



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

Αποκωδικοποίηση Αντίληψης Ομιλίας από Μη-Επεμβατικές Εγγραφές Εγκεφάλου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

της

ΔΕΣΠΟΙΝΑΣ ΠΟΣΤΑΝΤΖΟΓΛΟΥ

Επιβλέπων: Τασουλής Σωτήριος

Λαμία, Φεβρουάριος 2025



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

Αποκωδικοποίηση Αντίληψης Ομιλίας από Μη-Επεμβατικές Εγγραφές Εγκεφάλου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

της

ΔΕΣΠΟΙΝΑΣ ΠΟΣΤΑΝΤΖΟΓΛΟΥ

Επιβλέπων: Τασουλής Σωτήριος

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή, Φεβρουάριος 2025.

(Υπογραφή)

(Υπογραφή)

(Υπογραφή)

.....
Τασουλής Σωτήριος

.....
Πλαγιανάκος Βασίλειος

.....
Κακαρούντας Αθανάσιος

Λαμία, Φεβρουάριος 2025

Περίληψη

Η αποκωδικοποίηση της ομιλίας, μελετώντας την δραστηριότητα του εγκεφάλου παρουσιάζει μεγάλο ενδιαφέρον στον τομέα της υγείας και της νευροεπιστήμης. Οι τεχνολογικές εξελίξεις έχουν αναπτυχθεί ραγδαία και με την βοήθεια της μηχανικής μάθησης πραγματοποιούνται έρευνες, στόχος των οποίων είναι η παραγωγή ομιλίας από την εγκεφαλική δραστηριότητα μέσω επεμβατικών διαδικασιών [1], [2], [3], [4], [5], [6], [7]. Πρόσφατες έρευνες αποκαλύπτουν την ανάγκη αποκωδικοποίησης ομιλίας από εγκεφαλικά σήματα, οι αισθητήρες των οποίων εφαρμόζονται στην επιφάνεια του τριχωτού της κεφαλής, χωρίς να απαιτείται χειρουργική εμφύτευση αυτών [8], [9], [10], [11].

Στόχος της παρούσας έρευνας, είναι η κατασκευή ενός μοντέλου αποκωδικοποίησης ομιλίας από μη-επεμβατικά MEG εγκεφαλικά σήματα, χωρίς να θέτει τα άτομα σε κίνδυνο μέσω της χειρουργικής επέμβασης. Για την επίτευξη του συγκεκριμένου στόχου, χρησιμοποιούνται μέρη του "Whisper" μοντέλου, το οποίο σχεδιάστηκε για την μετατροπή ομιλίας σε κείμενο [12]. Τα μέρη που χρησιμοποιούνται είναι ο κωδικοποιητής και ο αποκωδικοποιητής του μοντέλου που αναφέρεται παραπάνω, τροποποιώντας κατάλληλα την αρχιτεκτονική τους προκειμένου να προσαρμόζονται στα MEG δεδομένα.

Η μελέτη αυτή, οδηγεί ένα βήμα πιο κοντά στην επίτευξη ενός ιδιαίτερα σημαντικού στόχου, την δυνατότητα επικοινωνίας ατόμων με περιορισμένη δυνατότητα αλληλεπίδρασης με το περιβάλλον τους.

Λέξεις Κλειδιά

Μηχανική Μάθηση, MEG Εγκεφαλικά Σήματα, Whisper Μοντέλο, Κωδικοποίηση, Αποκωδικοποίηση

Abstract

Decoding speech by analyzing brain activity is a significant area of interest in health and neuroscience. Technological advancements, particularly in machine learning, have accelerated research aimed at generating speech from human brain activity through invasive procedures [1], [2], [3], [4], [5], [6], [7]. However, recent studies emphasize the need for speech decoding from brain signals using sensors applied to the scalp's surface, eliminating the risks associated with surgical implantation [8], [9], [10], [11].

The objective of this research is to develop a model for decoding speech from non-invasive Magnetoencephalography (MEG) brain signals, thereby avoiding the risks of surgical procedures. To achieve this, components of the Whisper model—originally designed for speech-to-text conversion—are utilized [12]. Specifically, the model's encoder and decoder are adapted by modifying their architecture to suit MEG data.

This study marks a significant step toward an important goal, enabling communication for individuals with limited ability to interact with their environment.

Keywords

Machine Learning, MEG Brain Signals, Whisper Model, Encoding, Decoding

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή κ. Σωτήριο Τασουλή που μέσω των μαθημάτων του και των projects που υλοποίησα κατά την διεξαγωγή αυτών, με ώθησε να ασχοληθώ με τον τομέα της Μηχανικής Μάθησης. Επίσης, ευχαριστώ ιδιαίτερα τόσο τον κ. Σωτήριο Τασουλή όσο και την Δρ. Παρασκευή Τσακμάκη για την καθοδήγηση τους και την εξαιρετική συνεργασία που είχαμε.

Λαμία, Φεβρουάριος 2025

Δέσποινα Ποσταντζόγλου

Περιεχόμενα

Περίληψη	1
Abstract	2
Ευχαριστίες	3
1 Εισαγωγή	6
2 Συναφή Έργα	9
3 Μεθοδολογία	12
3.1 Σύνολο Δεδομένων	12
3.2 Προεπεξεργασία Συνόλου Δεδομένων	15
3.2.1 Ομαδοποίηση λέξεων σε προτάσεις	15
3.2.2 Επεξεργασία των MEG εγκεφαλικών σημάτων	16
3.2.3 Τμηματοποίηση των MEG εγκεφαλικών σημάτων	16
3.2.4 Έλεγχος ακραίων τιμών (outliers)	16
3.2.5 Κανονικοποίηση δεδομένων και Ανίχνευση επιπλέον ανωμαλιών	17
3.2.6 Διαχωρισμός δεδομένων σε δεδομένα εκπαίδευσης, επικύρωσης και ε- λέγχου	25
3.2.7 Εφαρμογή του “Whisper Encoder” για εξαγωγή χαρακτηριστικών από MEG σήματα	25
3.3 Σχεδιασμός μοντέλου	29
3.4 Εκπαίδευση μοντέλου	33
3.5 Αξιολόγηση μοντέλου	34
3.5.1 WER μετρική	35
3.5.2 CER μετρική	36
3.5.3 BLEU-N μετρική	37
3.5.4 ROUGE-N μετρική	41
4 Αποτελέσματα	45
4.1 WER μετρική	45
4.2 CER μετρική	47
4.3 BLEU-N μετρική	47
4.4 ROUGE-N μετρική	48
5 Σύγκριση Υπολογιστικής Πολυπλοκότητας	52



ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις (1), που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί χωρίς να τα περικλείω σε εισαγωγικά και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάστηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.
2. Δέχομαι ότι η αυτολεξεί παράθεση χωρίς εισαγωγικά, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.
3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια.
4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

(1) Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

(Υπογραφή)

.....

Δέσποινα Ποσταντζόγλου

Φεβρουάριος 2025

6 Συζήτηση	54
7 Συμπέρασμα	56
Βιβλιογραφία	59

Κεφάλαιο **1**

Εισαγωγή

Καθημερινά, χιλιάδες άνθρωποι ανά τον κόσμο αντιμετωπίζουν σοβαρές διαταραχές ομιλίας, έπειτα από ένα εγκεφαλικό επεισόδιο ή μία εγκεφαλική νόσο, χάνοντας την ικανότητα τους να επικοινωνούν. Τα τελευταία χρόνια, έχουν πραγματοποιηθεί έρευνες σχετικά με την παραγωγή ομιλίας από την εγκεφαλική δραστηριότητα των ανθρώπων τόσο με επεμβατική όσο και με μη-επεμβατική προσέγγιση. Ωστόσο, οι επεμβατικές καταγραφές εγκεφαλικών σημάτων, αν και προσφέρουν υψηλής ποιότητας δεδομένα, συνοδεύονται από προκλήσεις που δυσκολεύουν τη μακροχρόνια χρήση τους. Η φθορά των εμφυτευμάτων, οι ανοσολογικές αντιδράσεις του οργανισμού και η ανάγκη για αντικατάσταση αποτελούν σημαντικά εμπόδια. Επιπλέον, οι απαιτήσεις για επαναληπτικές επεμβάσεις αυξάνουν τους κινδύνους για τους ασθενείς. Για αυτόν τον λόγο, αρκετές ερευνητικές ομάδες έχουν αρχίσει να μελετούν το αντικείμενο της αποκωδικοποίησης ομιλίας από μη-επεμβατικές καταγραφές εγκεφαλικών σημάτων.

Παραδείγματος χάριν, οι Zhenhailong Wang και Heng Ji [8] ανέπτυξαν ένα σύστημα αποκωδικοποίησης EEG σε κείμενο με ανοιχτό λεξιλόγιο, αξιοποιώντας προεκπαιδευμένα γλωσσικά μοντέλα, όπως το BART [13]. Μάλιστα, ο Yiqun Duan και οι συνεργάτες του [9] βελτίωσαν την παραπάνω προσέγγιση, αποκωδικοποιώντας απευθείας τα ακατέργαστα EEG σήματα, χωρίς την χρήση χρονοσημείων. Ακόμη, ο Alexandre Défossez και η υπόλοιπη ομάδα [10] παρουσίασαν μια μέθοδο αποκωδικοποίησης τμημάτων ομιλίας διάρκειας 3 δευτερολέπτων, συνδέοντας τα MEG εγκεφαλικά σήματα με ακουστικά ερεθίσματα. Η συγκεκριμένη έρευνα εστιάζει στο πώς ο εγκέφαλος αντιλαμβάνεται και κατανοεί την ομιλία.

Μία ενδιαφέρουσα παράλληλη πτυχή της αποκωδικοποίησης ομιλίας από MEG εγκεφαλικά σήματα είναι η απευθείας δημιουργία κειμένου από τα σήματα αυτά. Η Yiqian Yang σε συνεργασία με άλλους ερευνητές [11] κατασκεύασαν το "NeuSpeech" μοντέλο, το οποίο καταφέρνει να μεταφράζει τα MEG εγκεφαλικά σήματα, τα οποία σχετίζονται με την κατανόηση της ομιλίας, απευθείας σε κείμενο. Ειδικότερα, εξετάζεται η συσχέτιση των MEG σημάτων με τη γλωσσική δραστηριότητα, ώστε να παραχθεί κατανοητό γραπτό κείμενο.

Το "NeuSpeech" μοντέλο [11] αποτελεί μια παραλλαγή του μοντέλου "Whisper". Γενικότερα, το "Whisper" μοντέλο είναι ένα μοντέλο αναγνώρισης ομιλίας και κατασκευάστηκε για να προσφέρει υψηλής ακρίβειας μετατροπή ομιλίας σε κείμενο (speech-to-text) με πολυγλωσσικές δυνατότητες. Διαχειρίζεται σύνθετα σενάρια, όπως περιβάλλοντα με θόρυβο και ποικιλία διαλέκτων. Διατίθεται ως εργαλείο ανοιχτού κώδικα, επιτρέποντας την εύκολη πρόσβαση, χρήση και προσαρμογή του από προγραμματιστές και ερευνητές.

Η βάση για τον σχεδιασμό της αρχιτεκτονικής του "Whisper" είναι ένας Transformer κωδικοποιητής - αποκωδικοποιητής (Transformer Encoder-Decoder). Ο κωδικοποιητής αποκτά αναπαραστάσεις της ομιλίας, δηλαδή μία σειρά κύριων χαρακτηριστικών της ομιλίας, ενώ ο αποκωδικοποιητής εξάγει κείμενο λαμβάνοντας τις αναπαραστάσεις της ομιλίας. Αναλυτικότερα, τα αρχεία ήχου υποβάλλονται σε διαδικασία αναδειγματοληψίας (Resampling) με συχνότητα δειγματοληψίας 16kHz, δηλαδή 16000 δείγματα το δευτερόλεπτο, προκειμένου να εξασφαλιστεί η ακριβής και αποτελεσματική αναπαράσταση του ήχου για περαιτέρω ανάλυση. Τα ηχητικά σήματα που συνθέτουν ομιλία μετατρέπονται σε Mel Spectrograms 80 καναλιών, μία μορφή αναπαράστασης του ήχου, η οποία χρησιμοποιείται συχνά σε εφαρμογές αναγνώρισης ομιλίας. Τα Mel Spectrograms βασίζονται στην κλίμακα Mel, η οποία μιμείται τον τρόπο που ο άνθρωπος αντιλαμβάνεται τις συχνότητες. Η χρήση των Mel Spectrograms στο συγκεκριμένο μοντέλο, δίνει μια ακριβή απεικόνιση των χαρακτηριστικών του ήχου, εστιάζοντας στα σημεία που είναι πιο σημαντικά για την κατανόηση της ομιλίας. Κάθε Mel Spectrogram λαμβάνει ηχητικά τμήματα των 30 δευτερολέπτων. Έπειτα, κάθε αναπαράσταση του ήχου με την μορφή που αναφέρθηκε, εισάγεται στο πρώτο μέρος του "Whisper" μοντέλου, τον κωδικοποιητή. Ο κωδικοποιητής περιλαμβάνει την βασική αρχιτεκτονική του "Transformer Encoder", η οποία ξεχωρίζει για την ικανότητα του να επικεντρώνεται ταυτόχρονα σε διαφορετικά σημεία των δεδομένων και θα αναλυθεί σε επόμενη ενότητα. Πριν από αυτόν προηγούνται δύο συνελκτικά στρώματα, τα οποία είναι αυτά που πρώτα λαμβάνουν τα φασματογράμματα. Συμπερασματικά, ο κωδικοποιητής είναι υπεύθυνος για την επεξεργασία του φασματογράμματος και την εξαγωγή κύριων χαρακτηριστικών από αυτόν. Ο αποκωδικοποιητής στη συνέχεια, λαμβάνει τα χαρακτηριστικά των σημάτων και προβλέπει την τελική έξοδο του μοντέλου, το κείμενο που αντιστοιχεί στο ηχητικό αρχείο. Η αρχιτεκτονική του βασίζεται όπως αναφέρθηκε παραπάνω στον "Transformer Decoder", ο οποίος θα αναλυθεί κατά τον σχεδιασμό του μοντέλου που προτείνεται στη παρούσα εργασία.

Διατίθενται έξι μεγέθη του "Whisper" μοντέλου, το καθένα με διαφορετικές απαιτήσεις μνήμης και ταχύτητας, προσφέροντας μια ποικιλία επιλογών για συμβιβασμούς μεταξύ ταχύτητας και ακρίβειας. Ειδικότερα, υπάρχει το μοντέλο "Whisper tiny" με 39 εκατομμύρια παραμέτρους και απαίτηση περίπου 1 GB μνήμης, προσφέροντας ταχύτητα περίπου 10 φορές μεγαλύτερη από το "Whisper large" μοντέλο, το οποίο περιλαμβάνει 1,55 δισεκατομμύρια παραμέτρους και 10 GB μνήμης. Επίσης, υπάρχει το "Whisper base" με 74 εκατομμύρια παραμέτρους, απαίτηση περίπου 1 GB μνήμης και ταχύτητα 7 φορές μεγαλύτερη από το "Whisper large" μοντέλο. Ακόμη, υπάρχει το "Whisper small" με 244 εκατομμύρια παραμέτρους, 2 GB μνήμης και ταχύτητα 4 φορές μεγαλύτερη από το "Whisper large" μοντέλο, το "Whisper medium" με 769 εκατομμύρια παραμέτρους, 5 GB μνήμης και ταχύτητα 2 φορές μεγαλύτερη από το "Whisper large" μοντέλο, καθώς και το "Whisper turbo" με 809 εκατομμύρια παραμέτρους, 6 GB μνήμης και ταχύτητα 8 φορές μεγαλύτερη από το "large" μοντέλο [12].

Αναφορικά με το "NeuSpeech" μοντέλο, εισάγονται τα ακατέργαστα MEG σήματα που καταγράφονται κατά την διάρκεια που οι συμμετέχοντες ακούν μία ιστορία και εξάγονται οι προτάσεις που άκουσαν. Το μοντέλο "Whisper" είναι σχεδιασμένο για να επεξεργάζεται ήχο και δεν είναι άμεσα συμβατό με τα MEG σήματα, τα οποία είναι διαφορετικού τύπου δεδομένα. Προκειμένου το μοντέλο να μπορεί να επεξεργαστεί τα MEG δεδομένα, το πρώτο

στρώμα συνέλιξης (Convolutional Layer) που είναι υπεύθυνο για την επεξεργασία των ηχητικών σημάτων, αντικαθίσταται με δύο νέα στρώματα συνέλιξης. Αυτή η τροποποίηση επιτρέπει στο μοντέλο να αναλύει καλύτερα τα εγκεφαλικά σήματα που δέχεται στην είσοδο του. Σχετικά με την εκπαίδευση του μοντέλου, οι ερευνητές επιλέγουν να εκπαιδεύσουν μόνο τον κωδικοποιητή τους ενώ "παγώνουν" το μέρος του αποκωδικοποιητή. Παράλληλα, εφαρμόζεται η τεχνική AdaLoRa κατά την προσαρμογή του "Whisper Encoder", προκειμένου να επιτευχθεί βελτιστοποίηση των παραμέτρων με αποδοτικό τρόπο. Συγκεκριμένα, στο AdaLoRa μοντέλο ενσωματώνονται μόνο ο κωδικοποιητής του "NeuSpeech" μοντέλου, συμπεριλαμβανομένων των συνελκτικών του επιπέδων. Αντί να πραγματοποιείται πλήρης ενημέρωση όλων των βαρών του μοντέλου κατά την εκπαίδευση, όπως επιτυγχάνεται με την χρήση του βελτιστοποιητή AdamW, η τεχνική LoRa (Low-Rank Adaptation) μειώνει τον αριθμό των παραμέτρων που πρέπει να προσαρμοστούν κατά την εκπαίδευση, εισάγοντας χαμηλής διάστασης (Low-Rank) προβολές των αρχικών βαρών. Αυτή η προσέγγιση οδηγεί σε σημαντική εξοικονόμηση μνήμης και υπολογιστικών πόρων κατά τη διάρκεια της εκπαίδευσης. Η τεχνική AdaLoRa (Adaptive LoRa) που χρησιμοποιείται στο "NeuSpeech" μοντέλο, αποτελεί μία βελτιωμένη εκδοχή της μεθόδου LoRa. Ειδικότερα, περιορίζει τη διαδικασία προσαρμογής στις πιο κρίσιμες παραμέτρους, εφαρμόζοντας μια προσαρμοστική κατανομή των υπολογιστικών πόρων. Αναγνωρίζει δηλαδή τις πιο σημαντικές παραμέτρους του μοντέλου και κατανέμει τους διαθέσιμους πόρους εκπαίδευσης σε αυτές, μειώνοντας έτσι τον υπολογιστικό φόρτο και διασφαλίζοντας παράλληλα την υψηλή απόδοση του μοντέλου [14].

Η παρούσα πτυχιακή εργασία στοχεύει στην αναπαραγωγή του πειράματος που περιγράφεται στην συγκεκριμένη έρευνα [11] και την βελτιστοποίηση της υπολογιστικής πολυπλοκότητας ενός τέτοιου σύνθετου μοντέλου. Η στρατηγική που ακολουθήθηκε περιλαμβάνει την ίδια ακριβώς διαχείριση της προεπεξεργασίας των δεδομένων, όπως στο πρωτότυπο άρθρο, με σημαντική διαφοροποίηση στην επιλογή του μοντέλου. Συγκεκριμένα, στη προτεινόμενη προσέγγιση επιλέγεται να εκπαιδευτεί ο αποκωδικοποιητής, με λιγότερα επίπεδα Transformer Decoder (από τέσσερα σε ένα) για λόγους μείωσης του υπολογιστικού κόστους. Αντίθετα, δεν εκπαιδεύεται ο κωδικοποιητής όπως στο "NeuSpeech", όμως αξιοποιείται για την εξαγωγή υψηλής ανάλυσης χαρακτηριστικών των εγκεφαλικών σημάτων και την εισαγωγή αυτών στον αποκωδικοποιητή. Όσον αφορά την αρχιτεκτονική του κωδικοποιητή, διατηρούνται οι προεπιλεγμένες επιλογές των παραμέτρων του "Whisper Encoder" χωρίς καμία τροποποίηση.

Κεφάλαιο 2

Συναφή Έργα

Τα τελευταία χρόνια, σημειώνονται αρκετές έρευνες σχετικές με την αποκωδικοποίηση των εγκεφαλικών σημάτων σε κείμενο, καθώς η πρόοδος στην νευροεπιστήμη, την μηχανική μάθηση και τις τεχνολογίες επεξεργασίας σήματος επιτρέπει την βελτίωση της ακρίβειας και της αποτελεσματικότητας τέτοιων συστημάτων.

Ειδικότερα, το 2019 οι ερευνητές Gopala Anumanchipalli, Josh Chartier και Edward Chang [1], ανέπτυξαν ένα μοντέλο ικανό να μεταφράζει ECoG σήματα (δηλαδή σήματα ενδοκρανιακής ηλεκτροεγκεφαλογραφίας τα οποία για να καταγραφούν απαιτείται κρανιοτομία για την εμφύτευση των ηλεκτροδίων) στις λεγόμενες κινηματικές αναπαραστάσεις της άρθρωσης. Οι κινηματικές αναπαραστάσεις της άρθρωσης αναφέρονται στην λεπτομερή αναπαράσταση των κινήσεων διάφορων μερών του φωνητικού συστήματος (όπως γλώσσα, χείλη, λάρυγγας, ουρανίσκος), τα οποία συνεργάζονται για να διαμορφώσουν τους ήχους της ανθρώπινης ομιλίας. Έπειτα, αυτές οι αναπαραστάσεις μετατρέπονται σε ηχητικά χαρακτηριστικά με την μορφή MFCCs (Mel-Frequency Cepstral Coefficients), τα οποία αναπαριστούν ηχητικά σήματα σε συχνότητες που τις αντιλαμβάνεται ο άνθρωπος. Τελικά, από το μοντέλο προκύπτει το ηχητικό κύμα της ομιλίας. Η προσέγγιση αυτή διατίθεται ελεύθερα στην ερευνητική κοινότητα για περαιτέρω εμφάθυνση του συγκεκριμένου ζητήματος.

Επιπλέον, το 2020 ο Ran Wang και οι συνεργάτες του [2], χρησιμοποίησαν ένα προεκπαιδευμένο μοντέλο GAN (Generative Adversarial Network) με απώτερο σκοπό την αποκωδικοποίηση ECoG σημάτων σε ομιλία, επιτυγχάνοντας υψηλή ακρίβεια και αποκαλύπτοντας τις δυναμικές αλληλεπιδράσεις μεταξύ διαφορετικών περιοχών του εγκεφάλου κατά την διάρκεια γνωστικών διεργασιών. Αρχικά, προεκπαιδεύεται ο γεννήτορας του GAN μοντέλου με δεδομένα ηχογραφημένων ομιλιών, έτσι ώστε να μάθει να παράγει φασματογράμματα, τα οποία αναπαριστούν κάθε ήχο της ομιλίας. Στη συνέχεια, ο αποκωδικοποιητής του δικτύου αναλαμβάνει να μετατρέψει τα σήματα του εγκεφάλου σε αναπαραστάσεις αυτού του τύπου, δηλαδή σε φασματογράμματα. Ο αποκωδικοποιητής επομένως, λειτουργεί ως μεταφραστής, μετατρέποντας τα νευρωνικά σήματα στον χώρο αναπαράστασης που έχει μάθει ο γεννήτορας. Αυτή η διαδικασία επιτρέπει την ανακατασκευή της ομιλίας από τα εγκεφαλικά σήματα, κάτι που είναι ιδιαίτερα χρήσιμο για άτομα με δυσκολίες στην ομιλία, αλλά με υγιή εγκεφαλική δραστηριότητα.

Το 2021, ο Francis R. Willett και οι συνεργάτες του [3] παρουσίασαν μια σημαντική πρόοδο στον τομέα της επικοινωνίας μέσω διεπαφών εγκεφάλου-υπολογιστή (Brain-Computer Interfaces, BCIs). Η ερευνητική ομάδα ανέπτυξε ένα σύστημα υψηλής απόδοσης

για την αποκωδικοποίηση εγκεφαλικών σημάτων που σχετίζονται με τη γραφή. Χρησιμοποιώντας δεδομένα από τον κινητικό φλοιό ενός ασθενούς με κάκωση νωτιαίου μυελού, κατά τη διάρκεια 10 ωρών συνεδριών γραφής, κατάφεραν να αποκωδικοποιήσουν 90 χαρακτήρες ανά λεπτό, με ακρίβεια που έφτανε το 94% (δηλαδή 15-18 λέξεις ανά λεπτό). Παράλληλα, ο David A. Moses και οι συνεργάτες του [4] αποκωδικοποίησαν 15,2 λέξεις ανά λεπτό από τα εγκεφαλικά σήματα ενός ασθενούς με αναρθρία, δηλαδή με πλήρη αδυναμία ομιλίας λόγω παράλυσης των μυών της ομιλίας. Χρησιμοποιώντας μια διεπαφή εγκεφάλου-υπολογιστή (BCI), εμφυτευμένη στον αισθητικο-κινητικό φλοιό του εγκεφάλου, μπόρεσαν να “μεταφράσουν” την εγκεφαλική δραστηριότητα σε λόγο με ρυθμό 15,2 λέξεων ανά λεπτό και ακρίβεια 74,4%. Η αποκωδικοποίηση βασίστηκε σε ένα λεξιλόγιο 50 λέξεων, καταγράφοντας δεδομένα από 48 συνεδρίες μέσα σε 22 ώρες.

Ακόμη, η έρευνα του Sean L. Metzger και των συνεργατών του [5] το 2022, επικεντρώθηκε σε έναν ασθενή με παράλυση των άκρων και του φωνητικού συστήματος, όπου ανέπτυξαν και εφάρμοσαν ένα σύστημα ορθογραφίας. Το σύστημα αυτό βασίζεται σε έναν κωδικό λέξεων. Κάθε γράμμα της αγγλικής αλφαβήτου αντιστοιχούσε σε μια συγκεκριμένη λέξη (π.χ. “άλφα” για το γράμμα “a”). Αυτή η προσέγγιση οδήγησε σε ποσοστό σφάλματος χαρακτήρων 6,13% και ταχύτητα 29,4 χαρακτήρων ανά λεπτό, προσφέροντας έναν δίαυλο επικοινωνίας σε ασθενείς με σοβαρή κινητική και φωνητική αναπηρία. Έναν χρόνο μετά, το 2023 ο Sean L. Metzger και άλλα μέλη εστίασαν στη δημιουργία ενός συστήματος που αποκωδικοποιεί σε πραγματικό χρόνο ECoG νευρικά σήματα. Η μελέτη ανέπτυξε και εκπαιδευσε ένα RNN μοντέλο που αναλύει δεδομένα εγκεφαλικής δραστηριότητας, καταγράφοντας προσπάθειες του συμμετέχοντα να μιλήσει σιωπηλά. Το σύστημα κατάφερε την ακριβή αποκωδικοποίηση σε τρεις διαφορετικές μορφές εξόδου: κείμενο, ηχητική ομιλία και κινούμενη εικόνα προσώπου (avatar), προσφέροντας μια πολυτροπική προσέγγιση που επιτρέπει την αποκατάσταση πλήρους επικοινωνίας για άτομα με παράλυση.

Την ίδια χρονιά (το 2023), ο Yan Liu και η υπόλοιπη ομάδα [7] επικεντρώθηκαν στην ανάπτυξη ενός μοντέλου που χρησιμοποιεί μια διαδικασία τριών σταδίων για την αποκωδικοποίηση λογόσυλλαβων γλωσσών, όπως τα Κινέζικα, από εγκεφαλικά σήματα που καταγράφονται μέσω του ECoG (Ηλεκτροκορτικογράφημα). Σε πρώτο στάδιο, τα εγκεφαλικά σήματα αναλύονται και αποκωδικοποιούνται με βάση την προφορά της Κινέζικης γλώσσας, η οποία περιλαμβάνει την προφορά των λέξεων και τις συλλαβές της γλώσσας. Έπειτα σε δεύτερο στάδιο, αναλύονται οι τόνοι και οι συλλαβές για την κατανόηση της σημασίας των λέξεων. Σε τρίτο στάδιο, τα δεδομένα συνδυάζονται και χρησιμοποιούνται για τη σύνθεση ομιλίας, αποδίδοντας τις λέξεις και τους τόνους με τρόπο που μπορεί να κατανοηθεί από τον ακροατή.

Όσον αφορά έρευνες με μη-επεμβατικές προσεγγίσεις, όπως αναφέρθηκε στην εισαγωγή, οι Zhenhailong Wang και Heng Ji [8], ο Yiqun Duan και οι συνεργάτες του [9], ο Alexandre Défossez και η υπόλοιπη ομάδα [10] καθώς και η Yiqian Yang σε συνεργασία με άλλους ερευνητές [11] εξήγαγαν ενδιαφέροντα αποτελέσματα για την ανάλυση και αποκωδικοποίηση ομιλίας από μη-επεμβατικές καταγραφές εγκεφαλικών σημάτων. Στις συγκεκριμένες μελέτες, χρησιμοποιήθηκαν εγκεφαλικά σήματα τα οποία καταγράφηκαν μέσω της μέθοδου EEG (Ηλεκτροεγκεφαλογραφία) και της MEG (Μαγνητοεγκεφαλογραφία). Οι έρευνες αυτές υπογραμμίζουν τη δυναμική των μη-επεμβατικών τεχνικών στην αποκωδικοποίηση

της ομιλίας, προσφέροντας πολύτιμες γνώσεις για τη νευρωνική επεξεργασία της γλώσσας. Παρόλο που η ακρίβεια και η γενίκευση των υπάρχοντων μοντέλων παραμένουν ανοιχτές προκλήσεις, η πρόοδος στην καταγραφή και ανάλυση εγκεφαλικών σημάτων σε συνδυασμό με ισχυρά γλωσσικά μοντέλα ενισχύει τις προοπτικές για την διευκόλυνση της επικοινωνίας ατόμων με επικοινωνιακές ή κινητικές δυσκολίες.

Κεφάλαιο **3**

Μεθοδολογία

Η συγκεκριμένη ενότητα, περιλαμβάνει διαδοχικά όλα τα βήματα που οδηγούν στον τελικό στόχο της έρευνας, την μετάφραση MEG σημάτων του εγκεφάλου σε κείμενο. Συγκεκριμένα, πραγματοποιείται αναφορά στο σύνολο δεδομένων που χρησιμοποιήθηκε για το πείραμα, ανάλυση της προεπεξεργασίας των δεδομένων καθώς και της αρχιτεκτονικής του μοντέλου.

3.1 Σύνολο Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιείται είναι το MEG-MASC [15], το οποίο περιέχει ένα σύνολο εγκεφαλικών σημάτων είκοσι επτά αγγλόφωνων συμμετεχόντων ενώ ακούν τέσσερις φανταστικές ιστορίες από το Manually Annotated Sub – Corpus (MASC) μαζί με τυχαίες λίστες λέξεων και ερωτήσεις κατανόησης. Πραγματοποιούνται δύο πανομοιότυπες συνεδρίες διάρκειας μίας ώρας. Το διάστημα μεταξύ των δύο συνεδριών καταγραφής των εγκεφαλικών σημάτων των συμμετεχόντων κατά την διάρκεια των ηχητικών ερεθισμάτων, είναι από τουλάχιστον μία ημέρα έως το πολύ δύο μήνες, ανάλογα με τη διαθεσιμότητα των συμμετεχόντων και των ερευνητών.

Κάθε συμμετέχων ακούει με διαφορετική σειρά τις τέσσερις αυτές ιστορίες, οι οποίες είναι οι εξής:

- **"Iw1"** - Η περιπέτεια ενός εξωγήινου διαστημόπλοιου στην αναζήτησή του να επιστρέψει στον πλανήτη του.
- **"cable_spool_fort"** - Δύο αδέρφια που διασκεδάζουν, παίζοντας μέσα στο δάσος.
- **"easy_money"** - Δύο φίλοι που χρησιμοποιούν ένα μαγικό τέχνασμα για να κερδίσουν χρήματα.
- **"The_black_widow"** - Προκλήσεις που συναντά ένας συγγραφέας κατά τη διάρκεια της συγγραφικής του διαδικασίας.

Προκειμένου να μειωθεί η συσχέτιση μεταξύ γλωσσικών χαρακτηριστικών και ακουστικών αναπαραστάσεων, χρησιμοποιούνται τρεις διαφορετικές φωνές ("Ava", "Samantha" και "Allison"), με διαφορετικό ρυθμό ομιλίας κάθε 5-20 προτάσεις που κυμαίνεται από 145 έως 205 λέξεις ανά λεπτό.

Η μελέτη για την συλλογή των δεδομένων, εγκρίθηκε από την Επιτροπή Ηθικής της Institutional Review Board (IRB) του New York University Abu Dhabi, όπως ανέφεραν η Laura Gwilliams και οι συνεργάτες της που σύλλεξαν τα δεδομένα [16].

Από τους συνολικά είκοσι επτά συμμετέχοντες, οι δεκαπέντε είναι γυναίκες και οι δώδεκα είναι άνδρες. Ο μέσος όρος ηλικίας των συμμετεχόντων είναι 24,8 χρόνια, ενώ η τυπική απόκλιση της ηλικίας, δηλαδή πόσο διαφέρει η ηλικία των συμμετεχόντων από τον μέσο όρο, είναι 6,4 χρόνια. Όλοι έχουν φυσιολογική ακοή και κανένα ιστορικό νευρολογικών διαταραχών. Είναι δεξιόχειρες, όπως αξιολογήθηκε μέσω του ερωτηματολογίου Edinburgh Handedness Inventory και μητρική τους γλώσσα είναι η Αγγλική. Κατ' εξαίρεση, ο sub-21 έχει ως μητρική γλώσσα τα Ινδικά, ωστόσο σε ηλικία 10 ετών έμαθε την Αγγλική γλώσσα. Ακόμη, πέντε συμμετέχοντες (sub-3, sub-12, sub-16, sub-20 και sub-21) δεν ολοκλήρωσαν και τις δύο συνεδρίες.

Η μέθοδος που χρησιμοποιήθηκε για την καταγραφή των εγκεφαλικών σημάτων των συμμετεχόντων είναι η μαγνητοεγκεφαλογραφία (MEG). Η συγκεκριμένη τεχνική χαρτογραφεί την δραστηριότητα του εγκεφάλου των συμμετεχόντων μέσω της καταγραφής των μαγνητικών πεδίων που δημιουργούνται από τα φυσικά ηλεκτρικά ρεύματα στον εγκέφαλο, χρησιμοποιώντας εξαιρετικά ευαίσθητους αισθητήρες μαγνητικών πεδίων [17]. Στη συγκεκριμένη περίπτωση, χρησιμοποιούνται 208 αισθητήρες (axial gradiometer), επιφανειακά κατανεμημένοι στο κεφάλι των συμμετεχόντων. Δηλαδή, από κάθε αισθητήρα/κανάλι λαμβάνεται η εγκεφαλική δραστηριότητα του εκάστοτε συμμετέχοντα. Αυτό το MEG σύστημα κατασκευάστηκε από το Kanazawa Institute of Technology (KIT) στην Ιαπωνία. Κάθε καταγραφή εγκεφαλικής δραστηριότητας πραγματοποιήθηκε με συχνότητα δειγματοληψίας 1000 Hz. [16]

Όλοι οι συμμετέχοντες κατέχουν ένα μοναδικό id της μορφής sub-*, όπου "*" είναι η σειρά κατάταξης του συμμετέχοντα κατά την πειραματική διαδικασία (δηλαδή sub-01, sub-02, sub-03, ..., sub-27). Όλα τα εγκεφαλικά σήματα των συμμετεχόντων που καταγράφηκαν σε κάθε συνεδρία και για κάθε ιστορία είναι αποθηκευμένα σε ανάλογα αρχεία με κατάληξη .con.

Εκτός από τα αρχεία που περιλαμβάνουν τα εγκεφαλικά σήματα των συμμετεχόντων καθώς και πληροφορίες για τους ίδιους τους συμμετέχοντες, το σύνολο δεδομένων περιλαμβάνει για όλους τους συμμετέχοντες σε κάθε συνεδρία και για όλες τις ιστορίες επιπλέον αρχεία ("events.tsv"), τα οποία περιλαμβάνουν τον χρονισμό κάθε φωνήματος και κάθε λέξης. Μάλιστα, κάθε λέξη εντάσσεται σε μία συγκεκριμένη πρόταση. Ο χρονισμός αυτός συνδέει την MEG εγγραφή του εκάστοτε συμμετέχοντα με τις αντίστοιχες στιγμές ομιλίας μίας συγκεκριμένης ιστορίας.

Στην εικόνα 3.1, απεικονίζεται ένα στιγμιότυπο του "events.tsv" αρχείου, το οποίο περιέχει κάθε είδους ερέθισμα που λαμβάνει ο συμμετέχοντας με id "sub-01", στην ιστορία "cable_spool_fort" της πρώτης συνεδρίας. Ειδικότερα, το πρώτο ερέθισμα που δέχεται ο συμμετέχοντας είναι ένας ήχος που διαρκεί στιγμιαία (duration: 0,0 sec). Το αμέσως επόμενο ακουστικό ερέθισμα είναι το φώνημα "dh_B", το οποίο αντιστοιχεί στο πρώτο φώνημα της λέξης "The". Η λέξη αυτή είναι η πρώτη λέξη (word_index: 0,0) της πρώτης πρότασης που ακούγεται (sequence_id: 0,0). Εμφανίζεται στα 0,0 δευτερόλεπτα (start: 0,0 sec), δηλαδή αμέσως μετά την έναρξη του ηχητικού ερεθίσματος. Η χρονική στιγμή που εμφανίζεται το

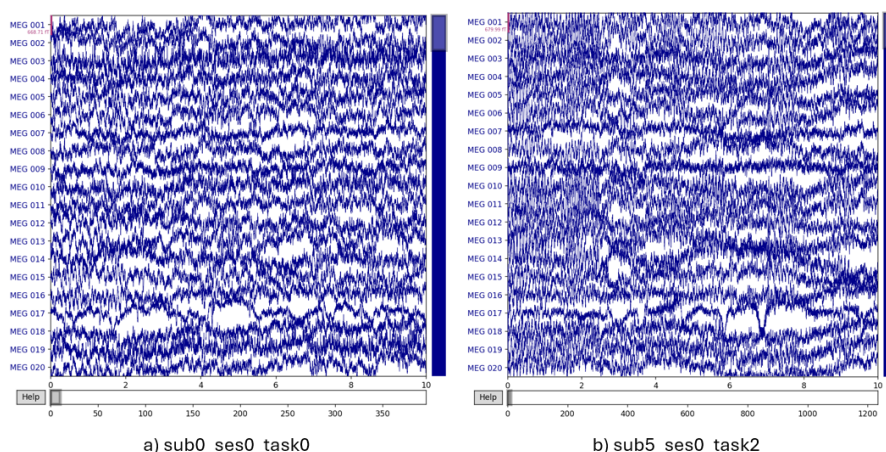
ερέθισμα είναι onset: 8,245 δευτερόλεπτα μετά την έναρξη της καταγραφής της εγκεφαλικής δραστηριότητας και διαρκεί για 0,05 δευτερόλεπτα (duration: 0,05 sec). Το ερέθισμα αυτό αντιστοιχεί στο sample 8245, δηλαδή το χρονικό σημείο της εγκεφαλικής δραστηριότητας στο οποίο εντοπίστηκε το ερέθισμα. Το επόμενο ερέθισμα είναι η ίδια η λέξη "The". Εμφανίζεται στα 0,0 δευτερόλεπτα (start: 0,0 sec), δηλαδή αμέσως μετά την έναρξη του ηχητικού ερεθίσματος. Η χρονική στιγμή που εμφανίζεται το ερέθισμα είναι onset: 8,245 δευτερόλεπτα μετά την έναρξη της καταγραφής της εγκεφαλικής δραστηριότητας και διαρκεί για 0,13 δευτερόλεπτα (duration: 0,13 sec). Το ερέθισμα αυτό αντιστοιχεί στο sample 8245 και όπως αναφέρθηκε παραπάνω, η λέξη αυτή είναι η πρώτη λέξη (word_index: 0,0) της πρώτης πρότασης που ακούγεται (sequence_id: 0,0). Το τελευταίο ερέθισμα που αφορά την συγκεκριμένη λέξη είναι το φώνημα "ah_E". Σύμφωνα με τα μεταδεδομένα του, ανήκει στη πρώτη λέξη ("word_index": 0,0) της πρώτης πρότασης που ακούγεται (sequence_id: 0,0). Εμφανίζεται στα 0,05 δευτερόλεπτα (start: 0,05 sec), δηλαδή 0,05 δευτερόλεπτα μετά την αρχή της ακρόασης. Η χρονική στιγμή που εμφανίζεται το ερέθισμα είναι onset: 8,295 δευτερόλεπτα μετά την έναρξη της καταγραφής της εγκεφαλικής δραστηριότητας και διαρκεί για 0,08 δευτερόλεπτα (duration: 0,08 sec). Το ερέθισμα αυτό αντιστοιχεί στο sample 8295. Αντίστοιχα γεγονότα υπάρχουν και στο υπόλοιπο αρχείο.

	onset (sec)	duration (sec)	start (sec)	kind	phoneme	word	sample
event 1	8,245	0,0	0,0	sound	-	-	8245
event 2	8,245	0,05	0,0	phoneme	dh_B	-	8245
event 3	8,245	0,13	0,0	word	-	The	8245
event 4	8,295	0,08	0,05	phoneme	ah_e	-	8295

Εικόνα 3.1: Στιγμιότυπο χρονισμού των φωνημάτων που συνθέτουν την λέξη "The" καθώς και της ίδιας της λέξης, στην πρώτη συνεδρία της ιστορίας "cable_spool_fort" για τον "sub-01".

Για την καλύτερη κατανόηση του συνόλου δεδομένων, απεικονίζονται παρακάτω τα ακατέργαστα (raw) εγκεφαλικά σήματα που λαμβάνονται από τα 208 MEG κανάλια σε συνάρτηση με τον χρόνο, κατά το άκουσμα μίας ιστορίας. Παρατηρούμε ότι κάθε MEG κανάλι εμφανίζει σήματα σε μορφή κυματομορφής που αντικατοπτρίζει την εγκεφαλική δραστηριότητα του ατόμου κάθε χρονική στιγμή. Κάθε MEG κανάλι, που σχετίζεται με μία συγκεκριμένη περιοχή του εγκεφάλου, έχει διαφορετική κυματομορφή, δεδομένου ότι κατά το άκουσμα μίας ιστορίας μπορεί να ενεργοποιούνται διαφορετικές περιοχές του εγκεφάλου, η κάθε μία από τις οποίες παρουσιάζει διαφορετική δραστηριότητα. Συγκεκριμένα, η παρακάτω εικόνα 3.2 απεικονίζει την εγκεφαλική δραστηριότητα του συμμετέχοντα με id "sub-01" απ' όλα τα κανάλια, καθώς ακούει την ιστορία "Iw1" κατά την πρώτη συνεδρία, καθώς και την εγκεφαλική δραστηριότητα του συμμετέχοντα με id "sub-05" απ' όλα τα κανάλια, καθώς ακούει την ιστορία "easy_money" πάλι κατά την πρώτη συνεδρία. Αντίστοιχες αναπαραστάσεις προκύπτουν και για τους υπόλοιπους συμμετέχοντες σε όλες τις συνεδρίες και για όλες τις ιστορίες.

Οι απεικονίσεις των ακατέργαστων εγκεφαλικών σημάτων που λαμβάνονται από τα 208



Εικόνα 3.2: α) Απεικόνιση εγκεφαλικής δραστηριότητας μέσω MEG καναλιών για τον συμμετέχοντα με id "sub-01" καθώς ακούει την ιστορία "lw1" κατά την πρώτη συνεδρία. β) Απεικόνιση εγκεφαλικής δραστηριότητας μέσω MEG καναλιών για τον συμμετέχοντα με id "sub-05" καθώς ακούει την ιστορία "easy_monkey" πάλι κατά την πρώτη συνεδρία.

MEG κανάλια, αποκαλύπτουν σημαντικές πτυχές της εγκεφαλικής δραστηριότητας κατά την ακρόαση μιας ιστορίας. Οι κυματομορφές που καταγράφονται διαφέρουν μεταξύ των καναλιών, αντανakλώντας τη μοναδική δραστηριότητα διάφορων περιοχών του εγκεφάλου και υποδεικνύοντας ότι διαφορετικά εγκεφαλικά δίκτυα ενεργοποιούνται καθώς εξελίσσεται η αφήγηση της ιστορίας. Επιπλέον, η σύγκριση μεταξύ συμμετεχόντων, όπως των "sub-01" και "sub-05" στη συγκεκριμένη περίπτωση, δείχνει ότι η εγκεφαλική απόκριση ποικίλει τόσο μεταξύ ατόμων όσο και μεταξύ ιστοριών. Οι διαφορετικές αποκρίσεις μπορεί να σχετίζονται με το πόσο συγκεντρωμένος είναι κάθε συμμετέχων, τις προηγούμενες εμπειρίες του που μπορεί να επηρεάζουν την κατανόηση ή τη συναισθηματική του απόκριση, καθώς και με το πόσο συναισθηματικά εμπλέκεται στην αφήγηση. Καθώς η ιστορία εκτυλίσσεται, παρατηρούνται στιγμές έντονης δραστηριότητας, πιθανώς αντιπροσωπεύοντας κρίσιμα αφηγηματικά σημεία ή συναισθηματικές κορυφώσεις. Συνολικά, οι απεικονίσεις προσφέρουν πολύτιμες πληροφορίες για τον τρόπο με τον οποίο ο εγκέφαλος επεξεργάζεται την ακουστική πληροφορία και αποκαλύπτουν τη δυναμική συνεργασία διαφορετικών περιοχών του για τη δημιουργία μιας ολοκληρωμένης εμπειρίας ακρόασης.

3.2 Προεπεξεργασία Συνόλου Δεδομένων

Η συγκεκριμένη υποενότητα περιλαμβάνει την εξαγωγή των λέξεων από τα αρχεία "events.tsv" και την ομαδοποίηση τους σε προτάσεις, την επεξεργασία των MEG εγκεφαλικών σημάτων και την δημιουργία MEG τμημάτων από τα αρχικά MEG σήματα, τον διαχωρισμό των δεδομένων σε δεδομένα εκπαίδευσης (train set), επικύρωσης (validation set) και ελέγχου (test set) καθώς και την εξαγωγή χαρακτηριστικών από τα σήματα.

3.2.1 Ομαδοποίηση λέξεων σε προτάσεις

Δεδομένου ότι το μοντέλο εκπαίδευσης έχει ως στόχο την μετάφραση των εγκεφαλικών σημάτων σε προτάσεις, καθίσταται αναγκαία η δημιουργία προτάσεων. Έχοντας τα αρχε-

ία "events.tsv" για κάθε συμμετέχοντα, σε κάθε συνεδρία και για κάθε ιστορία, τα οποία περιλαμβάνουν την χρονική στιγμή και την διάρκεια κάθε λέξης, καθώς και πληροφορίες για το ποιές λέξεις ανήκουν στην ίδια πρόταση (με βάση το sequence id κάθε πρότασης), πραγματοποιείται συνένωση των λέξεων που ανήκουν στην ίδια πρόταση. Η αρχή της πρότασης σηματοδοτείται με τον χρόνο έναρξης (onset) της πρώτης λέξης που ανήκει σε αυτή την πρόταση, ενώ το τέλος αυτής ορίζεται με βάση τον χρόνο έναρξης (onset) προστιθέμενο με την διάρκεια (duration) της τελευταίας λέξης. Παράλληλα, εκμεταλλευόμενοι επιπλέον πληροφορίες που υπάρχουν στα αρχεία, γνωρίζουμε τον τίτλο της ιστορίας στην οποία ανήκει η εκάστοτε πρόταση, τον ρυθμό ομιλίας και την φωνή του/της διηγητή/τριας.

3.2.2 Επεξεργασία των MEG εγκεφαλικών σημάτων

Ακολουθώντας την ίδια διαχείριση των δεδομένων με τους συγγραφείς του βασικού άρθρου που βασίζεται η πτυχιακή εργασία [11], πραγματοποιείται η τεχνική του φιλτραρίσματος συχνοτήτων (filtering) και της επαναδειγματοληψίας (resampling) των σημάτων.

Όσον αφορά το φιλτράρισμα των συχνοτήτων των MEG σημάτων, εφαρμόζεται ένα ζωνοπερατό φίλτρο (band-pass filter), το οποίο αφαιρεί συχνότητες εκτός του εύρους [1 Hz, 60 Hz]. Αναλυτικότερα, αποκόπτονται πολύ χαμηλές συχνότητες, μικρότερες του 1Hz (High-pass filter) καθώς και υψηλές συχνότητες, μεγαλύτερες των 60Hz (Low-pass filter). Σύμφωνα με τους συγγραφείς του άρθρου [11], είναι μία κοινή πρακτική προκειμένου να μειωθεί ο θόρυβος και να διασφαλιστεί ότι η ανάλυση των σημάτων θα περιοριστεί μόνο στις ενδιαφέρουσες συχνότητες που συνδέονται με την εγκεφαλική λειτουργία.

Όσον αφορά την επαναδειγματοληψία των MEG σημάτων, επιλέγεται νέα συχνότητα δειγματοληψίας, από 1000Hz (1000 δείγματα το δευτερόλεπτο) στα 200Hz (200 δείγματα το δευτερόλεπτο). Αυτό είναι ιδιαίτερα χρήσιμο, δεδομένου ότι τα υψηλής συχνότητας σήματα όπως στη δική μας περίπτωση, καταναλώνουν περισσότερη μνήμη και πόρους επεξεργασίας κατά την ανάλυση των δεδομένων. Η μείωση της δειγματοληψίας του σήματος λοιπόν, αποτελεί έναν αποτελεσματικό τρόπο εξοικονόμησης χρόνου.

3.2.3 Τμηματοποίηση των MEG εγκεφαλικών σημάτων

Η τμηματοποίηση των MEG εγκεφαλικών σημάτων επιτυγχάνεται λαμβάνοντας υπόψιν τον χρόνο έναρξης και λήξης των προτάσεων. Πολλαπλασιάζοντας τη χρονική στιγμή της έναρξης και της λήξης των προτάσεων με τη συχνότητα δειγματοληψίας, που είναι πλέον 200Hz, προκύπτει το σημείο εκκίνησης και τερματισμού του σήματος σε δείκτες δειγματοληψίας. Κάθε πρόταση, η οποία ανήκει σε μία από τις τέσσερις ιστορίες που άκουσαν οι συμμετέχοντες, συνδέεται με αυτόν τον τρόπο με το αντίστοιχο τμήμα εγκεφαλικού σήματος.

3.2.4 Έλεγχος ακραίων τιμών (outliers)

Σύμφωνα με την βιβλιογραφία [18], γενικά τα MEG σήματα είναι από τη φύση τους θορυβώδη, καθώς οι αισθητήρες ανιχνεύουν όλες τις μεταβολές του μαγνητικού πεδίου, είτε αυτές οφείλονται στην εγκεφαλική δραστηριότητα είτε σε φυσιολογικούς θορύβους (βλεφάρισμα ματιών, καρδιακοί παλμοί, κινήσεις) ή εξωτερικούς παράγοντες (παρεμβολές από τα

ηλεκτρομαγνητικά πεδία που παράγονται από τις γραμμές ρεύματος και τις ηλεκτρικές συσκευές). Έχοντας τα MEG τμήματα, τα οποία είναι σε μορφή πινάκων, με την χρήση της βιβλιοθήκης NumPy της γλώσσας Python, ακολουθείται ένας έλεγχος των δεδομένων για την ύπαρξη ακραίων τιμών ή αλλιώς outliers, προκειμένου να διασφαλιστεί ότι αυτά τα MEG τμήματα, που θα είναι είσοδοι στο μοντέλο, είναι καθαρά και έγκυρα. Γενικά, οι outliers είναι παρατηρήσεις στα δεδομένα που διαφέρουν σημαντικά από τις υπόλοιπες τιμές στο σύνολο των δεδομένων. Αυτές οι τιμές είναι είτε πολύ μεγαλύτερες είτε πολύ μικρότερες από τις περισσότερες τιμές του συνόλου και μπορεί να επηρεάσουν σημαντικά τα τελικά αποτελέσματα του μοντέλου εκπαίδευσης. Για τον εντοπισμό πιθανών outliers, χρησιμοποιείται ακριβώς ο ίδιος έλεγχος που επέλεξε η ερευνητική ομάδα του άρθρου [11]. Συγκεκριμένα, πραγματοποιείται μια σειρά από ελέγχους που αποσκοπούν στον εντοπισμό ειδικών περιπτώσεων, όπως μηδενικές τιμές, κενά δεδομένα στους πίνακες, μη αριθμητικές τιμές ή θετικά/αρνητικά άπειρα στοιχεία. Σε περίπτωση που εντοπιστεί κάποιο πρόβλημα κατά τον έλεγχο των αναφερόμενων συνθηκών, το εκάστοτε MEG τμήμα απορρίπτεται και δεν αποτελεί μέρος του συνόλου δεδομένων.

3.2.5 Κανονικοποίηση δεδομένων και Ανίχνευση επιπλέον ανωμαλιών

Η διαδικασία ξεκινά με την εφαρμογή του Robust Scaler, ενός στατιστικού κανονικοποιητή, ο οποίος σε αντίθεση με άλλες μεθόδους κανονικοποίησης (όπως η αφαίρεση του μέσου όρου και η διαίρεση με την τυπική απόκλιση), αυτός χρησιμοποιεί την διάμεσο (median) και το διατεταρτημοριακό εύρος (Interquartile Range, IQR). Η διάμεσος είναι η κεντρική τιμή μίας ταξινομημένης λίστας τιμών. Δεδομένου ότι η διάσταση κάθε πίνακα που αναπαριστά ένα MEG τμήμα είναι (δείγματα, MEG κανάλια), η διάμεσος είναι η κεντρική τιμή ενός καναλιού αφού τα δείγματα ταξινομηθούν κατά αύξουσα σειρά. Το διατεταρτημοριακό εύρος υπολογίζεται ως η διαφορά μεταξύ του 3ου τεταρτημορίου (75ο ποσοστιαίο σημείο) και του 1ου τεταρτημορίου (25ο ποσοστιαίο σημείο). Αυτό το εύρος περιέχει το 50% των τιμών γύρω από τη διάμεσο. Για κάθε κανάλι δεδομένων λοιπόν, υπολογίζονται οι δύο παραπάνω μετρικές, δηλαδή η διάμεσος και το διατεταρτημοριακό εύρος. Έπειτα, κάθε τιμή ενός δείγματος στο εκάστοτε κανάλι κανονικοποιείται σύμφωνα με τον παρακάτω τύπο:

$$X_{\text{scaled}} = \frac{X - \text{Median}}{\text{IQR}}$$

Η αφαίρεση της τιμής του δείγματος με την διάμεσο (κεντροποίηση), βοηθά ώστε η διάμεσος να μετατοπίζεται στο μηδέν, κάτι που βοηθά στη σταθεροποίηση της κατανομής των τιμών. Παράλληλα, η διαίρεση με το διατεταρτημοριακό εύρος (κλιμάκωση) βοηθά ώστε να περιορίζονται οι ακραίες τιμές και τα δεδομένα να αποκτούν μία πιο συγκρίσιμη κλίμακα μεταξύ διαφορετικών καναλιών.

Η χρήση του συγκεκριμένου κανονικοποιητή είναι ιδιαίτερα χρήσιμη στη δική μας περίπτωση όπου τα MEG δεδομένα υπάρχει πιθανότητα να περιέχουν εξαιρετικά μεγάλες ή μικρές τιμές (outliers) που θα μπορούσαν να επηρεάσουν την ανάλυση. Επειδή η διάμεσος και το διατεταρτημοριακό εύρος επηρεάζονται ελάχιστα από (outliers), αυτή η μέθοδος εξασφαλίζει ότι η κανονικοποίηση δεν αλλοιώνεται από τέτοιες τιμές. Αντίθετα, οι ακραίες τιμές μπορούν συχνά να επηρεάσουν αρνητικά τον μέσο όρο ή τη διακύμανση του δείγ-

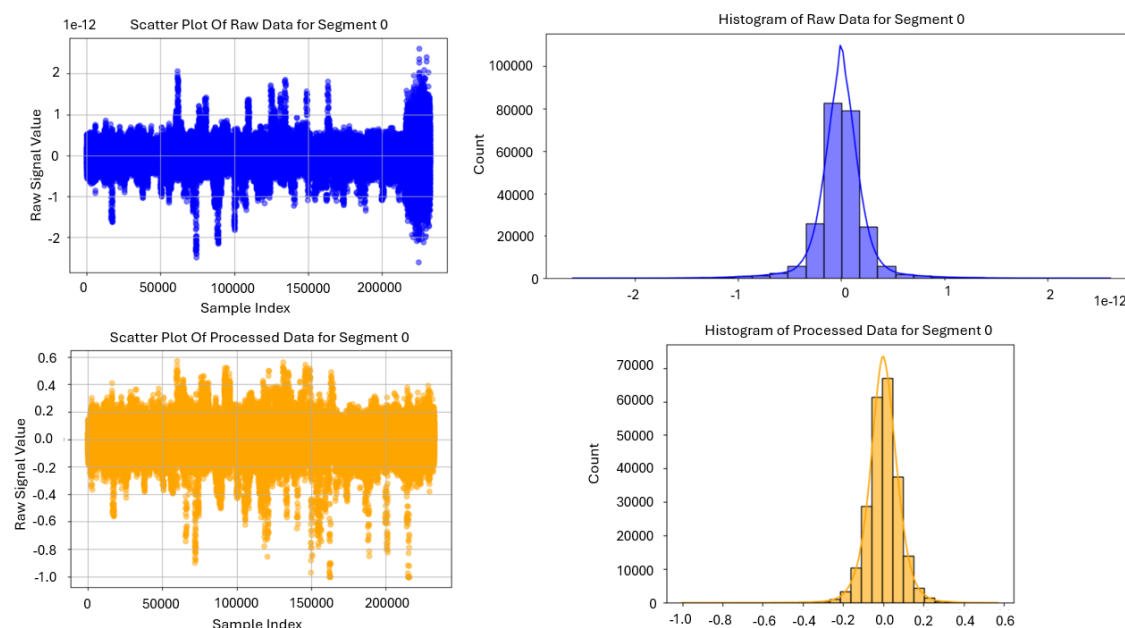
ματος. Για αυτόν τον λόγο, σε τέτοιες περιπτώσεις η χρήση της διαμέσου και του εύρους διατεταρτημορίων συχνά δίνει καλύτερα αποτελέσματα [19].

Το επόμενο βήμα περιλαμβάνει την εφαρμογή clipping στις τιμές που υπερβαίνουν ένα καθορισμένο κατώφλι (threshold). Αυτές οι υπερβολικές τιμές αντικαθίστανται με το όριο του κατωφλίου, ώστε να μην προκαλούν σφάλματα ή να διαταράσσουν την ανάλυση των δεδομένων. Αν το ποσοστό των τιμών που αποκόπηκαν υπερβαίνει το 20%, θεωρείται ότι τα δεδομένα είναι προβληματικά και απορρίπτονται.

Στη συνέχεια, τα δεδομένα κανονικοποιούνται περαιτέρω, διαιρούμενα με το κατώφλι, ώστε να διασφαλιστεί ότι οι τιμές παραμένουν εντός αποδεκτών ορίων. Η επιλογή του threshold στηρίχτηκε σε πειραματικές προσπάθειες της ερευνητικής ομάδας του άρθρου, προσπαθώντας να διατηρήσουν μία ισορροπία στα δεδομένα, δηλαδή ούτε να απορριφθεί μεγάλο ποσοστό των δεδομένων αλλά ούτε και να διατηρηθεί μεγάλο ποσοστό ακραίων τιμών.

Έπειτα από τα παραπάνω βήματα, εξετάζεται η παρουσία NaN (Not a Number) ή άπειρων τιμών (Inf) στα δεδομένα που θα μπορούσαν να προκαλέσουν σφάλματα κατά την ανάλυση. Αν εντοπιστούν τέτοιες ανωμαλίες, η διαδικασία διακόπτεται και το ανάλογο MEG τμήμα αφαιρείται από το σύνολο δεδομένων. Η διαδικασία ολοκληρώνεται με την επιστροφή των κανονικοποιημένων δεδομένων.

Στην παρακάτω εικόνα 3.3, μπορούμε να παρατηρήσουμε ενδεικτικά τα αποτελέσματα μετά την προεπεξεργασία του MEG τμήματος με id=0 του συμμετέχοντα "sub-01" καθώς ακούει την ιστορία task - 0: "Iw1" κατά την πρώτη συνεδρία (ses=0). Συγκεκριμένα, το συγκεκριμένο MEG τμήμα αντιστοιχεί στην πρόταση "Tara stood stock still waiting for the first tiny gleam from the scout craft to appear in the darkness of the" που ακούει ο συμμετέχων καθώς ξεκινάει η ιστορία.

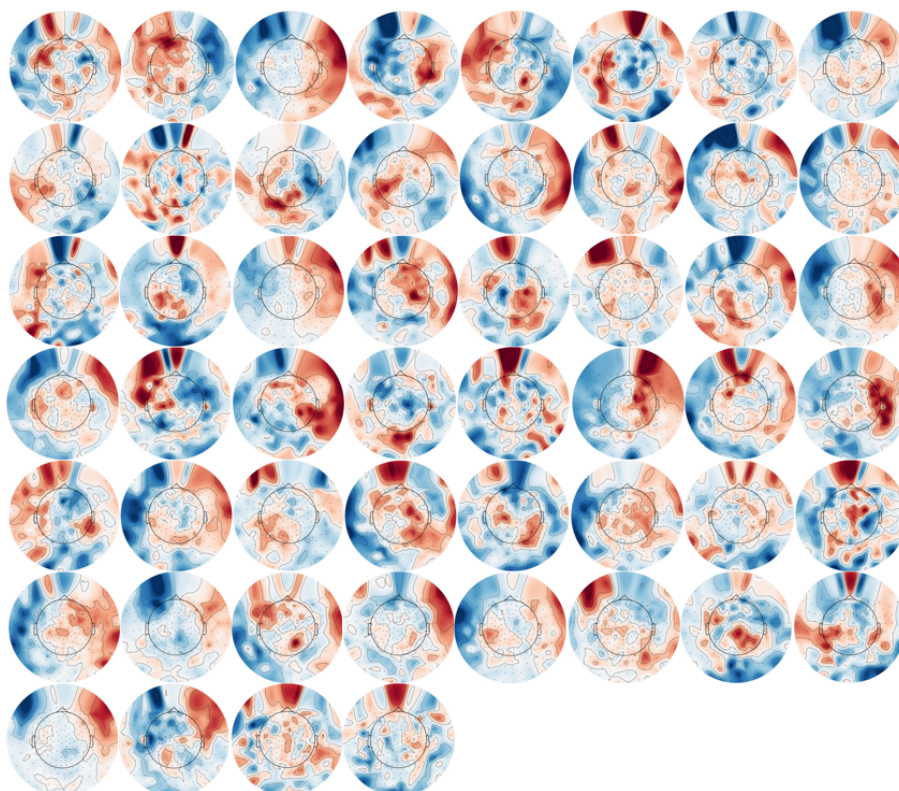


Εικόνα 3.3: Απεικόνιση τιμών MEG σήματος πριν και μετά την προεπεξεργασία.

Οι δύο απεικονίσεις στο πάνω μέρος της εικόνας, αντιστοιχούν στις τιμές των δειγμάτων του MEG σήματος πριν την προεπεξεργασία των δεδομένων. Αντίθετα, οι δύο απεικονίσεις

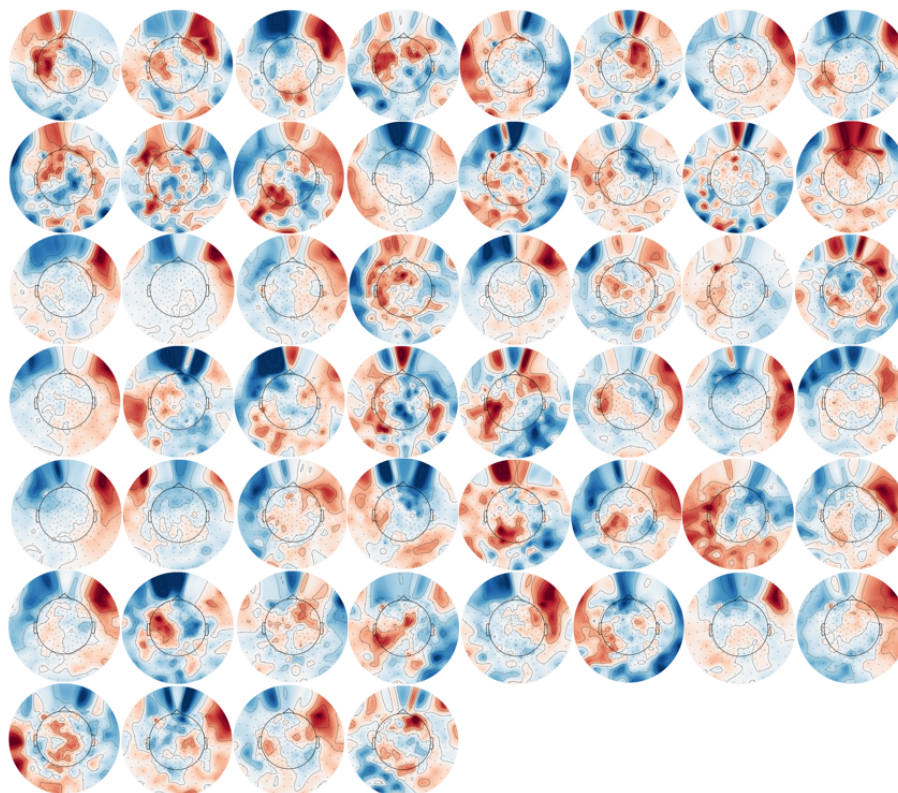
στο κάτω μέρος της εικόνας, αντιστοιχούν στις τιμές των δειγμάτων του MEG σήματος μετά την προεπεξεργασία των δεδομένων αυτών. Συγκρίνοντας τα γραφήματα, παρατηρούμε ότι μετά την προεπεξεργασία πράγματι έχει εφαρμοστεί κεντροποίηση καθώς οι περισσότερες τιμές είναι συγκεντρωμένες στο 0 καθώς και κλιμάκωση καθώς οι τιμές είναι στο εύρος [-1, 1]. Η κατανομή των τιμών φαίνεται καλύτερα στα δύο ιστογράμματα.

Έχοντας προεπεξεργαστεί τα MEG τμήματα όλων των συμμετεχόντων, παρακάτω απεικονίζεται ενδεικτικά η εγκεφαλική δραστηριότητα του συμμετέχοντα "sub-01" καθώς ακούει την ιστορία task - 0: "lw1" κατά την πρώτη συνεδρία, με την χρήση τοπογραφικών χαρτών [20]. Κάθε τοπογραφικός χάρτης απεικονίζει τη μέση δραστηριότητα των εγκεφαλικών σημάτων του συγκεκριμένου συμμετέχοντα για κάθε MEG κανάλι σε μια δεδομένη χρονική περίοδο. Η χρονική περίοδος αντιστοιχεί στο χρονικό διάστημα που ο συμμετέχοντας ακούει μία πρόταση. Η χρωματική κλίμακα δείχνει την ένταση της δραστηριότητας. Περιοχές με αποχρώσεις του κόκκινου αντιπροσωπεύουν πιο έντονες αποκρίσεις, δηλαδή μεγαλύτερη εγκεφαλική δραστηριότητα. Αντίθετα, περιοχές με αποχρώσεις του μπλε αντιπροσωπεύουν λιγότερο έντονες αποκρίσεις, δηλαδή μικρότερη εγκεφαλική δραστηριότητα. Με λίγα λόγια, κάθε τοπογραφικός χάρτης προσφέρει μια οπτική αναπαράσταση της κατανομής της νευρωνικής δραστηριότητας όπως ανιχνεύθηκε από τους MEG αισθητήρες για μία συγκεκριμένη πρόταση.



Εικόνα 3.4: Οπτική αναπαράσταση της κατανομής της νευρωνικής δραστηριότητας για κάθε πρόταση που ο συμμετέχων με id "sub-01" ακούει κατά την διάρκεια της ιστορίας "lw1". Οι περιοχές με κόκκινες αποχρώσεις υποδηλώνουν αυξημένη εγκεφαλική δραστηριότητα και ισχυρότερες αποκρίσεις, ενώ οι περιοχές με μπλε αποχρώσεις αντανakλούν μειωμένη εγκεφαλική δραστηριότητα και ασθενέστερες αποκρίσεις."

Παρακάτω, απεικονίζονται με ανάλογο τρόπο οι τοπογραφικοί χάρτες του συμμετέχοντα "sub-05" καθώς ακούει την ίδια ιστορία task - 0: "lw1" με τον "sub-01", πάλι κατά την πρώτη συνεδρία.

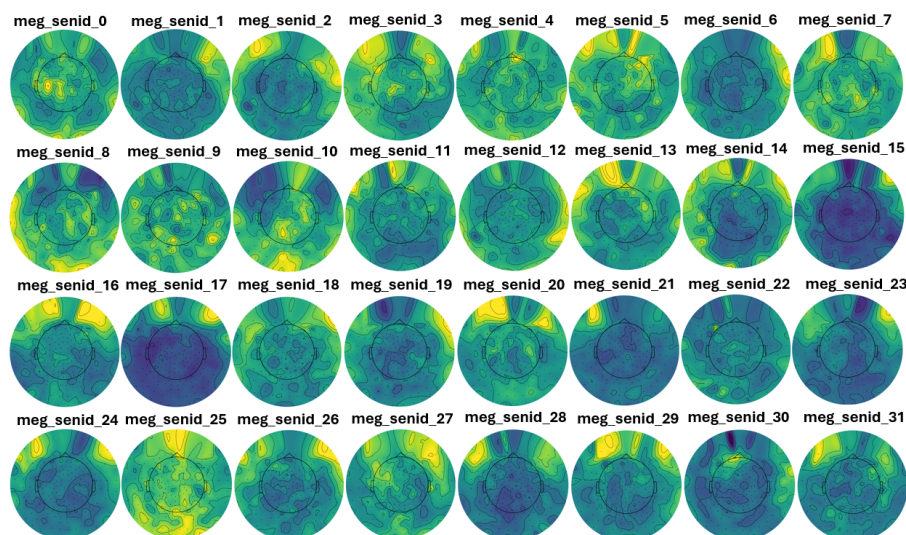


Εικόνα 3.5: Οπτική αναπαράσταση της κατανομής της νευρωνικής δραστηριότητας για κάθε πρόταση που ο συμμετέχων με id "sub-05" ακούει κατά την διάρκεια της ιστορίας "lw1". Οι περιοχές με κόκκινες αποχρώσεις υποδηλώνουν αυξημένη εγκεφαλική δραστηριότητα και ισχυρότερες αποκρίσεις, ενώ οι περιοχές με μπλε αποχρώσεις αντανακλούν μειωμένη εγκεφαλική δραστηριότητα και ασθενέστερες αποκρίσεις."

Παρατηρώντας όλους τους τοπογραφικούς χάρτες, μπορούμε διαισθητικά να παρατηρήσουμε ότι για κάθε πρόταση που ακούει ο συμμετέχοντας, ενεργοποιούνται διαφορετικές περιοχές του εγκεφάλου ανάλογα με τα ηχητικά ερεθίσματα που αντιλαμβάνεται.

Εύλογα μπορεί να αναρωτηθεί κανείς, αν οι δύο συμμετέχοντες ("sub-01" και "sub-05") παρουσιάζουν όμοια εγκεφαλική δραστηριότητα, δηλαδή αν ίδιες περιοχές του εγκεφάλου εμφανίζουν μεγαλύτερη ή μικρότερη δραστηριότητα, δεδομένου ότι ακούν την ίδια ακριβώς ιστορία (την "lw1"). Η απάντηση δίνεται μέσω της χρήσης διαφορικών χαρτών (Difference Maps), όπως φαίνεται στην εικόνα 3.6 και 3.7. Οι διαφορικοί χάρτες προκύπτουν από την αφαίρεση των τιμών κάθε MEG τμήματος του σήματος μεταξύ των δύο συμμετεχόντων. Η ακόλουθη εικόνα περιλαμβάνει διαφορικούς χάρτες, από τους οποίους ο καθένας απεικονίζει τις αποκλίσεις που μπορεί να υπάρχουν μεταξύ των εγκεφαλικών σημάτων των δύο συμμετεχόντων καθώς ακούν μία πρόταση. Για να εξάγουμε συμπεράσματα μέσω των διαφορικών χαρτών, πρέπει να λάβουμε υπόψιν ότι, οι κίτρινες περιοχές που διακρίνονται σε έναν χάρτη, αποκαλύπτουν αυξημένες αποκλίσεις στην εγκεφαλική δραστηριότητα μεταξύ των συμμετεχόντων. Οι πράσινες περιοχές αντιπροσωπεύουν αποκλίσεις αλλά όχι τόσο μεγάλες

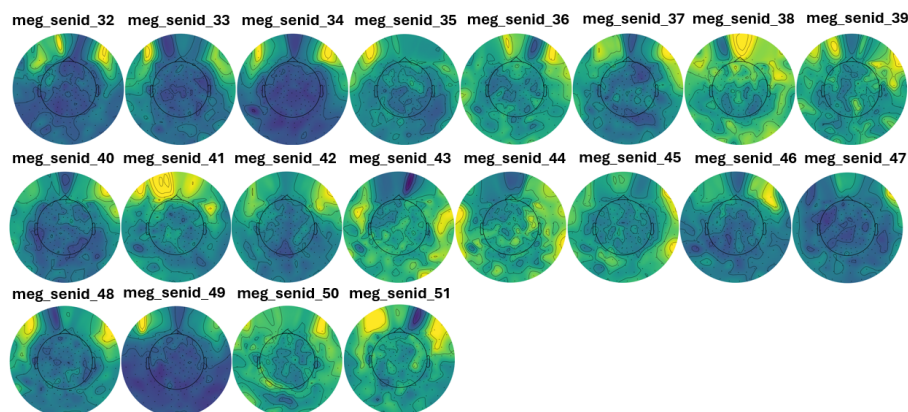
όσο στις κίτρινες περιοχές. Αντίθετα, οι μπλε περιοχές αποκαλύπτουν μικρές αποκλίσεις έως καθόλου των εγκεφαλικών σημάτων των δύο συμμετεχόντων. Καθώς η χρωματική κλίμακα κινείται από το σκούρο μπλε προς τις αποχρώσεις του πράσινου και έπειτα προς το κίτρινο, σημαίνει ότι οι περιοχές χαρακτηρίζονται από πολύ μικρές αποκλίσεις έως και πολύ μεγάλες.



Εικόνα 3.6: Κάθε διαφορικός χάρτης απεικονίζει τις διαφορές στην εγκεφαλική δραστηριότητα του sub-01 καθώς ακούει μία πρόταση της ιστορίας "Iw1" με την εγκεφαλική δραστηριότητα του sub-05 καθώς ακούει την ίδια ακριβώς πρόταση. Απεικονίζονται δηλαδή, οι αποκλίσεις μεταξύ του MEG τμήματος του "sub-01" που αντιστοιχεί σε μία πρόταση και του MEG τμήματος του "sub-05" που αντιστοιχεί στην ίδια πρόταση. Κάθε πρόταση έχει ένα συγκεκριμένο id, όπως senid_0, senid_1, senid_2. Στη συγκεκριμένη εικόνα, παρουσιάζονται οι διαφορικοί χάρτες για τις πρώτες 31 προτάσεις της ιστορίας. Η χρωματική κλίμακα, η οποία μεταβαίνει από το σκούρο μπλε προς το πράσινο και στη συνέχεια στο κίτρινο, αποτυπώνει τις κλίμακες διαφορών, αναδεικνύοντας περιοχές με ελάχιστες έως μεγάλες αποκλίσεις κατά την σύγκριση των σημάτων μεταξύ των δύο συμμετεχόντων.

Παρατηρώντας τους διαφορικούς χάρτες της εικόνας, αντιλαμβάνεται κανείς ότι κατά την ακρόαση των περισσότερων προτάσεων της ιστορίας, οι δύο συμμετέχοντες παρουσιάζουν σημαντικές διαφορές στην εγκεφαλική τους δραστηριότητα σε ορισμένες περιοχές του εγκεφάλου. Παραδείγματος χάριν, ο ένας συμμετέχων μπορεί να παρουσιάζει αυξημένη εγκεφαλική δραστηριότητα σε σχέση με τον άλλον, οδηγώντας σε αυξημένη απόκλιση μεταξύ της εγκεφαλικής τους δραστηριότητας (περιοχές με πράσινο και κίτρινο χρώμα). Παράλληλα, παρατηρούνται περιοχές με παρόμοια εγκεφαλική δραστηριότητα, όπως φαίνεται στις μπλε περιοχές. Οι διαφορές της εγκεφαλικής τους δραστηριότητας μπορεί να σχετίζεται με διαφορετική γλωσσική επεξεργασία, ατομικές διαφορές ή συγκεκριμένες γνωστικές λειτουργίες.

Αναλυτικότερα, η γλωσσική επεξεργασία σχετίζεται με τον τρόπο που κάθε συμμετέχων επεξεργάζεται τη γλώσσα. Αν ένας συμμετέχων είναι πιο εξοικειωμένος με τη γλώσσα, μπορεί να επεξεργάζεται τις λέξεις με μεγαλύτερη ευκολία, απαιτώντας λιγότερη γνωστική προσπάθεια και παρουσιάζοντας μειωμένη ενεργοποίηση σε ορισμένες εγκεφαλικές περιοχές. Αντίθετα, ένας λιγότερο εξοικειωμένος συμμετέχων ενδέχεται να καταβάλλει μεγαλύτερη προσπάθεια για την κατανόηση της ομιλίας, οδηγώντας σε αυξημένη εγκεφαλική δραστηριότητα σε περιοχές που σχετίζονται με τη σημασιολογική και φωνολογική επεξεργασία. Επιπλέον,

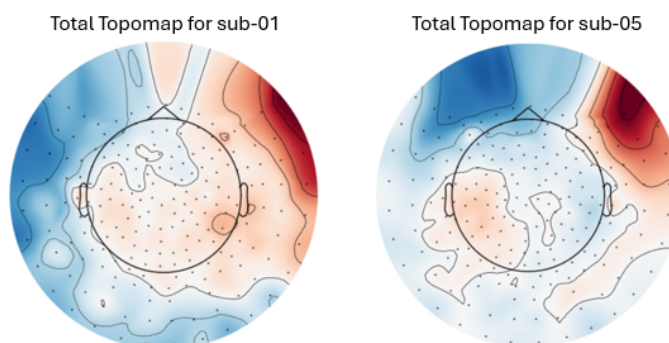


Εικόνα 3.7: Κάθε διαφορικός χάρτης αποτυπώνει τις αποκλίσεις μεταξύ των MEG τμημάτων των συμμετεχόντων "sub-01" και "sub-05", τα οποία αντιστοιχούν στις προτάσεις με *senid_31*, *senid_32*, *senid_33* έως και *senid_51*. Στη συγκεκριμένη εικόνα, παρουσιάζονται οι διαφορικοί χάρτες για τις τελευταίες 20 προτάσεις της ιστορίας. Η χρωματική κλίμακα, η οποία μεταβαίνει από το σκούρο μπλε προς το πράσινο και στη συνέχεια στο κίτρινο, αποτυπώνει τις κλίμακες διαφορών, αναδεικνύοντας περιοχές με ελάχιστες έως μεγάλες αποκλίσεις κατά την σύγκριση των σημάτων μεταξύ των δύο συμμετεχόντων.

οι διαφορές στην εγκεφαλική δραστηριότητα μεταξύ των ατόμων μπορεί να αντικατοπτρίζουν μοναδικά χαρακτηριστικά κάθε ατόμου. Η δομή και η λειτουργία του εγκεφάλου διαφέρουν μεταξύ των ανθρώπων, επηρεάζοντας τον τρόπο με τον οποίο επεξεργάζονται το ίδιο ακουστικό ερέθισμα. Η επεξεργασία της ομιλίας ακόμη, επηρεάζεται από παράγοντες όπως την μνήμη, την προσοχή και την συγκέντρωση. Ένα άτομο με υψηλή μνημονική ικανότητα μπορεί να ανακαλεί πληροφορίες πιο εύκολα, διευκολύνοντας τη συνολική κατανόηση του νοήματος μιας πρότασης. Ένα άτομο όμως με λιγότερη μνημονική ικανότητα μπορεί να χρειάζεται περισσότερη προσπάθεια για να συνδέσει τις πληροφορίες και να κατανοήσει το περιεχόμενο. Παράλληλα, ένα άτομο με σημαντικά επίπεδα προσοχής και συγκέντρωσης μπορεί να παρακολουθεί με ακρίβεια τις λεπτομέρειες της ομιλίας και να επεξεργάζεται τα νοήματα γρηγορότερα, παρουσιάζοντας πιθανόν λιγότερη εγκεφαλική δραστηριότητα.

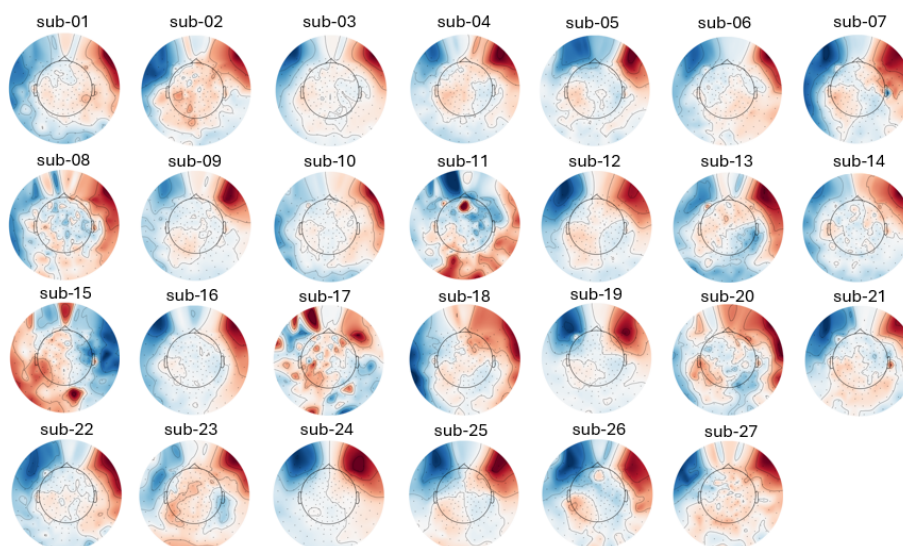
Στην εικόνα 3.8, απεικονίζεται αριστερά ένας τοπογραφικός χάρτης για τον συμμετέχοντα "sub-01" και ένας δεξιά για τον συμμετέχοντα "sub-05". Αυτοί οι τοπογραφικοί χάρτες απεικονίζουν την μέση εγκεφαλική δραστηριότητα για όλες τις προτάσεις που άκουσε ο καθένας μέχρι να ολοκληρωθεί η ιστορία "Iw1". Μέσω αυτών, δίνεται μία γενική εικόνα της συνολικής εγκεφαλικής δραστηριότητας για τον κάθε συμμετέχοντα. Ειδικότερα, παρατηρώντας τους δύο τοπογραφικούς χάρτες, μπορούμε να παρατηρήσουμε ότι το δεξιό ημισφαίριο φαίνεται να έχει πιο ενεργό ρόλο στην επεξεργασία της ομιλίας και για τους δύο συμμετέχοντες, ενώ το αριστερό ημισφαίριο παρουσιάζει λιγότερη δραστηριότητα.

Οι δύο αυτές απεικονίσεις, διέγειραν το ενδιαφέρον και εξετάστηκε η συνολική εγκεφαλική δραστηριότητα όλων των συμμετεχόντων που άκουγαν την ιστορία "Iw1" στη πρώτη συνεδρία. Το ερώτημα που κρίνεται απαραίτητο να απαντηθεί είναι εάν όλοι οι συμμετέχοντες παρουσιάζουν παρόμοια εγκεφαλική δραστηριότητα, δηλαδή σε όλους ενεργοποιείται περισσότερο το δεξιό ημισφαίριο. Εάν ναι, πρέπει να αναρωτηθούμε εάν το κοινό μοτίβο των εγκεφαλικών σημάτων όλων των συμμετεχόντων μπορεί να επηρεάσει την απόδοση του



Εικόνα 3.8: Σύγκριση της συνολικής εγκεφαλικής δραστηριότητας του συμμετέχοντα "sub-01" και "sub-05". Το δεξιό ημισφαίριο και των δύο παρουσιάζει πιο ενεργό ρόλο απ' όυ το αριστερό.

μοντέλου.



Εικόνα 3.9: Απεικόνιση της συνολικής εγκεφαλικής δραστηριότητας όλων των συμμετεχόντων καθώς άκουγαν την ιστορία "Iw1" στη πρώτη συνεδρία. Παρατηρείται ότι στους περισσότερους συμμετέχοντες το δεξιό ημισφαίριο παρουσιάζει μεγαλύτερη εγκεφαλική δραστηριότητα από το αριστερό.

Παρατηρώντας την εγκεφαλική δραστηριότητα όλων των ατόμων, όπως φαίνεται στην εικόνα 3.9, η πλειονότητα των συμμετεχόντων παρουσιάζει μεγαλύτερη εγκεφαλική δραστηριότητα στις περιοχές του δεξιού ημισφαιρίου. Αντίθετα, μικρότερη δραστηριότητα παρουσιάζουν στις περιοχές του αριστερού ημισφαιρίου. Ωστόσο, αποτελεί ενδιαφέρον η εγκεφαλική δραστηριότητα του συμμετέχοντα "sub-11", του "sub-15" και του "sub-17", καθώς ξεχωρίζουν από αυτή των υπολοίπων. Ειδικότερα, ο "sub-11" παρουσιάζει αυξημένη δραστηριότητα στη μετωπιαία περιοχή, λιγότερη στη περιοχή του δεξιού ημισφαιρίου και περισσότερη στο πίσω μέρος του εγκεφάλου. Όσον αφορά τον "sub-15" και τον "sub-17", πιο ενεργό ρόλο παρουσιάζει το αριστερό του ημισφαίριο αντί για το δεξιό.

Είναι σημαντικό να διερευνήσουμε ποια είναι η λειτουργική κατανομή μεταξύ του δεξιού και του αριστερού ημισφαιρίου και πώς αυτές οι διαφορετικές λειτουργίες επηρεάζουν τη συμπεριφορά και τις αντιδράσεις του εγκεφάλου όλων των συμμετεχόντων. Σύμφωνα

με τον Sahil Luthra [21], το αριστερό ημισφαίριο θεωρείται κυρίως υπεύθυνο για τις διαδικασίες που σχετίζονται με τη γλωσσική επεξεργασία. Είναι ενεργό στην αναγνώριση και κατηγοριοποίηση των φωνητικών στοιχείων καθώς και στην επεξεργασία της γραμματικής και της συντακτικής δομής του λόγου. Είναι σημαντικό για την κατανόηση των λέξεων και της σημασίας τους, ιδιαίτερα στη φυσική γλωσσική επικοινωνία και στον τρόπο που οι λέξεις συνδυάζονται για να δημιουργήσουν νοήματα. Αν και το αριστερό ημισφαίριο κυριαρχεί στην επεξεργασία του λόγου, το δεξιό ημισφαίριο επίσης παίζει κρίσιμο ρόλο στην αντίληψη του λόγου, κυρίως με την επεξεργασία των φωνητικών χαρακτηριστικών και των παραμέτρων του ήχου που δεν σχετίζονται άμεσα με τη σημασιολογική δομή. Το δεξιό ημισφαίριο είναι υπεύθυνο για την ανάλυση του τονισμού, του ρυθμού και της έντασης της ομιλίας καθώς και για την αναγνώριση των φωνητικών χαρακτηριστικών που συνδέονται με την ταυτότητα του ομιλητή. Αυτό σημαίνει ότι το δεξιό ημισφαίριο είναι σε θέση να αναγνωρίζει τις διαφορές στη φωνή μεταξύ των ομιλητών, όπως τον τόνο ή τον τρόπο ομιλίας τους. Αυτές οι διαφοροποιήσεις βοηθούν στη διάκριση μεταξύ διαφορετικών ομιλητών και είναι σημαντικές για την κατανόηση του λόγου καθώς επηρεάζουν το πλαίσιο μέσα στο οποίο χρησιμοποιούνται οι λέξεις.

Η συνεργασία συνεπώς μεταξύ των δύο ημισφαιρίων του εγκεφάλου είναι ιδιαίτερα σημαντική για την πλήρη κατανόηση και επεξεργασία του λόγου. Το αριστερό ημισφαίριο επικεντρώνεται στις δομικές πτυχές του λόγου, όπως τη γλωσσική σύνταξη και τη σημασία, διασφαλίζοντας την κατανόηση και τη σωστή χρήση των λέξεων. Από την αντίπερα όχθη, το δεξιό ημισφαίριο αναλαμβάνει την επεξεργασία των συναισθηματικών και κοινωνικών πτυχών του λόγου που αποκαλύπτουν την ταυτότητα και τη συναισθηματική κατάσταση του ομιλητή. Η αλληλεπίδραση των δύο ημισφαιρίων επιτρέπει την κατανόηση όχι μόνο του τι λέγεται, αλλά και του πώς και γιατί λέγεται.

Ο "sub-11", που παρουσιάζει αυξημένη δραστηριότητα στη μετωπιαία περιοχή και μειωμένη στο δεξιό ημισφαίριο, ενδεχομένως να επεξεργάζεται τον λόγο με έναν διαφορετικό μηχανισμό σε σχέση με τους υπόλοιπους συμμετέχοντες. Αντίθετα, ο "sub-15" και ο "sub-17", με αυξημένη δραστηριότητα στο αριστερό ημισφαίριο, φαίνεται να έχουν διαφορετική κατανομή της εγκεφαλικής δραστηριότητας, κάτι που υποδηλώνει διαφορετικούς τρόπους επεξεργασίας του λόγου, με έμφαση στις διαδικασίες που σχετίζονται με τη γλωσσική ανάλυση.

Επομένως, αν και φαίνεται ότι η πλειονότητα των συμμετεχόντων χαρακτηρίζεται με παρόμοια εγκεφαλική δραστηριότητα, υπάρχουν εγκεφαλικά δεδομένα διαφορετικής μεταβλητότητας. Αυτό σημαίνει ότι το μοντέλο δεν περιορίζεται σε έναν συγκεκριμένο τύπο εγκεφαλικής δραστηριότητας, αλλά μπορεί να ανιχνεύει και να ανταποκρίνεται σε πιο ποικιλόμορφα εγκεφαλικά σήματα. Με αυτόν τον τρόπο, καθίσταται πιο ευέλικτο και ικανό να γενικεύει, δηλαδή να εφαρμόζεται σε νέες καταστάσεις με αντικειμενικό τρόπο, χωρίς να επηρεάζεται από ατομικές διαφορές. Αυτό είναι σημαντικό για την αποτελεσματικότητα του μοντέλου σε ποικίλα σενάρια και αυξάνει την αξιοπιστία και χρηστικότητα του σε πραγματικές εφαρμογές.

3.2.6 Διαχωρισμός δεδομένων σε δεδομένα εκπαίδευσης, επικύρωσης και ελέγχου

Έπειτα από την προεπεξεργασία των MEG τμημάτων και την αντιστοίχηση τους με τις ανάλογες προτάσεις, πραγματοποιείται ο διαχωρισμός του συνόλου δεδομένων σε τρία σύνολα. Το σύνολο εκπαίδευσης, επικύρωσης και ελέγχου.

- Το σύνολο εκπαίδευσης (train set) χρησιμοποιείται κατά την εκπαίδευση του μοντέλου, το οποίο μαθαίνει από αυτά τα δεδομένα.
- Το σύνολο επικύρωσης (validation set) χρησιμοποιείται για να αξιολογηθεί η ικανότητα του μοντέλου να γενικεύσει σε δεδομένα που δεν έχει δει κατά τη διαδικασία εκπαίδευσης.
- Το σύνολο ελέγχου (test set) χρησιμοποιείται για να αξιολογηθεί η τελική απόδοση του εκπαιδευμένου πλέον μοντέλου με έναν αμερόληπτο τρόπο. Παρέχει μια αξιόπιστη εκτίμηση για το πώς αποδίδει το μοντέλο σε άγνωστα δεδομένα.

Και τα τρία σύνολα δημιουργήθηκαν με τυχαίο τρόπο έπειτα από τυχαίο ανακάτεμα του αρχικού συνόλου δεδομένων (shuffle). Το καθένα περιλαμβάνει τις διαδρομές των αρχείων που περιλαμβάνουν MEG τμήματα εγκεφαλικών σημάτων από διάφορους συμμετέχοντες, τα οποία καταγράφηκαν καθώς άκουγαν μία ιστορία σε μία συγκεκριμένη συνεδρία, καθώς και την ανάλογη πρόταση που αντιστοιχεί σε αυτό που άκουσαν στο συγκεκριμένο χρονικό διάστημα. Ο διαχωρισμός του αρχικού συνόλου δεδομένων σε train:validation:test έγινε με αναλογία 80:10:10, όπως ακριβώς και στο άρθρο [11]. Συγκεκριμένα, το train set περιέχει 23339 ζευγάρια δεδομένων "MEG τμήματα εγκεφαλικών σημάτων - αντίστοιχη πρόταση", το validation set 2917, ενώ το test set 2918.

3.2.7 Εφαρμογή του "Whisper Encoder" για εξαγωγή χαρακτηριστικών από MEG σήματα

Το τελευταίο μέρος της προεπεξεργασίας των MEG σημάτων αξιοποιεί την αρχιτεκτονική του "Whisper" κωδικοποιητή για την εξαγωγή χαρακτηριστικών των σημάτων αυτών. Ο "Whisper Encoder" προσφέρει πλούσιες αναπαραστάσεις της ομιλίας, όπως φωνητικά χαρακτηριστικά, φωνολογικά μοτίβα και γλωσσικές πληροφορίες, σε αντίθεση με πιο παραδοσιακές προσεγγίσεις όπως τα "MFCCs" (Mel-Frequency Cepstral Coefficients), γεγονός που το καθιστά πιο κατάλληλο για την κωδικοποίηση ομιλίας από εγκεφαλικά σήματα (όπως MEG δεδομένα). Όσον αφορά τα "MFCCs", είναι μία από τις κλασικές τεχνικές εξαγωγής χαρακτηριστικών από ηχητικά σήματα, η οποία απομονώνει τις φασματικές πληροφορίες του ήχου, δηλαδή συχνότητες που συνθέτουν το ηχητικό σήμα, οι οποίες αποκαλύπτουν σημαντικά χαρακτηριστικά όπως τονικότητα και χροιά, τα οποία είναι θεμελιώδη για την κατανόηση και αναγνώριση της ομιλίας. Εντούτοις, δεν διατηρεί πληροφορίες υψηλότερου επιπέδου, όπως φωνολογικά ή γλωσσικά μοτίβα, κάτι που είναι απαραίτητο για την κωδικοποίηση ομιλίας από εγκεφαλικά σήματα [22].

Ακόμη, ο "Whisper Encoder" είναι ιδανικός για την εξαγωγή χαρακτηριστικών από εγκεφαλικά σήματα που καταγράφηκαν κατά την διάρκεια ηχητικών ερεθισμάτων σε αντίθεση

με μοντέλα βασισμένα στον "BART" [8], [9], καθώς είναι ειδικά σχεδιασμένος για να επεξεργάζεται ηχητικά σήματα αλλά και ανθεκτικός σε θορυβώδη δεδομένα (όπως τα εγκεφαλικά σήματα). Αντίθετα, ο "BART" είναι ένα μοντέλο που χρησιμοποιείται κυρίως για επεξεργασία φυσικής γλώσσας (NLP), όπως για περιλήψεις, μετάφραση ή παραγωγή κειμένων και δεν είναι σχεδιασμένο για να επεξεργάζεται ακατέργαστα ηχητικά δεδομένα [23]. Συνεπώς, ένας από τους βασικούς λόγους επιλογής του "Whisper" είναι η ικανότητά του να μαθαίνει ισχυρές χρονικές και σημασιολογικές συσχετίσεις μεταξύ ακουστικών σημάτων και γλωσσικών αναπαραστάσεων. Δεδομένου ότι η μετάφραση των MEG σημάτων σε κείμενο παρουσιάζει ομοιότητες με την αναγνώριση ομιλίας, καθώς και στις δύο περιπτώσεις απαιτείται η αντιστοίχιση χρονικά εξαρτώμενων σημάτων σε γλωσσικές αναπαραστάσεις, το "Whisper" παρέχει μια αρχιτεκτονική ήδη προσαρμοσμένη στη ανάλυση πολύπλοκων ακολουθιακών δεδομένων.

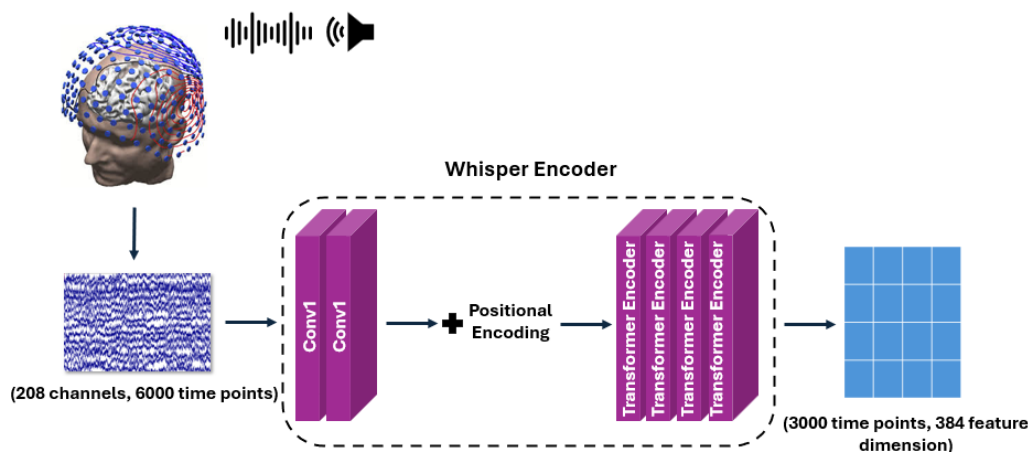
Προϋπόθεση για την είσοδο των MEG σημάτων στον κωδικοποιητή είναι να έχουν όλα το ίδιο μήκος καθώς και τον ίδιο αριθμό MEG καναλιών. Όσον αφορά το μήκος των σημάτων, το μέγιστο μήκος που επιτρέπεται να έχουν είναι 6000 χρονικά σημεία ή αλλιώς δείγματα. Ο αριθμός αυτός προκύπτει από τον πολλαπλασιασμό μεταξύ της μέγιστης διάρκειας του δείγματος (30 δευτερόλεπτα) και τη συχνότητα δειγματοληψίας του σήματος (200 Hz). Σε περίπτωση που τα σήματα έχουν μικρότερο μήκος εφαρμόζεται padding, δηλαδή προστίθενται μηδενικά μέχρι να συμπληρωθεί το μέγιστο μήκος δειγμάτων. Σε περίπτωση όμως που τα σήματα έχουν μεγαλύτερο μήκος, εφαρμόζεται αποκοπή του σήματος ώστε το μήκος να περιορίζεται στα 6000 δείγματα. Όσον αφορά τον αριθμό των καναλιών, διασφαλίζεται ότι τα δεδομένα έχουν τον σωστό αριθμό καναλιών, ελέγχοντας αν ο αριθμός είναι ήδη ίσος με τον επιθυμητό (208). Εάν είναι μικρότερος από τον επιθυμητό, προστίθεται padding ώστε να συμπληρωθούν τα κενά και να φτάσουν τον απαιτούμενο αριθμό καναλιών. Και στις δύο περιπτώσεις, πραγματοποιείται έλεγχος προκειμένου να διασφαλιστεί ότι τα δεδομένα είναι σε συμφωνία με τις απαιτήσεις του μοντέλου.

Έπειτα, τα σήματα εισάγονται σε δύο συνελκτικά επίπεδα. Το πρώτο μετατρέπει την διάσταση των χαρακτηριστικών εισόδου, δηλαδή τα 208 MEG κανάλια, σε διάσταση όση απαιτεί ο κωδικοποιητής και συγκεκριμένα σε 384. Στη συνέχεια, με το επόμενο συνελκτικό επίπεδο μειώνεται η χρονική διάσταση των σημάτων, δηλαδή το μήκος τους, κατά το ήμισυ (3000 χρονικά σημεία). Με αυτό τον τρόπο τα σήματα είναι συμβατά με βάση τις απαιτήσεις του κωδικοποιητή. Τα σήματα περνούν διαδοχικά στο επόμενο επίπεδο, το Positional Embedding Layer, στο οποίο προστίθεται πληροφορία για την σειρά των χρονικών σημείων των σημάτων, ενημερώνοντας το μοντέλο για το πόσο μακριά ή κοντά βρίσκεται ένα χρονικό σημείο του σήματος από ένα άλλο. Αυτό είναι και το τελευταίο στάδιο από το οποίο περνούν τα σήματα για να εισαχθούν στο βασικότερο κομμάτι με το οποίο επιτυγχάνεται η κωδικοποίηση τους σε χαρακτηριστικά, τον "Whisper" κωδικοποιητή. Ο "Whisper" κωδικοποιητής αποτελείται από 4 επίπεδα Transformer Encoder. Στο τέλος, από τον "Whisper" κωδικοποιητή εξάγεται ένας πίνακας διαστάσεων (3000 χρονικά σημεία, 384 χαρακτηριστικά), ο οποίος είναι οι λεγόμενες κρυφές καταστάσεις (hidden states) ή ενσωματώσεις (embeddings). Κάθε κελί του πίνακα περιέχει μία τιμή που αναπαριστά το βάρος του εκάστοτε χαρακτηριστικού για ένα δεδομένο χρονικό σημείο.

Τονίζεται ότι, η επιλογή των τιμών των παραμέτρων που συνθέτουν την αρχιτεκτονική του "Whisper Encoder" παραμένει ίδια με αυτή του "Whisper Encoder" που εκπαιδεύεται

στο "NeuSpeech" μοντέλο. Κατά την διαδικασία της εξαγωγής χαρακτηριστικών των εγκεφαλικών σημάτων, η προσέγγιση μας διαφοροποιείται μόνο στον τρόπο αποθήκευσης των αποτελεσμάτων που εξάγονται από τον κατά τα άλλα όμοιο σε αρχιτεκτονική κωδικοποιητή. Τα embeddings κάθε εγκεφαλικού σήματος αποθηκεύονται σε αρχεία μορφής H5 μαζί με τις ανάλογες προτάσεις. Η επιλογή αυτού του τύπου αρχείων βασίζεται στην ικανότητα τους να προσφέρουν έναν αποδοτικό και οργανωμένο τρόπο διαχείρισης μεγάλου όγκου δεδομένων, διευκολύνοντας την αποθήκευση τους σε συστήματα περιορισμένης χωρητικότητας. Η δυνατότητα συμπίεσης των δεδομένων μέσα σε τέτοιου τύπου αρχεία επιτρέπει την αποθήκευση πληροφοριών σε μικρότερο μέγεθος χωρίς απώλεια πληροφορίας, γεγονός που είναι ιδιαίτερα σημαντικό όταν διαχειριζόμαστε τεράστιες ποσότητες δεδομένων, όπως τα embeddings των εγκεφαλικών σημάτων. Επιτρέπουν γρήγορη πρόσβαση σε συγκεκριμένα τμήματα των δεδομένων χωρίς την ανάγκη φόρτωσης ολόκληρου του αρχείου στη μνήμη, βελτιώνοντας την αποτελεσματικότητα στην επεξεργασία μεγάλων και σύνθετων δεδομένων. Επιπλέον, η συγκεκριμένη μορφή αρχείων διευκολύνει την αποτελεσματική διαχείριση των δεδομένων, ιδιαίτερα κατά την εκπαίδευση του μοντέλου, καθώς επιτρέπει τη συγκέντρωση όλων των απαιτούμενων δεδομένων σε ενιαία αρχεία. Αυτή η συγκέντρωση μειώνει το υπολογιστικό κόστος, καθώς εξαιρείται η ανάγκη φόρτωσης των embeddings των εγκεφαλικών σημάτων και των αντίστοιχων προτάσεων από διαφορετικές πηγές, βελτιώνοντας έτσι την αποδοτικότητα της διαδικασίας.

Η εικόνα 3.10, αναπαριστά την διαδικασία εξαγωγής χαρακτηριστικών από τα εγκεφαλικά σήματα.

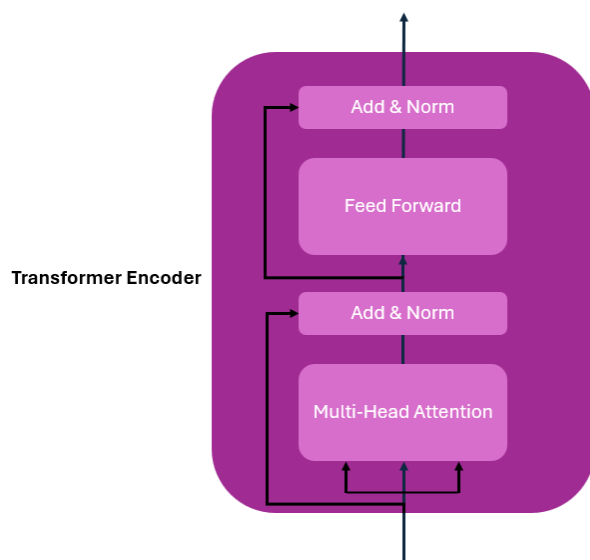


Εικόνα 3.10: Διαδικασία εξαγωγής χαρακτηριστικών από τα εγκεφαλικά σήματα. Η εικόνα είναι προσαρμοσμένη από την παρουσίαση του Joachim Gross [24] και το πλέγμα ηλεκτροδίων βασίζεται σε στιγμιότυπο αυτής της παρουσίασης.

Transformer Encoder

Το νευρωνικό δίκτυο Transformer είναι ένα ισχυρό μοντέλο βαθιάς μάθησης που παρουσιάστηκε το 2017 στο άρθρο "Attention Is All You Need" από τον Vaswani και τους συνεργάτες του [25]. Η καινοτομία του Transformer έγκειται στη χρήση του μηχανισμού Multi-Head Attention, ο οποίος επιτρέπει στο μοντέλο να επικεντρώνεται σε διαφορετικά

μέρη της εισόδου ταυτόχρονα, αντί να επεξεργάζεται την είσοδο σειριακά. Αυτή η παράλληλη επεξεργασία οδηγεί σε εξαιρετική ταχύτητα αλλά και απόδοση και γι' αυτό η συγκεκριμένη αρχιτεκτονική αποτελεί την βάση πολλών γλωσσικών μοντέλων.



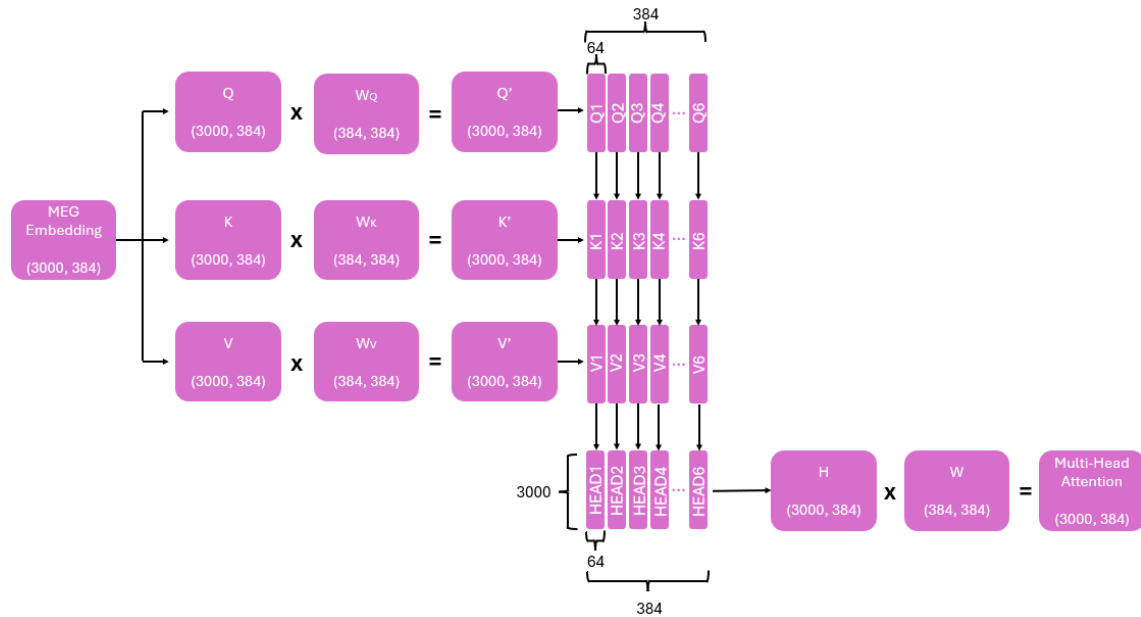
Εικόνα 3.11: *Transformer Encoder*

Αναφορικά με τον μηχανισμό Multi-Head Attention στη δική μας περίπτωση, αυτός επιτρέπει την συσχέτιση κάθε χρονικού σημείου ενός εγκεφαλικού σήματος με ένα άλλο. Έχοντας αρχικά 3000 χρονικά σημεία με 384 διαστάσεις, αυτό σημαίνει ότι κάθε χρονικό σημείο αντιπροσωπεύεται από ένα διάνυσμα μεγέθους 384. Όπως δείχνει η εικόνα 3.12, λαμβάνονται τρία αντίγραφα της εισόδου, δηλαδή των MEG Embeddings που προέκυψαν μετά το στάδιο του Positional Embedding, διαστάσεων (3000, 384) και εισάγονται στο Multi-Head Attention επίπεδο. Συγκεκριμένα, τα αντίγραφα αυτά είναι τρεις πίνακες Q , K και V . Ο καθένας από αυτούς πολλαπλασιάζεται με έναν πίνακα παραμέτρων, διαστάσεων (384, 384), με ονόματα W_Q , W_K και W_V αντίστοιχα. Οι πίνακες που προκύπτουν έχουν την ίδια διάσταση με αυτή της εισόδου. Έπειτα, οι τελικοί πίνακες χωρίζονται σε μικρότερους πίνακες.

Μία από τις παραμέτρους που έχει επιλεχθεί για τον κωδικοποιητή είναι οι κεφαλές (heads) και είναι 6 σε αριθμό. Αυτή είναι η παράμετρος που κάνει τον Transformer να ξεχωρίζει καθώς αναφέρεται σε πολλαπλούς μηχανισμούς προσοχής, οι οποίοι εκτελούνται παράλληλα. Κάθε κεφαλή προσοχής επικεντρώνεται σε διαφορετικές πτυχές των δεδομένων εισόδου. Οι μικρότεροι πίνακες λοιπόν αντικατοπτρίζουν τις κεφαλές, οπότε είναι 6. Κάθε κεφαλή έχει διάσταση $384/6 = 64$. Κάθε head βλέπει όλα τα χρονικά σημεία του εκάστοτε εγκεφαλικού σήματος, αλλά για ένα μόνο χαρακτηριστικό τους. Έτσι, η κάθε κεφαλή παρακολουθεί διαφορετικές πτυχές κάθε χρονικού σημείου.

Η προσοχή για κάθε κεφαλή ($HEAD_i$) υπολογίζεται από τον παρακάτω τύπο:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Εικόνα 3.12: *Multi-Head Attention*

όπου d_k οι διαστάσεις κάθε κεφαλής, δηλαδή 64. Η μαθηματική συνάρτηση softmax χρησιμοποιείται για να προκύψουν τα βάρη προσοχής (attention weights). Αυτά τα βάρη καθορίζουν τη σημασία της τιμής (Value) κάθε χρονικού σημείου για το τρέχον χρονικό σημείο. Έπειτα, όλα τα heads συνενώνονται σε έναν πίνακα H, ο οποίος πολλαπλασιάζεται με τον πίνακα παραμέτρων W για να προκύψει τελικά ο τελικός πίνακας προσοχής πολλαπλών κεφαλών που δείχνει τα βάρη ή αλλιώς την συσχέτιση μεταξύ δύο χρονικών σημείων.

Η έξοδος του Multi-Head Attention προστίθεται στην αρχική είσοδο και το αποτέλεσμα αυτών υπόκεινται σε διαδικασία κανονικοποίησης κατά την διάσταση των 384 χαρακτηριστικών (Layer Normalization), διατηρώντας τη σταθερότητα των αναπαραστάσεων των εγκεφαλικών σημάτων. Τελευταίο στάδιο είναι ένα πλήρως συνδεδεμένο δίκτυο (Feed Forward), μετασχηματίζοντας περαιτέρω τις αναπαραστάσεις των εγκεφαλικών σημάτων με στόχο την εξαγωγή πιο σύνθετων και εξειδικευμένων αναπαραστάσεων των εισόδων. Το Feed Forward Δίκτυο περιλαμβάνει δύο στρώματα νευρώνων με μια ReLU ανάμεσά τους, η οποία επιτρέπει στο δίκτυο να μοντελοποιήσει πιο σύνθετες και μη γραμμικές σχέσεις μεταξύ των εισόδων και των εξόδων. Η διαδικασία αυτή επαναλαμβάνεται 4 φορές καθώς όπως αναφέρθηκε παραπάνω χρησιμοποιούνται 4 Transformer Encoders, τα οποία συνθέτουν τον "Whisper Encoder". Η έξοδος του αποκωδικοποιητή είναι οι τελικές αναπαραστάσεις (embeddings) των εγκεφαλικών σημάτων.

3.3 Σχεδιασμός μοντέλου

Στην ενότητα αυτή, περιγράφεται και αναλύεται η αρχιτεκτονική του μοντέλου που πρόκειται να εκπαιδευτεί. Κατά τη διαδικασία σχεδιασμού του μοντέλου μετάφρασης MEG σημάτων σε κείμενο, πριν την τελική επιλογή της αρχιτεκτονικής του μοντέλου που πρόκειται να αναλυθεί παρακάτω, πραγματοποιήθηκε αρχικά εκπαίδευση ενός μοντέλου που η αρχιτεκτονική του βασίζεται στην αρχιτεκτονική του "Transformer", το οποίο αποτελείται από

έναν "Transformer Encoder" και από έναν "Transformer Decoder" [25]. Η αρχική επιλογή της συγκεκριμένης αρχιτεκτονικής βασίστηκε στο γεγονός ότι το "Transformer" μοντέλο έχει αποδειχθεί ιδιαίτερα αποτελεσματικό στην επίλυση προβλημάτων sequence-to-sequence, όπως η μετάφραση κειμένου από μία γλώσσα σε μία άλλη. Δεδομένης της ικανότητας του να αναλύει μακροχρόνιες εξαρτήσεις και να χειρίζεται ακολουθιακά δεδομένα, εκτιμήθηκε ότι θα μπορούσε να εφαρμοστεί και στη μετάφραση εγκεφαλικών σημάτων σε κείμενο, επιτρέποντας την αποτελεσματική εκμάθηση των συσχετίσεων μεταξύ των MEG σημάτων και των αντίστοιχων γλωσσικών αναπαραστάσεων. Αξίζει να αναφερθεί ότι έχει καθιερωθεί ως η βάση πολλών σύγχρονων γλωσσικών μοντέλων ("BART", "GPT", "Whisper"), τα οποία έχουν επιτύχει αξιοσημείωτα αποτελέσματα σε διάφορες φυσικές γλωσσικές διεργασίες.

Σε πρώτο στάδιο, τα MEG σήματα υποβάλλονταν σε επεξεργασία μέσω δύο συνελικτικών επιπέδων και έπειτα σε ένα στρώμα "Transformer Encoder" εξάγοντας χαρακτηριστικά των σημάτων. Έπειτα, η έξοδος του κωδικοποιητή εισάγονταν στην είσοδο ενός στρώματος "Transformer Decoder" σε συνδυασμό με τα tokens των προτάσεων. Το λεξιλόγιο που δημιουργούνταν για τις προβλέψεις αποτελούνταν μόνο από τις λέξεις που συναντούσε το μοντέλο κατά την εκπαίδευση. Παρότι οι προβλέψεις του μοντέλου ήταν ενθαρρυντικές καθώς σε σημαντικό βαθμό αντιστοιχούσαν στις προτάσεις που σχετίζονταν με τα εγκεφαλικά σήματα, το μοντέλο παρουσίαζε φαινόμενα υπερεκπαίδευσης. Η υπερεκπαίδευση πιθανώς οφειλόταν στο γεγονός ότι το μοντέλο απομνημόνευε τα δεδομένα εκπαίδευσης αντί να μαθαίνει να γενικεύει. Αυτό ενδέχεται να συνέβαινε λόγω ανεπαρκούς ανάλυσης των δεδομένων MEG και μη ορθής σύνδεσης τους με το αντίστοιχο κείμενο.

Η αρχική προσέγγιση με τον βασικό Transformer αποκάλυψε προκλήσεις όπως την υπερεκπαίδευση και την απομνημόνευση των δεδομένων εκπαίδευσης, καθιστώντας αναγκαία τη μετάβαση σε ένα μοντέλο με πιο ισχυρή γενίκευση. Επόμενη προσέγγιση για την μετάφραση των MEG σημάτων σε κείμενο περιλάμβανε την χρήση του "Whisper" μοντέλου, το οποίο χρησιμοποιείται και στην αρχιτεκτονική του "NeuSpeech" μοντέλου. Αρχικά, πραγματοποιήθηκαν προσπάθειες απλής αναπαραγωγής του "NeuSpeech" μοντέλου. Εξαιτίας όμως της πολυπλοκότητας του συγκεκριμένου μοντέλου και των περιορισμένων υπολογιστικών πόρων του συστήματος μας, το "NeuSpeech" μοντέλο διαχωρίστηκε σε δύο μέρη με απώτερο σκοπό την βελτιστοποίηση της υπολογιστικής πολυπλοκότητας του. Συγκεκριμένα, το τμήμα της κωδικοποίησης ενσωματώθηκε ως μέρος της προεπεξεργασίας των δεδομένων, ενώ για την τελική εκπαίδευση χρησιμοποιήθηκε το τμήμα του αποκωδικοποιητή. Μετά την εφαρμογή της συγκεκριμένης αρχιτεκτονικής του αποκωδικοποιητή, δεν πραγματοποιήθηκαν δοκιμές με άλλες αρχιτεκτονικές προεκπαιδευμένων μοντέλων, καθώς η ανάλυση επικεντρώθηκε αποκλειστικά στα στοιχεία που συνθέτουν το μοντέλο "NeuSpeech". Αυτή η προσέγγιση επιλέχθηκε προκειμένου να διασφαλιστεί μια πιο ακριβής και συγκρίσιμη αξιολόγηση των αποτελεσμάτων, εστιάζοντας στις επιδόσεις του "NeuSpeech" μοντέλου για την επίλυση του προβλήματος της μετάφρασης MEG σημάτων σε κείμενο.

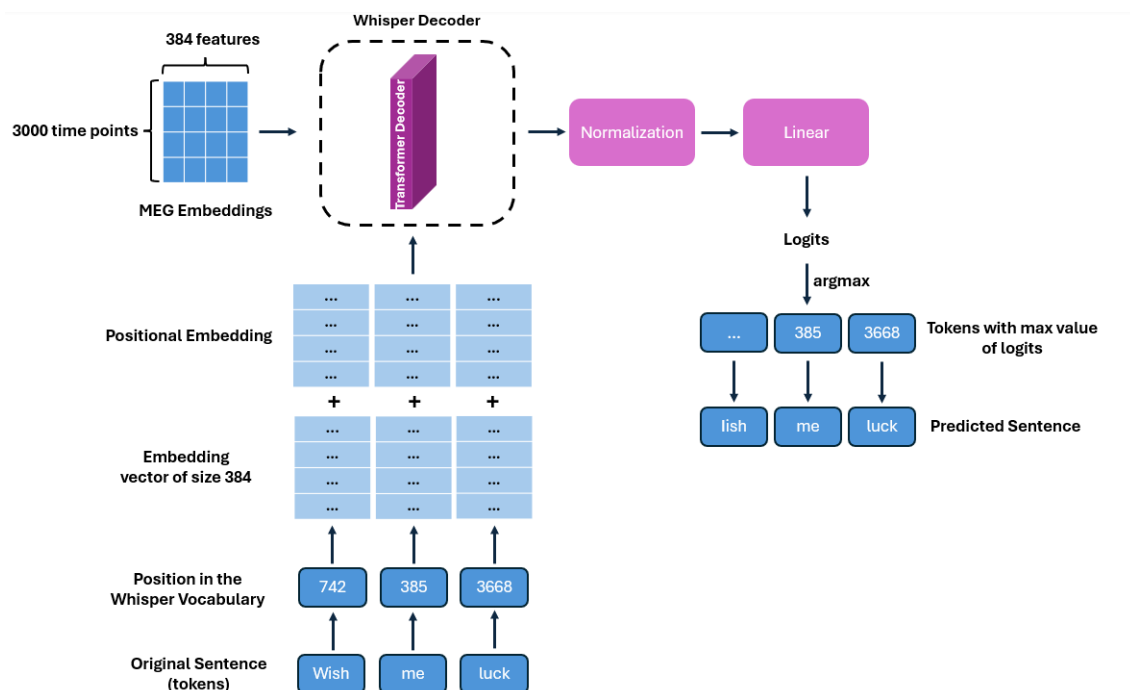
Αναφορικά με την αρχιτεκτονική του μοντέλου, στον αποκωδικοποιητή εισάγονται τα MEG embeddings που προέκυψαν από τον κωδικοποιητή που αξιοποιήθηκε κατά το στάδιο της προεπεξεργασίας των δεδομένων καθώς και οι πραγματικές προτάσεις που αντιστοιχίστηκαν με τα ανάλογα MEG embeddings.

Όσον αφορά τις προτάσεις, μία συνήθης τακτική σε γλωσσικά μοντέλα που βασίζονται

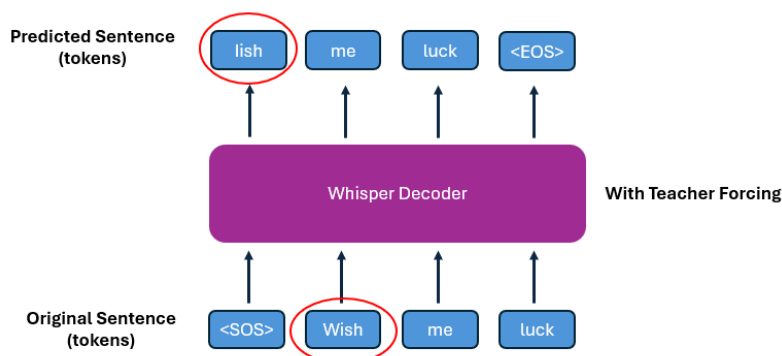
σε "Transformer" είναι ο διαχωρισμός τους αρχικά σε λέξεις ή διαφορετικά tokens. Όπως στο παράδειγμα της εικόνας 3.13, η πρόταση "Wish me luck" χωρίζεται στις λέξεις: "Wish", "me" και "luck". Χρησιμοποιώντας τον Whisper Tokenizer, τα tokens μετατρέπονται σε αριθμούς που είναι δείκτες σε λέξεις που περιέχονται στο Whisper λεξιλόγιο. Το συγκεκριμένο λεξιλόγιο δημιουργήθηκε κατά την εκπαίδευση του πολυγλωσσικού "Whisper" μοντέλου και περιέχει λέξεις ή τμήματα λέξεων, σημεία στίξης καθώς και ειδικά σύμβολα. Δύο από αυτά για παράδειγμα σηματοδοτούν την έναρξη και την λήξη της πρότασης. Κάθε λέξη, σημείο στίξης, ειδικό σύμβολο ή οτιδήποτε άλλο περιέχεται μέσα στο λεξιλόγιο, κατέχει το δικό του αριθμητικό id, που τελικά αυτό χρησιμοποιείται για το tokenization μίας λέξης της πρότασης. Εάν μία λέξη βρίσκεται παραπάνω από μία φορά σε μία πρόταση, κατά τον μετασχηματισμό της σε token λαμβάνει τον ίδιο αριθμό καθώς βρίσκεται στην ίδια θέση του λεξιλογίου. Το μέγεθος του λεξιλογίου είναι 51865.

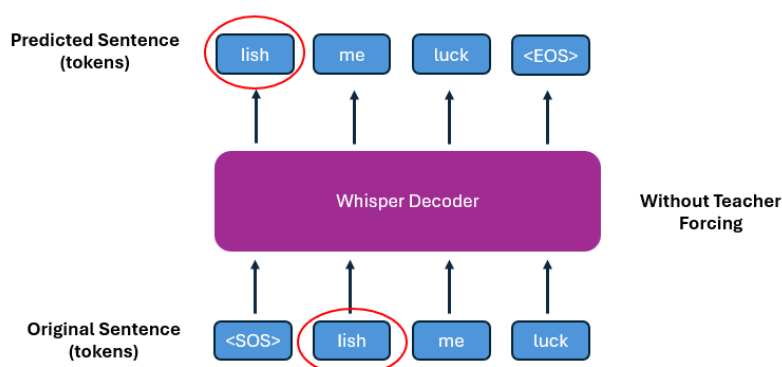
Έπειτα, κάθε token αντιστοιχίζεται σε ένα διάνυσμα μεγέθους 384 (Embedding), σύμφωνα με τις προδιαγραφές του "Whisper" αποκωδικοποιητή και έπειτα προστίθεται με ένα άλλο διάνυσμα ίδιου μεγέθους (για να μπορέσει να γίνει η πράξη της πρόσθεσης). Το δεύτερο διάνυσμα σχετίζεται με την θέση κάθε λέξης στην πρόταση (Positional Embedding). Ενώ το πρώτο διάνυσμα αποτυπώνει το νόημα της κάθε λέξης, το δεύτερο δίνει πληροφορίες για την θέση των λέξεων, δηλαδή πόσο κοντά ή μακριά βρίσκεται η μία από την άλλη. Μπορεί εμείς οπτικά να έχουμε την δυνατότητα να την διακρίνουμε αλλά το μοντέλο δεν μπορεί να την γνωρίζει. Το αποτέλεσμα της πρόσθεσης μεταξύ των δύο διανυσμάτων εισάγεται τελικά μαζί με τα MEG embeddings στον αποκωδικοποιητή, ο οποίος αποτελείται από 1 επίπεδο "Transformer Decoder". Έπειτα, τα αποτελέσματα των διεργασιών αυτών υπόκεινται ένα είδος κανονικοποίησης και περνούν από ένα γραμμικό στρώμα (Linear) από το οποίο εξάγονται τιμές, τα λεγόμενα logits. Τα logits στην πραγματικότητα είναι ένας πίνακας μεγέθους (80, 51865). Ο αριθμός 80 αντιστοιχεί στο μέγιστο μήκος που μπορεί να λάβει η πρόταση (αριθμός λέξεων), ενώ ο αριθμός 51865 αντιστοιχεί στο μέγεθος του λεξιλογίου. Ο πίνακας δηλαδή περιέχει 51865 τιμές (στήλες) για κάθε token (γραμμές) της πρότασης. Το κελί του πίνακα που περιέχει την μεγαλύτερη τιμή, αντικατοπτρίζει το token που προβλέπει το μοντέλο ότι είναι την δεδομένη στιγμή. Καθίσταται αναγκαίο να αναφερθεί πως ο αρχικός αποκωδικοποιητής αποτελείται από 4 επίπεδα Transformer Decoders και το μέγιστο μήκος της ακολουθίας είναι 448, ωστόσο επιλέχθηκαν για τον αποκωδικοποιητή οι παραπάνω τιμές των παραμέτρων για απλοποίηση της αρχιτεκτονικής, μείωση των υπολογιστικών απαιτήσεων καθώς και εξοικονόμηση χρόνου. Η λιγότερο βαθιά αρχιτεκτονική του "Whisper Decoder" (από τέσσερα επίπεδα Transformer Decoders σε ένα) συμβάλλει επίσης, στη μείωση της πιθανότητας υπερεκπαίδευσης και επιταχύνει τη διαδικασία σύγκλισης του μοντέλου. Έπειτα από πειραματικές δοκιμές, διαπιστώθηκε ότι η αύξηση του αριθμού των επιπέδων Transformer Decoders δεν οδήγησε σε βελτίωση της απόδοσης, αλλά συνέβαλε στην ενίσχυση των φαινομένων υπερεκπαίδευσης, ενώ παράλληλα αύξησε τις υπολογιστικές απαιτήσεις του μοντέλου.

Μία λεπτομέρεια που αξίζει να σημειωθεί για το μοντέλο είναι η χρήση μάθησης με καθοδήγηση ή ως γνωστόν Teacher Forcing. Κατά την διάρκεια εκπαίδευσης, το μοντέλο αντί να χρησιμοποιεί την πρόβλεψη του από το προηγούμενο βήμα ως είσοδο για το επόμενο βήμα, λαμβάνει την πραγματική τιμή από τα δεδομένα εκπαίδευσης. Για παράδειγμα, η δευτε-

Εικόνα 3.13: *Whisper Decoder*

ρη λέξη μίας πρότασης που προβλέπεται την τρέχουσα χρονική στιγμή, δεν βασίζεται στην πρόβλεψη της λέξης την προηγούμενη ακριβώς χρονική στιγμή, αλλά στην πραγματική λέξη της προηγούμενης χρονικής στιγμής. Η τεχνική αυτή εξασφαλίζει ότι το μοντέλο μαθαίνει με πιο αποτελεσματικό τρόπο, καθώς βασίζεται σε έγκυρα δεδομένα και όχι σε ενδεχομένως λανθασμένες προβλέψεις, οι οποίες θα μπορούσαν να επηρεάσουν αρνητικά την εκπαίδευση του μοντέλου. Ακολουθεί το παρακάτω παράδειγμα 3.14 για καλύτερη κατανόηση αυτής της μεθόδου. Αντίθετα, αν το μοντέλο δεν χρησιμοποιούσε Teacher Forcing, τότε η πρόβλεψη του προηγούμενου βήματος θα αποτελούσε είσοδο για την πρόβλεψη του επόμενου χρονικού βήματος, όπως φαίνεται στο παράδειγμα 3.15.

Εικόνα 3.14: *With Teacher Forcing*

Εικόνα 3.15: *Without Teacher Forcing*

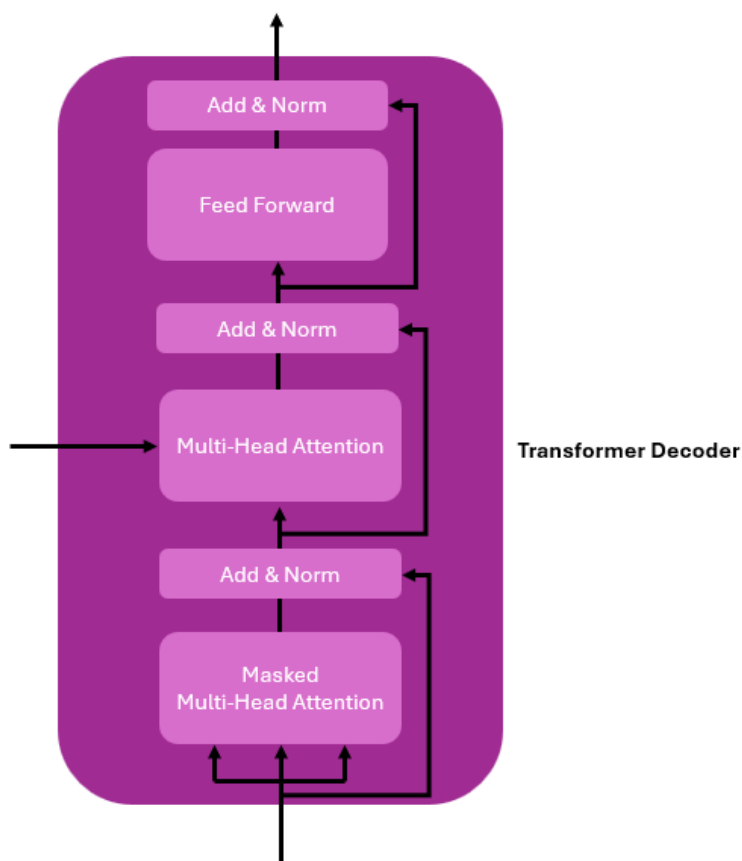
Transformer Decoder

Ο Transformer Decoder είναι το δεύτερο μέρος από το οποίο αποτελείται ο βασικός Transformer [25] και έχει παρόμοια δομή με τον Transformer Encoder. Σε αυτό που διαφέρει και αξίζει να αναφερθεί είναι στο επίπεδο Masked Multi-Head Attention. Το συγκεκριμένο επίπεδο περιλαμβάνει το Multi-Head Attention με τον ίδιο ακριβώς τρόπο όπως στον κωδικοποιητή αλλά με την εξής διαφορά: την χρήση μάσκας. Η μάσκα αυτή στην ουσία εμποδίζει το μοντέλο, κατά την διάρκεια που κάνει πρόβλεψη για μία συγκεκριμένη λέξη, να έχει πρόσβαση σε μελλοντικές λέξεις. Δηλαδή, το μοντέλο μπορεί να προβλέψει την επόμενη λέξη, βλέποντας μόνο τις προηγούμενες λέξεις της πρότασης αλλά όχι τις επόμενες. Χωρίς αυτή την μάσκα, το μοντέλο θα είχε πρόσβαση σε όλη την πρόταση και κατά την διάρκεια της πρόβλεψης μίας λέξης θα μπορούσε να δει ποια είναι η πραγματική λέξη που προσπαθεί να προβλέψει. Αυτό ωστόσο, δεν είναι ορθό για την εκπαίδευση του μοντέλου. Στην εικόνα 3.16 απεικονίζεται η δομή και τα στοιχεία από τα οποία αποτελείται ο Transformer Decoder.

3.4 Εκπαίδευση μοντέλου

Κατά την διαδικασία εκπαίδευσης του αποκωδικοποιητή, υπολογίζεται το σφάλμα (loss), με τη χρήση της Cross-Entropy Loss, συγκρίνοντας τις προβλέψεις με τα πραγματικά δεδομένα. Το σφάλμα χρησιμοποιείται για την ενημέρωση των βαρών του μοντέλου μέσω της αντίστροφης διάδοσης (backpropagation) και δείχνει κατά πόσο είναι σωστή η πρόβλεψη του μοντέλου για την εκάστοτε πρόταση που αντιστοιχεί σε ένα τμήμα εγκεφαλικού σήματος. Με την ενημέρωση των βαρών, στόχος είναι η ελαχιστοποίηση του σφάλματος ώστε τελικά οι προβλέψεις να πλησιάζουν όσο το δυνατόν γίνεται στα πραγματικά δεδομένα. Για την ενημέρωση των παραμέτρων του μοντέλου χρησιμοποιείται ο βελτιστοποιητής AdamW, όπως και στο άρθρο [11]. Ακόμη, το βήμα με το οποίο ενημερώνονται τα βάρη του μοντέλου μας μετά από κάθε υπολογισμό παραγώγων (backpropagation) είναι 0,0001 και η τιμή του αντιστοιχεί στην παράμετρο learning rate.

Μετά από κάθε εποχή εκπαίδευσης, γίνεται αξιολόγηση του μοντέλου με τα δεδομένα επικύρωσης (validation set). Εάν η απώλεια επικύρωσης (validation loss) μειώνεται, το μοντέλο αποθηκεύεται ως η καλύτερη έκδοση. Σε περίπτωση που η απώλεια δεν βελτιώνεται για έναν καθορισμένο αριθμό εποχών και συγκεκριμένα για 3 συνεχόμενες εποχές (early stop-

Εικόνα 3.16: *Transformer Decoder*

ping), γίνεται διακοπή της εκπαίδευσης για να αποφευχθεί η πιθανότητα υπερεκπαίδευσης του μοντέλου (overfitting). Η εκπαίδευση ολοκληρώθηκε έπειτα από 7 εποχές και διήρκησε περίπου 13 ώρες.

3.5 Αξιολόγηση μοντέλου

Για την εκτίμηση της απόδοσης του εκπαιδευμένου μοντέλου, χρησιμοποιείται το σύνολο ελέγχου (test set), το οποίο περιέχει δεδομένα που δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση ή την επικύρωση του μοντέλου. Με αυτόν τον τρόπο, επιτυγχάνεται η αξιολόγηση σχετικά με το πόσο καλά το μοντέλο γενικεύει σε άγνωστα δεδομένα. Το test set όπως αναφέρθηκε σε προηγούμενη ενότητα, περιέχει 2918 ζευγάρια δεδομένων “MEG τμήματα εγκεφαλικών σημάτων - αντίστοιχη πρόταση”.

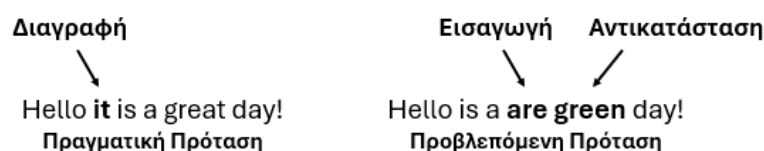
Για την αξιολόγηση των προβλέψεων του μοντέλου επιλέγονται οι μετρικές: WER (Word Error Rate), CER (Character Error Rate), BLEU-N (Bilingual Evaluation Understudy) και ROUGE-N (Recall-Oriented Understudy for Gisting Evaluation). Επιλέχθηκαν οι συγκεκριμένες μετρικές καθώς παρέχουν μια ολοκληρωμένη εκτίμηση της απόδοσης του μοντέλου σε διαφορετικά επίπεδα γλωσσικής ανάλυσης. Η WER και η CER μετρική είναι οι πιο διαδεδομένες μετρικές για την αξιολόγηση της απόδοσης συστημάτων αυτόματης αναγνώρισης ομιλίας (ASR) και επεξεργασίας της φυσικής γλώσσας (NLP), μετρώντας σφάλματα σε επίπε-

δο λέξεων και χαρακτήρων αντίστοιχα. Στη παρούσα έρευνα, είναι ιδιαίτερα χρήσιμες καθώς με αυτές αποκαλύπτονται πιθανά σφάλματα στα παραγόμενα κείμενα του μοντέλου τόσο σε επίπεδο λέξεων όσο και σε επίπεδο χαρακτήρων. Παράλληλα, η μετρική BLEU-N και ROUGE-N είναι δύο από τις πιο κοινές μεθόδους αξιολόγησης της ποιότητας των κειμένων που παράγονται από συστήματα μετάφρασης, περίληψης και αναγνώρισης κειμένων. Είναι ιδιαίτερα χρήσιμες για την αξιολόγηση του προτεινόμενου μοντέλου καθώς παρέχουν μετρήσεις για την ποιότητα των παραγόμενων κειμένων σε σχέση με των πραγματικών.

Καθίσταται αναγκαίο να αναφερθεί ότι, υπάρχουν και άλλες αξιολογικές μετρικής αξιολόγησης γλωσσικών μοντέλων, όπως η MER [26] που είναι παρόμοια της WER και η METEOR που είναι παρόμοια της BLEU-N και της ROUGE-N. Η METEOR [27] είναι μία πιο ευέλικτη μετρική καθώς λαμβάνει υπόψιν της όχι μόνο την ακρίβεια των N-grams μεταξύ των προβλεπόμενων και πραγματικών προτάσεων αλλά και εάν οι λέξεις είναι συνώνυμες. Ωστόσο, στη παρούσα εργασία επιλέχθηκαν οι παραπάνω μετρικές καθώς είναι ευρέως χρησιμοποιούμενες στην βιβλιογραφία. Με την επιλογή αυτών των μετρικών επιτυγχάνεται μια πιο αντικειμενική και αξιόπιστη σύγκριση των αποτελεσμάτων του προτεινόμενου μοντέλου με το "NeuSpeech", διευκολύνοντας την κατανόηση της συμβολής του παρόντος μοντέλου.

3.5.1 WER μετρική

Σύμφωνα με την Hugging Face [28], η συγκεκριμένη μετρική συγκρίνει λέξη προς λέξη την πραγματική πρόταση που αντιστοιχεί σε ένα σήμα εγκεφάλου με την πρόταση που προβλέπεται από το μοντέλο, προκειμένου να εντοπιστεί ο αριθμός των διαφορών μεταξύ τους. Οι δύο προτάσεις μπορεί να διαφέρουν α) εάν μία λέξη είναι μέρος της πραγματικής πρότασης αλλά όχι της προβλεπόμενης (Διαγραφή - Deletion), β) εάν μία λέξη δεν βρίσκεται στην πραγματική πρόταση αλλά έχει προστεθεί στη προβλεπόμενη (Εισαγωγή - Insertion) και γ) εάν μία λέξη της πραγματικής πρότασης αντικαθίστανται από μία άλλη στην προβλεπόμενη πρόταση (Αντικατάσταση - Substitution). Ακολουθεί το παρακάτω παράδειγμα στην εικόνα 3.17 που δόθηκε στο άρθρο του Ketan Doshi [29], για καλύτερη κατανόηση της Διαγραφής, Εισαγωγής και Αντικατάστασης.



Εικόνα 3.17: Διαγραφή, Εισαγωγή και Αντικατάσταση μεταξύ της πραγματικής και της προβλεπόμενης πρότασης. Η εικόνα είναι προσαρμοσμένη από 3.17.

Συγκρίνοντας την πραγματική και την προβλεπόμενη πρόταση του παραδείγματος, κατανοούμε ότι κατά την πρόβλεψη που έκανε το εκάστοτε μοντέλο, δεν κατάφερε να προβλέψει σωστά την επόμενη ακριβώς λέξη της λέξης "Hello", που είναι η λέξη "it" (Διαγραφή). Ακόμη, αντικατέστησε την λέξη "great" με την λέξη "green" (Αντικατάσταση), ενώ εισήγαγε μία λέξη που δεν παρατηρείται στην πραγματική πρόταση, την λέξη "are" (Εισαγωγή).

Η WER μετρική υπολογίζει το ποσοστό των λέξεων που χρειάζεται να αντικατασταθούν,

να διαγραφούν ή να προστεθούν ώστε η προβλεπόμενη πρόταση να ταυτίζεται με την πραγματική. Είναι μία μετρική που μετρά τα σφάλματα. Το εύρος των τιμών της είναι $[0,1]$ ή σε ποσοστιαία μορφή 0% έως 100%. Όσο πιο μικρή είναι η τιμή της, τόσο καλύτερη είναι η απόδοση του μοντέλου.

Για τον υπολογισμό της WER μετρικής, υπολογίζεται ο αριθμός των λέξεων που δεν πρόβλεψε το μοντέλο ενώ είναι μέρος της πραγματικής πρότασης (Deleted), ο αριθμός των λέξεων που πρόβλεψε το μοντέλο αλλά δεν είναι μέρος της πραγματικής πρότασης (Inserted) και ο αριθμός των λέξεων που αντικαταστάθηκαν κατά την πρόβλεψη από μία άλλη (Substituted). Έπειτα, όλοι οι αριθμοί προστίθενται μεταξύ τους και το αποτέλεσμα αυτών διαιρείται με τον συνολικό αριθμό των λέξεων της πραγματικής πρότασης. Παρακάτω παρουσιάζεται ο υπολογισμός της μετρικής στο συγκεκριμένο παράδειγμα.

$$WER = \frac{\text{Deleted} + \text{Inserted} + \text{Substituted}}{\text{Total words in original sentence}} = \frac{1 + 1 + 1}{6} = 0.5$$

Εικόνα 3.18: Υπολογισμός της WER μετρικής. Στο παράδειγμα, εμφανίζεται σφάλμα 50% κατά την πρόβλεψη της πρότασης.

Παρατηρώντας την εικόνα 3.17 που απεικονίζει το παράδειγμα και λαμβάνοντας υπόψιν τον τρόπο με τον οποίο υπολογίζεται η συγκεκριμένη μετρική, συμπεραίνουμε ότι μία λέξη (η λέξη "it") δεν υπάρχει στην προβλεπόμενη πρόταση, μία λέξη (η λέξη "are" προστέθηκε χωρίς να υπάρχει στη πραγματική πρόταση) και μία λέξη (η λέξη great) αντικαταστάθηκε κατά την πρόβλεψη με την λέξη "green". Ο συνολικός αριθμός των λέξεων που διαγράφηκαν, αντικαταστάθηκαν και προστέθηκαν κατά την πρόβλεψη είναι 3 και διαιρείται με τον συνολικό αριθμό λέξεων που συνθέτουν την πραγματική πρόταση, δηλαδή διαιρείται με τον αριθμό 6. Το αποτέλεσμα της πράξης είναι ο αριθμός 0,5 γεγονός που υποδεικνύει το μοντέλο απέτυχε να προβλέψει σωστά το 50% των λέξεων σε αυτήν την πρόταση.

3.5.2 CER μετρική

Η μετρική CER [30] λειτουργεί με παρόμοιο τρόπο με την μετρική WER καθώς υπολογίζει το ποσοστό των σφαλμάτων, όχι σε επίπεδο λέξεων αλλά σε επίπεδο χαρακτήρων, συγκρίνοντας την πραγματική με την προβλεπόμενη πρόταση. Η συγκεκριμένη μετρική δίνει μια πιο λεπτομερή εικόνα του σφάλματος καθώς μελετά τα σφάλματα σε επίπεδο χαρακτήρων. Το εύρος των τιμών της δεν είναι απαραίτητα πάντα $[0,1]$ ή σε ποσοστιαία μορφή 0% έως 100%, ειδικά σε περιπτώσεις που παρατηρείται αυξημένος αριθμός χαρακτήρων που εισάγονται κατά την πρόβλεψη. Χαμηλότερες τιμές της μετρικής υποδηλώνουν καλύτερη απόδοση του μοντέλου, με ιδανική ακρίβεια να επιτυγχάνεται όταν η τιμή της CER προσεγγίζει το μηδέν. Αυτό σημαίνει ότι το μοντέλο παράγει μια πρόταση ταυτόσημη με την πραγματική, χωρίς σφάλματα εισαγωγής, διαγραφής ή αντικατάστασης χαρακτήρων.

Για την καλύτερη κατανόηση αυτής της μετρικής, αναλύεται ακολούθως ο τρόπος υπολογισμού της. Υποθέτουμε ότι η πραγματική και η προβλεπόμενη πρόταση για ένα εγκεφαλικό τμήμα κάποιου συμμετέχοντα είναι η εξής:

- Πραγματική Πρόταση: "Hello it is a great day!"
- Προβλεπόμενη Πρόταση: "Hello **its** is a **grea** day!"

Η μετρική υπολογίζεται με τον εξής τρόπο:

$$\text{CER} = \frac{\text{Deleted} + \text{Inserted} + \text{Substituted}}{\text{Number of characters in original sentence}} = \frac{1 + 1 + 0}{24} = \frac{2}{24}$$

Παρατηρώντας το παράδειγμα που αναφέρεται παραπάνω, ο αριθμός των χαρακτήρων που είναι μέρος της πραγματικής αλλά όχι της προβλεπόμενης πρότασης (Deleted) είναι 1 (το γράμμα "t" από την λέξη "great"), ο αριθμός των χαρακτήρων που είναι μέρος της προβλεπόμενης αλλά όχι της πραγματικής πρότασης (Inserted) είναι πάλι 1 (το γράμμα "s" στη λέξη "it"), ενώ δεν παρατηρείται κάποια αντικατάσταση (Substituted) στο συγκεκριμένο παράδειγμα. Για τον υπολογισμό των συνολικών χαρακτήρων, πέραν των γραμμάτων συμπεριλαμβάνονται και οι κενοί χαρακτήρες καθώς και τα σημεία στίξης.

3.5.3 BLEU-N μετρική

Η BLEU-N μετρική [29] μετρά την ομοιότητα μεταξύ της πραγματικής και της προβλεπόμενης πρότασης κατά N-grams, όπου N-grams είναι ο αριθμός των λέξεων μίας πρότασης. Για παράδειγμα, εάν η πρόταση είναι: "The ball is bleu", τα N-grams παρουσιάζονται ως εξής:

- 1-gram (unigram): "The", "ball", "is", "bleu"
- 2-grams (bigram): "The ball", "ball is", "is bleu"
- 3-grams (trigram): "The ball is", "ball is bleu"
- 4-grams: "The ball is bleu"

Να αναφερθεί όμως ότι οι λέξεις πάντα λαμβάνονται με την σειρά, όπως είναι διατεταγμένες στην πρόταση. Δηλαδή, το "blue is The ball" δεν μπορεί να είναι έγκυρο.

Το εύρος των τιμών που μπορεί να πάρει η μετρική είναι [0,1] ή σε ποσοστιαία μορφή 0% έως 100%. Όσο πιο μεγάλη είναι η τιμή της, τόσο καλύτερη είναι η απόδοση του μοντέλου.

Για την καλύτερη κατανόηση αυτής της μετρικής, αναλύεται ακολούθως ο τρόπος υπολογισμού της. Υποθέτουμε ότι η πραγματική και η προβλεπόμενη πρόταση για ένα εγκεφαλικό τμήμα κάποιου συμμετέχοντα είναι η εξής:

- Πραγματική Πρόταση: "The guard arrived late because it was raining"
- Προβλεπόμενη Πρόταση: "The guard arrived late because of the rain"

Αρχικά, υπολογίζεται η μετρική "Precision" και για τα 4-grams, όπως φαίνεται παρακάτω:

1. Precision 1-gram

Η εικόνα 3.19, απεικονίζει την σύγκριση της πραγματικής με την προβλεπόμενη πρόταση κατά 1-gram, δηλαδή λέξη προς λέξη. Οι πρώτες πέντε λέξεις ("The", "guard", "arrived", "late", "because") προβλέπονται σωστά. Αντίθετα, προβλέπεται λανθασμένα η λέξη "of" αντί για την λέξη "it", η λέξη "the" αντί για την λέξη "was" και τέλος, η λέξη "raining" αντί για την λέξη "rain".



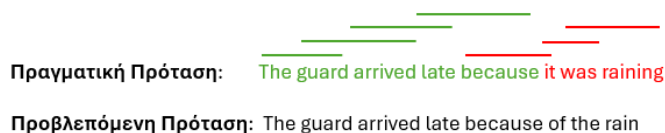
Εικόνα 3.19: Σύγκριση πραγματικής και προβλεπόμενης πρότασης κατά 1-gram, δηλαδή λέξη προς λέξη. Οι λέξεις που εμφανίζονται με πράσινο χρώμα υποδεικνύουν σωστές προβλέψεις, ενώ οι λέξεις που εμφανίζονται με κόκκινο χρώμα αντιπροσωπεύουν τις πραγματικές λέξεις που θα έπρεπε να προβλεφθούν αλλά προβλέφθηκαν λιάθος. Η εικόνα είναι προσαρμοσμένη από [29].

Η μετρική "Precision" κατά 1-gram υπολογίζεται διαιρώντας τον αριθμό των σωστά προβλεπόμενων λέξεων με τον συνολικό αριθμό των λέξεων της προβλεπόμενης πρότασης. Στο συγκεκριμένο παράδειγμα, ο αριθμός των σωστά προβλεπόμενων λέξεων είναι 5 ("The", "guard", "arrived", "late", "because") και ο συνολικός αριθμός των λέξεων της προβλεπόμενης πρότασης είναι 8 ("The", "guard", "arrived", "late", "because", "of", "the", "rain"). Ο τρόπος υπολογισμού παρουσιάζεται ακολούθως:

$$\text{Precision}_{1\text{-gram}} = p1 = \frac{\text{Number of correct predicted 1-gram}}{\text{Number of total predicted 1-gram}} = \frac{5}{8}$$

2. Precision 2-grams

Η εικόνα 3.20, απεικονίζει την σύγκριση της πραγματικής με την προβλεπόμενη πρόταση κατά 2-grams, δηλαδή ανά δύο λέξεις. Τα πρώτα τέσσερα ζευγάρια λέξεων ("The guard", "guard arrived", "arrived late", "late because") προβλέπονται σωστά. Αντίθετα, τα τελευταία τρία ζευγάρια λέξεων ("because it", "it was", "was raining") προβλέπονται λανθασμένα από τα ζευγάρια λέξεων "because of", "of the", "the rain" αντίστοιχα.



Εικόνα 3.20: Σύγκριση πραγματικής και προβλεπόμενης πρότασης κατά 2-grams, δηλαδή σύγκριση ανά δύο λέξεις. Τα ζευγάρια των λέξεων που εμφανίζονται με πράσινο χρώμα υποδεικνύουν σωστές προβλέψεις, ενώ εκείνα που εμφανίζονται με κόκκινο χρώμα αντιπροσωπεύουν τα πραγματικά ζευγάρια λέξεων που θα έπρεπε να προβλεφθούν αλλά προβλέφθηκαν λιάθος. Η εικόνα είναι προσαρμοσμένη από [29].

Ομοίως, η μετρική "Precision" κατά 2-grams υπολογίζεται διαιρώντας τον αριθμό των

σωστά προβλεπόμενων λέξεων με τον συνολικό αριθμό των λέξεων της προβλεπόμενης πρότασης, κατά 2-grams. Στο συγκεκριμένο παράδειγμα, ο αριθμός των σωστά προβλεπόμενων ανά δύο λέξεων είναι 4 ("The guard", "guard arrived", "arrived late", "late because") και ο συνολικός αριθμός των ανά δύο λέξεων της προβλεπόμενης πρότασης είναι 7 ("The guard", "guard arrived", "arrived late", "late because", "because of", "of the", "the rain"). Ο τρόπος υπολογισμού παρουσιάζεται ακολούθως:

$$\text{Precision}_{2\text{-grams}} = p2 = \frac{\text{Number of correct predicted 2-grams}}{\text{Number of total predicted 2-grams}} = \frac{4}{7}$$

3. Precision 3-grams

Η εικόνα 3.21, απεικονίζει την σύγκριση της πραγματικής με την προβλεπόμενη πρόταση κατά 3-grams, δηλαδή ανά τρεις λέξεις. Οι πρώτοι τρεις συνδυασμοί λέξεων ("The guard arrived", "guard arrived late", "arrived late because") προβλέπονται σωστά. Αντίθετα, οι τελευταίοι τρεις συνδυασμοί λέξεων ("late because it", "because it was", "it was raining") προβλέπονται λανθασμένα από το "late because of", "because of the", "of the rain" αντίστοιχα.

Πραγματική Πρόταση: The guard arrived late because it was raining

Προβλεπόμενη Πρόταση: The guard arrived late because of the rain

Εικόνα 3.21: Σύγκριση πραγματικής και προβλεπόμενης πρότασης κατά 3-grams, δηλαδή σύγκριση ανά τρεις λέξεις. Οι τριάδες λέξεων που εμφανίζονται με πράσινο χρώμα υποδεικνύουν σωστές προβλέψεις, ενώ εκείνες που εμφανίζονται με κόκκινο χρώμα αντιπροσωπεύουν τις πραγματικές τριάδες λέξεων που θα έπρεπε να προβλεφθούν αλλά προβλέφθηκαν λάθος. Η εικόνα είναι προσαρμοσμένη από [29].

Η μετρική "Precision" κατά 3-grams υπολογίζεται με παρόμοιο τρόπο, διαιρώντας τον αριθμό των σωστά προβλεπόμενων λέξεων με τον συνολικό αριθμό των λέξεων της προβλεπόμενης πρότασης, κατά 3-grams. Στο συγκεκριμένο παράδειγμα, ο αριθμός των σωστά προβλεπόμενων λέξεων, ανά τρεις λέξεις, είναι 3 ("The guard arrived", "guard arrived late", "arrived late because") και ο συνολικός αριθμός των λέξεων της προβλεπόμενης πρότασης, ανά τριάδες, είναι 6 ("The guard arrived", "guard arrived late", "arrived late because", "late because of", "because of the", "of the rain"). Ο τρόπος υπολογισμού παρουσιάζεται ακολούθως:

$$\text{Precision}_{3\text{-grams}} = p3 = \frac{\text{Number of correct predicted 3-grams}}{\text{Number of total predicted 3-grams}} = \frac{3}{6}$$

4. Precision 4-grams

Η εικόνα 3.22, απεικονίζει την σύγκριση της πραγματικής με την προβλεπόμενη πρόταση κατά 4-grams, δηλαδή ανά τέσσερις λέξεις. Οι πρώτοι δύο συνδυασμοί λέξεων ("The guard arrived late", "guard arrived late because") προβλέπονται σωστά. Αντίθετα, οι τελευταίοι τρεις συνδυασμοί λέξεων ("arrived late because it", "late because it was", "because it was")

raining") προβλέπονται λανθασμένα από το "arrived late because of", "late because of the", "because of the rain" αντίστοιχα.



Εικόνα 3.22: Σύγκριση πραγματικής και προβλεπόμενης πρότασης κατά 4-grams, δηλαδή σύγκριση ανά τέσσερις λέξεις. Οι τετράδες λέξεων που εμφανίζονται με πράσινο χρώμα υποδεικνύουν σωστές προβλέψεις, ενώ εκείνες που εμφανίζονται με κόκκινο χρώμα αντιπροσωπεύουν τις πραγματικές τετράδες λέξεων που θα έπρεπε να προβλεφθούν αλλά προβλέφθηκαν λάθος. Η εικόνα είναι προσαρμοσμένη από [29].

Η μετρική "Precision" κατά 4-grams υπολογίζεται με παρόμοιο τρόπο, διαιρώντας τον αριθμό των σωστά προβλεπόμενων λέξεων με τον συνολικό αριθμό των λέξεων της προβλεπόμενης πρότασης, κατά 4-grams. Στο συγκεκριμένο παράδειγμα, ο αριθμός των σωστά προβλεπόμενων λέξεων, ανά τέσσερις λέξεις, είναι 2 ("The guard arrived late", "guard arrived late because") και ο συνολικός αριθμός των λέξεων της προβλεπόμενης πρότασης, πάλι ανά τέσσερις, είναι 5 ("The guard arrived late", "guard arrived late because", "arrived late because it", "late because it was", "because it was raining"). Ο τρόπος υπολογισμού παρουσιάζεται παρακάτω:

$$\text{Precision}_{4\text{-grams}} = p_4 = \frac{\text{Number of correct predicted 4-grams}}{\text{Number of total predicted 4-grams}} = \frac{2}{5}$$

Συνεπώς ο γεωμετρικός μέσος ακρίβειας (Geometric Average Precision) υπολογίζεται ως εξής:

$$\text{Geometric Average Precision (N)} = \exp\left(\sum_{n=1}^N w_n \log p_n\right) = \prod_{n=1}^N p_n^{w_n} = (p_1)^{\frac{1}{4}} \cdot (p_2)^{\frac{1}{4}} \cdot (p_3)^{\frac{1}{4}} \cdot (p_4)^{\frac{1}{4}}$$

Στη συνέχεια, υπολογίζεται ο όρος "Brevity Penalty", ο οποίος εμποδίζει το μοντέλο να παράγει σύντομες προτάσεις που ενώ φαίνονται σωστές, στην πραγματικότητα δεν αποδίδουν το πλήρες κείμενο των πραγματικών προτάσεων. Αναλυτικότερα, η μετρική precision 1-gram υπολογίζει το ποσοστό των σωστά προβλεπόμενων λέξεων που ταυτίζονται με τις πραγματικές προτάσεις κατά 1-gram. Σε περίπτωση που το μοντέλο παράγει μια πρόταση που περιλαμβάνει μία μόνο λέξη, παραδείγματος χάριν "Hello" και αυτή η λέξη ταιριάζει με την αντίστοιχη λέξη της πραγματικής πρότασης, τότε το precision 1-gram είναι $1/1 = 1$. Δηλαδή, το μοντέλο έχει 100% ακρίβεια για αυτή την πρόταση. Ωστόσο, αυτό μπορεί να είναι παραπλανητικό, διότι σε περίπτωση που το μοντέλο συνεχίζει να παράγει μόνο μία λέξη σε κάθε πρόταση, τότε το precision 1-gram θα παραμείνει πολύ υψηλό, ακόμα κι αν η πρόταση είναι ατελής. Με λίγα λόγια, το μοντέλο θα παίρνει υψηλή βαθμολογία για τη χρήση λιγότερων λέξεων, ακόμα κι αν το περιεχόμενο της πρότασης είναι ανεπαρκές. Για να αποφευχθεί αυτό το ζήτημα, χρησιμοποιείται το "Brevity Penalty" που επιβραβεύει τις πιο πλήρεις προτάσεις, αποτρέποντας το υπερβολικά μικρό μήκος των προτάσεων. Ακολουθεί η παρακάτω εξίσωση:

$$\text{Brevity Penalty} = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c \leq r \end{cases}$$

όπου c : ο αριθμός των λέξεων στην προβλεπόμενη πρόταση και r : ο αριθμός των λέξεων στη πραγματική πρόταση.

Ο τελικός τύπος για τον υπολογισμό της μετρικής BLEU-N, εμφανίζεται παρακάτω:

$$\text{Bleu-N} = \text{Brevity Penalty} \cdot \text{Geometric Average Precision Scores (N)}$$

Για να υπολογιστεί το BLEU-1, χρησιμοποιείται μόνο η τιμή του p_1 πολλαπλασιασμένη με την τιμή του Brevity Penalty. Για το BLEU-2, πολλαπλασιάζεται η τιμή του Brevity Penalty με τον γεωμετρικό μέσο όρο της ακρίβειας p_1, p_2 . Για το BLEU-3, πολλαπλασιάζεται η τιμή του Brevity Penalty με τον γεωμετρικό μέσο όρο της ακρίβειας p_1, p_2 και p_3 , ενώ για το BLEU-4 με τον γεωμετρικό μέσο όρο της ακρίβειας p_1, p_2, p_3 και p_4 .

3.5.4 ROUGE-N μετρική

Σύμφωνα με τον Fabio Chiusano [31], η μετρική ROUGE-N συγκρίνει τις προβλεπόμενες με τις αντίστοιχες πραγματικές προτάσεις κατά N-grams, υπολογίζοντας την μετρική "Recall, Precision, F1-score". Το εύρος των τιμών που μπορεί να πάρει η μετρική είναι $[0,1]$ ή σε ποσοστιαία μορφή 0% έως 100%. Όσο πιο μεγάλη είναι η τιμή της, τόσο καλύτερη είναι η απόδοση του μοντέλου.

Για την καλύτερη κατανόηση αυτής της μετρικής, αναλύεται ακολούθως ο τρόπος υπολογισμού της. Υποθέτουμε ότι η πραγματική και η προβλεπόμενη πρόταση για ένα εγκεφαλικό τμήμα κάποιου συμμετέχοντα είναι η εξής:

- Πραγματική Πρόταση: "The cat is on the mat"
- Προβλεπόμενη Πρόταση: "The cat and the dog"

Υπολογίζεται η μετρική Precision, Recall και F1-score για όλα τα N-grams, όπου $N = [1,2,3,4]$.

1. ROUGE 1-gram

Η εικόνα 3.23, απεικονίζει την σύγκριση της πραγματικής με την προβλεπόμενη πρόταση κατά 1-gram, δηλαδή λέξη προς λέξη. Οι λέξεις ("The", "cat", "the") προβλέπονται σωστά, όπως φαίνεται με πράσινο χρώμα. Αντίθετα, προβλέπεται λανθασμένα η λέξη "and" αντί για την λέξη "is", καμία λέξη για την λέξη "on" και τέλος, η λέξη "dog" αντί για την λέξη "mat".

Η μετρική "Precision" κατά 1-gram υπολογίζεται κατά τον ίδιο τρόπο που υπολογίζεται στην BLEU-N μετρική, διαιρώντας τον αριθμό των σωστά προβλεπόμενων λέξεων με τον συνολικό αριθμό των λέξεων της προβλεπόμενης πρότασης. Στο συγκεκριμένο παράδειγμα, ο αριθμός των σωστά προβλεπόμενων λέξεων είναι 3 ("The", "cat", "the")



Εικόνα 3.23: Σύγκριση πραγματικής και προβλεπόμενης πρότασης κατά 1-gram, δηλαδή σύγκριση λέξη προς λέξη. Οι λέξεις που εμφανίζονται με πράσινο χρώμα υποδεικνύουν σωστές προβλέψεις, ενώ εκείνες που εμφανίζονται με κόκκινο χρώμα αντιπροσωπεύουν τις πραγματικές λέξεις που θα έπρεπε να προβλεφθούν αλλά προβλέφθηκαν λάθος. Η εικόνα είναι προσαρμοσμένη από [31].

και ο συνολικός αριθμός των λέξεων της προβλεπόμενης πρότασης είναι 5 ("The", "cat", "and", "the", "dog"). Ο τρόπος υπολογισμού παρουσιάζεται ακολούθως:

$$\text{ROUGE-1 precision} = \frac{\text{Number of correct predicted 1-gram}}{\text{Number of total predicted 1-gram}} = \frac{3}{5}$$

Για τον υπολογισμό του "Recall" κατά 1-gram, διαιρείται ο αριθμός των σωστά προβλεπόμενων λέξεων με τον συνολικό αριθμό αυτή την φορά των λέξεων της πραγματικής πρότασης, κατά 1-gram. Στη συγκεκριμένη περίπτωση, ο αριθμός των σωστά προβλεπόμενων λέξεων είναι 3 ("The", "cat", "the"), ενώ ο συνολικός αριθμός των λέξεων της πραγματικής πρότασης είναι 6 ("The", "cat", "is", "on", "the", "mat"). Ο τρόπος υπολογισμού παρουσιάζεται ακολούθως:

$$\text{ROUGE-1 recall} = \frac{\text{Number of correct predicted 1-gram}}{\text{Number of total original 1-gram}} = \frac{3}{6}$$

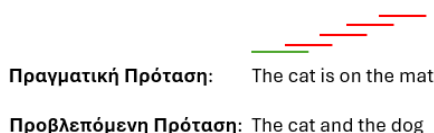
Για τον υπολογισμό του "F1-Score" κατά 1-gram, πολλαπλασιάζεται η τιμή του "Precision" και του "Recall" με τον αριθμό 2 και το αποτέλεσμα αυτών διαιρείται με το προστιθέμενο αποτέλεσμα της τιμής του "Precision" και του "Recall", όπως φαίνεται παρακάτω:

$$\text{ROUGE-1 F1-score} = \frac{2 \cdot (\text{precision} \cdot \text{recall})}{\text{precision} + \text{recall}}$$

2. ROUGE 2-grams

Η εικόνα 3.24, απεικονίζει την σύγκριση της πραγματικής με την προβλεπόμενη πρόταση κατά 2-grams, δηλαδή ανά δύο λέξεις. Μόνο το πρώτο ζευγάρι λέξεων ("The cat") προβλέπεται σωστά, όπως φαίνεται με πράσινο χρώμα. Αντίθετα, τα ζευγάρια λέξεων ("cat is", "is on", "on the", "the mat") δεν προβλέπονται σωστά, όπως φαίνεται στην εικόνα με κόκκινο χρώμα.

Ομοίως με τον τρόπο υπολογισμού της μετρικής "Precision" κατά 1-gram, η "Precision" κατά 2-grams υπολογίζεται διαιρώντας τον αριθμό των σωστά προβλεπόμενων λέξεων με τον συνολικό αριθμό των λέξεων της προβλεπόμενης πρότασης, κατά 2-grams. Στο συγκεκριμένο παράδειγμα, ο αριθμός των σωστά προβλεπόμενων ζευγαριών λέξεων είναι 1 ("The cat") και ο συνολικός αριθμός των λέξεων της προβλεπόμενης πρότασης είναι 4 ("The cat", "cat and", "and the", "the dog"). Ο τρόπος υπολογισμού παρουσιάζεται



Εικόνα 3.24: Σύγκριση πραγματικής και προβλεπόμενης πρότασης κατά 2-grams, δηλαδή σύγκριση ανά δύο λέξεις. Οι δυνάδες λέξεων που εμφανίζονται με πράσινο χρώμα υποδεικνύουν σωστές προβλέψεις, ενώ εκείνες που εμφανίζονται με κόκκινο χρώμα αντιπροσωπεύουν τις πραγματικές δυνάδες λέξεων που θα έπρεπε να προβλεφθούν αλληλά προβλεφθήκαν λάθος. Η εικόνα είναι προσαρμοσμένη από [31].

ακολουθώς:

$$\text{ROUGE-2 precision} = \frac{\text{Number of correct predicted 2-grams}}{\text{Number of total predicted 2-grams}} = \frac{1}{4}$$

Για τον υπολογισμό του "Recall" κατά 2-grams, διαιρείται ο αριθμός των σωστά προβλεπόμενων λέξεων με τον συνολικό αριθμό αυτή την φορά των λέξεων της πραγματικής πρότασης, κατά 2-grams. Στη συγκεκριμένη περίπτωση, ο αριθμός των σωστά προβλεπόμενων λέξεων είναι 1 ("The cat"), ενώ ο συνολικός αριθμός των λέξεων της πραγματικής πρότασης είναι 5 ("The cat", "cat is", "is on", "on the", "the mat"). Ο τρόπος υπολογισμού παρουσιάζεται ακολουθώς:

$$\text{ROUGE-2 recall} = \frac{\text{Number of correct predicted 2-grams}}{\text{Number of total original 2-grams}} = \frac{1}{5}$$

Ο υπολογισμός του "F1-Score" κατά 2-grams, πραγματοποιείται ακριβώς με τον ίδιο τρόπο με αυτόν του "F1-Score" κατά 1-gram, όπως φαίνεται παρακάτω:

$$\text{ROUGE-2 F1-score} = \frac{2 \cdot (\text{precision} \cdot \text{recall})}{\text{precision} + \text{recall}}$$

3. ROUGE L

Στηρίζεται στη μεγαλύτερη κοινή ακολουθία (Longest Common Subsequence-LCS) μεταξύ της προβλεπόμενης πρότασης και της πραγματικής, δηλαδή τη μεγαλύτερη ακολουθία λέξεων που είναι κοινή και στις δύο. Στο αναφερόμενο παράδειγμα, η μεγαλύτερη κοινή ακολουθία είναι η "The cat the", η οποία εμφανίζεται και στα δύο είδη προτάσεων. Να αναφερθεί ότι οι λέξεις δεν είναι απαραίτητα συνεχόμενες στη μεγαλύτερη κοινή ακολουθία.

Στη περίπτωση υπολογισμού της μετρικής "Precision", διαιρείται το μήκος της μεγαλύτερης κοινής ακολουθίας με τον συνολικό αριθμό των λέξεων της προβλεπόμενης πρότασης, κατά 1-gram. Σε αυτό το παράδειγμα, το μήκος της μεγαλύτερης κοινής ακολουθίας είναι 3 καθώς αποτελείται από τις λέξεις "The", "cat", "the" και ο αριθμός των λέξεων της προβλεπόμενης πρότασης, κατά 1-gram είναι 5 ("The", "cat", "and", "the", "dog").

$$\text{Rouge-L precision} = \frac{\text{Length of the LCS}}{\text{Number of 1-gram in predicted sentence}} = \frac{3}{5}$$

Στη περίπτωση υπολογισμού της μετρικής "Recall", διαιρείται το μήκος της μεγαλύτερης κοινής ακολουθίας με τον συνολικό αριθμό των λέξεων της πραγματικής πρότασης, κατά 1-gram. Σε αυτό το παράδειγμα, το μήκος της μεγαλύτερης κοινής ακολουθίας είναι 3 καθώς αποτελείται από τις λέξεις "The", "cat", "the" και ο αριθμός των λέξεων της πραγματικής πρότασης, κατά 1-gram είναι 6 ("The", "cat", "is", "on", "the", "mat").

$$\text{Rouge-L recall} = \frac{\text{Length of the LCS}}{\text{Number of 1-gram in original sentence}} = \frac{3}{6}$$

Ο τρόπος υπολογισμού του F1-Score είναι ίδιος όπως στις παραπάνω περιπτώσεις που αναφέρθηκαν.

$$\text{ROUGE-L F1-score} = \frac{2 \cdot (\text{precision} \cdot \text{recall})}{\text{precision} + \text{recall}}$$

4. Rouge S

Η μετρική αυτή, σε αντίθεση με την ROUGE-N που εξετάζει μόνο ακριβείς διαδοχικές αντιστοιχίες μεταξύ των πραγματικών και προβλεπόμενων προτάσεων, επιτρέπει την παράλειψη ενδιάμεσων λέξεων, γεγονός που την καθιστά πιο ευέλικτη για τη μέτρηση της ομοιότητας. Έχουμε το εξής παράδειγμα:

- Πραγματική Πρόταση: "The cat is on the mat"
- Προβλεπόμενη Πρόταση: "The gray cat and the dog"

Σε αυτό το παράδειγμα, η μετρική ROUGE-2 δεν θα θεωρούσε ως σωστά προβλεπόμενο 2-grams το "The cat" καθώς η προβλεπόμενη πρόταση περιλαμβάνει τη φράση "The gray cat". Αντίθετα, η ROUGE-S επιτρέπει την παράλειψη λέξεων. Επομένως, το "The cat" αντιστοιχίζεται με το "The gray cat".

Κεφάλαιο 4

Αποτελέσματα

4.1 WER μετρική

Αναφορικά με την εκτίμηση της απόδοσης του μοντέλου μας, χρησιμοποιώντας την WER μετρική, προκύπτει αποτέλεσμα **12,99%** σε σύγκριση με το μοντέλο "NeuSpeech" το οποίο χαρακτηρίζεται με ποσοστό σφαλμάτων 25,38%. Σημαίνει ότι, περίπου 12,99% των λέξεων στις προβλεπόμενες προτάσεις είναι λανθασμένες, δηλαδή 12,99% των λέξεων στις προβλεπόμενες προτάσεις είτε αντικαταστάθηκαν με άλλες, είτε διαγράφηκαν, είτε εισήχθησαν επιπλέον. Αναλυτικότερα, από τις συνολικά 37081 λέξεις των πραγματικών προτάσεων, οι 4542 λέξεις αντικαταστάθηκαν με λανθασμένες λέξεις. Επίσης, 71 λέξεις που βρίσκονται μεταξύ των πραγματικών προτάσεων δεν υπάρχουν στις προβλεπόμενες προτάσεις, ενώ 206 λέξεις προστέθηκαν στις προτάσεις που προέβλεψε το μοντέλο.

Στις εικόνες 4.1, 4.2 παρουσιάζονται κάποια παραδείγματα προτάσεων που προέβλεψε το μοντέλο καθώς και των αντίστοιχων πραγματικών, σημειώνοντας για κάθε πρόταση τις λέξεις που διαγράφηκαν, αντικαταστάθηκαν ή προστέθηκαν. Παρατηρώντας τα πρώτα 15 παραδείγματα των προβλέψεων, εντοπίζονται συγκεκριμένα πρότυπα στα σφάλματα των λέξεων. Αρχικά, παρατηρείται ότι το μοντέλο συχνά αντικαθιστά ονόματα με παρόμοια ακολουθία γραμμάτων, όπως Roy-Theoy (sentence 1), Tucker-Theucker (sentence 6), Chad-Iad (sentence 11). Παρόμοιες αντικαταστάσεις εντοπίζονται και σε άλλες προβλέψεις του μοντέλου. Αυτό υποδηλώνει ότι το μοντέλο παρουσιάζει μία δυσκολία στην ακριβή αναπαραγωγή ονομάτων, πιθανόν λόγω της χαμηλής συχνότητάς τους στο σύνολο εκπαίδευσης. Επίσης, το μοντέλο τείνει να αντικαθιστά λέξεις με το άρθρο "The" χωρίς να υπάρχει κάποια συσχέτιση, όπως Sorry-The (sentence 3), There-The (sentence 4), make-The (sentence 5), focused-The (sentence 7), He-The (sentence 8, sentence 14, sentence 15), Time-The (sentence 10), Anywhere-The (sentence 12), γεγονός που μπορεί να αποδοθεί στη προτίμηση του σε λεξικολογικές μονάδες υψηλής συχνότητας. Παράλληλα, παρατηρούνται αντικαταστάσεις λέξεων με σημασιολογικά όμοιες λέξεις, ωστόσο αυτή η αντικατάσταση οδηγεί σε τροποποίηση του αρχικού νοήματος της πρότασης. Χαρακτηριστικό παράδειγμα η αντικατάσταση της λέξης *oldest* από την λέξη *big* (sentence 6). Η αρχική πρόταση υποδηλώνει ότι ο Tucker είναι το μεγαλύτερο σε ηλικία αγόρι, ενώ η έκφραση *big boy* στην προβλεπόμενη πρόταση αλλοιώνει την αρχική έννοια, εστιάζοντας στη σωματική δύναμη του αγοριού. Ακόμη, παρατηρούνται αντικαταστάσεις ρημάτων, όπως *had-was* (sentence 4), *glared-ranared* (sentence 8), *sprinted-ran*(sentence 14), *knew-ran* (sentence 15), οι οποίες δεν έχουν κάποια νοη-

ματική συσχέτιση και δεν διατηρούν τη σημασιολογία των αρχικών λέξεων, πιθανόν λόγω αδυναμίας κατανόησης των λέξεων. Τέλος, οι περισσότερες τροποποιήσεις των αρχικών προτάσεων εμφανίζονται στην αρχή.

=== Sentence 1 === Reference: Roy Prediction: Theoy st Substitutions: [('Roy', 'Theoy')] Deletions: [] Insertions: ['st'] =====	=== Sentence 5 === Reference: make my stories thus but it is my nemesis confronting me anew at every turn Prediction: The my stories thus but it is my nemesis confronting me anew at every turn Substitutions: [('make', 'The')] Deletions: [] Insertions: [] =====
=== Sentence 2 === Reference: going to do it in two parts last part first so it take effect Prediction: I to do it in two parts last part first so it take effect Substitutions: [('going', 'I')] Deletions: [] Insertions: [] =====	=== Sentence 6 === Reference: Tucker was oldest boy in playground one who decided rules of game Prediction: Theucker was big boy in playground one who decided rules of game Substitutions: [('Tucker', 'Theucker'), ('oldest', 'big')] Deletions: [] Insertions: [] =====
=== Sentence 3 === Reference: Sorry Prediction: The Substitutions: [('Sorry', 'The')] Deletions: [] Insertions: [] =====	=== Sentence 7 === Reference: focused did after older oaths they sun oaths only complete iron forehead old nursery he doing Prediction: The did after older oaths they sun oaths only complete iron forehead old nursery he doing Substitutions: [('focused', 'The')] Deletions: [] Insertions: [] =====
=== Sentence 4 === Reference: There had to be more to her Prediction: The was to be more to her Substitutions: [('There', 'The'), ('had', 'was')] Deletions: [] Insertions: [] =====	

Εικόνα 4.1: Ανάλυση της WER μετρικής (διαγραφές, αντικαταστάσεις και προσθήκες) μεταξύ των προβλεπόμενων και πραγματικών προτάσεων του test set.

=== Sentence 8 === Reference: He glared at Allan eyes him and demanding a response Prediction: The ranared at Allan eyes him and demanding a response Substitutions: [('He', 'The'), ('glared', 'ranared')] Deletions: [] Insertions: [] =====	=== Sentence 12 === Reference: Anywhere Prediction: The crew he Substitutions: [('Anywhere', 'The')] Deletions: [] Insertions: ['crew', 'he'] =====
=== Sentence 9 === Reference: All that gold in Fort Knox was just too much for me to wrap my head around Prediction: Thean gold in Fort Knox was just too much for me to wrap my head around Substitutions: [('All', 'Thean'), ('that', 'gold'), ('gold', 'in'), ('in', 'Fort'), ('Fort', 'Knox'), ('Knox', 'was'), ('was', 'just'), ('just', 'too'), ('too', 'much'), ('much', 'for'), ('for', 'me'), ('me', 'to'), ('to', 'wrap'), ('wrap', 'my'), ('my', 'head'), ('head', 'around')] Deletions: ['around'] Insertions: [] =====	=== Sentence 13 === Reference: People will such stories they are worth publishing Prediction: I will such stories they are worth publishing Substitutions: [('People', 'I')] Deletions: [] Insertions: [] =====
=== Sentence 10 === Reference: Time out called not ready yet Prediction: The out called not ready yet Substitutions: [('Time', 'The')] Deletions: [] Insertions: [] =====	=== Sentence 14 === Reference: He sprinted for the corner of the building around which Roy had disappeared Prediction: The raned for the corner of the building around which Roy had disappeared Substitutions: [('He', 'The'), ('sprinted', 'raned')] Deletions: [] Insertions: [] =====
=== Sentence 11 === Reference: Chad just his head Prediction: lad was his head Substitutions: [('Chad', 'lad'), ('just', 'was')] Deletions: [] Insertions: [] =====	=== Sentence 15 === Reference: He knew it was too late to explain that he was trying to get a smile out of Roy Prediction: The ran it was too late to explain that he was trying to get a smile out of Roy Substitutions: [('He', 'The'), ('knew', 'ran')] Deletions: [] Insertions: [] =====

Εικόνα 4.2: Ανάλυση της WER μετρικής (διαγραφές, αντικαταστάσεις και προσθήκες) μεταξύ των προβλεπόμενων και πραγματικών προτάσεων του test set.

4.2 CER μετρική

Όσον αφορά την εκτίμηση της απόδοσης του προτεινόμενου μοντέλου, αξιοποιώντας την μετρική CER, προκύπτει αποτέλεσμα **8,89%**. Το συγκεκριμένο ποσοστό υποδεικνύει ότι το 8,89% των χαρακτήρων που παράγει το μοντέλο ως προβλέψεις, διαφέρουν από τους χαρακτήρες της αρχικής πρότασης, λόγω λαθών που σχετίζονται με αφαιρέσεις, προσθήκες ή αντικαταστάσεις χαρακτήρων. Είναι ένα χαμηλό ποσοστό λαθών σε επίπεδο χαρακτήρων και συνεπώς το μοντέλο έχει καλή ακρίβεια στην αναγνώριση των λέξεων, με μικρά σφάλματα σε επίπεδο χαρακτήρα, τα οποία δεν ασκούν σημαντική επιρροή στην ποιότητα της παραγόμενης εξόδου. Ενώ η συγκεκριμένη μετρική δεν έχει εφαρμοστεί στο μοντέλο "NeuSpeech", στην παρούσα μελέτη επιλέγεται για την ενδελεχή αξιολόγηση της απόδοσης του προτεινόμενου μοντέλου, επιτρέποντας την ανίχνευση ακόμη και των μικρότερων αποκλίσεων σε επίπεδο χαρακτήρων. Αυτή η προσέγγιση εξασφαλίζει μια λεπτομερή εκτίμηση της ποιότητας των προβλέψεων.

4.3 BLEU-N μετρική

Τα αποτελέσματα έδειξαν ότι, συγκρίνοντας τις προβλεπόμενες με τις πραγματικές προτάσεις:

- ανά μία λέξη (1-gram): το **BLEU-1 score** είναι **87,24%**.

Το BLEU-1 score αξιολογεί την αντιστοιχία μεταξύ των προβλεπόμενων και των πραγματικών προτάσεων σε επίπεδο λέξης (1-gram). Η τιμή του 87,24% υποδεικνύει ικανοποιητική ακρίβεια στην αναγνώριση και πρόβλεψη των λέξεων, χωρίς σημαντικές αποκλίσεις από τις λέξεις των πραγματικών προτάσεων. Συγκριτικά με το BLEU-1 score του "NeuSpeech" μοντέλου, το οποίο είναι 78,13%, το προτεινόμενο μοντέλο εμφανίζει βελτίωση κατά 9,11%.

- ανά δύο λέξεις (2-grams): το **BLEU-2 score** είναι **85,47%**.

Το BLEU-2 score υπολογίζει την αντιστοιχία για κάθε ζεύγος συνεχόμενων λέξεων (2-grams). Η τιμή 85,47% δείχνει ότι το μοντέλο εξακολουθεί να διατηρεί υψηλή ακριβή πρόβλεψη για δυάδες λέξεων, αλλά η τιμή είναι λίγο μικρότερη σε σχέση με το BLEU-1 score, υποδεικνύοντας μικρότερη ακριβή ταύτιση όταν εξετάζονται ζεύγη λέξεων. Σε σύγκριση με το BLEU-1 score του "NeuSpeech" μοντέλου, το οποίο είναι 71,25%, το προτεινόμενο μοντέλο συνεχίζει να εμφανίζει βελτίωση, στη περίπτωση αυτή κατά 14,22%.

- ανά τρεις λέξεις (3-grams): το **BLEU-3 score** είναι **84,72%**.

Το BLEU-3 score σχετίζεται με την σύγκριση των πραγματικών και των προβλεπόμενων προτάσεων ανά τριάδες λέξεων (3-grams). Η τιμή 84,72% δείχνει μια μικρή πτώση σε σχέση με τα προηγούμενα αποτελέσματα, γεγονός που υποδηλώνει ότι η πρόβλεψη του μοντέλου είναι πιο δύσκολη όταν πρόκειται για μεγαλύτερες αλληλουχίες λέξεων. Όσον αφορά την σύγκριση του με το BLEU-1 score του "NeuSpeech" μοντέλου, το μοντέλο μας υπερτερεί κατά 19,93%, δεδομένου ότι το "NeuSpeech" εμφανίζει 64,79%.

- ανά τέσσερις λέξεις (4-grams): το **BLEU-4 score** είναι **83,82%**.

Το BLEU-4 score, που συγκρίνει τις προβλεπόμενες και πραγματικές προτάσεις ανά τετράδες λέξεων (4-grams), είναι 83,82%, μια ελαφρώς μικρότερη τιμή από το BLEU-3 score, υποδηλώνοντας ότι το μοντέλο δυσκολεύεται περισσότερο να προβλέψει μεγαλύτερες αλληλουχίες λέξεων. Εξακολουθεί ωστόσο να εμφανίζει βελτίωση από το "NeuSpeech" κατά 24,96%, δεδομένου ότι το ποσοστό αυτού για το BLEU-4 score είναι 58,86%.

- Το συνολικό **BLEU score** είναι **85,30%**.

Το γεγονός ότι το Bleu score είναι **85,30%**, δείχνει ότι υπάρχει αρκετά υψηλή ταύτιση των λέξεων μεταξύ των πραγματικών προτάσεων και αυτών που προέβλεψε το μοντέλο καθώς και μικρές αποκλίσεις στην πρόβλεψη μεγαλύτερων αλληλουχιών λέξεων. Ο συνολικός αριθμός των λέξεων στις προβλεπόμενες προτάσεις είναι 37216 ενώ στις πραγματικές 37081. Για αυτόν τον λόγο, η τιμή του Brevity Penalty είναι 1,0.

Το μοντέλο που προτείνεται στην παρούσα εργασία παρουσιάζει σημαντική βελτίωση σε σχέση με το μοντέλο "NeuSpeech", με τις τιμές του BLEU score να είναι υψηλότερες σε όλες τις κλίμακες (N=1, N=2, N=3, N=4). Καθώς το N αυξάνεται από BLEU-1 σε BLEU-4, παρατηρούμε μια συνεχώς αυξανόμενη διαφορά στην απόδοση του μοντέλου μας σε σχέση με το "NeuSpeech". Συγκεκριμένα, η διαφορά μεταξύ των δύο μοντέλων αυξάνεται από 9.11% για BLEU-1 score σε 24.96% για BLEU-4 score. Η αυξανόμενη αυτή διαφορά δείχνει ότι το μοντέλο μας καθίσταται ικανό να διατηρεί την ποιότητα και τη συνέπεια της παραγωγής του σε σύνθετες προτάσεις, σε σύγκριση με το "NeuSpeech".

4.4 ROUGE-N μετρική

Στην περίπτωση του δικού μας μοντέλου, τα αποτελέσματα έδειξαν τα εξής:

- Το **ROUGE-1 score** είναι **78,04%**.

Η ROUGE-1 μετρική υπολογίζει την αντιστοιχία του μοντέλου με τις πραγματικές προτάσεις σε επίπεδο λέξης (1-gram). Το συγκεκριμένο ποσοστό υποδεικνύει ότι το προτεινόμενο μοντέλο καταφέρνει να προβλέψει σωστά το 78,04% των λέξεων που περιλαμβάνονται στις πραγματικές προτάσεις, υποδεικνύοντας ικανοποιητική ακρίβεια στην αναγνώριση των λέξεων. Ωστόσο, υπάρχει ακόμη περιθώριο βελτίωσης, δεδομένου ότι το υπάρχον μοντέλο "NeuSpeech" επιτυγχάνει υψηλότερο ποσοστό, προβλέποντας σωστά το 81,91% των λέξεων που περιλαμβάνονται στις πραγματικές προτάσεις.

- Το **ROUGE-2 score** είναι **72,78%**.

Η ROUGE-2 μετρική μετρά την αντιστοιχία των προβλεπόμενων και των πραγματικών προτάσεων, συγκρίνοντας ζεύγη συνεχόμενων λέξεων (2-grams). Η τιμή 72,78% υποδεικνύει μια ελαφρά μείωση σε σχέση με το ROUGE-1, υπογραμμίζοντας ότι το μοντέλο ενδέχεται να έχει λιγότερη ικανότητα να αναγνωρίζει με ακρίβεια ζεύγη συνεχόμενων λέξεων σε σχέση με τις μεμονωμένες λέξεις. Αυτό μπορεί να οφείλεται στο γεγονός ότι για την ακριβή πρόβλεψη δύο διαδοχικών λέξεων απαιτείται η κατανόηση της συντακτικής και σημασιολογικής τους σχέσης, καθώς και της θέσης τους εντός της

πρότασης. Η διαδικασία αυτή απαιτεί μεγαλύτερη γλωσσική κατανόηση και επεξεργασία σε επίπεδο δομής και νοήματος, κάτι που ενδέχεται να είναι πιο απαιτητικό για το μοντέλο σε σχέση με την απλή αντιστοιχία μεμονωμένων λέξεων, όπου η διαδικασία είναι πιο άμεση και βασίζεται σε λιγότερους γλωσσικούς περιορισμούς. Παρά τη μείωση του ποσοστού, η τιμή του δείχνει ότι το μοντέλο έχει ικανοποιητική απόδοση στην αναγνώριση των διαδοχικών λέξεων.

- Το **ROUGE-L score** είναι **78,03%**.

Η ROUGE-L μετρική αξιολογεί την ποιότητα των προβλέψεων μέσω της μεγαλύτερης κοινής ακολουθίας, η οποία καταγράφει την αντιστοιχία λέξεων μεταξύ της προβλεπόμενης και της πραγματικής πρότασης. Η συγκεκριμένη μετρική βασίζεται στην αντιστοιχία των λέξεων σε σωστή σειρά, χωρίς να απαιτεί οι λέξεις να είναι συνεχόμενες, αλλά να διατηρούν την ίδια σχετική ακολουθία. Η τιμή 78,03% υποδηλώνει ότι το μοντέλο καταφέρνει να διατηρεί την δομή και την συντακτική συνοχή των πραγματικών προτάσεων σε 78,03% των περιπτώσεων. Αυτό σημαίνει ότι οι λέξεις στην προβλεπόμενη πρόταση εμφανίζονται στην ίδια σειρά, διατηρώντας τη βασική νοηματική ροή του κειμένου.

- Το **ROUGE-S score** είναι **78,02%**.

Η ROUGE-S μετρική συγκρίνει τις προβλέψεις του μοντέλου με τις πραγματικές προτάσεις, αλλά με διαφορετικό τρόπο από τις μετρικές ROUGE-1 και ROUGE-2. Αντί να περιορίζεται στη σύγκριση συνεχόμενων λέξεων, εξετάζει την αντιστοιχία μεταξύ ομάδων λέξεων που δεν είναι απαραίτητα σε διαδοχικές θέσεις στην πρόταση. Ακόμα και όταν οι λέξεις δεν είναι συνεχόμενες, η συγκεκριμένη μετρική αξιολογεί τη σχέση μεταξύ τους, αναγνωρίζοντας αν αυτές οι λέξεις συνδέονται νοηματικά ή συντακτικά μέσα στο πλαίσιο της πρότασης. Το 78,02% υποδεικνύει ότι το μοντέλο αναγνωρίζει συσχετίσεις μεταξύ λέξεων που μπορεί να μην είναι διαδοχικές, δείχνοντας την ικανότητά του να κατανοεί και να διαχειρίζεται πιο περίπλοκες και σύνθετες σχέσεις λέξεων.

Η σημαντική μείωση της μετρικής WER που παρατηρείται στο μοντέλο μας (12,99%) σε σύγκριση με αυτή του "NeuSpeech" (25,38%) και γενικότερα η καλύτερη επίδοση του μοντέλου μας, αποδίδεται στη στρατηγική που ακολουθήθηκε με απώτερο σκοπό την μείωση του υπολογιστικού κόστους. Χρησιμοποιώντας τον προεκπαιδευμένο "Whisper Encoder" και προσαρμόζοντας τον κατάλληλα για εγκεφαλικά σήματα, εξάγονται υψηλής ποιότητας χαρακτηριστικά σχετικά με τα εγκεφαλικά δεδομένα, τα οποία αξιοποιεί το μοντέλο μας για την αποκωδικοποίηση εγκεφαλικών σημάτων σε κείμενο. Το γεγονός ότι στο μοντέλο εισάγονται απευθείας υψηλής ποιότητας χαρακτηριστικά των εγκεφαλικών σημάτων, χωρίς την ανάγκη εκπαίδευσης του "Whisper Encoder" για την εξαγωγή αυτών, όπως συμβαίνει στο "NeuSpeech", οδηγεί σε καλύτερης ποιότητας προβλέψεις. Παράλληλα, η χρήση λιγότερων Transformer Decoder επιπέδων και συγκεκριμένα η επιλογή μόνο ενός επιπέδου αντί τεσσάρων από τα οποία αποτελούνταν το αρχικό μοντέλο, βοήθησε στη μείωση της πιθανότητας υπερεκπαίδευσης και την επιτάχυνση της σύγκλισης του μοντέλου. Η μείωση της πολυπλοκότητας του μοντέλου επιτρέπει ταχύτερη προσαρμογή των παραμέτρων του και πιο γρήγορη σύγκλιση, καθώς το μοντέλο έχει λιγότερους υπολογιστικούς πόρους να διαχειριστεί. Με τη

χρήση λιγότερων επιπέδων, ο αριθμός των παραμέτρων μειώνεται, κάνοντας τη διαδικασία εκπαίδευσης πιο αποδοτική και επιταχύνοντας την επίτευξη του βέλτιστου αποτελέσματος χωρίς την ανάγκη πολλαπλών κύκλων εκπαίδευσης. Αυτό επιβεβαιώνεται μέσα από τις πειραματικές μας δοκιμές κατά τις οποίες παρατηρήθηκαν φαινόμενα υπερεκπαίδευσης με την χρήση περισσότερων του ενός Transformer Decoder επιπέδων.

Η εικόνα 4.3 περιλαμβάνει τις προβλέψεις του μοντέλου με ποιοτικό τρόπο. Για κάθε εγκεφαλικό σήμα, βλέπουμε την προβλεπόμενη πρόταση (Prediction) και κατά αντιστοιχία την πραγματική πρόταση (Reference). Φαίνεται ότι τα περισσότερα λάθη στις προβλέψεις του μοντέλου εμφανίζονται στην αρχή των προτάσεων. Κάποια από αυτά δεν σχετίζονται με την λέξη που θα έπρεπε να προβλεφθεί, ενώ άλλα φαίνεται να έχουν την ίδια κατάληξη.

Prediction: Theoy st Reference: Roy	Prediction: The was referring to your unhealthy for the fantastic Reference: I assume referring to your unhealthy for the fantastic
Prediction: I to do it in two parts last part first so it take effect Reference: going to do it in two parts last part first so it take effect	Prediction: The was out my palm phone looked up number Reference: I pulled out my palm phone looked up number
Prediction: The was to be more to her Reference: There had to be more to her	Prediction: The golden craft blew prepare color attractive crew Reference: than golden craft blew prepare color attractive crew
Prediction: Thean gold in Fort Knox was just too much for me to wrap my head around Reference: All that gold in Fort Knox was just too much for me to wrap my head around	Prediction: leed if still could could at any time follow advice seek a profession more grounded in reality Reference: Indeed if still could could at any time follow advice seek a profession more grounded in reality
Prediction: Theenty seven gates so far and none had turned up anything worth time Reference: twenty seven gates so far and none had turned up anything worth time	Prediction: The in name could have possessed you to do such a thing cried Arthur nearly raising his voice Reference: What in name could have possessed you to do such a thing cried Arthur nearly raising his voice
Prediction: The only pulled my leg any harder was coming off Reference: If he pulled my leg any harder was coming off	Prediction: The the the Reference: the the
Prediction: Thethough the concept was a simple one Allan thought it had potential Reference: Although the concept was a simple one Allan thought it had potential	Prediction: The rental and the books seemed troy to wake him up Reference: The drawings and the books seemed troy to wake him up
Prediction: The read in the refined world of print like a surly foreigner among a genteel crowd Reference: It appear in the refined world of print like a surly foreigner among a genteel crowd	Prediction: lahogany black curly hair cut short in the typical military style framed a dark skinned face Reference: Mahogany black curly hair cut short in the typical military style framed a dark skinned face
Prediction: I want to write as others do a tale set in our real world characters who might live therein not citizens some airy beyond the stars Reference: I want to write as others do a tale set in our real world characters who might live therein not citizens some airy beyond the stars	Prediction: The crew member that would jeopardize the project was meat Reference: Any crew member that would jeopardize the project was meat
Prediction: The that roots the sweet water of the stream grew a stand of black Reference: Before him roots the sweet water of the stream grew a stand of black	Prediction: The is more I wanted to show you Reference: This is what I wanted to show you

Εικόνα 4.3: Σύγκριση προβλεπόμενων και πραγματικών προτάσεων του test set.

Παρακάτω παρουσιάζεται ένας συγκεντρωτικός πίνακας 4.1 των αποτελεσμάτων τόσο του δικού μας μοντέλου όσο και του "NeuSpeech" μοντέλου του άρθρου [11]. Καθίσταται αναγκαίο να αναφερθεί ότι δεν υπάρχει διαθέσιμη σύγκριση μεταξύ του προτεινόμενου και του "NeuSpeech" μοντέλου όσον αφορά τις μετρικές CER, ROUGE-2, ROUGE-L και ROUGE-S λόγω έλλειψης δεδομένων στο αρχικό άρθρο.

Models	BLEU-N (%) ↑				ROUGE-N (%) ↑				WER (%) ↓	CER (%) ↓
	N=1	N=2	N=3	N=4	R-1	R-2	R-L	R-S		
NeuSpeech w/ tf	78.13	71.25	64.79	58.86	81.91	-	-	-	25.38	-
Our Model w/ tf	87.24	85.47	84.72	83.82	78.04	72.78	78.03	78.02	12.99	8.89

Πίνακας 4.1: Χρήση μετρικών για την αξιολόγηση του μοντέλου αποκωδικοποίησης των εγκεφαλικών σημάτων σε προτάσεις και αναφορά στο υπάρχον NeuSpeech μοντέλο. Το "w/ tf" σημαίνει "with teacher forcing". Το "NeuSpeech w/ tf" είναι μία από τις εκδοχές των πειραμάτων των ερευνητών.

Η εικόνα 4.4 απεικονίζει τις προβλέψεις τόσο του "NeuSpeech" μοντέλου όσο και του μοντέλου μας καθώς εισάγονται σε αυτά δύο διαφορετικά τμήματα εγκεφαλικών σημάτων,

τα οποία καταγράφηκαν κατά το άκουσμα δύο προτάσεων. Οι πραγματικές προτάσεις χαρακτηρίζονται στην εικόνα ως True.

True: My patience was long gone and I was back in the car to warming up when Acres tapped on the window and told me he had found whatever he was looking for Our Model: I patience was long gone and I was back in the car to warming up when Acres tapped on the window and told me he had found whatever he was looking for NeuSpeech Model: My patience was long gone and I was back in athe car.to warming up when Acres wastapped on thethe window and openedtold me he had found whatever he was looking for.
True: Very well said Arthur then take your story to that and let him defend you when no civilized publisher will approach you Our Model: I well said Arthur then take your story to that and let him defend you when no civilized publisher will approach you NeuSpeech Model: Very well said Arthur!then thattake your story to andthat and let him defend you and no civilized publisher will approach you when

Εικόνα 4.4: Παραγόμενες προτάσεις από το προτεινόμενο και το "NeuSpeech" μοντέλο έπειτα από την αποκωδικοποίηση δύο διαφορετικών εγκεφαλικών σημάτων.

Συγκρίνοντας τις παραγόμενες προτάσεις με την πραγματική πρόταση που αντιστοιχεί στο κάθε τμήμα εγκεφαλικού σήματος, διαπιστώνεται ότι και τα δύο μοντέλα καταφέρνουν να αναπαράγουν σε έναν αρκετά καλό βαθμό το κύριο περιεχόμενο των προτάσεων, γεγονός που δείχνει ότι είναι σε θέση να αποκωδικοποιήσουν με επιτυχία τη γλωσσική πληροφορία από τα εγκεφαλικά δεδομένα.

Επιπλέον, παρατηρείται ότι οι προτάσεις που παράγονται από το μοντέλο που προτείνεται στην παρούσα εργασία και το "NeuSpeech", παρουσιάζουν διαφορετικά είδη σφαλμάτων σε σύγκριση με την πραγματική πρόταση. Αναλυτικότερα, το μοντέλο μας διατηρεί σε μεγάλο βαθμό την συντακτική δομή της αρχικής πρότασης, ωστόσο εμφανίζει λάθη αντικατάστασης λέξεων, όπως η αντικατάσταση της λέξης "My" στην πρώτη πρόταση με την λέξη "I" και η αντικατάσταση της λέξης "Very" στην δεύτερη πρόταση πάλι με την λέξη "I". Αξίζει να παρατηρήσουμε ότι, στην πρώτη πρόταση η λέξη "My" και "I" παρουσιάζουν φωνολογική ομοιότητα, καθώς η κατάληξη και των δύο ηχεί με τον ίδιο τρόπο. Παράλληλα, παρατηρείται κάποιου είδους σημασιολογική συσχέτιση μεταξύ αυτών των δύο λέξεων καθώς και οι δύο σχετίζονται με το α' πρόσωπο. Στη δεύτερη πρόταση, η αντικατάσταση της λέξης "Very" με την λέξη "I" δεν φαίνεται να έχει κάποια σημασιολογική ή φωνολογική συσχέτιση. Αντίθετα, οι προτάσεις που παράγει το "NeuSpeech" μοντέλο περιλαμβάνουν περιττές επαναλήψεις, όπως στην πρώτη πρόταση με το "thethe" καθώς και ενωμένες λέξεις που το δεύτερο μέρος τους είναι η λέξη που θα έπρεπε να προβλεφθεί, όπως οι περιπτώσεις "athe", "wastapped", "openedtold" στην πρώτη πρόταση και η περίπτωση "thattake" στην δεύτερη πρόταση. Ακόμη παρατηρούνται προσθήκες λέξεων, όπως η λέξη "and" και "when" στη δεύτερη πρόταση.

Λαμβάνοντας υπόψιν τα δύο παραδείγματα, το "NeuSpeech" μοντέλο παρουσιάζει μεγαλύτερη δυσκολία στη σωστή τοποθέτηση των λέξεων ενώ το μοντέλο που προτείνεται εμφανίζεται πιο ακριβές στη σύνταξη και στη διατήρηση της βασικής σημασίας της πρότασης.

Σύγκριση Υπολογιστικής Πολυπλοκότητας

Στο άρθρο "Decode Neural Signal as Speech" [11], στο οποίο υλοποιήθηκε το μοντέλο "NeuSpeech" για διάφορες εκδοχές (όπως με Teacher Forcing ή χωρίς Teacher Forcing), δεν παρέχεται ξεχωριστή εκτίμηση του υπολογιστικού κόστους για κάθε εκδοχή του μοντέλου. Ωστόσο, αναφέρεται ότι όλα τα μοντέλα εκπαιδεύτηκαν για 500 epochs σε Nvidia H800 (80GB) GPUs και η εκπαίδευση διαρκεί περίπου 48 ώρες. Δεδομένου αυτού, χρησιμοποιούμε αυτόν τον χρόνο ως σημείο αναφοράς για τη σύγκριση του χρόνου εκπαίδευσης ανά εποχή. Εφόσον το προτεινόμενο μοντέλο χρησιμοποιεί την τεχνική Teacher Forcing, την οποία περιλαμβάνει επίσης μία εκδοχή του NeuSpeech, θεωρούμε ότι ο χρόνος εκπαίδευσης είναι πολύ κοντά σε αυτόν που αναφέρεται στο άρθρο. Συνεπώς, η ανάλυση της υπολογιστικής αποτελεσματικότητας γίνεται με βάση αυτή την εκτίμηση.

Αναλυτικότερα, δεδομένου ότι το "NeuSpeech" εκπαιδεύεται 48 ώρες για 500 εποχές, εκτιμούμε ότι χρειάζεται $48/500 = 0,096$ ώρες = $0,096 \text{ ώρες} * 60 \text{ λεπτά} = 5,7$ λεπτά ανά εποχή. Το γεγονός ότι το προτεινόμενο μοντέλο εκπαιδεύεται 13 ώρες για 7 εποχές, εκτιμούμε ότι χρειάζεται $13/7 = 1,86$ ώρες = $1,86 \text{ ώρες} * 60 \text{ λεπτά} = 111,6$ λεπτά ανά εποχή. Συνεπώς, το μοντέλο μας απαιτεί $1,86 \text{ ώρες ανά εποχή} / 0,096 \text{ ώρες ανά εποχή} = 19,4$ φορές περισσότερο χρόνο εκπαίδευσης ανά εποχή σε σύγκριση με το "NeuSpeech". Η περιγραφή αυτή αποτυπώνεται συγκεντρωτικά στον ακόλουθο πίνακα.

Models	Training Time	Epochs	Training Time per Epoch	Hardware
NeuSpeech w/ tf	48h	500	0.096 h = 5.7 min	Nvidia H500 (80 GB VRAM) GPUs
Our Model w/ tf	13h	7	1.86 h = 110.6 min	CPU (16GB RAM)

Πίνακας 5.1: Σύγκριση χρόνου εκπαίδευσης και αναφορά στο hardware που χρησιμοποιήθηκε κατά την εκπαίδευση του προτεινόμενου και του μοντέλου "NeuSpeech". Το προτεινόμενο μοντέλο απαιτεί 19,4 φορές περισσότερο χρόνο εκπαίδευσης ανά εποχή σε σύγκριση με το "NeuSpeech".

Αυτή η διαφορά δεν οφείλεται στην αρχιτεκτονική του μοντέλου, αλλά στις υπολογιστικές δυνατότητες του hardware που χρησιμοποιήθηκε. Συγκεκριμένα, το "NeuSpeech" εκπαιδεύτηκε σε μια ισχυρή Nvidia H800 GPU με 80GB VRAM, ενώ το προτεινόμενο μοντέλο εκπαιδεύτηκε σε CPU με 16GB RAM, γεγονός που αύξησε σημαντικά τον χρόνο εκπαίδευσης. Επιπλέον, η ανάγκη αποθήκευσης δεδομένων σε εξωτερικό HDD πιθανώς επηρέασε αρνητικά την απόδοση λόγω υψηλότερης καθυστέρησης εισόδου/εξόδου (I/O). Αν το ίδιο μοντέλο εκπαιδευόταν σε CPU, με βάση τις μετρήσεις μας, θα απαιτούσε περίπου 920 έως 1425 ώρες (38-59 ημέρες), λόγω της σημαντικής υπολογιστικής διαφοράς μεταξύ GPU και

CPU. Η εκτίμηση αυτή βασίζεται στον ρυθμό εκπαίδευσης του προτεινόμενου μοντέλου που είναι 19,4 φορές πιο αργός ανά εποχή συγκριτικά με το "NeuSpeech" που εκπαιδεύτηκε σε GPU. Αν εφαρμοζόταν ο ίδιος συντελεστής επιβράδυνσης για το "NeuSpeech" μοντέλο σε CPU, η εκπαίδευση του θα διαρκούσε $5,7 \text{ λεπτά} * 19,4 = 110,6 \text{ λεπτά}$ ανά εποχή, δηλαδή $110,6/60 = 1,84 \text{ ώρες}$ ανά εποχή και συνολικά για 500 εποχές, $1,84*500 = 920 \text{ ώρες} = 920/24 = 38,3 \text{ ημέρες}$. Αν όμως θεωρηθεί ένας πιο επιθετικός συντελεστής επιβράδυνσης 30 φορές πιο αργός, η εκπαίδευση θα απαιτούσε $5,7 \text{ λεπτά} * 30 = 171 \text{ λεπτά}$ ανά εποχή, δηλαδή $171/60 = 2,85 \text{ ώρες}$ ανά εποχή και συνολικά για 500 εποχές, $2,85*500 = 1425 \text{ ώρες} = 1425/24 = 59,4 \text{ ημέρες}$.

Model	Hardware	Deceleration Factor	Training Time per Epoch	500 Epochs
NeuSpeech w/tf	CPU (16GB RAM)	19.4x	110.6 min = 1.84 h	920 h = 38.3 days
NeuSpeech w/tf	CPU (16GB RAM)	30x	171 min = 2.85 h	1425 h = 59.4 days

Πίνακας 5.2: Σύγκριση χρόνου εκπαίδευσης του "NeuSpeech" μοντέλου σε CPU αντί για GPU, σε περίπτωση που η CPU είναι 19,4 φορές ή 30 φορές πιο αργή από την GPU. Με βάση τις μετρήσεις μας, θα απαιτούσε περίπου 920 έως 1425 ώρες (38-59 ημέρες) για να ολοκληρωθεί η εκπαίδευση του.

Συμπερασματικά, η εκπαίδευση του "NeuSpeech", που ολοκληρώνεται σε 48 ώρες (2 ημέρες) σε GPU, εκτιμάται ότι θα απαιτούσε 920 έως 1425 ώρες (38-59 ημέρες) σε CPU. Η αύξηση του χρόνου εκπαίδευσης κατά 872 έως 1377 ώρες σε σχέση με την εκπαίδευση σε GPU, καθιστά την χρήση της CPU μη πρακτική λόγω του υπερβολικού χρόνου εκπαίδευσης, των αυξημένων απαιτήσεων σε μνήμη και υπολογιστική ισχύ, καθώς και των δυσκολιών στην εκτέλεση επαναλαμβανόμενων πειραμάτων για τη βελτιστοποίηση του μοντέλου.

Δεδομένου ότι το μοντέλο μας αποτελεί μία απλοποιημένη εκδοχή του "NeuSpeech" μοντέλου και εκτελείται σε CPU με περιορισμένους υπολογιστικούς πόρους, η απόδοση του αναμενόταν να είναι χαμηλότερη συγκριτικά με την απόδοση του "NeuSpeech". Ωστόσο, τα αποτελέσματα δείχνουν ότι, παρά τους περιορισμούς, το μοντέλο μας παρουσιάζει εντυπωσιακή απόδοση, επιτυγχάνοντας μικρότερα ποσοστά σφαλμάτων (WER: 12,99%) έναντι του "NeuSpeech" (WER: 25,38%). Αυτό αποδεικνύει ότι το μοντέλο μας καταφέρνει να ανταγωνίζεται αποτελεσματικά το "NeuSpeech", προσφέροντας καλύτερη απόδοση σε πραγματικές συνθήκες με λιγότερους υπολογιστικούς πόρους. Η εκπαίδευση του αποκωδικοποιητή σε αντίθεση με την εκπαίδευση του κωδικοποιητή και παράλληλα η αξιοποίηση του κωδικοποιητή για την ανάλυση των εγκεφαλικών σημάτων σε συγκεκριμένα χαρακτηριστικά, τα οποία εισάγονται έπειτα στον αποκωδικοποιητή, συνέβαλαν σημαντικά στην βελτίωση των αποτελεσμάτων του προτεινόμενου μοντέλου. Συγχρόνως, η επιλογή της απλοποίησης της αρχιτεκτονικής πολυπλοκότητας του αποκωδικοποιητή μέσω της μείωσης των επιπέδων του Transformer Decoder από τέσσερα σε ένα, συνέβαλε στη βελτίωση της γενίκευσης του μοντέλου, περιορίζοντας τον κίνδυνο υπερεκπαίδευσης. Παράλληλα, διευκόλυνε τη μάθηση κρίσιμων γλωσσικών μοτίβων και επιτάχυνε τη σύγκλιση, καθιστώντας την εκπαίδευση πιο αποδοτική.

Κεφάλαιο 6

Συζήτηση

Η έρευνα με τίτλο "Decode Neural Signal as Speech" [11] αποτέλεσε σημαντική πηγή έμπνευσης για την εκπόνηση της παρούσας πτυχιακής εργασίας. Η μεγαλύτερη δοκιμασία θα μπορούσε να πεί κανείς, πως ήταν η υλοποίηση ενός μοντέλου μετάφρασης των MEG σημάτων σε κείμενο, με τέτοιο τρόπο ώστε να μπορεί να ανταποκριθεί σε υπολογιστικά συστήματα με όχι τόσο εξειδικευμένο hardware όπως αυτό που χρησιμοποιήθηκε στη συγκεκριμένη έρευνα. Το μοντέλο που αναπτύχθηκε, παρά την απλοποίηση του, παρουσιάζει ενθαρρυντικά αποτελέσματα σχετικά με την ανάλυση των MEG σημάτων και την μετατροπή τους σε κείμενο. Η πλειοψηφία των εγκεφαλικών σημάτων, που λήφθηκαν από τους συμμετέχοντες του πειράματος, μεταφράζεται σε έναν αρκετά ικανοποιητικό βαθμό στις αντίστοιχες προτάσεις που αυτοί άκουσαν.

Η προτεινόμενη προσέγγιση, για την αποκωδικοποίηση ομιλίας από εγκεφαλικά σήματα, έχει τη δυνατότητα να ενσωματωθεί σε πραγματικά συστήματα, οδηγώντας σε νέες δυνατότητες και καινοτόμες εφαρμογές σε πολλούς τομείς. Αναλυτικότερα, στον τομέα της ιατρικής, η προτεινόμενη μέθοδος μπορεί να προσφέρει σημαντική βοήθεια σε άτομα που αντιμετωπίζουν επικοινωνιακές δυσκολίες, όπως ασθενείς με εγκεφαλικά επεισόδια, παράλυση ή άλλες νευρολογικές παθήσεις, επιτρέποντας τους να επικοινωνούν μέσω των σκέψεων τους χωρίς την ανάγκη φυσικής ομιλίας. Επιπλέον, με την ενσωμάτωση αυτής της τεχνολογίας σε υπολογιστές και κινητά τηλέφωνα, άνθρωποι με κινητικά προβλήματα θα έχουν την δυνατότητα να πλοηγούνται εύκολα στο διαδίκτυο, να συγγράφουν κείμενα και να συνομιλούν με άτομα, χωρίς τη χρήση των χεριών ή της φωνής τους. Αυτή η τεχνολογία θα ενισχύσει την αυτονομία τους, προσφέροντας τους τη δυνατότητα να επιλέγουν και να εκτελούν ενέργειες με μεγαλύτερη ανεξαρτησία και χωρίς την ανάγκη εξωτερικής βοήθειας. Ακόμη, μία τέτοια τεχνολογία μπορεί να αξιοποιηθεί στην καθημερινότητα των ανθρώπων, παραδείγματος χάριν στην ασφάλεια της οδήγησης, αποτρέποντας πιθανά ατυχήματα. Με την παρακολούθηση της εγκεφαλικής δραστηριότητας των οδηγών, είναι δυνατόν να ανιχνευθούν πρώιμα σημάδια κόπωσης, γεγονός που επιτρέπει την έγκαιρη παρέμβαση για την αποφυγή επικίνδυνων για εκείνον καταστάσεων, όπως η υπνηλία ή η έλλειψη συγκέντρωσης. Γενικότερα, μία τέτοια τεχνολογία μπορεί να αξιοποιηθεί σε ποικίλες εφαρμογές που απαιτούν την ανάλυση της σκέψης. Πρέπει όμως να αναρωτηθούμε, σε τι βαθμό είναι έτοιμη η ανθρωπότητα να αξιοποιήσει ένα τέτοιο επίτευγμα μόνο για καλό σκοπό και με ποιόν τρόπο μπορεί να διασφαλιστεί η ιδιωτικότητα και η αυτονομία του ατόμου και να αποτρέπεται κάθε μορφή κατάχρησης ή ανεπιθύμητης πρόσβασης στα νευροδεδομένα του.

Η ιδιαίτερα υψηλή απόδοση του μοντέλου, παρά τους περιοριστικούς παράγοντες σχετικά με την ανάγκη μείωσης του υπολογιστικού κόστους, επιβεβαιώνει τη δυναμική αυτού του μοντέλου για την αποκωδικοποίηση ομιλίας από εγκεφαλικά σήματα. Ωστόσο, είναι απαραίτητο να πραγματοποιηθούν μελλοντικές βελτιώσεις στο μοντέλο μας, ώστε να ενισχυθεί η γενίκευση του και να αυξηθεί η αποτελεσματικότητά του σε ποικίλα πραγματικά σενάρια. Ειδικότερα, η αύξηση των δεδομένων εκπαίδευσης με την προσθήκη περισσότερων συνόλων δεδομένων που περιλαμβάνουν διαφορετικά άτομα, συνθήκες και περιβάλλοντα, θα επιτρέψει στο μοντέλο να κατανοήσει και να μάθει από μια μεγαλύτερη ποικιλία χαρακτηριστικών και διαφορών. Με αυτόν τον τρόπο, το μοντέλο θα ενισχύσει τη γενίκευσή του, καθώς μαθαίνει να αναγνωρίζει και να προσαρμόζεται σε νέες καταστάσεις και περιβάλλοντα που μπορεί να συναντήσει σε πραγματικά σενάρια. Η προσθήκη θορυβωδών συνθηκών, για παράδειγμα, μπορεί να καταστήσει το μοντέλο πιο ανθεκτικό στον θόρυβο και ικανό να αποδίδει καλύτερα, ακόμη και σε καταστάσεις με θορυβώδη δεδομένα, ενώ η εκπαίδευση με διαφορετικές γλώσσες θα ενισχύσει τη δυνατότητά του να αναγνωρίζει και να επεξεργάζεται ποικιλία φωνητικών και γλωσσικών χαρακτηριστικών. Με αυτές τις βελτιώσεις, το μοντέλο θα είναι πιο ευέλικτο και ικανό να λειτουργεί σε πραγματικές και ποικιλόμορφες συνθήκες. Ακόμη, η αφαίρεση της τεχνικής Teacher Forcing στο μοντέλο μπορεί να ενισχύσει τη γενίκευση και την ανθεκτικότητά του, καθώς αποφεύγεται η υπερβολική εξάρτηση από τις σωστές προβλέψεις κατά τη διάρκεια της εκπαίδευσης. Χωρίς τη χρήση αυτής της τεχνικής, το μοντέλο μαθαίνει να προσαρμόζεται σε πραγματικά σενάρια, όπου οι προβλέψεις του δεν βασίζονται πάντα σε σωστά δεδομένα, αλλά σε δικές του εκτιμήσεις. Αυτό αυξάνει την ικανότητά του να διαχειρίζεται αληθινές καταστάσεις με μεγαλύτερη ακρίβεια.

Συμπέρασμα

Η παρούσα πτυχιακή εργασία εστιάζει στη βελτιστοποίηση της υπολογιστικής πολυπλοκότητας του "NeuSpeech" μοντέλου, εφαρμόζοντας μια προσέγγιση που εξασφαλίζει ταυτόχρονα υψηλή απόδοση. Μέσω της εφαρμογής του κωδικοποιητή του "Whisper" μοντέλου για την αποτελεσματική ανάλυση των εγκεφαλικών σημάτων, καθώς και την εκπαίδευση του αποκωδικοποιητή του ίδιου μοντέλου για την αποκωδικοποίηση αυτών των σημάτων σε κείμενο, επιτυγχάνεται με υψηλή ακρίβεια η μετατροπή εγκεφαλικών σημάτων σε προτάσεις. Η προτεινόμενη προσέγγιση μειώνει την πολυπλοκότητα του αρχικού μοντέλου "NeuSpeech", επιτρέποντας γρηγορότερη σύγκλιση και αποδοτικότερη εκπαίδευση του μοντέλου αυτού, χωρίς να θυσιάζεται η ποιότητα των αποτελεσμάτων. Το κύριο όφελος αυτής της προσέγγισης είναι η ικανότητα του μοντέλου να μετατρέπει τα εγκεφαλικά σήματα σε προτάσεις με υψηλή ακρίβεια, ενώ συγχρόνως μειώνει την υπολογιστική επιβάρυνση, καθιστώντας τη ιδανική για φορητές συσκευές και συστήματα με περιορισμένες υπολογιστικές δυνατότητες. Τα εντυπωσιακά αποτελέσματα της προτεινόμενης προσέγγισης αποδεικνύουν την ισχυρή δυναμική του μοντέλου μας, ανοίγοντας νέους επιστημονικούς ορίζοντες στον τομέα της νευροεπιστήμης. Η εφαρμογή της τεχνολογίας αυτής, δημιουργεί τις κατάλληλες προϋποθέσεις για την ανάπτυξη καινοτόμων εφαρμογών που στοχεύουν στην ενίσχυση της ανθρώπινης επικοινωνίας και αλληλεπίδρασης μέσω εγκεφαλικής διεπαφής. Αυτές οι εφαρμογές έχουν τη δυνατότητα να συμβάλλουν στην αποκατάσταση και υποστήριξη ατόμων με κινητικές ή επικοινωνιακές δυσκολίες, βελτιώνοντας την καθημερινότητα και τη ποιότητα ζωής αυτών των ανθρώπων.

Βιβλιογραφία

- [1] G. K. Anumanchipalli, J. Chartier και E. F. Chang. *Speech synthesis from neural decoding of spoken sentences*. *Nature*, 568:493–498, 2019.
- [2] R. Wang, X. Chen, A. Khalilian-Gourtani, Z. Chen, L. Yu, A. Flinker και Y. Wang. *Stimulus Speech Decoding from Human Cortex with Generative Adversarial Network Transfer Learning*. σελίδες 390–394, 2020.
- [3] F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson και K. V. Shenoy. *High-performance brain-to-text communication via handwriting*. *Nature*, 593:249–254, 2021.
- [4] D. A. Moses, S. L. Metzger, J. R. Liu, G. K. Anumanchipalli, J. G. Makin, M. E. Dougherty, P. Z. Liu, G. M. Abrams και E. F. Chang. *Neuroprosthesis for decoding speech in a paralyzed person with anarthria*. *New England Journal of Medicine*, 385:217–227, 2021.
- [5] S. L. Metzger, F. R. Willett, S. Kellis, P. Nuyujukian, M. Aghagolzadeh, J. R. Liu, K. D. Anderson, K. V. Shenoy, L. R. Hochberg, J. M. Henderson, A. B. Ajiboye και C. E. Bouton. *Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis*. *Nature Communications*, 13:6510, 2022.
- [6] S.L. Metzger, K.T. Littlejohn, A.B. Silva, D. A. Moses, M. P. Seaton, R. Wang, M. E. Dougherty, J. R. Liu, P. Wu, M. A. Berger, I. Zhuravleva, A. Tu-Chan, K. Ganguly, G. K. Anumanchipalli και E. F. Chang. *A high-performance neuroprosthesis for speech decoding and avatar control*. *Nature*, 620:1037–1046, 2023.
- [7] Y. Liu, Z. Zhao, M. Xu, H. Yu, Y. Zhu, J. Zhang, L. Bu, Zhang X., J. Lu, Y. Li, D. Ming και J. Wu. *Decoding and synthesizing tonal language speech from brain activity*, 2023.
- [8] Z. Wang και H. Ji. *Open Vocabulary Electroencephalography-to-Text Decoding and Zero-Shot Sentiment Classification*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:5350–5358, 2022.
- [9] Y. Duan, C. Zhou, Z. Wang, Y. Wang και C. Lin. *DeWave: Discrete Encoding of EEG Waves for EEG-to-Text Translation*. *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

- [10] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli και J. King. *Decoding Speech Perception from Non-Invasive Brain Recordings. Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 5(10), 2023.
- [11] Y. Yang, Y. Duan, Zhang Q., H. Jo, J. Zhou, W. Lee, R. Xu και H. Xiong. *NeuSpeech: Decode Neural signal as Speech*, 2024.
- [12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey και I. Sutskever. *Robust Speech Recognition via Large-Scale Weak Supervision*, 2022.
- [13] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov και L. Zettlemoyer. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, σελίδες 7871–7880. Association for Computational Linguistics, 2020.
- [14] Q. Zhang, M. Chen, A. Bukharin, N. Karampatziakis, P. He, Y. Cheng, W. Chen και T. Zhao. *AdaLoRA: Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning*, 2023.
- [15] J. King και L. Gwilliams. *OSF Repository*. <https://osf.io/ag3kj/>.
- [16] L. Gwilliams, G. Flick, A. Marantz, L. Pylkkanen, D. Poeppel και J. King. *MEG-MASC: a high-quality magnetoencephalography dataset for evaluating natural speech processing. Scientific Data*, 10:862, 2023.
- [17] S. P. Singh. *Magnetoencephalography: Basic principles. Annals of Indian Academy of Neurology*, 17(Suppl 1):107–112, 2014.
- [18] P. Abry, P. Ciuciu, M. Dumeur, S. Jaffard και G. Saës. *Multifractal analysis based on the weak scaling exponent and applications to MEG recordings in neuroscience. HAL Open Archive*, 2024.
- [19] Scikitlearn developers. *RobustScaler - Scikit-learn*. <https://scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.RobustScaler.html>.
- [20] MNE Developers. *Evoked Topomap Visualization - MNE*. https://mne.tools/stable/auto_examples/visualization/evoked_topomap.html.
- [21] S. Luthra. *The Role of the Right Hemisphere in Processing Phonetic Variability Between Talkers. Neurobiology of Language*, 2(1):138–151, 2021.
- [22] S. Ali, S. Tanweer, S. Khalid και N. Rao. *Mel Frequency Cepstral Coefficient: A Review*. 2021.
- [23] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov και L. Zettlemoyer. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, 2019.

- [24] J. Gross. *Magnetoencephalography (MEG)*, 2024.
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, Gomez A. N., L. Kaiser και I. Polosukhin. *Attention Is All You Need*, 2017.
- [26] Kolena. *Metrics: WER, CER, MER*. <https://docs.kolena.com/metrics/wer-cer-mer/>.
- [27] Kolena. *Metric: METEOR*. <https://docs.kolena.com/metrics/meteor/>.
- [28] Hugging Face. *Evaluation - Hugging Face Audio Course*. <https://huggingface.co/learn/audio-course/chapter5/evaluation>.
- [29] Towards Data Science. *Foundations of NLP Explained: BLEU Score and WER Metrics*. <https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b>, 2021.
- [30] Hugging Face. *Metric: CER*. <https://huggingface.co/spaces/evaluate-metric/cer>.
- [31] NLPlanet. *Two Minutes NLP: Learn the ROUGE Metric by Examples*. <https://medium.com/nlplanet/two-minutes-nlp-learn-the-rouge-metric-by-examples-f179cc285499>, 2022.