ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

Αυτόματη Αναγνώριση Γλώσσας

Γιαννουρής Πολύδωρος - 9746
Καϊμακαμίδης Ανέστης - 9627
Γρηγοράκης Ευθύμιος - 9694
Μπαρμπουνάκης Κωνσταντίνος - 9759

Εργασία στο μάθημα

Τεχνολογία του Ήχου και της Εικόνας: Καταγραφή, επεξεργασία, μετάδοση

στην

Πολυτεχνική Σχολή

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Τηλεπικοινωνιών

Θεσσαλονική, 2023

Περίληψη

Η δυνατότητα αυτόματης αναγνώρισης της ομιλούμενης γλώσσας ενός ηχητικού σήματος συνιστά μια σημαντική εργασία με πολλές πρακτικές εφαρμογές όπως η αναγνώριση ομιλίας, η μετάφραση γλώσσας και η εκμάθηση γλωσσών. Το πρόβλημα αυτό χαρακτηρίζεται από αυξημένη πολυπλοκότητα λόγω του υψηλού βαθμού μεταβλητότητας των ακουστικών χαρακτηριστικών των σημάτων ομιλίας, τα οποία μπορούν να επηρεαστούν από διάφορους παράγοντες όπως η προφορά, ο τονισμός και ο θόρυβος του περιβάλλοντος του ομιλητή. Επιπλέον, ο αριθμός των γλωσσών που πρέπει να ταξινομηθούν μπορεί να είναι αρκετά μεγάλος. Στην παρούσα εργασία, διερευνάται η χρήση νευρωνικών δικτύων για την ταξινόμηση των ομιλούμενων γλωσσών με βάση τα χαρακτηριστικά του ακουστικού σήματος. Ακομή, προτείνεται μια καινοτόμος ιεραρχική προσέγγιση που αποτελείται από δύο επίπεδα νευρωνικών δικτύων. Η κεντρική ιδέα της προσέγγισης βασίζεται στο διαχωρισμό του αρχικού προβλήματος σε υποπροβλήματα. Ειδικότερα, προτείνεται η απόδοση του σήματος εισόδου σε μια συγκεκριμένη ομάδα με έναν αρχικό ταξινομητή ομάδων γλωσσών και στη συνέχεια, λαμβάνεται εσωτερικά της ομάδας με έναν δεύτερο ταξινομητή η τελική απόφαση της ταξινόμησης. Η ανάλυση αυτή είναι ιδιαίτερα ελπιδοφόρα, καθώς επιτρέπει την αποτελεσματική και ακριβή ταξινόμηση των γλωσσών ακόμη και με την παρουσία μεγάλου αριθμού γλωσσών. Για την εξαγωγή σχετικών ακουστικών χαρακτηριστικών από το σήμα εισόδου, χρησιμοποιούνται τυπικές τεχνικές όπως οι Τράπεζες Φίλτρων (Filter Banks) σε κλίμακα "Mel". Αυτά τα χαρακτηριστικά τροφοδοτούνται στα νευρωνικά δίκτυα, τα οποία εκπαιδεύονται χρησιμοποιώντας ένα μεγάλο ελεύθερο σύνολο δεδομένων με επισημασμένα (annotated) δείγματα ομιλίας. Συνολικά, η μελέτη καταδεικνύει την αποτελεσματικότητα της χρήσης των νευρωνικών δικτύων για την αναγνώριση της ομιλούμενης γλώσσας και συνάγονται ορισμένα αξιοσημείωτα συμπεράσματα.

Περιεχόμενα

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

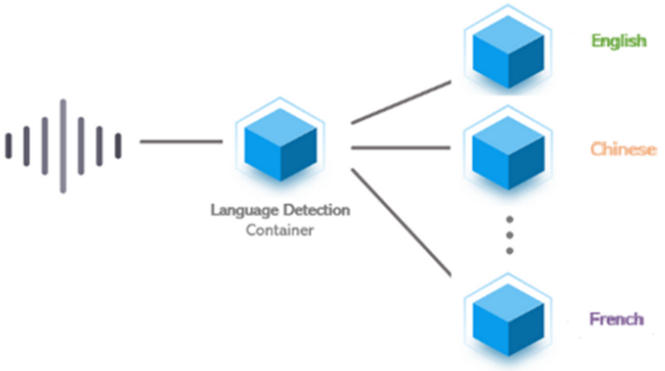
Στην παρούσα εργασία θα μελετηθεί το αντικείμενο της αυτόματης αναγνώρισης γλώσσας από αρχείο ήχου. Η ανάγκη ενός τέτοιου εργαλείου γίνεται εμφανής με την εξάπλωση των εικονικών βοηθών σε μη αγγλόφωνες χώρες και σε πολυγλωσσικά περιβάλλοντα όπου για την κατανόηση των εντολών του χρήστη θα πρέπει να προηγηθεί αναγνώριση γλώσσας. Ακόμη η εκθετική αύξηση των αρχείων ήχου στο διαδίκτυο καθιστά ασύμφορη την χειροκίνητη επισήμανση της γλώσσας πριν προβούμε σε αυτόματη παραγωγή υποτίτλων, ενώ τα υπάρχοντα συστήματα δεν καλύπτουν τις σύγχρονες ανάγκες.

- Σε αυτό το κεφάλαιο έγινε μία εισαγωγή στο πρόβλημα της ηχητικής αναγνώρισης γλώσσας.
- Στο κεφάλαιο 2 θα ορίσουμε το πρόβλημα καθώς και τις παραδοχές που κάνουμε.
- Στο κεφάλαιο 3 θα παρουσιάσουμε υπάρχουσες τεχνικές για αναγνώριση ήχου.
- Στο κεφάλαιο 4 θα προτείνουμε μία διαφορετική μεθοδολογία βασισμένη στον τρόπο που προσεγγίζουν οι άνθρωποι ένα τέτοιο πρόβλημα.
- Στο κεφάλαιο 5 θα αναλύσουμε την υλοποίηση του συστήματος στα δεδομένα μας.
- Στο κεφάλαιο 6 θα αξιολογήσουμε την προτεινόμενη προσέγγιση.
- Στο κεφάλαιο 7 θα εξάγουμε συμπεράσματα και θα συγκρίνουμε το σύστημά μας με τα υπάρχοντα ως προς την πολυπλοκότητα, την ακρίβεια, την ανάγκη σε δεδομένα εκπαίδευσης κ.α.

Κεφάλαιο 2

Διατύπωση του προβλήματος και παραδοχές

Το πρόβλημα που καλούμαστε να επιλύσουμε είναι η αναγνώριση της ομιλούμενης γλώσσας σε ένα αρχείο ήχου (κατηγοριοποίηση). Τέτοιου είδους προβλήματα είναι ιδανικά για χρήση τεχνικών μηχανικής μάθησης καθώς αφενός είναι δύσκολη η χρήση παραδοσιακού προγραμματισμού και κανόνων, αφετέρου υπάρχει μεγάλος όγκος ηχογραφήσεων σε πολλές γλώσσες. Η χρήση όμως τέτοιων μοντέλων ενέχει και πολλά μειονεκτήματα. Αρχικά θα πρέπει να εξασφαλίσουμε πως το δίκτυο θα μαθαίνει τις ειδοποιούς διαφορές των γλωσσών και όχι τα χαρακτηριστικά των ομιλητών που βρίσκονται στο σύνολο εκπαίδευσης. Ακόμη η αντιμετώπιση του προβλήματος αυτού με μεγαλύτερο σύνολο εκπαίδευσης συνεπάγεται αύξηση του απαιτούμενου αποθηκευτικού χώρου και σημαντικά περισσότερο χρόνο εκπαίδευσης. Τέλος για τις υπάρχουσες προσεγγίσεις είτε δε προσφέρεται ανοιχτός κώδικας, είτε ο αριθμός γλωσσών είναι μικρός οπότε αγνοείται το ζήτημα αναγνώρισης των συγγενικών (Guha et al., 2020). Για τα δεδομένα κάνουμε τις εξής παραδοχές. Πρώτον θεωρούμε ότι κάθε αρχείο ήχου περιέχει έναν ομιλητή μίας γλώσσας. Ο αριθμός γλωσσών που θα χρησιμοποιηθούν θα καθοριστεί κατά την υλοποίηση.



Σχήμα 2.1: Μπλοκ Διάγραμμα του Προβλήματος

Κεφάλαιο 3

Θεωρητικό υπόβαθρο και προσεγγίσεις

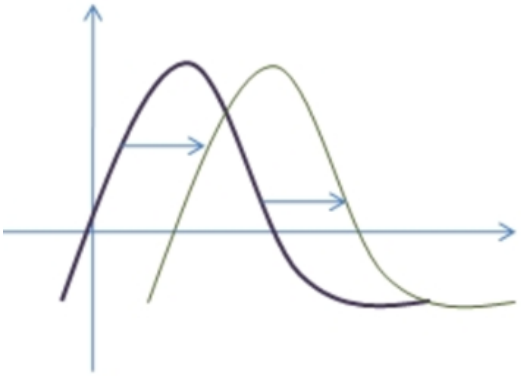
3.1 Επαύξηση δεδομένων

Η επαύξηση δεδομένων εφαρμόζεται με σκοπό την τεχνητή δημιουργία δεδομένων όταν δεν επαρκούν τα διαθέσιμα αλλά και για δημιουργία πιο robust δικτύων (Xie et al., 2020). Παρότι έχει εδραιωθεί στον χώρο της Μηχανικής Όρασης (Computer Vision), συχνά χρησιμοποιείται και στον ήχο. Παρακάτω παρουσιάζουμε συνοπτικά ορισμένες από τις μεθόδους επαύξησης που θα χρησιμοποιήσουμε.

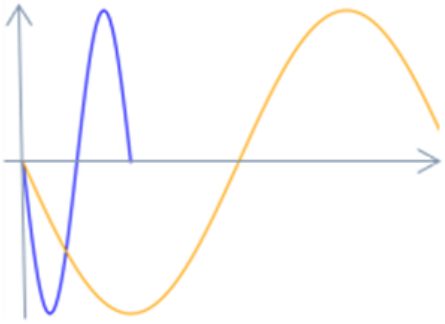
- Χρονική Μετατόπιση

  Στη χρονική μετατόπιση μετατοπίζουμε τον ήχο ως προς τον χρόνο. Αυτό δεν αλλοιώνει την υπάρχουσα πληροφορία αλλά βοηθάει το σύστημα να μην επηρεάζεται από τον χρόνο έναρξης της ομιλίας.

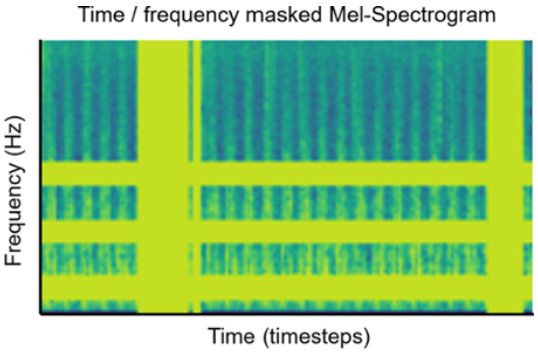
- Χρονική Διαστολή/Συστολή



Σχήμα 3.1: Χρονική μετατόπιση του μπλε σήματος. Το τελικό σήμα αναπαριστάται με πράσινο χρώμα.



Σχήμα 3.2: Παράδειγμα Χρονικής διαστολής σήματος



Σχήμα 3.3: Χρονική και συχνοτικής απόκρυψη σε φασματογράφημα.

Στη χρονική διαστολή/συστολή χωρίς αλλαγή του τόνου μπορούμε να αυξομειώσουμε την διάρκεια του ήχου.

- Αλλαγή τόνου

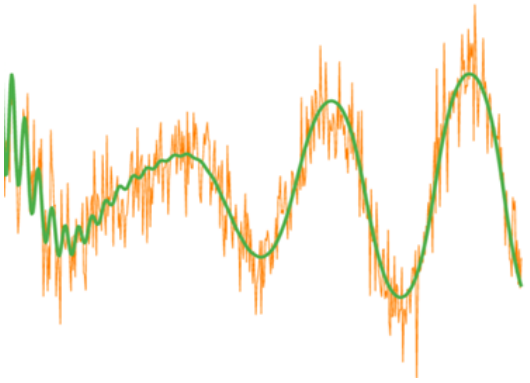
  Στην αλλαγή τόνου χρησιμοποιούνται τεχνικές επεξεργασίας σήματος, οι οποίες τροποποιούν το φασματικό περιεχόμενο με σκοπό την αύξηση ή μείωση του αρχικού τόνου (pitch). Οι αλλαγές που προκαλούνται δεν επηρεάζουν το ρυθμό και τη διάρκεια του ήχου στο πεδίο του χρόνου.

- Χρονο-Φασματική Απόκρυψη

  Στη Χρονο-Φασματική Απόκρυψη "κρύβουμε" από το σύστημα ορισμένες συχνότητες ή/και χρονικές στιγμές. Ο σκοπός είναι αυτό να οδηγηθεί στο να συμπεραίνει ποια είναι η γλώσσα χωρίς πλήρη γνώση, δηλαδή σε πιο ρεαλιστικές συνθήκες.

- Προσθήκη θορύβου

  Σε αυτήν την περίπτωση πραγματοποιείται επιτηδευμένα προσθήκη κάποιου



Σχήμα 3.4: Παράδειγμα προσθήκης θορύβου σε σήμα

είδους θορύβου στο αρχικό σήμα. Με αυτόν τον τρόπο, επιτυγχάνεται η προσομοίωση ρεαλιστικών καταστάσεων, ενισχύεται η γενίκευση των αποτελεσμάτων και ελέγχεται η ευρωστία του συστήματος.

3.2 Τεχνικές Επεξεργασίας Δεδομένων

- Φασματογράφημα (Spectogram)

  Το φασματόγραμμα συνιστά μια οπτική αναπαράσταση του συχνοτικού περιεχομένου ενός σήματος ως προς το χρόνο. Λαμβάνεται μέσω διαίρεσης του σήματος σε μικρά τμήματα, συνήθως χρησιμοποιώντας μια συνάρτηση παραθύρου, και στη συνέχεια υπολογίζοντας το φασματικό περιεχόμενο κάθε τμήματος χρησιμοποιώντας Μετασχηματισμό "Fourier" ή Μετασχηματισμό Κυματιδίων (Wavelet Transform). Το περιεχόμενο συχνότητας που προκύπτει σχεδιάζεται με την πάροδο του χρόνου για να δημιουργηθεί μια δισδιάστατη εικόνα, όπου ο άξονας x αντιπροσωπεύει το χρόνο και ο άξονας y αντιπροσωπεύει τη συχνότητα. Η τελική αναπαράσταση πραγματοποιείται συνήθως με χάρτες θερμότητας (heat maps). Χρησιμοποιείται ευρέως σε εφαρμογές που εμπλέκουν ηχητικά σήματα και ενδείκνυται για εργασίες όπως η αναγνώριση φωνητικών λέξεων.

- Κλίμακα "Mel"

  Η Κλίμακα "Mel" είναι ένας μετασχηματισμός της μονάδας μέτρησης της συχνότητας (Hz) σε μια πιο κατανοήτη κλίμακα για το ανθρώπινο αυτί. Η χρήση της στην επεξεργασία ήχου και ομιλίας επιτρέπει την ακριβέστερη μοντελοποίηση της ανθρώπινης αντίληψης του ήχου. Χρησιμοποιείται συνήθως στο σχεδιασμό τραπεζών φίλτρων (Filter Banks) για ανάλυση ομιλίας και ήχου, καθώς και στην ανάπτυξη αλγορίθμων για την αναγνώριση ομιλίας και την ανάκτηση πληροφοριών μουσικής. Η Κλίμακα "Mel" βασίζεται στην ιδέα ότι το ανθρώπινο αυτί αντιλαμβάνεται μικρότερες διαφορές μεταξύ ήχων χαμηλότερης συχνότητας

συγκριτικά με ήχους υψηλότερης συχνότητας. Επομένως, είναι μη γραμμική και συμπιέζει τον άξονα συχνότητας στις υψηλότερες συχνότητες. Η αντιστοίχιση της συχνότητας στην Κλίμακα "Mel" ορίζεται ως:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \tag{3.1}$$

- Τράπεζες Φίλτρων (Filter Banks)

  Οι τράπεζες φίλτρων συνιστούν μια τεχνική επεξεργασίας σήματος που χρησιμοποιείται για τη διαίρεση ενός σήματος σε πολλαπλές ζώνες συχνοτήτων. Ειδικότερα, χρησιμοποιούν συνήθως ένα σύνολο φίλτρων, τα οποία έχουν σχεδιαστεί έτσι ώστε να επιτρέπουν σε ορισμένες συχνότητες να περνούν ενώ μπλοκάρουν άλλες. Κάθε φίλτρο στην τράπεζα λειτουργεί σε διαφορετική ζώνη συχνοτήτων, επιτρέποντας στο σήμα να διαχωριστεί στα συστατικά στοιχεία του. Εξυπηρετούν ένα ευρύ φάσμα λειτουργιών όπως συμπίεση ήχου, εικόνας και αναγνώριση ομιλίας ενώ χρησιμοποιούνται επίσης στην επεξεργασία ψηφιακών σημάτων και στην υλοποίηση μετασχηματισμών κυματιδίων.

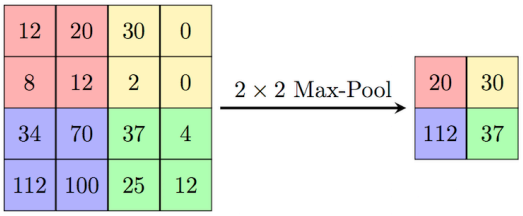
- Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA)

  Η Ανάλυση Κύριων Συνιστωσών αποτελεί μια από τις πιο διαδεδομένες τεχνικές προ-επεξεργασίας δεδομένων (pre-processsing). Αποτελεί μια στατιστική διαδικασία με την οποία αναπαριστούμε έναν πίνακα συμμεταβλητότητας ενός συνόλου μεταβλητών μέσα από ένα νέο σύνολο μεταβλητών, οι οποίες προκύπτουν από τον γραμμικό συνδυασμό των αρχικών μεταβλητών. Σκοπός της συγκεκριμένης μεθόδου είναι η μείωση διάστασης των δεδομένων, διατηρώντας παράλληλα ένα σημαντικό ποσοστό της αρχικής πληροφορίας.

- Μέγιστη Συγκέντρωση (Max Pooling)

  Η Μέγιστη Συγκέντρωση είναι μια τεχνική που χρησιμοποιείται στα Συνελικτικά Νευρωνικά Δίκτυα (CNN) για τη μείωση των χωρικών διαστάσεων των χαρτών χαρακτηριστικών (feature maps). Ο σκοπός της είναι η μείωση του δείγματος των χαρτών χαρακτηριστικών, διατηρώντας παράλληλα τις πιο σημαντικές πληροφορίες.

  Στη Μέγιστη Συγκέντρωση, ένα παράθυρο ή ένας πυρήνας σταθερού μεγέθους μετακινείται στον χάρτη χαρακτηριστικών και για κάθε θέση του παραθύρου, η μέγιστη τιμή εντός του παραθύρου λαμβάνεται ως έξοδος. Με αυτόν τον τρόπο, διατηρούνται τα πιο σημαντικά χαρακτηριστικά και μειώνεται η ευαισθησία σε μικρές διακυμάνσεις στην είσοδο. Επιπλέον, η μείωση διάστασης συνεπάγεται τη μείωση της υπολογιστικής πολυπλοκότητας και τη μείωση του κινδύνου



Σχήμα 3.5: Παράδειγμα Μέγιστης Συγκέντρωσης σε Χάρτη Χαρακτηριστικών 2 Διαστάσεων

υπερπροσαρμογής (overfitting). Τέλος, η παραπάνω μέθοδος δύναται να πραγματοποιηθεί με διάφορα μεγέθη παραθύρων και τιμές διασκελισμού ανάλογα με τον επιθυμητό βαθμό μείωσης του δείγματος.

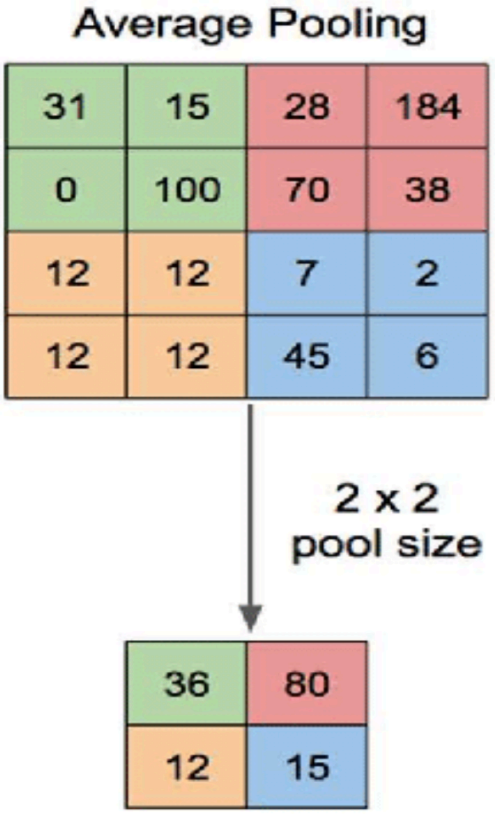
- Μεσοσταθμική Συγκέντρωση (Average Pooling)

  Η Μεσοσταθμική Συγκέντρωση αποτελεί επίσης τεχνική που χρησιμοποιείται στα Συνελικτικά Νευρωνικά Δίκτυα για τη μείωση των χωρικών διαστάσεων των χαρτών χαρακτηριστικών. Λειτουργεί με παρόμοιο τρόπο με τη Μέγιστη Συγκέντρωση. Η ειδοποιός διαφορά εντοπίζεται στην τιμή της εξόδου. Στην προκειμένη περίπτωση, λαμβάνεται ως έξοδος η μέση τιμή εντός του παραθύρου. Συγκριτικά με τη Μέγιστη Συγκέντρωση, η οποία επιλέγει τη μέγιστη τιμή, η Μεσοσταθμική Συγκέντρωση παρέχει πιο ομαλή έξοδο με λιγότερες απώλειες πληροφοριών, αλλά ενδέχεται να μην διατηρεί τις σημαντικότερες πληροφορίες της εισόδου. Η επιλογή ανάμεσα στις τεχνικές αυτές εξαρτάται συχνά από τις συγκεκριμένες απαιτήσεις της εφαρμογής και τα χαρακτηριστικά των δεδομένων εισόδου.

- Κανονικοποίηση Δέσμης (Batch Normalization)

  Η Κανονικοποίηση Δέσμης είναι μια τεχνική που χρησιμοποιείται στη βαθιά μάθηση (Deep Learning) για τη βελτίωση της ταχύτητας και της σταθερότητάς των δικτύων. Κανονικοποιεί την είσοδο κάποιου επιπέδου (layer) αφαιρώντας τον μέσο όρο και διαιρώντας με την τυπική απόκλιση της τρέχουσας δέσμης δεδομένων κατά τη διάρκεια της εκπαίδευσης.

  Ο κύριος σκοπός της τεχνικής αυτής είναι η αντιμετώπιση του ζητήματος της εσωτερικής μετατόπισης συμμεταβλητών, η οποία συμβαίνει όταν η κατανομή της εισόδου σε ένα επίπεδο αλλάζει κατά τη διάρκεια της εκπαίδευσης εξαιτίας της ανανέωσης των παραμέτρων στα προηγούμενα επίπεδα. Με την κανονικοποίηση της εισόδου σε κάθε επίπεδο, μειώνεται η εξάρτηση της κατανομής της εισόδου από τις παραμέτρους των προηγούμενων επιπέδων, επιταχύνεται η σύγκλιση και σταθεροποιείται η διαδικασία εκπαίδευσης. Συνήθως εισάγεται μετά τη συνάρτηση ενεργοποίησης κάθε επιπέδου και πριν από το επόμενο επίπεδο. Μπορεί να



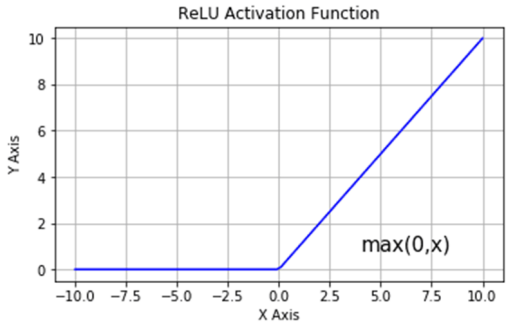
Σχήμα 3.6: Παράδειγμα Μεσοσταθμικής Συγκέντρωσης σε Χάρτη Χαρακτηριστικών 2 Διαστάσεων

εφαρμοστεί σε διάφορους τύπους αρχιτεκτονικών νευρωνικών δικτύων, συμπεριλαμβανομένων Πλήρως Συνδεδεμένων και Συνελικτικών Νευρωνικών Δικτύων.

- "Dropout"

  Η τεχνική "Dropout" αποτελεί μια εναλλακτική μορφή ομαλοποίησης (regularization) από την Κανονικοποίηση Δέσμης με σκοπό την αποφυγή της υπερπροσαρμογής στη διαδικασία εκπαίδευσης. Ειδικότερα, κατά τη διάρκεια της εκπαίδευσης, επιλέγεται τυχαία ένα υποσύνολο των νευρώνων σε ένα επίπεδο όπου αφαιρούνται προσωρινά από το δίκτυο, δημιουργώντας ουσιαστικά ένα μικρότερο δίκτυο. Συνεπώς, το νευρωνικό δίκτυο εμποδίζεται από το να βασιστεί εντονότερα σε κάποιο νευρώνα ή σύνολο νευρώνων και ενθαρρύνεται η εκμάθηση πιο ισχυρών και γενικεύσιμων χαρακτηριστικών. Συνήθως εισάγεται μετά τη συνάρτηση ενεργοποίησης κάθε επιπέδου και πριν από το επόμενο επίπεδο, ενώ παράλληλα δύναται να εφαρμοστεί σε διάφορους τύπους αρχιτεκτονικών νευρωνικών δικτύων.

- Συνάρτηση Ενεργοποίησης Ανορθωμένη Γραμμική Μονάδα (Rectified Linear Unit - ReLU)



Σχήμα 3.7: Γραφική Παράσταση Συνάρτησης Ενεργοποίησης Ανορθωμένης Γραμμικής Μονάδας

Εφαρμόζεται με σκοπό το μη γραμμικό μετασχηματισμό των δεδομένων εξόδου ένος επιπέδου στα νευρωνικά δίκτυα. Πρακτικά, μηδενίζει τις αρνητικές τιμές και αφήνει αναλλοίωτες τις θετικές τιμές. Η μαθηματική αναπαράσταση της συνάρτησης είναι:

$$f(x) = max(0, x) \tag{3.2}$$

- Συνάρτηση Ενεργοποίησης "Softmax"

  Αποτελεί μια ευρέως χρησιμοποιούμενη συνάρτηση ενεργοποίησης σε νευρωνικά δίκτυα, ιδιαίτερα για προβλήματα ταξινόμησης πολλαπλών κλάσεων. Παίρνει ως είσοδο ένα διάνυσμα αυθαίρετων πραγματικών αριθμών και εξάγει ένα άλλο διάνυσμα του ίδιου μήκους, όπου κάθε στοιχείο του διανύσματος εξόδου αντιπροσωπεύει την πιθανότητα της αντίστοιχης κλάσης. Ουσιαστικά, αποδίδει την πιθανοκρατική εκτίμηση του κατηγοριοποιητή. Το συνολικό άθροισμα της εξόδου ισούται με 1, ενώ ταυτόχρονα συνηθίζεται ο κατηγοριοποιητής να αποφασίζει την κλάση στην οποία αποδίδεται η μεγαλύτερη πιθανότητα. Η μαθηματική αναπαράσταση της συνάρτησης είναι:

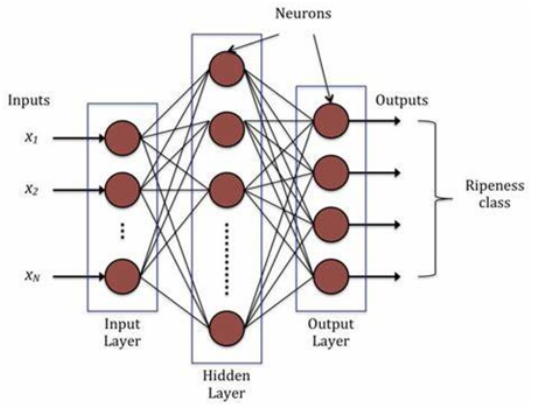
$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \qquad i = 1, 2, \ldots, K \tag{3.3}$$

3.3 Μοντέλα μηχανικής μάθησης

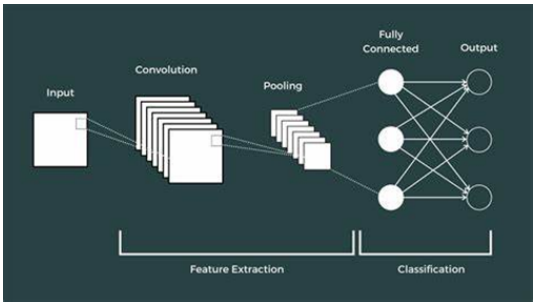
Για την κατηγοριοποίηση ήχου στη συγκεκριμένη εργασία εφαρμόζονται συνδυαστικά οι εξής αλγόριθμοι:

- Πολυστρωματικό "Perceptron" (Multilayer Perceptron - MLP)

  Αυτό που διακρίνει το Πολυστρωματικό "Perceptron" από τα κοινά νευρωνικά δίκτυα είναι η μη-γραμμική συνάρτηση ενεργοποίησης (nonlinear activation function) και η πλήρης σύνδεση των νευρώνων μεταξύ τους. Αποτελείται από ένα επίπεδο εισόδου (input layer), ένα ή περισσότερα κρυφά επίπεδα (hidden layers) και ένα επίπεδο εξόδου (output layer). Η έξοδος του τελευταίου επιπέδου



Σχήμα 3.8: Πολυστρωματικό "Perceptron" με ένα κρυφό επίπεδο.



Σχήμα 3.9: Παράδειγμα κατηγοριοποίησης δισδιάστατης εισόδου με συνελικτικό νευρωνικό δίκτυο.

συνιστά την πρόβλεψη του μοντέλου. Χρησιμοποιούνται τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης.

- Συνελικτικό Νευρωνικό Δίκτυο (Convolutional Neural Network - CNN)

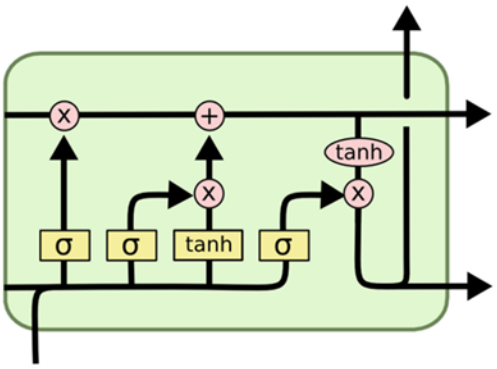
  Το Συνελικτικό Νευρωνικό Δίκτυο εφαρμόζει συνελίξεις της εισόδου με μάσκες, εξάγοντας χαρακτηριστικά. Μετά από ένα επίπεδο μείωσης της διάστασης των χαρακτηριστικών απομένουν τα κυριότερα, από τα οποία με ένα πλήρως συνδεδεμένο επίπεδο εξάγεται η απόφαση.

- Νευρωνικό Δίκτυο "Long-Short Term Memory" (LSTM)

  Πρόκειται για μια ειδική κατηγορία των Νευρωνικών Δικτύων με Ανάδραση (Recurrent Neural Networks - RNN). Το κύριο χαρακτηριστικό του είναι ότι περιέχει μια forget-gate η οποία αποφασίζει αν η πληροφορία είναι απαραίτητη και θα περάσει στο επόμενο επίπεδο. Έχει την ιδιότητα εξαγωγής μακρινών συσχετίσεων στα δεδομένα.

3.4 Μετρικές Αξιολόγησης και Αναπαράσταση Αποτελεσμάτων

- Πίνακας Σύγχυσης (Confusion Matrix)



Σχήμα 3.10: Παράδειγμα Βρόγχου Νευρωνικού Δικτύου LSTM

Ο Πίνακας Σύγχυσης χρησιμοποιείται για την αξιόλογηση της απόδοσης των μοντέλων σε προβλήματα ταξινόμησης. Συγκεκριμένα, συνοψίζεται η απόδοση του μοντέλου συγκρίνοντας τις προβλεπόμενες τιμές (predicted values) με τις πραγματικές τιμές (actual values) των δεδομένων.

Στον Πίνακα Σύγχυσης περιέχονται 4 διαφορετικές μετρικές:

(α΄) Αληθώς Θετικά (True Positives - TP): Αντιπροσωπεύουν τον αριθμό των περιπτώσεων που ταξινομήθηκαν σωστά ως θετικά από το μοντέλο.

(β΄) Ψευδώς Θετικά (False Positives - FP): Αντιπροσωπεύουν τον αριθμό των περιπτώσεων που ταξινομήθηκαν λανθασμένα ως θετικά.

(γ΄) Αληθώς Αρνητικά (True Negatives - TN): Αντιπροσωπεύουν τον αριθμό των περιπτώσεων που ταξινομήθηκαν σωστά ως αρνητικά.

(δ΄) Ψευδώς Αρνητικά (False Negatives - FN): Αντιπροσωπεύουν τον αριθμό των περιπτώσεων που ταξινομήθηκαν λανθασμένα ως αρνητικά.

Χρησιμοποιώντας τις παραπάνω τιμές μπορούν να υπολογιστούν διάφορες άλλες μετρικές αξιολόγησης, ενώ ταυτόχρονα ο Πίνακας Σύγχυσης αποτελεί ιδανικό τρόπο οπτικοποίησης των αποτελεσμάτων ταξινόμησης.

- Ορθότητα (Accuracy)

  Εκφράζει το λόγο του αριθμού των σωστών προβλέψεων του μοντέλου με τον συνολικό αριθμό δειγμάτων εισόδου. Όσο καλύτερα είναι διανεμημένα τα δεδομένα στις κλάσεις τόσο πιο αξιόπιστη είναι.

Κεφάλαιο 4

Προτεινόμενη προσέγγιση - Μεθοδολογία

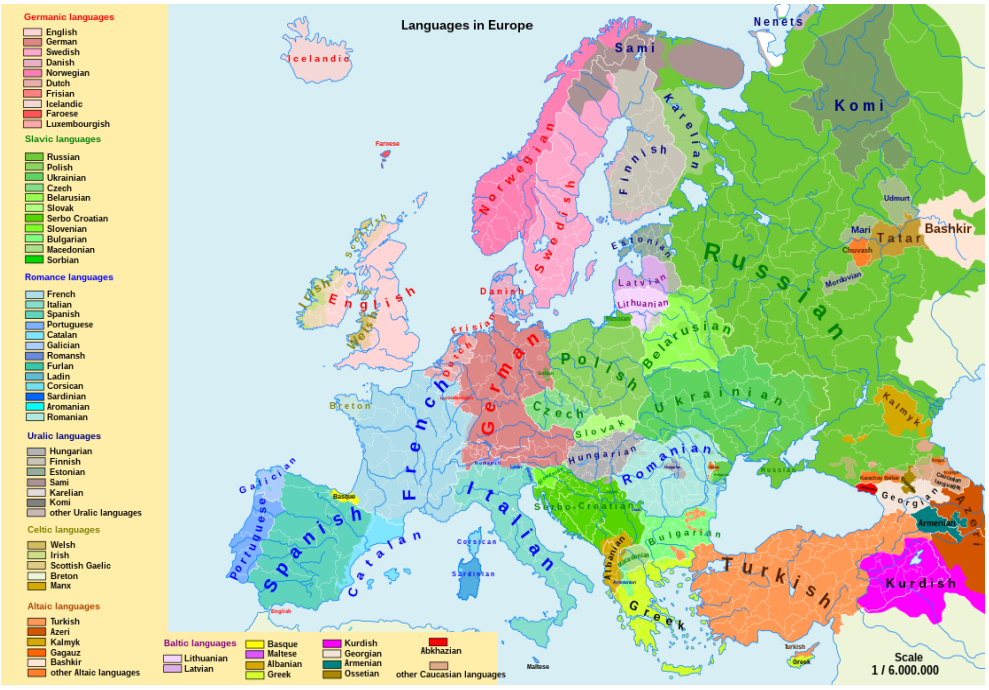
Ο τρόπος προσέγγισης του προβλήματος αναγνώρισης της ομιλούμενης γλώσσας ενός αρχείου ήχου διαδραματίζει σημαίνοντα ρόλο στο τελικό αποτέλεσμα. Είναι γεγονός ότι υπάρχει τεράστια ποικιλομορφία ομιλούμενων γλωσσών. Επιπλέον, με την πάροδο των γενεών, παρατηρούνται συχνά έντονες διαφορές ανάμεσα στον τρόπο ομιλίας μιας συγκεκριμένης γλώσσας από διαφορετικούς ομιλητές. Για παράδειγμα, οι διαφορετικές διάλεκτοι εντός μιας συγκεκριμένης γλώσσας, οι διαφοροποιήσεις στον τονισμό, την προφορά και τον τόνο από μια γεωγραφική περιοχή σε μια άλλη, αλλά και οι ιδιαιτερότητες στον τρόπο ομιλίας του εκάστοτε ομιλητή αυξάνουν την πολυπλοκότητα του αρχικού προβλήματος.

Όσον αφορά στις ιδιατερότητες μεταξύ των διαφορετικών γλωσσών, παρατηρείται επίσης έντονη ποικιλομορφία ως προς τον τονισμό, την προφορά, το ύφος, τον τόνο κ.ά. Η γεωγραφική περιοχή, το φύλο και η προσωπικότητα του ομιλητή επηρεάζουν σε μεγάλο βαθμό τον παραγόμενο ήχο. Στην παρούσα εργασία δοκιμάζονται και αντιπαραβάλλονται 2 διαφορετικές προσεγγίσεις.

- 1η Προσέγγιση - Απευθείας Κατηγοριοποίηση

  Αποτελεί τον πιο απλό τρόπο προσέγγισης του προβλήματος. Αφού καθοριστεί το σύνολο δεδομένων και το πλήθος των διαφορετικών γλωσσών (κλάσεων) προς ταξινόμηση, εκπαιδεύεται ένας κατηγοριοποιητής ο οποίος προβλέπει απευθείας την ομιλούμενη γλώσσα σύμφωνα με ορισμένα ακουστικά χαρακτηριστικά.

- 2η Προσέγγιση - Κατηγοριοποίηση Ομάδας και Εσωτερική Αναγνώριση Γλώσσας



Σχήμα 4.1: Παράδειγμα Ομαδοποίησης Ευρωπαϊκών Γλωσσών με βάση την προέλευση

Αποτελεί έναν καινοτόμο, πιο σύνθετο τρόπο προσέγγισης του προβλήματος. Σε αυτήν την εναλλακτική προσέγγιση, το αρχικό πρόβλημα ανάγεται στην επίλυση 2 υποπροβλημάτων. Αρχικά, επιχειρείται η κατηγοριοποίηση της ομιλούμενης γλώσσας σε μια ομάδα-οικογένεια γλωσσών με έναν αρχικό κατηγοριοποιητή (Επίπεδο 1). Στη συνέχεια, τα δεδομένα μεταβιβάζονται σε έναν δεύτερο ταξινομητή ο οποίος αποφασίζει την ομιλούμενη γλώσσα εντός μιας συγκεκριμένης ομάδας γλωσσών (Επίπεδο 2). Εάν N είναι το πλήθος των διαφορετικών ομάδων γλωσσών, τότε υπάρχουν N διαφορετικοί ταξινομητές στο δεύτερο επίπεδο κατηγοριοποίησης.

Ο καθορισμός των ομάδων γλωσσών αποτελεί από μόνος του ένα δισεπίλυτο πρόβλημα. Ενδεικτικοί τρόποι εντοπισμού των κλάσεων των Ομάδων Γλωσσών είναι οι εξής:

1. Γλωσσολογικές Αναλύσεις:

   Η ομαδοποίηση των γλωσσών βασίζεται στις επιστημονικές αναλύσεις των γλωσσολόγων με βάση την προέλευση, τη δομή, τη σύνταξη ή κάποιο άλλο χαρακτηριστικό των γλωσσών.

2. "Trial and Error":

   Αρχικά, χρησιμοποιείται ένας οποιοσδήποτε τρόπος ομαδοποίησης (τυχαία, μέσω γλωσσολογικών αναλύσεων κ.ά) επιλέγοντας το αρχικό πλήθος ομάδων-οικογενειών και τις επιμέρους γλώσσες σε κάθε ομάδα. Στη συνέχεια, εντοπίζονται τα συχνότερα λάθη του κατηγοριοποιητή και είτε εκτελούνται μεταβάσεις των γλωσσών από μια ομάδα σε μια άλλη είτε προστίθενται νέες ομάδες (κλάσεις) στον κατηγοριοποιητή με

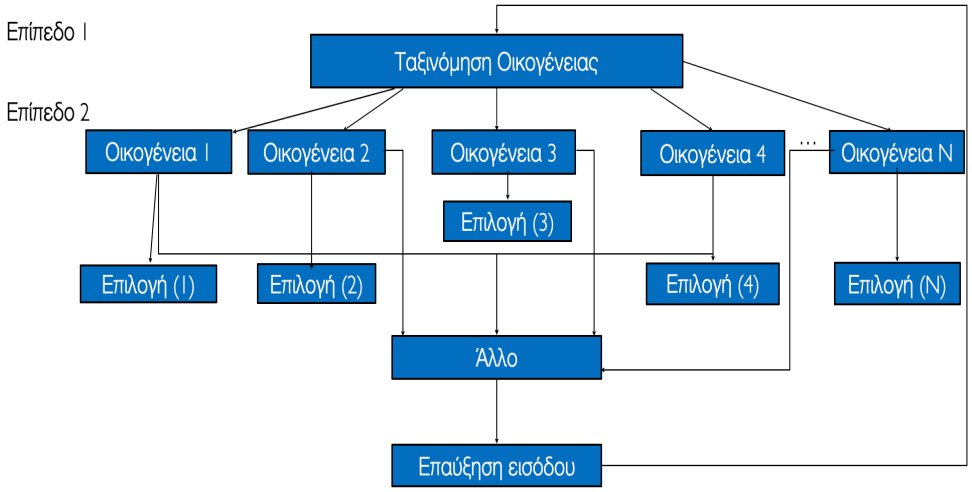
σκοπό την αύξηση της ορθότητας των αποτελεσμάτων. Σημειώνεται ότι η παραπάνω διαδικασία λειτουργεί επαναληπτικά.

3. Μη επιβλεπόμενη μάθηση (Unsupervised Learning):

   Σε αυτήν τη περίπτωση, η διαδικασία ομαδοποίησης των γλωσσών πραγματοποιείται απευθείας από κάποιο μοντέλο μηχανικής μάθησης χρησιμοποιώντας έναν αλγόριθμο βελτιστοποίησης.

Το πλεονέκτημα του διαχωρισμού του προβλήματος σε υπό-προβλήματα έγκειται στη δυνατότητα ανάπτυξης διαφορετικών κατηγοριοποιητών στο δεύτερο επίπεδο. Με αυτόν τον τρόπο, δίνεται η δυνατότητα σε κάθε κατηγοριοποιητή του δευτέρου επιπέδου να επικεντρωθεί στην αναγνώριση λεπτομερών χαρακτηριστικών γλωσσών εντός μιας ομάδας. Συνεπώς, προσφέρεται η δυνατότητα σωστής ταξινόμησης γλωσσών με παρόμοια χαρακτηριστικά, όπου ένας γενικός κατηγοριοποιητής (που πραγματοποιεί απευθείας ταξινόμηση) συνηθίζει να σφάλλει, αποφεύγοντας ταυτόχρονα το πρόβλημα της υπερπροσαρμογής.

Για τον περιορισμό μετάδοσης των σφαλμάτων κατηγοριοποίησης ομάδων γλωσσών στην τελική απόφαση του μοντέλου προτείνεται η ακόλουθη μεθοδολογία: Εάν M είναι το πλήθος των γλωσσών που αντιστοιχούν σε μια συγκεκριμένη ομάδα-οικογένεια, τότε ο ταξινομητής του δευτέρου επιπέδου εκπαιδεύεται να αναγνωρίζει M+1 διαφορετικές κλάσεις. Η πρόσθετη κλάση αντιστοιχεί στην απόφαση "Άλλο", δηλαδή στην αδυναμία ταξινόμησης της ομιλούμενης γλώσσας σε μια συγκεκριμένη κλάση. Σε περίπτωση που ο δεύτερος ταξινομητής αποφασίσει "Άλλο", τότε εφαρμόζονται τεχνικές επαύξησης δεδομένων εισόδου και ανατροφοδότηση στο πρώτο επίπεδο, δηλαδή στον κατηγοριοποιητή ομάδας γλωσσών. Τέλος, σημειώνεται ότι το πλήθος M των γλωσσών μιας ομάδας δεν είναι σταθερό και εξαρτάται από την εκάτοστοτε ομάδα και τον τρόπο προσέγγισης του προβλήματος στο πρώτο επίπεδο. Η παραπάνω μεθοδολογία που περιγράφηκε συνοψίζεται σχηματικά.



Σχήμα 4.2: Αναπαράσταση Προσέγγισης 2 επιπέδων

Κεφάλαιο 5

Το προτεινόμενο σύστημα

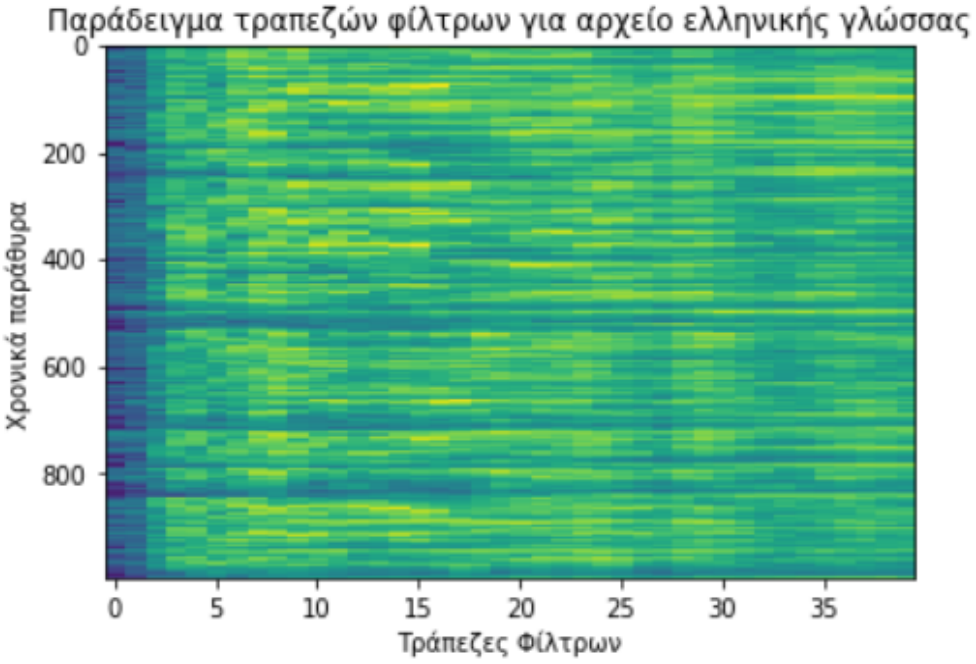
5.1 Σύνολο Δεδομένων

Για την εκπαίδευση και επικύρωση των μοντέλων χρησιμοποιείται το σύνολο δεδομένων "VoxLingua107" (Valk and Alumäe, 2021). Ειδικότερα, το σύνολο δεδομένων αποτελείται από μικρά τμήματα ομιλίας που εξάγονται αυτόματα από βίντεο στην πλατφόρμα του "YouTube" και επισημαίνονται σύμφωνα με τη γλώσσα του τίτλου και την περιγραφή του βίντεο, με ορισμένα βήματα μετά την επεξεργασία για το φιλτράρισμα των ψευδών θετικών στοιχείων. Τα αρχικά δεδομένα περιέχουν 107 γλώσσες. Η συνολική ποσότητα ομιλίας στο σετ εκπαίδευσης είναι 6628 ώρες, ενώ η μέση ποσότητα δεδομένων ανά γλώσσα είναι 62 ώρες. Ωστόσο, το πραγματικό ποσό ανά γλώσσα ποικίλλει πολύ. Σημειώνεται ότι η αυτόματη εξαγωγή δεδομένων ενέχει κινδύνους εγκυρότητας. Υπάρχει ένα ξεχωριστό σετ που περιέχει 1609 τμήματα ομιλίας από 33 γλώσσες, επικυρωμένα από τουλάχιστον δύο εθελοντές ώστε να περιέχουν πραγματικά τη δεδομένη γλώσσα.

Από το αρχικό σύνολο δεδομένων επιλέγονται να χρησιμοποιοηθούν για την ανάπτυξη των παρακάτω μοντέλων 8 γλώσσες: η ελληνική (el), η αγγλική (en), η γερμανική (de), η γαλλική (fr), η ισπανική (es), η ολλανδική (nl), η νορβηγική (no) και η σουηδική (sv). Η αρχική επιλογή γλωσσών ήταν ευρύτερη, μέρος όμως των δεδομένων δεν μπόρεσε να ανακτηθεί λόγω διεφθαρμένων αρχείων.

Επιπρόσθετα, εφαρμόζεται προ-επεξεργασία στα αρχεία ήχου των παραπάνω γλωσσών. Συγκεκριμένα, επιλέγονται αρχεία διάρκειας άνω των 10 δευτερολέπτων και "κόβεται" κάθε αρχείο έτσι ώστε το τελικό σήμα να αποτελείται από ένα τμήμα διάρκεια ακριβώς 10 δευτερόλεπτα. Η περίπτωση χωρισμού του σήματος σε τμήματα μικρότερης διάρκειας και η συνένωση των τμημάτων αυτών δεν ενδείκνυται εξαιτίας της έλλειψης πληροφορίας που δημιουργείται στα σημεία διακοπής, η οποία κατ' επέκταση οδηγεί σε μειωμένη απόδοση των ταξινομητών.

5.2 Μοντέλο Απευθείας Κατηγοριοποίησης



Σχήμα 5.1: Τράπεζες Φίλτρων σε δείγμα της Ελληνικής Γλώσσας

Για το πρόβλημα της απευθείας κατηγοριοποίησης των 8 προαναφερόμενων γλωσσών χρησιμοποιείται αρχιτεκτονική βαθιάς μάθησης. Συγκεκριμένα, αναπτύσσεται ένα ακολουθιακό (sequential) νευρωνικό δίκτυο 7 επιπέδων. Αρχικά, χρησιμοποιούνται 5 συνελικτικά επίπεδα εφαρμόζοντας κατά τη μετάβαση από το ένα επίπεδο στο άλλο Συνάρτηση Ενεργοποίησης Ανορθωμένης Γραμμικής Μονάδας, Μέγιστη Συγκέντρωση και Κανονικοποίηση Δέσμης. Στη συνέχεια τα δεδομένα εισάγονται σε ένα πλήρως συνδεμένο επίπεδο, όπου εφαρμόζεται και πάλι Συνάρτηση Ενεργοποίησης Ανορθωμένης Γραμμικής Μονάδας, Κανονικοποίηση Δέσμης και "Dropout" πριν εισαχθούν στο τελικό επίπεδο, ένα πυκνό επίπεδο (dense layer) 8 νευρώνων. Η Συνάρτηση Ενεργοποίησης του τελικού επιπέδου είναι η "Softmax".

Για την εκπαίδευση και επικύρωση του κατηγοριοποιητή χρησιμοποιούνται 7192 και 800 παραδείγματα αντίστοιχα. Σημειώνεται ότι το σύνολο δεδομένων είναι πλήρως ισορροπημένο, δηλαδή για κάθε κλάση περιέχονται 899 παραδείγματα εκπαίδευσης και 100 παραδείγματα για την επικύρωση.

Τα χαρακτηριστικά που εξάγονται για κάθε δείγμα είναι τράπεζες φίλτρων σε κλίμακα "Mel". Στη συνέχεια τα παραπάνω χαρακτηριστικά τροφοδοτούνται σε μορφή πίνακα στον κατηγοριοποιητή, ο οποίος εκπαιδεύεται για 60 επόχες.

5.3 Ομαδοποίηση με βάση τις Γλωσσολογικές Αναλύσεις

Στο συγκεκριμένο πρόβλημα, αξιοποιώντας τις γλωσσολογικές αναλύσεις, επιχειρούμε την κατηγοριοποίηση ομάδων γλωσσών στις εξής οικογένειες:

(α΄) 1η Οικογένεια: Ελληνικά (el)

(β΄) 2η Οικογένεια: Ισπανικά (es), Γαλλικά (fr)

(γ′) 3η Οικογένεια: Αγγλικά (en), Νορβηγικά (no), Σουηδικά, Γερμανικά (de), Ολλανδικά (nl)

Για την κατηγοριοποίηση χρησιμοποιείται η αρχιτεκτονική βαθιάς μάθησης που περιγράφηκε στην ενότητα 5.2, με μοναδική διαφορά τον αριθμό των νευρώνων στο επίπεδο εξόδου. Στο συγκεκριμένο πρόβλημα, διακρίνονται 3 ομάδες γλωσσών (κλάσεις), επομένως το επίπεδο εξόδου συνίσταται από 3 νευρώνες.

Για την εκπαίδευση και επικύρωση χρησιμοποιούνται 1000 δείγματα για κάθε ομάδα, δηλαδή συνολικά 3000 δείγματα. Το σύνολο εκπαίδευσης συνίσταται από 900 παραδείγματα για κάθε ομάδα (συνολικά 2700) και το σύνολο ελέγχου από 100 (συνολικά 300). Έαν N είναι το πλήθος των γλωσσών που περιέχονται σε μια ομάδα, τότε υπάρχουν σε αυτήν 900/N παραδείγματα εκπαίδευσης και 100/N παραδείγματα επικύρωσης από κάθε γλώσσα. Συνεπώς, διακρίνονται για το σύνολο εκπαίδευσης 900 παραδείγματα ελληνικής, 450 παραδείγματα ισπανικής και 450 γαλλικής, ενώ διατίθενται από 180 παραδείγματα αγγλικής, νορβηγικής, σουηδικής, γερμανικής και ολλανδικής γλώσσας.

Τα χαρακτηριστικά που εξάγονται για κάθε δείγμα είναι και πάλι τράπεζες φίλτρων σε κλίμακα "Mel" και τροφοδοτούνται στον κατηγοριοποιητή. Η εκπαίδευση πραγματοποιείται για 60 εποχές.

5.4 Ομαδοποίηση με βάση την Κατανόηση του Δικτύου

Λαμβάνοντας υπόψη τα λάθη του δικτύου στο πρόβλημα 5.3, επιλέγεται μια νέα κατανομή των γλωσσών σε ομάδες. Ειδικότερα, διακρίνουμε τις εξής οικόγενειες:

(α′) 1η Οικογένεια: Ελληνικά (el), Ισπανικά (es)

(β′) 2η Οικογένεια: Γερμανικά (de), Γαλλικά (fr), Αγγλικά (en), Ολλανδικά (nl)

(γ′) 3η Οικογένεια: Νορβηγικά (no), Σουηδικά

Για ακόμη μια φορά, χρησιμοποιούμε το νευρωνικό δίκτυο που περιγράφηκε στην υποενότητα 5.2 και εφαρμόστηκε επίσης στο πρόβλημα κατηγοριοποίησης της υποενότητας 5.3. Εφόσον, το πλήθος των ομάδων γλωσσών παραμένει ίσο με 3, το πλήθος των νευρώνων στο επίπεδο εξόδου ισούται με 3 και η αρχιτεκτονική του μοντέλου ταυτίζεται πλήρως με την αρχιτεκτονική της υποενότητας 5.3.

Το σύνολο των παρατηρήσεων εκπαίδευσης και επικύρωσης ανέρχεται σε 6000. Η κατανομή είναι ανάλογη με την προηγούμενη, μόνο που στην προκειμένη περίπτωση χρησιμοποιούνται 2000 δείγματα για κάθε ομάδα. Αξιοποιείται το 90% των δειγμάτων για την εκπαίδευση και το υπόλοιπο 10% για την επικύρωση του μοντέλου. Συγκεκριμένα, λαμβάνονται από 900 παραδείγματα εκπαίδεσυσης για την ελληνική, ισπανική, νορβηγική και σουηδική γλώσσα και από 450 για τη γερμανική, γαλλική, αγγλική και ολλανδική γλώσσα. Αντίστοιχα, τα παραδείγματα ελέγχου είναι 100 για καθεμία από τις 4 πρωτοαναφερθέντες γλώσσες και 50 για τις υπόλοιπες, συνθέτοντας ένα τελικό σύνολο επικύρωσης 600 παρατηρήσεων.

Τα χαρακτηριστικά που εξάγονται για κάθε δείγμα είναι τα ίδια με τις παραπάνω υλοποιήσεις και τροφοδοτούνται στον κατηγοριοποιητή. Ο αριθμός των κύκλων εκπαίδευσης (εποχών) είναι και πάλι ίσος με 60.

5.5 Κατηγοριοποίηση εντός των Ομάδων-Οικογενειών

Εφόσον έχουμε προβεί στην κατηγοριοποίηση του πρώτου επίπεδου, σύμφωνα με τη 2η προσέγγιση που περιγράφηκε στο Κεφάλαιο 4, προχωρούμε στην κατηγοριοποίηση των γλωσσών εντός των ομάδων. Η παρούσα εργασία επικεντρώνεται στην κατηγοριοποίηση του 2ου επιπέδου λαμβάνοντας υπόψη μια συγκεκριμένη κατανομή των γλωσσών σε ομάδες. Ύστερα από την ανάπτυξη, την εκπαίδευση, την αξιολόγηση και τη σύγκριση των επιδόσεων (βλ. Κεφάλαιο 6) μεταξύ των μοντέλων των υποενοτήτων 5.2 και 5.3, επιλέγεται η δεύτερη αρχιτεκτονική, δηλαδή η αρχιτεκτονική με βάση την κατανόηση του δικτύου.

Το 2ο επίπεδο κατηγοριόποιησης αποτελείται λοιπόν από 3 διαφορετικούς ταξινομητές. Οι αποφάσεις τους συνοψίζονται παρακάτω:

(α΄) Ταξινομητής 1: "Ελληνικά" ή "Ισπανικά" ή "Άλλο" (3 κλάσεις)

(β΄) Ταξινομητής 2: "Γερμανικά" ή "Γαλλικά" ή "Αγγλικά" ή "Ολλανδικά" ή "Άλλο" (5 κλάσεις)

(γ΄) Ταξινομητής 3: "Νορβηγικά" ή "Σουηδικά" ή "Άλλο" (3 κλάσεις)

Σε περίπτωση που κάποιος ταξινομητής αποφασίσει την κλάση "Άλλο" πραγματοποιείται επάυξηση δεομένων και αντροφοδότηση στο πρώτο επίπεδο κατηγοριοποίησης, δηλαδή στο μοντέλο της υποενότητας 5.3.

Για να υπάρχει κέρδος από την εκτέλεση αυτής της πολυεπίπεδης ταξινόμησης συστήνονται οι ταξινομήτες να συντελούνται από διαφορετικές αρχιτεκτονίκες ή/και να εισάγονται σε αυτούς διαφορετικά χαρακτηριστικά. Ακόμη, όπως περιγραφηκε παραπάνω για την αποφυγή μεταφοράς των λαθών χρησιμοποιείται σε κάθε ταξινομητή η κλάση "Άλλο" και στη συνέχεια πραγματοποιείται επαύξηση δεδομένων. Η επαύξηση αυτή πραγματοποιείται κατά την αξιολόγηση του μοντέλου και αποσκοπεί στη διόρθωση λαθών που προκύπτουν από χαρακτηριστικά των ομιλιτών. Για παράδειγμα, ένας ομιλητής με ιδιαίτερα μπάσα φωνή ενδέχεται να κατηγοριοποιηθεί σε κλάση "Άλλο". Έπειτα, κατά την επαύξηση η μετατροπή της φωνής του σε πιο πρίμα μπορεί να καταστήσει εφικτή την ορθή κατηγοριοποίησή του. Συνεπώς, η εκπαίδευση του μοντέλου πρέπει επίσης να πραγματοποιηθεί τόσο στα αυθεντικά όσο και στα επαυξημένα δεδομένα. Δυστυχώς, οι πόροι του συστήματος που χρησιμοποιήθηκε δεν επαρκούν για τη συγχωνευμένη εκπαίδευση αυθεντικών και επαυξημένων δεδομένων. Για την παραγωγή απτών αποτελεσμάτων χρησιμοποιήθηκε η αρχιτεκτονική που περιγράφηκε παραπάνω, δημιουργώντας 3 ταξινομητές. Σημειώνεται ότι δοκιμάστηκαν επίσης αρχιτεκτονικές Νευρωνικών Δικτύων τύπου "LSTM" και χρήση φασματογραφημάτων

στο διάνυσμα χαρακτηριστικών εισόδου χωρίς κάποιο κέρδος ορθότητας. Τέλος αναφέρεται ότι για την επαύξηση γίνονται με μία πιθανότητα οι εξής αλλαγές:

(α΄) Μικρή αλλαγή του τόνου.

(β΄) Μικρή αλλαγή της ταχύτητας αναπαραγωγής. Σε περίπτωση υπέρβασης των 10 δευτερολέπτων γίνεται "κόψιμο" του σήματος, ενώ σε περίπτωση μικρότερης διάρκειας το σήμα επαναλαμβάνεται.

(γ΄) Εναλλαγή μέρους του σήματος. Τότε το αρχείο χωρίζεται σε δύο μέρη και η σειρά αναραγωγής των ήχων αυτών αντιστρέφεται.

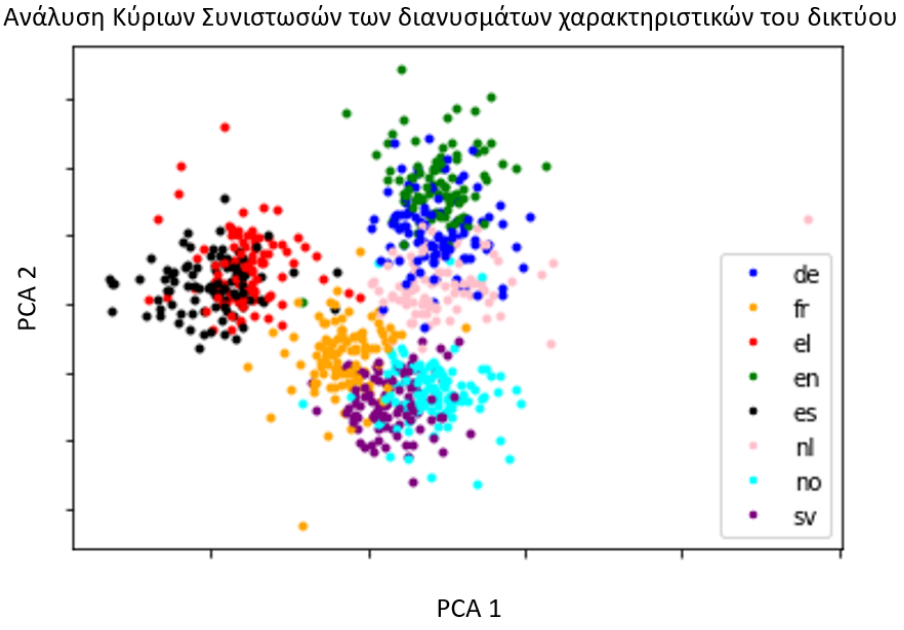
Κεφάλαιο 6

Αξιολόγηση συστήματος

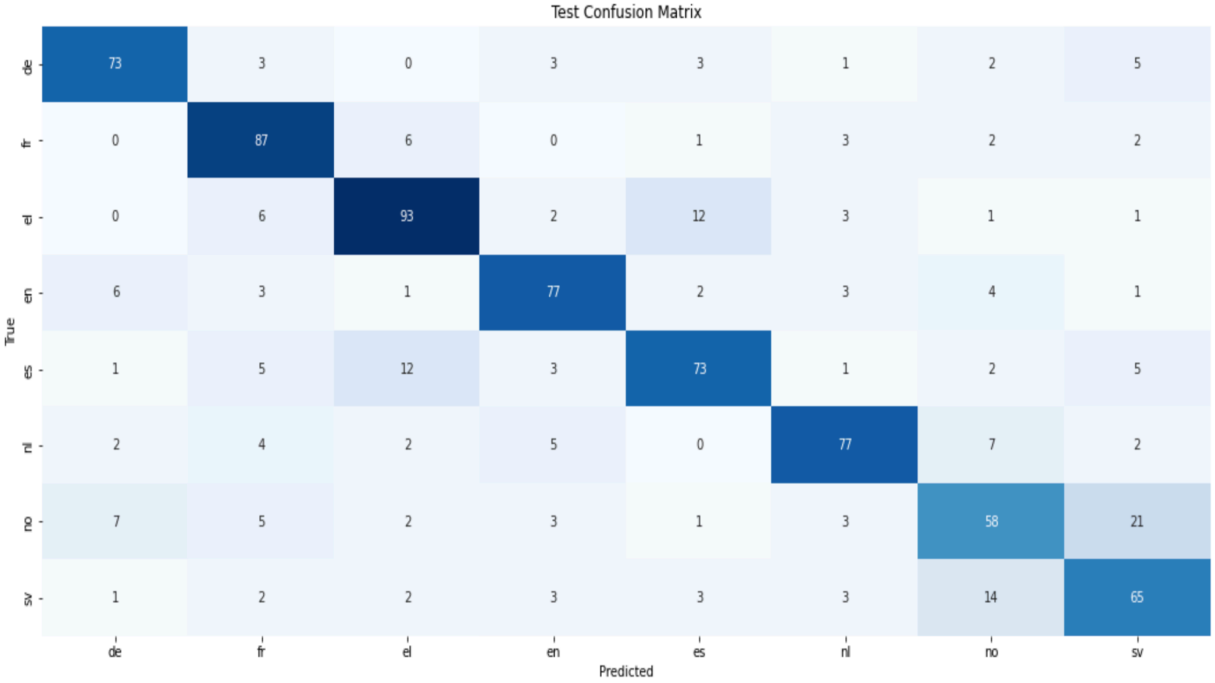
6.1 Αξιολόγηση Μοντέλου Απευθείας Κατηγοριοποίησης

Ο πίνακας σύγχυσης για το συγκεκριμένο πρόβλημα εμφανίζεται παρακάτω. Η ορθότητα του μοντέλου που υλοποιεί την απευθείας κατηγοριοποίηση 8 κλάσεων ανέρχεται στο 75,86%. Συμπληρωματικά, για την καλύτερη οπτικοποίηση των δεδομένων, αφαιρείται το επίπεδο εξόδου από την βασική αρχιτεκτονική και εφαρμόζεται Ανάλυση Κύριων Συνιστωσών στο Σύνολο Επικύρωσης. Με αυτόν τον τρόπο, αποκτάται μια εποπτική εικόνα των χαρακτηριστικών που εξάγει το δίκτυο μέσω της εκπαίδευσής του και εντοπίζονται σημεία στα οποία συγχέονται οι κλάσεις.

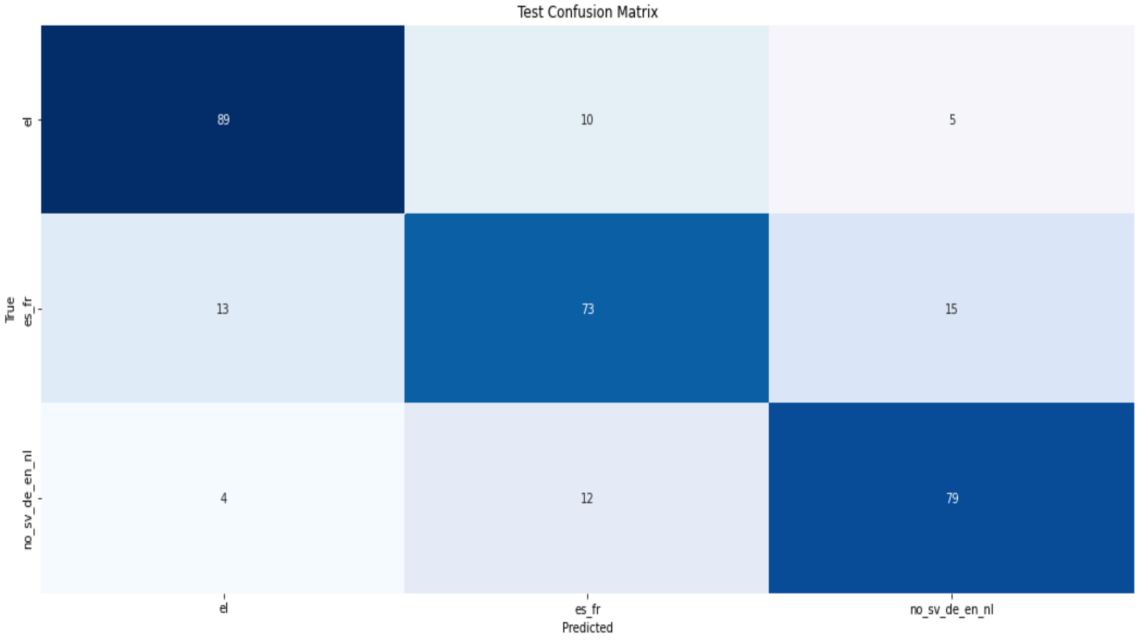
6.2 Αξιολόγηση Ομαδοποίησης με βάση Γλωσσολογικές Αναλύσεις



Σχήμα 6.1: Διδιάστατη Ανάλυση Κύριων Συνιστωσών στο Σύνολο Επικύρωσης



Σχήμα 6.2: Πίνακας Σύγχυσης Απευθείας Κατηγοριοποίησης

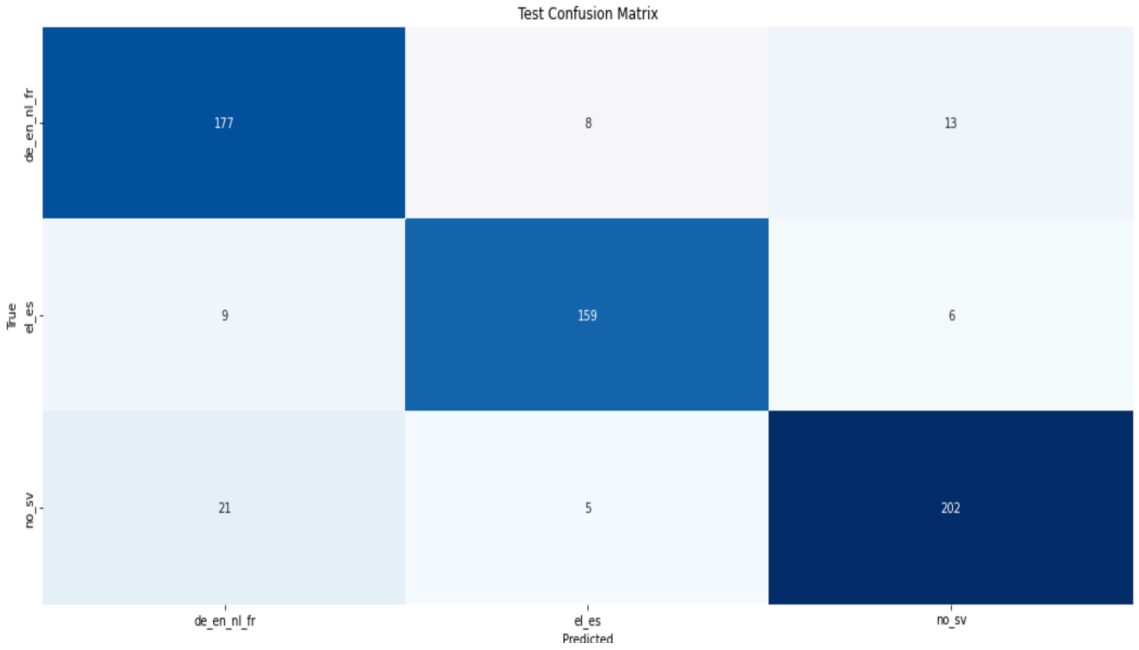


Σχήμα 6.3: Πίνακας Σύγχυσης Ομαδοποίησης βάσει Γλωσσολογίας

Η συγκεκριμένη επιλογή Ομαδοποίησης αποδίδει στο σύνολο επικύρωσης ορθότητα ίση με 80,33%. Οι παρατηρήσεις στον πίνακα σύγχυσης συνέβαλαν στην ανάπτυξη του επόμενου μοντέλου, αλλάζοντας κατάλληλα την κατανομή των γλωσσών στις ομάδες.

6.3 Αξιολόγηση Ομαδοποίησης με βάση την Κατανόηση του Δικτύου

Ο βελτιωμένος αύτος τρόπος Ομαδοποίησης επιτυγχάνει ορθότητα της τάξεως

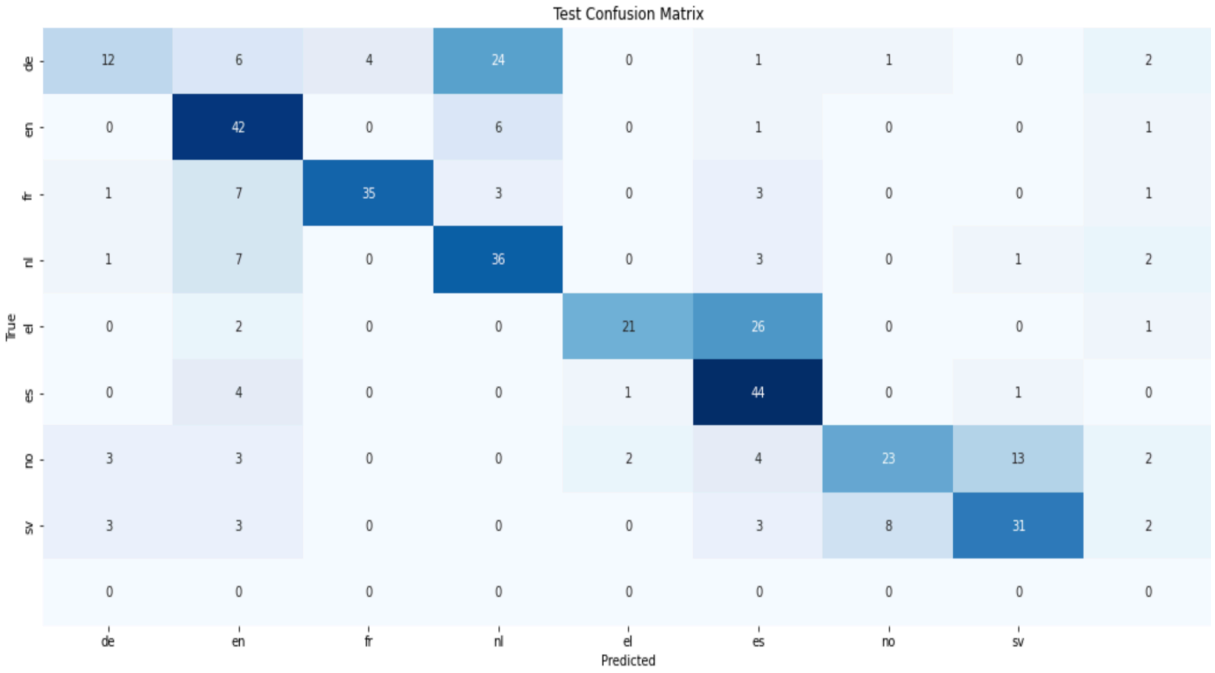


Σχήμα 6.4: Πίνακας Σύγχυσης Ομαδοποίησης βάσει Κατανόησης Δικτύου

του 89,67% σε επίπεδο ομάδων. Συμπεραίνεται λοιπόν, ότι η ομαδοποίηση με βάση τα φωνολογικά χαρακτηριστικά και τον τρόπο που μεταφράζονται από το Δίκτυο αποδίδει κάλυτερα από την ομαδοποίηση με βάση την προέλευση της γλώσσας.

6.4 Αξιολόγηση Κατηγοριοποίησης εντός των Ομάδων-Οικογενειών

Όπως αναφέρθηκε στο Κεφάλαιο 5, η επιθυμήτη υλοποίηση καθίσταται αδύνατη λόγω έλλειψης πόρων. Παρ' όλα αυτά προβαίνουμε κανονικά στην υλοποίηση και αξιολόγηση της ιεραρχικής αρχιτεκτονικής. Η διαφορά έγκειται στην εκπαίδευση χωρίς επαυξημένα δεδομένα. Η χρήση τους σε περίπτωση "Άλλο" υλοποιήθηκε κανονικά. Όπως ήταν αναμενόμενο, η απλοϊκή προσέγγιση εφαρμογής της ίδιας αρχιτεκτονικής και ίδιων διανυσμάτων χαρακτηριστικών στα δύο επίπεδα και σε όλες τις ομάδες οδήγησε σε μειωμένη απόδοση του τελικού αποτελεσμάτος. Ειδικότερα, η ορθότητα άγγιξε το ποσοστό της τάξεως του 61%.



Σχήμα 6.5: Πίνακας Σύγχυσης Ομαδοποίησης Τελικής Κατηγοριοποίησης 2 επιπέδων

Κεφάλαιο 7

Συμπεράσματα

Το πρόβλημα της αναγνώρισης της ομιλούμενης γλώσσας αποτελεί πρόκληση λόγω του υψηλού βαθμού μεταβλητότητας στα σήματα ομιλίας και του μεγάλου αριθμού γλωσσών που πρέπει να ταξινομηθούν. Σε κάθε περίπτωση η κατάλληλη εφαρμογή μεθόδων βαθιάς μάθησης επιτρέπει την αξιοποίηση του μεγάλου όγκου διαθέσιμων ηχητικών δεδομένων στο διαδίκτυο. Επιπλέον, η επισταμένη επιλογή προσέγγισης, αρχιτεκτονικών και διανυσμάτων χαρακτηρτικών καθορίζει σε μεγάλο βαθμό το τελικό αποτέλεσμα του προβλήματος. Ο λόγος που η προτεινόμενη προσέγγιση δεν παρουσίασε βελτιωμένη ορθότητα, οφείλεται στην αδυνάμια εκπόνησης της επιθυμητής εκπαίδευσης λόγω έλλειψης πόρων. Παρόλα αυτά, η μελέτη των φωνολογικών χαρακτηριστικών των ευρωπαϊκών γλωσσών που πραγματοποιήθηκε διευκολύνει το έργο για μελλοντικές υλοποιήσεις σε συστήματα με μεγαλύτερη υπολογιστική ισχύ και αφήνει ελπιδοφόρα σημάδια για τη βελτίωση της τελικής ορθότητας των αποτελεσμάτων.

Βιβλιογραφία

S. Guha, A. Das, P. K. Singh, A. Ahmadian, N. Senu, and R. Sarkar, ``Hybrid feature selection method based on harmony search and naked mole-rat algorithms for spoken language identification from audio signals,'' *IEEE Access*, vol. 8, pp. 182 868--182 887, 2020.

Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, ``Self-training with noisy student improves imagenet classification,'' in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 687--10 698.

J. Valk and T. Alumäe, ``VoxLingua107: a dataset for spoken language recognition,'' in *Proc. IEEE SLT Workshop*, 2021.