα χρησιμοποιείται για τη μεταφορά PCM δεδομένα ήχου από κάποιο μικρόφωνο σε μια MPU και το αντίστροφο αν στη θέση του μικροφώνου υπάρχει ένα ηχείο. Το πρωτόκολλο αυτό απαιτεί τρεις ζεύξεις μεταξύ του πομπού και του δέκτη. Η πρώτη γραμμή είναι αυτή που μεταφέρει το ρολόι σύμφωνα με το οποίο στέλνονται τα δεδομένα και γι αυτό ονομάζεται BCLK (Bit Clock). Έπειτα έχουμε την γραμμή επιλογής καναλιού η οποία καθορίζει ποιου από τα 2 κανάλια (σε περίπτωση στερεοφωνικού ήχου) η πληροφορία μεταφέρεται αυτή τη χρονική στιγμή και ονομάζεται WS (Word Select) ή LRCLK (Left-Right Clock). Τέλος, έχουμε τη ζεύξη που μεταφέρει την πληροφορία σε σειριακή μορφή και ονομάζεται SD (Serial Data). Ακολουθεί η γραφική που παρουσιάζει το ψηφιακό μικρόφωνο που επιλέχθηκε αρχικά για το συγκεκριμένο σύστημα.



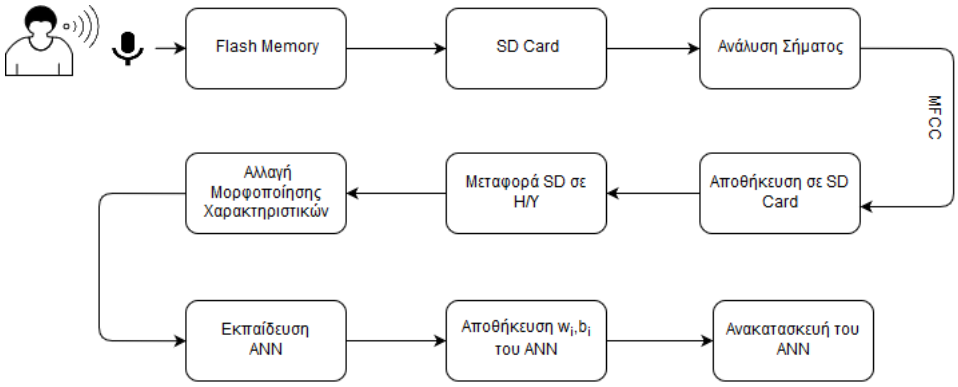
Σχήμα 3.12: Διάγραμμα Χρονισμού I2S Ψηφιακού Μικροφώνου του Συστήματος

4 Προσέγγιση Συστήματος

4.1 Περιγραφή Συστήματος

Θα περιγράψουμε το σύστημα με 2 αλγορίθμους όσες και οι διαδικασίες που εκτελέστηκαν με αυτό.

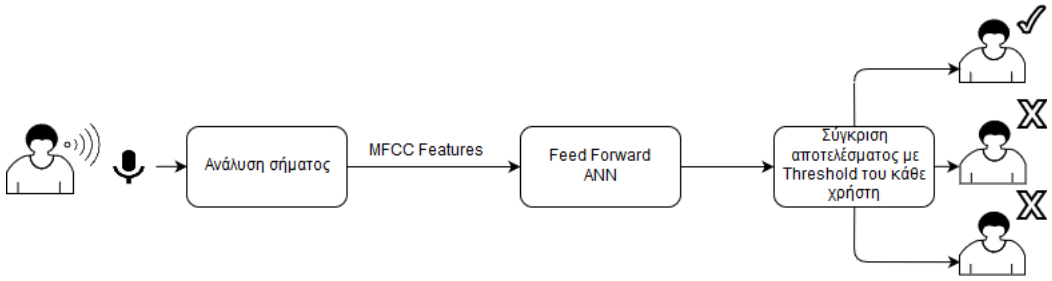
Στο πρώτο στάδιο με στόχο να εξορύξουμε αρκετά φωνητικά χαρακτηριστικά για τους χρήστες ηχογραφούσαμε, για ένα διάστημα είκοσι δευτερολέπτων, αποθηκεύοντας αρχικά τα δεδομένα στην Flash μνήμη μέσω DMA ώστε να μην υπάρχουν κενά στην ηχογράφηση. Μετά την ολοκλήρωση της ηχογράφησης μεταφέρουμε τα δεδομένα από την Flash στην SD Card, με σκοπό να κρατήσουμε την ηχογράφηση, έπειτα ακολουθούμε τον αλγόριθμο εξόρυξης των χαρακτηριστικών MFCC και τελικά τα γράφουμε κι αυτά στην SD Card. Στη συνέχεια, μεταφέρουμε τα δεδομένα της SD σε έναν υπολογιστή με εγκατεστημένη τη MATLAB και μετατρέπουμε τα δεδομένα σε κατάλληλη μορφή για την είσοδο αυτών στο Νευρωνικό Δίκτυο. Ακολουθεί, η εκπαίδευση του νευρωνικού αυτού, η οποία θα αναλυθεί σε επόμενη ενότητα, και η αποθήκευση των τελικών βαρών του για την ανακατασκευή αυτού στο σύστημα μας. Το παρακάτω σχήμα αναπαριστά ακριβώς αυτή τη διαδικασία.



Σχήμα 4.1: Λειτουργικό Διάγραμμα Εκπαίδευσης Συστήματος

Στο δεύτερο στάδιο και ουσιαστικά το τελικό σύστημα η λειτουργία εκκινεί με το πάτημα του αριστερού κουμπιού του συστήματος. Σε αυτή τη λειτουργία δεδομένου ότι δεν χρειαζόμαστε

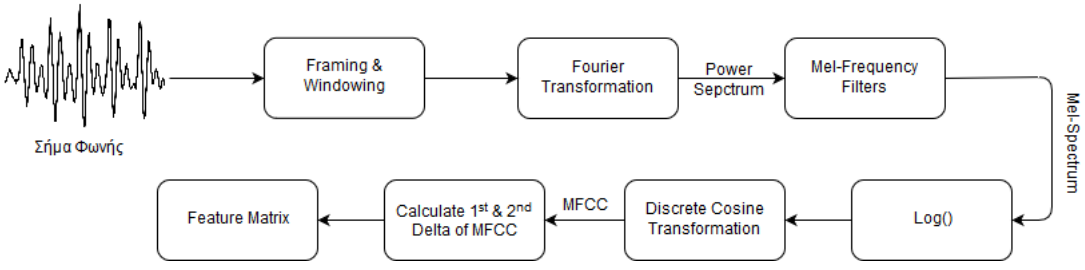
μεγάλη σε διάρκεια ηχογράφηση αρκεί η εσωτερική RAM του συστήματος. Επομένως, ηχογραφούμε για μερικά δευτερόλεπτα, αποθηκεύουμε την ηχογράφηση στην SD Card για μελλοντική χρήση, εφαρμόζουμε τον αλγόριθμο εξόρυξης των χαρακτηριστικών MFCC, περνάμε τις ομάδες εισόδων μέσα από το ANN, το οποίο για κάθε τέτοια ομάδα μας δίνει μια πρόβλεψη για το ποίος χρήστης είναι ο ομιλητής, και τελικά συγκρίνουμε το αποτέλεσμα με το όριο αναγνώρισης του κάθε χρήστη και αν τα αποτελέσματα του νευρωνικού ξεπερνούν κάποιο από αυτά τότε το σύστημα υποθέτει πως είμαστε αυτός ο χρήστης.



Σχήμα 4.2: Λειτουργικό Διάγραμμα Συστήματος

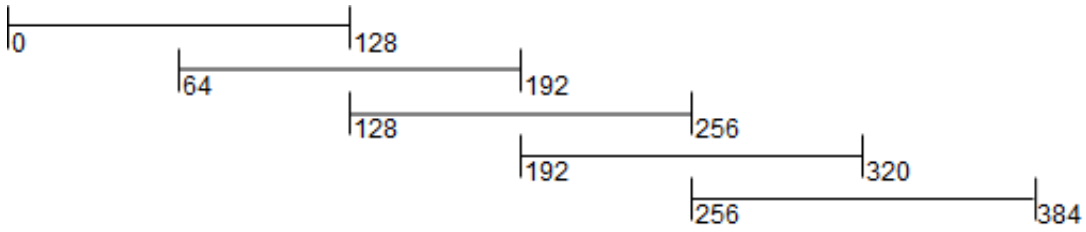
4.2 Αλγόριθμος Ανάλυσης Σήματος

Στα πλαίσια της ενότητας 2.2 έγινε μια ανάλυση του αλγορίθμου εξόρυξης χαρακτηριστικών MFCC και με δεδομένη αυτή την ανάλυση εδώ θα τονίσουμε τις βελτιώσεις που έγιναν με σκοπό την επιτάχυνση υπολογισμού των χαρακτηριστικών αυτών. Σαν αναφορά θα έχουμε το παρακάτω διάγραμμα ροής.



Σχήμα 4.3: Διάγραμμα Ροής για την Εξόρυξη Χαρακτηριστικών MFCC

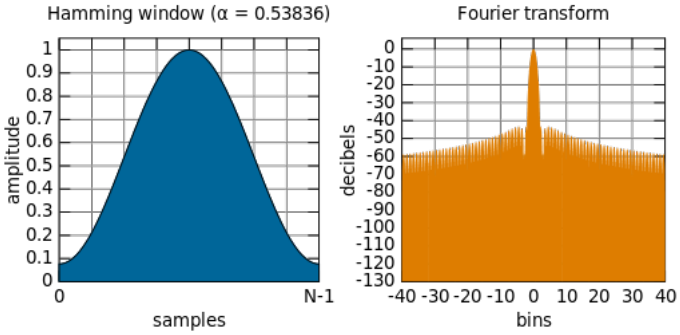
Framming:Για να κάνουμε στατικό το σήμα το "τεμαχίσαμε" σε επικαλυπτόμενα τμήματα διπλάσια των FFT bins.



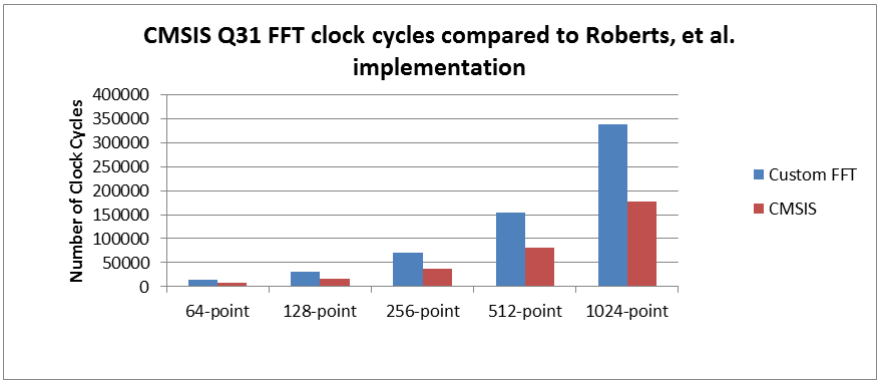
Σχήμα 4.4: Πλαισίωση του Σήματος σε Επικαλυπτόμενα Σήματα Σταθερού Μήκους

Windowing: Ακόμη, εφαρμόσαμε ένα παράθυρο Hamming σε αυτά με σκοπό να μειώσουμε, όσο είναι αυτό δυνατό, την φασματική διαρροή λόγω των ασυνεχειών . Για να αποφύγουμε τον υπολογισμό του παράθυρου Hamming σε κάθε επανάληψη οι τιμές αυτού έχουν αποθηκευθεί σε έναν πίνακα.

$$w(n) = \alpha - \beta cos(\frac{2\pi n}{N-1})$$



FFT: Για τον υπολογισμό του Fourier Transformation του εισερχόμενου σήματος με την ελάχιστη υπολογιστική ισχύ χρησιμοποιήθηκε η CMSIS-DSP βιβλιοθήκη για ARM επεξεργαστές.



Σχήμα 4.5: Σύγκριση CMSIS-FFT με άλλο αλγόριθμο FFT[10]

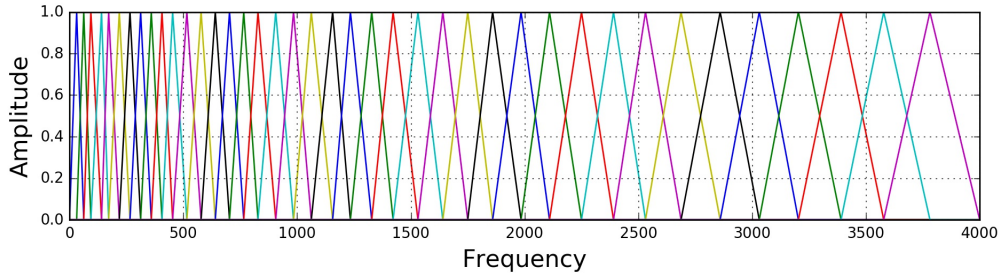
Mel Filters: Τα φίλτρα αυτά κατασκευάστηκαν σύμφωνα με τον παρακάτω μαθηματικό τύπο και αποθηκεύτηκαν σε πίνακα για να αποφύγουμε την επαναλαμβανόμενη δημιουργία τους. Ακολουθεί αναπαράσταση ενός τέτοιου συνόλου φίλτρων.

$$m = 2595 \log_{10}(1 + \frac{f}{700})$$
$$f = 700(10^{m/2595} - 1)$$

Τύποι μετατροπής Hertz σε Mel και αντίστροφα.

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k < f(m) \\ 1 & k = f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \leq f(m+1) \\ 0 & k > f(m-1) \end{cases}$$

Επομένως, είμαστε σε θέση να παράγουμε οποιουδήποτε μεγέθους φίλτρο όπως το παρακάτω.



Σχήμα 4.6: Mel-Filterbank με Fs = 8000 και n = 40

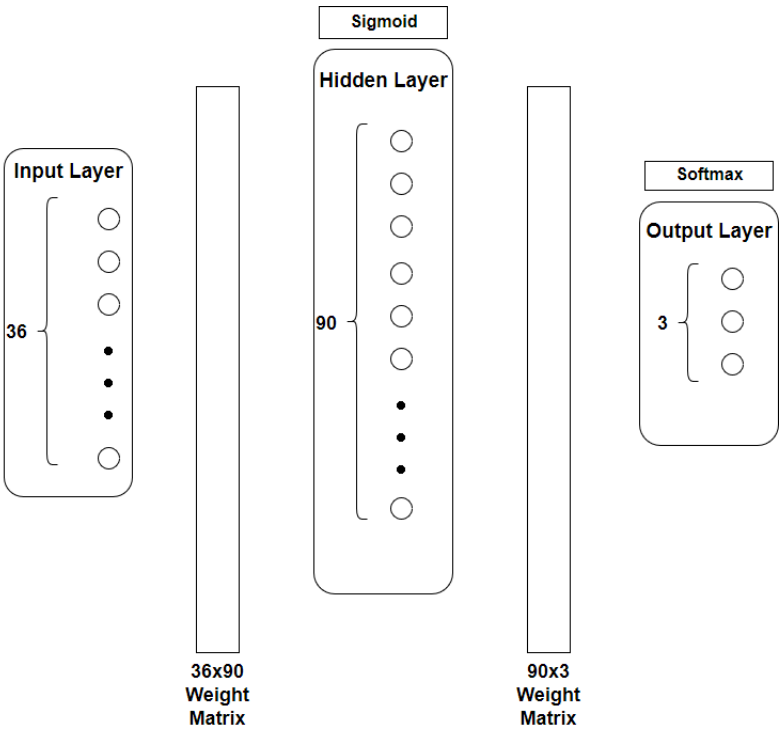
Όπου Fs είναι η συχνότητα δειγματοληψίας και n ο αριθμός των φίλτρων.
1^{st} ***και*** 2^{nd} ***Delta:*** Για τον υπολογισμό των Delta & Delta-Delta MFCC's χρησιμοποιήσαμε τη μέθοδο πεπερασμένων διαφορών.

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

4.3 Εκπαίδευση Νευρωνικού Δικτύου

Για την εκπαίδευση ενός νευρωνικού δικτύου υπάρχουν αρκετές μεθοδολογίες οι οποίες χρησιμοποιούν τον Backpropagation algorithm. Μια από τις πιο βασικές διαδικασίες που πρέπει να

κάνουμε ώστε το νευρωνικό να εκπαιδευτεί το ίδιο καλά για κάθε πιθανή έξοδο είναι το ανακάτεμα της σειράς των εισόδων πριν ξεκινήσει η νέα εποχή με σκοπό να αποφύγουμε το overfitting. Όταν λέμε ότι παρατηρούμε overfitting σημαίνει πως το νευρωνικό δίκτυο έχει προσαρμόσει τόσο πολύ τις τιμές του ώστε να αναγνωρίζει μόνο τα γνωστά σε αυτό πακέτα εισόδων. Αποτέλεσμα του overfitting είναι να δίνει λανθασμένο αποτέλεσμα τη στιγμή που οι είσοδοι απέχουν από τις τιμές του δείγματος εκπαίδευσης κατά μια πολύ μικρή απόσταση. Ακόμη, για την βέλτιστη εκπαίδευση του νευρωνικού πρέπει να ορίσουμε κάποιες συνθήκες εξόδου(όπως είδαμε και στο παράδειγμα λειτουργίας ενός νευρωνικού δικτύου στο Σχήμα 2.9). Κάποιες τέτοιες παράμετροι είναι ο αριθμός των εποχών, το επιθυμητό ποσοστό επιτυχίας και κάποιος αριθμός συνεχόμενων επιτυχιών.
Στα πλαίσια τις διπλωματικής χρησιμοποιήθηκε το toolbox της MATLAB για νευρωνικά δίκτυα. Συγκεκριμένα, σαν βάση χρησιμοποιήθηκε το template της matlab για προβλήματα pattern recognition με πολλές πιθανές εξόδους. Τελικά, καταλήξαμε σε ένα νευρωνικό δίκτυο με Cross-Entropy συνάρτηση σφάλματος, για τη μεταφορά του σφάλματος χρησιμοποιήθηκε η μέθοδος Scaled Conjurate Gradient και μοιράζοντας το δείγμα χαρακτηριστικών σε 70% δεδομένα εκπαίδευσης, 15% δεδομένα επιβεβαίωσης και 15% δεδομένα τελικής δοκιμής. Ως συνθήκες ελέγχου ορίσθηκαν οι 100 συνεχόμενες επιτυχείς δοκιμές επιβεβαίωσης. Ακόμη, ως συνάρτηση ενεργοποίησης του κρυφού επιπέδου χρησιμοποιήσαμε μια sigmoid ενώ για το επίπεδο εξόδου softmax.



Σχήμα 4.7: Γραφική Αναπαράσταση του ANN

4.4 Βασικοί Κανόνες Λειτουργίας

Για να πετύχουμε την μεγαλύτερη δυνατή αναγνώριση για το σύστημα θα πρέπει να ακολουθηθούν οι βασικοί κανόνες λειτουργίας. Αρχικά, για να επιτευχθεί αναγνώριση δεν πρέπει ο θόρυβος να ξεπερνάει τα λογικά επίπεδα ενός δωματίου. Ακόμη, πρέπει να μην μιλούν παραπάνω από ένα άτομα κατά την ηχογράφηση δεδομένου ότι δεν έχει γίνει μελέτη ώστε το σύστημα να εστιάζει μόνο στον ομιλητή με τη μεγαλύτερη ένταση. Έπειτα, πρέπει η απόσταση από το μικρόφωνο να είναι παρόμοια με την απόσταση που είχε ο ομιλητής κατά την ηχογράφηση εκπαίδευσης. Τέλος, ο ομιλητής καλό θα ήταν να έχει τον ίδιο τόνο με τον οποίο μιλούσε κατά την ηχογράφηση εκπαίδευσης και να μην αλλάζει τη φωνή του κατά τη διάρκεια της ηχογράφησης.

5 Πειράματα και Αποτελέσματα

5.1 Πειράματα

Έχοντας πλέον κατασκευάσει ένα σύστημα που μπορεί να ηχογραφήσει την ομιλία του χρήστη και να εξορύξει χαρακτηριστικά για τις βραχυπρόθεσμες φασματικές ιδιότητες της φωνής του πρέπει να διαλέξουμε τις βέλτιστες τιμές για το νευρωνικό δίκτυο:
(1) ***Αριθμός Εισόδων του ANN***
(2) ***Ποσοστό επιτυχημένων Δειγμάτων***

5.1.1 Αριθμός Εισόδων του ANN

Σύμφωνα με την παράμετρο αυτή καλούμαστε να επιλέξουμε αν θα αξιοποιήσουμε:
(α) Μόνο τις 12 βασικές MFCC
(β) Τις δύο βασικές αλλά και τις διαφορές αυτών με τις δύο χρονικά προηγούμενες ομάδες χαρακτηριστικών
(γ) Τον συνδυασμό περισσότερων από μια ομάδα MFCC
Δοκιμάζοντας διαφορετικές ομαδοποιήσεις εισόδων για το ίδιο σήμα εισόδου μπορούμε να παρατηρήσουμε, αρχικά τι ποσοστά επιτυχίας έχουμε μετά τη λήξη του σταδίου εκπαίδευσης, έπειτα τα ποσοστά επιτυχίας του κάθε χρήστη και τέλος δοκιμάζοντας το νευρωνικό για ένα δεύτερο σύνολο εισόδων να ελέγξουμε την ικανότητα του συστήματος να αναγνωρίζει σωστά τους χρήστες έχοντας ως εισόδους δεδομένα πάνω στα οποία δεν έχει εκπαιδευτεί.

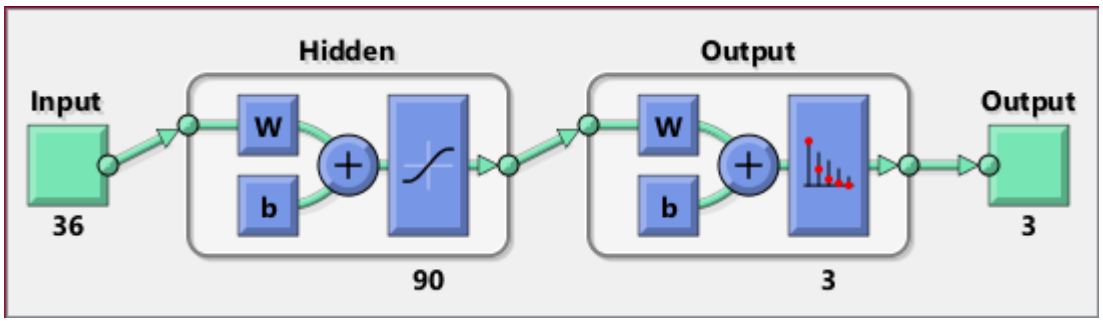
5.1.2 Ποσοστό επιτυχημένων Δειγμάτω

Το ποσοστό επιτυχημένων δειγμάτων για να ληφθεί απόφαση μπορεί να είναι σταθερό για όλους τους χρήστες είτε να αλλάζει ανάλογα με τον χρήστη και πόσο κοντά είναι η φωνή του στην αρχική του ηχογράφηση. Αλλάζοντας το ποσοστό αυτό ανάλογα με τον χρήστη θα πετύχουμε ένα προσαρμοσμένο σύστημα στους χρήστες με αποτέλεσμα να έχουμε τα καλύτερα δυνατά αποτελέσματα με την ελάχιστη προσπάθεια από τον χρήστη. Επομένως, στα πλαίσια αυτού του πειράματος θα πραγματοποιήσουμε ένα ορισμένο αριθμό δοκιμών με σκοπό να καθορίσουμε το ποσοστό αναγνώρισης

του κάθε χρήστη που μας προσφέρει ένα ικανοποιητικό επίπεδο σιγουριάς για να επιλέξουμε τον χρήστη.

5.2 Αποτελέσματα

Πραγματοποιώντας τα παραπάνω πειράματα στο περιβάλλον της MATLAB και με τις παραμέτρους που περιγράφηκαν στο κεφάλαιο 4 είμαστε σε θέση να δοκιμάσουμε τα αποτελέσματα της ανάλυσης σήματος στο ANN και να δούμε πόσο καλά μπορούμε να ταξινομήσουμε τον κάθε χρήστη σύμφωνα με αυτά. Πριν από την παρουσίαση των αποτελεσμάτων ας δούμε τη μορφή του ANN που χρησιμοποιήθηκε στην πειραματική διαδικασία με μόνη διαφορά την αλλαγή του αριθμού εισόδων.

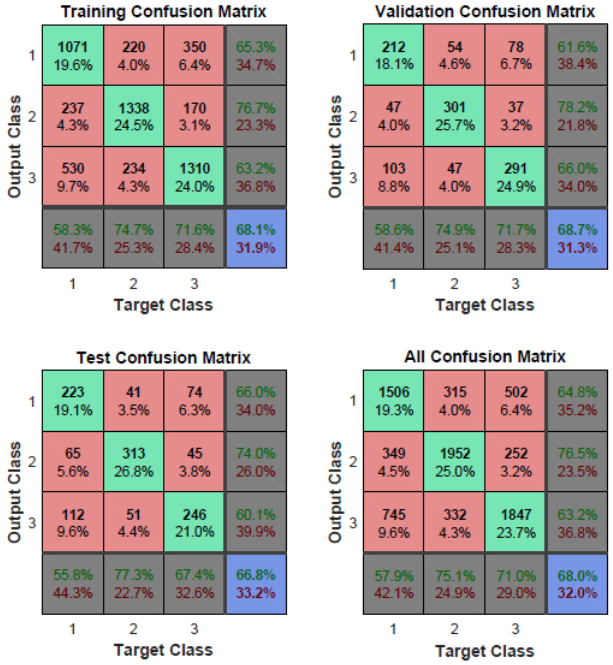


Σχήμα 5.1: Μορφή ANN Με Χρήση MATLAB Για 36 Εισόδους

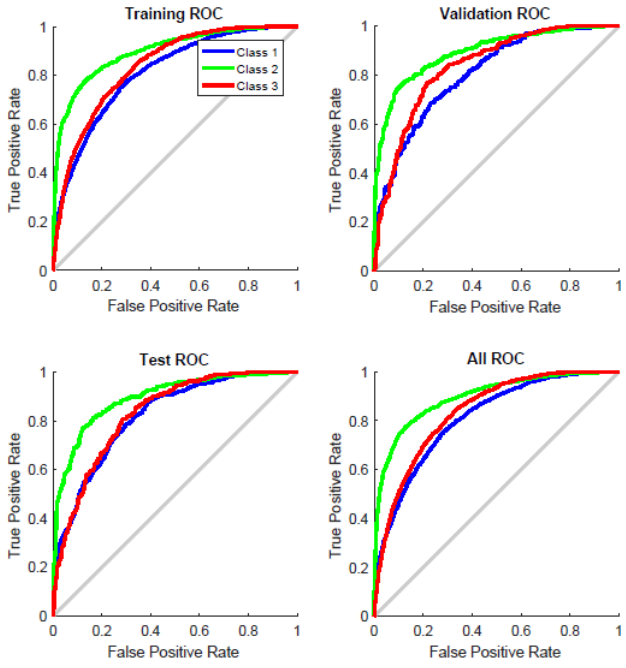
5.2.1 Αριθμός Εισόδων του ANN

Στην παράγραφο αυτή, παρουσιάζονται τα αποτελέσματα του ANN για 3 ομαδοποιήσεις των χαρακτηριστικών που έχουμε εξορύξει. Παρότι, η επιλογή των 72 εισόδων προσθέτει στο σύστημα μια αλληλεξάρτηση μεταξύ τον δειγμάτων, αλλάζοντας έτσι την φύση του συστήματος σε Λεκτικά Εξαρτώμενο, έχει ενδιαφέρον να δούμε τα αποτελέσματα του. Αξίζει να σημειωθεί ότι για κάθε μια από τις παρακάτω περιπτώσεις έγινε εκπαίδευση του ANN 5 φορές και κρατήθηκε η εκπαίδευση με το καλύτερο αποτέλεσμα για κάθε ομαδοποίηση των εισόδων.

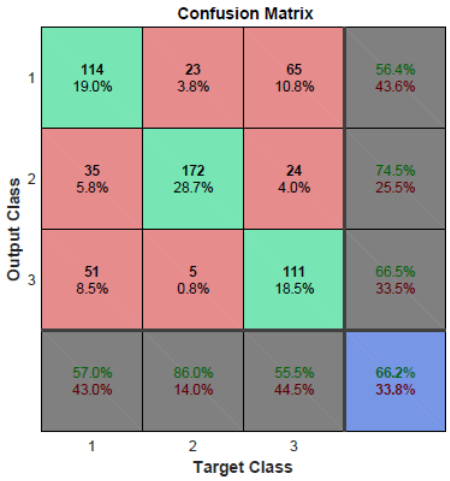
Χρήση μόνο των 12 βασικών MFCC



Σχήμα 5.2: Αποτελέσματα εκπαίδευσης ANN 12 Εισόδων



Σχήμα 5.3: Receiver Operating Characteristic Σε ANN 12 Εισόδων



Σχήμα 5.4: Δοκιμαστικό Δείγμα Για ANN 12 Εισόδων

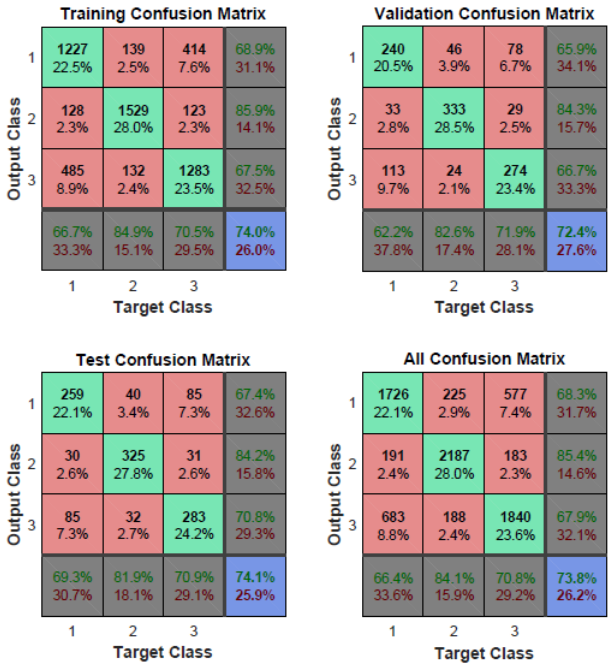
Ο παρακάτω πίνακας συνοψίζει τα αθροιστικά αποτελέσματα για την περίπτωση των 12 εισόδων.

Αποτελέσματα ANN με 12 Εισόδους				
	Training Dataset			Testing Dataset
	Training	Validation	Testing	Testing
Succes Rate (%)	68,1	68,7	66,8	66,2

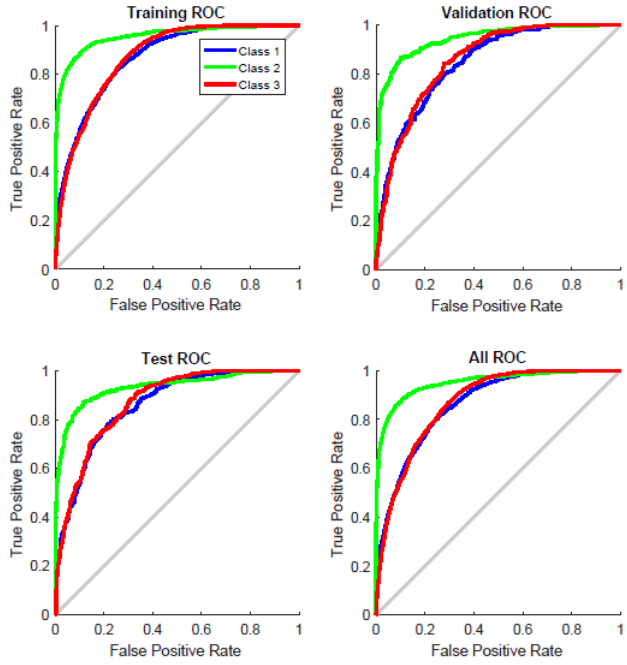
Πίνακας 5.1: Αποτελέσματα ANN με 12 Εισόδους

Η πρώτη μας προσέγγιση παρουσιάζει επίπεδα αναγνώρισης πάνω από 65% αξιοποιώντας τις βασικές MFCC και με την ελάχιστη επεξεργαστική ισχύ.

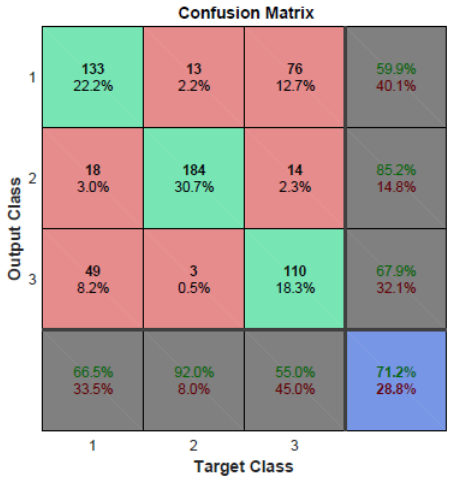
Χρήση των βασικών MFCC και των διαφορών τους με τις δύο προηγούμενες ομάδες MFCC



Σχήμα 5.5: Αποτελέσματα εκπαίδευσης ANN 36 Εισόδων



Σχήμα 5.6: Receiver Operating Characteristic Σε ANN 36 Εισόδων



Σχήμα 5.7: Δοκιμαστικό Δείγμα Για ANN 36 Εισόδων

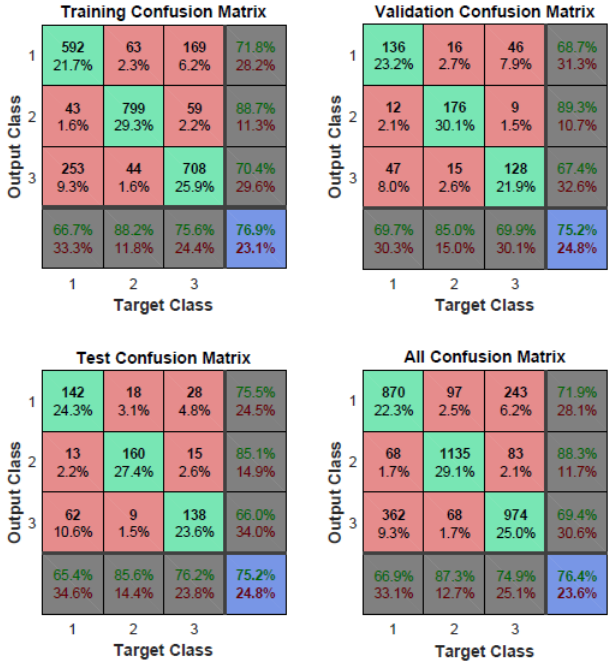
Ο παρακάτω πίνακας συνοψίζει τα αθροιστικά αποτελέσματα για την περίπτωση των 36 εισόδων.

Αποτελέσματα ANN με 36 Εισόδους				
	Training Dataset			Testing Dataset
	Training	Validation	Testing	Testing
Success Rate (%)	74,0	72,4	74,1	71,2

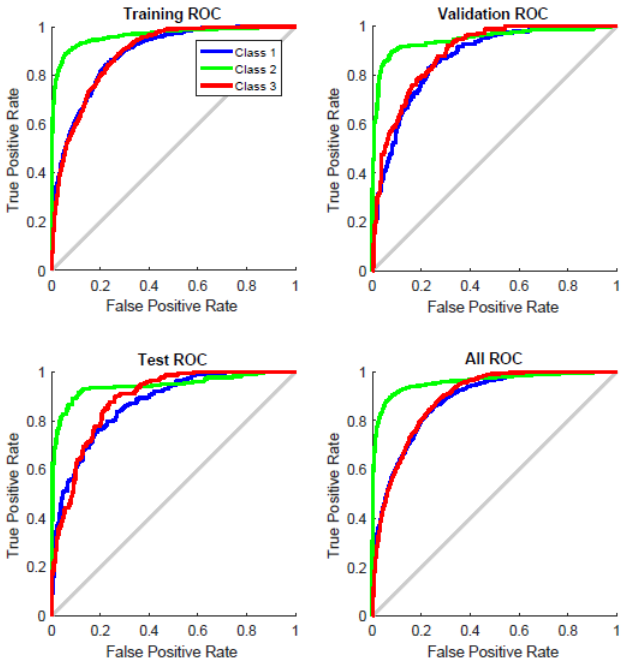
Πίνακας 5.2: Αποτελέσματα ANN με 36 Εισόδους

Στην παραπάνω προσέγγιση προσθέσαμε και τις σχέσεις μεταξύ των MFCC υπολογίζοντας τις διαφορές μεταξύ των τιμών. Οι διαφορές αυτές, έχοντας πολύ μικρό επεξεργαστικό κόστος στο σύστημα, προσέφεραν μεγάλη αύξηση στο ποσοστό επιτυχίας της τάξεως του 5%.

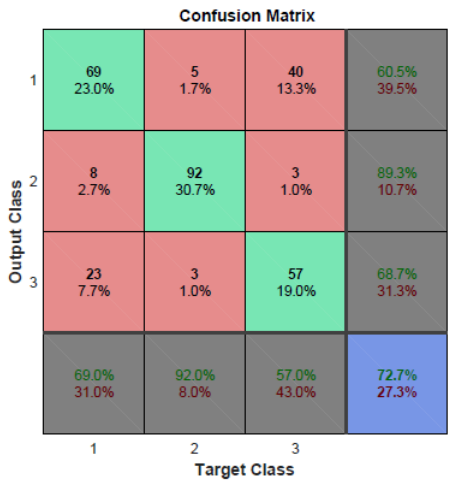
Χρήση δύο συνεχόμενων ομάδων MFCC



Σχήμα 5.8: Αποτελέσματα εκπαίδευσης ANN 72 Εισόδων



Σχήμα 5.9: Receiver Operating Characteristic Σε ANN 72 Εισόδων



Σχήμα 5.10: Δοκιμαστικό Δείγμα Για ANN 72 Εισόδων

Ο παρακάτω πίνακας συνοψίζει τα αθροιστικά αποτελέσματα για την περίπτωση των 72 εισόδων.

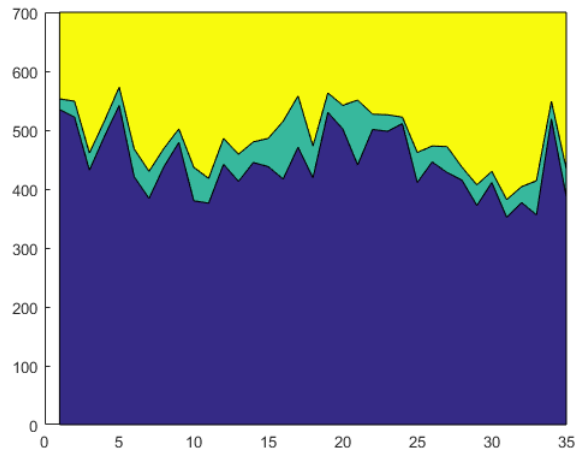
Αποτελέσματα ANN με 72 Εισόδους				
	Training Dataset			Testing Dataset
	Training	Validation	Testing	Testing
Success Rate (%)	76,9	75,2	75,2	72,7

Πίνακας 5.3: Αποτελέσματα ANN με 72 Εισόδους

Τέλος στην τρίτη μας προσέγγιση προσθέσαμε δύο συνεχόμενες ομάδες MFCC κάνοντας έτσι το σύστημα να πλησιάζει σε λεκτικά εξαρτώμενο αφού δημιουργείται μια χρονική εξάρτηση μεταξύ των δειγμάτων. Όπως έχουμε ορίσει και παραπάνω το σύστημα μας δεν είναι τέτοιου είδους και παρόλο που αύξησε το ποσοστό επιτυχίας κατά 1,5% που δεν είναι αρκετά μεγάλη βελτίωση και ταυτόχρονα αλλάζει τη μορφή του συστήματος μας, γι αυτό και δεν θα το χρησιμοποιήσουμε. Επομένως, όπως είναι φανερή η επιλογή των 36 εισόδων για το σύστημα αυτό λόγω αφενός του μεγαλύτερου ποσοστού επιτυχίας και αφετέρου δεν αλλάζει την φύση του συστήματος.

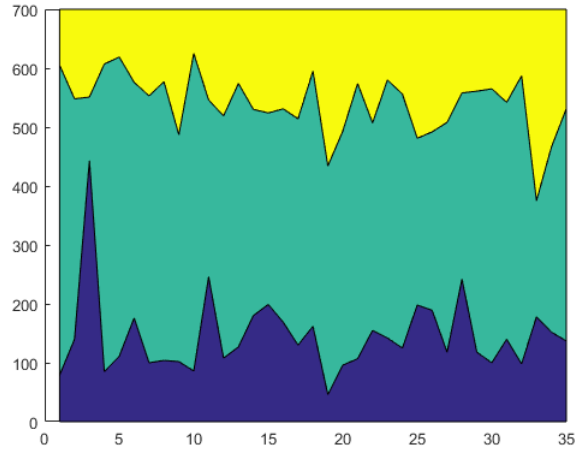
5.2.2 Ποσοστό επιτυχημένων Δειγμάτων

Για την πραγματοποίηση αυτού του πειράματος έπρεπε να χρησιμοποιήσουμε τα w_i,b_i του εκπαιδευμένου ANN και να το κατασκευάσουμε στο σύστημα μας. Έπειτα χρησιμοποιήσαμε το σύστημα με τον κάθε χρήστη για 35 δοκιμές με σκοπό να κρατήσουμε τον μέσο όρο επιτυχίας του κάθε χρήστη. Με βάση αυτό θα καθορίσουμε τις οριακές τιμές αναγνώρισης του κάθε χρήστη.



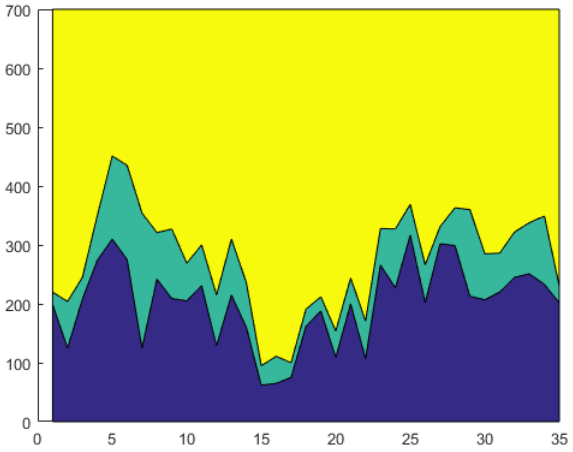
Σχήμα 5.11: Περιοχή Αποτελεσμάτων Χρήστη 1

Στην περίπτωση του πρώτου χρήστη παρατηρούμε έναν μέσο όρο σωστών αποτελεσμάτων κοντά στα 442 δείγματα. Επομένως ορίζοντας τα 400 ως ελάχιστο αριθμό επιτυχημένων δειγμάτων καταλήγουμε σε 77,14% ποσοστό επιτυχίας.



Σχήμα 5.12: Περιοχή Αποτελεσμάτων Χρήστη 2

Στην περίπτωση του δεύτερου χρήστη έχουμε μέσο όρο κοντά στα 394 δείγματα. Επομένως ορίζοντας τα 350 ως ελάχιστο αριθμό επιτυχημένων δειγμάτων καταλήγουμε σε 77,14 % ποσοστό επιτυχίας.



Σχήμα 5.13: Περιοχή Αποτελεσμάτων Χρήστη 3

Στην περίπτωση του τρίτου χρήστη έχουμε μέσο όρο κοντά στα 423 δείγματα. Επομένως ορίζοντας τα 380 ως ελάχιστο αριθμό επιτυχημένων δειγμάτων καταλήγουμε σε 68,6 % ποσοστό επιτυχίας.

Στον παρακάτω πίνακα έχουμε συγκεντρωμένα τα αποτελέσματα του παραπάνω πειράματος.

Πίνακας Παρουσίασης Αποτελεσμάτων Χρήσης Συστήματος			
Πιθανά Αποτελέσματα (%)	Χρήστες		
	Χρήστης 1	Χρήστης 2	Χρήστης 3
Επιτυχής Αναγνώριση	77,14	77,14	68,57
Αποτυχία Αναγνώρισης	22,86	20,00	31,42
Λανθασμένη Αναγνώριση	0,00	2,86	0,00

Πίνακας 5.4: Αποτελέσματα Χρήσης Συστήματος

6 Συμπεράσματα και Μελλοντικές Επεκτάσεις

6.1 Συμπεράσματα

Ολοκληρώνοντας αυτή τη μελέτη ενός λεκτικά ανεξάρτητου συστήματος αναγνώρισης ομιλητή βλέπουμε τις προοπτικές μιας τέτοιας εφαρμογής σε ένα σύστημα ασφαλείας και ειδικά σε ένα πολυπαραγοντικό σύστημα ασφαλείας αφού ως μόνος παράγοντας δεν προσφέρει τόσο ασφαλή αποτελέσματα.

Εστιάζοντας στα υλικά που χρησιμοποιήθηκαν περιοριστικός παράγοντας ήταν η ποιότητα του ήχου η οποία μειώθηκε αρκετά μετά την αντικατάσταση του ψηφιακού μικροφώνου με το αναλογικό και σίγουρα είχε ένα αντίκτυπο στα τελικά ποσοστά επιτυχίας. Από την άλλη, η μέθοδος εξόρυξης χαρακτηριστικών αποδείχθηκε ιδιαίτερα ικανή ακόμη και με την τελική ποιότητα ήχου. Ακόμη, παρατηρήθηκε πως οι σχέσεις μεταξύ των MFCC κρύβουν πληροφορίες για τη φωνή του ομιλητή και σαν αποτέλεσμα είναι σκόπιμη η χρήση αυτών σαν χαρακτηριστικά του ομιλητή. Έπειτα, αξίζει να αναφερθεί πως για να εκπαιδευτεί κατάλληλα το σύστημα πρέπει να μην υπάρχει θόρυβος στον χώρο ή τουλάχιστον η φωνή του ομιλητή να υπερισχύει όλων των υπόλοιπον ήχων. Κατά την ηχογράφηση του Χρήστη 2 ο ήχος που υπερισχύει δεν είναι η φωνή του ομιλητή κάνοντας έτσι το σύστημα να εστιάζει στην ένταση της φωνής του Χρήστη 2 και όχι τις τιμές των χαρακτηριστικών. Αυτός είναι και ο βασικός λόγος που το νευρωνικό έχει ταξινομήσει τόσο καλά τον Χρήστη 2 και αυτό δεν παρουσιάζεται στις δοκιμές του συστήματος.

Τέλος, παρατηρώντας τα αποτελέσματα της χρήσης του ANN για την αναγνώριση ομιλητή και συγκρίνοντας το με άλλες μεθόδους ταξινόμησης(GMM) παρατηρούμε μικρότερο ποσοστό επιτυχίας, ίσως κάποιο άλλο είδος νευρωνικού δικτύου να είχε καλύτερα αποτελέσματα, για παράδειγμα ένα Convolutional Neural Network(CNN). Βέβαια η σύγκριση του ANN δεν είναι πραγματική δεδομένου ότι δεν δοκιμάστηκε με κοινό δείγμα ώστε να έχουμε πραγματική εικόνα του αν είναι καλύτερο από κάποια άλλη μέθοδο ταξινόμησης.

6.2 Μελλοντικές Επεκτάσεις

Το σύστημα σε αυτή του τη μορφή έχει πολλές δυνατές επεκτάσεις που θα βελτίωναν τόσο τα ποσοστά επιτυχίας του συστήματος όσο και την χρηστικότητα του σε διαφόρων ειδών εφαρμογές. Αρχικά, αν θέλουμε να εξαλείψουμε τα προβλήματα ποιότητας ήχου πρέπει να αντικαταστήσουμε το μικρόφωνο που υπάρχει αυτή τη στιγμή στο σύστημα με ένα ψηφιακό ικανό να μας δώσει ήχο με τουλάχιστον 16-bit ανάλυσης.
Έπειτα, από άποψη αναγνώρισης μια καλή ιδέα είναι η μετατροπή του ηχογραφημένου μηνύματος σε κείμενο με σκοπό να προσθέσουμε έναν ακόμη παράγοντα ταυτοποίησης στο σύστημα μας. Συγκεκριμένα, εμφανίζοντας στον χρήστη μια τυχαία λέξη την οποία πρέπει να διαβάσει ώστε να αναγνωριστεί από το σύστημα πράγμα που μπορεί να εξαλείψει τον κίνδυνο αναγνώρισης του χρήστη από μια ηχογράφηση του όπως είπαμε σε προηγούμενο κεφάλαιο (Ενότητα 2.3). Επεκτείνοντας το παραπάνω σκεπτικό ίσως και η προσθήκη κάποιου άλλου αισθητήρα όπως ένας αισθητήρας δακτυλικού αποτυπώματος και η δημιουργία ενός αρχείου που θα αποθηκεύει πότε και ποιοι χρήστες είχαν πετυχημένη χρήση του συστήματος. Ακόμη, με τη σύνδεση του συστήματος με ένα κεντρικό υπολογιστή μπορούμε να μεταφέρουμε τις νέες επιτυχημένες ηχογραφήσεις στον ηλεκτρονικό υπολογιστή με σκοπό την επανεκπαίδευση του νευρωνικού δικτύου με σκοπό τα w_i,b_i να ανανεώνονται και το σύστημα να εκπαιδεύεται καλύτερα με τη χρήση του.
Με στόχο την σύγκριση του συστήματος με άλλα αντίστοιχα συστήματα χρήσιμο θα ήταν η εκπαίδευση του συστήματος με κάποιο κοινώς αποδεκτό δείγμα(dataset)ήχου με αποτέλεσμα να αξιολογηθεί επί ίσοις όροις με άλλα συστήματα αναγνώρισης ομιλητή.

Βιβλιογραφία

[1] Titze I. R. ***Principles of Voice Production***. Prentice-Hall, 1994.

[2] Laukkanen A. & Leino T. ***Ihmeellinen ihmisääni***. Gaudeamus, 1999.

[3] Story B. H. "An overview of the physiology, physics and modeling of the sound source for vowels". In: ***Acoustical Science and Technology,5*** 23.4 (2002), pp. 195–206.

[4] Hannu Pulakka. "Analysis of Human Voice Production Using Inverse Filtering, High-Speed Imaging, and Electroglottography". In: (2005).

[5] Karjalainen M. ***Kommunikaatioakustiikka***. Otamedia Oy, 2000.

[6] Bosch L. & Compernolle D. V. Claes T. Dologlou I. "A novel feature transformation for vocal tract length normalization in automatic speech recognition". In: ***IEEE Transactions on Speech and Audio Processing*** 6.6 (), pp. 549–557.

[7] Abhinav Anand, Thomas H. Lee, and Fabio Scotti. "Text-Independent Speaker Recognition for Ambient Intelligence Applications by Using Information Set Features". In: ***IEEE Journal of Solid-State Circuits*** 38.12 (2003), pp. 2269–2279.

[8] F. K. Soong, A. E. Rosenberg, L. R. Rabiner and B. H. Juang. "A Vector Quantization Approach to Speaker Recognition". In: ***International Conference on Acoustics, Speech and Signal Processing*** (1985).

[9] Michael Nielsen. ***Neural Networks and Deep Learning***. Free Online book, 2017.

[10] Malcolm Slaney Tom Roberts and Dimitrios P. Bouras. ***FFTs using fixed point arithmetic in C***. https://gist.github.com/Tomwi/3842231. 1989.