ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Ανάλυση αρχείων ήχου (podcasts) ως προς το περιεχόμενο με χρήση Μηχανικής Μάθησης»

Του φοιτητή
Αλκιβιάδη Μυρόβαλη
Αρ. Μητρώου: 175093

Επιβλέπων
Κωνσταντίνος Διαμαντάρας
Καθηγητής

Ιανουάριος 2025

Ανάλυση αρχείων ήχου (podcasts) ως προς το περιεχόμενο με χρήση Μηχανικής Μάθησης

Κωδικός Π.Ε.: 24223

Αλκιβιάδης Μυρόβαλης

Κωνσταντίνος Διαμαντάρας

Ημερομηνία ανάληψης Π.Ε.: 11/9/2024

Ημερομηνία περάτωσης Π.Ε.: 26/1/2025

«*Αφιέρωση»*

"Θα ήθελα να εκφράσω την ευγνωμοσύνη στην οικογένειά μου και στους φίλους μου για την υποστήριξη και ενθάρρυνσή τους κατά τη διάρκεια των σπουδών μου. Επίσης, ευχαριστώ θερμά τους καθηγητές για την καθοδήγησή τους. Ιδιαίτερα, θα ήθελα να ευχαριστήσω το European School Radio, το μαθητικό ραδιόφωνο που αποτέλεσε μέρος της πτυχιακής μου εργασίας, για την ευκαιρία να αναπτύξω τις γνώσεις μου στο πεδίο της τεχνητής νοημοσύνης, ενισχύοντας τις δεξιότητές μου και τη συνεργατική μου εμπειρία."

Πρόλογος

Η συγκεκριμένη πτυχιακή εργασία προήλθε από τον ενθουσιασμό μου για τον συνδιασμό της τεχνολογίας του ήχου και της μηχανικής μάθησης. Η ραγδαία ανάπτυξη των αλγορίθμων επεξεργασίας ήχου και φυσικής γλώσσας, καθώς και η ανάγκη για βελτίωση της αναγνώρισης και κατηγοριοποίησης ηχητικού περιεχομένου, αποτέλεσαν το κίνητρο για την ενασχόλησή μου με το συγκεκριμένο αντικείμενο. Η μελέτη και η εφαρμογή αλγορίθμων όπως ο YamNet, ο Whisper, και ο DeepFilterNet, ο BART, συνέβαλαν στην κατανόησή μου για το πώς οι σύγχρονοι αλγόριθμοι μπορούν να βελτιώσουν την επεξεργασία ηχητικών δεδομένων και φυσικής γλώσσας, επιτρέποντας την αυτόματη ανάλυση και κατηγοριοποίηση περιεχομένου. Η εργασία εστιάζει στην ανάλυση των podcasts του European School Radio και το πως οι συγκεκριμένοι αλγόριθμοι λειτουργούν σε αυτα τα ηχητικά δεδομένα. Ο στόχος της εργασίας είναι να να βρεθούν οι κατάλληλοι αλγόριθμοι με τους οποίους θα φτιαχτεί ένα ολοκληρωμένο σύστημα επεξεργασίας ηχητικού περιεχομένου που θα βελτιώσει την αναγνώριση,την ταξινόμηση και αναζήτηση περιεχομένου στο European School Radio, ενισχύοντας την εμπειρία των χρηστών του.

Περίληψη

Η πτυχιακή αυτή εξετάζει την εφαρμογή αλγορίθμων μηχανικής μάθησης σε δεδομένα ήχου, με ιδιαίτερη έμφαση στα podcast. Στόχος της εργασίας είναι η διερεύνηση αλγορίθμων, όπως τα μοντέλα YamNet, Whisper, Deep FilterNet και BART, στην αυτόματη ανάλυση και κατανόηση του περιεχομένου podcast. Αρχικά, διατυπώνεται μια σύντομη ιστορική αναδρομή της εξέλιξης της τεχνολογίας ήχου και του ρόλου που έπαιξε στην ανάπτυξη της ψηφιακής επεξεργασίας. Στη συνέχεια, αναλύονται οι τέσσερις αλγόριθμοι που αναφέρονται παραπάνω, YamNet για κατηγοριοποίηση συμβάντων ήχου, Deep FilterNet για καθαρισμό σήματος ήχου, Whisper για ακριβή μετατροπή ομιλίας σε κείμενο και BART για ανάλυση κειμένου. Μέσω της χρήσης αυτών των αλγορίθμων, στόχος είναι η αυτοματοποίηση των διαφόρων ετικετών, ο καθαρισμός ήχου, η μετατροπή ομιλίας σε κείμενο και στη συνέχεια η ανάλυση του περιεχομένου των podcast, με απώτερο στόχο τη βελτίωση της προσβασιμότητας, της ευρετηρίασης και της κατανόησής τους.

«Analysis of audio files (podcasts) in terms of content using Machine Learning»

«Alkiviadis Mirovalis»

Abstract

This thesis examines the application of machine learning algorithms to audio data, with a particular focus on podcasts. The aim of the thesis is to investigate algorithms, such as YamNet, Whisper, Deep FilterNet and BART, in the automatic analysis and understanding of podcast content. First, a brief historical review of the evolution of audio technology and the role it played in the development of digital processing is formulated. Then, the four algorithms mentioned above are discussed, YamNet for audio event categorization, Deep FilterNet for audio signal cleaning, Whisper for accurate speech-to-text conversion, and BART for text analysis. Through the use of these algorithms, the goal is to automate the various tagging, audio cleaning, speech-to-text conversion and then content analysis of podcasts, with the ultimate goal of improving accessibility, indexing and comprehension.

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον Καθηγητή και Αντιπρύτανη του Διεθνούς Πανεπιστημίου, Κωνσταντίνο Διαμαντάρα, για την καθοδήγηση και την πολύτιμη υποστήριξή του κατά τη διάρκεια εκπόνησης της παρούσας εργασίας. Η γνώση και η εμπειρία του στον τομέα της μηχανικής μάθησης υπήρξαν καθοριστικής σημασίας για την επιτυχή ολοκλήρωσή της. Επιπλέον, ευχαριστώ θερμά την ομάδα του European School Radio για την παροχή των απαραίτητων δεδομένων και την άψογη συνεργασία τους. Τέλος, θα ήθελα να ευχαριστήσω από καρδιάς την οικογένεια και τους φίλους μου, για την στήριξη και την κατανόησή τους καθ' όλη τη διάρκεια αυτής της απαιτητικής προσπάθειας.

Περιεχόμενα

Κατάλογος Σχημάτων

Συντομογραφίες

ΔΙΠΑΕ	Διεθνές Πανεπιστήμιο Ελλάδος
Π.Ε.	Πτυχιακή Εργασία
ESR	European School Radio
DAW	Digital Audio Workstation
FM	Frequency Modulation
AM	Amplitude Modulation
ML	Machine Learning
DSP	Digital Signal Processing
AI	Artificial Intelligence
NLP	Natural Language Processing
ID3	Iterative Dichotomiser 3
MP3	MPEG Audio Layer 3
FLAC	Free Lossless Audio Codec
WAV	Waveform Audio File Format
STT	Speech To Text
HMM	Hidden Markov Models
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
API	Application Programming Interface
LLC	Library Of Congress Classification
DDC	Dewey Decimal Classification
MODS	Metadata Object Description Schema
XAI	Explainable Artificial Intelligence
ASR	Automatic Speech Recognition
BART	Bidirectional and Auto-regressive Transformers
IDE	Integrated Development Environment

Κεφάλαιο 1ο: Βασικές Αρχές της Επεξεργασίας Ήχου και της Μηχανικής Μάθησης

1.1 Εισαγωγή

Η επεξεργασία ήχου και η μηχανική μάθηση αποτελούν δύο από τους πιο δυναμικούς και ταχέως εξελισσόμενους τομείς της σύγχρονης τεχνολογίας. Η συνδυαστική χρήση αυτών των τεχνολογιών έχει οδηγήσει σε σημαντικές καινοτομίες, επιτρέποντας την ανάπτυξη προηγμένων συστημάτων αναγνώρισης και κατηγοριοποίησης ηχητικού περιεχομένου. Η παρούσα πτυχιακή εργασία επικεντρώνεται στην εισαγωγή στις βασικές αρχές της επεξεργασίας ήχου και της μηχανικής μάθησης, καθώς και στην εφαρμογή τους σε πραγματικά προβλήματα. Στόχος της εργασίας είναι να αναδείξει τη σημασία της διασύνδεσης αυτών των δύο πεδίων και να παρουσιάσει τις δυνατότητες που προσφέρουν για την ανάλυση και επεξεργασία ηχητικών δεδομένων καθώς και με ποιον τρόπο λειτουργούν. Μέσω της μελέτης και της εφαρμογής αλγορίθμων, η εργασία αυτή φιλοδοξεί να συμβάλει στην κατανόηση και την περαιτέρω ανάπτυξη των τεχνολογιών αυτών.

1.2 Ιστορική Αναδρομή στην Επεξεργασία Ήχου και η Σύνδεσή της με την Υπολογιστική Τεχνολογία

Η επεξεργασία του ήχου ξεκινά από τις πρώτες προσπάθειες καταγραφής και αναπαραγωγής ήχου τον 19ο αιώνα. Η εξέλιξη της επεξεργασίας του έχει επηρεάσει βαθιά τον τρόπο με τον οποίο ακούμε και δημιουργούμε μουσική, ταινίες και άλλα μέσα.Η ιστορία της ηχογράφησης ξεκινά με την εφεύρεση του φωνογράφου από τον Thomas Edison το 1877 [1]. Η συσκευή αυτή χρησιμοποιούσε ένα μεγάλο κωνικό κέρας το οποίο στην ουσία, είναι ένας σωλήνας με σχήμα κώνου που λειτουργεί ως ηχητικός αγωγός για να συλλέγει και να εστιάζει τα ηχητικά κύματα, τα οποία στη συνέχεια καταγράφονταν σε ένα κινούμενο μέσο, όπως ένας κύλινδρος επικαλυμμένος με κερί. Οι πρώτες ηχογραφήσεις ήταν χαμηλής πιστότητας και έντασης, αλλά άνοιξαν τον δρόμο για μελλοντικές εξελίξεις.Σημαντική ήταν και η συμβολή του Édouard-Léon Scott de Martinville με τον φωνοαυτόγραφο, που κατέγραψε ηχητικά κύματα σε χαρτί επικαλυμμένο με αιθάλη η οποία ειναί ουσιαστικά ένας τύπος καπνού που μπορεί να κατακάθεται σε επιφάνειες και να δημιουργεί ένα μαύρο, καπνισμένο επίστρωμα [2].

Η εισαγωγή των ηλεκτρικών μικροφώνων και ενισχυτών τη δεκαετία του 1920 έφερε μια επανάσταση στην ηχογράφηση [3]. Οι συγκεκριμένες τεχνολογίες επέτρεψαν την καταγραφή ήχου με μεγαλύτερη πιστότητα και ευκρίνεια, ενώ η χρήση περισσότερων μικροφώνων και ηλεκτρονικών φίλτρων βελτίωσε σημαντικά την ποιότητα των ηχογραφήσεων. Η τεχνολογία αυτή επέτρεψε την ανάπτυξη νέων μουσικών ειδών και την εξέλιξη της ραδιοφωνίας. Μετά τον Β' Παγκόσμιο Πόλεμο, η τεχνολογία της μαγνητικής κυριάρχησε στον τομέα ηχογράφησης [4]. Το γεγονός αυτό συνέβη διότι παρείχε υψηλότερη πιστότητα και μεγαλύτερη ευελιξία στην επεξεργασία ήχου, επιτρέποντας την επανεγγραφή και τη επεξεργασία των ηχογραφήσεων με καλύτερη ακρίβεια. Η ανάπτυξη των πολυκάναλων ηχογραφήσεων, επέτρεψαν την καταγραφή και την ανάμειξη πολλαπλών πηγών ήχου. Η μαγνητική ταινία έφερε επανάσταση στη μουσική βιομηχανία, επιτρέποντας την ανάπτυξη της στερεοφωνικής ηχογράφησης και την εισαγωγή νέων μορφών, όπως οι κασέτες και τα 8-track. Η εισαγωγή της

ψηφιακής τεχνολογίας τη δεκαετία του 1970 έφερε μια νέα επανάσταση στην επεξεργασία ήχου [5]. Οι ψηφιακές ηχογραφήσεις προσφέρουν εξαιρετική πιστότητα και ευελιξία, επιτρέποντας την επεξεργασία και την ανάμειξη ήχου με ακρίβεια που δεν ήταν δυνατή με τις αναλογικές τεχνολογίες.

Η ανάπτυξη των προσωπικών υπολογιστών και των λογισμικών επεξεργασίας ήχου έχει καταστήσει την επεξεργασία ήχου προσιτή σε όλους, από επαγγελματίες μουσικούς μέχρι ερασιτέχνες δημιουργούς. Η ψηφιακή τεχνολογία έχει επίσης επιτρέψει την ανάπτυξη νέων μορφών διανομής μουσικής, όπως τα CD, τα MP3 και οι υπηρεσίες streaming. Η ιστορία της επεξεργασίας ήχου είναι μια ιστορία συνεχούς εξέλιξης και καινοτομίας. Από τις πρώτες μηχανικές συσκευές μέχρι τις σύγχρονες ψηφιακές τεχνολογίες, η επεξεργασία ήχου έχει επηρεάσει βαθιά τον τρόπο με τον οποίο ακούμε και δημιουργούμε μουσική, ταινίες και άλλα μέσα. Η σύνδεση της επεξεργασίας ήχου με την υπολογιστική τεχνολογία έχει ανοίξει νέους ορίζοντες και έχει καταστήσει δυνατή τη δημιουργία και την επεξεργασία ήχου με τρόπους που ήταν αδιανόητη στο παρελθόν. Η συνεχής εξέλιξη της τεχνολογίας υπόσχεται ακόμη περισσότερες καινοτομίες στο μέλλον, καθιστώντας την επεξεργασία ήχου ένα από τα πιο δυναμικά και συναρπαστικά πεδία της σύγχρονης τεχνολογίας.

1.2.1 Σημαντικά Ορόσημα στην Ιστορία του Ήχου και της Ηχογράφησης

Η ιστορία της ηχογράφησης είναι γεμάτη με σημαντικά ορόσημα που έχουν διαμορφώσει τον τρόπο με τον οποίο καταγράφουμε και αναπαράγουμε ήχο. Από τις πρώτες μηχανικές συσκευές μέχρι τις σύγχρονες ψηφιακές τεχνολογίες, η εξέλιξη της ηχογράφησης έχει επηρεάσει βαθιά τη μουσική βιομηχανία και την καθημερινή μας ζωή.

Η Ακουστική Εποχή (1877–1925)

Η πρώτη πρακτική ηχογράφηση έγινε το 1877 με την εφεύρεση του φωνογράφου από τον Thomas Edison [6]. Αυτές οι πρώιμες ηχογραφήσεις ήταν χαμηλής πιστότητας και έντασης, αλλά άνοιξαν τον δρόμο για μελλοντικές εξελίξεις.

Η Ηλεκτρική Εποχή (1925–1945)

Το 1925, η εισαγωγή του ολοκληρωμένου συστήματος ηλεκτρικών μικροφώνων και ενισχυτών από την Western Electric βελτίωσε σημαντικά την ποιότητα της ηχογράφησης [7]. Αυτή η εποχή είδε την ανάπτυξη του επαγγελματία μηχανικού ήχου και την εισαγωγή της ηλεκτρονικής ενίσχυσης.

Η Μαγνητική Εποχή (1945–1975)

Μετά τον Β' Παγκόσμιο Πόλεμο, η τεχνολογία της μαγνητικής ταινίας έγινε το πρότυπο για την ηχογράφηση ήχου [8]. Η μαγνητική ταινία επέτρεψε μεγαλύτερη πιστότητα και ευελιξία στην επεξεργασία του ήχου, οδηγώντας στην ανάπτυξη της ηχογράφησης πολλαπλών καναλιών.

Η Ψηφιακή Εποχή (1975–σήμερα)

Η ψηφιακή ηχογράφηση ξεκίνησε τη δεκαετία του 1970 και γρήγορα αντικατέστησε τις αναλογικές τεχνολογίες [9]. Η εισαγωγή του CD από τη Sony και τη Philips το 1982 έφερε επανάσταση στην καταναλωτική αγορά ήχου, ενώ οι ψηφιακοί σταθμοί εργασίας ήχου (DAW) έχουν γίνει το πρότυπο στα επαγγελματικά στούντιο.

Η εξέλιξη της ηχογράφησης από τις μηχανικές συσκευές του 19ου αιώνα μέχρι τις σύγχρονες ψηφιακές τεχνολογίες έχει αλλάξει ριζικά τον τρόπο με τον οποίο καταγράφουμε και αναπαράγουμε ήχο. Κάθε εποχή έφερε νέες καινοτομίες που βελτίωσαν την ποιότητα και την ευελιξία της ηχογράφησης.

1.2.2 Η Επίδραση της Υπολογιστικής Τεχνολογίας στην Επεξεργασία Ήχου

Η Επίδραση της Υπολογιστικής Τεχνολογίας στην Επεξεργασία Ήχου έχει αλλάξει ριζικά τον τρόπο με τον οποίο επεξεργαζόμαστε και χειριζόμαστε τον ήχο. Από την εγγραφή και την επεξεργασία μέχρι την αναπαραγωγή και τη διανομή. Αυτές οι αλλαγές επηρέασαν βαθιά τόσο τη μουσική βιομηχανία όσο και τον τρόπο με τον οποίο καταναλώνουμε μουσική στην καθημερινή μας ζωή.

Ψηφιακή Επεξεργασία Ήχου

Η έλευση της ψηφιακής τεχνολογίας στη δεκαετία του 1970 επέτρεψε την ακριβή καταγραφή και αναπαραγωγή ήχου. Οι ψηφιακοί σταθμοί εργασίας ήχου (DAW) κατέστησαν δυνατή τη λεπτομερή επεξεργασία ηχητικών κομματιών, προσφέροντας εργαλεία για κοπή, αντιγραφή, μίξη και εφαρμογή εφέ. Αυτές οι διαδικασίες που άλλοτε απαιτούσαν εξειδικευμένες αναλογικές συσκευές, τώρα μπορούν να γίνουν σε υπολογιστή με λογισμικό ψηφιακής επεξεργασίας ήχου [10].

Η Ποιότητα και η Πιστότητα του Ήχου

Η υπολογιστική τεχνολογία επέτρεψε τη δημιουργία ηχητικών αρχείων υψηλής πιστότητας, όπως είναι για παράδειγμα τα αρχεία FLAC και WAV, τα οποία διατηρούν την ποιότητα του αρχικού ήχου χωρίς απώλειες. Ταυτόχρονα, οι προηγμένοι αλγόριθμοι συμπίεσης, όπως το MP3 και το AAC, επιτρέπουν την αποθήκευση μεγάλων ποσοτήτων δεδομένων ήχου σε μικρότερα μεγέθη αρχείων, καθιστώντας την αναπαραγωγή και τη διανομή πιο αποδοτική [10].

Επαγγελματικά Στούντιο Επεξεργασίας Ήχου

Η υπολογιστική τεχνολογία έχει μειώσει σημαντικά το κόστος και την πολυπλοκότητα της δημιουργίας επαγγελματικών στούντιο ήχου. Σήμερα, ακόμη και μικρά στούντιο ή ανεξάρτητοι καλλιτέχνες μπορούν να έχουν πρόσβαση σε εργαλεία και τεχνολογίες που παλαιότερα ήταν διαθέσιμες μόνο σε μεγάλες εταιρείες παραγωγής [10].

Τεχνολογία Τεχνητής Νοημοσύνης και Μηχανικής Μάθησης

Η τεχνητή νοημοσύνη και η μηχανική μάθηση έχουν αρχίσει να παίζουν σημαντικό ρόλο στην επεξεργασία ήχου. Εργαλεία βασισμένα σε AI μπορούν να αναλύσουν και να βελτιώσουν την ποιότητα του ήχου, να αφαιρέσουν τον θόρυβο, να εφαρμόσουν αυτόματη ρύθμιση τόνου και να δημιουργήσουν νέα ηχητικά εφέ που μέχρι πρότινος ήταν αδύνατα να επιτευχθούν [10].

Ηλεκτρονική Μουσική και Παραγωγή

Η υπολογιστική τεχνολογία έχει επίσης οδηγήσει στην ανάπτυξη της ηλεκτρονικής μουσικής. Λογισμικά σύνθεσης και επεξεργασίας ήχου επιτρέπουν στους μουσικούς να δημιουργούν και να επεξεργάζονται ήχους που δεν υπάρχουν στη φύση, ανοίγοντας νέους ορίζοντες στη μουσική δημιουργία [10].

1.2.3 Η Εξέλιξη του Ραδιοφώνου

Η εξέλιξη της ραδιοτεχνολογίας αντιπροσωπεύει μια παραδειγματική αλλαγή στην ανθρώπινη επικοινωνία, σηματοδοτώντας μια επαναστατική πρόοδο στη διάδοση της πληροφορίας και της πολιτιστικής ανταλλαγής. Από τα θεωρητικά θεμέλιά της στις προβλέψεις ηλεκτρομαγνητικών κυμάτων του Maxwell έως την πειραματική επικύρωση του Heinrich Hertz, η ραδιοτεχνολογία προέκυψε μέσω καθοριστικών συνεισφορών από εφευρέτες όπως ο Guglielmo Marconi και ο Nikola Tesla στις αρχές του 20ού αιώνα. Το επίτευγμα της υπερατλαντικής ραδιοφωνικής μετάδοσης το 1901 προανήγγειλε μια νέα εποχή παγκόσμιας συνδεσιμότητας, ενώ οι επακόλουθες τεχνολογικές καινοτομίες -κυρίως η ανάπτυξη της διαμόρφωσης πλάτους (AM) και της διαμόρφωσης συχνότητας (FM) βελτίωσαν σημαντικά την ποιότητα του σήματος και τις δυνατότητες εκπομπής. Η μετάβαση από τους σωλήνες κενού στα τρανζίστορ στα μέσα του 20ου αιώνα έφερε επανάσταση στην προσβασιμότητα και τη φορητότητα του ραδιοφώνου, ενώ η μετέπειτα εμφάνιση των ψηφιακών συστημάτων ραδιοφωνικής μετάδοσης βελτίωσε περαιτέρω την ακρίβεια μετάδοσης και την αποτελεσματικότητα του φάσματος [11].

Πέρα από την τεχνική του εξέλιξη, το ραδιόφωνο μεταμόρφωσε βαθιά την κοινωνία, λειτουργώντας ως κρίσιμο μέσο για τη διάδοση ειδήσεων κατά τη διάρκεια μεγάλων ιστορικών γεγονότων, ενισχύοντας την πολιτιστική ομογενοποίηση μέσω της μαζικής ψυχαγωγίας και δημιουργώντας πρωτοφανείς πλατφόρμες για δημόσιο λόγο και εμπορική διαφήμιση. Αυτή η τεχνολογική πρόοδος, από την πειραματική της αρχή έως την τρέχουσα ψηφιακή της ενσάρκωση, αποτελεί παράδειγμα της περίπλοκης σχέσης μεταξύ της επιστημονικής καινοτομίας και του κοινωνικού μετασχηματισμού, τοποθετώντας το ραδιόφωνο ως θεμελιώδες στοιχείο στο σύγχρονο επικοινωνιακό τοπίο. Η τροχιά της εξέλιξης του ραδιοφώνου συνεχίζει να επιταχύνεται στην ψηφιακή εποχή. Η εμφάνιση του διαδικτυακού ραδιοφώνου, της δορυφορικής μετάδοσης και της ψηφιακής εκπομπής ήχου (DAB) έχει επεκτείνει την εμβέλεια του ραδιοφώνου πέρα από τους παραδοσιακούς επίγειους περιορισμούς, διατηρώντας παράλληλα τη βασική του λειτουργία ως άμεσο και προσβάσιμο μέσο [11].

Οι κινητές συσκευές και οι πλατφόρμες ροής έχουν αλλάξει τον τρόπο με τον οποίο το κοινό αλληλεπιδρά με το ραδιοφωνικό περιεχόμενο, επιτρέποντας την ακρόαση κατ' απαίτηση παράλληλα με τις ζωντανές μεταδόσεις. Παρά τις προβλέψεις για την απαρχαιότητά του εν όψει των νεότερων τεχνολογιών μέσων, το ραδιόφωνο έχει επιδείξει αξιοσημείωτη ανθεκτικότητα και προσαρμοστικότητα. Η οικειότητά του ως μέσου που καθοδηγείται από τη φωνή, η ικανότητά του να εξυπηρετεί τις τοπικές κοινότητες και ο ρόλος του ως πηγή πληροφοριών έκτακτης ανάγκης παραμένουν μοναδικά πολύτιμοι. Η ενσωμάτωση των μέσων κοινωνικής δικτύωσης και των διαδραστικών ψηφιακών πλατφορμών έχει βελτιώσει αντί να μειώσει τη σύνδεση του ραδιοφώνου με τους ακροατές, δημιουργώντας νέες μορφές δέσμευσης, διατηρώντας παράλληλα τα ιδιαίτερα χαρακτηριστικά αμεσότητας και προσβασιμότητας. Κοιτάζοντας το μέλλον, η εξέλιξη του ραδιοφώνου δείχνει προς την αυξημένη εξατομίκευση, τη βελτιωμένη ψηφιακή ολοκλήρωση και τη συνεχή τεχνολογική καινοτομία, διατηρώντας παράλληλα τον ουσιαστικό ρόλο του ως ζωτικού καναλιού για ανθρώπινη σύνδεση και επικοινωνία [11].

Η συνεχής εξέλιξη της ραδιοτεχνολογίας έχει επηρεαστεί θετικά από τεχνικές καινοτομίες που έχουν βελτιώσει τις δυνατότητές της και έχουν διευρύνει την εμβέλειά της. Η ανάπτυξη του ραδιοφώνου που ορίζεται από λογισμικό (SDR) επέτρεψε μεγαλύτερη ευελιξία στην επεξεργασία σήματος και στα σχήματα διαμόρφωσης, ενώ τα γνωστικά ραδιοφωνικά συστήματα μπορούν να προσαρμοστούν δυναμικά στη διαθεσιμότητα του φάσματος και στις συνθήκες παρεμβολής. Αυτές οι εξελίξεις έχουν κάνει τις ραδιοεπικοινωνίες πιο αποτελεσματικές και ανθεκτικές. Η εμφάνιση του ψηφιακού ραδιοφώνου mondiale (DRM) έφερε ήχο ψηφιακής ποιότητας στις παραδοσιακές ζώνες AM/SW/MW, ενώ το HD Radio έχει βελτιώσει την πιστότητα των εκπομπών FM. Στον τομέα του δορυφορικού ραδιοφώνου, υπηρεσίες όπως το SiriusXM έχουν δημιουργήσει νέα μοντέλα που βασίζονται σε συνδρομές για την παροχή περιεχομένου σε ηπειρωτικές περιοχές χωρίς περιορισμούς επίγειας υποδομής. Ο πολιτιστικός αντίκτυπος του ραδιοφώνου συνεχίζει να εξελίσσεται παράλληλα με αυτές τις τεχνικές προόδους [11].

Οι κοινοτικοί ραδιοφωνικοί σταθμοί έχουν καταστεί ζωτικής σημασίας πλατφόρμες για τη διατήρηση των τοπικών γλωσσών, παραδόσεων και μουσικής κληρονομιάς, ιδιαίτερα στις αναπτυσσόμενες περιοχές. Τα εκπαιδευτικά ραδιοφωνικά προγράμματα προσφέρουν ευκαιρίες εξ αποστάσεως εκπαίδευσης σε απομακρυσμένες περιοχές, ενώ οι εξειδικευμένες μορφές ραδιοφώνου καλύπτουν όλο και πιο διαφορετικά ενδιαφέροντα ακροατών. Η άνοδος του podcasting, αν και τεχνικά διαφέρει από το παραδοσιακό ραδιόφωνο, αντιπροσωπεύει μια εξέλιξη της θεμελιώδους γοητείας του ραδιοφώνου μεταξύ ομιλητή και ακροατή. Η ραδιοφωνική δημοσιογραφία παραμένει μια κρίσιμη πηγή άμεσης, αξιόπιστης πληροφόρησης κατά τη διάρκεια κρίσεων και φυσικών καταστροφών, με την υποδομή της να αποδεικνύεται συχνά πιο ανθεκτική από άλλα δίκτυα επικοινωνίας. Στον εμπορικό τομέα, το ραδιόφωνο έχει προσαρμοστεί στις μεταβαλλόμενες συνθήκες της αγοράς μέσω της αυξημένης εξειδίκευσης σε μορφή, του στοχευμένου προγραμματισμού και της ενσωμάτωσης με ψηφιακές πλατφόρμες [11].

Η ραδιοφωνική διαφήμιση συνεχίζει να εξελίσσεται, αγκαλιάζοντας την αγορά διαφημίσεων μέσω προγραμματισμού, βελτιωμένα αναλυτικά στοιχεία κοινού και ενσωμάτωση καμπανιών μεταξύ πλατφορμών. Ο κλάδος γνώρισε επίσης σημαντική ενοποίηση, με μεγάλους ραδιοτηλεοπτικούς ομίλους να επιτυγχάνουν οικονομίες κλίμακας ενώ εκφράζουν ανησυχίες για την ποικιλομορφία των μέσων ενημέρωσης και τον τοπικισμό. Παρά αυτές τις προκλήσεις, ανεξάρτητες και εναλλακτικές ραδιοφωνικές εξόδους συνεχίζουν να ευδοκιμούν, αξιοποιώντας συχνά τη ροή στο Διαδίκτυο για να προσεγγίσουν παγκόσμιο κοινό, διατηρώντας παράλληλα ισχυρές τοπικές συνδέσεις. Κοιτάζοντας το μέλλον, το ραδιόφωνο αντιμετωπίζει ευκαιρίες και προκλήσεις. Η δυνατότητα των δικτύων 5G να επιτρέψουν υβριδικές ευρυζωνικές υπηρεσίες εκπομπής θα μπορούσε να μεταμορφώσει την παράδοση και την αλληλεπίδραση ραδιοφώνου. Η τεχνητή νοημοσύνη και η μηχανική μάθηση μπορεί να φέρουν επανάσταση στη δημιουργία περιεχομένου, τον προγραμματισμό και την αφοσίωση του κοινού. Ωστόσο, παραμένουν ερωτήματα σχετικά με την κατανομή του φάσματος, τη διαχείριση ψηφιακών δικαιωμάτων και την ισορροπία μεταξύ τοπικισμού και ενοποίησης [11].

Καθώς το ραδιόφωνο μπαίνει στον δεύτερο αιώνα του, τα θεμελιώδη πλεονεκτήματά του, η αμεσότητα, η προσβασιμότητα και η σύνδεση με την κοινότητα το τοποθετούν ως ένα ζωτικό και εξελισσόμενο μέσο στην ψηφιακή εποχή.

1.2.4 Το European School Radio

Το Ευρωπαϊκό Σχολικό Ραδιόφωνο (ESR) ξεκίνησε στην Ελλάδα το 2009 και επίσημα το 2011. Από μια εκπαιδευτική πρωτοβουλία, έχει εξελιχθεί σε πανευρωπαϊκό εκπαιδευτικό ραδιόφωνο. Όπως και επίσημη ονομασία του φορέα μου το λειτουργεί, προάγει το διαθεματικό, διαπολιτισμικό περιεχόμενο χωρίς διακρίσεις και προσφέρει ισάξια σε κάθε μαθητή/τρια το δικαίωμα να συμμετέχει μέσω του σχολείου ή του οργανισμού του. Από το 2021 έχει έδρα στο Διεθνές Πανεπιστήμιο της Ελλάδος στη Θεσσαλονίκη, διασφαλίζοντας την πρόοδο στην επιστήμη μαζί με την πράξη ενός Μέσου που εξελίσσει συνεχώς του διαδικτυακά του εργαλεία, τις υποδομές και τις υπηρεσίες στο ευρύ εκπαιδευτικό κοινό. Οι βασικοί του στόχοι είναι η ανάπτυξη δεξιοτήτων στα μέσα επικοινωνίας, η πολιτιστική ανταλλαγή, η πρακτική στη ραδιοφωνική παραγωγή και η προώθηση της δημοκρατικής συμμετοχής. Παράλληλα, αναπτύσσει δράσεις παραγωγής ποιοτικού περιεχομένου μαζί με το Τμήμα Δημοσιογραφίας και ΜΜΕ του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης και της Ένωσης Συντακτών Μακεδονίας- Θράκης. (ΕΣΗΕΜ-Θ), στο πεδίο της μαθητικής δημοσιογραφίας. Το ESR προσφέρει ζωντανές και ηχογραφημένες εκπομπές, διάφορα είδη προγραμματισμού και έχει ένα αρχείο podcast με πάνω από 10.000 παραγωγές, διασφαλίζοντας ότι η μουσική που μεταδίδεται είναι νόμιμη και ηθική. Η τεχνολογική υποδομή του περιλαμβάνει δυνατότητες για ζωντανή και καταγεγραμμένη μετάδοση, προσφέροντας πλούσιο περιεχόμενο για τους ακροατές. Οι μαθητές συμμετέχουν ενεργά στην παραγωγή και παρουσίαση εκπομπών, αποκτώντας πρακτική εμπειρία στη ραδιοφωνική παραγωγή και ανάπτυξη δεξιοτήτων. Το ESR έχει λάβει διακρίσεις σε ευρωπαϊκό, όπως τα Medea Awards το 2014 για το εκπαιδευτικό περιεχόμενο που δημιουργείται από τους χρήστες του (user-generated content type ή UGC type), τον ειδικό έπαινο στα EU Web Awards το 2020 για τις διαδικτυακές υπηρεσίες σχεδιασμένες αποκλειστικά για τη νεολαία.

Στο πλαίσιο ευρωπαϊκών προγραμμάτων, συνεργάζεται με εκπαιδευτικά ιδρύματα σε όλη την Ευρώπη, όπως το Παιδαγωγικό Ινστιτούτο Κύπρου, το Πανεπιστήμιο Aalborg της Δανίας, τον Οργανισμό Les Francas στη Γαλλία, τον Οργανισμό Technische Jugendfreizeit- und Bildungsgesellschaft (tjfbg) gemeinnützige GmbH στη Γερμανία, διευκολύνοντας την ανταλλαγή βέλτιστων πρακτικών και καινοτόμων εκπαιδευτικών μεθοδολογιών. Επιπλέον, το ESR φιλοξενεί διαδραστικά εργαστήρια, εκπαιδευτικές συνεδρίες, καλλιτεχνικές εκδηλώσεις, συνεργατικές παραγωγές, ετήσια μαθητικά φεστιβάλ ραδιοφώνου και διεθνείς μαθητικούς μουσικο-ραδιοφωνικούς διαγωνισμούς. Αυτές οι δραστηριότητες ενισχύουν την ψηφιακή ενσωμάτωση, επεκτείνουν τη διεθνή εμβέλεια του ESR, αναπτύσσουν νέα διαδραστικά εργαλεία και προωθούν τις διαπολιτισμικές εκπαιδευτικές σχέσεις. Το ESR στοχεύει να προωθήσει τον γραμματισμό στα μέσα επικοινωνίας και την ψηφιακή ιθαγένεια, ενθαρρύνοντας τη μάθηση μέσω έργων, την ανάπτυξη δημιουργικότητας και την προώθηση της συμμετοχής των μαθητών σε δημοκρατικές διαδικασίες. Η επιτυχία του ESR δείχνει πώς η παραδοσιακή εκπομπή μπορεί να προσαρμοστεί στις σύγχρονες εκπαιδευτικές ανάγκες, υποστηρίζοντας τη μάθηση, προωθώντας τον ψηφιακό γραμματισμό και διευκολύνοντας πολιτιστικές ανταλλαγές. Με αυτές τις πρωτοβουλίες, το ESR συμβάλλει στην ολοκληρωμένη ανάπτυξη των μαθητών, προετοιμάζοντάς τους για ενεργή και ενημερωμένη συμμετοχή σε έναν ψηφιακά διασυνδεδεμένο κόσμο.[12].

1.3 Ιστορικη αναδρομη στην μηχανική μάθηση

Η μηχανική μάθηση (ML) είναι η επιστήμη της ανάπτυξης και χρήσης συστημάτων υπολογιστών που μαθαίνουν και προσαρμόζονται χωρίς να ακολουθούν σαφείς οδηγίες. Χρησιμοποιεί αλγόριθμους και

στατιστικά μοντέλα για την ανάλυση και την παραγωγή προγνωστικών αποτελεσμάτων από μοτίβα σε δεδομένα [15].

Η ιστορία της μηχανικής μάθησης ξεκινά με τις πρώτες προσπάθειες δημιουργίας νευρωνικών δικτύων και προχωρά σε σύγχρονες εφαρμογές τεχνητής νοημοσύνης. Το 1943, οι νευροεπιστήμονες Walter Pitts και Warren McCulloch δημοσίευσαν το πρώτο μαθηματικό μοντέλο ενός νευρωνικού δικτύου, το οποίο θεωρείται το έτος γέννησης της μηχανικής μάθησης [14]. Το 1950, ο Alan Turing δημοσίευσε το *Computing Machinery and Intelligence*, εισάγοντας το Test Turing και ανοίγοντας το δρόμο για την τεχνητή νοημοσύνη. Το 1952, ο Arthur Samuel δημιούργησε το πρώτο πρόγραμμα παιχνιδιών αυτομάθησης, το Samuel Checkers-Playing Program [16]. Στη δεκαετία του 1960 αναπτύχθηκαν τα πρώτα προγράμματα που αναγνώρισαν μοτίβα και χαρακτήρες σε χειρόγραφο κείμενο και το 1967 εισήχθη ο αλγόριθμος του πλησιέστερου γείτονα. Το 1970, ο Seppo Linnainmaa δημοσίευσε τον αλγόριθμο αντίστροφης διάδοσης, ο οποίος αποτελεί τη βάση για την εκπαίδευση των νευρωνικών δικτύων [17]. Η δεκαετία του 1980 έφερε την ανακάλυψη των επαναλαμβανόμενων νευρωνικών δικτύων (RNNs) από τον John Hopfield, ενώ το 1989 οι Yann LeCun, Yoshua Bengio και Patrick Haffner έδειξαν πώς τα συνελικτικά νευρωνικά δίκτυα (CNN) μπορούσαν να χρησιμοποιηθούν για την αναγνώριση χειρόγραφων χαρακτήρων [18].

Το 1997, οι Sepp Hochreiter και Jürgen Schmidhuber πρότειναν δίκτυα μακροπρόθεσμης και βραχυπρόθεσμης μνήμης (LSTM), τα οποία μπορούν να επεξεργάζονται ακολουθίες δεδομένων όπως ομιλία ή βίντεο [19]. Στη δεκαετία του 2000, η μηχανική μάθηση έγινε πιο δημοφιλής με την ανάπτυξη του ImageNet το 2009, το οποίο παρείχε μια τεράστια βάση δεδομένων εικόνων για την εκπαίδευση μοντέλων αναγνώρισης αντικειμένων. Το 2012, το AlexNet, ένα νευρωνικό δίκτυο, πέτυχε σφάλμα 15,3% στο ImageNet, πολύ υψηλότερο από οποιοδήποτε άλλο μοντέλο εκείνη την εποχή. Η δεκαετία του 2010 χαρακτηρίστηκε από ταχεία ανάπτυξη και καινοτομία στον τομέα της μηχανικής μάθησης. Τα νευρωνικά δίκτυα άρχισαν να γίνονται εξαιρετικά δημοφιλή, με εφαρμογές σε πολλούς τομείς, από την υγειονομική περίθαλψη μέχρι την αυτοκινητοβιομηχανία. Η βαθιά μάθηση, η οποία χρησιμοποιεί πολυεπίπεδα νευρωνικά δίκτυα, έγινε μια από τις πιο σημαντικές μεθόδους στη μηχανική μάθηση. Αυτά τα μοντέλα είναι ικανά να αναλύουν τεράστιες ποσότητες δεδομένων, να μαθαίνουν από αυτά και να βελτιώνουν συνεχώς την απόδοσή τους [15].

Έχουν επιτευχθεί εντυπωσιακά αποτελέσματα στους τομείς της αναγνώρισης εικόνας, της αυτόματης αναγνώρισης ομιλίας και της επεξεργασίας φυσικής γλώσσας. Η μηχανική μάθηση είναι ένα βασικό υποσύνολο της τεχνητής νοημοσύνης. Οι σύγχρονες εφαρμογές της τεχνητής νοημοσύνης περιλαμβάνουν συστήματα όπως αυτόνομα οχήματα, εικονικούς βοηθούς και προγνωστικές αναλύσεις [14]. Οι εταιρείες τεχνολογίας επενδύουν δισεκατομμύρια σε έρευνα και ανάπτυξη για να προωθήσουν την τεχνητή νοημοσύνη και να την ενσωματώσουν σε καταναλωτικά προϊόντα. Η ενισχυτική μάθηση κερδίζει έδαφος ως σημαντική τεχνολογία στη μηχανική μάθηση. Αυτή η προσέγγιση περιλαμβάνει συστήματα που μαθαίνουν να λαμβάνουν αποφάσεις μέσω δοκιμής και λάθους, επιβραβεύοντας τις σωστές ενέργειες και αποθαρρύνοντας τις λανθασμένες [17].

Σημαντικές επιτυχίες έχουν σημειωθεί στη ρομποτική και στα βιντεοπαιχνίδια. Η ανάπτυξη του cloud computing και των τεχνολογιών μεγάλων δεδομένων έχει επιταχύνει την πρόοδο της μηχανικής μάθησης. Οι πλατφόρμες cloud παρέχουν απεριόριστη υπολογιστική ισχύ και αποθήκευση, ενώ τα μεγάλα δεδομένα προσφέρουν τον απαραίτητο όγκο πληροφοριών για την εκπαίδευση πολύπλοκων μοντέλων [18]. Παρά τη σημαντική πρόοδο, η μηχανική μάθηση αντιμετωπίζει προκλήσεις όπως η εξήγηση των αποφάσεων των αλγορίθμων (το πρόβλημα του μαύρου κουτιού), η διασφάλιση του

απορρήτου των δεδομένων και η ηθική χρήση της τεχνολογίας. Η συνεχής έρευνα και καινοτομία αναμένεται να αντιμετωπίσουν αυτές τις προκλήσεις και να ανοίξουν νέους δρόμους για τη μηχανική μάθηση και την τεχνητή νοημοσύνη στο μέλλον.

1.4 Η Εφαρμογή της Μηχανικής Μάθησης στην Ανάλυση Ηχητικών Δεδομένων

Η μηχανική μάθηση (ML) έχει αναδειχθεί ως ένα ισχυρό εργαλείο για την ανάλυση δεδομένων ήχου, επιτρέποντας την εξαγωγή πολύτιμων πληροφοριών και την πρόβλεψη των αποτελεσμάτων σε διάφορους τομείς. Η ανάλυση δεδομένων ήχου περιλαμβάνει την επεξεργασία και ερμηνεία ηχητικών σημάτων για αναγνώριση, ταξινόμηση και πρόβλεψη προτύπων. Η εφαρμογή της ML στην ανάλυση δεδομένων ήχου έχει επεκταθεί σε πολλούς τομείς, όπως η υγεία, η ασφάλεια, η ψυχαγωγία και η βιομηχανία [20],[21].

Η διαδικασία ανάλυσης δεδομένων ήχου ξεκινά με τη συλλογή και την προεπεξεργασία των ηχητικών σημάτων. Αυτά τα σήματα μπορούν να προέρχονται από διάφορες πηγές, όπως μικρόφωνα, ηχογραφήσεις και αισθητήρες. Η προεπεξεργασία περιλαμβάνει την αφαίρεση του θορύβου, την κανονικοποίηση και τη μετατροπή των σημάτων σε μορφή κατάλληλη για ανάλυση. Ένα σημαντικό βήμα είναι ο μετασχηματισμός του σήματος από το πεδίο του χρόνου στο πεδίο της συχνότητας χρησιμοποιώντας τεχνικές όπως ο μετασχηματισμός Fourier [22].

Η εξαγωγή χαρακτηριστικών είναι ένα κρίσιμο στάδιο στην ανάλυση δεδομένων ήχου. Αυτά τα χαρακτηριστικά μπορεί να περιλαμβάνουν ενέργεια, συχνότητα, φάσμα και άλλα στατιστικά χαρακτηριστικά του σήματος. Η χρήση τεχνικών όπως οι εγκεφαλικοί συντελεστές Mel-Frequency Cepstral Coefficients (MFCCs) και τα φασματογράμματα επιτρέπει την αναπαράσταση ηχητικών σημάτων σε μορφή που μπορεί να χρησιμοποιηθεί από μοντέλα μηχανικής μάθησης [23],[24].

Τα μοντέλα μηχανικής μάθησης που χρησιμοποιούνται για την ανάλυση δεδομένων ήχου περιλαμβάνουν αλγόριθμους ταξινόμησης και παλινδρόμησης. Για παράδειγμα, τα συνελικτικά νευρωνικά δίκτυα (CNN) είναι ιδιαίτερα αποτελεσματικά στην ανάλυση φασματογραμμάτων, ενώ τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) και τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM) χρησιμοποιούνται για την ανάλυση ακολουθιών δεδομένων [25]. Η εφαρμογή της μηχανικής μάθησης στην ανάλυση δεδομένων ήχου έχει πολλές πρακτικές εφαρμογές. Στην υγειονομική περίθαλψη, μπορεί να χρησιμοποιηθεί για τον εντοπισμό παθολογιών της φωνής και την παρακολούθηση της ψυχικής υγείας μέσω ανάλυσης ομιλίας [21],[24]. Στην ασφάλεια, μπορεί να χρησιμοποιηθεί για ανίχνευση ανωμαλιών και αναγνώριση ήχου σε περιβάλλοντα επιτήρησης [22]. Στην ψυχαγωγία, μπορεί να χρησιμοποιηθεί για την αναγνώριση ειδών μουσικής και την ανάλυση συναισθημάτων στην ομιλία [20],[23].

Η συνεχής ανάπτυξη των αλγορίθμων μηχανικής μάθησης και η αύξηση της υπολογιστικής ισχύος αναμένεται να επεκτείνουν περαιτέρω τις εφαρμογές της ανάλυσης δεδομένων ήχου στο μέλλον. Η ικανότητα εξαγωγής πολύτιμων πληροφοριών από ηχητικά σήματα ανοίγει νέους ορίζοντες για την κατανόηση και τη χρήση αυτών των δεδομένων σε διάφορους τομείς [25].

1.5 Σκοπός, Στόχοι και Μεθοδολογία της Εργασίας

Η παρούσα πτυχιακή εργασία έχει ως στόχο τη διερεύνηση και εφαρμογή τεχνικών μηχανικής μάθησης για την ανάλυση και επεξεργασία ηχητικών δεδομένων, με έμφαση στη χρήση τους σε podcasts. Σκοπός

της εργασίας είναι η εύρεση αλγορίθμων που θα επιτρέπει την αυτοματοποιημένη επισήμανση, κατηγοριοποίηση και ανάλυση του περιεχομένου των podcasts, βελτιώνοντας την προσβασιμότητα και την κατανόησή τους.

Η εργασία επικεντρώνεται στη συλλογή και προεπεξεργασία δεδομένων από podcasts του European School Radio, την έρευνα και ανάπτυξη αλγορίθμων για την αναγνώριση και κατηγοριοποίηση ηχητικών γεγονότων, την εφαρμογή τεχνικών μετατροπής ομιλίας σε κείμενο και την ανάλυση του προκύπτοντος κειμένου, καθώς και τη βελτίωση της ποιότητας του ηχητικού σήματος μέσω τεχνικών αποθορυβοποίησης. Η μεθοδολογία της εργασίας εκτός από την τη συλλογή ηχητικών δεδομένων από τα podcasts του European School Radio παρουσιάζει και αναλύει μοντέλα ML όπως το YamNet,το Whisper,το Deep FilterNet και το BART. Η συλλογή δεδομένων περιλαμβάνει την αποθήκευση ηχητικών σημάτων από διάφορα podcasts του ESR, τα οποία προεπεξεργάζονται για την εξαγωγή χρήσιμων χαρακτηριστικών. Η προεπεξεργασία περιλαμβάνει την απομάκρυνση θορύβου και τη μετατροπή των σημάτων σε κατάλληλη μορφή για ανάλυση. Οι αλγόριθμοι που αναπτύσσονται περιλαμβάνουν τεχνικές για την αναγνώριση ηχητικών γεγονότων, τη μετατροπή ομιλίας σε κείμενο και την ανάλυση κειμένου. Η εργασία αυτή φιλοδοξεί να συμβάλει στην κατανόηση και την περαιτέρω ανάπτυξη των τεχνολογιών επεξεργασίας ήχου και του ML που χρησιμοποιείται, προσφέροντας νέες δυνατότητες για την ανάλυση και την αξιοποίηση ηχητικών δεδομένων. Επιπλέον, αναμένεται να προσφέρει σημαντικά οφέλη στην εκπαιδευτική κοινότητα, επιτρέποντας την καλύτερη κατανόηση και αξιοποίηση του περιεχομένου των podcasts.

1.6 Δομή της Εργασίας

Η παρούσα πτυχιακή εργασία είναι δομημένη σε τέσσερα κύρια κεφάλαια, καθένα από τα οποία καλύπτει διαφορετικές πτυχές της έρευνας και της ανάλυσης που πραγματοποιήθηκε. Αρχικά, παρέχεται μια εισαγωγή στις βασικές αρχές της επεξεργασίας ήχου και της μηχανικής μάθησης, περιλαμβάνοντας μια ιστορική αναδρομή στην εξέλιξη της τεχνολογίας ήχου και τη σύνδεσή της με την υπολογιστική τεχνολογία, καθώς και μια σύντομη ανασκόπηση της ιστορίας της μηχανικής μάθησης. Στη συνέχεια παρουσιάζεται η εφαρμογή της μηχανικής μάθησης στην ανάλυση ηχητικών δεδομένων και οι στόχοι της εργασίας. Η συγκεκριμένη πτυχιακή επικεντρώνεται στην ηχητική ετικετοποίηση (audio tagging), την ανάλυση φυσικής γλώσσας και την μετατροπή της σε κείμενο, την αποθορυβοποίηση ηχητικών και την ανάλυση κειμένου, αναλύοντας την ιστορία και την σημασία τους στη σημερινή εποχή. Αναλύονται οι αλγόριθμοι YamNet, Whisper, Deep FilterNet και BART, οι οποίοι χρησιμοποιούνται για την κατηγοριοποίηση ηχητικών γεγονότων, τη μετατροπή ομιλίας σε κείμενο, τον καθαρισμό του ηχητικού σήματος και την ανάλυση του κειμένου αντίστοιχα. Επίσης, παρουσιάζονται οι εφαρμογές των αποτελεσμάτων αυτών των αλγορίθμων. Ακολουθεί η περιγραφή της διαδικασίας δημιουργίας και προετοιμασίας του dataset των podcasts του European School Radio. Αναλύεται η εφαρμογή των αλγορίθμων YamNet, Whisper, Deep FilterNet και BART στην ανάλυση των podcasts και παρουσιάζονται τα ποσοτικά και ποιοτικά αποτελέσματα της εφαρμογής τους. Τέλος, περιλαμβάνονται τα συμπεράσματα από την εφαρμογή των αλγορίθμων, οι περιορισμοί της εργασίας και οι προκλήσεις που αντιμετωπίστηκαν. Επιπλέον παρουσιάζονται προτάσεις για μελλοντική έρευνα, επεκτάσεις της εργασίας, καθώς και οι δυνατότητες της εφαρμογής αυτής.

Κεφάλαιο 2ο: Αλγόριθμοι και Τεχνικές για την Ανάλυση Ήχου και Φυσικής Γλώσσας

2.1 Εισαγωγή

Η ανάλυση ήχου και της φυσικής γλώσσας αποτελούν έναν από τους πιο συναρπαστικούς και ταχέως αναπτυσσόμενους τομείς της τεχνητής νοημοσύνης και της μηχανικής μάθησης. Οι αλγόριθμοι και οι τεχνικές που χρησιμοποιούνται για την ανάλυση αυτών, έχουν τη δυνατότητα να μετασχηματίσουν τον τρόπο με τον οποίο αλληλεπιδρούμε με την τεχνολογία και να βελτιώσουν την κατανόηση και την επεξεργασία της ανθρώπινης επικοινωνίας. Η εργασία αυτή, εξετάζει διάφορους αλγόριθμους και τεχνικές που χρησιμοποιούνται για την ανάλυση ήχου και φυσικής γλώσσας, δίνοντας έμφαση στους αλγόριθμους YamNet, DeepFilterNet, Whisper και BART. Αυτοί οι αλγόριθμοι έχουν αποδειχθεί ιδιαίτερα αποτελεσματικοί σε ποικίλες εφαρμογές και προσφέρουν σημαντικές ευκαιρίες για καινοτομία και βελτίωση των υπαρχόντων συστημάτων. Η κατανόηση και η επεξεργασία του ήχου και της φυσικής γλώσσας είναι κρίσιμη για την ανάπτυξη προηγμένων συστημάτων τεχνητής νοημοσύνης που μπορούν να αλληλεπιδρούν με τους ανθρώπους με φυσικό και αποτελεσματικό τρόπο. Η εργασία αυτή στοχεύει να αναδείξει τη σημασία της συνδυαστικής χρήσης της επεξεργασίας ήχου και της μηχανικής μάθησης για την ανάλυση και κατηγοριοποίηση ηχητικών δεδομένων. Θα αναλύσουμε τη λειτουργία, τα πλεονεκτήματα και τις προκλήσεις των αλγορίθμων YamNet, DeepFilterNet, Whisper και BART, καθώς και τις δυνατότητες που προσφέρουν για μελλοντική έρευνα και ανάπτυξη. Σε αυτή την ενότητα, παρουσιάζεται η ιστορία της εξέλιξης της ανάλυσης ήχου και της φυσικής γλώσσας. Θα εξετάσουμε την εξέλιξη του Audio Tagging, την ανάπτυξη των τεχνολογιών Denoisers και τη βελτίωση ποιότητας ήχου, τις τεχνικές μετατροπής ομιλίας σε κείμενο, καθώς και την κατηγοριοποίηση κειμένου.

2.2 Ιστορική Αναδρομή στην Ανάλυση Ήχου και Φυσικής Γλώσσας

Η μελέτη της ακουστικής ανάλυσης και της επεξεργασίας φυσικής γλώσσας έχει εξελιχθεί τις τελευταίες δεκαετίες, αποτελώντας ένα σημαντικό πεδίο στο ευρύτερο πεδίο της ιστορίας. Στις αρχές του 20ου αιώνα, η εμφάνιση των τεχνολογιών εγγραφής σηματοδότησε την αρχή της συστηματικής ανάλυσης ήχου. Η εφεύρεση του φωνογράφου από τον Thomas Edison και η επακόλουθη ανάπτυξη του γραμμοφώνου από τον Emile Berliner έθεσαν τα θεμέλια για ηχογράφηση και αναπαραγωγή ήχου [26].

Στα μέσα του 20ου αιώνα, σημειώθηκαν σημαντικές εξελίξεις στα οπτικοακουστικά μέσα με την εισαγωγή του ραδιοφώνου και της τηλεόρασης. Στα τέλη του 20ου αιώνα παρατηρήθηκε η άνοδος των ψηφιακών τεχνολογιών, οι οποίες έφεραν επανάσταση στην καταγραφή και την ανάλυση. Η εισαγωγή ψηφιακών συσκευών εγγραφής, όπως οι δίσκοι συμπαγούς δίσκου (CD), επέτρεψε την αποθήκευση και αναπαραγωγή ήχου υψηλής ποιότητας. Η μετάβαση από την αναλογική στην ψηφιακή εγγραφή σηματοδότησε ένα σημαντικό ορόσημο στην ανάλυση ήχου. Οι ψηφιακές τεχνολογίες επέτρεψαν πιο ακριβή και αποτελεσματική εγγραφή, αποθήκευση και ανάλυση δεδομένων ήχου. Η ανάπτυξη αλγορίθμων λογισμικού για την αυτοματοποιημένη ανίχνευση και ταξινόμηση ήχων ενίσχυσε περαιτέρω τις δυνατότητες ανάλυσης ήχου [26].

Η επεξεργασία φυσικής γλώσσας (NLP) αναδείχθηκε ως κρίσιμος τομέας έρευνας, εστιάζοντας στην αλληλεπίδραση μεταξύ των υπολογιστών και της ανθρώπινης γλώσσας. Τα πρώιμα συστήματα NLP βασίζονταν σε κανόνες, αλλά οι εξελίξεις στη μηχανική μάθηση και την τεχνητή νοημοσύνη οδήγησαν

στην ανάπτυξη πιο εξελιγμένων μοντέλων ικανών να κατανοούν και να δημιουργούν ανθρώπινη γλώσσα [27],[29].

Η ενοποίηση της ακουστικής ανάλυσης και των τεχνολογιών γλώσσας έχει επηρεάσει βαθιά τη βιομηχανία της ψυχαγωγίας και των μέσων ενημέρωσης. Οι οπτικοακουστικές αναπαραστάσεις της ιστορίας, όπως τα ντοκιμαντέρ και τα ραδιοφωνικά χαρακτηριστικά, έχουν γίνει δημοφιλείς μορφές για τη μετάδοση της ιστορικής γνώσης. Τα ψηφιακά παιχνίδια και οι ταινίες μεγάλου μήκους χρησιμοποιούν επίσης αυτές τις τεχνολογίες για να δημιουργήσουν καθηλωτικές εμπειρίες. Στον τομέα της επικοινωνίας, οι τεχνολογίες NLP έχουν επιτρέψει την ανάπτυξη εικονικών βοηθών, chatbots και άλλων εφαρμογών που βασίζονται σε AI. Αυτές οι τεχνολογίες έχουν αλλάξει τον τρόπο με τον οποίο τα άτομα αλληλεπιδρούν με τις ψηφιακές συσκευές, καθιστώντας την επικοινωνία πιο φυσική και διαισθητική [28].

Η ιστορική τροχιά της ακουστικής ανάλυσης και των τεχνολογιών επεξεργασίας φυσικής γλώσσας υπογραμμίζει τη σημαντική συμβολή τους σε διάφορους τομείς. Από πρώιμες συσκευές εγγραφής ήχου έως προηγμένα συστήματα NLP, αυτές οι τεχνολογίες εξελίσσονται συνεχώς, διαμορφώνοντας τον τρόπο με τον οποίο κατανοούμε και αλληλεπιδρούμε με τον ήχο και τη γλώσσα.

2.2.1 Εξέλιξη του Audio Tagging

Η προσθήκη ετικετών ήχου έχει γίνει μια ζωτική πτυχή της διαχείρισης ψηφιακής μουσικής, δίνοντας τη δυνατότητα στους χρήστες να οργανώνουν, να κατηγοριοποιούν και να ανακτούν αποτελεσματικά αρχεία μουσικής. Η έννοια των ετικετών ID3(Iterative Dichotomiser 3) χρονολογείται από τις αρχές της δεκαετίας του 1990, όταν εισήχθη για πρώτη φορά η μορφή MP3. Μέχρι εκείνη τη στιγμή, δεν υπήρχε τυποποιημένος τρόπος για να συμπεριληφθούν μεταδεδομένα σε αρχεία ήχου, ωθώντας τους προγραμματιστές να πειραματιστούν με τρόπους ενσωμάτωσης πληροφοριών σε MP3 [36].

Το 1996, ο Eric Kemp ανέπτυξε την πρώτη έκδοση των ετικετών ID3, επιτρέποντας στους χρήστες να περιλαμβάνουν βασικές πληροφορίες για τη μουσική τους στα αρχεία MP3 τους. Το κίνητρο του Kemp ήταν να δημιουργήσει ένα σύστημα που θα μπορούσε να αποθηκεύει μεταδεδομένα μέσα στο ίδιο το αρχείο ήχου, καθιστώντας ευκολότερη τη διαχείριση και την οργάνωση ψηφιακών μουσικών συλλογών [37]. Με τα χρόνια, η μορφή ID3 έχει εξελιχθεί ώστε να περιλαμβάνει περισσότερες πληροφορίες και υποστήριξη για διαφορετικές μορφές ήχου. Σήμερα, οι πιο συχνά χρησιμοποιούμενες εκδόσεις των ετικετών ID3 είναι οι ID3v1, ID3v2.3 και ID3v2.4, καθεμία με συγκεκριμένα χαρακτηριστικά και περιορισμούς.

Η εξέλιξη των ετικετών ID3 καθοδηγείται από την ανάγκη για πιο ολοκληρωμένα μεταδεδομένα για τη βελτίωση της εμπειρίας χρήστη και της διαχείρισης μουσικής. Η μετάβαση από την αναλογική στην ψηφιακή εγγραφή σηματοδότησε ένα σημαντικό ορόσημο στην προσθήκη ετικετών ήχου. Οι ψηφιακές τεχνολογίες επέτρεψαν την πιο ακριβή και αποτελεσματική εγγραφή, αποθήκευση και ανάλυση δεδομένων ήχου [38]. Η ανάπτυξη αλγορίθμων λογισμικού για αυτοματοποιημένη ανίχνευση και ταξινόμηση ήχων ενίσχυσε περαιτέρω τις δυνατότητες σήμανσης ήχου.

Οι ετικέτες ID3 προσφέρουν πιο προηγμένες δυνατότητες από τα βασικά μεταδεδομένα. Ένα τέτοιο χαρακτηριστικό είναι το εξώφυλλο, το οποίο επιτρέπει την ενσωμάτωση ενός αρχείου εικόνας στο αρχείο ήχου που αντιπροσωπεύει το άλμπουμ ή το τραγούδι. Ένα άλλο προηγμένο χαρακτηριστικό είναι οι στίχοι, επιτρέποντας την προσθήκη στίχων τραγουδιών απευθείας σε ετικέτες ID3. Η λειτουργία BPM (beats per minute) χρησιμοποιείται συνήθως από DJ για τη δημιουργία λιστών αναπαραγωγής με σταθερό ρυθμό [39]. Οι ετικέτες ID3 παρέχουν έναν απλό και αποτελεσματικό τρόπο οργάνωσης ψηφιακών αρχείων μουσικής. Με την ενσωμάτωση περιγραφικών πληροφοριών στο αρχείο ήχου, όπως όνομα καλλιτέχνη, τίτλος άλμπουμ και όνομα κομματιού, γίνεται ευκολότερο να παρακολουθείτε μεμονωμένα τραγούδια και άλμπουμ.

Αυτό είναι ιδιαίτερα χρήσιμο για άτομα με μεγάλες μουσικές βιβλιοθήκες, καθώς εξαλείφει την ανάγκη να θυμάστε ονόματα αρχείων ή δομές φακέλων. Οι ετικέτες ID3 μπορούν επίσης να διευκολύνουν την ανακάλυψη και την αναζήτηση μουσικής. Παρέχοντας περιγραφικές πληροφορίες σχετικά με αρχεία ήχου, οι ετικέτες ID3 διευκολύνουν την εύρεση μουσικής που ταιριάζει με τα ενδιαφέροντα ή τη διάθεση του ακροατή [40]. Για παράδειγμα, εάν ένας ακροατής αναζητά μουσική από έναν συγκεκριμένο καλλιτέχνη, μπορεί να χρησιμοποιήσει μια λειτουργία αναζήτησης για να βρει γρήγορα όλα τα αρχεία ήχου με το όνομα αυτού του καλλιτέχνη ενσωματωμένο στην ετικέτα ID3.

Το μέλλον της τεχνολογίας σήμανσης ήχου είναι πιθανό να καθοδηγείται από τεχνολογίες τεχνητής νοημοσύνης και μηχανικής μάθησης. Αυτές οι τεχνολογίες μπορούν να βελτιώσουν την ακρίβεια και την αποτελεσματικότητα της προσθήκης ετικετών ήχου αυτοματοποιώντας τη διαδικασία παραγωγής μεταδεδομένων και βελτιώνοντας την ταξινόμηση αρχείων ήχου. Εταιρείες και ερευνητές εργάζονται συνεχώς για την ανάπτυξη πιο εξελιγμένων αλγορίθμων και εργαλείων για περαιτέρω προώθηση της τεχνολογίας σήμανσης ήχου.

Η ενσωμάτωση της τεχνητής νοημοσύνης και της μηχανικής μάθησης θα επιτρέψει πιο εξατομικευμένα και διαισθητικά συστήματα διαχείρισης μουσικής, διευκολύνοντας τους χρήστες να οργανώνουν και να διαχειρίζονται τις μουσικές τους συλλογές. Συμπερασματικά, οι ετικέτες ID3 έχουν παίξει καθοριστικό ρόλο στην οργάνωση και διαχείριση ψηφιακών αρχείων μουσικής. Με την τεράστια ποσότητα μουσικής που είναι διαθέσιμη σήμερα, μπορεί να είναι συντριπτική η παρακολούθηση μεμονωμένων αρχείων. εδώ μπαίνουν στο παιχνίδι οι ετικέτες ID3. Αυτές οι ετικέτες παρέχουν έναν τρόπο προσθήκης βασικών πληροφοριών σε ψηφιακά αρχεία ήχου, διευκολύνοντας την ταξινόμηση και την ανάκτηση.

Η χρήση των ετικετών ID3 περιλαμβάνει βελτιωμένη οργάνωση ψηφιακών αρχείων, βελτιωμένη εμπειρία αναπαραγωγής μουσικής, αυξημένη ευκολία ανακάλυψης και αναζήτησης μουσικής και καλύτερη συμβατότητα σε διαφορετικές συσκευές αναπαραγωγής μουσικής και πλατφόρμες [6]. Καθώς η τεχνολογία συνεχίζει να εξελίσσεται, το μέλλον της προσθήκης ετικετών ήχου φαίνεται πολλά υποσχόμενο, με τις εξελίξεις στην τεχνητή νοημοσύνη και τη μηχανική μάθηση να ενισχύουν περαιτέρω αυτό το βασικό εργαλείο.

2.2.2 Ανάπτυξη Τεχνολογιών Denoisers: Ιστορική ανασκόπηση με έμφαση στη βελτίωση ποιότητας ήχου.

Η ιστορία της μείωσης του θορύβου στον ήχο χρονολογείται από τις δεκαετίες του 1930 και του 1940, με την ανάπτυξη αναλογικών ηλεκτρονικών κυκλωμάτων με στόχο τη βελτίωση της σαφήνειας της τηλεφωνικής επικοινωνίας. Οι πρώτες προσπάθειες επικεντρώθηκαν στη μείωση του θορύβου γραμμής και των παρεμβολών χρησιμοποιώντας απλά αναλογικά φίλτρα. Η έλευση της ψηφιακής επεξεργασίας σήματος (DSP) στις δεκαετίες του 1960 και του 1970 σηματοδότησε ένα σημαντικό άλμα προς τα εμπρός. Το DSP επέτρεψε την ανάλυση σε πραγματικό χρόνο των σημάτων ήχου και την εφαρμογή πιο εξελιγμένων τεχνικών μείωσης θορύβου. Κατά τη διάρκεια αυτής της περιόδου, αναπτύχθηκαν επίσης αλγόριθμοι ικανοί να αναγνωρίσουν και να αφαιρέσουν θόρυβο από ένα σήμα, κάτι που ήταν ζωτικής σημασίας για τις τηλεπικοινωνίες και την εγγραφή [30].

Στις δεκαετίες του 1980 και του 1990, η τεχνολογία μείωσης θορύβου έγινε πιο προηγμένη και διαδεδομένη. Οι αλγόριθμοι που έγιναν πιο εξελιγμένοι, ικανοί να αναγνωρίσουν και να μείνουν ένα ευρύτερο φάσμα θορύβων, συμπεριλαμβανομένων των διακοπτόμενων και μεταβλητών ήχων. Σήμερα, η ενσωμάτωση τεχνολογιών ακύρωσης θορύβου σε ηλεκτρονικές συσκευές όπως τα κινητά τηλέφωνα και τα ακουστικά είναι συνηθισμένη. Ο 21ος αιώνας έχει δει συνεχείς προόδους που οδηγούνται από τις αυξημένες δυνατότητες επεξεργασίας δεδομένων και την τεχνητή νοημοσύνη. Οι σύγχρονοι αλγόριθμοι μείωσης θορύβου μηχανικής τεχνικής, δίνοντάς τους τη δυνατότητα να προσαρμόζονται και να βελτιώνονται με την πάροδο του χρόνου. Αυτά τα συστήματα μπορούν τώρα να διαφοροποιήσουν μια τεράστια γκάμα ήχων και είναι αποτελεσματικά ακόμη και στα πιο απαιτητικά ακουστικά περιβάλλοντα [31].

Η πρώτη ευρέως χρησιμοποιούμενη τεχνική μείωσης θορύβου αναπτύχθηκε από τον Ray Dolby το 1966. Αυτό το σύστημα προοριζόταν για επαγγελματική χρήση και περιλάμβανε μια διαδικασία κωδικοποίησης/αποκωδικοποίησης που αύξησε την αναλογία σήματος προς θόρυβο στην ταινία έως και 10 dB. Οι καινοτομίες της Dolby έθεσε τις βάσεις για μελλοντικές αλλαγές στην τεχνολογία μείωσης θορύβου. Η ενσωμάτωση της μηχανικής μάθησης και της τεχνητής νοημοσύνης (AI) στις τεχνολογίες μείωσης του θορύβου έχει φέρει επανάσταση στον τομέα. Αυτές οι προηγμένες τεχνολογίες έχουν πιο ακριβή και προσαρμοστική μείωση θορύβου, βελτιώνοντας σημαντικά τον ήχο του σε διάφορες εφαρμογές [32].

Οι αλγόριθμοι μηχανικής μάθησης μπορούν να αναλύσουν τεράστια σύνολα δεδομένα εγγραφών για να εντοπίσουν μοτίβα και χαρακτηριστικά θορύβου. Μαθαίνοντας από αυτά τα σύνολα δεδομένα, οι αλγόριθμοι μπορούν να διακριθούν αποτελεσματικά μεταξύ του θορύβου και των επιθυμητών σημάτων ήχου. Αυτή η δυνατότητα είναι ιδιαίτερα χρήσιμη σε δυναμικά περιβάλλοντα όπου τα χαρακτηριστικά θορύβου μπορούν να αλλάξουν γρήγορα. Τα συστήματα ακύρωσης θορύβου που βασίζονται σε AI μπορούν επίσης να προσαρμοστούν σε πραγματικό χρόνο σε διάφορες συνθήκες θορύβου. Για, στα τηλεπικοινωνιακά συστήματα, η τεχνητή νοημοσύνη μπορεί να παρακολουθεί συνεχώς το ηχητικό σήμα και να προσαρμόζει τις παραμέτρους μείωσης θορύβου για να διατηρήσει τη βέλτιστη ποιότητα ήχου [33]. Αυτή η προσαρμοστικότητα διασφαλίζει ότι ο ήχος παραμένει καθαρός και κατανοητός ακόμη και σε δύσκολα ακουστικά περιβάλλοντα.

Μία από τις πιο σημαντικές στη μείωση του θορύβου που βασίζεται στην τεχνητή νοημοσύνη είναι η χρήση τεχνικών βαθιάς εκμάθησης. Τα μοντέλα βαθιάς μάθησης, όπως τα συνελικτικά νευρωνικά δίκτυα (CNN) και τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNNs), έχουν επιδείξει αξιοσημείωτη απόδοση σε εργασίες μείωσης θορύβου [34]. Αυτά τα μοντέλα δεν μπορούν να επεξεργασθούν πολύπλοκα ηχητικά σήματα και να αφαιρέσουν τον θόρυβο με ελάχιστη επίδραση στην επιθυμητή ποιότητα ήχου. Η εφαρμογή της τεχνητής νοημοσύνης στη μείωση του θορύβου εκτείνεται πέρα από την παραδοσιακή επεξεργασία ήχου. αναγνώριση ομιλίας, η μείωση θορύβου που βασίζεται σε AI ενισχύει την ακρίβεια των φωνητικών εντολών και των υπηρεσιών μεταγραφής. Ομοίως, στα ακουστικά βαρηκοΐας, οι αλγόριθμοι AI μπορούν να βελτιώσουν την ευκρίνεια της ομιλίας για τους χρήστες σε θορυβώδη περιβάλλοντα, παρέχοντας καλύτερη εμπειρία ακρόασης.

Καθώς οι τεχνολογίες τεχνητής νοημοσύνης και μηχανικής μάθησης συνεχίζουν να εξελίσσονται, ο αντίκτυπος τους στις τεχνικές μειώσεις του θορύβου να αυξηθεί. Οι μελλοντικές αλλαγές μπορούν να περιλαμβάνουν πιο αποτελεσματικούς αλγόριθμους, καλύτερες δυνατότητες επεξεργασίας σε πραγματικό χρόνο και βελτιωμένη ενσωμάτωση με άλλες τεχνολογίες επεξεργασίας ήχου [35]. Η συνεχιζόμενη έρευνα και ανάπτυξη σε αυτόν τον τομέα αναμένεται να προσφέρει ακόμα πιο αποτελεσματικές λύσεις μείωσης του θορύβου, βελτιώνοντας περαιτέρω την ποιότητα των εγγραφών και των επικοινωνιών.

Συμπερασματικά, η ανάπτυξη τεχνολογιών μείωσης θορύβου έχει προχωρήσει πολύ από τα πρώτα αναλογικά φίλτρα μέχρι τα σημερινά εξελιγμένα συστήματα που βασίζονται σε τεχνητή νοημοσύνη. Οι συνεχείς εξελίξεις στη μηχανική μάθηση και την τεχνητή νοημοσύνη οδηγούν την εξέλιξη των τεχνικών μείωσης θορύβου, διασφαλίζοντας καθαρότερο και πιο κατανοητό σε ένα ευρύ φάσμα εφαρμογών.

2.2.3 Από την Ομιλία στο Κείμενο: Πώς εξελίχθηκαν οι τεχνικές μετατροπής ομιλίας σε κείμενο

Η μετατροπή ομιλίας σε κείμενο (STT), με στόχο τη μετατροπή της προφορικής γλώσσας σε ακριβές κείμενο, έχει σημειώσει σημαντικές προόδους κατά τη διάρκεια των δεκαετιών. Η έννοια του STT χρονολογείται από τις πρώτες μέρες της πληροφορικής. Στη δεκαετία του 1950, η Bell Labs πρωτοστάτησε στην ανάπτυξη των πρώιμων συστημάτων αναγνώρισης ομιλίας. Η δημιουργία της «Audrey» το 1952, ικανή να αναγνωρίζει ψηφία που εκφωνούνται με μία φωνή, έθεσε τις βάσεις για μελλοντικές καινοτομίες στον τομέα, αποδεικνύοντας ότι οι μηχανές μπορούσαν να ερμηνεύσουν προφορικές λέξεις. Αν και το "Audrey" μπορούσε να αναγνωρίσει μόνο αριθμούς που μιλούσε ο χειριστής του, έθεσε τις βάσεις για πιο εξελιγμένα συστήματα. Αυτή η περίοδος σημαδεύτηκε από την αρχική εξερεύνηση ακουστικών μοντέλων και φωνητικής ανάλυσης, τα οποία ήταν κρίσιμα για την ανάπτυξη των τεχνολογιών STT.

Η δεκαετία του 1960 είδε την είσοδο της IBM στο χώρο με την ανάπτυξη του "Shoebox", το οποίο μπορούσε να αναγνωρίσει 16 προφορικές λέξεις και ψηφία. Αυτή η συσκευή επέκτεινε το λεξιλόγιο που μπορούσαν να κατανοήσουν οι μηχανές, ξεπερνώντας απλώς τα αριθμητικά ψηφία. Το Shoebox της IBM παρουσιάστηκε στην Παγκόσμια Έκθεση της Νέας Υόρκης το 1964, σηματοδοτώντας μια σημαντική δημόσια έκθεση τεχνολογίας αναγνώρισης ομιλίας. Κατά τη διάρκεια αυτής της εποχής, οι ερευνητές άρχισαν να κατανοούν την πολυπλοκότητα της ανθρώπινης ομιλίας, οδηγώντας στην

ανάπτυξη πιο προηγμένων αλγορίθμων αναγνώρισης. Στη δεκαετία του 1970, το σύστημα «Harpy» του Πανεπιστημίου Carnegie Mellon εισήγαγε την έννοια της «αναζήτησης δέσμης», μια μέθοδο για τη βελτίωση της ακρίβειας της αναγνώρισης προφορικών λέξεων περιορίζοντας πιθανές ερμηνείες με βάση το πλαίσιο. Το "Harpy" μπορούσε να καταλάβει πάνω από 1.000 λέξεις και αντιπροσώπευε ένα άλμα προς τα εμπρός τόσο στο μέγεθος του λεξιλογίου όσο και στην ακρίβεια.

Η ανάπτυξη του Harpy ήταν μέρος των πρωτοβουλιών Advanced Research Projects Agency Network (ARPANET), το οποίο ώθησε περαιτέρω τα όρια της υπολογιστικής γλωσσολογίας. Η εισαγωγή της αναζήτησης δέσμης σηματοδότησε ένα σημαντικό ορόσημο, επιτρέποντας στα συστήματα να χειρίζονται αποτελεσματικά μεγαλύτερα λεξιλόγια με μεγαλύτερη ακρίβεια. Κατά την περίοδο των δεκαετιών του 1980 και του 1990, η τεχνολογία αναγνώρισης ομιλίας άρχισε να γίνεται εμπορικά διαθέσιμη. Το σύστημα «Tangora» της IBM, που παρουσιάστηκε το 1987, μπορούσε να αναγνωρίσει έως και 20.000 λέξεις, καθιστώντας το ένα πιο πρακτικό εργαλείο για εφαρμογές πραγματικού κόσμου. Το Tangora πήρε το όνομά του από τον Albert Tangora, έναν παγκοσμίου φήμης δακτυλογράφο ταχύτητας, συμβολίζοντας την ικανότητα του συστήματος να χειρίζεται γρήγορη ομιλία. Κατά τη διάρκεια αυτής της περιόδου, τα Hidden Markov Models (HMMs) έγιναν η κυρίαρχη τεχνική για την αναγνώριση ομιλίας, βελτιώνοντας σημαντικά τις στατιστικές μεθόδους που χρησιμοποιούνται για την πρόβλεψη προφορικών λέξεων. Τα HMM επέτρεψαν τη μοντελοποίηση της χρονικής δυναμικής στην ομιλία, η οποία ήταν μια σημαντική ανακάλυψη για την επίτευξη πιο φυσικής και ακριβούς αναγνώρισης.

Το 1997, το Dragon NaturallySpeaking σημείωσε μια σημαντική ανακάλυψη στη συνεχή αναγνώριση ομιλίας, επιτρέποντας στους χρήστες να υπαγορεύουν κείμενο με κανονικό ρυθμό ομιλίας χωρίς να χρειάζεται να κάνουν παύση μεταξύ των λέξεων. Αυτή η εξέλιξη σηματοδότησε την αρχή ενός εμπορικά βιώσιμου λογισμικού αναγνώρισης ομιλίας για προσωπικούς υπολογιστές. Το Dragon NaturallySpeaking χρησιμοποίησε τόσο HMM όσο και εξελιγμένα μοντέλα γλώσσας για να επιτύχει σημαντική μείωση στα ποσοστά σφαλμάτων. Η ικανότητα του λογισμικού να προσαρμόζεται στις ατομικές φωνές των χρηστών μέσω εκπαιδευτικών συνεδριών συνέβαλε επίσης στην εμπορική επιτυχία του, καθιστώντας την αναγνώριση ομιλίας πιο προσιτή στο ευρύ κοινό. Τα τελευταία χρόνια, η ενσωμάτωση της μηχανικής μάθησης έχει φέρει επανάσταση στις τεχνολογίες ομιλίας σε κείμενο. Οι λύσεις STT που υποστηρίζονται από AI, όπως αυτές που αναπτύχθηκαν από την Google, τη Microsoft και την IBM, χρησιμοποιούν μοντέλα βαθιάς μάθησης για να μετατρέψουν αποτελεσματικά τις προφορικές λέξεις σε κείμενο.

Αυτά τα μοντέλα, ιδιαίτερα τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) και τα συνελικτικά νευρωνικά δίκτυα (CNN), μπορούν να μάθουν από τεράστιες ποσότητες δεδομένων και να βελτιώνονται με την πάροδο του χρόνου. Αυτές οι καινοτομίες έχουν βρει εφαρμογές σε διάφορους κλάδους, συμπεριλαμβανομένων των τηλεπικοινωνιών, της υγειονομικής περίθαλψης και της εξυπηρέτησης πελατών. Για παράδειγμα, το STT API της Google, που παρουσιάστηκε το 2016, αξιοποιεί την τεχνολογία WaveNet της DeepMind για την παραγωγή μεταγραφών με πολύ φυσικό ήχο. Η χρήση της τεχνητής νοημοσύνης επέτρεψε στα συστήματα αναγνώρισης ομιλίας να χειρίζονται διάφορες προφορές, διαλέκτους και γλώσσες με μεγαλύτερη ακρίβεια. Με τη βοήθεια των API, η τεχνολογία αναγνώρισης ομιλίας έχει γίνει πιο προσιτή στο ευρύτερο κοινό, επιτρέποντας στους προγραμματιστές να ενσωματώνουν εύκολα εξελιγμένες δυνατότητες STT στις εφαρμογές τους. Αυτό επέτρεψε ένα ευρύ φάσμα περιπτώσεων χρήσης, από εικονικούς βοηθούς έως υπηρεσίες μεταγραφής σε πραγματικό χρόνο.

API όπως η υπηρεσία Azure Speech Service της Microsoft και η Watson Speech to Text της IBM παρέχουν ισχυρά εργαλεία για προγραμματιστές χωρίς να απαιτείται βαθιά εξειδίκευση στη μηχανική μάθηση ή στην επεξεργασία φυσικής γλώσσας. Η διαθεσιμότητα αυτών των API έχει ωθήσει την καινοτομία, επιτρέποντας σε μικρότερες εταιρείες και μεμονωμένους προγραμματιστές να δημιουργήσουν εφαρμογές που αξιοποιούν την προηγμένη τεχνολογία αναγνώρισης ομιλίας. Η παγκόσμια αγορά API ομιλίας σε κείμενο έχει γνωρίσει αξιοσημείωτες τεχνολογικές προόδους την τελευταία δεκαετία, βελτιώνοντας την ακρίβεια, την αποτελεσματικότητα και την προσβασιμότητα των τεχνολογιών αναγνώρισης ομιλίας.

Οι τεχνολογίες για την απομόνωση της ομιλίας από τον θόρυβο του περιβάλλοντος βελτιώνουν τη σαφήνεια και την ακρίβεια. Η επεξεργασία σε πραγματικό χρόνο και ο υπολογισμός άκρων επιτρέπουν την αναγνώριση ομιλίας σε πραγματικό χρόνο με χαμηλό λανθάνοντα χρόνο, απαραίτητη για εφαρμογές όπως η ζωντανή μεταγραφή και οι συσκευές που ενεργοποιούνται με φωνή. Η εκπαίδευση προσαρμοσμένων μοντέλων STT για συγκεκριμένους κλάδους ή ορολογία βελτιώνει την ακρίβεια του συστήματος σε εξειδικευμένα περιβάλλοντα. Ο συνδυασμός του STT με την ανάλυση συναισθήματος, την εξαγωγή λέξεων κλειδιών και τα βιομετρικά στοιχεία φωνής οδηγεί σε πιο ολοκληρωμένα συστήματα. Η ενσωμάτωση ισχυρών μέτρων ασφαλείας διασφαλίζει την προστασία των δεδομένων και τη συμμόρφωση με τους κανονισμούς περί απορρήτου. Το μέλλον των τεχνολογιών ομιλίας σε κείμενο έγκειται στην περαιτέρω πρόοδο στην τεχνητή νοημοσύνη και τη μηχανική μάθηση.

Επιπλέον, η ενοποίηση των τεχνολογιών STT με άλλες εφαρμογές που βασίζονται στο AI, όπως η επεξεργασία φυσικής γλώσσας και η ανάλυση συναισθήματος, θα συνεχίσει να οδηγεί την καινοτομία σε αυτόν τον τομέα. Οι νέες τάσεις περιλαμβάνουν τη δημιουργία πιο εξελιγμένων μοντέλων που μπορούν να επεξεργαστούν την αυθόρμητη ομιλία, τις διάφορες διακοπές που πιθανώς να συμβαίνουν και τους πολλαπλούς ομιλητές. Υπάρχει μια αυξανόμενη εστίαση στην ενίσχυση της εμπειρίας του χρήστη, ελαχιστοποιώντας την ανάγκη για περαιτέρο εκπαίδευση και προσαρμογές. Από τα πρώιμα μηχανικά συστήματα έως τις εξελιγμένες λύσεις που βασίζονται στην τεχνητή νοημοσύνη του σήμερα, αυτές οι τεχνολογίες έχουν γίνει απαραίτητα εργαλεία σε διάφορους τομείς, ενισχύοντας την παραγωγικότητα και την προσβασιμότητα. Οι συνεχιζόμενες προσπάθειες έρευνας και ανάπτυξης σε αυτόν τον τομέα αναμένεται να οδηγήσουν σε ακόμη πιο εντυπωσιακές δυνατότητες, καθιστώντας την τεχνολογία STT ακρογωνιαίο λίθο της αλληλεπίδρασης ανθρώπου-υπολογιστή.

2.2.4 Κατηγοριοποίηση Κειμένου: Μια ιστορική προοπτική της ανάπτυξης συστημάτων κατηγοριοποίησης

Η έννοια της κατηγοριοποίησης κειμένων έχει εξελιχθεί δραματικά με το πέρασμα των χρόνων. Αρχικά, βασιζόμασταν σε χειροκίνητα συστήματα ευρετηρίασης και ταξινόμησης, τα οποία απαιτούσαν μεγάλη εργασία και ήταν επιρρεπή σε λάθη. Οι πρώτες προσπάθειες για την αυτοματοποιημένη κατηγοριοποίηση κειμένων μπορούν να εντοπιστούν στις δεκαετίες του 1950 και του 1960, με την έλευση της τεχνολογίας των υπολογιστών. Στη δεκαετία του 1960, ερευνητές όπως ο Gerard Salton πρωτοστάτησαν στη χρήση μοντέλων διανυσματικού χώρου για ανάκτηση πληροφοριών, θέτοντας τις βάσεις για την αυτοματοποιημένη κατηγοριοποίηση κειμένων. Η εργασία του Salton στο Σύστημα Ανάκτησης Πληροφοριών SMART εισήγαγε την έννοια της στάθμισης του όρου συχνότητα-αντίστροφη συχνότητα εγγράφων (TF-IDF), η οποία έγινε ακρογωνιαίος λίθος στην κατηγοριοποίηση

και ανάκτηση κειμένου [47]. Την συγκεκριμένο διάστημα πραγματοποιήθηκε επίσης η ανάπτυξη των πειραμάτων Cranfield, τα οποία καθιέρωσαν μεθοδολογίες για την αξιολόγηση των συστημάτων ανάκτησης πληροφοριών [49].

Τα χειροκίνητα συστήματα κατηγοριοποίησης, όπως η Ταξινόμηση της Βιβλιοθήκης του Κογκρέσου (LCC) και η δεκαδική ταξινόμηση (DDC), ήταν από τις πρώτες μεθόδους οργάνωσης δεδομένων κειμένου. Στη δεκαετία του 1980 εισήχθησαν οι τεχνικές μηχανικής μάθησης, με αλγόριθμους όπως ο Naive Bayes και τα δέντρα αποφάσεων να εφαρμόζονται σε εργασίες ταξινόμησης κειμένου [50]. Αυτές οι μέθοδοι αξιοποίησαν τις στατιστικές ιδιότητες των δεδομένων κειμένου για να βελτιώσουν την ακρίβεια κατηγοριοποίησης. Πέρα από τα Naive Bayes και τα δέντρα αποφάσεων, οι ερευνητές διερεύνησαν επίσης τους k-Nearest Neighbors (k-NN) και ταξινομητές βάσει κανόνων για την κατηγοριοποίηση κειμένων. Κατά τη διάρκεια αυτής της περιόδου, η ανάπτυξη του συνόλου δεδομένων Reuters-21578 παρείχε ένα τυπικό σημείο αναφοράς για την αξιολόγηση των αλγορίθμων κατηγοριοποίησης κειμένου [50]. Επιπλέον, η εισαγωγή του πιθανοτικού μοντέλου από τους Robertson και Sparck Jones προώθησε περαιτέρω το πεδίο ενσωματώνοντας πιθανοτικό συλλογισμό στην κατηγοριοποίηση κειμένων [48].

Η εισαγωγή του συστήματος International Standard Book Number (ISBN) διευκόλυνε επίσης την αναγνώριση και ταξινόμηση των βιβλίων παγκοσμίως. Η δεκαετία του 1990 σηματοδότησε ένα σημαντικό ορόσημο με την ανάπτυξη των Υποστήριξης Διανυσματικών Μηχανών (SVM) και την εφαρμογή της θεωρίας της στατιστικής μάθησης στην κατηγοριοποίηση κειμένων [50]. Η άνοδος του διαδικτύου οδήγησε σε μια έκρηξη δεδομένων ψηφιακών κειμένων, δημιουργώντας την ανάγκη για πιο εξελιγμένες τεχνικές κατηγοριοποίησης [47]. Η ανάπτυξη του σχήματος περιγραφής αντικειμένου μεταδεδομένων (MODS) παρείχε έναν τυποποιημένο τρόπο περιγραφής των ψηφιακών πόρων, βοηθώντας στην κατηγοριοποίηση και ανάκτησή τους. Η εισαγωγή του μοντέλου bag-of-words και η χρήση n-grams ενίσχυσαν περαιτέρω την ικανότητα λήψης πληροφοριών με βάση τα συμφραζόμενα σε κείμενο [48].

Επιπλέον, οι διαγωνισμοί TREC (Text REtrieval Conference), που είχαν ξεκινήσει στις αρχές της δεκαετίας του 1990, παρείχαν μια πλατφόρμα για τους ερευνητές να αξιολογήσουν και να συγκρίνουν τα συστήματα κατηγοριοποίησης κειμένων τους σε κοινό έδαφος [49]. Στη δεκαετία του 2000, η εστίαση μετατοπίστηκε στη βελτίωση της ακρίβειας και της αποτελεσματικότητας των συστημάτων κατηγοριοποίησης κειμένου. Οι ερευνητές διερεύνησαν διάφορες μεθόδους επιλογής και εξαγωγής χαρακτηριστικών, όπως το TF-IDF και το Latent Semantic Indexing (LSI). Η ανάπτυξη μεθόδων συνόλου, όπως το boosting και το bagging, συνέβαλε επίσης στη βελτίωση της απόδοσης. Επιπλέον, η διαθεσιμότητα συνόλων δεδομένων μεγάλης κλίμακας, όπως το σύνολο δεδομένων των 20 ομάδων συζήτησης, διευκόλυνε την εκπαίδευση και την αξιολόγηση πιο περίπλοκων μοντέλων [50]. Η εισαγωγή μεθόδων πυρήνα, ιδιαίτερα σε SVM, επέτρεψε τον χειρισμό μη γραμμικών σχέσεων σε δεδομένα κειμένου.

Η εισαγωγή της βαθιάς μάθησης τη δεκαετία του 2010 έφερε επανάσταση στον τομέα, με τα νευρωνικά δίκτυα και τις ενσωματώσεις λέξεων να βελτιώνουν σημαντικά την απόδοση των συστημάτων κατηγοριοποίησης κειμένου [50]. Τεχνικές όπως τα συνελικτικά νευρωνικά δίκτυα (CNN) και τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) επέτρεψαν τη σύλληψη περίπλοκων μοτίβων και εξαρτήσεων σε δεδομένα κειμένου. Τα δίκτυα Long Short-Term Memory (LSTM) και οι Gated

Recurrent Units (GRU) χρησιμοποιήθηκαν επίσης για την καταγραφή εξαρτήσεων μεγάλης εμβέλειας σε δεδομένα κειμένου. Η αρχιτεκτονική του μετασχηματιστή, που εισήχθη από τους Vaswani et al. το 2017, έφερε επανάσταση στην κατηγοριοποίηση κειμένων με τον μηχανισμό αυτοπροσοχής, οδηγώντας σε μοντέλα όπως το BERT, το GPT-3 και το T5 [48]. Η ανάπτυξη προ-εκπαιδευμένων μοντέλων γλώσσας, όπως τα Word2Vec, GloVe και BERT, παρείχαν ισχυρά εργαλεία για την αναπαράσταση και ταξινόμηση κειμένων. Αυτά τα μοντέλα αξιοποίησαν μεγάλα σώματα δεδομένων κειμένου για να μάθουν πλούσιες σημασιολογικές αναπαραστάσεις, οι οποίες θα μπορούσαν να προσαρμοστούν με ακρίβεια για συγκεκριμένες εργασίες κατηγοριοποίησης κειμένου.

Σήμερα, η κατηγοριοποίηση κειμένων συνεχίζει να εξελίσσεται, με συνεχή έρευνα που διερευνά νέους αλγόριθμους, αναπαραστάσεις χαρακτηριστικών και εφαρμογές σε διάφορους τομείς [50]. Η εμφάνιση μοντέλων που βασίζονται σε μετασχηματιστές, όπως το GPT-3 και το T5, έχει ωθήσει τα όρια του τι είναι δυνατό στην κατηγοριοποίηση κειμένου. Οι ερευνητές διερευνούν επίσης τη χρήση της μάθησης μεταφοράς και την προσαρμογή τομέα για τη βελτίωση της γενίκευσης των μοντέλων κατηγοριοποίησης σε διαφορετικά σύνολα δεδομένων και εργασίες [49]. Επιπλέον, υπάρχει ένα αυξανόμενο ενδιαφέρον για το εξηγήσιμο AI (XAI) ώστε τα μοντέλα κατηγοριοποίησης κειμένου να γίνουν πιο ερμηνεύσιμα και διαφανή. Η ενοποίηση της κατηγοριοποίησης κειμένου με πολυτροπικά δεδομένα, όπως εικόνες και ήχο, είναι ένας αναδυόμενος τομέας έρευνας [47]. Οι ερευνητές διερευνούν τη χρήση τεχνικών μάθησης μηδενικής λήψης και λίγων λήψεων για να επιτρέψουν στα μοντέλα κατηγοριοποίησης κειμένου να γενικευτούν σε νέες κατηγορίες με ελάχιστα δεδομένα εκπαίδευσης.

Η ανάπτυξη πλαισίων ηθικής τεχνητής νοημοσύνης γίνεται όλο και πιο σημαντική για τη διασφάλιση της δικαιοσύνης, της υπευθυνότητας και της διαφάνειας στα συστήματα κατηγοριοποίησης κειμένων. Η ανάπτυξη συστημάτων κατηγοριοποίησης κειμένων ήταν μια συλλογική προσπάθεια, με συνεισφορές από ερευνητές σε πολλούς κλάδους, όπως η επιστήμη των υπολογιστών, η γλωσσολογία και η επιστήμη της πληροφορίας. Καθώς ο τομέας συνεχίζει να προοδεύει, υπόσχεται να επιτρέψει πιο ακριβή και αποτελεσματική οργάνωση και ανάκτηση πληροφοριών σε έναν συνεχώς αναπτυσσόμενο ψηφιακό κόσμο. Η ενοποίηση της κατηγοριοποίησης κειμένων με άλλες τεχνολογίες, όπως η επεξεργασία φυσικής γλώσσας (NLP) και η ανάκτηση πληροφοριών (IR), αναμένεται να ενισχύσει περαιτέρω τις δυνατότητες και τις εφαρμογές της.

2.3 YamNet: Αναγνώριση Ηχητικών Συμβάντων με Βάση τη Μηχανική Μάθηση

Η αναγνώριση ηχητικών συμβάντων αποτελεί ένα σημαντικό πεδίο έρευνας στην τεχνητή νοημοσύνη και τη μηχανική μάθηση. Η δυνατότητα ανίχνευσης και ταξινόμησης ηχητικών γεγονότων έχει εφαρμογές σε πολλούς τομείς, όπως η ασφάλεια, η παρακολούθηση περιβάλλοντος και η αλληλεπίδραση ανθρώπου υπολογιστή. Οι κύριες προκλήσεις περιλαμβάνουν την ποικιλία των ηχητικών σημάτων, τον θόρυβο περιβάλλοντος και την ανάγκη για ακριβή και γρήγορη ανίχνευση. Το YamNet, ένα προεκπαιδευμένο νευρωνικό δίκτυο που αναπτύχθηκε από την Google, έχει σχεδιαστεί για την αναγνώριση ηχητικών συμβάντων [51].

Βασίζεται στην αρχιτεκτονική MobileNetV1 και χρησιμοποιεί depthwise separable convolutions για την εξαγωγή χαρακτηριστικών από ηχητικά σήματα. Το μοντέλο έχει εκπαιδευτεί σε ένα μεγάλο σύνολο δεδομένων, το AudioSet, το οποίο περιλαμβάνει πάνω από 2 εκατομμύρια ηχητικά κλιπ από το YouTube. Αξιοποιεί προηγμένες τεχνικές μηχανικής μάθησης για την ανάλυση και ταξινόμηση

ηχητικών συμβάντων. Δέχεται ως είσοδο ένα ηχητικό κύμα και εξάγει χαρακτηριστικά μέσω των συνελικτικών στρώσεων του νευρωνικού δικτύου. Στη συνέχεια, αυτά τα χαρακτηριστικάχρησιμοποιούνται για την πρόβλεψη της κατηγορίας του ηχητικού συμβάντος. Το YamNet μπορεί να αναγνωρίσει 521 διαφορετικές κατηγορίες ηχητικών συμβάντων [52].

Η εκπαίδευση του μοντέλου βασίζεται σε τεχνικές μεταφοράς μάθησης, όπου οι γνώσεις που αποκτήθηκαν από την εκπαίδευση σε ένα συγκεκριμένο σύνολο δεδομένων, όπως το AudioSet, χρησιμοποιούνται ως βάση για την εκπαίδευση του μοντέλου σε νέα δεδομένα. Το YamNet έχει εφαρμογές σε διάφορους τομείς. Για παράδειγμα, μπορεί να χρησιμοποιηθεί σε συστήματα παρακολούθησης για την ανίχνευση επικίνδυνων ηχητικών συμβάντων, όπως πυροβολισμοί ή κραυγές [55]. Επίσης, μπορεί να ενσωματωθεί σε συσκευές έξυπνου σπιτιού για την αναγνώριση καθημερινών ήχων, όπως το άνοιγμα μιας βρύσης ή το χτύπημα μιας πόρτας. Επιπλέον, μπορεί να χρησιμοποιηθεί σε τεχνολογίες υποβοήθησης, όπως ειδοποιήσεις σε πραγματικό χρόνο για άτομα με προβλήματα ακοής.

Σε σύγκριση με άλλα συστήματα αναγνώρισης ηχητικών συμβάντων, όπως το VGGish ή το OpenL3, το YamNet προσφέρει μοναδικά πλεονεκτήματα, όπως η ελαφριά αρχιτεκτονική και η υψηλή ακρίβεια [54]. Ωστόσο, παρουσιάζει και περιορισμούς, όπως η εξάρτηση από το σύνολο δεδομένων AudioSet και η δυσκολία στην αναγνώριση ηχητικών συμβάντων σε θορυβώδη περιβάλλοντα. Μελλοντικές βελτιώσεις θα μπορούσαν να περιλαμβάνουν την ενσωμάτωση μεγαλύτερων συνόλων δεδομένων, την προσαρμογή σε περιβάλλοντα υπολογισμού αιχμής και την ανάπτυξη τεχνικών αυτοεπιβλεπόμενης μάθησης για την αναγνώριση ήχων [53].

Επιπλέον, η μείωση της καθυστέρησης και η βελτίωση της γενίκευσης του μοντέλου είναι σημαντικοί τομείς για περαιτέρω έρευνα. Συνοψίζοντας, το YamNet αποτελεί ένα ισχυρό εργαλείο για την αναγνώριση ηχητικών συμβάντων, προσφέροντας σημαντικές δυνατότητες και εφαρμογές. Η ενσωμάτωση της μηχανικής μάθησης επιτρέπει την ακριβή και αποδοτική ανίχνευση ηχητικών συμβάντων, συμβάλλοντας στην πρόοδο του πεδίου της τεχνητής νοημοσύνης και της μηχανικής μάθησης. Η ανάλυση των δυνατοτήτων και των περιορισμών του YamNet, καθώς και οι προτάσεις για μελλοντική έρευνα, υπογραμμίζουν τη σημασία του στην προώθηση της αναγνώρισης ηχητικών συμβάντων.

2.3.1 Αρχιτεκτονική, Λειτουργία και Χαρακτηριστικά του YamNet

Το YamNet αποτελεί ένα προεκπαιδευμένο νευρωνικό δίκτυο βαθιάς μάθησης, ειδικά σχεδιασμένο για την αυτόματη αναγνώριση ηχητικών συμβάντων. Στοχεύει στην παροχή μιας αποδοτικής και ακριβούς μεθόδου ανάλυσης ήχου, ικανής να ταξινομεί ένα ευρύ φάσμα ηχητικών σημάτων, από καθημερινούς θορύβους του περιβάλλοντος, όπως το θρόισμα των φύλλων ή το κελάηδισμα των πουλιών, έως πιο εξειδικευμένα ηχητικά συμβάντα, όπως μουσική, ομιλία, ή ακόμα και ήχους μηχανών ή ζώων. Η αρχιτεκτονική του YamNet βασίζεται στο MobileNetV1, ένα ελαφρύ και υπολογιστικά αποδοτικό νευρωνικό δίκτυο, ιδανικό για χρήση σε συσκευές με περιορισμένους πόρους, όπως κινητά τηλέφωνα, ενσωματωμένα συστήματα, IoT συσκευές και άλλες edge devices. Αυτή η επιλογή αρχιτεκτονικής το καθιστά ιδιαίτερα χρήσιμο σε εφαρμογές πραγματικού χρόνου, όπου η ταχύτητα επεξεργασίας είναι κρίσιμη, και σε περιβάλλοντα με περιορισμένες υπολογιστικές δυνατότητες, όπως σε αισθητήρες ήχου χαμηλής κατανάλωσης. Το MobileNetV1 αξιοποιεί διαχωρισμένες συνελικτικές στρώσεις (depthwise separable convolutions), μια τεχνική που μειώνει σημαντικά τον αριθμό των παραμέτρων και τις

υπολογιστικές απαιτήσεις σε σύγκριση με τις παραδοσιακές συνελικτικές στρώσεις, διατηρώντας παράλληλα ικανοποιητική ακρίβεια και απόδοση. Αυτή η αρχιτεκτονική επιτρέπει στο YamNet να είναι αρκετά μικρό σε μέγεθος, γεγονός που διευκολύνει την ανάπτυξή του σε διάφορες πλατφόρμες [56]. Η λειτουργία του YamNet περιλαμβάνει μια σειρά διαδοχικών βημάτων για την ανάλυση του ήχου. Αρχικά, η είσοδος δεδομένων στο YamNet γίνεται με τη μορφή ενός ηχητικού κύματος, δηλαδή ενός ψηφιακού αρχείου ήχου σε μορφή όπως .wav. Αυτό το σήμα αναπαριστά τις διακυμάνσεις της πίεσης του αέρα με την πάροδο του χρόνου, τις οποίες αντιλαμβανόμαστε ως ήχο. Στη συνέχεια, το ηχητικό κύμα μετατρέπεται σε Mel φασματογράφημα. Αυτή η μετατροπή είναι κρίσιμη και γίνεται για δύο βασικούς λόγους: πρώτον, για να αναπαραστήσει το σήμα στο πεδίο της συχνότητας, αναλύοντας την ένταση του ήχου σε κάθε συχνότητα αντί για κάθε χρονική στιγμή, κάτι που είναι πιο χρήσιμο για την αναγνώριση ήχων, καθώς οι διάφοροι ήχοι έχουν διαφορετικές συχνοτικές συνιστώσες, δημιουργώντας μοναδικά "δακτυλικά αποτυπώματα" στο φασματογράφημα. Δεύτερον, για να προσομοιώσει την ανθρώπινη ακοή. Η κλίμακα Mel είναι μια ψυχοακουστική κλίμακα που προσομοιάζει τον τρόπο με τον οποίο αντιλαμβάνεται ο άνθρωπος τις συχνότητες, δίνοντας έμφαση σε αυτές που είναι πιο σημαντικές για την ανθρώπινη ακοή και μειώνοντας την σημασία συχνοτήτων που δεν αντιλαμβανόμαστε εύκολα. Το αποτέλεσμα αυτής της μετατροπής είναι μια "εικόνα", όπου ο οριζόντιος άξονας αναπαριστά τον χρόνο, ο κάθετος άξονας αναπαριστά τις συχνότητες στην κλίμακα Mel και η φωτεινότητα κάθε σημείου αναπαριστά την ένταση του ήχου σε αυτή τη συχνότητα και χρονική στιγμή. Η διαδικασία αυτή συχνά περιλαμβάνει και προεπεξεργασία του ηχητικού σήματος, όπως φιλτράρισμα θορύβου για την απομάκρυνση ανεπιθύμητων συχνοτήτων και κανονικοποίηση για την ομαλοποίηση της έντασης του ήχου, βελτιώνοντας έτσι την ακρίβεια της ανάλυσης [57].

Αυτό το Mel φασματογράφημα εισέρχεται στο νευρωνικό δίκτυο MobileNetV1. Οι συνελικτικές στρώσεις του MobileNetV1 μαθαίνουν να αναγνωρίζουν μοτίβα στην "εικόνα" του φασματογραφήματος, τα οποία αντιστοιχούν σε χαρακτηριστικά του ήχου. Για παράδειγμα, κάποιες στρώσεις μπορεί να μαθαίνουν να αναγνωρίζουν τις συχνότητες που αντιστοιχούν σε ένα γαύγισμα σκύλου, ενώ άλλες σε ένα θόρυβο αυτοκινήτου, αναλύοντας την χρονική και συχνοτική εξέλιξη των ήχων. Τα χαρακτηριστικά που εξάγονται από το MobileNetV1 στη συνέχεια τροφοδοτούνται σε πλήρως συνδεδεμένες στρώσεις. Αυτές οι στρώσεις πραγματοποιούν την τελική ταξινόμηση, δηλαδή προβλέπουν την πιθανότητα ο ήχος να ανήκει σε κάθε μία από τις 521 κατηγορίες ήχων που έχει μάθει το YamNet (αυτές οι κατηγορίες προέρχονται από το AudioSet). Η έξοδος του YamNet είναι ένα διάνυσμα με 521 τιμές, όπου κάθε τιμή αναπαριστά την πιθανότητα ο ήχος να ανήκει στην αντίστοιχη κατηγορία. Είναι σημαντικό να σημειωθεί ότι το YamNet δεν παράγει μία μόνο πρόβλεψη για ολόκληρο το ηχητικό σήμα, αλλά μια σειρά από προβλέψεις για μικρά χρονικά τμήματα του σήματος, επιτρέποντας την ανίχνευση αλλαγών στους ήχους με την πάροδο του χρόνου και παρέχοντας μια πιο λεπτομερή ανάλυση της χρονικής εξέλιξης των ηχητικών συμβάντων. Αυτό το χαρακτηριστικό είναι ιδιαίτερα χρήσιμο σε περιπτώσεις όπου οι ήχοι αλλάζουν με την πάροδο του χρόνου, όπως για παράδειγμα σε μια συζήτηση όπου εναλλάσσονται οι ομιλητές [57].

Ένα κρίσιμο στοιχείο της επιτυχίας του YamNet είναι η χρήση transfer learning από το AudioSet, ένα τεράστιο σύνολο δεδομένων με εκατομμύρια ηχητικά κλιπ και αντίστοιχες ετικέτες για διάφορα ηχητικά συμβάντα. Η προεκπαίδευση στο AudioSet επιτρέπει στο YamNet να έχει ήδη μάθει γενικά χαρακτηριστικά των ήχων, τα οποία στη συνέχεια χρησιμοποιεί για να αναγνωρίζει πιο συγκεκριμένους ήχους, επιταχύνοντας σημαντικά τη διαδικασία εκπαίδευσης και βελτιώνοντας τη γενίκευση του μοντέλου σε νέα, άγνωστα ηχητικά περιβάλλοντα. Κατά τη διάρκεια της εκπαίδευσης, χρησιμοποιούνται διάφορες τεχνικές βελτιστοποίησης, όπως η κανονικοποίηση (π.χ., batch

normalization) για τη σταθεροποίηση της εκπαίδευσης και την αποφυγή υπερπροσαρμογής (π.χ., dropout, weight decay) για τη βελτίωση της ικανότητας του μοντέλου να γενικεύει σε νέα δεδομένα, καθώς και συναρτήσεις απώλειας όπως η Cross-Entropy, η οποία μετρά τη διαφορά μεταξύ των προβλέψεων του μοντέλου και των πραγματικών ετικετών, και αλγόριθμοι βελτιστοποίησης όπως ο Adam optimizer, ο οποίος προσαρμόζει τους συντελεστές μάθησης κατά τη διάρκεια της εκπαίδευσης για ταχύτερη σύγκλιση, με τη στρατηγική mini-batch gradient descent, όπου τα δεδομένα εκπαίδευσης χωρίζονται σε μικρές παρτίδες για πιο αποδοτική εκπαίδευση [58]. Η απόδοση του YamNet αξιολογείται με μετρικές όπως η ακρίβεια (accuracy), η ευστοχία (precision), η ανάκληση (recall) και το F1-score, οι οποίες παρέχουν μια ολοκληρωμένη εικόνα της ικανότητας του μοντέλου να αναγνωρίζει σωστά τους ήχους.Λόγω της ελαφριάς αρχιτεκτονικής του, το YamNet είναι ιδιαίτερα προσαρμόσιμο σε εφαρμογές πραγματικού χρόνου, όπως η ανίχνευση ήχων σε έξυπνες συσκευές, η αυτόματη μεταγραφή ομιλίας, η παρακολούθηση του περιβάλλοντος και η ανάλυση ηχητικού περιεχομένου σε διάφορα πεδία, όπως η βιομηχανία, η ιατρική και η ασφάλεια. Για παράδειγμα, μπορεί να χρησιμοποιηθεί για την ανίχνευση ασυνήθιστων θορύβων σε μηχανές, προειδοποιώντας για πιθανές βλάβες, ή για την παρακολούθηση του ήχου σε νοσοκομεία, ανιχνεύοντας κρίσεις πανικού ή άλλες καταστάσεις έκτακτης ανάγκης. Στον τομέα της ασφάλειας, μπορεί να χρησιμοποιηθεί για την ανίχνευση πυροβολισμών ή άλλων επικίνδυνων θορύβων. Επιπλέον, για χρήση σε συσκευές με ακόμα πιο περιορισμένους πόρους, όπου η κατανάλωση ενέργειας και η υπολογιστική ισχύς είναι περιορισμένες, το YamNet μπορεί να συμπιεστεί περαιτέρω με τεχνικές όπως η κβαντοποίηση (quantization). Η κβαντοποίηση μειώνει την ακρίβεια των αριθμών που χρησιμοποιούνται στο μοντέλο (π.χ., από 32-bit floating point σε 8-bit integers), μειώνοντας το μέγεθος του μοντέλου και τις υπολογιστικές απαιτήσεις, με ελάχιστη ή και αμελητέα απώλεια στην ακρίβεια. Αυτό καθιστά το YamNet ιδανικό για ανάπτυξη σε μικροελεγκτές και άλλες ενσωματωμένες συσκευές [59]. Επιπλέον, αξίζει να αναφερθεί ότι το YamNet, εκτός από την ταξινόμηση ηχητικών συμβάντων, μπορεί να χρησιμοποιηθεί και για την εξαγωγή ενδιάμεσων αναπαραστάσεων του ήχου, γνωστών ως embeddings. Αυτά τα embeddings είναι διανύσματα που αναπαριστούν τα χαρακτηριστικά του ήχου σε έναν υψηλότερο επίπεδο αφαίρεσης και μπορούν να χρησιμοποιηθούν ως είσοδος σε άλλα μοντέλα μηχανικής μάθησης για περαιτέρω ανάλυση ή για την επίλυση πιο σύνθετων προβλημάτων. Για παράδειγμα, τα embeddings του YamNet μπορούν να χρησιμοποιηθούν για την ομαδοποίηση παρόμοιων ήχων, την ανίχνευση αλλαγών στο ηχητικό περιβάλλον ή την αναγνώριση μουσικών οργάνων. [57]

Συνοψίζοντας, το YamNet αποτελεί ένα ισχυρό και ευέλικτο εργαλείο για την αναγνώριση ηχητικών συμβάντων, προσφέροντας υψηλή ακρίβεια, αποδοτικότητα και προσαρμοστικότητα σε μια ευρεία γκάμα εφαρμογών, από απλή ανίχνευση θορύβων έως πιο σύνθετες εργασίες ανάλυσης ήχου. Η ελαφριά αρχιτεκτονική του, σε συνδυασμό με τη χρήση transfer learning και την δυνατότητα συμπίεσης, το καθιστούν ιδανικό για ανάπτυξη σε διάφορες πλατφόρμες, συμπεριλαμβανομένων των συσκευών με περιορισμένους πόρους, ανοίγοντας νέους δρόμους για την ενσωμάτωση της τεχνητής νοημοσύνης στην ανάλυση ήχου σε διάφορους τομείς.

2.3.2 Ανάλυση των τεχνικών που χρησιμοποιούνται για ανίχνευση ηχητικών γεγονότων.

Η διαδικασία ανάλυσης ήχου από το YamNet περιλαμβάνει μια σειρά από διαδοχικά βήματα. Αρχικά, το YamNet λαμβάνει ως είσοδο ένα ψηφιακό σήμα ήχου, συνήθως αποθηκευμένο σε μορφή αρχείου «.wav». Αυτό το σήμα αντιπροσωπεύει τις διακυμάνσεις της πίεσης του αέρα που αντιλαμβανόμαστε

ως ήχο. Πριν τροφοδοτηθεί το σήμα στο νευρωνικό δίκτυο, υποβάλλεται σε μια διαδικασία προεπεξεργασίας, η οποία συνήθως περιλαμβάνει τα ακόλουθα:

Επαναδειγματοληψία: Εάν ο ρυθμός δειγματοληψίας του εισερχόμενου σήματος είναι διαφορετικός από τον επιθυμητό (συνήθως 16 kHz), εκτελείται νέα δειγματοληψία για την προσαρμογή του [61],[62].

Κανονικοποίηση: Η ένταση του σήματος προσαρμόζεται σε ένα συγκεκριμένο εύρος τιμών προκειμένου να βελτιωθεί η σταθερότητα και η ακρίβεια της ανάλυσης [60],[62].

Φιλτράρισμα θορύβου: Εφαρμόζονται διάφορες τεχνικές φιλτραρίσματος για την αφαίρεση ανεπιθύμητων συχνοτήτων ή θορύβου που μπορεί να επηρεάσουν αρνητικά την ανάλυση [60],[62].

Το προεπεξεργασμένο ηχητικό σήμα στη συνέχεια μετατρέπεται σε φασματόγραμμα Mel. Αυτή η μετατροπή είναι ζωτικής σημασίας για την αποτελεσματική αναγνώριση ήχου. Το φασματογράφημα Mel είναι μια οπτική αναπαράσταση του ήχου στον τομέα της συχνότητας, η οποία προσομοιώνει τον τρόπο με τον οποίο οι άνθρωποι αντιλαμβάνονται τις συχνότητες. Η δημιουργία του φασματογράμματος Mel περιλαμβάνει τα ακόλουθα βήματα:

Μετασχηματισμός Fourier βραχείας διάρκειας (STFT): Το ηχητικό σήμα χωρίζεται σε μικρά, επικαλυπτόμενα χρονικά τμήματα (πλαίσια) και για κάθε τμήμα, ο μετασχηματισμός Fourier υπολογίζεται για την ανάλυση των συχνοτήτων που περιέχονται σε αυτό το τμήμα [61],[62].

Μετατροπή στην κλίμακα Mel: Οι συχνότητες που προκύπτουν από το STFT μετατρέπονται στην κλίμακα Mel, μια ψυχοακουστική κλίμακα που δίνει μεγαλύτερη έμφαση στις χαμηλές συχνότητες, όπου η ανθρώπινη ακοή είναι πιο ευαίσθητη [61],[62].

Λογαριθμική κλίμακα: Η ένταση κάθε συχνότητας στην κλίμακα Mel μετατρέπεται σε λογαριθμική κλίμακα, συμπιέζοντας το δυναμικό εύρος εντάσεων και κάνοντας πιο διακριτές τις μικρές διαφορές στην ένταση [61],[62].

Το τελικό αποτέλεσμα είναι μια δισδιάστατη «εικόνα», το φασματόγραμμα Mel, όπου ο οριζόντιος άξονας αντιπροσωπεύει το χρόνο, ο κατακόρυφος άξονας αντιπροσωπεύει συχνότητες στην κλίμακα Mel και η φωτεινότητα κάθε σημείου αντιστοιχεί στην ένταση του ήχου στη συγκεκριμένη συχνότητα και του χρόνου.

Αυτό το φασματόγραμμα Mel στη συνέχεια τροφοδοτείται στο νευρωνικό δίκτυο MobileNetV1. Το δίκτυο, χρησιμοποιώντας τα συνελικτικά στρώματά του (και συγκεκριμένα τα συνελικτικά στρώματα που μπορούν να διαχωριστούν σε βάθος, ένα χαρακτηριστικό του MobileNetV1 που βελτιώνει την απόδοση), αναλύει το φασματόγραμμα και μαθαίνει να αναγνωρίζει χαρακτηριστικά μοτίβα που αντιστοιχούν σε διαφορετικούς ήχους [58],[62]. Μετά την επεξεργασία από τα συνελικτικά επίπεδα, τα εξαγόμενα χαρακτηριστικά τροφοδοτούνται σε πλήρως συνδεδεμένα επίπεδα, τα οποία λαμβάνουν την τελική ταξινόμηση του ήχου σε μία από τις 521 προκαθορισμένες κατηγορίες που προέρχονται από το AudioSet, ένα τεράστιο σύνολο δεδομένων ήχου που χρησιμοποιείται για την εκπαίδευση του YamNet [58],[62]. Η έξοδος του YamNet είναι ένα διάνυσμα με 521 τιμές, όπου κάθε τιμή αντιστοιχεί στην πιθανότητα ο ήχος να ανήκει σε μια συγκεκριμένη κατηγορία. Είναι σημαντικό να τονιστεί ότι το YamNet δεν παράγει μία μόνο πρόβλεψη για ολόκληρο το ηχητικό σήμα, αλλά μάλλον μια σειρά προβλέψεων για μικρά, διαδοχικά χρονικά τμήματα του σήματος. Αυτό επιτρέπει την ανίχνευση αλλαγών στον ήχο με την πάροδο του χρόνου και παρέχει μια χρονική σειρά προβλέψεων [58],[62].

Η αποτελεσματικότητα του YamNet βελτιώνεται σημαντικά με τη χρήση της εκμάθησης μεταφοράς από το AudioSet. Αυτό σημαίνει ότι το μοντέλο έχει ήδη εκπαιδευτεί σε ένα τεράστιο και ποικίλο σύνολο δεδομένων ήχου με πάνω από 2 εκατομμύρια αποσπάσματα ήχου 10 δευτερολέπτων με ετικέτα, μαθαίνοντας τα γενικά χαρακτηριστικά των ήχων [58],[62]. Αυτή η προ-εκπαίδευση επιτρέπει στο YamNet να αναγνωρίζει πιο εύκολα και με ακρίβεια νέους ήχους, ακόμα κι αν διαφέρουν ελαφρώς από εκείνους που είχε δει κατά την αρχική του εκπαίδευση.

2.3.3 Εφαρμογές του YamNet στο Audio Tagging

Η προεκπαίδευση του μοντέλου αυτού επιτρέπει να αναγνωρίζει ένα ευρύ φάσμα ήχων, καθιστώντας το ένα ισχυρό εργαλείο για διάφορες εφαρμογές Audio Tagging. Στον τομέα των έξυπνων συσκευών, το YamNet μπορεί να χρησιμοποιηθεί για την αναγνώριση συγκεκριμένων ήχων σε ένα έξυπνο σπίτι, όπως ο ήχος ενός συναγερμού (πυρκαγιά, διάρρηξη), το κλάμα ενός μωρού, το σπάσιμο του γυαλιού ή ο θόρυβος μιας συσκευής που δεν λειτουργεί. Αυτή η δυνατότητα μπορεί να βελτιώσει σημαντικά την ασφάλεια και την άνεση του σπιτιού, ενεργοποιώντας αυτόματα ειδοποιήσεις στους χρήστες ή προκαθορισμένες ενέργειες από άλλες έξυπνες συσκευές [57]. Για παράδειγμα, η ανίχνευση ενός μωρού που κλαίει μπορεί να ενεργοποιήσει απαλή μουσική ή να στείλει μια ειδοποίηση στους γονείς.

Στον τομέα της αυτόματης μεταγραφής ομιλίας, το YamNet μπορεί να βοηθήσει στη βελτίωση της ακρίβειας των συστημάτων αναγνωρίζοντας τα ηχητικά πλαίσια μέσα στα οποία είναι ενσωματωμένη η ομιλία και φιλτράροντας ανεπιθύμητους θορύβους, όπως μουσική υπόκρουση ή περιβαλλοντικούς θορύβους. Αυτό είναι ιδιαίτερα χρήσιμο για την καταγραφή και την ανάλυση συναντήσεων, διαλέξεων, συνεντεύξεων ή τηλεφωνικών κλήσεων, όπου η σαφής μεταγραφή της ομιλίας είναι ζωτικής σημασίας [60].

Στην περιβαλλοντική παρακολούθηση, το YamNet μπορεί να χρησιμοποιηθεί για την ανίχνευση συγκεκριμένων ήχων σε διάφορα περιβάλλοντα, όπως ήχους ζώων σε φυσικό καταφύγιο (για παρακολούθηση βιοποικιλότητας), ήχους κυκλοφορίας σε πόλη (για ανάλυση ηχορύπανσης) ή ήχους που υποδεικνύουν μια επικείμενη φυσική καταστροφή. όπως ένας σεισμός ή μια πλημμύρα. Αυτές οι δυνατότητες επιτρέπουν την έγκαιρη προειδοποίηση και την αποφυγή κινδύνων [58].

Στην ανάλυση περιεχομένου ήχου, το YamNet μπορεί να χρησιμοποιηθεί για την αυτόματη ταξινόμηση και ευρετηρίαση αρχείων ήχου με βάση το περιεχόμενό τους, διευκολύνοντας την αναζήτηση, την ανάκτηση και την οργάνωση μεγάλων βάσεων δεδομένων ήχου. Παραδείγματα περιλαμβάνουν μουσικές βιβλιοθήκες, αρχεία ήχου ιστορικού ενδιαφέροντος ή ηχητικά αποσπάσματα από ταινίες και τηλεοπτικά προγράμματα [57].

Στον τομέα της υγειονομικής περίθαλψης, η ανάλυση των ηχητικών σημάτων του σώματος όπως οι καρδιακοί παλμοί, οι ήχοι των πνευμόνων ή οι εντερικοί θόρυβοι μπορεί να βοηθήσει στην έγκαιρη διάγνωση και παρακολούθηση της υγείας των ασθενών. Για παράδειγμα, το YamNet μπορεί να χρησιμοποιηθεί για τον εντοπισμό καρδιακών αρρυθμιών ή αναπνευστικών προβλημάτων, παρέχοντας πολύτιμες πληροφορίες στους επαγγελματίες υγείας.

Η χρήση προεκπαιδευμένων μοντέλων όπως το YamNet επιτρέπει την αναγνώριση ήχου υψηλής ακρίβειας ακόμη και σε θορυβώδη περιβάλλοντα. Η τεχνική μεταφοράς μάθησης επιτρέπει την εκπαίδευση νέων μοντέλων για πιο εξειδικευμένες εργασίες με λιγότερα δεδομένα και σε λιγότερο

χρόνο [58]. Το YamNet είναι ένα ισχυρό και ευέλικτο εργαλείο για την ανάλυση και ταξινόμηση δεδομένων ήχου, προσφέροντας πολλές δυνατότητες και πλεονεκτήματα σε διάφορους τομείς.

2.4 Whisper: Αυτόματη Μετατροπή Ομιλίας σε Κείμενο

Η τεχνολογία μετατροπής ομιλίας σε κείμενο έχει εξελιχθεί σημαντικά τις τελευταίες δεκαετίες, μεταβαίνοντας από απλά συστήματα με περιορισμένο λεξιλόγιο σε προηγμένα μοντέλα που μπορούν να κατανοούν και να μεταγράφουν πολλές γλώσσες με υψηλή ακρίβεια. Η αυξανόμενη ζήτηση για ακριβείς και πολυγλωσσικές τεχνολογίες μεταγραφής οφείλεται στην ανάγκη για προσβασιμότητα, αποτελεσματική επικοινωνία και αυτοματοποίηση σε διάφορους τομείς.

Το Whisper, που αναπτύχθηκε από την OpenAI, αποτελεί μια σημαντική πρόοδο στην αυτόματη αναγνώριση ομιλίας (ASR). Έχει εκπαιδευτεί σε 680.000 ώρες πολυγλωσσικών και πολυλειτουργικών δεδομένων, καθιστώντας το ανθεκτικό σε προφορές, θορύβους περιβάλλοντος και τεχνική γλώσσα. Τα βασικά χαρακτηριστικά του Whisper περιλαμβάνουν την αρχιτεκτονική end-to-end Transformer encoder-decoder, που του επιτρέπει να διαχειρίζεται διάφορες εργασίες όπως αναγνώριση γλώσσας, χρονικές σημάνσεις και πολυγλωσσική μεταγραφή ομιλίας[63]. Σε σχέση με τα προηγούμενα μοντέλα, το Whisper προσφέρει υψηλότερη ακρίβεια και μεγαλύτερη δυνατότητα κλιμάκωσης, καθιστώντας το μια σημαντική πρόοδο στον τομέα της ASR.

2.4.1 Αρχιτεκτονική και Λειτουργία του Whisper

Η τεχνολογία ομιλίας σε κείμενο έχει εξελιχθεί σημαντικά τις τελευταίες δεκαετίες, μεταβαίνοντας από απλά συστήματα με περιορισμένο λεξιλόγιο σε προηγμένα μοντέλα που μπορούν να κατανοήσουν και να μεταγράψουν πολλές γλώσσες με υψηλή ακρίβεια. Η αυξανόμενη ζήτηση για ακριβείς και πολύγλωσσες τεχνολογίες μεταγραφής καθοδηγείται από την ανάγκη για προσβασιμότητα, αποτελεσματική επικοινωνία και αυτοματισμό σε διάφορους τομείς [65].

Τα βασικά χαρακτηριστικά του Whisper περιλαμβάνουν την αρχιτεκτονική του κωδικοποιητή-αποκωδικοποιητή Transformer από άκρο σε άκρο, η οποία του επιτρέπει να χειρίζεται εργασίες όπως η αναγνώριση γλώσσας, η χρονοσφραγίδα και η πολυγλωσσική μεταγραφή ομιλίας [64],[66]. Το Whisper λαμβάνει το ηχητικό σήμα, το οποίο μπορεί να είναι οποιοσδήποτε ήχος ομιλίας. Ο ήχος χωρίζεται σε τμήματα 30 δευτερολέπτων για ευκολία επεξεργασίας. Κάθε τμήμα του ήχου μετατρέπεται σε φασματογράμματα log-Mel, τα οποία είναι οπτικές αναπαραστάσεις του ήχου που δείχνουν πώς η ενέργεια του ήχου αλλάζει με την πάροδο του χρόνου [64]. Ο κωδικοποιητής Whisper επεξεργάζεται τα φασματογράμματα log-Mel και εξάγει ενδιάμεσες αναπαραστάσεις που καταγράφουν τα βασικά χαρακτηριστικά του ηχητικού σήματος [64]. Ο αποκωδικοποιητής λαμβάνει τις ενδιάμεσες αναπαραστάσεις από τον κωδικοποιητή και τις μετατρέπει σε κείμενο. Ο αποκωδικοποιητής χρησιμοποιεί ειδικά διακριτικά για την αναγνώριση γλώσσας, τις χρονικές σημάνσεις και άλλες εργασίες [64].

Η αρχιτεκτονική Whisper βασίζεται σε ένα μοντέλο Transformer, το οποίο είναι γνωστό για την ικανότητά του να μαθαίνει από μεγάλα και διαφορετικά σύνολα δεδομένων [66]. Το Whisper έχει εκπαιδευτεί σε 680.000 ώρες δεδομένων, που εκτείνονται σε πολλές γλώσσες και διαφορετικά περιβάλλοντα, καθιστώντας το ανθεκτικό στο θόρυβο και τις διαφορετικές προφορές [64]. Ο κωδικοποιητής Whisper αποτελείται από πολλαπλά στρώματα μπλοκ Transformer. Κάθε μπλοκ

περιλαμβάνει μηχανισμούς αυτοπροσοχής και πλήρως συνδεδεμένα δίκτυα (δίκτυα τροφοδοσίας). Ο κωδικοποιητής λαμβάνει τα φασματογράμματα log-Mel ως είσοδο και παράγει ενδιάμεσες αναπαραστάσεις που περιέχουν τις σημαντικές πληροφορίες του σήματος ήχου [64],[66].

Ο αποκωδικοποιητής Whisper βασίζεται επίσης σε μπλοκ Transformer και λαμβάνει τις ενδιάμεσες αναπαραστάσεις από τον κωδικοποιητή. Ο αποκωδικοποιητής προβλέπει το αντίστοιχο κείμενο, χρησιμοποιώντας μηχανισμούς αυτοπροσοχής και προσοχής στον κωδικοποιητή (encoder-decoder). Επιπλέον, ο αποκωδικοποιητής χρησιμοποιεί ειδικά διακριτικά για την αναγνώριση γλώσσας, τις χρονικές σημάνσεις και άλλες εργασίες [64],[66].

Το Whisper επεξεργάζεται τις εισερχόμενες πληροφορίες ήχου μετατρέποντάς τις πρώτα από κυματομορφές σε φασματογράμματα log-Mel, τα οποία στη συνέχεια τροφοδοτούνται στον κωδικοποιητή. Ο κωδικοποιητής δημιουργεί ενδιάμεσες αναπαραστάσεις που καταγράφουν τα κύρια χαρακτηριστικά του σήματος ήχου. Αυτές οι αναπαραστάσεις περνούν στον αποκωδικοποιητή, ο οποίος προβλέπει το αντίστοιχο κείμενο, ενσωματώνοντας ειδικά διακριτικά για την αναγνώριση γλώσσας, χρονικές σημάνσεις και εργασίες μετάφρασης [64].

Οι δυνατότητες πολυγλωσσικής μεταγραφής του Whisper αποτελούν σημαντική πρόοδο, επιτρέποντάς του να χειρίζεται πολλές γλώσσες και διαλέκτους με υψηλή ακρίβεια [64]. Η ανθεκτικότητα του μοντέλου σε θορυβώδη περιβάλλοντα επιτυγχάνεται μέσω εκτεταμένης εκπαίδευσης σε μια ποικιλία δεδομένων ήχου, συμπεριλαμβανομένου του θορύβου περιβάλλοντος και των διαφορετικών τόνων. Αυτή η εκπαίδευση επιτρέπει στον Whisper να διατηρεί υψηλές επιδόσεις ακόμη και σε απαιτητικά σενάρια του πραγματικού κόσμου [64],[67].

Η χρήση ενός μεγάλου και διαφορετικού συνόλου δεδομένων εκπαίδευσης, σε συνδυασμό με την αρχιτεκτονική που βασίζεται σε Transformer, έχει ως αποτέλεσμα ανώτερη ακρίβεια μεταγραφής σε σύγκριση με τα παραδοσιακά μοντέλα [64],[66]. Η ικανότητα του Whisper να χειρίζεται πολλές γλώσσες και διαλέκτους χωρίς εκτεταμένη προσαρμογή το καθιστά εξαιρετικά ευέλικτο και κατάλληλο για παγκόσμιες εφαρμογές [64]. Η ανθεκτικότητα σε θορυβώδη περιβάλλοντα και διαφορετικές πινελιές εξασφαλίζει αξιόπιστη απόδοση σε πρακτικές εφαρμογές, όπως ζωντανή μεταγραφή και εργαλεία προσβασιμότητας [64],[67].

Η αρχιτεκτονική του Whisper επιτρέπει ανώτερη απόδοση σε διάφορα σενάρια. Για παράδειγμα, σε εφαρμογές ζωντανής μεταγραφής, το Whisper μπορεί να μεταγράψει την ομιλία σε πραγματικό χρόνο με ακρίβεια, ακόμη και σε θορυβώδη περιβάλλοντα όπως συνέδρια ή δημόσιες εκδηλώσεις [64]. Επιπλέον, οι πολύγλωσσες δυνατότητές του το καθιστούν ένα ανεκτίμητο εργαλείο προσβασιμότητας, παρέχοντας ζωντανές μεταγραφές και μεταφράσεις για χρήστες με προβλήματα ακοής ή γλωσσικά εμπόδια [64].

Συμπερασματικά, η προηγμένη αρχιτεκτονική και οι μηχανισμοί λειτουργίας του Whisper το καθιστούν ένα ισχυρό και ευέλικτο σύστημα ASR. Η ικανότητά του να χειρίζεται πολύγλωσσες μεταγραφές και μεταφράσεις με υψηλή ακρίβεια και στιβαρότητα το ξεχωρίζει από άλλα μοντέλα, καθιστώντας το πολύτιμο εργαλείο για ένα ευρύ φάσμα εφαρμογών στον πραγματικό κόσμο [64].

2.4.2 Πώς Μετατρέπει την Ομιλία σε Κείμενο

Το Whisper είναι ένα αυτόματο σύστημα αναγνώρισης ομιλίας (ASR) που αναπτύχθηκε από την OpenAI. Μετατρέπει την ομιλία σε κείμενο χρησιμοποιώντας ένα μοντέλο νευρωνικού δικτύου εκπαιδευμένο σε ένα μεγάλο σύνολο δεδομένων από πολύγλωσσα και εποπτικά δεδομένα πολλαπλών εργασιών [64]. Οι κρυφές αναπαραστάσεις από τον κωδικοποιητή τροφοδοτούνται σε έναν αποκωδικοποιητή. Ο αποκωδικοποιητής είναι εκπαιδευμένος να προβλέπει τις αντίστοιχες λεζάντες κειμένου από τις λειτουργίες ήχου. Δημιουργεί διακριτικά διαδοχικά κειμένου, υπό την προϋπόθεση ότι τα προηγούμενα διακριτικά και οι κρυφές αναπαραστάσεις από τον κωδικοποιητή [64]. Ο αποκωδικοποιητής επίσης χρησιμοποιεί ειδικά διακριτικά για να εκτελέσει διάφορες εργασίες, όπως αναγνώριση γλώσσας, χρονικές σημειώσεις σε επίπεδο φράσης, πολυγλωσσική μεταγραφή ομιλίας και μετάφραση από άλλες γλώσσες στα αγγλικά [64]. Αυτά τα διακριτικά βοηθούν το μοντέλο να κατανοήσει το πλαίσιο και τη συγκεκριμένη εργασία που πρέπει να εκτελεστεί. Η τελική έξοδος είναι μια ακολουθία διακριτικών κειμένου που αντιπροσωπεύουν τη μεταγραφόμενη ομιλία. Το μοντέλο μπορεί επίσης να παρέχει χρονικές σημειώσεις για κάθε φράση, υποδεικνύοντας πότε εκφωνείται κάθε λέξη στον ήχο [64].

Ο Whisper εκπαιδεύτηκε σε 680.000 ώρες πολύγλωσσων και εποπτευόμενων δεδομένων πολλαπλών εργασιών που συλλέχθηκαν από τον Ιστό. Αυτό το μεγάλο και ποικίλο σύνολο δεδομένων βοηθά το μοντέλο να επιτύχει ανθεκτικότητα σε διαφορετικές προφορές, θόρυβο φόντου και τεχνικής γλώσσας [64]. Η αρχιτεκτονική του Whisper του επιτρέπει να εκτελεί πολλαπλές εργασίες όπως η αναγνώριση ομιλίας, η μετάφραση και η αναγνώριση γλώσσας χρησιμοποιώντας ένα μόνο μοντέλο [64]. Η ικανότητα του Whisper να κατανοεί πολλές γλώσσες προέρχεται από την εκτεταμένη εκπαίδευσή του σε ένα ποικίλο σύνολο δεδομένων που περιλαμβάνει ήχο και μεταγραφές σε πολλές γλώσσες [64]. Το μοντέλο μαθαίνει να αναγνωρίζει και να επεξεργάζεται διαφορετικά φωνητικά και γλωσσικά μοτίβα μεταξύ των γλωσσών. Επιπλέον, η χρήση ειδικών διακριτικών για αναγνώριση γλώσσας βοηθά το μοντέλο να εναλλάσσεται μεταξύ γλωσσών απρόσκοπτα [64].

Η εκπαίδευση μεγάλης κλίμακας δεδομένων περιλαμβάνει διάφορες γλώσσες, διαλέκτους και προφορές, επιτρέποντας στο Whisper να γενικεύει καλά σε νέες και μη εμφανείς γλώσσες [64]. Το Whisper χρησιμοποιεί μια αρχιτεκτονική Transformer, η οποία είναι εξαιρετικά αποτελεσματική για εργασίες αλληλουχίας σε ακολουθία (sequence-to-sequence) [64]. Το μοντέλο του Transformer αποτελείται από έναν κωδικοποιητή (encoder) και έναν αποκωδικοποιητή (decoder), τα οποία αποτελούνται από πολλαπλά επίπεδα νευρωνικών δικτύων αυτοπροσοχής (self-attention) και τροφοδοσίας (feed-forward) [64]. Ο μηχανισμός self-attention επιτρέπει στο μοντέλο να σταθμίζει τη σημασία διαφορετικών τμημάτων της ακολουθίας εισόδου όταν κάνει προβλέψεις, επιτρέποντάς του να συλλαμβάνει εξαρτήσεις μεγάλης εμβέλειας και πληροφορίες συμφραζομένων (contextual information) [64]. Ο μηχανισμός self-attention στο μοντέλο Transformer υπολογίζει ένα σύνολο βαρών προσοχής (attention weights) για κάθε θέση στην ακολουθία εισόδου. Αυτά τα βάρη καθορίζουν πόση επιρροή πρέπει να έχει κάθε θέση στην αναπαράσταση της τρέχουσας θέσης [64].

Αυτό επιτρέπει στο μοντέλο να εστιάζει σε σχετικά μέρη της εισόδου όταν κάνει προβλέψεις, βελτιώνοντας την ικανότητα του να χειρίζεται πολύπλοκα και ποικίλα μοτίβα ομιλίας [64]. Εφόσον το μοντέλο Transformer δεν έχει ενσωματωμένη έννοια της σειράς ακολουθίας, προστίθενται κωδικοποιήσεις θέσης στις ενσωματώσεις εισόδου για να παρέχουν πληροφορίες σχετικά με τις σχετικές θέσεις των διακριτικών στην ακολουθία [64]. Αυτό βοηθά το μοντέλο να κατανοήσει τη σειρά

των λέξεων και των φράσεων στον ήχο εισόδου [64]. Το Whisper εκπαιδεύεται χρησιμοποιώντας supervised learning, όπου το μοντέλο παρέχεται με ζεύγη εισόδου ήχου και αντίστοιχες μεταγραφές κειμένου [64]. Ο στόχος της εκπαίδευσης είναι να ελαχιστοποιηθεί η διαφορά μεταξύ του προβλεπόμενου κειμένου και της βασικής μεταγραφής της αλήθειας [64]. Αυτό γίνεται συνήθως χρησιμοποιώντας μια συνάρτηση απώλειας, όπως απώλεια cross-entropy, η οποία μετρά την απόκλιση μεταξύ της προβλεπόμενης κατανομής πιθανότητας σε διακριτικό κείμενο και της πραγματικής κατανομής [64].

Οι παράμετροι του μοντέλου βελτιστοποιούνται χρησιμοποιώντας αλγόριθμους βελτιστοποίησης που βασίζονται σε κλίση όπως ο AdamW [64]. Κατά τη διάρκεια της εκπαίδευσης, τεχνικές όπως ο προγραμματισμός του ρυθμού εκμάθησης (learning rate scheduling), η αποκοπή κλίσης (gradient clipping) και οι μέθοδοι τακτοποίησης (regularization method, π.χ. dropout) χρησιμοποιούνται για την αποφυγή υπερβολικής προσαρμογής (overfitting) [64]. Κατά τη διάρκεια της εξαγωγής συμπερασμάτων (inference), το εκπαιδευμένο μοντέλο επεξεργάζεται νέες εισόδους ήχου για να δημιουργήσει μεταγραφές κειμένου [64]. Η αναζήτηση δέσμης (beam search) χρησιμοποιείται συχνά για την εύρεση της πιο πιθανής ακολουθίας διακριτικών κειμένου, λαμβάνοντας υπόψη πολλαπλές πιθανές ακολουθίες και επιλέγοντας αυτή με τη μεγαλύτερη πιθανότητα [64]. Για τη βελτίωση της ευρωστίας του μοντέλου, πρόσθετη αύξηση δεδομένων (data augmentation) όπως το SpecAugment χρησιμοποιείται [64]. Το SpecAugment περιλαμβάνει την εφαρμογή τυχαίας στρέβλωσης χρόνου (time warping), κάλυψης συχνότητας (frequency masking) και κάλυψης χρόνου (time masking) στα φασματογράμματα εισόδου κατά τη διάρκεια της εκπαίδευσης [64]. Αυτό βοηθά το μοντέλο να γενικεύει καλύτερα σε διαφορετικούς τύπους παραμορφώσεων και θορύβου ήχου [64].

Χρησιμοποιείται μέθοδος τακτοποίησης (regularization method) όπως το Dropout και το Stochastic Depth για την αποφυγή υπερβολικής προσαρμογής (overfitting) [64]. Το Dropout ρίχνει τυχαία μονάδες από το νευρωνικό δίκτυο κατά τη διάρκεια της εκπαίδευσης, ενώ το Stochastic Depth παρακάμπτει τυχαία ολόκληρα στρώματα [64]. Αυτές οι τεχνικές βοηθούν το μοντέλο να μάθει πιο ισχυρά χαρακτηριστικά και να βελτιώσει τη γενίκευση (generalization) [64]. Η απόδοση του μοντέλου Whisper αξιολογείται χρησιμοποιώντας μετρήσεις όπως Word Error Rate (WER) και Character Error Rate (CER) [64]. Το WER μετρά το ποσοστό των λέξεων που έχουν μεταγραφεί λανθασμένα, ενώ το CER μετρά το ποσοστό των χαρακτήρων που έχουν μεταγραφεί λανθασμένα [64]. Αυτές οι μετρήσεις παρέχουν ένα ποσοτικό μέτρο της ακρίβειας του μοντέλου [64]. Η ικανότητα του Whisper να εκτελεί zero-shot learning είναι ένα βασικό χαρακτηριστικό [64]. Αυτό σημαίνει ότι το μοντέλο μπορεί να γενικευτεί σε νέες γλώσσες και εργασίες χωρίς να απαιτούνται πρόσθετα δεδομένα [64].

Η εκτεταμένη προεκπαίδευση σε ένα ποικίλο σύνολο δεδομένων επιτρέπει στο μοντέλο να χειρίζεται ένα ευρύ φάσμα εργασιών αναγνώρισης ομιλίας και μετάφρασης εκτός συσκευασιών [64]. Το Whisper έχει σχεδιαστεί για να είναι ανθεκτικό σε διάφορους τύπους θορύβου και παραμορφώσεις ήχου [64]. Το μεγάλο και ποικίλο σύνολο δεδομένων εκπαίδευσης περιλαμβάνει ήχο με διαφορετικά επίπεδα θορύβου φόντου, τόνους και συνθήκες καταγραφής [64]. Αυτό βοηθά το μοντέλο να αποδίδει καλά σε σενάρια πραγματικού κόσμου όπου η ποιότητα ήχου μπορεί να διαφέρει [64]. Το Whisper χρησιμοποιεί ειδικά διακριτικά για να αναγνωρίσει τη γλώσσα του ήχου εισόδου [64]. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο μαθαίνει να συσχετίζει αυτά τα διακριτικά με συγκεκριμένες γλώσσες, επιτρέποντάς του να εναλλάσσεται μεταξύ των γλωσσών απρόσκοπτα [64]. Αυτή η δυνατότητα είναι ζωτικής σημασίας για εργασίες αναγνώρισης και μετάφρασης σε πολλές γλώσσες [64].

Η εκπαίδευση του Whisper σε ένα διαφορετικό σύνολο δεδομένων του επιτρέπει να μαθαίνει διαφορετικά φωνητικά και γλωσσικά μοτίβα σε όλες τις γλώσσες [64]. Αυτό βοηθά το μοντέλο να αναγνωρίζει και να επεξεργάζεται την ομιλία σε διάφορες γλώσσες, ακόμη και σε αυτές με διαφορετικές φωνητικές δομές και γραμματικούς κανόνες [64]. Το μοντέλο δημιουργεί πολυγλωσσικές ενσωματώσεις (multilingual embeddings) που καταγράφουν τις σημασιολογικές και συντακτικές πληροφορίες διαφορετικών γλωσσών [64]. Αυτές οι ενσωματώσεις βοηθούν το μοντέλο να κατανοεί και να μεταφράζει την ομιλία από τη γλώσσα στην άλλη, βελτιώνοντας την απόδοσή του σε πολύγλωσσες εργασίες [64]. Οι μηχανισμοί προσοχής (attention mechanisms) στο μοντέλο Transformer βρίσκεται στο Whisper να εστιάζει σε διαφορετικά μέρη της ακολουθίας εισόδου όταν κάνει προβλέψεις [64]. Αυτό είναι ιδιαίτερα χρήσιμο για τον χειρισμό εξαρτήσεων μεγάλης εμβέλειας (long-range dependencies) και πληροφορίες με βάση τα συμφραζόμενα (contextual information) στην ομιλία [64]. Το μοντέλο μπορεί να προσαρμόσει δυναμικά την εστίασή του με βάση τη σημασία διαφορετικών τμημάτων της εισόδου, βελτιώνοντας την ικανότητα του να μεταγράφει πολύπλοκα και ποικίλα μοτίβα ομιλίας [64]. Κατά τη διάρκεια της εξαγωγής συμπερασμάτων (inference), το Whisper χρησιμοποιεί την αναζήτηση δέσμης (beam search) για να βρει την πιο πιθανή ακολουθία διακριτικών κειμένου [64].

Η αναζήτηση δέσμης (beam search) είναι ένας ευρετικός αλγόριθμος αναζήτησης που εξερευνά πολλαπλές πιθανές ακολουθίες και επιλέγει αυτή με τη μεγαλύτερη πιθανότητα [64]. Αυτό βοηθά το μοντέλο να δημιουργήσει πιο ακριβείς και συνεκτικές μεταγραφές, ειδικά σε περιπτώσεις όπου υπάρχουν πολλές ερμηνείες του ήχου εισόδου [64]. Η εκπαίδευση πολλαπλών εργασιών (multi-task training approach) του Whisper του επιτρέπει να εκτελεί διάφορες εργασίες που σχετίζονται με τον ομιλία χρησιμοποιώντας ένα μόνο μοντέλο [64]. Με την εκπαίδευση σε πολλαπλές εργασίες ταυτόχρονα, το μοντέλο μαθαίνει να γενικεύει καλύτερα και να αξιοποιεί την κοινή γνώση (shared knowledge) σε όλες τις εργασίες [64]. Αυτό βελτιώνει την απόδοσή του σε μεμονωμένες εργασίες και του δίνει τη δυνατότητα να χειριστεί ένα ευρύ φάσμα εφαρμογών επεξεργασίας ομιλίας [64]. Το Whisper αξιοποιεί τη μεταφορά μάθησης (transfer learning) για να βελτιώσει την απόδοσή του σε νέες εργασίες και γλώσσες [64]. Με την προεκπαίδευση σε ένα μεγάλο και ποικίλο σύνολο δεδομένων, το μοντέλο μαθαίνει γενικά χαρακτηριστικά και μοτίβα που μπορούν να μεταφερθούν σε νέες εργασίες με ελάχιστη πρόσθετη εκπαίδευση [64].

Αυτό καθιστά το Whisper εξαιρετικά προσαρμόσιμο και ικανό να χειρίζεται ένα ευρύ φάσμα εργασιών αναγνώρισης ομιλίας και μετάφρασης (speech recognition and translation tasks) [64]. Η ικανότητα του Whisper να κατανοεί το πλαίσιο (context) είναι ζωτικής σημασίας για την ακριβή αναγνώριση και τη μετάφραση ομιλίας [64]. Το μοντέλο μπορεί να αξιοποιήσει τις πληροφορίες συμφραζομένων (contextual information) από τον ήχο εισόδου και το προηγούμενο κείμενο για να κάνει πιο ενημερωμένες προβλέψεις [64]. Αυτό το βοηθά να χειρίζεται διφορούμενα ή πολύπλοκα μοτίβα ομιλίας και να δημιουργήσει πιο ακριβείς μεταγραφές [64]. Η αρχιτεκτονική του Whisper έχει σχεδιαστεί για να κλιμακώνεται αποτελεσματικά με τον όγκο των δεδομένων εκπαίδευσης και το μέγεθος του μοντέλου [64]. Αυξάνοντας το μέγεθος του μοντέλου (model size) και τον όγκο των δεδομένων εκπαίδευσης (training data volume), το Whisper μπορεί να επιτύχει μεγαλύτερη ακρίβεια και στιβαρότητα (robustness) [64]. Αυτή η επεκτασιμότητα το καθιστά κατάλληλο για μεγάλη κλίμακα εφαρμογών αναγνώρισης ομιλίας και μετάφρασης [64]. Το Whisper είναι ικανό για αναγνώριση και

μετάφραση ομιλίας σε πραγματικό χρόνο (real-time speech recognition and translation), καθιστώντας τις κατάλληλες για εφαρμογές που απαιτούν άμεση ανατροφοδότηση [64].

Η αποτελεσματική αρχιτεκτονική του μοντέλου και η βελτιστοποιημένη διαδικασία συμπερασμάτων (inference process) του φαίνεται να επεξεργάζεται γρήγορα εισόδους ήχου και να δημιουργεί ακριβείς μεταγραφές σε πραγματικό χρόνο (real-time) [64]. Συνοπτικά, το Whisper μετατρέπει την ομιλία σε κείμενο επεξεργάζοντας χαρακτηριστικά ήχου μέσω ενός μοντέλου μετασχηματιστή κωδικοποιητή-αποκωδικοποιητή (encoder-decoder transformer model), χρησιμοποιώντας ένα μεγάλο και ποικίλο σύνολο δεδομένων εκπαίδευσης (training data) για την υψηλή ακρίβεια και ευρωστία (robustness) σε διαφορετικές γλώσσες και εργασίες [64]. Το μοντέλο αξιοποιεί προηγμένες τεχνικές μηχανικής εκμάθησης (machine learning techniques), όπως αυτοπροσοχή (self-attention), κωδικοποίηση θέσης (positional encoding), εκμάθηση πολλαπλών εργασιών (multi-task learning), αύξηση δεδομένων (data augmentation) και τακτοποίηση (regularization) για να χειριστεί την πολυπλοκότητα της πολυγλωσσικής και πολλαπλής αναγνώρισης ομιλίας [64]. Η ικανότητα του Whisper να κατανοεί πολλές γλώσσες από την εκτεταμένη εκπαίδευσή του σε ένα ποικίλο σύνολο δεδομένων, τη χρήση ειδικών διακριτικών (special tokens) για την αναγνώριση της γλώσσας και την εκμάθηση φωνητικών και γλωσσικών προτύπων (patterns) μεταξύ των γλωσσών [64].

Επιπλέον, οι μηχανισμοί προσοχής του Whisper, η αναζήτηση δέσμης, η εκπαίδευση πολλαπλών εργασιών, η μάθηση μεταφοράς, η κατανόηση των συμφραζομένων, η επεκτασιμότητα και οι δυνατότητες επεξεργασίας σε πραγματικό χρόνο συμβάλλουν στην αποτελεσματικότητά του στο χειρισμό ενός ευρέος φάσματος εργασιών αναγνώρισης ομιλίας και μετάφρασης [64].

2.4.3 Εφαρμογές του Whisper στην Ανάλυση Φυσικής Γλώσσας

Μία από τις κύριες εφαρμογές του Whisper στο NLP είναι η ανάπτυξη εικονικών βοηθών που ενεργοποιούνται με φωνή. Αυτοί οι βοηθοί βασίζονται στην ακριβή αναγνώριση ομιλίας για να κατανοούν και να ανταποκρίνονται στις εντολές του χρήστη. Οι πολυγλωσσικές δυνατότητες και η επεξεργασία σε πραγματικό χρόνο του Whisper το καθιστούν ιδανικό για τη δημιουργία εικονικών βοηθών που μπορούν να λειτουργούν σε διαφορετικά γλωσσικά περιβάλλοντα [63].

Μια άλλη σημαντική εφαρμογή είναι στη μεταγραφή περιεχομένου ήχου. Το Whisper μπορεί να χρησιμοποιηθεί για τη μεταγραφή συνεντεύξεων, podcast και συσκέψεων, καθιστώντας το περιεχόμενο αναζητήσιμο και ευκολότερο στην ανάλυση. Αυτό είναι ιδιαίτερα χρήσιμο σε τομείς όπως η δημοσιογραφία, η έρευνα και οι επιχειρήσεις, όπου η ακριβής μεταγραφή είναι απαραίτητη για την τεκμηρίωση και την ανάλυση [63].

Το Whisper παίζει επίσης κρίσιμο ρόλο στη γλωσσική μετάφραση. Με τη μεταγραφή της ομιλίας σε μια γλώσσα και τη μετάφρασή της σε μια άλλη, το Whisper διευκολύνει την επικοινωνία μεταξύ ομιλητών διαφορετικών γλωσσών. Αυτή η εφαρμογή είναι πολύτιμη στις διεθνείς επιχειρήσεις, τα ταξίδια και την εξυπηρέτηση πελατών, όπου τα γλωσσικά εμπόδια μπορεί να είναι μια σημαντική πρόκληση [63].

Επιπλέον, η ικανότητα του Whisper να παρέχει χρονικές σημάνσεις σε επίπεδο φράσης είναι ευεργετική για τη δημιουργία υπότιτλων για βίντεο ή podcasts. Αυτό βελτιώνει την προσβασιμότητα για άτομα με προβλήματα ακοής και βελτιώνει την εμπειρία θέασης για τους μη μητρικούς ομιλητές. Επιπλέον, η ισχυρή απόδοση του Whisper σε θορυβώδη περιβάλλοντα το καθιστά κατάλληλο για χρήση σε αυτοματοποιημένα συστήματα εξυπηρέτησης πελατών. Τα συστήματα αυτά μπορούν να χειριστούν τα ερωτήματα των πελατών μέσω φωνητικών αλληλεπιδράσεων, παρέχοντας γρήγορη και αποτελεσματική υποστήριξη χωρίς την ανάγκη ανθρώπινης παρέμβασης [63].

Το Whisper έχει επίσης εφαρμογές στους παρακάτω τομείς όπως η εκπαίδευση και διδασκαλία, όπου μπορεί να χρησιμοποιηθεί για τη μετατροπή διαλέξεων και μαθημάτων σε κείμενο, καθιστώντας το περιεχόμενο ευκολότερο στην ανάγνωση και μελέτη για τους μαθητές. Στην Υγειονομική Περίθαλψη, η ακριβής μεταγραφή ιατρικών συνομιλιών μπορεί να είναι εξαιρετικά χρήσιμη για τη διατήρηση ιατρικών αρχείων και την ανάλυση ασθενών. Στην Εξυπηρέτηση Πελατών, οι πολυγλωσσικές δυνατότητες του Whisper το καθιστούν ιδανικό για χρήση σε κέντρα εξυπηρέτησης πελατών που χρειάζονται επικοινωνία σε πολλές γλώσσες. Στη Διαφήμιση και Μάρκετινγκ, η μεταγραφή και ανάλυση συνεντεύξεων και συζητήσεων με πελάτες για την καλύτερη κατανόηση των αναγκών και των επιθυμιών τους. Τέλος, στην Απομαγνητοφώνηση Συναντήσεων, η μεταγραφή εταιρικών συναντήσεων και συμβουλίων για καλύτερη τεκμηρίωση και αναφορά [63].

Συνοπτικά, το Whisper μετατρέπει την ομιλία σε κείμενο επεξεργάζοντας χαρακτηριστικά ήχου μέσω ενός μοντέλου μετασχηματιστή κωδικοποιητή-αποκωδικοποιητή, χρησιμοποιώντας ένα μεγάλο και ποικίλο σύνολο δεδομένων εκπαίδευσης για την επίτευξη υψηλής ακρίβειας και ευρωστίας σε διαφορετικές γλώσσες και εργασίες. Το μοντέλο αξιοποιεί προηγμένες τεχνικές μηχανικής εκμάθησης, όπως αυτοπροσοχή, κωδικοποίηση θέσης, εκμάθηση πολλαπλών εργασιών, αύξηση δεδομένων και τακτοποίηση για να χειριστεί την πολυπλοκότητα της πολυγλωσσικής και πολλαπλής αναγνώρισης ομιλίας [63].

2.5 Deep FilterNet: Αποθορυβοποίηση και Βελτίωση Ήχου

Το Deep FilterNet είναι ένα πλαίσιο βελτίωσης ήχου που χρησιμοποιεί βαθιά μάθηση (deep learning) για την αποθορυβοποίηση (denoising) και τη βελτίωση της ποιότητας του ήχου. Η προσέγγιση αυτή βασίζεται στη χρήση φίλτρων (filters) που εφαρμόζονται σε πολλαπλά χρονικά και συχνοτικά διαστήματα (temporal and spectral domains), επιτρέποντας την ανάκτηση του καθαρού σήματος από θορυβώδεις ηχογραφήσεις. Το DeepFilterNet εκμεταλλεύεται τις ιδιότητες της ανθρώπινης ακουστικής αντίληψης (human auditory perception) και χρησιμοποιεί τεχνικές όπως οι διαχωριστικές συνελίξεις (separable convolutions) και οι επαναλαμβανόμενες νευρωνικές μονάδες (recurrent neural units) για να επιτύχει χαμηλή πολυπλοκότητα (low complexity) και υψηλή απόδοση σε πραγματικό χρόνο (real-time performance) [68].

Η αρχιτεκτονική του Deep FilterNet περιλαμβάνει δύο στάδια. Το πρώτο στάδιο βελτιώνει τον φασματικό φάκελο του σήματος χρησιμοποιώντας κέρδη κλιμακωμένα με βάση την ισοδύναμη ορθογώνια ζώνη (Equivalent Rectangular Bandwidth, ERB), ενώ το δεύτερο στάδιο εφαρμόζει βαθιά φίλτρα (deep filters) για την ενίσχυση των περιοδικών συνιστωσών του λόγου. Η προσέγγιση αυτή επιτρέπει την αποτελεσματική αποθορυβοποίηση (denoising) και την ενίσχυση της ποιότητας του ήχου, καθιστώντας το Deep FilterNet κατάλληλο για εφαρμογές όπως η αναγνώριση ομιλίας (speech

recognition), τα συστήματα τηλεδιάσκεψης (teleconferencing systems) και οι βοηθητικές συσκευές ακρόασης (assistive listening devices) [68].

Το πρώτο στάδιο της αρχιτεκτονικής του Deep FilterNet χρησιμοποιεί ένα φίλτρο ισοδύναμης ορθογώνιας ζώνης (Equivalent Rectangular Bandwidth, ERB) για να μειώσει τις διαστάσεις εισόδου και εξόδου σε 32 ζώνες, επιτρέποντας τη χρήση ενός υπολογιστικά φθηνού δικτύου κωδικοποιητή/αποκωδικοποιητή (encoder/decoder network). Το δεύτερο στάδιο εφαρμόζει βαθιά φίλτρα (deep filters) για την ενίσχυση των περιοδικών συνιστωσών του λόγου, χρησιμοποιώντας γραμμικά φίλτρα με σύνθετους συντελεστές (linear filters with complex coefficients) που εφαρμόζονται στις αντίστοιχες συχνοτικές ζώνες (frequency bands). Αυτή η προσέγγιση επιτρέπει την ανάκτηση του καθαρού σήματος από θορυβώδεις ηχογραφήσεις με υψηλή απόδοση (high performance) και χαμηλή πολυπλοκότητα (low complexity) [68].

Η χρήση διαχωριστικών συνελίξεων (separable convolutions) και εκτεταμένης ομαδοποίησης (extensive pooling) σε γραμμικά και επαναλαμβανόμενα στρώματα (linear and recurrent layers) επιτρέπει την επίτευξη χαμηλής πολυπλοκότητας (low complexity) και υψηλής απόδοσης σε πραγματικό χρόνο. Το DeepFilterNet εκμεταλλεύεται τις ιδιότητες της ανθρώπινης ακουστικής αντίληψης (human auditory perception) για να επιτύχει βέλτιστα αποτελέσματα αποθορυβοποίησης (denoising) και βελτίωσης της ποιότητας του ήχου (audio quality enhancement) [68].

Συνολικά, το DeepFilterNet προσφέρει μια καινοτόμο λύση για την αποθορυβοποίηση και τη βελτίωση του ήχου, συνδυάζοντας την υψηλή απόδοση με τη χαμηλή πολυπλοκότητα, καθιστώντας το ιδανικό για χρήση σε ενσωματωμένες συσκευές και εφαρμογές πραγματικού χρόνου [68].

2.5.1 Μηχανισμοί Αποθορυβοποίησης στο Deep FilterNet

Το Deep FilterNet χρησιμοποιεί προηγμένους μηχανισμούς αποθορυβοποίησης για να βελτιώσει την ποιότητα του ήχου σε θορυβώδη περιβάλλοντα. Η αρχιτεκτονική του περιλαμβάνει δύο κύρια στάδια. Στο πρώτο στάδιο, το DeepFilterNet βελτιώνει τον φασματικό φάκελο του σήματος (spectral envelope) χρησιμοποιώντας κέρδη κλιμακωμένα με βάση την ισοδύναμη ορθογώνια ζώνη (Equivalent Rectangular Bandwidth, ERB). Αυτό επιτρέπει τη μείωση των διαστάσεων εισόδου και εξόδου σε 32 ζώνες, καθιστώντας δυνατή τη χρήση ενός υπολογιστικά φθηνού δικτύου κωδικοποιητή/αποκωδικοποιητή (encoder/decoder network) [68]. Στο δεύτερο στάδιο, το Deep FilterNet εφαρμόζει βαθιά φίλτρα (deep filters) για την ενίσχυση των περιοδικών συνιστωσών του λόγου. Χρησιμοποιεί γραμμικά φίλτρα με σύνθετους συντελεστές (linear filters with complex coefficients) που εφαρμόζονται στις αντίστοιχες συχνοτικές ζώνες (frequency bands). Αυτή η προσέγγιση επιτρέπει την ανάκτηση του καθαρού σήματος από θορυβώδεις ηχογραφήσεις με υψηλή απόδοση και χαμηλή πολυπλοκότητα [68].

Επιπλέον, η χρήση διαχωριστικών συνελίξεων και εκτεταμένης ομαδοποίησης σε γραμμικά και επαναλαμβανόμενα στρώματα επιτρέπει την επίτευξη χαμηλής πολυπλοκότητας και υψηλής απόδοσης σε πραγματικό χρόνο. Το Deep FilterNet εκμεταλλεύεται τις ιδιότητες της ανθρώπινης ακουστικής αντίληψης για να επιτύχει βέλτιστα αποτελέσματα αποθορυβοποίησης και βελτίωσης της ποιότητας του ήχου [68].

Η καινοτομία του DeepFilterNet έγκειται στην ικανότητά του να προσαρμόζεται σε διάφορους τύπους θορύβων και να εξάγει καθαρό ήχο ακόμη και σε ιδιαίτερα δύσκολα περιβάλλοντα. Οι αλγόριθμοι

αποθορυβοποίησης του έχουν εκπαιδευτεί σε μεγάλο πλήθος δεδομένων που περιέχουν ποικίλα είδη θορύβων, επιτρέποντας την αποτελεσματική διαχείριση διαφορετικών σεναρίων [68]. Ενσωματώνει επίσης προηγμένες τεχνικές μάθησης μηχανών, όπως βαθιά νευρωνικά δίκτυα, για την ακριβή ανάλυση και διόρθωση των σημάτων [68].

2.5.2 Αρχιτεκτονική και Τεχνικές για Καθαρισμό Ήχου

Η αποθορυβοποίηση και ο καθαρισμός ήχου αποτελούν σημαντικές προκλήσεις στον τομέα της επεξεργασίας σήματος, ιδιαίτερα σε εφαρμογές όπως η αναγνώριση ομιλίας, η παραγωγή μουσικής και οι βοηθητικές συσκευές ακρόασης. Οι παραδοσιακές μέθοδοι βασίζονται σε φίλτρα και τεχνικές επεξεργασίας σήματος, αλλά συχνά δεν επιτυγχάνουν ικανοποιητικά αποτελέσματα χωρίς να εισάγουν παραμορφώσεις [69].

Τα τελευταία χρόνια, οι τεχνικές βαθιάς μάθησης έχουν αναδειχθεί ως μια πολλά υποσχόμενη λύση για τον καθαρισμό ήχου. Αυτές οι μέθοδοι εκμεταλλεύονται την ισχύ των βαθιών νευρωνικών δικτύων (deep neural networks) για να εξάγουν σημαντικά χαρακτηριστικά (features) από θορυβώδη σήματα και να τα χρησιμοποιούν για την αποθορυβοποίηση (denoising) του ήχου [70]. Μια από τις πιο δημοφιλείς αρχιτεκτονικές για τον καθαρισμό ήχου είναι το UNet, το οποίο χρησιμοποιείται ευρέως για την αποθορυβοποίηση και την ενίσχυση της ποιότητας του ήχου. Το UNet αποτελείται από έναν κωδικοποιητή (encoder) και έναν αποκωδικοποιητή (decoder), οι οποίοι συνεργάζονται για να μάθουν την αντιστοίχιση μεταξύ θορυβωδών και καθαρών σημάτων. Ο κωδικοποιητής εξάγει χαρακτηριστικά από το θορυβώδες σήμα, ενώ ο αποκωδικοποιητής χρησιμοποιεί αυτά τα χαρακτηριστικά για να ανακατασκευάσει το καθαρό σήμα [71].

Εκτός από το UNet, άλλες αρχιτεκτονικές όπως τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNNs) και τα υβριδικά δίκτυα έχουν επίσης προταθεί για την αποθορυβοποίηση του ήχου. Τα RNNs είναι ιδιαίτερα αποτελεσματικά στην επεξεργασία χρονικών ακολουθιών, ενώ τα υβριδικά δίκτυα συνδυάζουν τα πλεονεκτήματα διαφορετικών αρχιτεκτονικών για να επιτύχουν καλύτερα αποτελέσματα [70].

Οι τεχνικές βαθιάς μάθησης για τον καθαρισμό ήχου μπορούν να κατηγοριοποιηθούν σε επιβλεπόμενες (supervised) και μη επιβλεπόμενες μεθόδους (unsupervised methods). Οι επιβλεπόμενες μέθοδοι εκπαιδεύονται σε σύνολα δεδομένων που περιλαμβάνουν ζεύγη θορυβωδών και καθαρών σημάτων (pairs of noisy and clean signals), ενώ οι μη επιβλεπόμενες μέθοδοι (unsupervised methods) προσπαθούν να μάθουν την κατανομή του θορύβου (noise distribution) απευθείας από το θορυβώδες σήμα χωρίς να απαιτείται καθαρό σύνολο εκπαίδευσης [72].

Η διαδικασία λειτουργίας των τεχνικών βαθιάς μάθησης για τον καθαρισμό ήχου περιλαμβάνει τα εξής βήματα. Αρχικά, συλλέγονται ζεύγη θορυβωδών και καθαρών σημάτων για την εκπαίδευση του μοντέλου. Τα δεδομένα αυτά μπορεί να περιλαμβάνουν διάφορους τύπους θορύβου και διαφορετικές πηγές ήχου [70]. Στη συνέχεια, τα δεδομένα προεπεξεργάζονται για να δημιουργηθούν τα χαρακτηριστικά που θα χρησιμοποιηθούν ως είσοδος στο νευρωνικό δίκτυο. Αυτό μπορεί να περιλαμβάνει τη μετατροπή των σημάτων σε φασματικά χαρακτηριστικά, όπως τα φασματογραφήματα [71]. Κατόπιν, το νευρωνικό δίκτυο εκπαιδεύεται χρησιμοποιώντας τα προεπεξεργασμένα δεδομένα. Κατά την εκπαίδευση, το μοντέλο μαθαίνει να αντιστοιχίζει τα θορυβώδη σήματα με τα καθαρά σήματα, προσαρμόζοντας τα βάρη του για να ελαχιστοποιήσει το σφάλμα [70]. Μετά την εκπαίδευση, το μοντέλο μπορεί να χρησιμοποιηθεί για την αποθορυβοποίηση νέων θορυβωδών σημάτων. Το θορυβώδες σήμα εισάγεται στο μοντέλο, το οποίο εξάγει το καθαρό σήμα [71]. Τέλος, η απόδοση του

μοντέλου αξιολογείται χρησιμοποιώντας μετρικές όπως το Mean Squared Error (MSE) και το Perceptual Evaluation of Speech Quality (PESQ), για να διασφαλιστεί ότι το μοντέλο επιτυγχάνει την επιθυμητή ποιότητα ήχου [72].

Συνολικά, οι τεχνικές βαθιάς μάθησης προσφέρουν σημαντικά πλεονεκτήματα στον καθαρισμό ήχου, καθώς μπορούν να μάθουν περίπλοκες αναπαραστάσεις των θορυβωδών σημάτων και να επιτύχουν αποτελεσματική αποθορυβοποίηση σε διάφορους τύπους θορύβου. Η συνεχής έρευνα και ανάπτυξη σε αυτόν τον τομέα αναμένεται να οδηγήσει σε ακόμα πιο προηγμένες λύσεις για την αποθορυβοποίηση και την ενίσχυση της ποιότητας του ήχου [70].

2.5.3 Εφαρμογές στην Επεξεργασία Ήχου

Η επεξεργασία ήχου αποτελεί έναν από τους πιο σημαντικούς τομείς της σύγχρονης τεχνολογίας, με εφαρμογές που εκτείνονται από την αναγνώριση ομιλίας και την παραγωγή μουσικής, μέχρι τις βοηθητικές συσκευές ακρόασης και τα συστήματα τηλεδιάσκεψης. Οι τεχνικές επεξεργασίας ήχου χρησιμοποιούνται για τη βελτίωση της ποιότητας του ήχου, την αποθορυβοποίηση, την ενίσχυση της ομιλίας και την ανάλυση των ηχητικών σημάτων [69].

Η αναγνώριση ομιλίας είναι μια από τις πιο διαδεδομένες εφαρμογές της επεξεργασίας ήχου. Χρησιμοποιείται σε συστήματα φωνητικής αναγνώρισης, όπως οι ψηφιακοί βοηθοί (π.χ. Siri, Google Assistant), για την κατανόηση και την εκτέλεση εντολών από τον χρήστη. Οι τεχνικές αποθορυβοποίησης και ενίσχυσης της ομιλίας βελτιώνουν την ακρίβεια των συστημάτων αναγνώρισης ομιλίας, επιτρέποντας την καλύτερη κατανόηση της ομιλίας σε θορυβώδη περιβάλλοντα [70]. Επιπλέον, η επεξεργασία ήχου μπορεί να χρησιμοποιηθεί για τη βελτίωση της ποιότητας των ηχογραφήσεων σε συστήματα αυτόματης μεταγραφής και μετάφρασης [73].

Στην παραγωγή μουσικής, οι τεχνικές επεξεργασίας ήχου χρησιμοποιούνται για την αφαίρεση ανεπιθύμητων θορύβων από ηχογραφήσεις, την ενίσχυση της ποιότητας του ήχου και την προσθήκη εφέ. Οι επαγγελματίες του ήχου χρησιμοποιούν εργαλεία όπως οι ισοσταθμιστές, οι συμπιεστές και τα φίλτρα για να βελτιώσουν την ηχητική εμπειρία και να δημιουργήσουν υψηλής ποιότητας μουσικά κομμάτια [71]. Επιπλέον, η επεξεργασία ήχου χρησιμοποιείται για τη δημιουργία ειδικών ηχητικών εφέ και την ανάμιξη πολλαπλών ηχητικών πηγών για την παραγωγή τελικών μουσικών κομματιών [69].

Οι βοηθητικές συσκευές ακρόασης, όπως τα ακουστικά βαρηκοΐας, χρησιμοποιούν τεχνικές επεξεργασίας ήχου για να βελτιώσουν την ακουστική εμπειρία των ατόμων με προβλήματα ακοής. Η αποθορυβοποίηση και η ενίσχυση της ομιλίας επιτρέπουν στους χρήστες να ακούνε καθαρότερα και πιο ευκρινή ήχο, βελτιώνοντας την ποιότητα ζωής τους [70]. Επιπλέον, οι τεχνικές επεξεργασίας ήχου μπορούν να χρησιμοποιηθούν για την προσαρμογή των ακουστικών συσκευών στις ατομικές ανάγκες των χρηστών, προσφέροντας εξατομικευμένες λύσεις ακρόασης [73].

Στα συστήματα τηλεδιάσκεψης, η επεξεργασία ήχου χρησιμοποιείται για την αποθορυβοποίηση και την ενίσχυση της ομιλίας, επιτρέποντας την καθαρή και ευκρινή επικοινωνία μεταξύ των συμμετεχόντων. Οι τεχνικές επεξεργασίας ήχου μπορούν να μειώσουν τον θόρυβο περιβάλλοντος και να βελτιώσουν την ποιότητα του ήχου, καθιστώντας τις τηλεδιασκέψεις πιο αποτελεσματικές και ευχάριστες [71]. Επιπλέον, η επεξεργασία ήχου μπορεί να χρησιμοποιηθεί για την αυτόματη ανίχνευση και καταστολή της ηχούς, βελτιώνοντας την εμπειρία των χρηστών [69].

Η ανάλυση ηχητικών σημάτων περιλαμβάνει την εξαγωγή και την ανάλυση χαρακτηριστικών από ηχητικά σήματα για διάφορες εφαρμογές, όπως η αναγνώριση συναισθημάτων, η ανίχνευση ανωμαλιών και η ανάλυση βιομετρικών δεδομένων. Οι τεχνικές επεξεργασίας ήχου μπορούν να χρησιμοποιηθούν για την ανάλυση της φωνής και την εξαγωγή πληροφοριών σχετικά με την υγεία και την ευεξία των ατόμων, καθώς και για την ανίχνευση και την αναγνώριση ηχητικών γεγονότων σε περιβάλλοντα παρακολούθησης και ασφάλειας [70].

Πέρα από αυτά, η επεξεργασία ήχου βρίσκει εφαρμογή και στις τεχνολογίες εικονικής και επαυξημένης πραγματικότητας (VR και AR), βελτιώνοντας την αίσθηση του ήχου στον εικονικό κόσμο και προσφέροντας πιο ρεαλιστικές και καθηλωτικές εμπειρίες. Οι τεχνικές επεξεργασίας ήχου μπορούν να χρησιμοποιηθούν για τη δημιουργία περιβαλλοντικού ήχου, την αναπαραγωγή τρισδιάστατων ηχητικών εφέ και την προσαρμογή του ήχου στις κινήσεις του χρήστη μέσα στον εικονικό χώρο [73]. Αυτό προσδίδει μια νέα διάσταση στην εμπειρία της εικονικής και επαυξημένης πραγματικότητας, καθιστώντας τη ακόμα πιο συναρπαστική και αληθοφανή [71].

Η επεξεργασία ήχου έχει επίσης σημαντικό ρόλο στην ανάπτυξη συστημάτων υποβοηθούμενης διαβίωσης και έξυπνων σπιτιών. Οι τεχνικές ανίχνευσης και αναγνώρισης ήχου μπορούν να χρησιμοποιηθούν για την ανίχνευση επικίνδυνων καταστάσεων, όπως η πτώση ηλικιωμένων ή ηχητικά σήματα επείγουσας ανάγκης, επιτρέποντας την έγκαιρη παρέμβαση και βελτιώνοντας την ασφάλεια και την ποιότητα ζωής των χρηστών [69]. Επιπλέον, τα έξυπνα ηχεία και οι φωνητικοί βοηθοί μπορούν να χρησιμοποιούν επεξεργασία ήχου για την αναγνώριση φωνητικών εντολών και την αλληλεπίδραση με τους χρήστες, προσφέροντας μια πιο φυσική και άνετη εμπειρία χρήσης των έξυπνων συσκευών [70].

Συνολικά, οι εφαρμογές της επεξεργασίας ήχου καλύπτουν ένα ευρύ φάσμα τομέων και προσφέρουν σημαντικά πλεονεκτήματα στην ποιότητα και την αποτελεσματικότητα της επικοινωνίας, της ψυχαγωγίας και της ανάλυσης δεδομένων [73].

2.6 BART: Προηγμένος Μετασχηματιστής για Ανάλυση Κειμένου

Αυτό το μοντέλο γλωσσικής επεξεργασίας άλλαξε την ανάλυση κειμένου με τη συνδυασμένη χρήση bidirectional και autoregressive τεχνικών, πετυχαίνοντας εξαιρετικά αποτελέσματα σε εργασίες όπως δημιουργία κειμένου, περίληψη και μετάφραση [74],[75]. Η αρχιτεκτονική του ενδυναμώνει την ικανότητα κατανόησης και παραγωγής κειμένου με ακρίβεια [74]. Βασίζεται σε προεκπαίδευση σε τεράστιες συλλογές κειμένων όπως βιβλία, άρθρα και διαδικτυακό περιεχόμενο, ετσι ώστε να γίνει εξειδίκευση για να χρησιμοποιηθεί πανω σε συγκεκριμένες εφαρμογές [74],[75]. Διαδικασίες όπως η αποθορυβοποίηση και η γενετική μοντελοποίηση συμβάλλουν στην εκμάθηση πλούσιων αναπαραστάσεων πληροφορίας με βάση το περιεχόμενο [74],[75].

Η υψηλή απόδοσή του καθιστά το μοντέλο ένα από τα πιο σημαντικά εργαλεία στη γλωσσική τεχνολογία [74],[73]. Έχει διαπιστωθεί η αποτελεσματικότητά του σε εφαρμογές όπως ταξινόμηση κειμένων, απάντηση ερωτήσεων και δημιουργία διαλόγων [74],[73]. Οι δυνατότητές του στην κατανόηση και παραγωγή φυσικής γλώσσας το κάνουν ιδανικό για έξυπνους συνομιλητές και ψηφιακούς βοηθούς [74],[73]. Παρά τα πλεονεκτήματά του, απαιτεί εκτεταμένα εκπαιδευτικά δεδομένα

και σημαντικούς υπολογιστικούς πόρους [74],[75]. Σε εξέλιξη βρίσκεται έρευνα για τη βελτιστοποίησή του και την αντιμετώπιση αυτών των προκλήσεων [74],[75].

2.6.1 Λειτουργία και αρχιτεκτονική του BART

Ο BART αντιπροσωπεύει μια πρωτοποριακή πρόοδο στην επεξεργασία φυσικής γλώσσας, ενσωματώνοντας τις δυνατότητες αμφίδρομης (bidirectional) και αυτόματης παλινδρόμησης (autoregressive) στην αρχιτεκτονική του. Αποτελούμενο από δύο κύρια στοιχεία — κωδικοποιητή (encoder) και αποκωδικοποιητή (decoder) — αξιοποιεί την αμφίδρομη επεξεργασία στον πρώτο για να συλλάβει πληροφορίες από τα συμφραζόμενα τόσο από προηγούμενα όσο και από μελλοντικά διακριτικά, ενώ ο δεύτερος παράγει έξοδο αυτοπαλινδρομικά, βασιζόμενος σε λέξεις που είχαν προβλεφθεί προηγουμένως [74],[75].
Στη προεκπαίδευση, περιλαμβάνονται δύο βασικές φάσεις, η αποθορυβοποίηση (denoising) και δημιουργική μοντελοποίηση (generative modeling). Στην πρώτη, το σύστημα εκπαιδεύεται να ανακατασκευάζει το αρχικό κείμενο από κατεστραμμένες εκδόσεις, ενώ η δεύτερη εστιάζει στην παραγωγή συνεκτικού κειμένου με βάση πληροφορίες συμφραζομένων [74],[75].

Στη φάση της αποθορυβοποίησης, εισάγεται θόρυβος (noise) στο κείμενο, επιτρέποντας στην αρχιτεκτονική να μάθει να ανακατασκευάζει την αρχική είσοδο. Η δημιουργία κειμένου γίνεται με διαδοχική πρόβλεψη επόμενων λέξεων, όπου κάθε νέα λέξη εξαρτάται από το προηγούμενο κείμενο και το γενικότερο πλαίσιο. [74],[75]. Η εκπαίδευση σε μεγάλα σώματα κειμένων διευκολύνει την απόκτηση πλούσιων γλωσσικών αναπαραστάσεων (rich language representations), οι οποίες στη συνέχεια ρυθμίζονται με ακρίβεια για συγκεκριμένες εργασίες, χρησιμοποιώντας εξειδικευμένα σύνολα δεδομένων [74],[75].

Ο μηχανισμός αυτοπροσοχής (self-attention) του encoder υπολογίζει τα βάρη της προσοχής (attention weights) μεταξύ λέξεων, ενσωματώνοντας πληροφορίες από ολόκληρο το κείμενο. Η τεχνική αυτή ενισχύει την ακρίβεια και τη συνοχή του παραγόμενου υλικού, αποδίδοντας προτεραιότητα σε σημαντικές λέξεις ή φράσεις [74],[75]. Στη λεπτή ρύθμιση (fine-tuning), εφαρμόζονται τεχνικές βελτιστοποίησης υπερπαραμέτρων (hyperparameter optimization) και έγκαιρης διακοπής (early stopping) για τον μετριασμό της υπερπροσαρμογής (overfitting), βελτιώνοντας την απόδοση σε εργασίες επεξεργασίας γλώσσας [74],[75].

Η εκτεταμένη χρήση δεδομένων κατά την προεκπαίδευση επιτρέπει την ανάπτυξη γενικευμένων γλωσσικών αναπαραστάσεων, μεταφερόμενων σε διάφορες εργασίες με υψηλή απόδοση ακόμη και με περιορισμένα δεδομένα [74],[75]. Η αρχιτεκτονική ενσωματώνει τεχνικές αύξησης δεδομένων (data augmentation) και προσομοίωσης (synthetic data), ενισχύοντας την απόδοσή της σε εξειδικευμένες εφαρμογές. Μέσω αυτών, το μοντέλο γενικεύει αποτελεσματικά σε νέα δεδομένα, βελτιώνοντας την εφαρμοστικότητά του σε πραγματικές συνθήκες [74],[75].

2.6.2 Δυνατότητες του BART

Το μοντέλο συνδυάζει διπλής κατεύθυνσης κωδικοποιητή (bidirectional encoder) και αυτόματης παλινδρόμησης αποκωδικοποιητή (autoregressive decoder), προσφέροντας ευελιξία σε εργασίες

επεξεργασίας φυσικής γλώσσας (NLP) [75]. Προσφέρει μια σειρά από δυνατότητες, όπως τη δημιουργία κειμένου (άρθρων, email ή απαντήσεων για chatbots), τη σύνοψη κειμένων με τεχνικές abstractive summarization και την ακριβή μετάφραση γλωσσών με σημασιολογική κατανόηση [75],[76],[77]. Επιπλέον, χρησιμοποιείται για την ταξινόμηση email (spam vs. μη-spam), την αξιολόγηση συναισθημάτων και την οργάνωση επιστημονικών άρθρων, καθώς και για την ανάλυση της δημόσιας γνώμης σε κοινωνικά δίκτυα [75],[77].

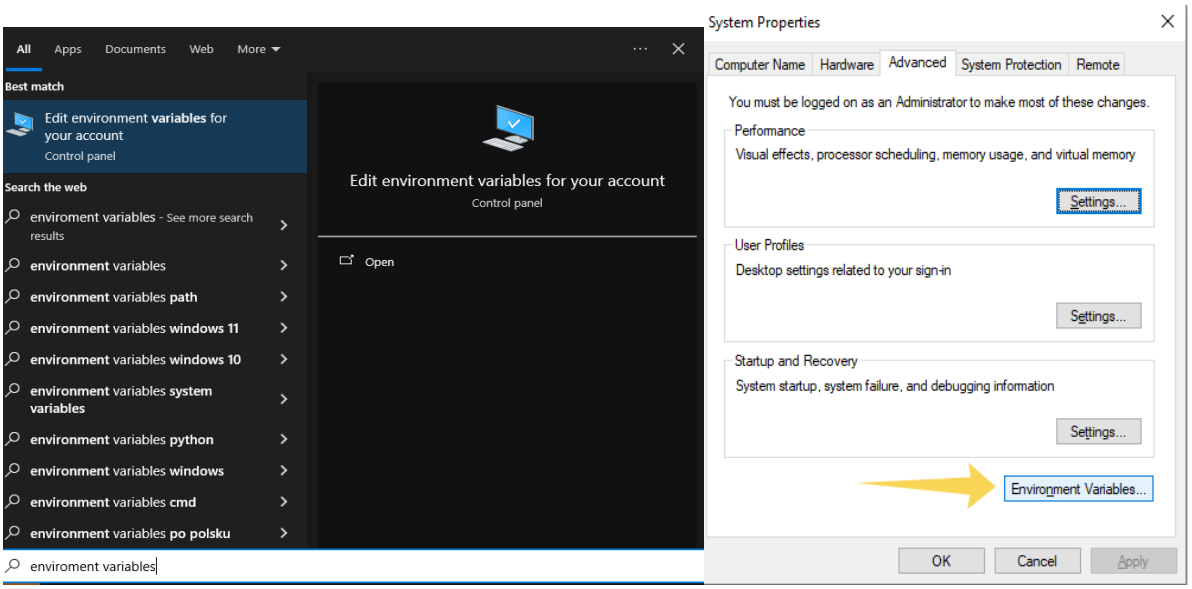
Το μοντέλο βρίσκει εφαρμογές σε πολλούς τομείς. Στη δημοσιογραφία και το ψηφιακό μάρκετινγκ, υποστηρίζει την αυτόματη συγγραφή άρθρων, τη δημιουργία email campaigns και την παροχή σύντομων συνοψέων για γρήγορη ανάλυση [75],[76]. Στην επιχειρησιακή αποτελεσματικότητα, συμβάλλει μέσω φιλτραρίσματος spam, ανάλυσης κριτικών προϊόντων και αυτοματοποιημένης εξυπηρέτησης πελατών, ενώ στη διεθνή επικοινωνία προσφέρει ακριβή μετάφραση εγγράφων για πολυγλωσσική συνεργασία [75],[77]. Επίσης, χρησιμοποιείται στην επιστημονική έρευνα για την εξαγωγή πληροφοριών από ακαδημαϊκά άρθρα και την αναγνώριση θεματικών τάσεων, καθώς και στην κοινωνική ανάλυση για την παρακολούθηση συναισθημάτων στα social media [75],[77].

Τέλος, το μοντέλο διακρίνεται για τα τεχνικά του χαρακτηριστικά. Χρησιμοποιεί μηχανισμούς αυτοπροσοχής (self-attention) για την ανάλυση σχέσεων μεταξύ λέξεων και είναι προεκπαιδευμένο με τεχνικές αποθορυβοποίησης (denoising) και γενετικής μοντελοποίησης [75],[76]. Επιπλέον, υποστηρίζει λεπτή ρύθμιση (fine-tuning) για εξειδικευμένες εργασίες, ακόμη και με ελάχιστα δεδομένα, γεγονός που το καθιστά ιδιαίτερα ευέλικτο και αποτελεσματικό σε ποικίλες εφαρμογές [75].

Κεφάλαιο 3ο: Εφαρμογή των Αλγορίθμων σε Podcasts του European School Radio

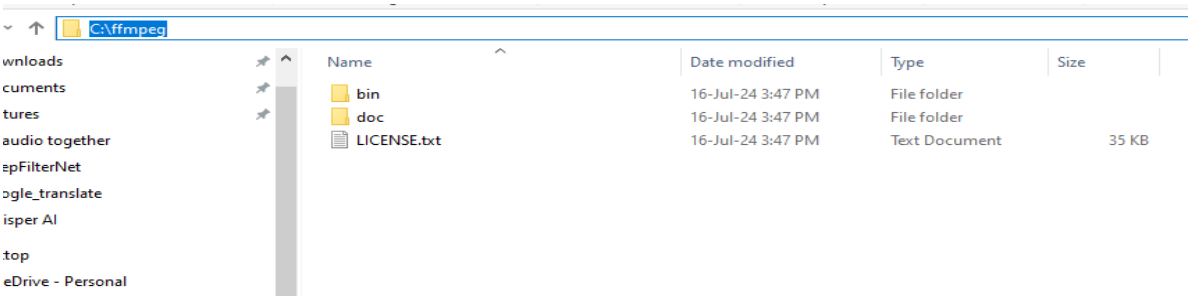
Σ αυτό το σημείο της εργασίας θα παρουσιαστεί η υλοποίηση των αλγορίθμων YamNet, DeepFilterNet, Whisper και bart σε κάποια ηχητικά podcasts του European School Radio. Οι αλγόριθμοι χρησιμοποιούν την Python 3.12.8 και ο καθένας ξεχωριστά έχει το δικό του περιβάλλον(venv). Οι συγκεκριμένοι κώδικες εφαρμόστηκαν σε λειτουργικο windows στο IDE VScode(Visual Studio Code). Για να υλοποιηθούν σε άλλα λογισμικά θα πρέπει να γίνουν κάποιες μικρές αλλαγές.

Για την εισαγωγή των μοντέλων YamNet, του DeepFilterNet και του Whisper τοπικά στον υπολογιστή πρέπει να γίνει η εισαγωγή της βιβλιοθήκης ffmpeg συγκεκριμένα το αρχείο (ffmpeg-git-full.7z). Αν εχουμε windows πρεπει να την δηλώσουμε στις μεταβλητές του υπολογιστή μας(environment variables) μετα την εγκατάστασή της.



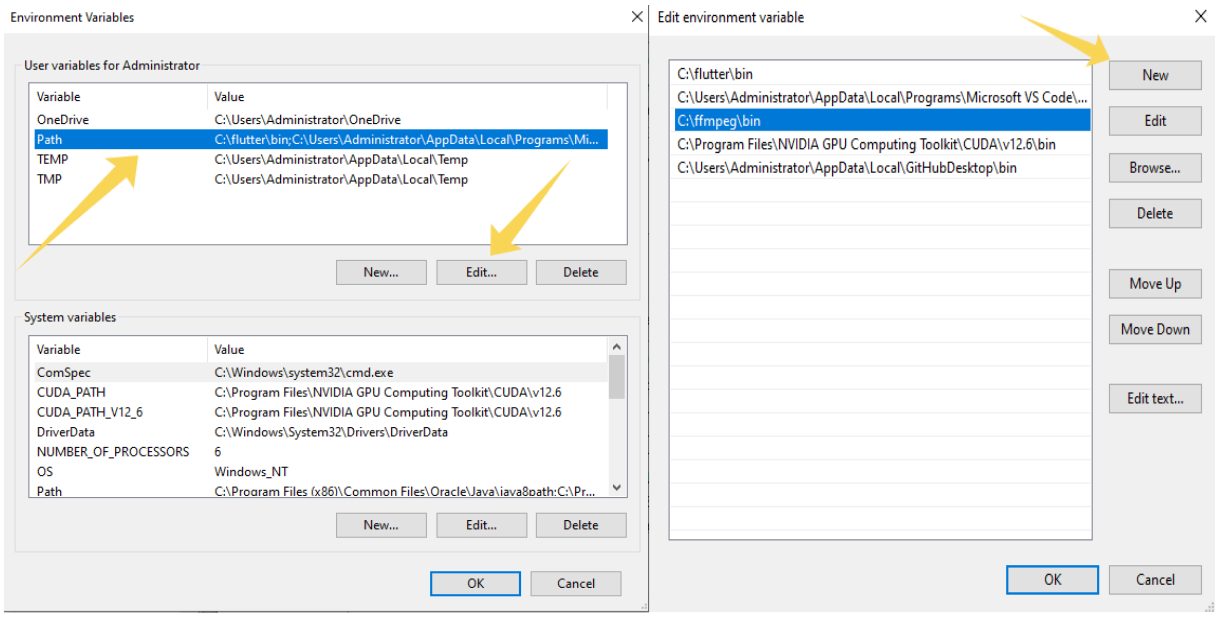
Εικόνα 3.1: Αλλαγή μεταβλητών περιβάλλοντος σε windows

Η εγκατάσταση γινεται κανοντας αποσυμπίεση τον φακελο του αρχειου (ffmpeg-git-full.7z) στον local disk C:/ του υπολογιστή μετανομάζοντας τον φάκελο που περιέχει τα αρχεία σε ffmpeg.



Εικόνα 3.2: Η διαδρομή και η δομή του φακέλου ffmpeg

Στην συνέχεια ανοίγοντας το παράθυρο των μεταβλητών πατόντας new τοποθετούμε την διαδρομή του φακέλου bin στο path.



Εικόνα 3.3: Επεξεργασία μεταβλητών περιβάλλοντος στα windows

Για να ελεγχθεί αν έγινε σωστά η εγκατάσταση ανοίγουμε τερματικό(CMD) και πληκτρολογόντας ffmpeg. Αν όλα έγιναν σωστά εμφανίζεται το παρακάτω αποτέλεσμα.

```
C:\Users\Administrator>ffmpeg
ffmpeg version N-116312-g635f7c0f6c-20240716 Copyright (c) 2000-2024 the FFmpeg developers
  built with gcc 14.1.0 (crosstool-NG 1.26.0.93_a87bf7f)
  configuration: --prefix=/ffbuild/prefix --pkg-config-flags=--static --pkg-config=pkg-config --cross-prefix=x86_64-w64-mingw32- --arch=x86_64 --target-os=mingw32
--enable-gpl --enable-version3 --disable-debug --disable-w32threads --enable-pthreads --enable-iconv --enable-zlib --enable-libfreetype --enable-libfribidi --enabl
e-gmp --enable-libxml2 --enable-fontconfig --enable-libharfbuzz --enable-libvorbis --enable-opencl --disable-libpulse --enable-libvmaf --disable-libxcb --disable-x
lib --enable-amf --enable-libaom --enable-libaribb24 --enable-avisynth --enable-chromaprint --enable-libdav1d --enable-libdavs2 --enable-libdvdread --enable-libdvd
nav --disable-libfdk-aac --enable-ffnvcodec --enable-cuda-llvm --enable-frei0r --enable-libgme --enable-libkvazaar --enable-libaribcaption --enable-libass --enable
-libbluray --enable-libjxl --enable-libmp3lame --enable-libopus --enable-librist --enable-libssh --enable-libtheora --enable-libvpx --enable-libwebp --enable-lv2 -
-enable-libvpl --enable-openal --enable-libopencore-amrnb --enable-libopencore-amrwb --enable-libopenh264 --enable-libopenjpeg --enable-libopenmpt --enable-librav1
e --enable-librubberband --enable-schannel --enable-sdl2 --enable-libsoxr --enable-libsrt --enable-libsvtav1 --enable-libtwolame --enable-libuavs3d --disable-libdr
m --enable-vaapi --enable-libvidstab --enable-vulkan --enable-libshaderc --enable-libplacebo --enable-libx264 --enable-libx265 --enable-libxavs2 --enable-libxvid -
-enable-libzimg --enable-libzvbi --extra-cflags=-DLIBTWOLAME_STATIC --extra-cxxflags= --extra-libs=-lgomp --extra-ldflags=-pthread --extra-ldexeflags= --cc=x86_64-
w64-mingw32-gcc --cxx=x86_64-w64-mingw32-g++ --ar=x86_64-w64-mingw32-gcc-ar --ranlib=x86_64-w64-mingw32-gcc-ranlib --nm=x86_64-w64-mingw32-gcc-nm --extra-version=2
0240716
  libavutil      59. 28.100 / 59. 28.100
  libavcodec     61. 10.100 / 61. 10.100
  libavformat    61.  5.101 / 61.  5.101
  libavdevice    61.  2.100 / 61.  2.100
  libavfilter    10.  2.102 / 10.  2.102
  libswscale      8.  2.100 /  8.  2.100
  libswresample   5.  2.100 /  5.  2.100
  libpostproc    58.  2.100 / 58.  2.100
Universal media converter
usage: ffmpeg [options] [[infile options] -i infile]... {[outfile options] outfile}...

Use -h to get full help or, even better, run 'man ffmpeg'
```

Εικόνα 3.4: Επιτυχής εγκατάσταση του ffmpeg απο το τερματικό των windows

3.1 Εφαρμογή του Αλγορίθμου YamNet

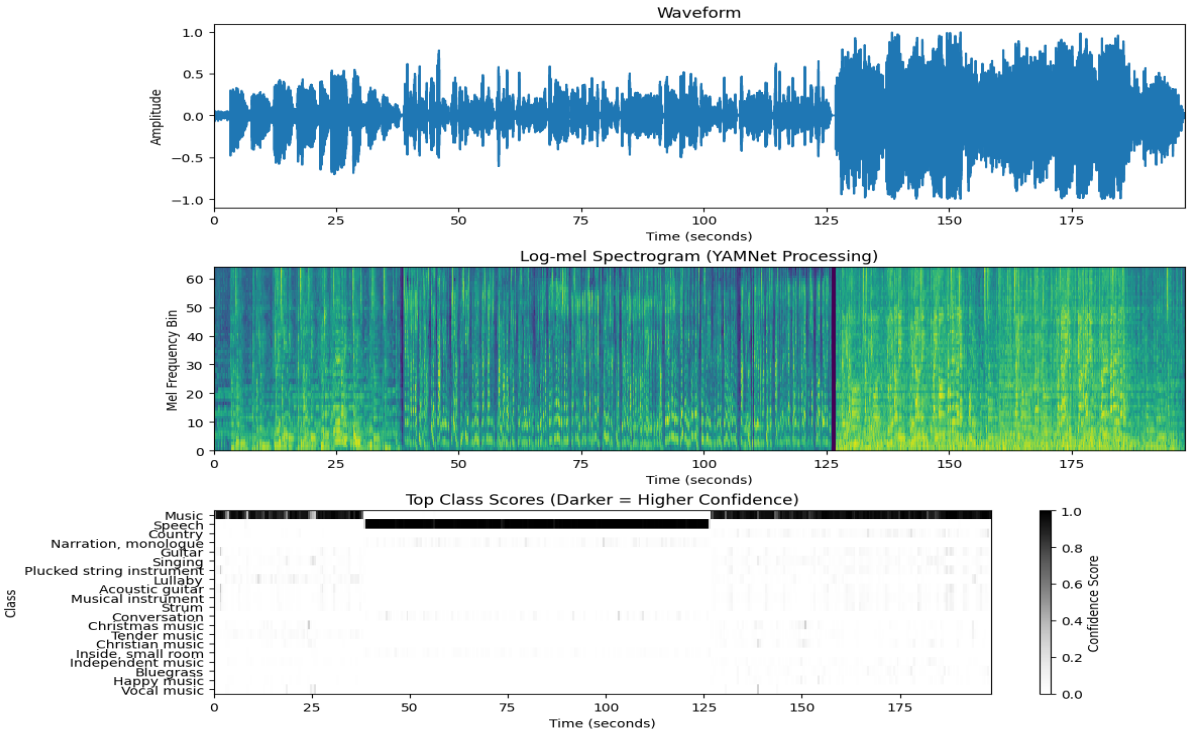
Εφόσον έχει γίνει η εγκατάσταση του ffmpeg δημιουργούμε το περιβάλλον(venv) στο οποίο θα εγκατασταθούν οι υπόλοιπες βιβλιοθήκες που χρειάζεται το μοντέλο ανοίγοντας το τερματικό(terminal) που υπάρχει και πληκτρολογόντας την παρακάτω εντολή θα ξεκινησει η λήψη και εισαγωγή των απαραίτητων βιβλιοθηκών στο τοπικό μας περιβάλλον.

pip install -r requirements.txt

Μετά απο αυτές τις διαδικασίες το μοντέλο είναι έτοιμο να τρέξει και να υλοποιήσει την ανάλυση των ηχητικών. Επιλέχθηκαν πέντε ηχητικά για τις ανάγκες της παρουσίασης της εφαρμογής αυτης.

Στα παρακάτω αποτελέσματα της εφαρμογής του Yamnet μπορούμε να δούμε τρία γραφήματα στα οποία μπορούμε να δούμε την κυματομορφή, τις συχνότητες(mel Frequency) και τα ηχητικά γεγονότα που συνέβησαν μέσα στα συγκεκριμένα. Η κυματομορφή προσφέρει πληροφορίες όπως η ένταση του ηχητικού σήματος, προσδιορίζοντας τα πιθανα ηχητικά χαρακτηριστικά με βάση τις διακυμάνσεις που πιθανόν να υπάρχουν μεταξύ διαφορετικών ήχων. Οι mel frequencies στο δεύτερο διάγραμμα μπορούν να αποτυπώσουν τις διάφορες συχνότητες που περιέχονται στα ηχητικά, δίνοντας έτσι ακόμα μια πληροφορία για να προβλέψουμε το ηχητικό γεγονός που είναι πιθανό να υπάρχει κάθε στιγμή. Στο τρίτο διάγραμμα φαίνεται η πρόβλεψη των ηχητικών γεγονότων που έχει κάνει το μοντέλο, δείχνοντας με σκούρο γκρι χρώμα τις κλάσεις που έχουν την μεγαλύτερη πιθανότητα και μεγαλύτερη διάρκεια ενω με απαλο χρωμα τις κλάσεις εκείνες που έχουν μικρή πιθανότητα.

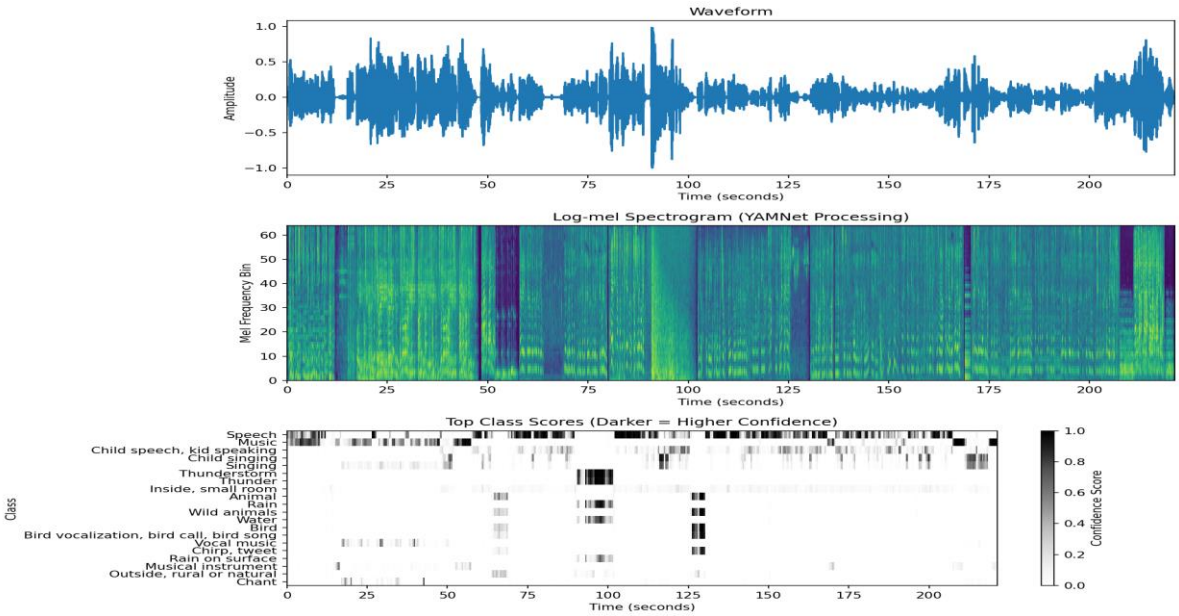
Audio 1



Εικόνα 3.5: Αποτελέσματα του μοντέλου YamNet για το Audio 1

Για το audio 1 παρατηρούμε απο την κυματομορφή του οτι στα 125 δευτερόλεπτα υπάρχει αύξηση της έντασης κατι το οποίο υποδηλώνει την ύπαρξη μουσικής ή κάποιου θορύβου. Το μοντέλο όπως βλέπουμε έχει ανιχνεύσει την μουσική και την ομιλία ως τις πιο πιθανές κλάσεις στο συγκεκριμένο ηχητικό, ωστόσο προβλέπει και κάποιες άλλες με μικρότερη πιθανότητα οι οποίες εμφανίζονται για μικρά χρονικά διαστήματα. Όπως φαίνεται στις κλάσεις μας δίνονται και καποιες πληροφορίες που αφορούν το ειδος της μουσικης και τα οργανα που μπορει να περιέχονται σ αυτην. Το συμπέρασμα για το συγκεκριμένο ηχητικό προβλέπεται να είναι podcast στο οποιο υπάρχει μουσική στην αρχή ομιλία στη μέση και στο τέλος ξανά μουσική.

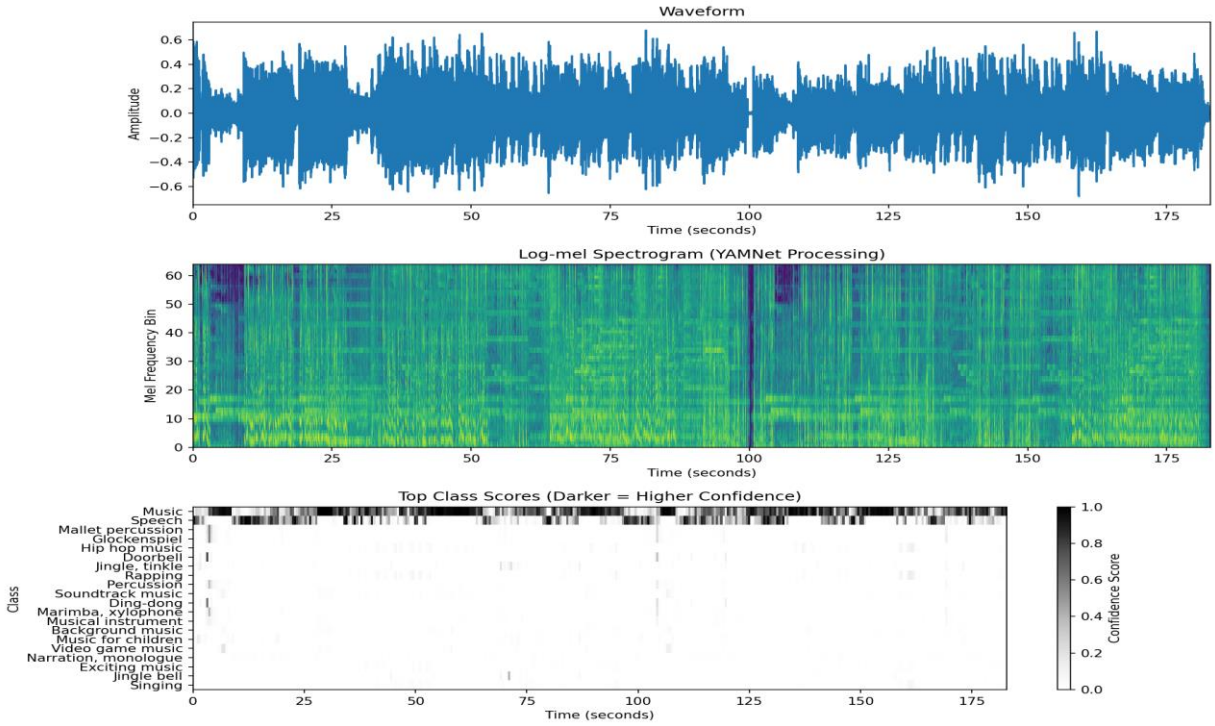
Audio 2



Εικόνα 3.6: Αποτελέσματα του μοντέλου YamNet για το Audio 2

Για το Audio 2, όπως φαίνεται από την ανάλυση, το μοντέλο YamNet έχει ανιχνεύσει την ομιλία και τη μουσική ως τις κύριες κλάσεις, με κάποιες άλλες κλάσεις να εμφανίζονται για μικρά χρονικά διαστήματα οπως το Thunder, Thunderstorm, Rain, Water, Bird, οι οποίες εξίσου έχουν μεγάλες πιθανότητες. Οι πληροφορίες που παρέχονται για τις κλάσεις περιλαμβάνουν πιθανά είδη μουσικής και όργανα που μπορεί να ακούγονται. Το συμπέρασμα για την πρόβλεψη του μοντέλου στο συγκεκριμένο ηχητικό ειναι οτι προκειται για podcasts στο οποίο υπάρχει μουσική και ομιλίες παιδιών που εναλλάσσονται καθώς όμως και εφε τα οποία εμφανίζονται σε μικρα χρονικά διαστήματα.

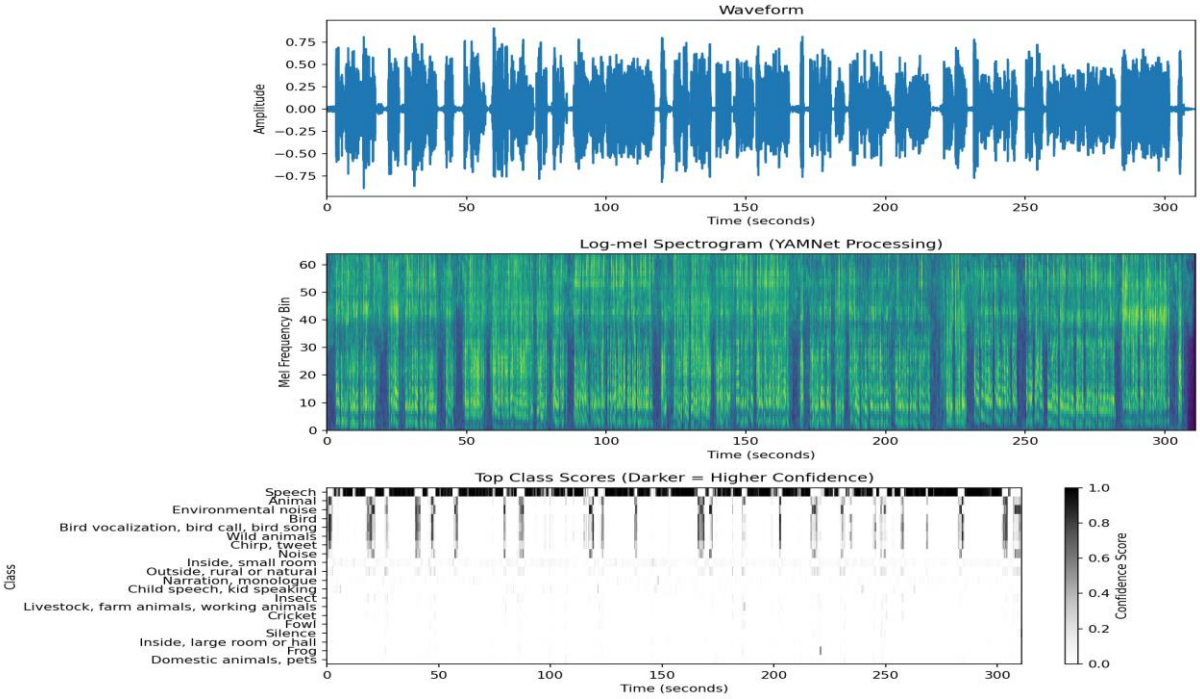
Audio 3



Εικόνα 3.7: Αποτελέσματα του μοντέλου YamNet για το Audio 3

Για το audio 3, παρατηρούμε από την κυματομορφή καθόλη την διάρκεια υπάρχουν μεγάλες εντάσεις κατι το οποίο εκφράζει την ύπαρξη μουσικής . Το log-mel spectrogram δείχνει έντονες και σκοτεινές περιοχές του φάσματος, υποδεικνύοντας την ένταση των διαφόρων συχνοτήτων με την πάροδο του χρόνου. Οι κορυφαίες κατηγορίες που ανιχνεύθηκαν από το μοντέλο περιλαμβάνουν μουσική και ομιλία ενω οι υπόλοιπες που εμφανίζονται φαίνονται να έχουν μικρές πιθανότητες. Το συγκεκριμένο ηχητικό πιθανόν να είναι επίσης ένα podcast όπου οι εναλλαγές μεταξύ μουσικής και ομιλίας είναι έντονες.

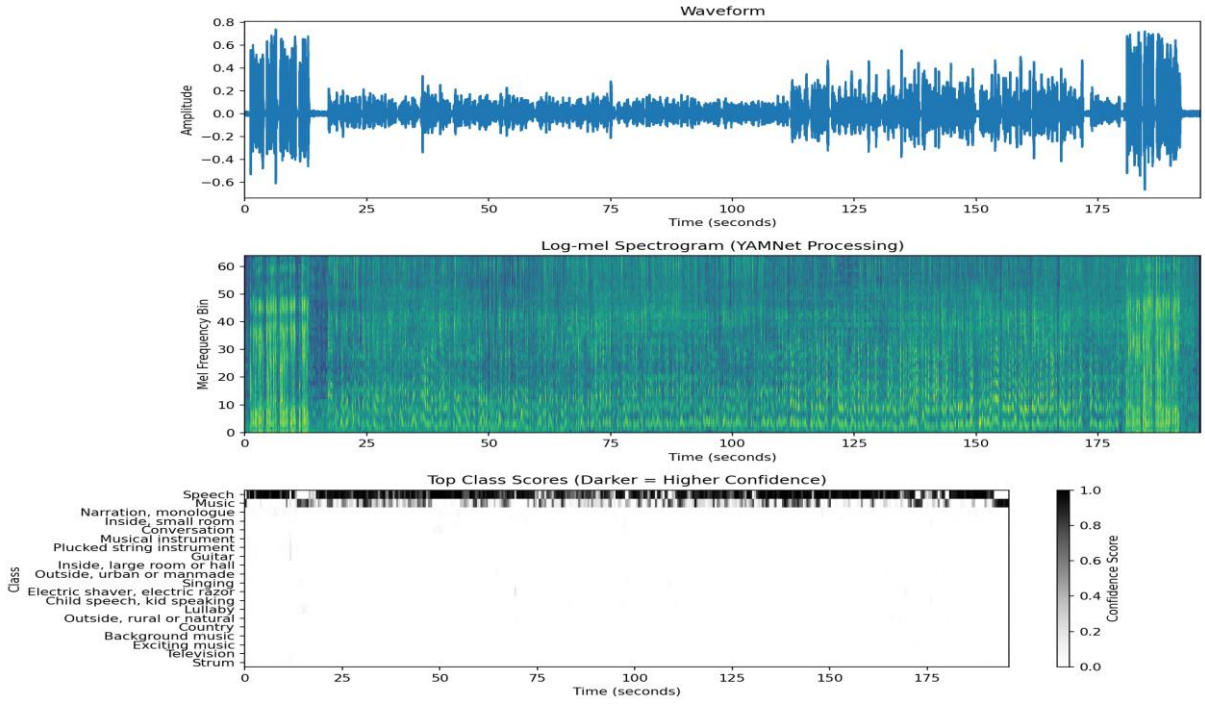
Audio 4



Εικόνα 3.8: Αποτελέσματα του μοντέλου YamNet για το Audio 4

Παρατηρούμε από την κυματομορφή του audio 4 ότι καθ' όλη τη διάρκεια υπάρχουν μεγάλες εντάσεις, κάτι που υποδηλώνει την ύπαρξη μουσικής ή θορύβου. Οι κορυφαίες κατηγορίες που ανιχνεύθηκαν σ αυτο το ηχητικό, περιλαμβάνουν ομιλία και θορύβους τις φύσεις συμφωνα με την ανάλυση του μοντέλου. Το συγκεκριμένο ηχητικό πιθανόν να είναι επίσης ένα podcast, όπου η ηχογράφηση του έγινε σε εξωτερικό χώρο ή έχουν προστεθεί ήχοι για να δημιουργήσουν μια τέτοια αισθητική ατμόσφαιρα.

Audio 5



Εικόνα 3.9: Αποτελέσματα του μοντέλου YamNet για το Audio 5

Παρατηρούμε ότι στο audio οτι υπάρχει μεγάλο confidence σε ορισμένες κλάσεις, όπως φαίνεται από τις σκοτεινότερες περιοχές στο διάγραμμα βαθμολογίας κλάσεων. Οι κύριες κατηγορίες που εντοπίστηκαν σε αυτό το ηχητικό περιλαμβάνουν ομιλία και μουσική με τις υπόλοιπες να έχουν ελάχιστες πιθανότητες.

3.2 Καθαρισμός του θορύβου με DeepFilterNet

Εκτος απο την εγκατάσταση του FFmpeg το DeepFilterNet χρειαζεται την εγκατάσταση του git(Git-2.47.1.2-64-bit) και της γλώσσας προγραμματισμού Rust(rustup-init). Πρέπει επίσης να γίνουν τα βήματα που έγιναν και με το ffmpeg, δηλαδή να δηλώσουμε τον φάκελο της cargo και της rust στις μεταβλητές όπως είχαμε δείξει στην αρχή του κεφαλαίου αυτού με την μόνη διαφορά οτι ο φάκελος αυτος βρίσκεται σ αυτά τα μονοπάτια.

C:\Users\<yourUsername>.cargo

C:\Users\<yourUsername>.rustup

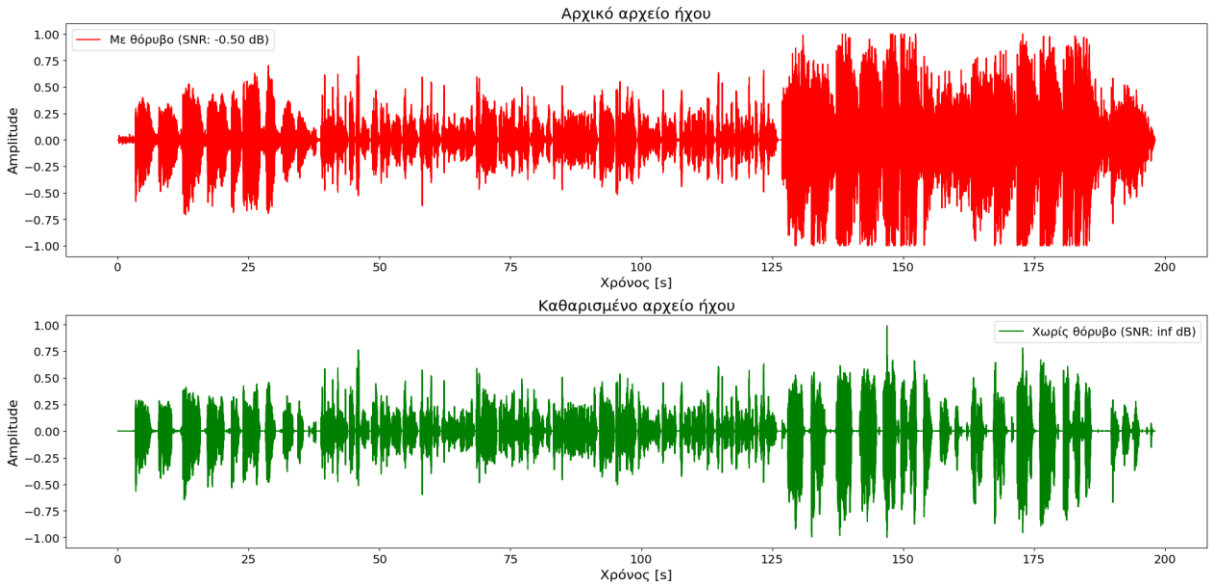
Το επόμενο βήμα είναι η δημιουργία ενός εικονικού περιβάλλοντος (venv) για την εγκατάσταση των απαραίτητων βιβλιοθηκών. Τρέχοντας παρακάτω τις εντολές και κανοντας cd στον φάκελο(DeepFilterNet/DeepFilterNet) μεσα απο το terminal του VSCode, μπορούμε να τρέξουμε τις παρακάτω εντολές.

pip install .

Εφόσον έγινε σωστα η εγκατάσταση το μοντέλο είναι έτοιμο για να λειτουργήσει πάνω στα ηχητικά τα οποία επιλέξουμε.

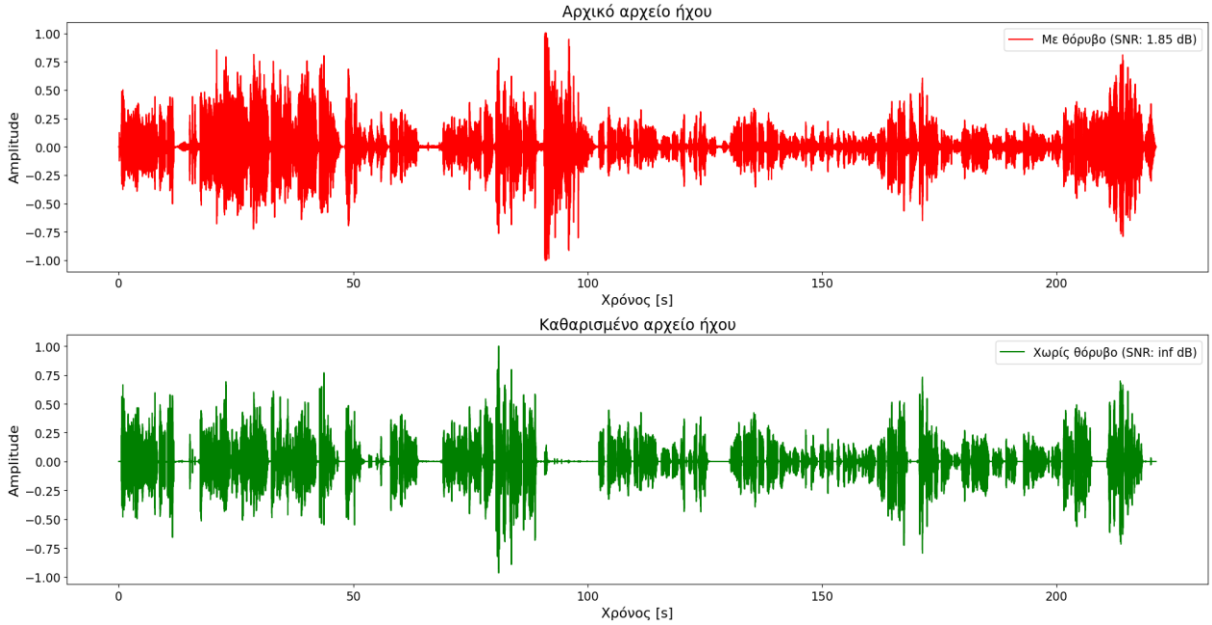
Απο την εφαρμογή του καθαρισμού προέκυψαν τα παρακάτω γραφήματα τα οποία δείχνουν πως ήταν το αρχικό αρχείο και πως έγινε μετά από την επεξεργασία. Το μοντέλο έχει ρυθμιστεί να ανιχνεύει την φωνή και να αποκόβει οποιοδήποτε άλλο ήχο έτσι ώστε στο τέλος της επεξεργασίας του ηχητικού θα υπάρχει μόνο η ομιλία. Ο δείκτης SNR(Signal to Noise Ratio) έχει προστεθεί στα γραφήματα για να ανιχνεύει το ποσοστό του θορύβου που υπάρχει στο αρχικό ηχητικό σε σύγκριση με το καθαρό σήμα που παράγεται μετά την επεξεργασία. Παρατηρουμε πως η ενταση του ηχου στα καθαρισμένα αρχεία είναι μικρότερη, ενω το SNR βρίσκεται στο infinity διότι το αρχείο πλέον δεν περιέχει θόρυβο. Στα ηχητικά τα οπόια το SNR εχει θετικές τιμές οπως το audio 2, audio 4 και audio 5, σημαίνει ότι ο θόρυβος ή μουσική είχαν πιο μεγάλη ενταση απο την την ομιλία.

Audio 1



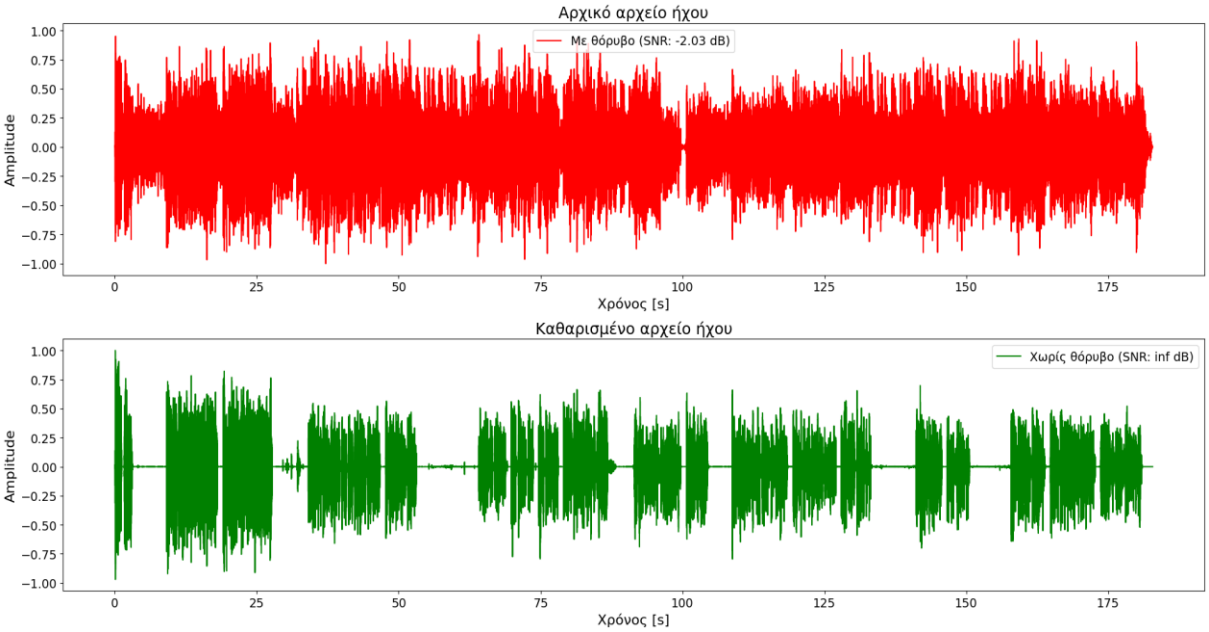
Εικόνα 3.10: Σύγκριση καθαρού και θορυβώδους σήματος για το Audio 1

Audio 2



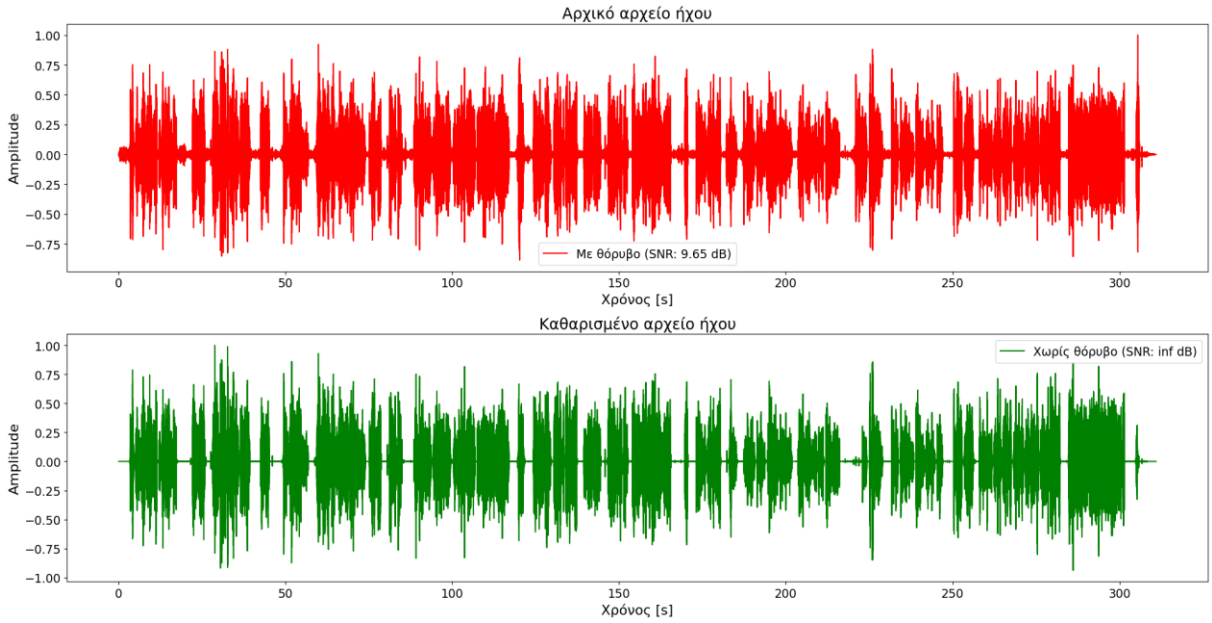
Εικόνα 3.11: Σύγκριση καθαρού και θορυβώδους σήματος για το Audio 2

Audio 3



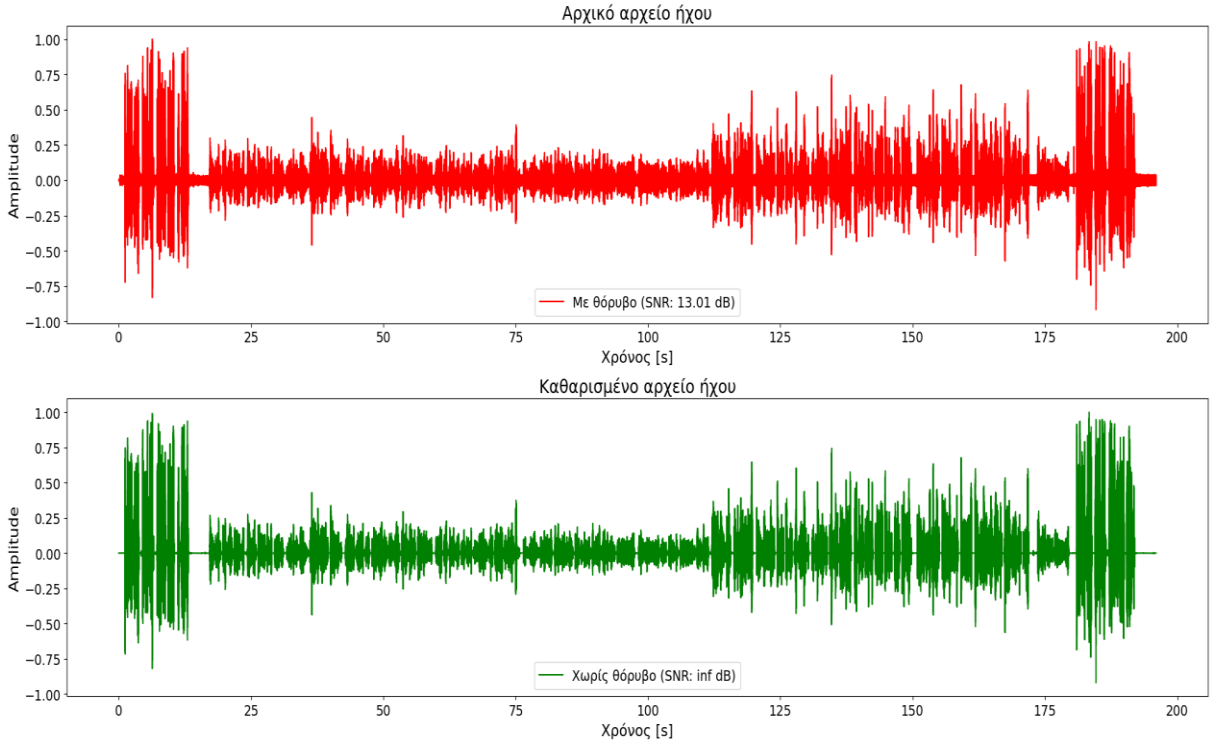
Εικόνα 3.12: Σύγκριση καθαρού και θορυβώδους σήματος για το Audio 3

Audio 4



Εικόνα 3.13: Σύγκριση καθαρού και θορυβώδους σήματος για το Audio 4

Audio 5



Εικόνα 3.15: Σύγκριση καθαρού και θορυβώδους σήματος για το Audio 5

3.3 Εφαρμογή του Αλγορίθμου Whisper

Το ffmpeg χρειάζεται για την εφαρμογή του μοντέλου Whisper. Αφου εγκατασταθεί, δημιουργούμε το περιβάλλον(venv) στο οποίο θα εγκατασταθούν και οι υπόλοιπες βιβλιοθήκες που χρειάζεται το μοντέλο ανοίγοντας το τερματικό(terminal) που υπάρχει και πληκτρολογόντας την παρακάτω εντολή.

pip install -r requirements.txt

Αν όλα πήγαν καλα με την εγκατάσταση του ffmpeg και των βιβλιοθηκών, το Whisper επιλέγοντας την διαδρομή του αρχείου είναι έτοιμο να μετατρέψει τον ήχο σε κείμενο αν βέβαια υπάρχει ομιλία σ αυτο.

Αρχικά τα ηχητικά δοκιμάστηκαν χωρίς την προσθήκη του DeepFilterNet και στη συνέχεια μετά την αφαίρεση του θορύβου. Η συγκεκριμένη έκδοση που χρησιμοποιήθηκε είναι η turbo.

Με βάση τα συγκεκριμένα αποτελέσματα παρατηρούμε οτι η πολυγλωσσική αναγνώριση ομιλίας του Whisper επιδεικνύει αξιοσημείωτη προσαρμοστικότητα, ικανή να επεξεργάζεται ηχητικό υλικό με δίγλωσσο περιεχόμενο (π.χ. Αγγλικά-Ελληνικά) με υψηλή συνοχή και ελάχιστη υποβάθμιση ακρίβειας, ακόμη και σε συνθήκες μεικτής γλωσσικής χρήσης. Επιδόσεις σε ηχητικά σήματα διαφορετικής ποιότητας (αρχικό vs. αποθορυβοποιημένο) δείχνουν σταθερή απόδοση, με τα Denoised audios να είναι καλύτερα ως αναφορά την ευκρίνεια μεταγραφής, ιδίως σε περιπτώσεις που υπάρχει αρκετός θόρυβος. Η χρονική σήμανση (timestamps) διατηρεί υψηλό βαθμό ακρίβειας και συνέπειας ανεξαρτήτως εισόδου, χαρακτηριστικό κρίσιμο για εφαρμογές υποτιτλισμού. Ωστόσο, παρατηρούνται συστηματικές αδυναμίες στην ορθογραφική απόδοση ελληνικών λέξεων, επαναλαμβανόμενες φράσεις σε τμήματα με ασάφειες, και ασυνέπειες στην απόδοση τεχνικών όρων, ζητήματα που υπογραμμίζουν την ανάγκη μετα-επεξεργασίας. Παρά τους περιορισμούς, το σύστημα προσφέρει αξιόπιστα αποτελέσματα για αυτοματοποιημένη δημιουργία υποτίτλων σε πολυπολιτισμικά πλαίσια.

Πιο συγκεκριμένα η συγκριτική ανάλυση των αρχικών (original) και αποθορυβοποιημένων (denoised) ηχητικών μεταγραφών αποκάλυψε διαφορές σε σχέση με την ακρίβεια, τη δομή και την ερμηνευτικότητα του περιεχομένου. Σε όλες τις περιπτώσεις, η διαδικασία αποθορυβοποίησης βελτίωσε τη σαφήνεια του ηχητικού σήματος, ιδιαίτερα σε αγγλόφωνα τμήματα (π.χ. τραγούδι "Sound of Silence"), όπου η ροή και η συνοχή του κειμένου ενισχύθηκαν. Ωστόσο, σε ελληνόφωνα περιεχόμενα, παρατηρήθηκαν αστοχίες στην αναγνώριση λέξεων με παρόμοια φωνητική δομή και παραλλαγές στη χρονική αντιστοίχιση διαλόγων, με συνέπεια την τροποποίηση της δομικής ακεραιότητας του κειμένου. Επιπλέον, σε ορισμένες μεταγραφές (π.χ., Audio 3), η διαδικασία οδήγησε σε σημασιολογικές αλλοιώσεις, όπως η παραποίηση ονομάτων χαρακτήρων ("Ζμή" αντί "Σμιθ") ή γραμματικά λάθη ("ναβαγούς" αντί "ναυαγούς"), υπογραμμίζοντας την ευαισθησία των αλγορίθμων σε γλωσσικές λεπτομέρειες.

Audio 1 (Original Sound)

[0.00s - 10.50s] Hello darkness my old friend, I've come to talk with you again
[10.50s - 19.94s] Because a vision softly creeping left its seeds while I was sleeping
[19.94s - 35.94s] And the vision that was planted in my brain still remains within the sound of silence
[35.94s - 37.94s] In rest
[37.94s - 40.94s] Συνέχεια μιλάμε για τις επανάστασεις του παρελθόντος
[40.94s - 43.94s] Στις μέρες μας δεν είχε νόημα η Επανάσταση
[43.94s - 44.94s] Πώς και δεν θα είχε
[44.94s - 47.94s] Η Επανάσταση είναι έννοια διαχρονική
[47.94s - 49.94s] Και τι σημαίνει η Επανάσταση
[49.94s - 53.94s] Η Επανάσταση είναι η δικαίωση που επιτυχάνεται μέσα από τον αγώνα
[53.94s - 56.94s] Η Επανάσταση είναι η ανάγκη για τη δικαιοσύνη
[56.94s - 60.94s] Η Επανάσταση είναι η προσπάθεια των ανθρώπων να ακουστεί η γνώμη τους
[60.94s - 64.94s] Επανάσταση είναι η αντίσταση στον άδικο θάνατο
[64.94s - 67.94s] τον κατακτητή, τον πλούσιο, τον άδικο
[67.94s - 71.94s] Επανάσταση είναι η ανάγκη για απελευθέρωση και η ανάγκη για αλλαγή
[71.94s - 73.94s] Επανάσταση είναι η αναγέννηση
[73.94s - 77.94s] Γιατί κατά τη διάρκεια της αφήνεις πίσω κάτι παλιό και πηγαίνεις σε κάτι νέο
[77.94s - 80.94s] Υπάρχουν λόγοι σήμερα να επαναστατούν οι άνθρωποι
[80.94s - 82.94s] Πώς δεν υπάρχουν
[82.94s - 84.94s] Οι άνθρωποι επαναστατούν για πολλούς λόγους
[84.94s - 88.94s] Αναμφίβολα, βασικότερους από αυτούς αποτεμούν η ελευθερία, η δικαιοσύνη
[88.94s - 90.94s] και η κατοχή των ανθρωπίνων δικαιωμάτων
[90.94s - 94.94s] Η ελευθερία είναι δικαίωμα κάθε ζωντανού οργανισμού
[94.94s - 98.94s] Γι' αυτό εξίζει κάποιος να επαναστατήσει και ο κόπος του δεν θα πάει ποτέ χαμένος
[98.94s - 103.94s] Η επανάσταση μπορεί να γίνει ώστε οι άνθρωποι να καταχειρώνουν τα ατομικά τους δικαιώματα
[103.94s - 106.94s] Οι άνθρωποι επαναστατούν όταν νιώθουν αδικημένοι
[106.94s - 109.94s] Η επανάσταση γίνεται για πολλούς λόγους
[109.94s - 113.94s] Κλιματική αλλαγή, παιδική εργασία, ανισότητα, πολλοί λόγοι χωρίς απαντήσεις
[113.94s - 117.58s] Επανάσταση για τον κόσμο που καταστρέφεται από την κλιματική αλλαγή
[117.58s - 122.22s] Την ώρα που άλλοι ρίχνουν βόμβες, εκμεταλλεύονται παιδιά, ανθρώπους, ζώα
[122.22s - 125.94s] Που ριπαίνουν τον κόσμο, ο οποίος δεν θα μας αντέξει για πολλήν
[125.94s - 155.92s] Υπότιτλοι AUTHORWAVE
[155.94s - 185.92s] Υπότιτλοι AUTHORWAVE
[185.94s - 195.94s] Υπότιτλοι AUTHORWAVE

Audio 1 (Denoised Sound)

[0.00s - 10.28s] Hello darkness my old friend I've come to talk with you again

[10.28s - 19.82s] Because a vision softly creeping Left its seeds while I was sleeping

[19.82s - 35.64s] And the vision that was planted in my brain Still remains within the sound of silence

[35.64s - 41.56s] Συνέχεια μιλάω για τις επανάστασεις του παρελθόντος

[41.56s - 43.90s] Τις μέρες μας δεν είχε νόημα η επανάσταση

[43.90s - 45.22s] Πώς και δεν θα είχε

[45.22s - 47.88s] Η επανάσταση είναι έννοια διαχρονική

[47.88s - 49.88s] Και τι σημαίνει η επανάσταση

[49.88s - 54.10s] Η επανάσταση είναι η δικαίωση που επιτυχάνεται μέσα από τον αγώνα

[54.10s - 57.24s] Η επανάσταση είναι η ανάγκη για τη δικαιοσύνη

[57.24s - 61.54s] Η επανάσταση είναι η προσπάθεια των ανθρώπων να ακουστεί η γνώμη τους

[61.54s - 65.18s] Επανάσταση είναι η αντίσταση στον άδικο θάνατο

[65.18s - 68.06s] Τον κατακτητή, τον πλούσιο, τον άδικο

[68.06s - 72.30s] Επανάσταση είναι η ανάγκη για απελευθέρωση και η ανάγκη για αλλαγή

[72.30s - 78.30s] Επανάσταση είναι η αναγέννηση γιατί κατά τη διάρκεια της αφήνεις πίσω κάτι παλιό και πηγαίνεις σε κάτι νέο

[78.96s - 81.64s] Υπάρχουν λόγοι σήμερα να επαναστατούν οι άνθρωποι

[81.64s - 82.72s] Πώς δεν υπάρχουν

[82.72s - 85.24s] Οι άνθρωποι επαναστατούν για πολλούς λόγους

[85.24s - 91.48s] Αναμφίβολα βασικότερους από αυτούς αποτεμούν η ελευθερία, η δικαιοσύνη και η κατοχή των ανθρωπίνων δικαιωμάτων

[91.48s - 94.76s] Η ελευθερία είναι δικαίωμα κάθε ζωντανού οργανισμού

[94.76s - 98.76s] Γι' αυτό εξίζει κάποιος να επαναστατήσει και ο κόπος του δεν θα πάει ποτέ χαμένος

[99.38s - 103.60s] Η επανάσταση μπορεί να γίνει ώστε οι άνθρωποι να κατοχυρώνουν τα ντομικά τους δικαιώματα

[103.60s - 106.76s] Οι άνθρωποι επαναστατούν όταν νιώθουν αδικημένοι

[106.76s - 109.50s] Η επανάσταση γίνεται για πολλούς λόγους

[109.50s - 112.16s] Κλιματική αλλαγή, κατατική εργασία, ανισότητα

[112.16s - 113.96s] Πολλοί λόγοι χωρίς απαντήσεις

[113.96s - 117.62s] Επανάσταση για τον κόσμο που καταστρέφεται από την κλιματική αλλαγή

[117.62s - 119.50s] Την ώρα που άλλοι έρχονται βόμβες

[119.50s - 122.24s] Εκμεταλλεύονται παιδιά, ανθρώπους, ζώα

[122.24s - 123.74s] Πουρυπαίνουν τον κόσμο

[123.74s - 125.78s] Ο οποίες δεν θα μας αντέξει για πολλοί

[125.78s - 130.70s] And in the naked light I saw

[130.70s - 134.94s] 10,000 people, maybe more

[134.94s - 140.04s] People talking without speaking

[140.04s - 144.58s] People hearing without listening

[144.58s - 148.70s] People writing songs

[148.70s - 152.52s] That voices never share

[152.52s - 160.76s] No one dare disturb the sound of silence

[160.76s - 165.28s] Fools that I do not know

[165.28s - 169.38s] Silence like a cancer grow

[169.38s - 174.80s] Hear my words that I might teach you

[174.80s - 179.18s] Take my arms that I might beat you

[179.18s - 183.94s] But my words like silence

[183.94s - 186.06s] Ringer up spell

[186.06s - 195.08s] Echo in the wells of silence

[195.08s - 197.48s] And

Audio 2 (Original Sound)

[0.00s - 8.18s] Το δέντρο που έδινε έγινε η αφορμή για τους μαθητές του 28ου υπηγωγείου ηλίου να δημιουργήσουν μια καινούργια ιστορία.

[8.70s - 11.72s] Λίγο διαφορετική, μα με μεγάλη σημασία.

[11.72s - 41.70s] Υπότιτλοι AUTHORWAVE

[41.72s - 71.70s] Υπότιτλοι AUTHORWAVE

[71.72s - 101.70s] Υπότιτλοι AUTHORWAVE

[101.72s - 109.12s] Αντίθετα, όταν είχε πολύ ήλιο, καθόταν οι ανθρώποι στη σκιά του.

[109.36s - 114.96s] Έδινε τους καλούς του δελνές για να φτιάξει ένα σπίτι.

[114.96s - 125.44s] Τα χελιδόνια πήγαιναν σε αυτό να βρουν νέες φωλιές.

[125.44s - 136.12s] Τα μπερπέγια και τα ζώα βρίσκανε το φαγητό τους στο δάσο.

[136.48s - 140.62s] Έδινε στις ακόρδες για να κοιμηθούν.

[140.62s - 147.34s] Όλοι οι νοχιάν, καλούμανοι και ύλεμοι.

[148.02s - 152.28s] Ακούδαντας, θα το δει, θα το δει από την κουκουδιά.

[152.28s - 161.44s] Το δάσος έδινε και μέρα στα ζώα, στα στεφαλόπω και στα φυτά για να ζήσουν.

[161.76s - 168.12s] Μαμισσιαμέρα, κακή κακή άθρωπη, έβαλα φωτιά το δάσος.

[168.12s - 174.60s] Και ξεκίνησε μια μεγάλη πυλκαγιά.

[175.36s - 179.18s] Τα έχουν τα ζώα και άσο και τα γλυτός.

[179.18s - 185.52s] Οι πυλκοδοσπές δεν έβησαν την φωτιά, αλλά ήταν όλοι καμένα.

[186.34s - 191.18s] Ένα φωλάκι, μάλιστα. Έφαγε κάτι καμένο καλό.

[192.06s - 196.54s] Οι άνθρωποι φοβήθη, κάνονται ο κόσμος τα καταστραφή,

[196.54s - 200.84s] γιατί το δάσος του σε δίνε ζώη.

[201.38s - 205.34s] Όμως σκέφτηκαν να φυτρέψουν μέρα από την αρχή,

[205.34s - 207.22s] για να φτιάξουν ότι είχε χαθεί.

[210.82s - 218.22s] Και ζήσαν τα φτυγκάλα και το δάσος καλύτερα.

Audio 2 (Denoised Sound)

[0.00s - 5.64s] Το δέντρο που έδινε έγινε η αφορμή για τους μαθητές του 28ου υπεογείου ηλίου

[5.64s - 11.72s] να δημιουργήσουν μια καινούργια ιστορία, λίγο διαφορετική, μα με μεγάλη σημασία.

[17.00s - 27.88s] Ο κοινίκλος στήδε με, στις ανέμι τυλιγμένοι,

[27.88s - 46.88s] Δω στις κλώτσο να γυρνήσει, παραμύθι να αρχινήσει.

[47.88s - 51.88s] Τόνομα σας που έδινε.

[57.88s - 63.88s] Μια φορά και έναν καιρό ήταν την δάσος που έδινε πράγματα σε όποιον το χρειαζόταν.

[68.56s - 72.00s] Δίνε ο ξηγόνος σε όποιον το ζητούσε.

[72.28s - 76.88s] Και όποιον δεν είχε φαγητό έδινε τους καρπούς του για να επιζείς.

[77.92s - 78.88s] Ήταν πολύ φιλικό.

[78.88s - 89.04s] Όταν έβλεσε και κάποιος και κάποιος περίσσε μέσα από αυτό, το δάσος ζήνε τον Ιό Μπρέρ.

[89.04s - 109.12s] Αντίθετα, όταν είχε πολύ ήλιο, καθόταν οι ανθρώποι στη σκιά του.

[109.12s - 114.92s] Έδινε τους καρμούς του δελού για να πτιάξει ένα σπίτι.

[115.60s - 125.40s] Τα χελιδόνια πήγαιναν σε αυτό να βρούνε νέες φορές.

[125.40s - 136.20s] Τα πυρβήλια και τα ζώα βρίσκανε το φάγητό τους στο δάσος.

[136.50s - 140.60s] Έδινε στις ακόρδες για να κοιμηθούν.

[141.04s - 147.30s] Όλοι γνωστιά, χαλούμανοι και ήλεμοι.

[147.88s - 152.28s] Ακούνταντας, θα το δει τις δύο από το κουλιά.

[152.28s - 161.38s] Το δάσος έδινε και μέρα στα ζώα, στο σπίτι όπου εσπίρφη τα γερμαζίσουν.

[161.80s - 168.10s] Μεμισμιαμέα, κακή κακή άνθρωποι, έβαλα φωτιά το δάσος.

[170.56s - 174.64s] Και ξεκίνησε μια μεγάλη πυρκαγιά.

[175.36s - 177.74s] Τα έχουν τα ζώα και άνθρωποι.

[177.74s - 179.26s] Τα κοιτάζουν.

[179.26s - 185.50s] Οι πυρκοτράσβης πείες έβησαν την φωτιά, αλλά ήταν όλοι καμένα.

[186.38s - 188.14s] Ένα κλάξ, μάλιστα.

[189.22s - 191.24s] Έφαγε κάτι καμένο καλό.

[191.24s - 207.24s] Οι άνθρωποι φοβήθηκαν να φτιάξουν ότι είχε χαθεί.

[207.24s - 218.08s] Και έφυσα να φτιάξει καλά και το δάσος κάνει η πέρα.

[218.08s - 248.06s] Υπότιτλοι AUTHORWAVE

Audio 3 (Original Sound)

[0.00s - 1.14s] Το Μυστηριώδες Νησί

[1.14s - 3.08s] Συγγραφέας Ιούλιος Βέρν

[3.08s - 17.86s] Το Μυστηριώδες Νησίφ του Ιουλίου Βέρν είναι ένα κλασικό έργο της λογοτεχνίας επιστημονικής φαντασίας που δημοσιεύτηκε για πρώτη φορά το 1874.

[19.08s - 27.70s] Το βιβλίο αφηγείται την ιστορία πέντε ανδρών και ενός σκύλου που βρίσκονται ναυαγεί σε ένα κατοίκητο νησί στον Ειρηνικό Ωκεανό κατά τη διάρκεια του Αμερικανικού Εμφυλίου Πολέμου.

[30.00s - 46.70s] Η ιστορία ξεκινά με την απαγωγή των πέντε ανδρών, ο μηχανικός κύρος Σμιθ, ο δημοσιογράφος Γκίδεων Σπίλετ, ο ναυτικός Μπον, ο υπηρέτης Ναμ, ο νεαρός Χάρμπερτ Μπράουν και ο σκύλος Τοφ, με ένα αερόστατο που παρασύρεται από καταιγίδα.

[47.62s - 53.04s] Το αερόστατο συντρίβεται και οι άνδρες βρίσκονται ναυαγεί σε ένα άγνωστο νησί, το οποίο ονομάζουν νησί Λίγκων.

[60.00s - 68.86s] Οι ναυαγοί, με τη βοήθεια των γνώσεων του κύρου Σμιθ, αρχίζουν να φτίζουν μια νέα ζωή στο νησί.

[69.82s - 73.68s] Ανακαλύπτουν νερό, φτιάχνουν εργαλεία και καταφήγια και καλλιεργούν τη γη.

[74.64s - 78.02s] Με την πάροδο του χρόνου, ανακαλύπτουν ότι το νησί κρύβει πολλά μυστικά.

[79.02s - 86.82s] Ανεξήγητα γεγονότα συμβαίνουν, όπως η εμφάνιση προμηθειών και η διάσωση των ναυαγών από διάφορους κινδύνους, που

τους κάνουν να πιστεύουν ότι δεν είναι μόνοι.

[86.82s - 99.66s] Στην πορεία, οι ναυαγοί ανακαλύπτουν έναν ναυαγισμένο πειρατικό θησαυρό και συναντούν τον Αιρτόν, έναν πρώην πειρατή που είχε εγκαταληφθεί στο νησί από τους συντρόφους του.

[100.66s - 104.42s] Ο Αιρτόν εντάσσεται στην ομάδα και τους βοηθά στις προσπάθειές τους για επιβίωση.

[104.42s - 118.46s] Το αποκορύφωμα της ιστορίας έρχεται όταν ανακαλύπτουν ότι ο μυστηριώδης προστάτης τους είναι ο κάπτεν Νέμο, ο οποίος ζει στο υποβρύχιο του, το ναυτίλιος, το οποίο είχε προσάραξη στη σπηλιά του νησιού.

[118.46s - 127.00s] Ο Νέμο αποκαλύπτει την ταυτότητά τους στους ναυαγούς και τους εξηγεί ότι είχε καταφύγει στο νησί μετά από την απογοήτευση του με τον κόσμο και τις πολιτικές του φιλοδοξίες.

[127.94s - 133.26s] Ο Νέμο είναι βαριά άρρωστος και λίγο αργότερα πεθαίνει, αφήνοντας τους ναυαγούς να φάψουν το σώμα τους στη θάλασσα.

[140.72s - 145.74s] Οι ναυαγοί, αφού περνούν δύο χρόνια στο νησί, τελικά σώζονται από ένα διερχόμενο πλοίο που τους εντοπίζει.

[145.74s - 150.56s] Το πλοίο τους επιστρέφει στον πολιτισμό, κλείνοντας έτσι την περιπέτειά τους με ευτυχισμένο τέλος.

[157.34s - 163.80s] Το Μυστηριώδες Νησί είναι ένα μυθιστόρημα που συνδυάζει την περιπέτεια, την επιστήμη και την ανθρώπινη επινοητικότητα.

[164.80s - 172.64s] Μέσα από τις περιγραφές του Βέρν, οι αναγνώστες μπορούν να θαυμάσουν την ικανότητα του ανθρώπου να προσαρμόζεται και να επιβιώνει ακόμα και στις πιο δύσκολες συνθήκες.

[172.64s - 181.02s] Η ιστορία αναδεικνύει την αξία της συνεργασίας, της φιλίας και της αλληλεγγύης και υπογραμμίζει τη σημασία της γνώσης και της εφευρετικότητας.

Audio 3 (Denoised Sound)

[0.00s - 1.14s] Το Μυστηριώδες Νησί

[1.14s - 3.10s] Συγγραφέας Ιούλιος Βέρν

[3.10s - 17.86s] Το Μυστηριώδες Νησίφ του Ιουλίου Βέρν είναι ένα κλασικό έργο της λογοτεχνίας επιστημονικής φαντασίας που δημοσιεύτηκε για πρώτη φορά το 1874.

[19.10s - 27.70s] Το βιβλίο αφηγείται την ιστορία πέντε ανδρών και ενός σκύλου που βρίσκονται ναυαγεί σε ένα ακατοίκητο νησί στον Ειρηνικό ωκεανό κατά τη διάρκεια του Αμερικανικού Εμφυλίου Πολέμου.

[30.00s - 59.98s] Υπότιτλοι AUTHORWAVE

[60.00s - 68.86s] Οι αβαγοί, με τη βοήθεια των γνώσεων του κύρου Ζμή, αρχίζουν να φτίζουν μια νέα ζωή στον Εσί.

[69.82s - 73.72s] Ανακαλύπτουν νερό, φτιάχνουν εργαλεία και καταφύγια και καλλιεργούν τη γη.

[74.62s - 78.04s] Με την πάροδο του χρόνου, ανακαλύπτουν ότι τον Εσί κρύβει πολλά μυστικά.

[78.04s - 86.86s] Ανεξήγητα γεγονότα συμβαίνουν, όπως η εμφάνιση προμητιών και η διάσωση των αγαγών από διάφορους κινδύνους, που τους κάνουν να πιστεύουν ότι δεν είναι μόνοι.

[86.86s - 99.66s] Στην πορεία, οι ναβαγοί ανακαλύπτουν έναν αναβαγισμένο πειρατικό θεσσαυρό και συναντούν των Ερτών, έναν πρώην πειρατή που είχε εγκαταλειφθεί στο νησί από τους συντρόφους του.

[100.68s - 103.86s] Ο Ερτών εντάσσεται στην ομάδα και τους βοηθά στις προσπάθειές τους για επιβίωση.

[103.86s - 118.48s] Το αποκορύφωμα της ιστορίας έρχεται όταν ανακαλύπτουν ότι ο μυστηριώδης προστάτης τους είναι ο κάπτεν Νέμο, ο οποίος ζει στο υποβρύχιο του,

το ναυτίλος, το οποίο είχε προσάραξη στη σπηλιά του νησιού.

[119.14s - 126.48s] Ο Νέμο αποκαλύπτει την αυτοτητά τους στους ναβαγούς και τους εξηγεί ότι είχε καταφύγει στο νοσί μετά από την απογοήτευση του με τον κόσμο και τις πολιτικές του φιλοδοξίες.

[126.48s - 133.24s] Ο Νέμο είναι βαριά άρρωστος και λίγο αργότερα πεθαίνει, αφήνοντας τους ναβαγούς να φάσουν το σώμα τους στη θάνασσα.

[140.70s - 145.76s] Οι ναβαγοί, αφού περνούν δύο χρόνια στο νησί, τελικά σώζονται από ένα διερπόμενο πλοίο που τους εντοπίζει.

[146.62s - 150.56s] Το πλοίο τους επιστρέφει στον πολιτικό, λύνοντας έτσι την περιπέτειά τους με εκτυχισμένο τέλος.

[156.48s - 163.80s] Το Μυστηριώδες Νησί είναι ένα μυτιστόρημα που συνδυάζει την περιφέτεια, την επιστήμη και την ανθρώπινη επινοητικότητα.

[164.80s - 172.66s] Μέσα από τις περιγραφές του Βέρν, οι αναγνώστες μπορούν να θαυμάσουν την ικανότητα του ανθρώπου να προσαρμόζεται και να επιβιώνει ακόμα και

στις πιο δύσκολες συνθήκες.

[173.16s - 180.66s] Η ιστορία αναδεικνύει την αξία της συνεργασίας, της φιλίας και της αλληλεγγύης και υπογραμμίζει τη σημασία της γνώσης και της εφευρετικότητας.

Audio 4 (Original Sound)

[0.00s - 9.94s] Hello everybody, you are listening to the EU Econius podcast for people who love learning

[9.94s - 17.38s] about biodiversity. Today we are talking about biodiversity loss in Europe and all around the world.

[21.76s - 25.36s] So, to begin with, what do you know about biodiversity?

[25.36s - 39.18s] Biodiversity is a name we give to the variety of all life on Earth, but they add to bubbles, plants to people, so the range of life on our planet is incredible.

[42.36s - 45.04s] And why is biodiversity loss a big problem?

[45.04s - 55.04s] Biodiversity is essential for the process that support our life on Earth, including humans.

[55.04s - 73.04s] Without a wide range of animals, plants and microorganisms, we cannot have the healthy ecosystems that we rely on to provide us with the air we breathe and the food we eat.

[73.04s - 78.04s] And people also value nature for itself.

[80.04s - 84.04s] What about marine biodiversity? Why is it so important?

[84.04s - 85.04s] What about marine biodiversity?

[85.04s - 86.04s] What about marine biodiversity?

[86.04s - 87.04s] What about marine biodiversity?

[87.04s - 88.04s] What about marine biodiversity?
[88.04s - 89.04s] What about marine biodiversity?
[89.04s - 90.04s] What about marine biodiversity?
[90.04s - 91.04s] What about marine biodiversity?
[91.04s - 92.04s] What about marine biodiversity?
[92.04s - 93.04s] What about marine biodiversity?
[93.04s - 94.04s] What about marine biodiversity?
[94.04s - 95.04s] What about marine biodiversity?
[95.04s - 96.04s] What about marine biodiversity?
[96.04s - 97.04s] What about marine biodiversity?
[97.04s - 98.04s] What about marine biodiversity?
[98.04s - 99.04s] What about marine biodiversity?
[99.04s - 100.04s] What about marine biodiversity?
[100.04s - 101.04s] What about marine biodiversity?
[101.04s - 102.04s] What about marine biodiversity?
[102.04s - 103.04s] What about marine biodiversity?
[103.04s - 104.04s] What about marine biodiversity?
[104.04s - 105.04s] What about marine biodiversity?
[105.04s - 106.04s] What about marine biodiversity?
[106.04s - 107.04s] What about marine biodiversity?
[107.04s - 108.04s] What about marine biodiversity?
[108.04s - 109.04s] What about marine biodiversity?
[109.04s - 110.04s] What about marine biodiversity?
[110.04s - 111.04s] What about marine biodiversity?
[111.04s - 112.04s] What about marine biodiversity?
[112.04s - 113.04s] What about marine biodiversity?
[113.04s - 114.04s] What about marine biodiversity?
[114.04s - 115.04s] What about marine biodiversity?
[115.04s - 116.04s] What about marine biodiversity?
[116.04s - 117.04s] What about marine biodiversity?
[117.04s - 118.04s] What about marine biodiversity?
[118.04s - 119.04s] What about marine biodiversity?
[119.04s - 120.04s] What about marine biodiversity?
[120.04s - 121.04s] What about marine biodiversity?
[121.04s - 122.04s] What about marine biodiversity?
[122.04s - 123.04s] What about marine biodiversity?
[123.04s - 124.04s] What about marine biodiversity?
[124.04s - 125.04s] What about marine biodiversity?
[125.04s - 126.04s] What about marine biodiversity?
[126.04s - 127.04s] What about marine biodiversity?
[127.04s - 128.04s] What about marine biodiversity?
[128.04s - 129.04s] What about marine biodiversity?

[129.04s - 130.04s] What about marine biodiversity?
[130.04s - 131.04s] What about marine biodiversity?
[131.04s - 132.04s] What about marine biodiversity?
[132.04s - 133.04s] What about marine biodiversity?
[133.04s - 134.04s] What about marine biodiversity?
[134.04s - 135.04s] What about marine biodiversity?
[135.04s - 136.04s] What about marine biodiversity?
[136.04s - 137.04s] What about marine biodiversity?
[137.04s - 138.04s] What about marine biodiversity?
[138.04s - 139.04s] What about marine biodiversity?
[139.04s - 140.04s] What about marine biodiversity?
[140.04s - 141.04s] What about marine biodiversity?
[141.04s - 142.04s] What about marine biodiversity?
[142.04s - 143.04s] What about marine biodiversity?
[143.04s - 144.04s] What about marine biodiversity?
[144.04s - 145.04s] What about marine biodiversity?
[145.04s - 146.04s] What about marine biodiversity?
[146.04s - 147.04s] What about marine biodiversity?
[147.04s - 148.04s] What about marine biodiversity?
[148.04s - 149.04s] What about marine biodiversity?
[149.04s - 150.04s] What about marine biodiversity?
[150.04s - 151.04s] What about marine biodiversity?
[151.04s - 152.04s] What about marine biodiversity?
[152.04s - 153.04s] What about marine biodiversity?
[153.04s - 154.04s] What about marine biodiversity?
[154.04s - 155.04s] What about marine biodiversity?
[155.04s - 156.04s] What about marine biodiversity?
[156.04s - 157.04s] What about marine biodiversity?
[157.04s - 158.04s] What about marine biodiversity?
[158.04s - 159.04s] What about marine biodiversity?
[159.04s - 160.04s] What about marine biodiversity?
[160.04s - 161.04s] What about marine biodiversity?
[161.04s - 162.04s] What about marine biodiversity?
[162.04s - 163.04s] What about marine biodiversity?
[163.04s - 164.04s] What about marine biodiversity?
[164.04s - 165.04s] What about marine biodiversity?
[165.04s - 166.04s] What about marine biodiversity?
[166.04s - 167.04s] What about marine biodiversity?
[167.04s - 168.04s] What about marine biodiversity?
[168.04s - 169.04s] What about marine biodiversity?
[169.04s - 170.04s] What about marine biodiversity?
[170.04s - 171.04s] What about marine biodiversity?

[171.04s - 172.04s] What about marine biodiversity?

[172.04s - 173.04s] What about marine biodiversity?

[173.04s - 178.04s] Yes, and a recent report from European Parliament indicated that one million species

[178.04s - 181.04s] could be threatened with extinction.

[181.04s - 186.04s] And who is responsible for this big loss?

[186.04s - 192.04s] Human activities are taking a heavy toll on biodiversity, threatening the delicate balance

[192.04s - 194.04s] of life in the Earth.

[194.04s - 200.04s] Overfishing, pollution, habitat destruction and climate change are among the most pressing threats

[200.04s - 202.04s] facing ecosystems today.

[202.04s - 207.04s] We will also understand its importance if we know that the work of one important Europeans

[207.04s - 212.04s] is directly related to the natural environment.

[212.04s - 217.04s] Also, that pastures cover about 25% of the Earth's surface.

[221.04s - 224.04s] And what can we do to protect biodiversity?

[224.04s - 231.04s] Is there anything we can do to stop biodiversity loss?

[231.04s - 233.04s] Of course, to protect biodiversity.

[233.04s - 238.04s] We should produce food much more efficiently using less land and real-less waste.

[238.04s - 245.04s] We should also change how and more we urbanize and industrialize the landscape and how we produce energy,

[245.04s - 247.04s] according to royal society.

[247.04s - 248.04s] So, we must change our daily lives.

[248.04s - 249.04s] Yes, everyone must make an effort for biodiversity.

[249.04s - 262.04s] Also, we will tell the people around us and ways around us.

[262.04s - 268.04s] If we can get permission from our school, we will open a lesson and explain biodiversity.

[268.04s - 271.04s] Then we can share an informative video about it.

[271.04s - 276.04s] We need to tell as many people as possible about biodiversity loss.

[276.04s - 282.04s] And we must encourage the people around us to take action and support the conservation of biodiversity.

[284.04s - 290.04s] We hope this helped you to understand the biodiversity loss and how we can protect biodiversity.

[290.04s - 292.04s] That's all we have time for.

[292.04s - 295.04s] Thanks for turning into this eco-news podcast.

[295.04s - 300.04s] And remember that we can all do our part to reduce our impact on biodiversity.

[300.04s - 302.04s] Goodbye.

[302.04s - 305.04s] Goodbye.

[305.04s - 306.04s] Goodbye.

Audio 4 (Denoised Sound)

[0.00s - 9.92s] Hello everybody, you are listening to the EU Econews podcast for people who love learning

[9.92s - 16.60s] about biodiversity. Today, we are talking about biodiversity loss in Europe and all

[16.60s - 17.36s] around the world.

[21.80s - 25.38s] So, to begin with, what do you know about biodiversity?

[25.38s - 34.38s] Biodiversity is a name we give to the variety of all life on earth, but they add to bubbles,

[34.38s - 39.38s] plants to people, so the amount of life on our planet is incredible.

[42.38s - 45.38s] And why is biodiversity loss a big problem?

[45.38s - 55.38s] Biodiversity is essential for the process that support our life on earth, including humans.

[55.38s - 68.38s] Without a wide range of animals, plants and microorganisms, we cannot have the healthy ecosystems that

[68.38s - 76.38s] we rely on to provide us with the air we breathe and the food we eat. And people also value nature

[76.38s - 77.38s] for itself.

[77.38s - 83.38s] What about marine biodiversity? Why is it so important?

[83.38s - 85.38s] Why is it so important?

[88.38s - 94.38s] Merian biodiversity is essential for the health of our planet and the wall bank of human societies. Protecting

[94.38s - 100.38s] its species is crucial because each of them plays a significant role in our ecosystem. For example,

[100.38s - 106.38s] coral reefs, often called the rainforests of the sea, are among the most biodiverse habitats on earth. They

[106.38s - 113.38s] provide the home for a country species of fish, invertebrates and other marine organisms forming complex

[113.38s - 117.38s] ecosystems that rival to spawn in tropical rainforests.

[119.38s - 121.38s] Wow, that's amazing!

[124.38s - 133.38s] Yes, but even the smallest creatures, like plankton and krill, play essential roles in the marine food web,

[133.38s - 137.38s] supporting the entire ecosystem from the bottom up.

[139.38s - 144.38s] That's incredible! Talking about biodiversity, what is the scale of biodiversity loss?

[146.38s - 152.38s] Plant and animal species are disappearing at an even faster rate due to human activity.

[153.38s - 160.38s] Over half of Europe's endemic trees, including the host Cessna, Hiberthenia excelsa and the Sorbus,

[160.38s - 165.38s] the human species are at risk and about one-fifth of amphibians and reptiles are endangered.

[169.38s - 170.38s] That's a lot!

[173.38s - 179.38s] Yes, and a recent report from European Parliament indicated that one million species could be threatened with extinction.

[179.38s - 184.38s] Who is responsible for this big loss?

[185.38s - 193.38s] Human activities are taking a heavy toll on biodiversity, threatening the delicate balance of life in the air.

[194.38s - 201.38s] Overfishing, pollution, habitat destruction and climate change are among the most present facing ecosystems today.

[201.38s - 210.38s] We will also understand its importance if we know that the work of one in quarantine Europeans is directly related to the natural environment.

[211.38s - 215.38s] Also, that pastures cover about 25% of the air's surface.

[215.38s - 228.38s] What can we do to protect biodiversity? Is there anything we can do to stop biodiversity loss?

[228.38s - 237.38s] Of course, to protect biodiversity. We should produce food much more efficiently using less light and real-less waste.

[238.38s - 246.38s] We should also change how and where we urbanize and industrialize the landscape and how we produce energy, according to royal society.

[246.38s - 252.38s] So, we must change our daily lives.

[253.38s - 256.38s] Yes! Everyone must make an effort for biodiversity.

[257.38s - 261.38s] Also, we will tell the people around us and the ways around us.

[262.38s - 266.38s] If we can get permission from our school, we will open a lesson and explain biodiversity.

[267.38s - 270.38s] Then, we can share an informative video about it.

[270.38s - 275.38s] We need to tell as many people as possible about biodiversity loss.

[276.38s - 282.38s] And we must encourage the people around us to take action and support the conservation of biodiversity.

[283.38s - 289.38s] We hope this helped you to understand the biodiversity loss and how we can protect biodiversity.

[290.38s - 291.38s] That's all we have time for.

[291.38s - 294.38s] Thanks for tuning into this EcoNews Podcast.

[295.38s - 299.38s] And remember that we can all do our part to reduce our impact on biodiversity.

[300.38s - 301.38s] Goodbye!

[304.38s - 305.38s] Goodbye!

Audio 5 (Original Sound)

[0.00s - 4.00s] Ακούτε την εκπομπή «Η Φρουρή του Διαδικτύου»

[4.00s - 7.00s] Είμαστε από το Γυμνάσιο της Νέας Αρτάκης

[7.00s - 11.00s] και όσα μεταδίδουμε βρίσκονται στο εγχειρίδιο του μαθητή

[11.00s - 14.00s] του Safer Internet for Kids.

[17.00s - 18.00s] Γεια σας!

[18.00s - 21.00s] Σήμερα θα σας μιλήσουμε για ένα πρόβλημα σημερινή εποχή

[21.00s - 24.00s] για την εικόνα σώματος στο social media.

[24.00s - 26.00s] Σερφάρατε στα κοινωνικά δίκτυα,

[26.00s - 29.00s] έχετε πίποτε στον εαυτό σας, κάτι από τα παρακάτω.

[29.00s - 31.00s] Πολύ θα ήθελα να είχα αυτό το σώμα.

[31.00s - 34.00s] Μα τι τέλεια εμφάνιση που έχει αυτός αυτή,

[34.00s - 36.00s] εγώ ποτέ δεν θα γίνω έτσι.

[36.00s - 39.00s] Είναι τόσο όμορφος όμορφη, γι' αυτό είναι τόσο χαρούμενο.

[39.00s - 42.00s] Πρέπει να χάσω οπωσδήποτε κοιλά.

[42.00s - 45.00s] Η πυλοψηφία των φωτογραφιών που κυκλοφωνούν

[45.00s - 47.00s] στα κοινωνικά δίκτυα απεικονίζουν.

[47.00s - 50.00s] Έψεγάδια στα σώματα δίνοντας έτσι έμφαση

[50.00s - 52.00s] στο τέλειο σώμα και στην άψεγη εμφάνιση

[52.00s - 53.00s] με ποικίλης τρόπους.

[53.00s - 56.00s] Τα πρότυπα που προβάλλονται επιχεί το τέλειο δίνατο σώμα.

[56.00s - 59.00s] Διάφερει σίγουρα από το σώμα που έχουμε οι άνθρωποι στην πυλεψηφία μας.

[59.00s - 62.00s] Οι φωτογραφίες στα κοινωνικά δίκτυα πολλές φορές

[62.00s - 64.00s] είναι επεξεργασμένες με σκοπό να φαίνεται

[64.00s - 66.00s] απικονιζόμενος περισσότερον λειτουργιστικός.

[66.00s - 70.00s] Οι χρήστες πολύ συχνά ψάχνουν να βρουν την προσωπική τους αξία

[70.00s - 76.00s] στην ποσότητα των likes και στα θετικά σχόλια που δέχονται.

[76.00s - 82.00s] Οι σέλφοι συχνά δίνουν το μήνυμο ότι αποκλειστικά η ομορφιά προσδιορίζει την αξία μας.

[82.00s - 86.00s] Έτσι το σώμα μας γίνεται αντικείμενο μόνιμης παρατήρησης

[86.00s - 89.00s] τόσο από εμάς τους ίδιους όσο και από τους άλλους

[89.00s - 94.00s] χωρίς να αναλαμβάνονται υπόψη άλλα πολύ θετικά στοιχεία του χαρακτήρα μας

[94.00s - 98.00s] που συντελούν στο να είμαστε και να φαινόμαστε ακόμα πιο όμορφοι.

[98.00s - 104.00s] Η ίδια η φύση των κοινωνικών δικτύων οδηγεί στο να κάνουμε συγκρίσεις με τους άλλους χρήστες

[104.00s - 109.00s] κρίνοντας αυστηρά τον εαυτό μας ενώ συχνά νιώθουμε ότι δεν είμαστε τόσο χαρούμενοι,

[109.00s - 112.00s] όμορφοι και επιτυχημένοι όπως οι άλλοι.

[112.00s - 116.00s] Εδώ λοιπόν μερικές συμβουλές για θετική εικόνα σώματος.

[116.00s - 120.00s] Η προσωπική μας αξία δεν καθορίζεται από τον αριθμό των likes.

[120.00s - 124.00s] Ας είμαστε προσεκτικοί ποιον ακολουθούμε στα κοινωνικά μας δίκτυα.

[124.00s - 132.00s] Μπορούμε να θέλουμε να ακολουθήσουμε σελίδες ή διάσημους με συμβουλές ομορφιάς και οδηγίες για γυμναστική,

[132.00s - 134.00s] αλλά ασπρασέχουμε αυτές τις σελίδες.

[134.00s - 139.00s] Τα προβάλλουν υγιεί πρότυπα για να μας κάνουν να νιώθουμε καλά με τον εαυτό μας.

[139.00s - 143.00s] Δεν χρειάζεται να ανταποκρινόμαστε σε συγκεκριμένα πρότυπα ομορφιάς.

[143.00s - 150.00s] Π.χ. αδύνατο σώμα, ραμωμένη κοιλιακή, μικροσκοπικό τζιν, προκειμένου να είμαστε όμορφοι.

[150.00s - 155.00s] Η ψυχική μας και η σωματική μας υγεία είναι πάντα προτεραιότητα.

[155.00s - 159.00s] Ασχολούμαστε με δραστηριότητες στην πραγματική ζωή.

[159.00s - 161.00s] Το αδύνατο δεν είναι και το ιδανικό.

[161.00s - 167.00s] Το υγιές είναι το ιδανικό και έρχεται σε όλα τα σχήματα, χρώματα και μεγέθη.

[167.00s - 172.00s] Είναι ένας τρόπος ζωής που δίνει έμφαση στο πνεύμα, την ψυχή και στο σώμα.

[172.00s - 180.00s] Ο λόγος αυτός που παρουσιάστηκε δημιουργήθηκε από το saferinternet4kids.gr

[180.00s - 184.00s] Ακούσατε την εκπομπή «Η Φρουρή του Διαδικτύου».

[184.00s - 186.00s] Είμαστε από το Γυμνάσιο της Νέας Αρτάκης.

[186.00s - 192.00s] Και όσα μεταδώσαμε βρίσκονται στο εγχειρίδιο του μαθητή, του saferinternet4kids.

[192.00s - 196.00s] Υπότιτλοι AUTHORWAVE

Audio 5 (Denoised Sound)

[0.00s - 4.00s] Ακούτε την εκπομπή «Η φρουρή του διαδικτύου»
[4.00s - 7.00s] Είμαστε από το Γυμνάσιο της Νέας Αρτάκης
[7.00s - 11.00s] και όσα μεταδίδουμε βρίσκονται στο εγχειρίδο του μαθητή
[11.00s - 14.00s] του Safer Internet for Kids.
[17.00s - 19.00s] Γεια σας! Σήμερα θα σας μιλήσουμε
[19.00s - 21.00s] για ένα πρόβλημα σημερινή εποχή
[21.00s - 24.00s] για την εικόνα σώματος στο social media.
[24.00s - 26.00s] Σε εφάρατε στα κοινωνικά δίκτυα.
[26.00s - 29.00s] Έχετε πει ποτέ στον εαυτό σας κάτι από τα παραγκάτω.
[29.00s - 31.00s] Πουλήθα ήθελα να είχα φωτεσόμα.
[31.00s - 34.00s] Μα τι τέλεια εμφάνιση που έχει αυτός αυτή.
[34.00s - 36.00s] Εγώ ποτέ δεν θα γίνω έτσι.
[36.00s - 38.00s] Είναι τόσο όμορφος όμορφη.
[38.00s - 40.00s] Γι' αυτό είναι τόσο χαρούμενο.
[40.00s - 42.00s] Πρέπει να χάσω οπωσδήποτε κυλάβη.
[42.00s - 44.00s] Η πλειοψηφία τα φωτογραφιών
[44.00s - 46.00s] που κυκλοφωνεί στα κοινωνικά δίκτυα
[46.00s - 47.00s] απεικονίζουν.
[47.00s - 50.00s] Έξιγαν διεστασσώματα δίνοντας έτσι έμφαση
[50.00s - 52.00s] στο τέλειο σώμα και στην άψογη εμφάνιση
[52.00s - 53.00s] με ποικίλης τρόπους.
[53.00s - 55.00s] Τα πρότυπα που προβάλλονται επιχεί
[55.00s - 56.00s] το τέλειο δίνοντο σώμα.
[56.00s - 58.00s] Διαφέρει σίγουρα από το σώμα που έχουμε οι άνθρωποι
[58.00s - 59.00s] στην πλειοψηφία μας.
[59.00s - 61.00s] Οι φωτογραφίες τα κοινωνικά δίκτυα
[61.00s - 63.00s] πολλές φορές είναι επεξεργασμένες
[63.00s - 65.00s] με σκοπό να φαίνονται
[65.00s - 67.00s] απικονιζόμενος και ισοτερικιστικός.
[67.00s - 69.00s] Οι χρήστες πολύ συχνά
[69.00s - 71.00s] ψάφνουν να βρουν την προσωπική τους αξία
[71.00s - 73.00s] στην ποσότητα των likes
[73.00s - 76.00s] σταθερικά σχόλια που δέχονται.
[76.00s - 78.00s] Οι σέλθοι συχνά δίνουν το μήνυμο
[78.00s - 80.00s] ότι αφουκλιστικά η ομορφιά
[80.00s - 82.00s] προσδιορίζει την αξία μας.
[82.00s - 84.00s] Έτσι το σώμα μας γίνεται αντικείμενο
[84.00s - 86.00s] μόνιμης παρατήρησης

[86.00s - 88.00s] τόσο από εμάς τους ίδιους
[88.00s - 89.00s] όσο και από τους άλλους
[89.00s - 91.00s] χωρίς να αναλαμβάνονται υπόψη
[91.00s - 93.00s] άλλα πολύ θετικά στοιχεία
[93.00s - 95.00s] του φαρακτήρα μας που συντελούν
[95.00s - 97.00s] στο να είμαστε και να φαινόμαστε
[97.00s - 98.00s] ακόμα πιο όμορφοι.
[98.00s - 100.00s] Η ίδια η φύση των κοινωνικών δικτύων
[100.00s - 102.00s] οδηγεί στο να κάνουμε συγκρίσεις
[102.00s - 104.00s] με τους άλλους φρίστες
[104.00s - 106.00s] κρίνοντας αυστηρά τον εαυτό μας
[106.00s - 108.00s] ενώ συχνά νιώθουμε ότι δεν είμαστε
[108.00s - 110.00s] τόσο χαρούντοι, όμορφοι και επιτυχημένοι
[110.00s - 112.00s] όπως οι άλλοι.
[112.00s - 114.00s] Εδώ λοιπόν μερικές συμβουλές
[114.00s - 116.00s] διευθετική εικόνα σώματος.
[116.00s - 118.00s] Η προσωπική μας αξία
[118.00s - 120.00s] δεν καθορίζεται από τον αριθμό των likes.
[120.00s - 122.00s] Ας είμαστε προσεκτικοί
[122.00s - 124.00s] ποιον ακολουθούμε στα κοινωνικά μας δίκτυα.
[124.00s - 128.00s] Μπορούμε να θέλουμε να ακολουθήσουμε σελίδες
[128.00s - 130.00s] ή διάσημους με συμβουλές ομορφιάς και οδηγίες
[130.00s - 132.00s] για γυμναστική,
[132.00s - 134.00s] αλλά ασπρασέχουμε αυτές τις σελίδες
[134.00s - 136.00s] τα προβάλλουν υγιείοι πρότυπα
[136.00s - 138.00s] και να μας κάνουν να νιώθουμε καλά με τον εαυτό μας.
[138.00s - 140.00s] Δεν χρειάζεται να ανταποκρινόμαστε
[140.00s - 142.00s] σε συγκεκριμένα πρότυπα ομορφιάς.
[142.00s - 144.00s] Π.χ. αδύνατο σώμα,
[144.00s - 146.00s] ραμωμένη κοιλιακή,
[146.00s - 148.00s] μικροσκοπικό τζίν,
[148.00s - 150.00s] προκειμένου να είμαστε όμορφοι.
[150.00s - 154.00s] Η ψυχική μας και η σωματική μας υγεία
[154.00s - 156.00s] είναι πάντα προτεραιότητα.
[156.00s - 158.00s] Ασχολούμαστε με δραστηριότητες
[158.00s - 160.00s] στην πρασματική ζωή.
[160.00s - 162.00s] Το αδύνατο δεν είναι και το ιδανικό.
[162.00s - 164.00s] Το υγιές είναι το ιδανικό
[164.00s - 166.00s] και έρχεται σε όλα τα σχήματα,
[166.00s - 168.00s] χρώματα και μεγέθη.
[168.00s - 170.00s] Λόγος ζωής που δίνει έμφαση στο πνεύμα,
[170.00s - 172.00s] την ψυχή και στο σώμα.

[172.00s - 174.00s] Ο λόγος αυτός που παρουσιάστηκε
[174.00s - 176.00s] δημιουργήθηκε
[176.00s - 180.00s] από το safer internet for kids.gr.
[180.00s - 182.00s] Ακούσατε την εκπομπή
[182.00s - 184.00s] «Η Φρουρή του Διαδικτύου»
[184.00s - 186.00s] Είμαστε από το Γυμνάσιο της Νέας Αρτάκης
[186.00s - 188.00s] και όσα μεταδώσαμε βρίσκονται
[188.00s - 190.00s] στο εγχειρίδιο του μαθητή
[190.00s - 192.00s] του safer internet for kids.
[192.00s - 194.00s] Ευχαριστώ.

3.4 Εφαρμογή του BART για Ανάλυση Κειμένου που Προκύπτει από την Μετατροπή Ομιλίας

Το μόνο που χρειάζεται για να τρέξουμε την εφαρμογή του BART είναι να φτιάξουμε το περιβάλλον του φακέλου(venv) και να τρέξουμε την παρακάτω εντολή.

pip install -r requirements.txt

Μετα την ολοκλήρωση της εγκατάστασης είμαστε έτοιμοι να εφαρμόσουμε τον BART στις αναλύσεις των audio, δηλαδή τα κείμενα που προέκυψαν απο το μοντέλο Whisper.

Στη συγκεκριμένη ανάλυση χρησιμοποιήθηκε το μοντέλο Bart Large Mnli (Zero Shot Classification) με οποίο μπορούμε να βάλουμε τις δικές μας κλάσεις για ταξινόμηση. Για την συγκεκριμένη εφαμοργή χρησιμοποιήκαν οι κλάσεις του European School Radio. Αυτές οι κλάσεις είναι οι εξής, Environment & health, Art & culture, Music,Books, Literature & Poetry, Sports,Local news,School news, Social issues, Reuse of Educational Resources, Pedagogical Audio Material for Teachers, Science and Technology, European and International News, Digital Storytelling, Discussions Interviews Documentaries, Active Citizenship Issues, Theater & Sound Drama, Cinema,Human Relations, Poetry, Pedagogical issues, Fairy tales, Educational subjects, Interviews.

Πιο αναλυτικά, στο Audio 1 συνδυάζοντας το τραγούδι "Sound of Silence" με μια συζήτηση για επαναστάσεις, το BART αναγνώρισε τόσο τη μουσική διάσταση του τραγουδιού όσο και την κοινωνικοπολιτική του προοπτική. Επικεντρώθηκε σε λέξεις-κλειδιά όπως "επανάσταση" και "δικαιοσύνη", ταξινομώντας το περιεχόμενο σε Κοινωνικά Θέματα (0.7354) και Μουσική (0.6003), γεγονός που υπογραμμίζει τη συμβολή του στη διασύνδεση τέχνης και κοινωνικής κριτικής. Στο Audio 2, πρόκειται για ένα παραμύθι που διηγείται την ιστορία ενός δέντρου που βοηθά ζώα και ανθρώπους μετά από πυρκαγιά. Το μοντέλο εντόπισε την αφηγηματική δομή και την έμφαση σε συνεργασία και

περιβαλλοντική συνείδηση, ταξινομώντας το σε Παραμύθια (0.9465) και Εκπαιδευτικά Θέματα (0.8601), αναδεικνύοντας τη διδακτική του αξία. Στο Audio 3, η ανάλυση του βιβλίου "Το Μυστηριώδες Νησί" του Ιούλιου Βερν ανέδειξε θέματα επιστημονικής φαντασίας και ανθρώπινης προσαρμοστικότητας. Το BART κατέταξε το περιεχόμενο σε Εκπαιδευτικά Θέματα (0.8798) και Επιστήμη & Τεχνολογία (0.8306), υπογραμμίζοντας τη συμβολή του κλασικού έργου στη διερεύνηση τεχνολογικών και φιλοσοφικών ερωτημάτων. Στο Audio 4, σε ένα podcast για την απώλεια βιοποικιλότητας και την κλιματική αλλαγή, το μοντέλο επικεντρώθηκε σε επιστημονικές έννοιες και πρακτικές συστάσεις. Η ταξινόμηση σε Περιβάλλον & Υγεία (0.8249) και Ευρωπαϊκά & Διεθνή Νέα (0.8965) αντανακλά τον διττό του ρόλο: τόσο στην ενημέρωση όσο και στην προβολή παγκόσμιων προκλήσεων. Τέλος, στο Audio 5, μια εκπομπή για την επίδραση των κοινωνικών δικτύων στην αυτοεκτίμηση αναλύθηκε με έμφαση σε ψυχοκοινωνικά φαινόμενα και στρατηγικές ψυχικής υγείας. Το BART κατέταξε το θέμα σε Κοινωνικά Θέματα (0.9123) και Ανθρώπινες Σχέσεις (0.7079), καταγράφοντας τη σχέση μεταξύ τεχνολογίας και ψυχικής ευημερίας.

Audio 1

Predicted Classes:

- Social issues with score: 0.7354
- Music with score: 0.6003
- Art & culture with score: 0.5455
- Active Citizenship Issues with score: 0.4343
- Human Relations with score: 0.4099
- Environment & health with score: 0.3094
- Digital Storytelling with score: 0.3002

Audio 2

Predicted Classes:

- Fairy tales with score: 0.9465
- Educational subjects with score: 0.8601
- Reuse of Educational Resources with score: 0.8374
- Social issues with score: 0.7386
- Digital Storytelling with score: 0.6945
- Pedagogical issues with score: 0.6884

- Art & culture with score: 0.6497
- School news with score: 0.6209
- Music with score: 0.6160
- Human Relations with score: 0.6135
- Pedagogical Audio Material for Teachers with score: 0.5945
- Local news with score: 0.5883
- Theater & Sound Drama with score: 0.5683
- Literature & Poetry with score: 0.5137
- Active Citizenship Issues with score: 0.4302
- Environment & health with score: 0.4294
- Cinema with score: 0.3507
- Interviews with score: 0.3450
- Poetry with score: 0.3374

Audio 3

Predicted Classes:

- Educational subjects with score: 0.8798
- Human Relations with score: 0.8384
- Science and Technology with score: 0.8306
- Social issues with score: 0.8289
- Environment & health with score: 0.8052
- Digital Storytelling with score: 0.6614
- Books with score: 0.6535
- Pedagogical Audio Material for Teachers with score: 0.6522
- Art & culture with score: 0.6444
- Reuse of Educational Resources with score: 0.6258
- Local news with score: 0.5877
- Pedagogical issues with score: 0.5574
- Active Citizenship Issues with score: 0.5019
- Discussions Interviews Documentaries with score: 0.4215
- European and International News with score: 0.4092
- School news with score: 0.3505

Audio 4

Predicted Classes:

- Educational subjects with score: 0.9050
- European and International News with score: 0.8965
- Environment & health with score: 0.8249
- Digital Storytelling with score: 0.7510

- Human Relations with score: 0.7480
- Active Citizenship Issues with score: 0.6930
- Social issues with score: 0.6802
- School news with score: 0.6779
- Local news with score: 0.6543
- Interviews with score: 0.6278
- Pedagogical issues with score: 0.5872
- Science and Technology with score: 0.5372
- Reuse of Educational Resources with score: 0.4949
- Discussions Interviews Documentaries with score: 0.4890
- Pedagogical Audio Material for Teachers with score: 0.4578
- Cinema with score: 0.3997
- Sports with score: 0.3335
- Music with score: 0.3203

Audio 5

Predicted Classes:

- Social issues with score: 0.9123
- Educational subjects with score: 0.8781
- Human Relations with score: 0.7079
- Digital Storytelling with score: 0.7071
- Environment & health with score: 0.6954
- Pedagogical issues with score: 0.6864
- Interviews with score: 0.6788
- European and International News with score: 0.6465
- School news with score: 0.6443
- Reuse of Educational Resources with score: 0.6409
- Local news with score: 0.5994
- Active Citizenship Issues with score: 0.5994
- Discussions Interviews Documentaries with score: 0.5171
- Art & culture with score: 0.4923
- Music with score: 0.4854
- Cinema with score: 0.4753
- Science and Technology with score: 0.4699
- Pedagogical Audio Material for Teachers with score: 0.4653
- Books with score: 0.3771
- Theater & Sound Drama with score: 0.3629
- Sports with score: 0.3597
- Literature & Poetry with score: 0.3012

Κεφάλαιο 4ο: Συμπεράσματα, Περιορισμοί και Προτάσεις για Μελλοντική Έρευνα

4.1 Συμπεράσματα από την Εφαρμογή

Η εφαρμογή των αλγορίθμων YamNet, Deep FilterNet, Whisper και BART στην ανάλυση των podcasts του European School Radio οδήγησε σε σημαντικά συμπεράσματα σχετικά με τις δυνατότητες και την αποτελεσματικότητα των σύγχρονων τεχνικών μηχανικής μάθησης στην επεξεργασία ηχητικών δεδομένων.

Αρχικά, ο αλγόριθμος YamNet αποδείχθηκε ιδιαίτερα αποτελεσματικός στην αναγνώριση και κατηγοριοποίηση ηχητικών συμβάντων. Η χρήση της αρχιτεκτονικής MobileNetV1 και των depth-wise separable convolutions επέτρεψε την αποδοτική επεξεργασία των ηχητικών δεδομένων, ενώ η προεκπαίδευση στο σύνολο δεδομένων AudioSet προσέφερε υψηλή ακρίβεια στην αναγνώριση διαφόρων τύπων ήχων σε συγκεκριμένα χρονικά πεδία. Η εφαρμογή του Deep FilterNet στη βελτίωση της ποιότητας των ηχητικών δεδομένων έδειξε σημαντικά οφέλη. Η χρήση των τεχνικών αποθορυβοποίησης και καθαρισμού ήχου βελτίωσε σημαντικά την ποιότητα των ηχογραφήσεων, διευκολύνοντας την περαιτέρω ανάλυση και επεξεργασία τους. Ο αλγόριθμος Whisper επέδειξε εξαιρετική απόδοση στη μετατροπή ομιλίας σε κείμενο, με ιδιαίτερη ανθεκτικότητα σε διαφορετικές προφορές και θορυβώδη περιβάλλοντα. Η ικανότητά του να διαχειρίζεται πολυγλωσσικό περιεχόμενο και η υψηλή ακρίβεια στη μεταγραφή αποδείχθηκαν πολύτιμες για την ανάλυση των podcasts. Ο BART προσέφερε αποτελεσματικές λύσεις στην ανάλυση και κατανόηση του κειμένου που προέκυψε από τη μεταγραφή των ηχητικών. Η ικανότητά του στην κατηγοριοποίηση συνέβαλε σημαντικά στην κατανόηση του περιεχομένου των podcasts.

Η συνδυαστική χρήση των τεσσάρων αλγορίθμων μπορεί να δημιούργησει ένα σύστημα για την αυτοματοποιημένη ανάλυση και κατηγοριοποίηση των podcasts. Η εφαρμογή των αλγορίθμων στα podcasts του European School Radio ανέδειξε τη δυνατότητα αυτοματοποίησης της διαδικασίας επισήμανσης και κατηγοριοποίησης του περιεχομένου, προσφέροντας νέες δυνατότητες για την οργάνωση και αξιοποίηση του ηχητικού υλικού. Τα αποτελέσματα δείχνουν ότι η χρήση τεχνικών μηχανικής μάθησης μπορεί να βελτιώσει σημαντικά την προσβασιμότητα και την αξιοποίηση του εκπαιδευτικού περιεχομένου.

4.2 Περιορισμοί της Εργασίας και Προκλήσεις που Αντιμετωπίστηκαν

Παρά τα θετικά αποτελέσματα της έρευνας, η εργασία αντιμετώπισε διάφορους περιορισμούς και προκλήσεις που είναι σημαντικό να αναφερθούν. Οι περιορισμοί αυτοί σχετίζονται τόσο με τεχνικές πτυχές όσο και με ζητήματα υλοποίησης των αλγορίθμων. Στην περίπτωση του YamNet, ο κύριος περιορισμός ήταν η εξάρτηση από το προκαθορισμένο σύνολο των 521 κατηγοριών ήχου. Αυτό περιόρισε την ευελιξία του συστήματος στην αναγνώριση εξειδικευμένων ηχητικών γεγονότων που δεν περιλαμβάνονταν στις προκαθορισμένες κατηγορίες. Επιπλέον, η απόδοση του αλγορίθμου επηρεαζόταν σημαντικά από την ποιότητα των ηχητικών δεδομένων και την παρουσία θορύβου στο περιβάλλον αλλά ακόμη και απο την απουσία ήχου. Ο DeepFilterNet, αν και αποτελεσματικός στην αποθορυβοποίηση, παρουσίασε προκλήσεις στην επεξεργασία σε πραγματικό χρόνο λόγω των υψηλών υπολογιστικών απαιτήσεων. Η ανάγκη για ισχυρό υπολογιστικό εξοπλισμό περιόρισε τη δυνατότητα εφαρμογής του σε συστήματα με περιορισμένους πόρους. Επιπλέον, σε περιπτώσεις έντονου θορύβου, η διατήρηση της ποιότητας του αρχικού σήματος αποτέλεσε μια πρόκληση.

Ο αλγόριθμος Whisper, παρά την υψηλή ακρίβεια στη μετατροπή ομιλίας σε κείμενο, αντιμετώπισε δυσκολίες με συγκεκριμένες προφορές και διαλέκτους. Επίσης, μια σημαντική πρόκληση ήταν η τάση του αλγορίθμου να εισάγει τη λέξη "authorwave" ή "author" όταν αντιμετώπιζε αβεβαιότητα στην αναγνώριση, ιδιαίτερα σε περιπτώσεις θορυβώδους περιβάλλοντος ή χαμηλής ποιότητας ηχογράφησης αλλα ακομη και παρατεταμένης σιγής στο ηχητικό.

Τέλος, το BART, παρότι προσέφερε αποτελεσματικές λύσεις στην ανάλυση κειμένου, αντιμετώπισε προκλήσεις στην επεξεργασία μεγάλων κειμένων λόγω περιορισμών μνήμης και υπολογιστικών πόρων. Επιπλέον, η ποιότητα των παραγόμενων περιλήψεων και αναλύσεων εξαρτιόταν σημαντικά από την ποιότητα του αρχικού κειμένου που προέκυπτε από τη μετατροπή ομιλίας.

4.3 Προτάσεις για Μελλοντική Έρευνα και Πιθανές Επεκτάσεις

Η παρούσα έρευνα έχει αναδείξει διάφορες κατευθύνσεις για μελλοντική έρευνα και επέκταση στον τομέα της ανάλυσης ηχητικών δεδομένων με χρήση μηχανικής μάθησης. Οι προτάσεις αυτές αφορούν τη βελτίωση αλγορίθμων και μοντέλων, εφαρμογές σε νέους τομείς, ενσωμάτωση με άλλες τεχνολογίες και βελτίωση της προσβασιμότητας.

Αν υπήρχε περισσότερος χρόνος για την περάτωση αυτής της εργασίας, θα μπορούσε να συνεχιστεί περαιτέρω η έρευνα για την άντληση περισσότερων και ακριβέστερων πληροφοριών από τα ηχητικά podcasts. Μια κύρια προτεραιότητα θα ήταν η βελτίωση της ποιότητας του ήχου μέσω ειδικών διαδικασιών, ώστε να επιτευχθεί η αναβάθμιση της αρχικά χαμηλής ηχητικής απόδοσης, προσαρμοσμένη στις ανάγκες του κάθε συγκεκριμένου ηχητικού ώστε να βελτιωθεί ακόμη περισσότερο η ακρίβεια στην ανάλυση των αλγορίθμων.

Μια αλλη κατεύθυνση θα ήταν η ανάλυση της μουσικής με χρήση τεχνολογιών αναγνώρισης, όπως το Shazam, θα μπορούσε να αναγνωρίζει τη μουσική που υπάρχει στο ηχητικό υλικό και να αναλύει την κατηγορία της μουσικής. Αυτή η προσέγγιση θα μπορούσε να επεκτείνει την κατανόηση και την ανάλυση των ηχητικών δεδομένων, βελτιώνοντας την ακρίβεια και την ποικιλία των πληροφοριών που μπορούν να εξαχθούν από τα podcasts. Αυτό θα επέτρεπε την κατηγοριοποίηση των μουσικών κομματιών σε διάφορα είδη, παρέχοντας έτσι λεπτομερείς πληροφορίες σχετικά με το ηχητικό περιεχόμενο των podcasts.

Στον τομέα της βελτίωσης αλγορίθμων και μοντέλων, θα ήταν σημαντικό να αναπτυχθούν πιο αποδοτικές αρχιτεκτονικές για το YamNet που θα απαιτούν λιγότερους υπολογιστικούς πόρους. Επιπλέον, η βελτίωση της ικανότητας του Whisper να αναγνωρίζει ομιλία σε θορυβώδη περιβάλλοντα είναι κρίσιμη για την εφαρμογή του σε πιο απαιτητικά σενάρια. Η ενίσχυση των δυνατοτήτων του DeepFilterNet για καλύτερη αποθορυβοποίηση σε πραγματικό χρόνο αποτελεί επίσης προτεραιότητα, όπως και η επέκταση των δυνατοτήτων του BART για καλύτερη κατανόηση και ανάλυση κειμένου.

Όσον αφορά τις εφαρμογές σε νέους τομείς, θα ήταν ενδιαφέρον να εφαρμοστούν οι αλγόριθμοι σε άλλους τύπους ηχητικών δεδομένων, όπως μουσική και φυσικούς ήχους. Η ανάπτυξη συστημάτων για την ανάλυση ηχητικών δεδομένων σε πραγματικό χρόνο σε τομείς όπως η ασφάλεια και η υγεία θα μπορούσε να προσφέρει σημαντικές βελτιώσεις. Επιπλέον, η εφαρμογή των τεχνολογιών αυτών στον τομέα της εκπαίδευσης θα μπορούσε να βελτιώσει την εμπειρία μάθησης, παρέχοντας εξατομικευμένη ανάλυση και ανατροφοδότηση στους μαθητές.

Η ενσωμάτωση με άλλες τεχνολογίες προσφέρει επίσης πολλές δυνατότητες. Ο συνδυασμός των αλγορίθμων με τεχνολογίες επαυξημένης και εικονικής πραγματικότητας μπορεί να προσφέρει βελτιωμένη εμπειρία χρήστη. Η ενσωμάτωση με συστήματα IoT για την ανάλυση ηχητικών δεδομένων σε έξυπνα σπίτια και πόλεις θα μπορούσε επίσης να προσφέρει σημαντικά πλεονεκτήματα. Για παράδειγμα, η ανάλυση ηχητικών δεδομένων σε πραγματικό χρόνο θα μπορούσε να βελτιώσει την ασφάλεια και την άνεση των κατοίκων, ανιχνεύοντας ανωμαλίες ή επείγουσες καταστάσεις και ειδοποιώντας τους χρήστες έγκαιρα.

Η βελτίωση της προσβασιμότητας αποτελεί έναν σημαντικό τομέα. Η ανάπτυξη εργαλείων που θα βοηθήσουν άτομα με προβλήματα ακοής να κατανοούν καλύτερα το περιβάλλον τους μέσω ανάλυσης ηχητικών δεδομένων είναι κρίσιμη. Η δημιουργία συστημάτων που θα παρέχουν ζωντανές μεταγραφές και μεταφράσεις σε πραγματικό χρόνο θα μπορούσε να βελτιώσει σημαντικά την ποιότητα ζωής αυτών των ατόμων. Επιπλέον, οι τεχνολογίες αυτές θα μπορούσαν να χρησιμοποιηθούν για τη δημιουργία εκπαιδευτικών εργαλείων που θα βοηθήσουν τα άτομα με μαθησιακές δυσκολίες να κατανοούν καλύτερα και να επεξεργάζονται τις πληροφορίες.

Αρκετά ενδιαφέρουσα επίσης είναι η ενσωμάτωση τεχνολογιών ανάλυσης επεξεργασίας εικόνας που μπορεί να οδηγήσει στην ανάπτυξη προηγμένων εφαρμογών οι οποίες θα ενισχύσουν την αυτονομία και την ποιότητα ζωής ατόμων με προβλήματα όρασης. Η προσέγγιση συνδυάζει τις δυνατότητες των δύο τεχνολογιών, παρέχοντας ουσιαστική υποστήριξη στους χρήστες. Η χρήση αλγορίθμων επεξεργασίας εικόνας για την ανάλυση του οπτικού περιβάλλοντος επιτρέπει τη δημιουργία φωνητικών περιγραφών σε πραγματικό χρόνο. Οι αλγόριθμοι αναγνωρίζουν αντικείμενα, εμπόδια και άλλες σημαντικές πληροφορίες, οι οποίες μετατρέπονται σε φωνητικές περιγραφές και μεταδίδονται στους χρήστες μέσω ακουστικών. Αυτό το σύστημα προσφέρει στα άτομα με προβλήματα όρασης τη δυνατότητα να κινούνται με μεγαλύτερη αυτοπεποίθηση και ασφάλεια. Επίσης θα μπορούσε να δημιουργηθεί ξεχωριστά plugin ή να ενσωματωθεί με τα παραπάνω η νοηματική γλώσσα, η οποία θα μπορούσε να δέχεται περιεχόμενο απο τους browsers σε οποιαδήποτε γλώσσα και να μεταγλωτιζεται με την μορφή κινουμένων σχεδίων.

Με αυτες τις προσθήκες οι χρήστες θα μπορούν να διαχειρίζονται την καθημερινότητά τους χωρίς την ανάγκη συνεχούς βοήθειας από άλλους, ενισχύοντας την αυτοπεποίθηση και την αίσθηση αυτονομίας τους. Με τις φωνητικές περιγραφές και την ανάγνωση κειμένων, οι χρήστες θα έχουν άμεση πρόσβαση σε σημαντικές πληροφορίες που θα ήταν δύσκολο να προσλάβουν διαφορετικά. Αναγνωρίζοντας πρόσωπα και αντικείμενα, οι χρήστες θα μπορούν να συμμετέχουν ενεργά σε κοινωνικές δραστηριότητες και να διατηρούν καλύτερες σχέσεις με τους γύρω τους.

Επιπλέον, αυτές οι τεχνολογίες μπορούν να ωφελήσουν άτομα με μαθησιακές δυσκολίες. Η συνδυασμένη χρήση ηχητικών και οπτικών δεδομένων μπορεί να προσφέρει περιγραφές και πληροφορίες με τρόπο που να είναι κατανοητός και προσαρμοσμένος στις ανάγκες τους. Αυτό θα επιτρέψει στα άτομα με μαθησιακές δυσκολίες να βελτιώσουν την κατανόηση και την επεξεργασία των πληροφοριών, διευκολύνοντας την εκπαιδευτική τους διαδικασία και την καθημερινή τους ζωή. Οι τεχνολογίες αυτές μπορούν να παρέχουν εξατομικευμένη ανατροφοδότηση και υποστήριξη, συμβάλλοντας στη βελτίωση της μαθησιακής εμπειρίας και της γενικότερης ποιότητας ζωής τους.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1]Wikipedia Contributors, “History of sound recording,” Wikipedia, May 19, 2019. https://en.wikipedia.org/wiki/History_of_sound_recording

[2]A. R. Reppert, “HISTORY OF AUDIO, Analog to Digital Production | Timeline,” www.travsonic.com, Oct. 26, 2023. https://www.travsonic.com/history-of-audio-recording-analog-digital/

[3]T. Smith, “The History of Recording Sound/Music (Timeline),” Artloft Media, Nov. 19, 2022. https://artloftmedia.com/the-history-of-recording-sound-music-timeline/

[4]“Audio Media Timeline,” Museum of Obsolete Media. https://obsoletemedia.org/audio/audio-timeline/

[5]“An Audio Timeline,” www.aes.org. https://www.aes.org/aeshc/docs/audio.history.timeline.html

[6]Wikipedia Contributors, “Phonograph,” Wikipedia, Jan. 07, 2019. https://en.wikipedia.org/wiki/Phonograph

[7]“Western Electric,” Wikipedia, Dec. 22, 2022. https://en.wikipedia.org/wiki/Western_Electric

[8]“Audio Media Timeline,” Museum of Obsolete Media. https://obsoletemedia.org/audio/audio-timeline/

[9]Wikipedia Contributors, “Compact disc,” Wikipedia, Jul. 07, 2019. https://en.wikipedia.org/wiki/Compact_disc

[10]M. K. Mandal, “Digital Audio Processing,” Jan. 01, 2003. https://www.researchgate.net/publication/302279223_Digital_Audio_Processing

[11]“Listening in: Radio and the American imagination - DOKUMEN.PUB,” dokumen.pub, 2025. https://dokumen.pub/listening-in-radio-and-the-american-imagination.html (accessed Jan. 10, 2025).

[12]“Who we are - European School Radio Radio Community,” European School Radio Radio Community -, Dec. 11, 2023. https://community.europeanschoolradio.eu/who-we-are (accessed Jan. 2, 2025).

[13]Wikipedia Contributors, “Machine learning,” Wikipedia, Apr. 29, 2019. https://en.wikipedia.org/wiki/Machine_learning

[14]R. Karjian, “History and evolution of machine learning: A timeline,” WhatIs, 2024. https://www.techtarget.com/whatis/feature/History-and-evolution-of-machine-learning-A-timeline

[15]A. Jacinto, “Machine Learning History: The Complete Timeline,” Startechup Inc, Sep. 09, 2022. https://www.startechup.com/blog/machine-learning-history/

[16]Wikipedia Contributors, “Timeline of machine learning,” Wikipedia, Jan. 17, 2019. https://en.wikipedia.org/wiki/Timeline_of_machine_learning

[17]A. Jacinto, “Machine Learning History: The Complete Timeline,” Startechup Inc, Sep. 09, 2022. https://www.startechup.com/blog/machine-learning-history/

[18]K. D. Foote, “A Brief History of Machine Learning - DATAVERSITY,” DATAVERSITY, Dec. 03, 2021. https://www.dataversity.net/a-brief-history-of-machine-learning/

[19]R. Koch, “History of Machine Learning - A Journey through the Timeline,” clickworker.com, Sep. 01, 2022. https://www.clickworker.com/customer-blog/history-of-machine-learning/

[20]Suhas Maddali, “How can Machine Learning be used in Audio Analysis?,” Medium, Jan. 10, 2023. https://towardsdatascience.com/how-can-machine-learning-be-used-in-audio-analysis-847ebbefeb6?gi=4a73af07c4dbsdatascience.com/how-can-machine-learning-be-used-in-audio-analysis-847ebbefeb6 (accessed Jan. 10, 2025).

[21]“Audio Analytics,” Microsoft Research. https://www.microsoft.com/en-us/research/project/audio-analytics/

[22]“What Is Audio Data Collection? Importance & Benefits Explained,” Sapien.io, 2024. https://www.sapien.io/blog/what-is-audio-data-collection-and-why-is-it-important

[23]“Learn about the best tools and techniques for audio analysis with machine learning and how they can help you improve your audio editing skills.,” Linkedin.com, Mar. 13, 2023. https://www.linkedin.com/advice/3/what-best-tools-techniques-audio-analysis-machine (accessed Jan. 10, 2025).

[24]“Audio Analysis With Machine Learning: Building AI-Fueled Sound Detection App,” AltexSoft, May 12, 2022. https://www.altexsoft.com/blog/audio-analysis/

[25]“Deep Learning for Audio Applications - MATLAB & Simulink,” www.mathworks.com. https://www.mathworks.com/help/audio/gs/intro-to-deep-learning-for-audio-applications.html

[26]“An Introduction to Audio, Speech, and Language Processing,” www.appen.com. https://www.appen.com/blog/an-introduction-to-audio-speech-and-language-processing

[27]Wikipedia Contributors, “Natural language processing,” Wikipedia, Jan. 20, 2025.

[28]K. D. Foote, “A Brief History of Natural Language Processing (NLP),” DATAVERSITY, May 22, 2019. https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/

[29]P. Johri, S. K. Khatri, A. T. Al-Taani, M. Sabharwal, S. Suvanov, and A. Kumar, “Natural Language Processing: History, Evolution, Application, and Future Work,” Lecture Notes in Networks and Systems, pp. 365–375, 2021, doi: https://doi.org/10.1007/978-981-15-9712-1_31.

[30]S. Zhang and Y. Cheng, “Masking and noise reduction processing of music signals in reverberant music,” Journal of Intelligent Systems, vol. 31, no. 1, pp. 420–427, Jan. 2022, doi: https://doi.org/10.1515/jisys-2022-0024.

[31]J. Kaur, S. Baghla, and S. Kumar, “A REVIEW: AUDIO NOISE REDUCTION AND VARIOUS TECHNIQUES,” International Journal of Advances in Science Engineering and Technology, no. 3, pp. 2321–9009, 2015, Available: http://www.iraj.in/journal/journal_file/journal_pdf/6-162-1440572779132-135.pdf

[32]P. Akademia Baru, “A Survey of Filter Design for Audio Noise Reduction,” Journal of Advanced Review on Scientific Research ISSN, vol. 12, no. 1, pp. 26–44, 2015, Available: https://www.akademiabaru.com/doc/ARSRV12_N1_P26_44.pdf

[33]A. O. M. Salih, “Audio Noise Reduction Using Low Pass Filters,” OALib, vol. 04, no. 11, pp. 1–7, 2017, doi: https://doi.org/10.4236/oalib.1103709.

[34]D. Sinha, S. Saeed, and A. Ferreira, “A Novel Automatic Noise Removal Technique for Audio and Speech Signals,” vol. 3, Jan. 2007, Available:

https://www.researchgate.net/publication/260321897_A_Novel_Automatic_Noise_Removal_Technique_for_Audio_and_Speech_Signals

[35]R. Bentler and L.-K. Chiou, “Digital Noise Reduction: An Overview,” Trends in Amplification, vol. 10, no. 2, pp. 67–82, Jun. 2006, doi: https://doi.org/10.1177/1084713806289514.

[36]“Home - ID3.org,” Id3.org, 2020. https://id3.org

[37]“Adding ID3 Tags to MP3 Music and Podcast Automatically,” Cinch Solutions, Dec. 09, 2023. https://www.cinchsolution.com/adding-id3-tags/

[38]“How to Edit ID3 Tags for MP3,” Sage Audio, 2025. https://www.sageaudio.com/articles/edit-id3-tags-mp3

[39]Wikipedia Contributors, “ID3,” Wikipedia, Dec. 15, 2024.

[40]Florian Heidenreich, “Mp3tag - the universal Tag Editor (ID3v2, MP4, OGG, FLAC, ...),” Mp3tag.de, 2019. https://www.mp3tag.de/en/

[41]“Speech recognition,” www.ibm.com. https://www.ibm.com/history/voice-recognition

[42] Wikipedia Contributors, “IBM Shoebox,” Wikipedia, Aug. 25, 2024.

[43]“A Brief History of Speech Recognition,” www.linkedin.com. https://www.linkedin.com/pulse/brief-history-speech-recognition-cronan-mcnamara (accessed Sep. 13, 2021).

[44]“The early history of voice technologies in 6 short chapters,” Dasha.ai, Sep. 25, 2020. https://dasha.ai/en-us/blog/voice-technology-early-history

[45]“A brief history of speech recognition | Sonix,” Sonix.ai, 2017. https://sonix.ai/history-of-speech-recognition (accessed Sep. 24, 2019).

[46]an, “‘Shoebox’: an artificial intelligence history project,” dale lane, Jan. 11, 2025. https://dalelane.co.uk/blog/?p=5463 (accessed Jan. 7, 2025).

[47]Y. Yang and T. Joachims, “Text categorization,” Scholarpedia, vol. 3, no. 5, p. 4242, 2008, doi: https://doi.org/10.4249/scholarpedia.4242.

[48]H. Zhu and L. Lei, “The Research Trends of Text Classification Studies (2000–2020): A Bibliometric Analysis,” SAGE Open, vol. 12, no. 2, p. 215824402210899, Apr. 2022, doi: https://doi.org/10.1177/21582440221089963.

[49]Y. Aphinyanaphongs, “Text Categorization Models for High-Quality Article Retrieval in Internal Medicine,” Journal of the American Medical Informatics Association, vol. 12, no. 2, pp. 207–216, Nov. 2004, doi: https://doi.org/10.1197/jamia.m1641.

[50]D. Shen, “Text Categorization,” Encyclopedia of Database Systems, pp. 3041–3044, 2009, doi: https://doi.org/10.1007/978-0-387-39940-9_414.

[51]“Unified Audio Event Detection,” Arxiv.org, 2023. https://arxiv.org/html/2409.08552v1 (accessed Jan. 5, 2025).

[52]A. Kumar and B. Raj, “Audio Event Detection using Weakly Labeled Data,” Proceedings of the 24th ACM international conference on Multimedia, Oct. 2016, doi: https://doi.org/10.1145/2964284.2964310.

[53]E. Babaee, N. B. Anuar, A. W. Abdul Wahab, S. Shamshirband, and A. T. Chronopoulos, “An Overview of Audio Event Detection Methods from Feature Extraction to Classification,” Applied Artificial Intelligence, vol. 31, no. 9–10, pp. 661–714, Nov. 2017, doi: https://doi.org/10.1080/08839514.2018.1430469.

[54]C. Clavel, T. Ehrette, and G. Richard, “Events Detection for an Audio-Based Surveillance System,” IEEE Xplore, Jul. 01, 2005. https://ieeexplore.ieee.org/document/1521669 (accessed Mar. 24, 2021).

[55]A. Mesaros, T. Heittola, Antti Eronen, and Tuomas Virtanen, “Acoustic event detection in real-life recordings,” 18th European Signal Processing Conference, Jul. 2014, Available: https://www.researchgate.net/publication/228880528_Acoustic_event_detection_in_real-life_recordings

[56]C. Mesa-Cantillo et al., “A Sound Events Detection and Localization System based on YAMNet Model and BLE Beacons.” Accessed: Jan. 5, 2025. [Online]. Available: https://accedacris.ulpgc.es/bitstream/10553/121504/1/icwmc_2023_1_10_20007.pdf

[57]“Transfer learning with YAMNet for environmental sound classification | TensorFlow Core,” TensorFlow. https://www.tensorflow.org/tutorials/audio/transfer_learning_audio

[58]“Transfer Learning Yamnet Overview | Restackio,” Restack.io, 2025. https://www.restack.io/p/transfer-learning-answer-yamnet-cat-ai (accessed Jan. 5, 2025).

[59]I. Kuzminykh, “Audio Interval Retrieval using Convolutional Neural Networks.” Accessed: Jan. 5, 2025. [Online]. Available: https://pure.port.ac.uk/ws/portalfiles/portal/26773756/Audio_interval_retrieval.pdf

[60]“Audio Classification Using Google’s YAMnet,” GeeksforGeeks, Feb. 22, 2024. https://www.geeksforgeeks.org/audio-classification-using-googles-yamnet/

[61]“Preprocess audio for YAMNet classification - MATLAB yamnetPreprocess,” www.mathworks.com. https://www.mathworks.com/help/audio/ref/yamnetpreprocess.html

[62]“Transfer Learning for Audio Data with YAMNet.” https://blog.tensorflow.org/2021/03/transfer-learning-for-audio-data-with-yamnet.html

[63]A. Radford, J. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” Dec. 2022. Available: https://cdn.openai.com/papers/whisper.pdf

[64] OpenAI. (2022). Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. OpenAI Research.

[65]D. Amodei et al., “Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin.” Available: https://proceedings.mlr.press/v48/amodei16.pdf

[66]A. Vaswani et al., “Attention Is All You Need,” 2017. Available: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[67]A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.” Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[68]M. Milling, S. Liu, A. Triantafyllopoulos, I. Aslan, and B. W. Schuller, “Audio Enhancement for Computer Audition—An Iterative Training Paradigm Using Sample Importance,” Journal of Computer Science and Technology, vol. 39, no. 4, pp. 895–911, Jul. 2024, doi: https://doi.org/10.1007/s11390-024-2934-x.

[69]F. P. Romero, D. C. Piñol, and C. R. Vázquez-Seisdedos, “DeepFilter: An ECG baseline wander removal filter using deep learning techniques,” Biomedical Signal Processing and Control, vol. 70, p. 102992, Sep. 2021, doi: https://doi.org/10.1016/j.bspc.2021.102992.

[70]I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning,” Deeplearningbook.org, 2016. https://www.deeplearningbook.org/

[71]O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” May 2015. Available: https://arxiv.org/pdf/1505.04597

[72]P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and Composing Robust Features with Denoising Autoencoders.” Available: https://www.cs.toronto.edu/~larocheh/publications/icml-2008-denoising-autoencoders.pdf

[73]“IEEE/ACM Transactions on Audio, Speech, and Language Processing | IEEE Xplore,” Ieee.org, 2025. https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6570655 (accessed Jan. 11, 2025).

[74]“BART,” huggingface.co. https://huggingface.co/docs/transformers/model_doc/bart

[75]M. Lewis et al., “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” arXiv:1910.13461 [cs, stat], Oct. 2019, Available: https://arxiv.org/abs/1910.13461

[76]J.O. Schneppat, “BART (Bidirectional and Auto-Regressive Transformers),” Schneppat AI, 2020. https://schneppat.com/bart.html (accessed Jan. 23, 2025).

[77]“BART — transformers 4.1.1 documentation,” Huggingface.co, 2019. https://huggingface.co/transformers/v4.1.1/model_doc/bart.html (accessed Jan. 13, 2025).

ΠΑΡΑΡΤΗΜΑ Α: Κώδικες υλοποίησης των Μοντέλων

Κώδικας Yamnet

```
# Εισαγωγή απαραίτητων βιβλιοθηκών
import numpy as np
import pandas as pd
import soundfile as sf
import os
import matplotlib.pyplot as plt
import params as yamnet_params
import yamnet as yamnet_model
import tensorflow as tf

# pip install pydub

from pydub import AudioSegment

# Καθορισμός της διαδρομής προς το εκτελέσιμο ffmpeg
AudioSegment.converter = "C:/ffmpeg/bin/ffmpeg.exe"
# pip install resampy
import resampy

# Συνάρτηση για μετατροπή MP3 σε WAV
def convert_mp3_to_wav(mp3_file, wav_file):
    """
    # Μετατρέπει ένα αρχείο MP3 σε WAV
    """
    try:
        # Φόρτωση του MP3 αρχείου
        audio = AudioSegment.from_mp3(mp3_file)

        # Εξαγωγή σε μορφή WAV
        audio.export(wav_file, format="wav")

        print(f"Η μετατροπή ολοκληρώθηκε: {mp3_file} -> {wav_file}")
    except Exception as e:
```

```
        print(f"Η μετατροπή απέτυχε: {e}")

# Ορισμός του αρχείου εισόδου MP3
mp3_file = "18031996-the_power_of_friendship-9982.mp3"



wav_file = "output.wav"

# Κλήση της συνάρτησης μετατροπής
convert_mp3_to_wav(mp3_file, wav_file)

# Συνάρτηση για μετατροπή WAV σε 16kHz
def convert_wav_to_16k(input_filename, output_filename):
    """
    # Μετατρέπει ένα αρχείο WAV σε μορφή 16 kHz mono PCM
    """
    # Ανάγνωση του WAV αρχείου
    wav_data, sample_rate = sf.read(input_filename)

    # Μετατροπή σε μονοφωνικό αν χρειάζεται
    if wav_data.ndim > 1:
        wav_data = np.mean(wav_data, axis=1, dtype=wav_data.dtype)

    # Επαναδειγματοληψία στα 16 kHz
    wav_data_16k = resampy.resample(wav_data, sample_rate, 16000)

    # Προσαρμογή μήκους ήχου
    expected_length = int(len(wav_data) * 16000 / sample_rate)
    wav_data_16k = np.resize(wav_data_16k, expected_length)

    # Αποθήκευση του νέου WAV αρχείου
    sf.write(output_filename, wav_data_16k, 16000, format='WAV', subtype='PCM_16')

# Μετατροπή του αρχείου σε 16kHz
convert_wav_to_16k('output.wav', 'output_16k.wav')

# Ανάγνωση του ήχου και προετοιμασία για το μοντέλο
wav_file_name = 'output_16k.wav'
```

```
wav_data, sr = sf.read(wav_file_name, dtype=np.int16)
waveform = wav_data / 32768.0
# Το γράφημα είναι σχεδιασμένο για ρυθμό δειγματοληψίας 16 kHz, αλλά υψηλότεροι ρυθμοί θα λειτουργήσουν επίσης
# Επίσης δημιουργούμε βαθμολογίες με ρυθμό καρέ 10 Hz
params = yamnet_params.Params(sample_rate=sr, patch_hop_seconds=0.1)

print("Ρυθμός δειγματοληψίας =", params.sample_rate)

# Προετοιμασία του μοντέλου YAMNet
class_names = yamnet_model.class_names('yamnet_class_map.csv')
yamnet = yamnet_model.yamnet_frames_model(params)
yamnet.load_weights('yamnet.h5')
# Εκτέλεση του μοντέλου
scores, embeddings, spectrogram = yamnet(waveform)
scores = scores.numpy()
spectrogram = spectrogram.numpy()
# Δημιουργία γραφικών παραστάσεων
plt.figure(figsize=(10, 8))

# Σχεδίαση της κυματομορφής
plt.subplot(3, 1, 1)
plt.plot(waveform)
plt.xlim([0, len(waveform)])

# Σχεδίαση του λογαριθμικού-mel φασματογραφήματος (που επιστρέφεται από το μοντέλο)
plt.subplot(3, 1, 2)
plt.imshow(spectrogram.T, aspect='auto', interpolation='nearest', origin='lower')

# Σχεδίαση και επισήμανση των βαθμολογιών εξόδου του μοντέλου για τις κορυφαίες κατηγορίες
mean_scores = np.mean(scores, axis=0)
top_N = 10
top_class_indices = np.argsort(mean_scores)[::-1][:top_N]
plt.subplot(3, 1, 3)
plt.imshow(scores[:, top_class_indices].T, aspect='auto', interpolation='nearest', cmap='gray_r')

# Αντιστάθμιση για το παράθυρο context patch_window_seconds (0.96s) για ευθυγράμμιση με το φασματογράφημα
patch_padding = (params.patch_window_seconds / 2) / params.patch_hop_seconds
plt.xlim([-patch_padding, scores.shape[0] + patch_padding])
```

```
# Επισήμανση των top_N κατηγοριών
yticks = range(0, top_N, 1)
plt.yticks(yticks, [class_names[top_class_indices[x]] for x in yticks])

# Προσαρμογή των ορίων y και αποθήκευση του γραφήματος
plt.ylim(-0.5 + np.array([top_N, 0]))

plt.tight_layout()

# Αποθήκευση του γραφήματος ως αρχείο .png
plt.savefig("output_plot.png", dpi=200)

# Προαιρετική εμφάνιση του γραφήματος άμεσα
plt.show()

top_class_names = [class_names[i] for i in top_class_indices]
top_class_scores = scores[:, top_class_indices]

# Plot and label the model output scores for the top-scoring classes.
 mean_scores = np.mean(scores, axis=0)
 top_n = 30
 top_class_indices = np.argsort(mean_scores)[::-1][:top_n]
 plt.figure(figsize=(15,50))
plt.imshow(scores[:, top_class_indices].T, aspect='auto', interpolation='nearest', cmap='gray_r')

patch_padding = (PATCH_WINDOW_SECONDS / 2) / PATCH_HOP_SECONDS
values from the model documentation
patch_padding = (0.025 / 2) / 0.01
 plt.xlim([-patch_padding-0.5, scores.shape[0] + patch_padding-0.5])
Label the top_N classes.
 yticks = range(0, top_n, 1)
 plt.yticks(yticks, [class_names[top_class_indices[x]] for x in yticks])
 _ = plt.ylim(-0.5 + np.array([top_n, 0]))

audio = AudioSegment.from_file(wav_file)

# Λήψη της διάρκειας σε δευτερόλεπτα
```

```
duration = len(audio)/1000
print(duration)

# Εξασφάλιση ότι το PATCH_HOP_SECONDS ταιριάζει με την ακριβή διάρκεια του ήχου
audio_duration = duration  # Πραγματική διάρκεια ήχου σε δευτερόλεπτα
PATCH_HOP_SECONDS = audio_duration / (scores.shape[0] - 1)  # Προσαρμογή για να καλύψει όλη τη διάρκεια

# Υπολογισμός χρόνων για κάθε πλαίσιο
times = np.arange(scores.shape[0]) * PATCH_HOP_SECONDS

# Δημιουργία DataFrame για να συμπεριλάβει όλες τις λεπτομέρειες
data = {
    'Time (s)': times
}

# Προσθήκη των κορυφαίων βαθμολογιών κατηγοριών στα δεδομένα
for idx, class_name in enumerate(top_class_names):
    data[class_name] = top_class_scores[:, idx]

# Μετατροπή σε DataFrame
df = pd.DataFrame(data)

# Αποθήκευση σε CSV
df.to_csv('top_classes_with_times.csv', index=False)

print("Το αρχείο CSV 'top_classes_with_times.csv' δημιουργήθηκε με επιτυχία.")

# Διαγραφή των αρχείων output.wav και output_16k.wav για σάρωση νέων αρχείων ήχου

file_path = 'output.wav'
file_path2='output_16k.wav'

try:
    os.remove(file_path)
    print(f"Το αρχείο {file_path} διαγράφηκε με επιτυχία.")
except FileNotFoundError:
    print(f"Το αρχείο {file_path} δε βρέθηκε.")
except PermissionError:
```

```
    print(f"Άρνηση πρόσβασης: {file_path}.")
except Exception as e:
    print(f"Σφάλμα: {e}")


try:
    os.remove(file_path2)
    print(f"Το αρχείο {file_path2} διαγράφηκε με επιτυχία.")
except FileNotFoundError:
    print(f"Το αρχείο {file_path2} δε βρέθηκε.")
except PermissionError:
    print(f"Άρνηση πρόσβασης: {file_path2}.")
except Exception as e:    print(f"Σφάλμα: {e}")
```

Κωδικας DeepFilterNet

```
# Εισαγωγή απαραίτητων βιβλιοθηκών
import os
from pydub import AudioSegment
from df.enhance import main as enhance_main

# Παίρνουμε το τρέχον directory του αρχείου
current_dir = os.path.dirname(__file__)
print(current_dir)

def delete_file(file_path):
    """Διαγράφει ένα αρχείο εάν υπάρχει."""
    if os.path.exists(file_path):
        os.remove(file_path)
        print(f"Το αρχείο '{file_path}' διαγράφηκε.")
    else:
        print(f"Το αρχείο '{file_path}' δεν υπάρχει.")

def convert_flac_to_wav(flac_file_path, wav_file_path):
    # Φόρτωση του αρχείου FLAC
    audio = AudioSegment.from_file(flac_file_path, format="flac")

    # Εξαγωγή του ήχου σε μορφή WAV
    audio.export(wav_file_path, format='wav')
    print(f"Η μετατροπή από '{flac_file_path}' σε '{wav_file_path}' ολοκληρώθηκε επιτυχώς.")

def convert_mp3_to_flac(mp3_file_path, flac_file_path):
    # Φόρτωση του αρχείου MP3
    audio = AudioSegment.from_mp3(mp3_file_path)

    # Εξαγωγή του ήχου σε μορφή FLAC
    audio.export(flac_file_path, format='flac')
    print(f"Η μετατροπή από '{mp3_file_path}' σε '{flac_file_path}' ολοκληρώθηκε επιτυχώς.")
```

```
def enhance_audio(input_file, output_dir):
    # Δημιουργία του φακέλου εξόδου εάν δεν υπάρχει
    os.makedirs(output_dir, exist_ok=True)

    # Ρύθμιση παραμέτρων command-line για την ενίσχυση του ήχου
    class Args:
        def __init__(self, input_file, output_dir):
            self.input_file = input_file  # Αρχείο εισόδου
            self.output_dir = output_dir  # Φάκελος εξόδου
            self.model_base_dir = None    # Φάκελος μοντέλου (προαιρετικό)
            self.sr = 16000               # Ρυθμός δειγματοληψίας
            self.output_format = 'wav'    # Μορφή αρχείου εξόδου
            self.noise_reduction = True    # Ενεργοποίηση μείωσης θορύβου
            self.pf = True                # Ενεργοποίηση μετα-φιλτραρίσματος
            self.log_level = "INFO"       # Επίπεδο καταγραφής
            self.epoch = "best"           # Επιλογή καλύτερης εποχής
            self.no_df_stage = False      # Απενεργοποίηση σταδίου βαθιού φιλτραρίσματος
            self.suffix = None            # Επίθημα αρχείου
            self.noisy_dir = None         # Φάκελος θορυβωδών αρχείων
            self.noisy_audio_files = [input_file]  # Λίστα θορυβωδών αρχείων
            self.compensate_delay = True   # Αντιστάθμιση καθυστέρησης
            self.atten_lim = None         # Όρια εξασθένησης

    # Δημιουργία αντικειμένου παραμέτρων
    args = Args(input_file=input_file, output_dir=output_dir)

    # Κλήση της συνάρτησης ενίσχυσης
    enhance_main(args)

if __name__ == "__main__":
    # Ορισμός διαδρομών εισόδου και εξόδου
    mp3_file = r"C:\Users\Administrator\Desktop\DeepFilterNet\DeepFilterNet\output23.wav"
    flac_file = os.path.join(current_dir, "output.flac")
    wav_file = os.path.join(current_dir,  "output.wav")
```

```
# Μετατροπή από MP3 σε FLAC
convert_mp3_to_flac(mp3_file, flac_file)

input_flac_path = flac_file
output_directory = os.path.join(current_dir)
# Ενίσχυση ήχου
enhance_audio(input_flac_path, output_directory)

# Μετατροπή από FLAC σε WAV
convert_flac_to_wav(flac_file, wav_file)

# Διαγραφή του προσωρινού αρχείου FLAC
delete_file(flac_file)
```

Κωδικας Whisper

```
import whisper
import torch
import threading
import itertools
import time
import os

# Ρύθμιση χρήσης της κάρτας γραφικών
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

# Έλεγχος διαθεσιμότητας CUDA
print("CUDA Available:", torch.cuda.is_available())

# Φόρτωση του μοντέλου Whisper (στην σωστή συσκευή)
model = whisper.load_model("turbo").to(device)  # Αντικαταστήστε το 'base' με 'turbo' αν χρειάζεται

# Φόρτωση του αρχείου ήχου
audio_path = r"output_audio_eq.wav"

# Ρυθμίσεις αποκωδικοποίησης για το Whisper
options = whisper.DecodingOptions(
    task="transcribe",
    #language="en",
    temperature=0.23,          # Ντετερμινιστική έξοδος για ταχύτερα αποτελέσματα
    sample_len=None,
    best_of=2,                # Μόνο 1 υπόθεση για γρηγορότερη επεξεργασία
    beam_size=None,           # Μονή ακτίνα για γρήγορη και αποτελεσματική αποκωδικοποίηση
    patience=None,            # Μέτρια υπομονή για χειρισμό ασαφούς ομιλίας χωρίς επιβράδυνση
    length_penalty=0.5,       # Μικρή ποινή για αποφυγή περικοπής χρήσιμων πληροφοριών
    prompt="Transcribe the audio with accuracy, focusing on unclear speech and also looking for audio inside the silences",
    prefix=None,
    suppress_blank=False,     # Παράλειψη κενών εξόδων για εστίαση στην ομιλία
    without_timestamps=False,  # Συμπερίληψη χρονικών σημάνσεων για πλήρη μεταγραφή
    max_initial_timestamp=0.5,
    fp16=torch.cuda.is_available()  # Χρήση GPU αν είναι διαθέσιμη για ταχύτερη εκτέλεση
)

# Συνάρτηση για την εμφάνιση του δείκτη προόδου
def spinner_task():
    for char in itertools.cycle(['|', '/', '-', '\\']):
        if not spinner_running:
            break
        print(f"\rΜεταγραφή ήχου... {char}", end="", flush=True)
```

```
        time.sleep(0.1)

# Εκκίνηση του δείκτη προόδου
spinner_running = True
spinner_thread = threading.Thread(target=spinner_task)
spinner_thread.start()

# Μεταγραφή του ήχου
try:
    print("\nΈναρξη μεταγραφής...")
    result = model.transcribe(audio_path, **vars(options))
    print("\rΗ μεταγραφή ολοκληρώθηκε!\n")

finally:
    # Τερματισμός του δείκτη προόδου
    spinner_running = False
    spinner_thread.join()

# Εκτύπωση των αποτελεσμάτων με χρονικές σημάνσεις
for segment in result["segments"]:
    start_time = segment["start"]
    end_time = segment["end"]
    text = segment["text"]
    print(f"[{start_time:.2f}s - {end_time:.2f}s] {text}")
```

Κώδικας Bart

```
# Εισάγουμε το pipeline από τη βιβλιοθήκη transformers
from transformers import pipeline

# Φορτώνουμε το pipeline ταξινόμησης zero-shot χρησιμοποιώντας το μοντέλο BART
classifier = pipeline("zero-shot-classification", model="facebook/bart-large-mnli")

# Οι προσαρμοσμένες ετικέτες υποψηφίων (κατηγορίες)
candidate_labels = [
"Environment & health",
"Art & culture",
"Music",
"Books",
"Literature & Poetry",
"Sports",
"Local news",
"School news",
"Social issues",
"Reuse of Educational Resources",
"Pedagogical Audio Material for Teachers",
"Science and Technology",
"European and International News",
"Digital Storytelling",
"Discussions Interviews Documentaries",
"Active Citizenship Issues",
"Theater & Sound Drama",
"Cinema",
"Human Relations",
"Poetry",
"Pedagogical issues",
"Fairy tales",
"Educational subjects",
"Interviews",
]

# Δείγμα κειμένων προς ταξινόμηση
texts = [
    '''    Good morning, we are the theater group of the first general high school of Kavalla, which has been active for ten years, presenting theatrical performances to the public of our city.'''
]

# Ταξινόμηση κειμένων
for text in texts:
    # Εκτελούμε την ταξινόμηση με δυνατότητα πολλαπλών ετικετών
    result = classifier(text, candidate_labels, multi_label=True)
```

```
# Φιλτράρουμε τις κατηγορίες με βάση ένα όριο βαθμολογίας (π.χ. 0.3)
threshold = 0.3
predicted_classes = [(label, score) for label, score in zip(result['labels'], result['scores']) if score >= threshold]

# Εκτύπωση των πρώτων 50 χαρακτήρων του κειμένου για συντομία
print(f"Text: {text[:50]}...")

# Εκτύπωση των προβλεπόμενων κατηγοριών
print("Predicted Classes:")
for label, score in predicted_classes:
    print(f"  - {label} with score: {score:.4f}")
```