Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Αυτόματη Αναγνώριση Ομιλίας Σχολιαστών Ποδοσφαίρου με Χρήση Τεχνικών Βαθιάς Μάθησης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΛΕΝΗ Κ. ΤΣΙΛΙΓΙΑΝΝΗ

Επιβλέπων : Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Θεόδωρος Γιαννακόπουλος
Ερευνητής Β', ΙΙΤ Δημόκριτος

Αθήνα, Σεπτέμβριος 2023

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Αυτόματη Αναγνώριση Ομιλίας Σχολιαστών Ποδοσφαίρου με Χρήση Τεχνικών Βαθιάς Μάθησης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΛΕΝΗ Κ. ΤΣΙΛΙΓΙΑΝΝΗ

Επιβλέπων : Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Θεόδωρος Γιαννακόπουλος
Ερευνητής Β', ΙΙΤ Δημόκριτος

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Σεπτεμβρίου 2023.

.....	.....	.....
Γεώργιος Στάμου Καθηγητής Ε.Μ.Π.	Θεόδωρος Γιαννακόπουλος Ερευνητής Β', ΙΙΤ Δημόκριτος	Αθανάσιος Βουλόδημος Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2023

Ελένη Κ. Τσιλιγιάννη

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Σκοπός της παρούσας μεταπτυχιακής διπλωματικής εργασίας είναι η μελέτη συστημάτων αυτόματης αναγνώρισης ομιλίας και η δημιουργία ενός τέτοιου συστήματος ειδικού σκοπού που εφαρμόζεται πάνω σε δεδομένα περιγραφών ποδοσφαιρικών αγώνων. Μελετήθηκαν κλασικές μέθοδοι κατασκευής ASR συστημάτων που χρησιμοποιούνται στην παραγωγή αλλά και οι πιο σύγχρονες, όπως τα συστήματα από άκρη-σε-άκρη που χρησιμοποιούν μετασχηματιστές, ως επί το πλείστον, όπως και το δικό μας σύστημα.

Λάβαμε υπόψιν 3 σετ δεδομένων εκ των οποίων τα 2 ήταν γενικού σκοπού και περιεχομένου ενώ το τρίτο αφορούσε τον ειδικό τομέα. Η πρωτόλεια μορφή αυτού του σετ δεδομένων μας ελήφθη αρχικά ως βίντεο, ενώ στη συνέχεια έγινε εξαγωγή του ηχητικού σήματος, ώστε να παραχθεί το κατάλληλο σύνολο χαρακτηριστικών που να μπορεί να τροφοδοτήσει με ορθό και αποτελεσματικό τρόπο μία διάταξη βαθιάς μάθησης που αποτελεί το ακουστικό μοντέλο. Στο κείμενο αναφοράς που παράχθηκε έγινε διόρθωση και υποσημείωση χειροκίνητα, για να υπολογιστεί σωστά η επίδοση για καθένα από τα πειράματα και η σύγκριση μεταξύ τους.

Στα πειράματα που ακολούθησαν αναλύθηκαν η διαδικασία τόσο της εκπαίδευσης όσο και του fine-tuning του μοντέλου για τους διαφορετικούς συνδυασμούς των σετ δεδομένων. Η εργασία ολοκληρώνεται με προτάσεις βελτίωσης της απόδοσης αλλά και μελλοντικών επεκτάσεων χρήσης του συστήματος.

Λέξεις κλειδιά

Αυτόματη Αναγνώριση Ομιλίας από άκρο-σε-άκρο, Ακουστικό Μοντέλο, Γλωσσικό Μοντέλο, Wav2Vec 2.0, XLS-R, Βαθιά Μάθηση, Μετασχηματιστές, MFCCs, Συνδετική Χρονική Ταξινόμηση, Κωδικοποιητής-Αποκωδικοποιητής, Επαναλαμβανόμενος Μετατροπέας Νευρωνικού Δικτύου, Αυτο-εποπτευόμενη μάθηση, Αντιθετική (Contrastive) μάθηση, Αναζήτηση Δέσμης, KenLM, Φώνημα, N-grams, Ποσοστό Σφάλματος Λέξης

Abstract

The purpose of this master's thesis is the study of Automatic Speech Recognition systems and the development of such a system, that is domain specific and trained on football commentator speech data. We examined the standard techniques of constructing an ASR system that is used in production as well as the well-performing state of the art ones, so called End-to-End ASR systems that rely on Transformers. The model that we built in this diploma thesis is an End-to-End Transformer based model, that also uses a Language Model.

We are considering 3 datasets, 2 of them of general context and various domains, while the third was targeting a specific domain, the football related speech. The raw data of the latter set is downloaded in video format so we extracted the audio signal, in order to acquire the necessary features that will be properly and efficiently fed into a deep learning network, which is our acoustic model. The ground truth text that was produced had to pass through a manual error correction and annotation process, in order to calculate the performance of the system and apply comparisons between the various experiments.

In the experiments that followed we analyzed both training and fine-tuning processes of the model for each dataset combination. In the end the thesis is completed with proposals on performance improvement and some scenario cases on which the E2E ASR system could be used and ported in the future.

Key words

End-to-End Automatic Speech Recognition, Acoustic Model, Language Model, Wav2Vec 2.0, XLS-R, Deep Learning, Transformers, MFCCs, Connectionist Temporal Classification (CTC), Encoder - Decoder, Recurrent Neural Network-Transducer (RNN-T), Self-Supervides Learning, Contrastive Learning, Beam Search, KenLM, Phoneme, N-grams, Word Error Rate (WER)

Ευχαριστίες

Ας μου επιτραπεί αυτό το σημαντικό κομμάτι να γραφτεί σε πιο ελεύθερη γλώσσα, καθώς είναι προσωπικό και αφιερώνεται στους παρόντες των ολίγων αυτών μηνών που διήρκησε η εκπόνηση της εργασίας και χωρίς αυτούς η ολοκλήρωσή της δεν θα ήταν δυνατή (είστε πάρα πολλοί, κι ενώ πολύ θα ήθελα να αναφερθώ στον καθένα ξεχωριστά, για να μην πλατειάζω - εξάλλου ξέρετε ποιοι είστε - θα γίνει ομαδοποίηση).

Αρχικά λοιπόν, κι όπως είθισται άλλωστε, θα ήθελα να ευχαριστήσω τον επιβλέποντα της εργασίας Δρ. Θεόδωρο Γιαννακόπουλο για το πολύ ενδιαφέρον θέμα με το οποίο ασχοληθήκαμε, τις γνώσεις, τη θέληση και την καθοδήγησή του που βοήθησαν στο να έχουμε μία άψογη συνεργασία και να κυλήσουν όλα ομαλά και σε εύλογο χρονικό διάστημα. Επίσης ευχαριστώ θερμά τον Καθηγητή κ. Γιώργο Στάμου για την πολύτιμη βοήθεια και εκτιμώ απεριόριστα τον χρόνο που αφιέρωσε στη μετάδοση γνώσης καθ' όλη τη διάρκεια των μεταπτυχιακών μου σπουδών.

Στη συνέχεια ένα τεράστιο ευχαριστώ στους ανθρώπους που με είδαν και με άντεξαν στις ακραίες στιγμές μου, τα υπερόπλα μου όπως τους ονομάζω, για όλες τις συμβουλές και τη συμπαράστασή τους που είμαι τυχερή να έχω κατά τη διάρκεια της ζωής μου και δε θα μπορούσαν να λείπουν κι από αυτό το διάστημα.

Ευχαριστώ τις "αδερφές" μου με τις οποίες μοιράστηκα προβληματισμούς επί της εργασίας αλλά και πολύ βαθιές μου σκέψεις που, όπως πάντα, οδηγούν σε ζόρικες αλλά καθ' όλα εποικοδομητικές κουβέντες. Έπειτα δεν θα παρέλειπα να ευχαριστήσω τις φίλες και φίλους, παλιούς και νέους, από την ομάδα και εκτός, με τους οποίους πέρασα αυτούς τους μήνες πολλές ευχάριστες στιγμές που εξισορρόπησαν κι έδωσαν νόημα σε κάθε κόπο, κούραση ή απογοήτευση που εμφανίστηκε κατά τη διάρκεια του δρόμου.

Τέλος, και ίσως πάνω από όλα, οφείλω απέραντη ευγνωμοσύνη στον MVP αυτής της περιόδου, για τις "καλές αντάμωσες", για τα "πανέμορφα μέρη και τα ομορφότερα που θα έρθουν", για τις αναμνήσεις που θα με κρατούν πάντα πιστή, αισιόδοξη και αφρενάριστη, κυρίως όμως για τη συμβολή του στην ανάκτηση της, έως πρότινος, κλεμμένης σπίθας μου.

Ελένη Κ. Τσιλιγιάννη,

Αθήνα, 15η Σεπτεμβρίου 2023

Περιεχόμενα

Κατάλογος σχημάτων

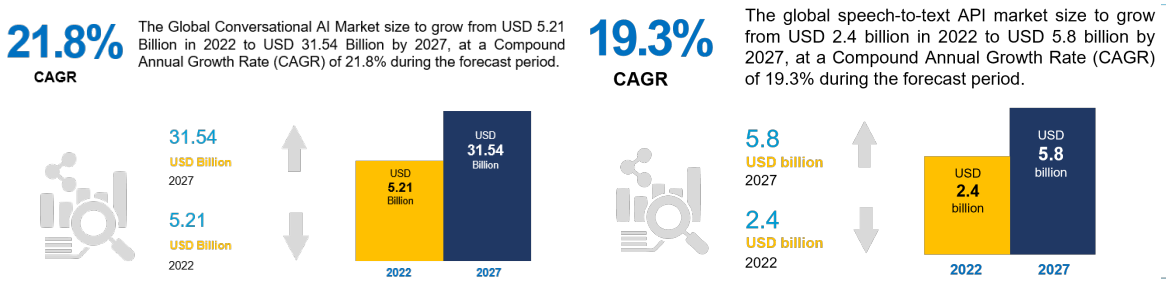
Κατάλογος πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγή

Το μέγεθος της αγοράς διαλογικών συστημάτων παγκοσμίως ξεπερνά τα 6 δισεκατομμύρια δολλάρια το 2022 και αναμένεται να έχει περίπου 22% αύξηση των κερδών σύμφωνα με τον δείκτη του σύνθετου ρυθμού ετήσιας ανάπτυξης (Compound Annual Growth Rate - CAGR) από το 2022 έως το 2027. Ένα υποσύνολο των εργασιών που περιλαμβάνει ένα διαλογικό σύστημα είναι η μεταγραφή ομιλίας σε κείμενο (Speech-to-Text), όπου η αντίστοιχη πρόβλεψη φτάνει τα 5.8 δισεκατομμύρια δολλάρια, από τα 2,4 που ήταν το 2022, όπως δείχνει το Σχήμα 1.1. Η αυξανόμενη υιοθέτηση βιομετρικών στοιχείων φωνής και η δυνατότητα ανάλυσης της συνομιλίας σε πραγματικό χρόνο σε διάφορους κλάδους της βιομηχανίας έχει τονώσει τη ζήτηση της αγοράς. Οι πιο γνωστές εταιρείες που δραστηριοποιούνται στον κλάδο είναι μεταξύ άλλων οι κολοσσοί Amazon, Apple, Google LLC, IBM Corporation, Microsoft, οι οποίες προσφέρουν εξατομικευμένες λύσεις που ενσωματώνουν τεχνολογίες με συστήματα αναγνώρισης φωνής. Τα προϊόντα των παραπάνω διευκολύνουν άλλες εταιρείες και οργανισμούς να υιοθετήσουν λύσεις που ενισχύουν τα κέντρα επικοινωνίας τους με τους πελάτες, παρέχοντας ”ψηφιακό” εργατικό δυναμικό που βασίζεται στην τεχνητή νοημοσύνη αυξάνοντας κατά συνέπεια την παραγωγικότητά τους.



(a) Πρόβλεψη CAGR σε διαλογικά συστήματα **(b)** Πρόβλεψη CAGR σε συστήματα Speech-to-Text

Σχήμα 1.1: Πρόβλεψη δείκτη σύνθετου ρυθμού ετήσιας ανάπτυξης σε διαλογικά συστήματα [8] και αντίστοιχα σε συστήματα (σε μορφή API) μεταγραφής ομιλίας σε κείμενο 2022 - 2027 [9]

Τα πεδία χρήσης των συστημάτων αυτόματης αναγνώρισης ομιλίας (ASR) σήμερα είναι πολλά:

- Σε συστήματα αυτοκινήτου, όπου απλές φωνητικές εντολές μπορούν να χρησιμοποιηθούν για την έναρξη τηλεφωνικών κλήσεων, την επιλογή ραδιοφωνικών σταθμών ή την αναπαραγωγή μουσικής από συμβατό έξυπνο τηλέφωνο, συσκευή αναπαραγωγής mp3 ή μονάδα flash που περιέχει μουσική. Οι δυνατότητες αναγνώρισης φωνής διαφέρουν μεταξύ μάρκας και μοντέλου αυτοκινήτου. Μερικά από τα πιο πρόσφατα μοντέλα αυτοκινήτων προσφέρουν αναγνώριση ομιλίας σε φυσική γλώσσα αντί για ένα σταθερό σύνολο εντολών, επιτρέποντας στον οδηγό να χρησιμοποιεί πλήρεις προτάσεις και κοινές φράσεις. Επομένως, με τέτοια συστήματα δεν χρειάζεται ο χρήστης να απομνημονεύει ένα σύνολο σταθερών λέξεων εντολών.

- Στον τομέα της υγείας η αναγνώριση ομιλίας μπορεί να εφαρμοστεί στη διαδικασία της ιατρικής τεκμηρίωσης. Ο κλινικός γιατρός έχει τη δυνατότητα να υπαγορεύει στο σύστημα και να γίνεται από αυτό η καταγραφή του ιστορικού ενός ασθενούς, η εισαγωγή μετρήσεων από εξετάσεις (π.χ. αριθμητικές τιμές ή κώδικες από μια λίστα ή ένα ελεγχόμενο λεξιλόγιο) κι έτσι γίνεται πιο εύκολη η διατήρηση ενός σημαντικού όγκου δεδομένων αλλά και η πρόσβαση σε δομημένα ιατρικά έγγραφα. Τα περισσότερα ηλεκτρονικά αρχεία που υπάρχουν στα ιατρικά κέντρα δεν έχουν ρητά προσαρμοστεί για να εκμεταλλεύονται τις δυνατότητες αναγνώρισης φωνής, οπότε ένα μεγάλο μέρος της αλληλεπίδρασης των γιατρών με τέτοια συστήματα περιλαμβάνει πλοήγηση μέσω της διεπαφής χρήστη χρησιμοποιώντας μενού και κλικ καρτελών/κουμπιών και εξαρτάται σε μεγάλο βαθμό από το πληκτρολόγιο και το ποντίκι.

- Στον τομέα της στρατιωτικής άμυνας την τελευταία δεκαετία έχουν αφιερωθεί ουσιαστικές προσπάθειες για τη δοκιμή και την αξιολόγηση της αναγνώρισης ομιλίας σε μαχητικά αεροσκάφη και ελικόπτερα. Σε αυτά τα προγράμματα, οι συσκευές αναγνώρισης ομιλίας έχουν λειτουργήσει με επιτυχία με εφαρμογές που περιλαμβάνουν τη ρύθμιση ραδιοσυχνοτήτων, την εντολή ενός συστήματος αυτόματου πιλότου, τη ρύθμιση των συντεταγμένων του σημείου διεύθυνσης και των παραμέτρων απελευθέρωσης όπλων και τον έλεγχο της οθόνης πτήσης. Επιπλέον η εκπαίδευση για ελεγκτές εναέριας κυκλοφορίας (ATC) αντιπροσωπεύει μια εξαιρετική εφαρμογή για συστήματα αναγνώρισης ομιλίας. Πολλά συστήματα εκπαίδευσης ATC απαιτούν επί του παρόντος από ένα άτομο να ενεργεί ως «ψευδο-πιλότος», συμμετέχοντας σε φωνητικό διάλογο με τον εκπαιδευόμενο ελεγκτή, ο οποίος προσομοιώνει το διάλογο που θα έπρεπε να διεξάγει ο ελεγκτής με τους πιλότους σε μια πραγματική κατάσταση ATC. Οι τεχνικές αναγνώρισης ομιλίας και σύνθεσης προσφέρουν τη δυνατότητα να εξαλειφθεί η ανάγκη ενός ατόμου να ενεργεί ως ψευδοπιλότος, μειώνοντας έτσι το εκπαιδευτικό και το προσωπικό υποστήριξης. Θεωρητικά, οι εργασίες του ελεγκτή αέρα χαρακτηρίζονται επίσης από εξαιρετικά δομημένη ομιλία ως την κύρια έξοδο του ελεγκτή, επομένως θα πρέπει να είναι δυνατή η μείωση της δυσκολίας της εργασίας αναγνώρισης ομιλίας. Στην πράξη, αυτό συμβαίνει σπάνια αφού ο αριθμός των φράσεων που υποστηρίζονται από ένα από τα συστήματα αναγνώρισης ομιλίας προμηθευτών προσομοίωσης υπερβαίνει τις 500.000.

- Η αυτόματη αναγνώριση ομιλίας είναι πλέον κοινός τόπος στον τομέα της τηλεφωνίας και γίνεται όλο και πιο διαδεδομένος στον τομέα των ηλεκτρονικών παιχνιδιών και της προσομοίωσης. Στα συστήματα τηλεφωνίας τα ASR συστήματα χρησιμοποιούνται πλέον κυρίως σε τηλεφωνικά κέντρα ενσωματώνοντας και διαδραστικές φωνητικές πύλες. Η βελτίωση των ταχυτήτων του επεξεργαστή για κινητά έχει κάνει την αναγνώριση ομιλίας πρακτική στα στα έξυπνα τηλέφωνα, ενώ η ομιλία χρησιμοποιείται κυρίως ως μέρος μιας διεπαφής χρήστη, για τη δημιουργία προκαθορισμένων ή προσαρμοσμένων εντολών ομιλίας.

- Η αναγνώριση ομιλίας μπορεί να είναι χρήσιμη και στην εκπαίδευση για εκμάθηση γλώσσας, για να διδάξει τη σωστή προφορά, εκτός από το να βοηθήσει ένα άτομο να αναπτύξει ευχέρεια στις ομιλητικές του δεξιότητες. Ακόμη οι μαθητές που είναι τυφλοί ή έχουν πολύ χαμηλή όραση μπορούν να επωφεληθούν από τη χρήση της τεχνολογίας για την υπαγόρευση λέξεων και στη συνέχεια την απαγγελία από τον υπολογιστή, ή δίνοντας εντολές με τη φωνή τους, αντί να πρέπει να κοιτούν την οθόνη και το πληκτρολόγιο. Οι μαθητές μπορούν να γράφουν σχολικές εργασίες χρησιμοποιώντας προγράμματα ομιλίας σε κείμενο καθώς να χρησιμοποιήσουν την τεχνολογία αναγνώρισης ομιλίας για να πραγματοποιούν αναζητήσεις στο διαδίκτυο.

Αφού είδαμε ένα ευρύ φάσμα χρήσης των συστημάτων αυτόματης αναγνώρισης ομιλίας σε διαφορετικούς τομείς, επιλέξαμε στην παρούσα διπλωματική να κατασκευάσουμε ένα ASR σύστημα που βρίσκει εφαρμογή στον τομέα του αθλητισμού και συγκεκριμένα του ποδοσφαίρου. Το ποδόσφαιρο είναι ένα από τα πιο δημοφιλή αθλήματα παγκοσμίως και η ποσότητα του περιεχομένου που σχετίζεται με το άθλημα σε όλο τον κόσμο, συμπεριλαμβανομένων βίντεο, σχολίων ήχου, στατιστικών ομάδων/παικτών και βαθμολογιών είναι τεράστια και ταχέως αναπτυσσόμενη. Οι διάφορες μορφές

που συναντώνται (multimodality), ήχος, εικόνα και κείμενο παρουσιάζουν τεράστιο ενδιαφέρον για τους ραδιοτηλεοπτικούς φορείς αλλά και τους θαυμαστές, καθώς ένα μεγάλο ποσοστό κοινού προτιμά να ακολουθεί μόνο τα κύρια σημεία ενός παιχνιδιού. Ωστόσο, ο σχολιασμός σημαντικών γεγονότων και η παραγωγή κειμένου συχνά απαιτεί μεγάλο κόστος, εξοπλισμό και πολλή κουραστική, δυσκίνητη, χειρωνακτική εργασία. Ο λόγος για τον οποίο αξίζει να δημιουργήσουμε κείμενο απευθείας από ένα βίντεο ή ένα ηχητικό κομμάτι έγγειται στο ότι η γλώσσα έχει εκφραστικά πλεονεκτήματα, που βοηθούν τη δημιουργία ειδησεογραφικών άρθρων. Χρησιμοποιώντας λοιπόν τεχνολογίες Speech to Text (STT) παράγουμε κείμενο από ήχο με ελάχιστη προσπάθεια. Οι υπότιτλοι είναι πολύ πιο εύκολο να υποστούν επεξεργασία από ότι οι ροές βίντεο μεγαλύτερου μεγέθους και στη συνέχεια να χρησιμοποιηθούν για να τροφοδοτήσει άλλους σκοπούς, όπως αναζήτηση περιεχομένου και δημιουργία περιλήψεων.

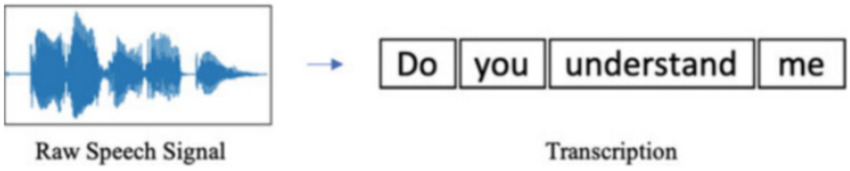
Κεφάλαιο 2

Συστήματα Αυτόματης Αναγνώρισης Ομιλίας (ASR)

2.1 Ορισμός

Στην επιστήμη της πληροφορικής, η αναγνώριση ομιλίας αναφέρεται στην μετατροπή προφερόμενων λέξεων σε κείμενο. Είναι επίσης αλλιώς γνωστή και ως αυτόματη αναγνώριση ομιλίας, υπολογιστική αναγνώριση ομιλίας ή speech-to-text (STT), δηλαδή (μετατροπή) από-ομιλία-σε-κείμενο. Κάποια συστήματα αναγνώρισης ομιλίας χρησιμοποιούν αναγνώριση ομιλίας ανεξάρτητη από τον ομιλήτη, ενώ άλλα χρησιμοποιούν ”εξάσκηση”, όπου ένα άτομο διαβάζει κομμάτια κειμένου σε ένα σύστημα αναγνώρισης ομιλίας. Τότε, αυτού του είδους τα συστήματα αναλύουν τη φωνή ενός ομιλητή και την χρησιμοποιούν για να προσαρμόσουν την αναγνώριση της ομιλίας του συγκεκριμένου ατόμου από τον υπολογιστή, με αποτέλεσμα την πιο ακριβή καταγραφή της. Συστήματα που δεν χρησιμοποιούν εξάσκηση ονομάζονται συστήματα ανεξάρτητα από τον ομιλητή. Οι εφαρμογές αναγνώρισης ομιλίας συμπεριλαμβάνουν τα Φωνητικά Περιβάλλοντα Χρήστη (Voice User Interfaces) όπως η φωνητική πληκτρολόγηση, ο έλεγχος των οικιακών ηλεκτρονικών συσκευών και συστημάτων, η διαδικτυακή αναζήτηση και πολλά άλλα. Ο όρος αναγνώριση ομιλίας δεν αναφέρεται τόσο στο ”ποιος” μιλάει, αλλά στο ”τι” λέει. Αναγνωρίζοντας επιπλέον όμως και την ταυτότητα του ομιλητή είναι δυνατόν να διευκολυνθεί η διαδικασία της μετάφρασης της ομιλίας του σε συστήματα που έχουν προηγουμένως εξασκηθεί στην φωνή του συγκεκριμένου ατόμου ή ακόμη είναι δυνατόν να πιστοποιηθεί ή αναγνωριστεί η ταυτότητα του ομιλητή σε συστήματα ασφαλείας.

Με απλά λόγια, ο ορισμός της αυτόματης αναγνώρισης λόγου (ASR) μπορεί να περιγραφεί ως εξής: με δεδομένη την είσοδο δειγμάτων ήχου X από ένα ηχογραφημένο σήμα ομιλίας εφαρμόζουμε μια συνάρτηση f ώστε να γίνει η αντιστοίχηση σε μια ακολουθία λέξεων W που αντιπροσωπεύουν τη μεταγραφή όσων ειπώθηκαν.



Σχήμα 2.1: Μετατροπή ψηφιοποιημένου σήματος ομιλίας σε αναγνώσιμο κείμενο (μεταγραφή)

$$\mathcal{W} = f(x) \tag{2.1}$$

Ωστόσο, η εύρεση μιας τέτοιας συνάρτησης είναι αρκετά δύσκολη και απαιτεί διαδοχικά μοντέλα για την παραγωγή της ακολουθίας των λέξεων. Αυτά τα μοντέλα πρέπει να είναι ανθεκτικά στο θόρυβο από τον περιβάλλοντα χώρο αλλά και το νοηματικό πλαίσιο. Για παράδειγμα, η ανθρώπινη ομιλία μπορεί να έχει οποιονδήποτε συνδυασμό χρονικής διακύμανσης (ταχύτητα), άρθρωση, προφορά, ένταση και φωνητικές παραλλαγές (ρασμώδης ή ρινική ομιλία) αλλά καταλήγουν στην ίδια μεταγραφή. Γλωσσικά, συναντώνται πρόσθετες μεταβλητές όπως η προσωδία (ανεβαίνει σε τονισμό

όταν θέτουμε μια ερώτηση), ο αυθόρμητος λόγος, που περιλαμβάνει συμπληρωματικές λέξεις (”χμ” ή ”εεε”), όπου όλες μπορούν να υποδηλώνουν διαφορετικά συναισθήματα ή υπονοούμενα, παρόλο που λέγονται με τα ίδια λόγια. Συνδυάζοντας αυτές τις μεταβλητές με πολλά περιβαλλοντικά σενάρια όπως ποιότητα ήχου, απόσταση μικροφώνου, αντήχηση συμπαιραίνουμε ότι όλα αυτά αυξάνουν εκθετικά την πολυπλοκότητα του την εργασία αναγνώρισης.

Η αυτόματη αναγνώριση ομιλίας είναι δύσκολη λόγω μιας ποικιλίας παραγόντων στον ακατέργαστο ήχο όπως η μοναδικότητα στην εκφορά του λόγου, των διφορούμενων ορίων των λέξεων και της έλλειψης συμφραζόμενων (context). Ας ρίξουμε μια ματιά σε αυτές τις προκλήσεις. Ο θόρυβος αναφέρεται σε τυχαίες διακυμάνσεις που αποκρύπτουν ή δεν περιέχουν σημαντικά δεδομένα ή άλλες πληροφορίες. Κατά την αναγνώριση θα πρέπει να μπορούν να απομονωθούν οι περιοχές των ηχητικών σημάτων από τις περιοχές άχρηστου θορύβου. Ως θόρυβος λογίζονται συνομιλίες στο παρασκήνιο, ”μικροφωνισμοί”, ήχοι όπως για παράδειγμα αεροπλάνα που πετούν, σκυλιά που γαβγίζουν και ούτω καθεξής. Η έκφραση/εκφορά αναφέρεται στον τρόπο με τον οποίο ένα άτομο προφέρει και αρθρώνει τις λέξεις. Διαφορετικοί άνθρωποι έχουν διαφορετικές εκφωνήσεις, όπως μεταβλητότητα στον τόνο, μεταβλητότητα στον όγκο και μεταβλητότητα στην ταχύτητα των λέξεων, και επομένως είναι δύσκολο να ληφθούν υπόψη αυτές οι πολλές διαφορές. Για παράδειγμα, οπτικά και ακουστικά μπορούμε να δούμε και να ακούσουμε ότι η «ταχύτητα» και η «ταχύυυυτητα» εκφωνούνται διαφορετικά. Είναι πιθανό το ύψος και ο όγκος να είναι επίσης διαφορετικά, οπότε αυτές οι αποκλίσεις πρέπει να ευθυγραμμιστούν και να αντιστοιχιστούν. Σε αντίθεση με τον γραπτό λόγο, ο προφορικός λόγος δεν έχει σαφή και καθορισμένα όρια μεταξύ λέξεων. Το γραπτό κείμενο έχει κενά, κόμματα και άλλες μορφές διαχωρισμού, στην ομιλία, οι λέξεις φαίνεται να επικαλύπτονται η μία μετά την άλλη. Γενικά, οι συνομιλίες ρέουν επειδή οι άνθρωποι είναι σε θέση να συμπληρώσουν τα κενά ανάλογα με τα συμφραζόμενα. Χωρίς πλαίσιο, είναι δύσκολο να γίνει διάκριση μεταξύ δύο τελείως διαφορετικών προτάσεων που μπορεί να ακούγονται όμοια.

2.2 Ιστορική Αναδρομή

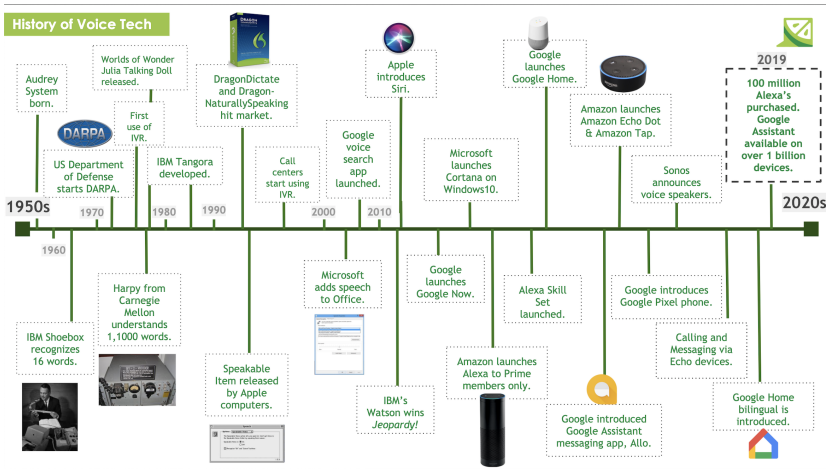
Παρόλο που πολλοί πιστεύουν ότι η τεχνολογία φωνής συγκαταλέγεται στις καινοτομίες της εποχής, η αλήθεια είναι ότι η μελέτη, ανάπτυξη και εφαρμογή τεχνολογιών αναγνώρισης φωνής και ομιλίας είναι ενεργή τα τελευταία 70 χρόνια. Στο Σχήμα 2.2 γίνεται επισκόπηση της ιστορίας της τεχνολογίας φωνής και του τρόπου με τον οποίο αναπτύχθηκε από τα πρώιμα στάδιά της μέχρι πρόσφατα. Το 1952, το πρώτο σύστημα αναγνώρισης ομιλίας που σχεδιάστηκε από τα Bell Laboratories ήταν γνωστό ως σύστημα «Audrey» και μπορούσε να αναγνωρίσει μόνο μεμονωμένα φωνητικά ψηφία που εκφωνούνταν δυνατά. Η μηχανή μπορούσε να καταλάβει τα ψηφία 0-9 όμως έπρεπε να προσαρμοστεί σε κάθε χρήστη προτού μπορέσει να καταγράψει την ομιλία του με ακρίβεια. Θεωρήθηκε ότι η Audrey θα μπορούσε να χρησιμοποιηθεί για τηλεφωνικές κλήσεις, ωστόσο δεν είχε τελικά απήχηση στη μαζική αγορά λόγω του μεγάλου μεγέθους της, των απαιτήσεων ισχύος και του κόστους παραγωγής και συντήρησης. Έπειτα από σχεδόν δέκα χρόνια η IBM παρουσίασε το ”Shoebox”, το οποίο ήταν σε θέση να κατανοήσει και να ανταποκριθεί σε 16 ομιλούμενες λέξεις στα αγγλικά καθώς και να κατανοήσει τους αριθμούς 0-9. Όπως η Audrey, έτσι κι αυτή η συσκευή προσπάθησε να αναγνωρίσει και να ενεργήσει με βάση τη συγκεκριμένη συχνότητα των φωνηέντων σε κάθε προφορικό ψηφίο.

Στη δεκαετία του 1970, η Υπηρεσία Προηγμένων Ερευνητικών Έργων του Υπουργείου Άμυνας των ΗΠΑ (DARPA) ξεκίνησε το πρόγραμμα έρευνας κατανόησης ομιλίας (SUR) που επικεντρώθηκε στην ανάπτυξη και έρευνα τεχνολογίας αναγνώρισης ομιλίας στο Πανεπιστήμιο Carnegie Mellon. Ο στόχος της DARPA ήταν να αναπτύξει μια τεχνολογία αναγνώρισης ομιλίας που θα μπορούσε να κατανοήσει έως και 1.000 λέξεις. Ως αποτέλεσμα της έρευνας και της εργασίας που διεξήχθη από το SUR κατά τη διάρκεια της δεκαετίας του 1970, ο Carnegie Mellon μπόρεσε να αναπτύξει το σύστημα ομιλίας «Harpy» το 1976, το οποίο κατανοούσε πάνω από 1.000 αγγλικές προφορικές λέξεις. Το Harpy μπορούσε να επεξεργαστεί ομιλία που ακολουθούσε προϋπάρχον λεξιλόγιο, προφορά και γραμματικούς κανόνες. Όπως και οι φωνητικοί βοηθοί που έγιναν διαθέσιμοι το 2018, το Harpy επέστρεφε ένα μήνυμα ”Δεν ξέρω τι είπατε, παρακαλώ επαναλάβετε” όταν δεν μπορούσε να καταλάβει

τον ομιλητή. Για ακόμη μια φορά, το Harpy είχε περιορισμένη ικανότητά να κατανοεί τη φυσική γλώσσα. Επιπλέον, στα τέλη της δεκαετίας του 1970, ξεκίνησε η πρώτη εμπορική εφαρμογή (διαδραστικής φωνητικής απόκρισης) IVR, που σχεδιάστηκε και αναπτύχθηκε από τον Steven Shmidt. Τα συστήματα IVR είναι αυτοματοποιημένα τηλεφωνικά συστήματα υπολογιστή που χρησιμοποιούν εξειδικευμένο υλικό για τον χειρισμό ψηφιοποιημένης φωνής σε κλήσεις.

Κατά τη διάρκεια της δεκαετίας του 1980 η ανάπτυξη των Κρυφών Μαρκοβιανών Μοντέλων (Hidden Markov Model - HMM) βοήθησε στην περαιτέρω έρευνα και ανάπτυξη της τεχνολογίας φωνής χρησιμοποιώντας στατιστικά στοιχεία για τον προσδιορισμό της πιθανότητας μιας λέξης να προέρχεται από έναν άγνωστο ήχο. Αυτή η στατιστική μέθοδος ήταν μια σημαντική ανακάλυψη επειδή αντί να χρησιμοποιεί απλώς λέξεις και να ψάχνει για ηχητικά μοτίβα, το HHM εκτίμησε την πιθανότητα οι άγνωστοι ήχοι να είναι λέξεις. Στη δεκαετία του 1990 σημειώθηκαν πολλές τεχνολογικές εξελίξεις, συμπεριλαμβανομένων των καταναλωτών που είχαν ευρύτερη πρόσβαση τόσο σε προσωπικούς υπολογιστές όσο και σε τεχνολογίες αναγνώρισης ομιλίας. Το DragonDictate, που αναπτύχθηκε από τον Δρ. James Baker, είναι το πρώτο προϊόν αναγνώρισης ομιλίας καταναλωτή που χρησιμοποιεί διακριτές μεθόδους υπαγόρευσης, οι οποίες απαιτούσαν από τον χρήστη να κάνει παύση μεταξύ κάθε προφορικής λέξης. Αργότερα το 1997, το Dragon NaturallySpeaking, το πρώτο προϊόν συνεχούς αναγνώρισης ομιλίας που διατίθεται για τους καταναλωτές, εισήλθε στην αγορά. Το Dragon NaturallySpeaking ήταν σε θέση να αναγνωρίσει και να μεταγράψει τη φυσική ανθρώπινη ομιλία, με ρυθμό περίπου 100 λέξεων ανά λεπτό. Το Dragon NaturallySpeaking δεν απαιτούσε από τους χρήστες να κάνουν παύση μεταξύ κάθε λέξης όπως έκανε το Dragon Dictate, έτσι με τη συνεχή πρόοδο στην αναγνώριση ομιλίας η Dragon έκανε πρακτική για πρώτη φορά τη χρήση της αναγνώρισης ομιλίας για τη δημιουργία εγγράφω. Το Dragon NaturallySpeaking είναι ακόμα διαθέσιμο για λήψη και χρησιμοποιείται περισσότερο από επαγγελματίες του ιατρικού τομέα. Επιπλέον, στη δεκαετία του 1990, τα τηλεφωνικά κέντρα άρχισαν να επενδύουν στην ενοποίηση τηλεφωνίας υπολογιστών (CTI) με συστήματα IVR, που ήταν η γέννηση της αυτοματοποιημένης τηλεφωνικής κλήσης. Αντικείμενα με δυνατότητα φωνητικής παραγωγής αποτελούν την πρώτη προσπάθεια ενσωματωμένης αναγνώρισης ομιλίας, οπότε δημιουργήθηκε και το αντίστοιχο λογισμικό ελέγχου με δυνατότητα φωνής για υπολογιστές Apple. Στις αρχές της δεκαετίας του 2000, η Microsoft κυκλοφόρησε υπολογιστές που παρείχαν παρόμοια δυνατότητα.

Από τη δεκαετία του 2010 έως τώρα, η ανάπτυξη της έρευνας, ανάπτυξης και εφαρμογής της τεχνολογίας φωνής έχει εκτοξευθεί στα ύψη. Η δεκαετία ξεκίνησε με το Watson της IBM, ένα σύστημα τηλεφωνητή/υπολογιστή ικανό να κατανοήσει τη φυσική γλώσσα, που κατάφερε να νικήσει στο παιχνίδι γνώσεων Jeopardy τον μεγάλο πρωταθλητή Ken Jennings. Αργότερα το ίδιο έτος, η Apple παρουσίασε το Siri σε όλες τις κινητές συσκευές της κι έτσι έγινε η αρχή για όλες τις εταιρείες να κυκλοφορήσουν τις δικές τους τεχνολογίες και συσκευές αναγνώρισης ομιλίας και φυσικής γλώσσας. Το 2013, η Microsoft παρουσίασε την Cortana, έναν εικονικό βοηθό παρόμοιο με το Siri, που και την έθεσε σε εφαρμογή σε όλες τις συσκευές Windows. Το 2015 η Amazon παρουσίασε τη συσκευή Alexa στις Ηνωμένες Πολιτείες, ενώ ένα χρόνο αργότερα η Google έβγαλε στην αγορά το Google Home. Τώρα, βλέπουμε ότι οι εικονικοί βοηθοί με ομιλία είναι συνηθισμένοι τόσο στα σπίτια όσο και στα αυτοκίνητά μας. Το 2020, η βάση χρηστών έξυπνων ηχείων στις ΗΠΑ αυξήθηκε κατά 32% από το προηγούμενο έτος, φθάνοντας σε 87,7 εκατομμύρια ενήλικες ενώ οι συνολικοί χρήστες φωνητικών βοηθών στο αυτοκίνητο ανέρχονται συνολικά σε σχεδόν 130 εκατομμύρια στις ΗΠΑ, με 83,8 εκατομμύρια χρήστες ενεργούς μηνιαίως.



Σχήμα 2.2: Χρονοδιάγραμμα στην τεχνολογία αναγνώρισης φωνής [10]

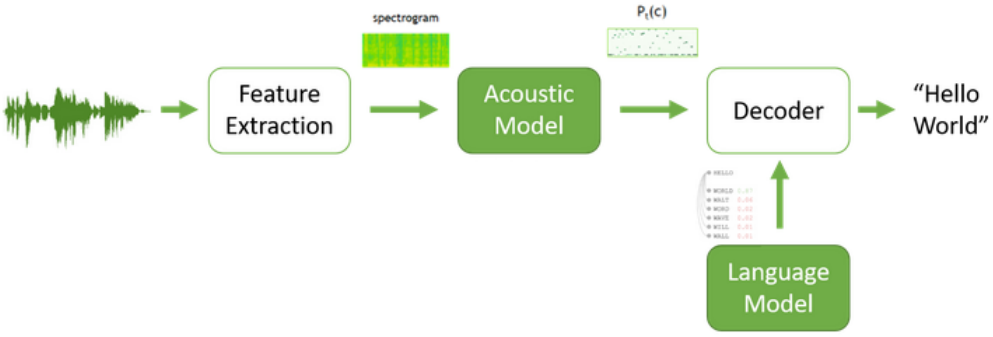
2.3 Μέρη ενός συστήματος ASR

Όπως είπαμε ο κύριος στόχος ενός συστήματος ASR είναι να μεταμορφώσει ένα σήμα εισόδου $x = (x_1, x_2, ...x_T)$ με συγκεκριμένο μήκος σε μια αλληλουχία λέξεων ή χαρακτήρων (ταμπέλες - labels) $y = (y_1, y_2, ...y_T)$ με $y_n \in \mathbb{V}$, όπου $\mathbb{V}$ είναι το λεξιλόγιο. Τα labels μπορεί να είναι σε επίπεδο χαρακτήρα, για παράδειγμα γράμματα, ή σε επίπεδο ολόκληρων λέξεων. Η πιο πιθανή συμβολοσειρά δίνεται από τη φόρμουλα:

$$\hat{y} = \underset{y \in V}{\operatorname{argmax}}\, p(y|x) \tag{2.2}$$

Ένα τυπικό σύστημα ASR έχει τα ακόλουθα βήματα κατασκευής, που φαίνονται και γραφικά στο σχήμα:

- Προεπεξεργασία και Εξαγωγή χαρακτηριστικών
- Ακουστικό Μοντέλο
- Αποκωδικοποίηση
- Γλωσσικό Μοντέλο



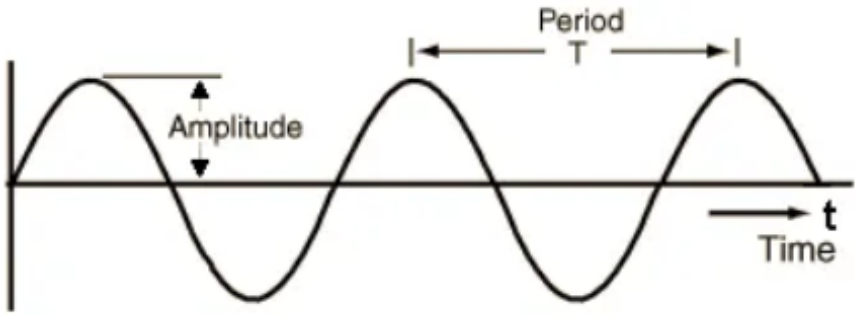
Σχήμα 2.3: Διακριτά μέρη ενός ASR [12]

2.3.1 Προεπεξεργασία και Εξαγωγή χαρακτηριστικών

Το βήμα προεπεξεργασίας στοχεύει στη βελτίωση του σήματος ήχου μειώνοντας την αναλογία σήματος προς θόρυβο, μειώνοντας το θόρυβο και φιλτράροντας το σήμα. Γενικά, τα χαρακτηριστικά που χρησιμοποιούνται εξάγονται με συγκεκριμένο αριθμό τιμών ή συντελεστών (coefficients), οι οποίοι δημιουργούνται με την εφαρμογή διαφόρων μεθόδων στην είσοδο. Αυτό το βήμα είναι κρίσιμο, όσον αφορά διάφορους ποιοτικούς παράγοντες, όπως ο θόρυβος ή το φαινόμενο ηχούς. Έτσι τα ακουστικά χαρακτηριστικά πρέπει να είναι αρκετά περιγραφικά ώστε να παρέχουν χρήσιμες πληροφορίες σχετικά το σήμα, καθώς και αρκετά ανθεκτικά στις πολλές διαταραχές που μπορεί να προκύψουν στο περιβάλλον. Η πλειοψηφία των μεθόδων που χρησιμοποιούνται στην κατασκευή ASR υιοθετούν τις 2 τεχνικές εξαγωγής: i) συντελεστές Mel (MFCCs) ii) διακριτός μετασχηματισμός κυμάτων (DWT).

Παραγωγή ήχου

Ένα ηχητικό σήμα παράγεται από τις διακυμάνσεις της πίεσης του αέρα. Μπορούμε να μετρήσουμε την ένταση των διακυμάνσεων της πίεσης και να σχεδιάσουμε αυτές τις μετρήσεις με την πάροδο του χρόνου. Τα ηχητικά σήματα επαναλαμβάνονται συχνά σε τακτά χρονικά διαστήματα, έτσι ώστε κάθε κύμα να έχει το ίδιο σχήμα. Το ύψος δείχνει την ένταση του ήχου και είναι γνωστό ως πλάτος. Ο χρόνος που απαιτείται για να ολοκληρώσει το σήμα ένα πλήρες κύμα είναι η περίοδος, όπως φαίνεται και στο σχήμα 2.4. Ο αριθμός των κυμάτων που δημιουργεί το σήμα σε ένα δευτερόλεπτο ονομάζεται συχνότητα. Η συχνότητα είναι η αντίστροφη της περιόδου και η μονάδα συχνότητας είναι τα Hertz. Η πλειοψηφία των ήχων που συναντάμε μπορεί να μην ακολουθούν τόσο απλά και κανονικά περιοδικά μοτίβα. Αλλά σήματα διαφορετικών συχνοτήτων μπορούν να προστεθούν μαζί για να δημιουργήσουν σύνθετα σήματα με πιο πολύπλοκα επαναλαμβανόμενα μοτίβα. Όλοι οι ήχοι που ακούμε, συμπεριλαμβανομένης της ανθρώπινης φωνής μας, αποτελούνται από τέτοιες κυματομορφές. Το ανθρώπινο αυτί είναι σε θέση να διακρίνει διαφορετικούς ήχους με βάση την «ποιότητα» του ήχου που είναι επίσης γνωστός ως ηχόχρωμα.



Σχήμα 2.4: Επαναλαμβανόμενο σήμα με πλάτος συναρτήσει του χρόνου

Η ανθρώπινη ομιλία δημιουργείται από τη φωνητική οδό και διαμορφώνεται με τη γλώσσα, τα δόντια και τα χείλη (συχνά αναφέρονται ως αρθρωτές) ως εξής:

- Ο αέρας ωθείται προς τα πάνω από τους πνεύμονες και δονεί τις φωνητικές χορδές παράγοντας σχεδόν περιοδικούς ήχους.
- Ο αέρας ρέει στον φάρυγγα, τη ρινική και στοματική κοιλότητα.
- Διάφοροι αρθρωτές ρυθμίζουν τα κύματα του αέρα.
- Ο αέρας διαφεύγει από το στόμα και τη μύτη.

Η ανθρώπινη ομιλία συνήθως περιορίζεται στο εύρος 85 Hz–8 kHz, ενώ η ανθρώπινη ακοή περιορίζεται στην περιοχή των 20 Hz με 20 kHz.

Πρωτόλεια κυματομορφή

Τα κύματα της πίεσης του αέρα που παράγονται μετατρέπονται σε τάση μέσω μικροφώνου και πραγματοποιείται δειγματοληψία με μετατροπέα αναλογικού σε ψηφιακό. Η έξοδος της διαδικασίας εγγραφής είναι ένας 1-διάστατος πίνακας αριθμών που αντιπροσωπεύει τα διακριτά δείγματα της μετατροπής. Το ψηφιοποιημένο σήμα έχει τρεις κύριες ιδιότητες: ρυθμός δειγματοληψίας, αριθμός των καναλιών και ακρίβεια (μερικές φορές αναφέρεται ως βάθος bit). Ο ρυθμός δειγματοληψίας είναι η συχνότητα με την οποία γίνεται δειγματοληψία του αναλογικού σήματος (σε Hertz), ο αριθμός των καναλιών αναφέρεται στη λήψη ήχου με πολλαπλές πηγές μικροφώνου. Για παράδειγμα ήχος μονού καναλιού αναφέρεται ως μονοφωνικός ήχος, ενώ ο όρος stereo αναφέρεται σε ήχο δύο καναλιών. Πρόσθετα κανάλια - πολυκαναλικόw ήχοw - μπορεί να είναι χρήσιμα για φιλτράρισμα σήματος σε απαιτητικά ακουστικά περιβάλλοντα. Η ακρίβεια ή το βάθος είναι ο αριθμός των bit ανά δείγμα, που αντιστοιχεί στην ανάλυση της πληροφορίας. Ο τυπικός ήχος τηλεφώνου έχει ρυθμό δειγματοληψίας 8 kHz και ακρίβεια 16 bit, ενώ η ποιότητα ενός δίσκου (CD) είναι 44,1 kHz, με ακρίβεια 16 bit, η σύγχρονη επεξεργασία ομιλίας εστιάζει στα 16 kHz ή υψηλότερα. Μερικές φορές ο ρυθμός bit χρησιμοποιείται για τη μέτρηση της συνολικής ποιότητας του ήχου που υπολογίζεται ως:

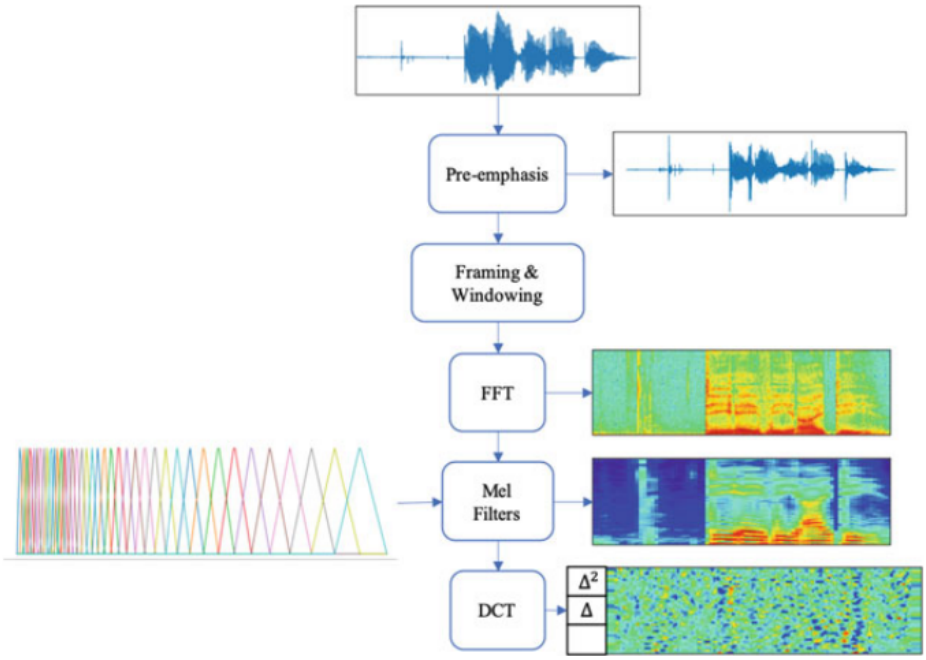
$$bit\,rate = sample\,rate * precision * number\,of\,channels \tag{2.3}$$

Το μη επεξεργασμένο σήμα ομιλίας είναι υψηλών διαστάσεων και είναι δύσκολο να μοντελοποιηθεί. Τα περισσότερα ASR συστήματα βασίζονται σε χαρακτηριστικά που εξάγονται από το ηχητικό σήμα για να μειώσουν τις διαστάσεις και φιλτράροντας τα ανεπιθύμητα σήματα. Πολλά από αυτά τα χαρακτηριστικά προέρχονται από κάποια μορφή φασματικής ανάλυσης που μετατρέπει το ηχητικό σήμα σε ένα σύνολο χαρακτηριστικών και ενισχύουν τα σήματα που μιμούνται το ανθρώπινο αυτί. Πολλές από αυτές τις μεθόδους εξαρτώνται από τον μετασχηματισμό Fourier (Short Time Fourier transform - STFT) του ηχητικού σήματος χρησιμοποιώντας τον αλγόριθμο FFT (Fast Fourier transform), διάφορα φίλτρων ή κάποιο συνδυασμό των δύο.

Συντελεστές Συχνότητας Mel (Mel Frequency Cepstral Coefficients - MFCC)

Οι συντελεστές συχνότητας Mel είναι οι πιο συχνά χρησιμοποιούμενοι σε ASR συστήματα και η επιτυχία τους βασίζεται στην ικανότητά τους να εκτελούν παρόμοιους τύπους φιλτραρίσματος με το ανθρώπινο ακουστικό σύστημα και τη χαμηλή του διάσταση. Υπάρχουν επτά βήματα για τον υπολογισμό των χαρακτηριστικών MFCCs, όπως φαίνεται και στο Σχήμα 2.5 και είναι παρόμοια για τις περισσότερες τεχνικές δημιουργίας χαρακτηριστικών, με κάποια μεταβλητότητα στους τύπους των φίλτρων που χρησιμοποιούνται. Ακολουθούνται τα εξής βήματα:

1. Προέμφαση
2. Framing
3. Windowing
4. Fast Fourier Transform
5. Mel Filter Bank
6. Διακριτός μετασχηματισμός συνημιτόνου (Discrete Cosine Transform - DCT)
7. Δέλτα ενέργεια και φάσμα δέλτα (Delta Energy and Delta Spectrum)



Σχήμα 2.5: Βήματα εξαγωγής συντελεστών MFCC

Προ-έμφαση (Pre-emphasis)

Η προ-έμφαση είναι το πρώτο βήμα στη δημιουργία χαρακτηριστικών MFCC. Στην παραγωγή λόγου (και της επεξεργασίας σήματος γενικά), η ενέργεια των σημάτων υψηλότερης συχνότητας τείνει να είναι χαμηλότερη. Κατά τη διαδικασία της προ-έμφασης εφαρμόζεται ένα φίλτρο στο σήμα εισόδου που δίνει έμφαση στα πλάτη των υψηλότερων συχνοτήτων και μειώνει τα πλάτη των χαμηλότερων ”ομάδων” συχνοτήτων.

Framing

Το ακουστικό σήμα αλλάζει διαρκώς στην ομιλία. Η μοντελοποίηση αυτού του μεταβαλλόμενου σήματος γίνεται με διαχωρισμό μικρών τμημάτων που λαμβάνονται από τον ήχο και λογίζονται ως στατικά. Framing είναι η διαδικασία διαχωρισμού των δειγμάτων από τον ακατέργαστο ήχο σε τμήμα σταθερού μήκους και ονομάζονται πλαίσια (frames). Αυτά τα τμήματα μετατρέπονται στη διάσταση της συχνότητας με χρήση του αλγορίθμου FFT, που αποδίδει μια αναπαράσταση της ισχύος των συχνοτήτων όσο διαρκεί το κάθε πλαίσιο. Τα τμήματα δηλώνουν τα όρια μεταξύ των φωνητικών αναπαραστάσεων του λόγου. Οι φωνητικοί ήχοι που σχετίζονται με την ομιλία τείνουν να είναι στο εύρος των 5–100 ms, επομένως το μήκος των καρέ επιλέγεται συνήθως ώστε να λαμβάνει αυτό υπόψιν. Τυπικά, τα πλαίσια είναι περίπου 20 ms για τα περισσότερα συστήματα ASR, με επικάλυψη 10 ms για τα καρέ μας.

Windowing

Με την ”παραθυροποίηση” τα δείγματα πολλαπλασιάζονται με μια συνάρτηση κλιμάκωσης. με σκοπό την εξομάλυνση των δυνητικά απότομων επιπτώσεων του framing που μπορεί να προκαλέσουν έντονες διαφορές στις ”άκρες” των πλαισίων. Εφαρμόζοντας συναρτήσεις παραθύρου στα δείγματα όμως μειώνει τις αλλαγές σε όλο το τμήμα ώστε να μειώσει τα σήματα κοντά στις άκρες του πλαισίου που μπορεί να έχουν όχι και τόσο επιθυμητά αποτελέσματα μετά την εφαρμογή του FFT. Πολλές λειτουργίες παραθύρου μπορούν να εφαρμοστούν σε ένα σήμα, η πιο συχνά χρησιμοποιούμενη σε

συστήματα ASR είναι το παράθυρο Hann:

$$w(n) = 0.5(1 - \cos(\frac{2\pi n}{N-1})) = \sin^2 \frac{\pi n}{N-1} \tag{2.4}$$

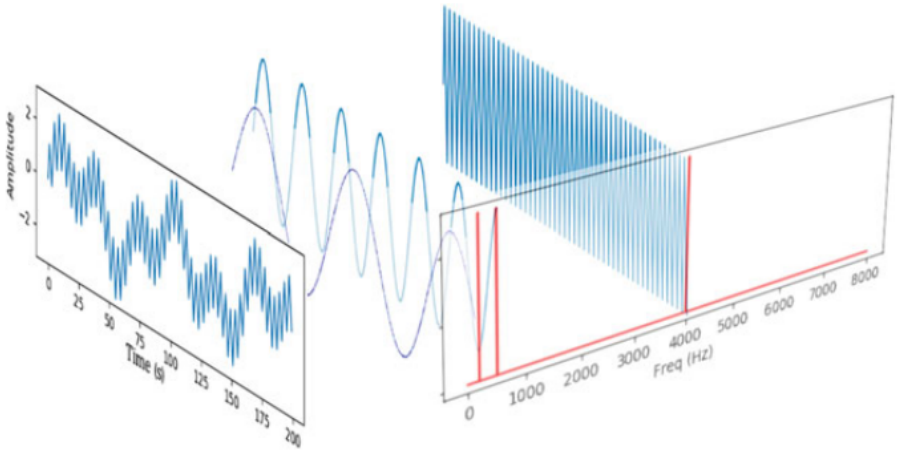
και το παράθυρο Hamming:

$$w(n) = 0.54 - 0.46 \cos(\frac{2\pi n}{N-1}) \tag{2.5}$$

όπου N είναι το μέγεθος του παραθύρου και $0 \leq n \leq N-1$.

Fast Fourier Transform

Ο βραχυπρόθεσμος(short-time) μετασχηματισμός Fourier (STFT) μετατρέπει το μονοδιάστατο σήμα από τον τομέα του χρόνου στον τομέα συχνότητας χρησιμοποιώντας τα πλαίσια και εφαρμόζοντας τον διακριτό μετασχηματισμό Fourier (DFT) στο καθένα από αυτά. Μια απεικόνιση της μετατροπής DFT εμφανίζεται στο σχήμα 2.6. Οι γρήγοροι μετασχηματισμοί Fourier (FFT) είναι ένας αποτελεσματικός αλγόριθμος για τον υπολογισμό του DFT υπό κατάλληλες συνθήκες και είναι σύνηθες για ASR συστήματα.



Σχήμα 2.6: Το επιθυμητό αποτέλεσμα του FFT σε ένα σήμα εισόδου (αριστερά) και η κανονικοποιημένη έξοδος του FFT στη διάσταση της συχνότητας (δεξιά)

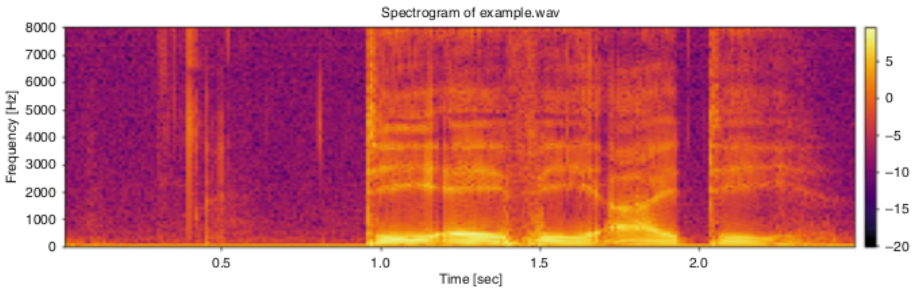
Το φασματογράφημα είναι ένας τρισδιάστατος οπτικός μετασχηματισμός FFT του ακουστικού σήματος και είναι συχνά ένα πολύτιμο σύνολο χαρακτηριστικών. Η αναπαράσταση STFT πλεονεκτεί γιατί κάνει τις λιγότερες υποθέσεις σχετικά με το σήμα ομιλίας (εκτός από την ακατέργαστη κυματομορφή). Για ορισμένα συστήματα από άκρο σε άκρο (end-to-end) το φασματογράφημα χρησιμοποιείται ως είσοδος, επειδή περιγράφει τη συχνότητα με υψηλότερη ανάλυση.Το Σχήμα 2.7, αναπαριστά το χρόνο κατά μήκος του άξονα x, τα τμήματα (buckets) της συχνότητας στον άξονα y, και την ένταση αυτής της συχνότητας στον άξονα z, που συνήθως αναπαρίσταται χρωματικά. Το πλάτος υπολογίζεται ως εξής:

$$S_m = |FFT(x_i)|^2 \tag{2.6}$$

Η ενέργεια του φασματογραφήματος είναι μερικές φορές πιο χρήσιμη γιατί κανονικοποιεί το πλάτος με βάση τον αριθμό των σημείων που λαμβάνονται υπόψιν:

$$S_p = \frac{|FFT(x_i)|^2}{N} \tag{2.7}$$

όπου N είναι ο αριθμός των σημείων από τον υπολογισμό του FFT, κι είναι συνήθως 256 ή 512.



Σχήμα 2.7: Φασματογράφημα αρχείου ήχου σε λογαριθμική κλίμακα

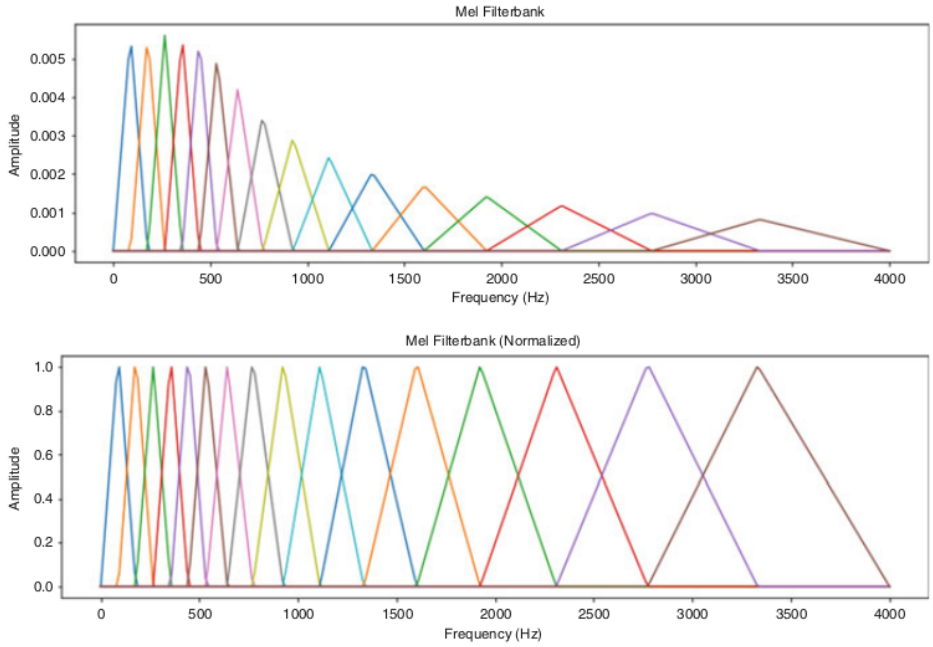
Οι περισσότερες από τις σημαντικές συχνότητες βρίσκονται στο χαμηλότερο τμήμα του φάσματος, επομένως το φασματογράφημα αντιστοιχίζεται τυπικά στην λογαριθμική κλίμακα.

Mel Filter Bank

Τα χαρακτηριστικά που δημιουργούνται από τον μετασχηματισμό STFT του ήχου στοχεύουν στην προσομοίωση των μετατροπών που γίνονται από το ανθρώπινο ακουστικό σύστημα. Η τράπεζα φίλτρων Mel είναι ένα σύνολο ζωνοπερατών φίλτρων που μιμούνται το ανθρώπινο ακουστικό σύστημα. Αντί να ακολουθούν γραμμική κλίμακα, αυτά τα τριγωνικά φίλτρα δρουν λογαριθμικά σε υψηλότερες συχνότητες και γραμμικά σε χαμηλότερες συχνότητες, κάτι που είναι χαρακτηριστικό στα σήματα ομιλίας, όπως φαίνεται και στο Σχήμα 2.8. Η τράπεζα φίλτρων έχει συνήθως 40 φίλτρα. Η μετατροπή μεταξύ των περιοχών Mel (m) και Hertz (f) γίνεται μέσω των παρακάτω εξισώσεων:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.8}$$

$$f = 700(10^{\frac{m}{2595}} - 1) \tag{2.9}$$



Σχήμα 2.8: Mel filter bank shown with 16 filters. The filters are applied to the input signal to produce the Mel-scale output

Κάθε ένα από τα φίλτρα παράγει μια έξοδο που είναι το σταθμισμένο άθροισμα των φασματικών συχνοτήτων που αντιστοιχούν σε κάθε φίλτρο. Αυτές οι τιμές αντιστοιχίζουν τις συχνότητες εισόδου στην κλίμακα Mel.

Διακριτός μετασχηματισμός συνημιτόνου (Discrete Cosine Transform - DCT)

Ο διακριτός μετασχηματισμός συνημιτόνου (DCT) αντιστοιχίζει τα χαρακτηριστικά της κλίμακας Mel στο χρόνο. Η συνάρτηση DCT είναι παρόμοια με έναν μετασχηματισμό Fourier αλλά χρησιμοποιεί μόνο πραγματικούς αριθμούς (ο μετασχηματισμός Fourier παράγει μιγαδικούς αριθμού) συμπιέζει τα δεδομένα εισόδου σε ένα σύνολο συνημιτονικών συντελεστών που περιγράφουν τις ταλαντώσεις στη συνάρτηση. Η έξοδος αυτής της μετατροπής αναφέρεται ως MFCC.

Δέλτα ενέργεια και φάσμα δέλτα (Delta Energy and Delta Spectrum)

Η ενέργεια δέλτα και το φάσμα δέλτα (επίσης γνωστά ως «δέλτα δέλτα» ή «διπλό δέλτα») είναι χαρακτηριστικά που παρέχουν πληροφορίες σχετικά με την κλίση της μετάβασης μεταξύ των πλαισίων. Τα χαρακτηριστικά της δέλτα ενέργειας είναι η διαφορά μεταξύ των συντελεστών διαδοχικών πλαισίων (το τρέχον και το προηγούμενο πλαίσιο). Τα χαρακτηριστικά του φάσματος δέλτα είναι η διαφορά μεταξύ διαδοχικών χαρακτηριστικών της ενέργειας δέλτα (το τρέχον και το προηγούμενο δέλτα ενεργειας των χαρακτηριστικών). Οι εξισώσεις για τον υπολογισμό της ενέργειας δέλτα και του φάσματος δέλτα είναι:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{2.10}$$

$$dd_t = \frac{\sum_{n=1}^{N} n(d_{t+n} - d_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{2.11}$$

Άλλα ακουστικά χαρακτηριστικά

Πολλά ακουστικά χαρακτηριστικά έχουν προταθεί όλα αυτά τα χρόνια, εφαρμόζοντας διαφορετικά φίλτρα και μετασχηματισμούς για να τονίσουν διάφορες πτυχές του ακουστικού φάσματος. Πολλές από αυτές τις προσεγγίσεις βασίστηκαν σε χαρακτηριστικά κατασκευασμένα χειροκίνητα, όπως τα MFCCs, τα gammatone χαρακτηριστικά και οι γραμμικοί προγνωστικοί συντελεστές, ωστόσο τα MFCCs παραμένουν οι πιο δημοφιλείς. Ένα από τα μειονεκτήματα των MFCCs (ή οποιουδήποτε σετ χαρακτηριστικών που έχει σχεδιαστεί με μη αυτόματο τρόπο) είναι η ευαισθησία στο θόρυβο λόγω της εξάρτησής τους από τη φασματική μορφή. Σε χαμηλές διάστασεις του χώρου χαρακτηριστικών ήταν ιδιαίτερα ευεργετική με τις παλαιότερες τεχνικές μηχανικής μάθησης όμως με προσεγγίσεις βαθιάς μάθησης, όπως τα συνελικτικά νευρωνικά δίκτυα, χαρακτηριστικά υψηλότερης ανάλυσης μπορούν να χρησιμοποιηθούν ή ακόμα και να μαθευτούν. Γενικά τα MFCCs είναι εύκολα υπολογίσιμα, εφαρμόζουν χρήσιμα φίλτρα σε ASR συστήματα και απο-συσχετίζουν τα χαρακτηριστικά μεταξύ τους. Μερικές φορές συνδυάζονται με επιπλέον χαρακτηριστικά που σχετίζονται με τα ηχεία (συνήθως i-vectors) για τη βελτίωση της ευρωστίας του μοντέλου.

Αυτόματη εκμάθηση

Έχουν γίνει διάφορες προσπάθειες να μάθουμε τις αναπαραστάσεις των χαρακτηριστικών άμεσα, αντί να βασιζόμαστε σε χαρακτηριστικά που εξάγουμε, τα οποία μπορεί να μην είναι τα καλύτερα για την ελαχιστοποίηση της μετρικής αξιολόγησης Word Error Rate που χρησιμοποιούμε στα συστήματα

αναγνώρισης ομιλίας. Μερικές από τις προσεγγίσεις περιλαμβάνουν: επιβλεπόμενη μάθηση χαρακτηριστικών με βαθιά νευρωνικά δίκτυα (DNN), συνελικτικά δίκτυα (CNN) σε ακατέργαστη ομιλία, συνδυασμός CNN-DNN ή ακόμη και μάθηση χωρίς επίβλεψη με χρήση Restricted Boltzmann Machines (RBM). Οι λειτουργίες αυτόματης εκμάθησης βελτιώνουν την ποιότητα σε συγκεκριμένα σενάρια, αλλά μπορούν επίσης να είναι περιοριστικά σε πολλές περιπτώσεις. Τα χαρακτηριστικά που παράγονται με την επιβλεπόμενη εκπαίδευση μαθαίνουν να διακρίνουν μεταξύ των παραδειγμάτων στο σύνολο δεδομένων και μπορεί να περιοριστεί σε περιβάλλοντα που δεν υπάρχουν παρατηρήσεις. Με την εισαγωγή των end-to-end μοντέλων για ASR συστήματα αυτά τα χαρακτηριστικά συντονίζονται, κάνοντας ευκολότερη τη διαδικασία εκπαίδευσης δύο σταδίων.

Φωνήματα και Γραφήματα (phonemes & graphemes)

Στη Φυσική Επεξεργασία Γλώσσας (NLP) η πιο λογική γλωσσική αναπαράσταση στο μετασχηματισμό της ομιλίας σε κείμενο είναι οι λέξεις, ακριβώς επειδή η μεταγραφή σε επίπεδο λέξης είναι ο επιθυμητός στόχος και υπάρχει νόημα που άπτεται σε αυτό το επίπεδο. Ωστόσο στην πράξη τα σύνολα δεδομένων ομιλίας τείνουν να έχουν λίγα μεταγραφόμενα παραδείγματα ανά λέξη, καθιστώντας δύσκολη τη μοντελοποίηση. Μια κοινή αναπαράσταση είναι επιθυμητή ώστε να αποκτηθούν επαρκή δεδομένα εκπαίδευσης για την ποικιλία των λέξεων, όσο αυτό είναι δυνατόν.

Ο ρυθμός, η προφορά και το περιβάλλον του ομιλητή παίζουν σημαντικό ρόλου στον τρόπο αντιστοίχισης της ροής ήχου σε μια ακολουθία εξόδου. Το σύστημα ASR εστιάζει στην αναγνώριση παρά στην ερμηνεία, δηλαδή στην ακριβή αναγνώριση των προφορικών λέξεων παρά της ακολουθίας λέξεων που εξαρτάται από τα συμφραζόμενα, η οποία όμως είναι μια σημαντική πτυχή. Στην περίπτωση των ομόφωνων, δηλαδή δύο λέξεων με την ίδια φωνητική παράσταση αλλά διαφορετική ορθογραφία, η πρόβλεψη της σωστής λέξης βασίζεται εξ' ολοκλήρου στο νοηματικό πλαίσιο. Σε αυτήν την περίπτωση, μερικά από τα ζητήματα μπορούν να ξεπεραστούν με ένα γλωσσικό μοντέλο, που θα εξετάσουμε αργότερα. Οι λάθος φωνητικές αντικαταστάσεις περιπλέκουν περαιτέρω τα πράγματα, όπως για παράδειγμα, στα αγγλικά οι αναπαραστάσεις του pin (πείρος) [P IH N] και του pen (στυλό) [P EH N] είναι διακριτές. Ωστόσο, αν και αυτές οι λέξεις έχουν διαφορετικές φωνητικές αναπαραστάσεις, συχνά γίνεται λάθος ή προφέρονται παρόμοια, με αποτέλεσμα η σωστή επιλογή να εξαρτάται από το νόημα περισσότερο από τα ίδια τα φωνήματα. Με τις διαφορετικές προφορές οι φωνητικές αναπαραστάσεις μπορεί να είναι ακόμα περισσότερο προβληματικές, απαιτώντας εναλλακτικές μεθόδους για τον προσδιορισμό τους. Αυτού του είδους τα σενάρια είναι ζωτικής σημασίας στα συστήματα ASR, γιατί ενώ πολλές φορές οι άνθρωποι μπορεί να πουν τη λάθος λέξη όμως το πλαίσιο και η πρόθεση μπορούν ακόμα να ερμηνευθούν. Όλοι αυτοί οι παράγοντες του προφορικού λόγου συμβάλλουν στην πολυπλοκότητα της αυτόματης αναγνώρισης ομιλίας.

Παρόλο λοιπόν που υπάρχουν δυσκολίες στην αναγνώριση ομιλίας, υπάρχουν φυσικές ιδιότητες που βοηθούν, για παράδειγμα, μπορούμε να χωρίσουμε την ομιλία σε βασικές μονάδες όπως γραφήματα και φωνήματα. Τα φωνήματα είναι διακριτοί ήχοι στη γλώσσα, που υποδηλώνονται με κάθετες, π.χ. /ah/. Με άλλα λόγια, είναι η μικρότερη μονάδα ήχου. Υπάρχουν περίπου 40 φωνήματα για τα αγγλικά, (44 φωνήματα για τα αγγλικά του Ηνωμένου Βασιλείου). Τα φωνήματα βοηθούν τον κάθε ομιλητή να προφέρει τους φθόγγους οι οποίοι, καθώς εκφέρονται ο ένας μετά τον άλλο σχηματίζουν τις λέξεις. Έτσι, για να πούμε στα νέα ελληνικά τη λέξη «πόδι», χρησιμοποιούμε τους φθόγγους [p][ό][δ][i], ενώ, για να πούμε τη λέξη «παιδί», χρησιμοποιούμε τους φθόγγους [p][e][δ][í].

Συνολικά οι φθόγγοι της νέας ελληνικής είναι 23:

```
[α] [ε] [ι] [ο] [ου]

[κ] [π] [τ] [γκ] [μπ] [ντ] [ν] [μ] [σ] [ζ]
[ρ] [λ] [θ] [φ] [χ] [δ] [β] [γ]
```

Στους παραπάνω φθόγγους μερικοί μελετητές προσθέτουν και τους [τσ][τζ] κι έτσι οι φθόγγοι γίνονται 25. Τα διεθνή σύμβολα για τους ελληνικούς φθόγγους είναι τα εξής:

```
[α=a] [ε=e] [ι=i] [ο=o] [ου=u]

[κ=k] [π=p] [τ=t] [γκ=ng] [μπ=mb/b] [ντ=nd/d]
[ν=n] [μ=m] [σ=s] [ζ=z] [ρ=r] [λ=l] [θ=θ]
[φ=φ] [χ=x] [δ=δ] [β=v] [γ=γ] [τσ=ts][τζ=dz]
```

Η εναλλαγή ενός φωνήματος με ένα άλλο αλλάζει τη σημασία της λέξης, αν και αυτό μπορεί να μην ισχύει για τα ίδια φωνήματα σε άλλη γλώσσα. Για παράδειγμα στα αγγλικά, εάν το τρίτο φώνημα στη λέξη sweet [swit] (γλυκός) αλλάξει από [i] σε [E], η σημασία ολόκληρης της λέξης αλλάζει: [swEt] (ιδρώτας). Τα γραφήματα είναι διακριτοί χαρακτήρες στη γλώσσα, που υποδηλώνονται μέσα σε αγκύλες, π.χ. <a>. Είναι η μικρότερη μονάδα ενός συστήματος γραφής. Στα αγγλικά, υπάρχουν 250 γραφήματα, αλλά το μικρότερο γράφημα είναι το σύνολο των αλφαβητικών γραμμάτων (A-Z) συν ο χαρακτήρας διαστήματος <space>.

Δεν υπάρχει αντιστοίχιση 1 προς 1 μεταξύ φωνημάτων → γραφηματων ή γραφηματων → φωνημάτων. Πολλαπλά φωνήματα αντιστοιχίζονται στο ίδιο γράφημα και πολλαπλά γραφήματα αντιστοιχίζονται στο ίδιο φώνημα. Στα αγγλικά το γράμμα c αντιστοιχεί σε διαφορετικούς ήχους ανάλογα με τη λέξη, δηλαδή:

```
<C> → /K/ για τη λέξη cat (γάτα)
<C> → /CH/ για τη λέξη chat (συνομιλία)
<C> → /S/ για τη λέξη ceremony (τελετή)
```

Επίσης, ο ήχος «IY» στη λέξη «speech», μπορεί να σχηματιστεί με διαφορετικές ορθογραφίες, δηλαδή:

```
/IY/ → <EI> για τη λέξη receive (λαμβάνω)
/IY/ → <IE> για τη λέξη believe (πιστεύω)
/IY/ → <EE> για τη λέξη speech (ομιλία)
```

Αν και τα φωνήματα δεν μπορούν να αντιστοιχιστούν απευθείας στα γραφήματα, αποτελούν ένα χρήσιμο ενδιάμεσο βήμα για την επεξεργασία της ομιλίας, που ονομάζεται λεκτική αποκωδικοποίηση. Εάν μπορούμε να αποκωδικοποιήσουμε επιτυχώς την έξοδο ενός ακουστικού μοντέλου σε φωνήματα, τότε μπορούμε να αντιστοιχίσουμε αυτά τα φωνήματα στις λέξεις τους χρησιμοποιώντας ένα λεξικό (χαρτογράφηση λεξικού) στα βήματα μετά την επεξεργασία για να σχηματίσουμε λέξεις και προτάσεις. Επιπλέον, αυτό το βήμα μετά την επεξεργασία μας επιτρέπει να μειώσουμε την πολυπλοκότητα αντιστοιχίζοντας σε ένα σταθερό σύνολο τιμών. Για προβλήματα υψηλής διάστασης και προβλήματα με μεγάλο λεξιλόγιο, τα φωνήματα μπορεί να είναι χρήσιμα για να μειώσουν δραστικά τον αριθμό των συγκρίσεων μεταξύ των λέξεων. Ωστόσο, εάν το πρόβλημα έχει μικρό μέγεθος λεξιλογίου, μπορεί κανείς να επιλέξει να χρησιμοποιήσει το ακουστικό μοντέλο για απευθείας αντιστοίχιση με λέξεις και να παραλείψει αυτό το ενδιάμεσο βήμα.

2.3.2 Ακουστικά Μοντέλα

Το ακουστικό μοντέλο χρησιμοποιείται στην αυτόματη αναγνώριση ομιλίας για να αναπαραστήσει τη σχέση μεταξύ ενός ηχητικού σήματος και των φωνημάτων ή άλλων γλωσσικών μονάδων που συνθέτουν την ομιλία. Το μοντέλο μαθαίνει από ένα σύνολο ηχογραφήσεων και τις αντίστοιχες μεταγραφές τους και χρησιμοποιώντας λογισμικό για τη δημιουργία στατιστικών αναπαραστάσεων των ήχων που απαρτίζουν κάθε λέξη.Τα κλασικά συστήματα αναγνώρισης ομιλίας κατασκευάζονται χρησιμοποιώντας ακουστικά μοντέλα όπως τα Γκαουσιανά Μεικτά Μοντέλα (Gaussian Mixture Model - GMM) και τα Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Model - HMM). Ένα τυπικό σύστημα αναγνώρισης ομιλίας που βασίζεται σε ακουστικά μοντέλα HMM ή GMM έχει δύο σημαντικές εργασίες, την αναγνώριση φωνημάτων και την αποκωδικοποίηση λέξεων. Αυτή η διαδικασία δύο βημάτων εξάγει χρήσιμα χαρακτηριστικά από τη βάση του σήματος ομιλίας σε προηγούμενα δεδομένα και

χρησιμοποιεί διακριτικά μοντέλα για να εκτιμήσει την πιθανότητα κάθε φωνήματος. Οι εταιρείες που δραστηριοποιούνται και παράγουν ASR προϊόντα χρησιμοποιούν έναν συνδυασμό Αναγνώρισης ομιλίας και Επεξεργασίας Φυσικής Γλώσσας για να αναλύσουν τη φωνή των πελατών και να αποκριθούν σε αυτήν με τη μορφή ήχου και κειμένου. Η αναγνώριση ομιλίας προετοιμάζει τη φωνητική εισαγωγή (ομιλία), έτσι ώστε τα εκπαιδευμένα μοντέλα NLP να μπορούν να εφαρμοστούν στα δεδομένα και να μπορούν να επιτύχουν τα αναμενόμενα αποτελέσματα. Τα συστήματα βαθιάς μάθησης μπορούν να αντιστοιχίσουν τα ακουστικά χαρακτηριστικά στα προφορικά φωνήματα απευθείας και έτσι είναι γνωστά ως συστήματα από άκρη σε άκρη (end-to-end ASR). Τα ακουστικά μοντέλα που δημιουργούνται με χρήση βαθιών νευρωνικών δικτύων παρέχουν καλύτερη ακρίβεια ταξινόμησης. Οι τρεις κύριοι τύποι τέτοιων συστημάτων από άκρο σε άκρο είναι: με χρήση δικτύων που βασίζονται στην προσοχή (attention-based), τα δίκτυα που εκπαιδεύονται μέσω της μεθόδου της συνδετικής χρονικής ταξινόμησης (Connectionist Temporal Classification - CTC) και τα συνελικτικά μοντέλα (CNN).

2.3.3 Γλωσσικά Μοντέλα

Η μοντελοποίηση γλώσσας (Language Modeling - LM) είναι η χρήση διαφόρων στατιστικών και πιθανολογικών τεχνικών για τον προσδιορισμό της πιθανότητας μιας δεδομένης ακολουθίας λέξεων να εμφανίζεται σε μια πρόταση. Τα γλωσσικά μοντέλα αναλύουν δεδομένα κειμένου για να παρέχουν μια βάση για τις προβλέψεις των λέξεων. Χρησιμοποιούνται σε εφαρμογές επεξεργασίας φυσικής γλώσσας (NLP), ιδιαίτερα σε αυτές που δημιουργούν κείμενο ως έξοδο. Μερικές από αυτές τις εφαρμογές περιλαμβάνουν την αυτόματη μετάφραση (machine translation) και την απάντηση ερωτήσεων (question answering). Τα γλωσσικά μοντέλα λοιπόν καθορίζουν την πιθανότητα λέξης αναλύοντας δεδομένα κειμένου. Ερμηνεύουν αυτά τα δεδομένα τροφοδοτώντας τα μέσω ενός αλγόριθμου που καθορίζει κανόνες για το νοηματικό πλαίσιο στη φυσική γλώσσα. Στη συνέχεια, το μοντέλο εφαρμόζει αυτούς τους κανόνες για να προβλέψει ή να δημιουργήσει με ακρίβεια νέες προτάσεις. Το μοντέλο ουσιαστικά μαθαίνει τα χαρακτηριστικά της βασικής γλώσσας και τα χρησιμοποιεί για να κατανοήσει νέες φράσεις. Υπάρχουν πολλές διαφορετικές πιθανολογικές προσεγγίσεις για τη μοντελοποίηση της γλώσσας, οι οποίες ποικίλλουν ανάλογα με τον σκοπό του γλωσσικού μοντέλου. Από τεχνική άποψη, οι διάφοροι τύποι διαφέρουν ως προς τον όγκο των δεδομένων κειμένου που αναλύουν και τα μαθηματικά που χρησιμοποιούν για την ανάλυσή τους. Για παράδειγμα, ένα μοντέλο γλώσσας που έχει σχεδιαστεί για τη δημιουργία προτάσεων για ένα αυτοματοποιημένο ρομπότ στο Twitter μπορεί να χρησιμοποιεί διαφορετικά μαθηματικά και να αναλύει δεδομένα κειμένου με διαφορετικό τρόπο από ένα μοντέλο γλώσσας που έχει σχεδιαστεί για τον προσδιορισμό της πιθανότητας ενός ερωτήματος αναζήτησης.

Μερικοί συνήθεις τύποι στατιστικής μοντελοποίησης γλώσσας είναι:

- Unigram: είναι ο απλούστερος τύπος γλωσσικού μοντέλου. Δεν εξετάζει κανένα νοηματικό πλαίσιο ή υπολογισμό συνθήκης πιθανότητας στους υπολογισμούς του, αντίθετα αξιολογεί κάθε λέξη ή όρο ανεξάρτητα. Τα μοντέλα unigram συνήθως χρησιμοποιούνται σε εργασίες επεξεργασίας γλώσσας, όπως η ανάκτηση πληροφοριών (information retrieval - IR). Το unigram είναι το θεμέλιο μιας πιο συγκεκριμένης παραλλαγής μοντέλου που ονομάζεται μοντέλο πιθανότητας ερωτήματος (query likelihood model), το οποίο χρησιμοποιεί την ανάκτηση πληροφοριών για να εξετάσει μια ομάδα εγγράφων και να αντιστοιχίσει το πιο σχετικό με ένα συγκεκριμένο ερώτημα.

- N-grams: είναι μια σχετικά απλή προσέγγιση στα γλωσσικά μοντέλα. Δημιουργούν μια κατανομή πιθανότητας για μια ακολουθία n, όπου το n μπορεί να είναι οποιοσδήποτε αριθμός και ορίζει το μέγεθος της ακολουθίας λέξεων στις οποίες εκχωρείται μια πιθανότητα. Για παράδειγμα, εάν n = 5, ένα 5-gram μπορεί να μοιάζει με αυτό: ”μπορείτε να με καλέσετε παρακαλώ”. Στη συνέχεια, το μοντέλο εκχωρεί πιθανότητες χρησιμοποιώντας ακολουθίες μεγέθους n, άρα ουσιαστικά, το n μπορεί να θεωρηθεί ως το μέγεθος του ”πλαισίου” που καλείται να εξετάσει το μοντέλο. Μερικοί τύποι n-gram είναι το μονόγραμμα, που ειδαμε παραπάνω, τα διγράμματα, τα τρίγραμμα και ούτω καθεξής.

- Αμφίδρομος τύπος (bidirectional): Σε αντίθεση με τα μοντέλα n-gram, τα οποία αναλύουν κείμενο προς μία κατεύθυνση (προς τα πίσω), τα αμφίδρομα μοντέλα αναλύουν κείμενο και προς τις δύο κατευθύνσεις, προς τα πίσω και προς τα εμπρός. Αυτά τα μοντέλα μπορούν να προβλέψουν οποιαδήποτε λέξη σε μια πρόταση ή σώμα κειμένου χρησιμοποιώντας κάθε άλλη λέξη στο κείμενο. Η αμφίδρομη εξέταση κειμένου αυξάνει την ακρίβεια των αποτελεσμάτων. Αυτός ο τύπος χρησιμοποιείται συχνά σε εφαρμογές μηχανικής εκμάθησης και παραγωγής ομιλίας. Για παράδειγμα, η Google χρησιμοποιεί ένα αμφίδρομο μοντέλο για την επεξεργασία ερωτημάτων αναζήτησης.

- Εκθετικός τύπος (exponential): γνωστό και ως μοντέλο μέγιστης εντροπίας, αυτός ο τύπος είναι πιο περίπλοκος από τον n-grams. Με απλά λόγια, το μοντέλο αξιολογεί το κείμενο χρησιμοποιώντας μια εξίσωση που συνδυάζει συναρτήσεις χαρακτηριστικών με n-grams. Ουσιαστικά καθορίζει χαρακτηριστικά και παραμέτρους των επιθυμητών αποτελεσμάτων και σε αντίθεση με τον n-grams, αφήνει τις παραμέτρους ανάλυσης να είναι πιο διφορούμενες, δηλαδή για παράδειγμα δεν καθορίζει μεμονωμένα μέγεθος του n. Το μοντέλο βασίζεται στην αρχή της εντροπίας, η οποία δηλώνει ότι η κατανομή πιθανοτήτων με τη μεγαλύτερη εντροπία είναι η καλύτερη επιλογή. Με άλλα λόγια, το μοντέλο με τη μεγαλύτερη αταξία (ή χάος) και το μικρότερο περιθώριο για υποθέσεις, είναι το πιο ακριβές. Τα εκθετικά μοντέλα σχεδιάζονται μεγιστοποιώντας τη διασταυρούμενη εντροπία (cross entropy), η οποία ελαχιστοποιεί στατιστικά τον αριθμό των υποθέσεων που μπορούν να γίνουν. Αυτό επιτρέπει στους χρήστες να εμπιστεύονται καλύτερα τα αποτελέσματα που έχουν από αυτά τα μοντέλα.

- Τύπος συνεχούς χώρου (continuous space): αυτός ο τύπος μοντέλου αναπαριστά τις λέξεις ως έναν μη γραμμικό συνδυασμό βαρών σε ένα νευρωνικό δίκτυο. Η διαδικασία της ανάθεσης βάρους σε μια λέξη είναι επίσης γνωστή ως ενσωματωμένα διανύσματα λέξης (word embeddings). Αυτός ο τύπος γίνεται ιδιαίτερα χρήσιμος καθώς τα σύνολα δεδομένων γίνονται ολοένα και μεγαλύτερα, επειδή τα μεγαλύτερα σύνολα δεδομένων συχνά περιλαμβάνουν περισσότερες μοναδικές λέξεις. Η παρουσία πολλών μοναδικών ή σπάνια χρησιμοποιούμενων λέξεων μπορεί να προκαλέσει προβλήματα σε ένα γραμμικό μοντέλο όπως το n-gram. Αυτό συμβαίνει επειδή ο αριθμός των πιθανών ακολουθιών λέξεων αυξάνεται και τα μοτίβα που ενημερώνουν τα αποτελέσματα γίνονται πιο αδύναμα. Με τη στάθμιση των λέξεων με μη γραμμικό, κατανεμημένο τρόπο, αυτό το μοντέλο μπορεί να «μάθει» να προσεγγίζει λέξεις και επομένως να μην παραπλανηθεί από άγνωστες τιμές. Η «κατανόησή» του για μια δεδομένη λέξη δεν είναι τόσο στενά συνδεδεμένη με τις άμεσες γύρω λέξεις όπως είναι στα μοντέλα n-gram.

Τα μοντέλα που αναφέρονται παραπάνω είναι πιο γενικές στατιστικές προσεγγίσεις από τις οποίες προκύπτουν πιο συγκεκριμένες παραλλαγές γλωσσικών μοντέλων. Για παράδειγμα, όπως αναφέρεται στην περιγραφή n-gram, το μοντέλο πιθανοτήτων ερωτήματος είναι ένα πιο συγκεκριμένο ή εξειδικευμένο μοντέλο που χρησιμοποιεί την προσέγγιση n-gram. Οι τύποι μοντέλων μπορούν να χρησιμοποιηθούν σε συνδυασμό μεταξύ τους. Τα μοντέλα που παρατίθενται διαφέρουν επίσης σημαντικά ως προς την πολυπλοκότητα. Σε γενικές γραμμές, τα πιο σύνθετα γλωσσικά μοντέλα είναι καλύτερα στις εργασίες που έχουν να κάνουν με Επεξεργασίας Φυσικής Γλώσσας, επειδή η ίδια η γλώσσα είναι εξαιρετικά περίπλοκη και συνεχώς εξελισσόμενη. Επομένως, ένα εκθετικό μοντέλο ή ένα μοντέλο συνεχούς χώρου μπορεί να είναι καλύτερο από ένα n-gram, επειδή έχουν σχεδιαστεί για να εξηγούν την ασάφεια και την ποικιλομορφία στη γλώσσα. Ένα καλό γλωσσικό μοντέλο θα πρέπει επίσης να μπορεί να επεξεργάζεται μακροπρόθεσμες εξαρτήσεις, να χειρίζεται λέξεις που μπορεί να αντλούν το νόημά τους από άλλες λέξεις που εμφανίζονται σε μακρινά, ανόμοια μέρη του κειμένου. Θα πρέπει, επιπλέον, να είναι σε θέση να καταλάβει πότε μια λέξη αναφέρεται σε μια άλλη λέξη που απέχει μεγάλη απόσταση, σε αντίθεση με το να βασίζεται πάντα σε εγγύς λέξεις μέσα σε ένα συγκεκριμένο σταθερό ιστορικό, πράγμα που απαιτεί ένα πιο σύνθετο μοντέλο.

Η μοντελοποίηση γλώσσας είναι ζωτικής σημασίας στις σύγχρονες εφαρμογές NLP, αφού είναι ο λόγος που οι μηχανές μπορούν να κατανοήσουν ποιοτικές πληροφορίες. Κάθε τύπος γλωσσικού

μοντέλου, με τον ένα ή τον άλλο τρόπο, μετατρέπει τις ποιοτικές πληροφορίες σε ποσοτικές πληροφορίες. Αυτό επιτρέπει στους ανθρώπους να επικοινωνούν με τις μηχανές όπως κάνουν μεταξύ τους σε περιορισμένο βαθμό. Χρησιμοποιείται απευθείας σε μια ποικιλία βιομηχανιών, όπως η τεχνολογία, τα οικονομικά, η υγειονομική περίθαλψη, οι μεταφορές, τα νομικά, τα στρατιωτικά κ.ά. Επιπλέον, οι περισσότεροι άνθρωποι έχουν αλληλεπιδράσει με κάποιο γλωσσικό μοντέλο κάποια στιγμή της ημέρας, είτε μέσω της αναζήτησης Google, είτε μέσω μιας λειτουργίας αυτόματης συμπλήρωσης κειμένου είτε μέσω της ενασχόλησης με έναν φωνητικό βοηθό. Οι ρίζες της μοντελοποίησης της γλώσσας όπως υπάρχει σήμερα μπορούν να εντοπιστούν πίσω στο 1948. Εκείνη τη χρονιά, ο Claude Shannon δημοσίευσε μια εργασία με τίτλο "A Mathematical Theory of Communication". Σε αυτό, περιέγραψε λεπτομερώς τη χρήση ενός στοχαστικού μοντέλου που ονομάζεται αλυσίδα Markov για τη δημιουργία ενός στατιστικού μοντέλου για τις ακολουθίες γραμμάτων στο αγγλικό κείμενο. Αυτή η εργασία είχε μεγάλο αντίκτυπο στη βιομηχανία των τηλεπικοινωνιών, έθεσε τις βάσεις για τη θεωρία της πληροφορίας και τη μοντελοποίηση της γλώσσας. Το μοντέλο Markov εξακολουθεί να χρησιμοποιείται σήμερα και τα n-grams είναι συνδεδεμένα πολύ στενά με την ιδέα.

Χρήση και παραδείγματα Γλωσσικών Μοντέλων

Τα γλωσσικά μοντέλα αποτελούν τη ραχοκοκαλιά της επεξεργασίας φυσικής γλώσσας (NLP). Ορισμένες εφαρμογές των γλωσσικών μοντέλων είναι οι εξής:

- η αναγνώριση ομιλίας (speech recognition) όπου μια μηχανή που μπορεί να επεξεργάζεται ήχο και χρησιμοποιείται συνήθως από βοηθούς φωνής όπως η Siri και η Alexa.

- η αυτόματη μετάφραση (machine translation) που περιλαμβάνει τη μετάφραση μιας γλώσσας σε μια άλλη, για παράδειγμα το Google Translate και το Microsoft Translator αλλά και το SDL Government, το οποίο χρησιμοποιείται για τη μετάφραση ροών κοινωνικής δικτύωσης σε πραγματικό χρόνο για την κυβέρνηση των ΗΠΑ.

- η προσθήκη ετικετών σε μέρη του λόγου (parts-of-speech-tagging), που περιλαμβάνει τη σήμανση και την κατηγοριοποίηση των λέξεων με βάση ορισμένα γραμματικά χαρακτηριστικά. Χρησιμοποιείται στη μελέτη της γλωσσολογίας και είναι γνωστή εφαρμογή από τη μελέτη του Brown Corpus, ενός σώματος κειμένου που αποτελείται από τυχαία αγγλική πεζογραφία που σχεδιάστηκε ώστε να μελετηθεί από υπολογιστές. Αυτό το σώμα έχει χρησιμοποιηθεί για την εκπαίδευση πολλών σημαντικών μοντέλων γλώσσας, συμπεριλαμβανομένου ενός που χρησιμοποιείται από την Google για τη βελτίωση της ποιότητας αναζήτησης.

- η ανάλυση κειμένου (parsing) οποιασδήποτε σειράς δεδομένων ή πρότασης που συμμορφώνεται με τυπικούς κανόνες γραμματικής και σύνταξης. Στη μοντελοποίηση γλώσσας, αυτό μπορεί να λάβει τη μορφή διαγραμμάτων προτάσεων που απεικονίζουν τη σχέση κάθε λέξης με τις άλλες. Οι εφαρμογές ορθογραφικού ελέγχου χρησιμοποιούν μοντελοποίηση και ανάλυση γλώσσας.

- η ανάλυση συναισθήματος (sentiment analysis) που περιλαμβάνει τον προσδιορισμό του συναισθήματος πίσω από μια δεδομένη φράση. Συγκεκριμένα, μπορεί να χρησιμοποιηθεί για την κατανόηση απόψεων και στάσεων που εκφράζονται σε ένα κείμενο. Οι επιχειρήσεις μπορούν να το χρησιμοποιήσουν για να αναλύσουν κριτικές προϊόντων ή γενικές αναρτήσεις σχετικά με το προϊόν τους, καθώς και να αναλύσουν εσωτερικά δεδομένα, όπως έρευνες εργαζομένων και συνομιλίες υποστήριξης πελατών. Το μοντέλο BERT της Google χρησιμοποιείται επίσης για ανάλυση συναισθήματος.

- η οπτική αναγνώριση χαρακτήρων (optical character recognition - OCR) περιλαμβάνει τη χρήση ενός μηχανήματος για τη μετατροπή εικόνων σε κείμενο κωδικοποιημένο από μηχανή. Η εικόνα μπορεί να είναι ένα σαρωμένο έγγραφο ή φωτογραφία με κείμενο κάπου, σε μια πινακίδα για παράδειγμα. Χρησιμοποιείται συχνά στην εισαγωγή δεδομένων κατά την επεξεργασία παλαιών

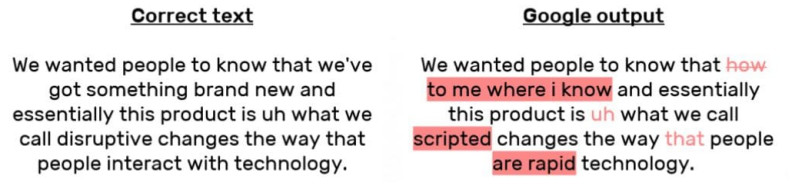
εγγράφων σε χαρτί που πρέπει να ψηφιοποιηθούν, ενώ μπορεί να χρησιμοποιηθεί για την ανάλυση και την αναγνώριση δειγμάτων γραφής.

- η ανάκτηση πληροφοριών (information retrieval - IR) που περιλαμβάνει την αναζήτηση πληροφοριών σε ένα έγγραφο, την αναζήτηση εγγράφων γενικά και την αναζήτηση μεταδεδομένων που αντιστοιχούν σε ένα έγγραφο. Τα προγράμματα περιήγησης ιστού είναι οι πιο κοινές εφαρμογές ανάκτησης πληροφοριών.

2.4 Αξιολόγηση και Μέτρηση Απόδοσης

Για την αξιολόγηση ενός συστήματος αναγνώρισης φωνής χρησιμοποιούνται διάφορες μετρικές, με την πιο κοινή να είναι το ποσοστό σφάλματος λέξης (Word Error Rate - WER). Η γενική δυσκολία μέτρησης της απόδοσης έγκειται στο γεγονός ότι η αναγνωρισμένη ακολουθία λέξεων μπορεί να έχει διαφορετικό μήκος από την ακολουθία αναφοράς. Η μετρική WER έχει συνάφεια με την απόσταση Levenshtein, όμως διαφέρει διότι υπολογίζεται στο επίπεδο της λέξης αντί για το επίπεδο του φωνήματος. Η απόσταση Levenshtein μετρά τις διαφορές μεταξύ δύο λέξεων και είναι ο ελάχιστος αριθμός επεξεργασιών ενός χαρακτήρα (εισαγωγές, διαγραφές ή αντικαταστάσεις) που απαιτούνται για την αλλαγή μιας λέξης σε μια άλλη, άρα η μετρική WER ορίζεται ως η κανονικοποιημένη απόσταση επεξεργασίας Levenshtein. Η κανονικοποιημένη απόσταση επεξεργασίας μεταξύ X και Y, d(X, Y) ορίζεται ως το ελάχιστο των W(P) / L (P), όπου P είναι μια διαδρομή επεξεργασίας μεταξύ X και Y, W (P) είναι το άθροισμα των βαρών των στοιχειωδών πράξεων επεξεργασίας του P, και το L(P) είναι ο αριθμός αυτών των πράξεων (μήκος του P).

Το WER είναι ένα πολύτιμο εργαλείο για τη σύγκριση διαφορετικών συστημάτων καθώς και για την αξιολόγηση των βελτιώσεων σε ένα σύστημα. Αυτό το είδος μέτρησης, ωστόσο, δεν παρέχει λεπτομέρειες σχετικά με τη φύση των μεταφραστικών σφαλμάτων και ως εκ τούτου απαιτείται περαιτέρω εργασία για τον εντοπισμό της κύριας πηγής τους. Αυτό το πρόβλημα επιλύεται ευθυγραμμίζοντας πρώτα την αναγνωρισμένη ακολουθία λέξεων με την ακολουθία λέξεων αναφοράς χρησιμοποιώντας δυναμική στοίχιση συμβολοσειρών. Επίσης δεν υπάρχει διαφοροποίηση μεταξύ των λέξεων που είναι απαραίτητες για το νόημα της πρότασης και εκείνων που δεν είναι τόσο σημαντικές. Δεν λαμβάνεται υπόψη εάν δύο λέξεις διαφέρουν μόνο κατά έναν χαρακτήρα ή αν διαφέρουν εντελώς.



Σχήμα 2.9: Παράδειγμα μέτρησης του Word Error Rate[24]

Το WER ορίζεται ως ο αριθμός των σφαλμάτων διαιρεμένος με τον συνολικό αριθμό των λέξεων στην ακολουθία αναφοράς. Όσο χαμηλότερη είναι η τιμή τόσο μεγαλύτερη η ακρίβεια του συστήματος, για παράδειγμα, ένα WER 20% σημαίνει ότι η μεταγραφή είναι 80% ακριβής. Τα σφάλματα παρουσιάζονται σε τρεις μορφές: αντικαταστάσεις, εισαγωγές και διαγραφές.

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{2.12}$$

όπου S είναι ο αριθμός των αντικαταστάσεων, D των διαγραφών, I των εισαγωγών, C των σωστών λέξεων και N ο αριθμός των λέξεων στο κείμενο αναφοράς (N = S + D + C).

Κατά την αναφορά της απόδοσης ενός συστήματος αναγνώρισης ομιλίας, μερικές φορές χρησιμοποιείται αντ' αυτού η ακρίβεια λέξης (WAcc):

$$WA_{cc} = 1 - WER = \frac{N - S - D - I}{N} = \frac{C - I}{N} \tag{2.13}$$

Πριν από τον υπολογισμό του WER, είναι σύνηθες να εφαρμόζουμε μερικούς μετασχηματισμούς στην πρόβλεψη του μοντέλου και/ή στο κείμενο αναφοράς για να προσπαθήσουμε να ελαχιστοποιήσουμε τις αρνητικές επιπτώσεις των διαφορών μορφοποίησης μεταξύ του σώματος εκπαίδευσης του μοντέλου και των δεδομένων δοκιμής. Αυτή η διαδικασία είναι γνωστή ως ”κανονικοποίηση κειμένου”. Τυπικοί μετασχηματισμοί μπορεί να είναι η αφαίρεση πεζών γραμμάτων και σημείων στίξης, επιπλέον των πιο προφανών αντιστοιχίσεων πολλών προς ένα που λειτουργούν σε εκφωνήσεις κειμένου όπως ψηφία, διευθύνσεις, νομίσματα κ.λπ. Με αυτούς τους μετασχηματισμούς παρατηρείται σημαντική αύξηση στην απόδοση των μοντέλων. Υπάρχουν πολλοί λόγοι που προκαλούν τα σφάλματα μεταγραφής ομιλίας σε κείμενο, εκ των οποίων οι 5 πιο κοινές που επηρεάζουν το WER στα σύγχρονα συστήματα ASR είναι:

1. Προφορές και παραλλαγές στο ρυθμό ομιλίας.
2. Ομόφωνα, ομόγραφα και ομώνυμα.
3. Crosstalk γνωστός και ως επικαλυπτόμενος διάλογος.
4. Ποιότητα ήχου και θόρυβος στον περιβάλλοντα χώρο.
5. Ακρωνύμια και ειδική ορολογία του κλάδου.

Κεφάλαιο 3

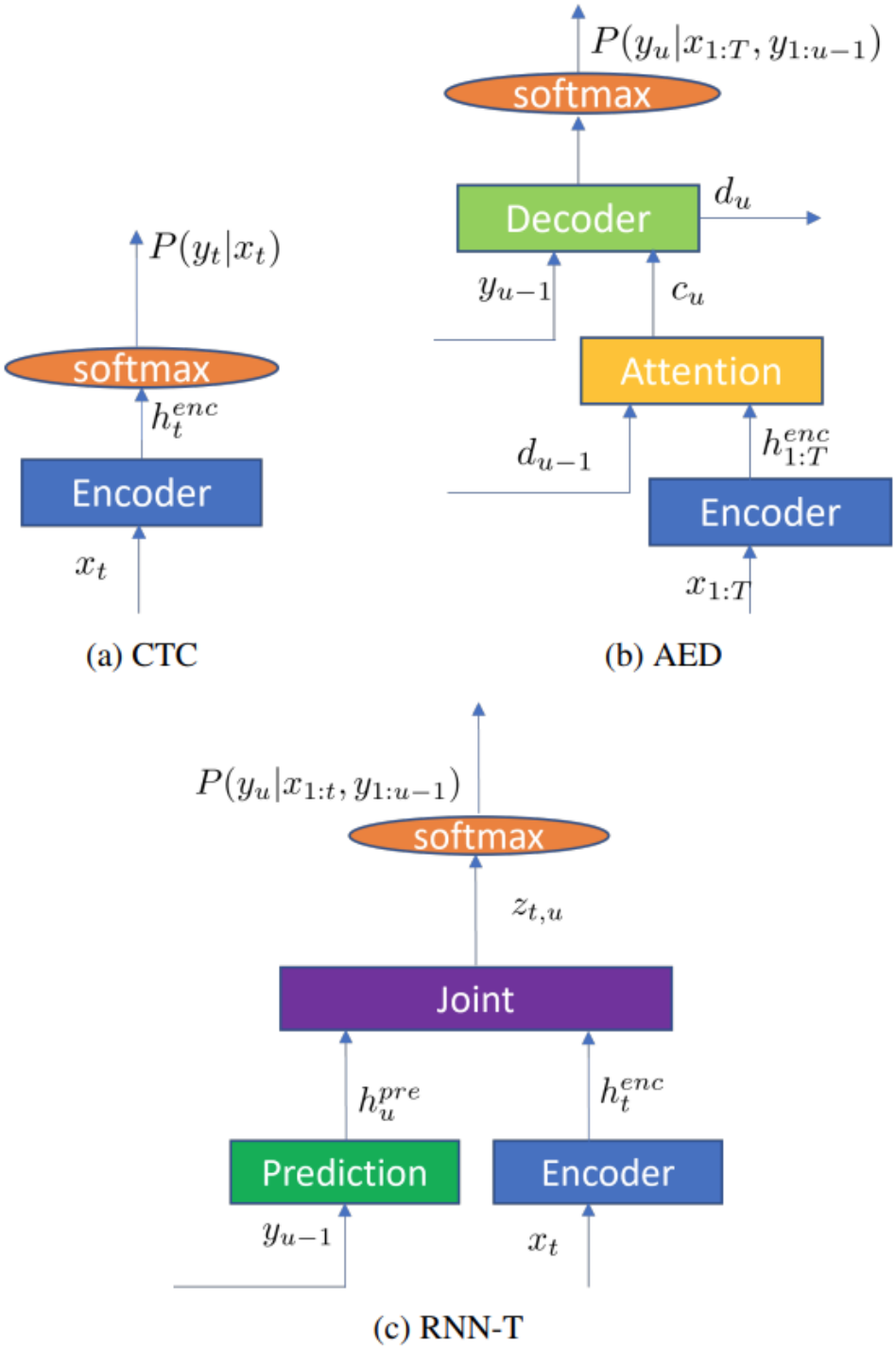
Σύγχρονες Τεχνικές Κατασκευής Συστημάτων End-to-End (E2E) ASR

3.1 Πρόσφατες εξελίξεις (state-of-the-art)

Πρόσφατα, στην αναγνώριση ομιλίας υπάρχει μια σημαντική τάση μετάβασης από την υβριδική μοντελοποίηση που βασίζεται σε βαθιά νευρωνικά δίκτυα σε μοντελοποίηση από άκρο σε άκρο (E2E). Ενώ τα μοντέλα E2E επιτυγχάνουν εντυπωσιακά αποτελέσματα όσον αφορά την ακρίβεια, τα υβριδικά μοντέλα εξακολουθούν να χρησιμοποιούνται σε μεγάλο ποσοστό εμπορικών συστημάτων ASR. Υπάρχουν πολλοί πρακτικοί παράγοντες που επηρεάζουν την απόφαση για το είδος του μοντέλου στην παραγωγή. Παραδοσιακά τα υβριδικά μοντέλα, που έχουν βελτιστοποιηθεί για την παραγωγή για δεκαετίες, είναι συνήθως επιτυχημένα με βάσει αυτούς τους παράγοντες, οπότε χωρίς να παρέχουν άριστα λύσεις είναι δύσκολο για τα μοντέλα E2E να διατεθούν ευρέως στο εμπόριο. Θα κάνουμε μια επισκόπηση των πρόσφατων εξελίξεις στα μοντέλα E2E, εστιάζοντας σε τεχνολογίες που αντιμετωπίζουν αυτές τις προκλήσεις από τη σκοπιά του κλάδου.

Η ακρίβεια των συστημάτων αυτόματης αναγνώρισης ομιλίας ενισχύθηκε σημαντικά με τα βαθιά νευρωνικά δίκτυα (DNN), η οποία υιοθετήθηκε μια δεκαετία πριν. Αυτή η ανακάλυψη λοιπόν χρησιμοποίησε το DNN για να αντικαταστήσει το παραδοσιακό Γκαουσιανό Μεικτό Μοντέλο (Gaussian Mixture Model - GMM) για την αξιολόγηση της ακουστικής πιθανότητας, διατηρώντας παράλληλα όλα τα στοιχεία όπως το ακουστικό μοντέλο, το γλωσσικό μοντέλο και το μοντέλο λεξιλογίου. Πρόσφατα, στην κοινότητα της ομιλίας έγινε μια νέα σημαντική ανακάλυψη με τη μετάβαση από την υβριδική μοντελοποίηση σε μοντελοποίηση από άκρο σε άκρο, που μεταφράζει άμεσα μια ακολουθία ομιλίας εισόδου σε μια ακολουθία εξόδου χρησιμοποιώντας ένα ενιαίο δίκτυο. Μια τέτοια σημαντική ανακάλυψη είναι ακόμη πιο επαναστατική γιατί ανατρέπει όλα τα στοιχεία μοντελοποίησης στα παραδοσιακά συστήματα ASR, τα οποία έχουν χρησιμοποιηθεί για δεκαετίες.

Υπάρχουν πολλά σημαντικά πλεονεκτήματα των μοντέλων E2E σε σχέση με τα παραδοσιακά υβριδικά μοντέλα. Αρχικά, τα μοντέλα E2E χρησιμοποιούν μια ενιαία αντικειμενική συνάρτηση που είναι συνεπής με το στόχο που είναι η βελτιστοποίηση ολόκληρου του δικτύου, ενώ τα υβριδικά μοντέλα βελτιστοποιούν μεμονωμένα τα ξεχωριστά τμήματα, που δεν μπορεί να εγγυηθεί το ολικό βέλτιστο. Επομένως, τα μοντέλα E2E έχουν αποδειχθεί ότι ξεπερνούν τα παραδοσιακά υβριδικά μοντέλα όχι μόνο στην έραυνα αλλά και στη βιομηχανία. Δεύτερον, επειδή τα E2E μοντελοποιούν άμεσα τους χαρακτήρες ή ακόμα και τις λέξεις στην έξοδο, απλοποιεί πολύ τη ροή. Αντίθετα, ο σχεδιασμός του παραδοσιακού υβριδικού μοντέλου είναι πολύπλοκος και απαιτεί πολλές ειδικές γνώσεις και πολυετή εμπειρία σε συστήματα ASR. Τέλος, επειδή χρησιμοποιείται ένα μόνο δίκτυο για το ASR, τα μοντέλα E2E είναι πολύ πιο συμπαγή από τα παραδοσιακά υβριδικά, επομένως τα E2E μπορούν να εξελιχθούν σε συσκευές με υψηλή ακρίβεια. Οι πιο δημοφιλείς τεχνικές E2E για τα συστήματα ASR είναι: (α) η Συνδετική Χρονική Ταξινόμηση (Connectionist Temporal Classification - CTC) (β) ο Κωδικοποιητής-Αποκωδικοποιητής βασισμένος στην προσοχή (Attention-based Encoder-Decoder - AED) και (γ) ο Επαναλαμβανόμενος Μετατροπέας Νευρωνικού Δικτύου (Recurrent Neural Network-Transducer - RNN-T). Ανάμεσα τους, το RNN-T παρέχει μια λύση για εφαρμογές ASR ροής (streaming) με υψηλή ακρίβεια και χαμηλή καθυστέρηση, ιδανικό για εμπορική χρήση. Θα κάνουμε μια σύντομη επισκόπηση σε αυτές τις τρεις πιο δημοφιλείς τεχνικές E2E, που φαίνονται στο πρακάτω σχήμα.



Σχήμα 3.1: Αρχιτεκτονικές των τριών πιο δημοφιλών E2E τεχνικών

3.1.1 Συνδετική Χρονική Ταξινόμηση (CTC)

Η τεχνική Connectionist Temporal Classification (CTC) σχεδιάστηκε για να αντιστοιχίσει την ακολουθία ομιλίας εισόδου σε μια ακολουθία ετικετών εξόδου. Επειδή το μήκος των ετικετών εξόδου είναι μικρότερο από αυτό της ομιλίας εισόδου, μια κενή ετικέτα <b> εισάγεται μεταξύ των ετικετών εξόδου, με επιτρεπόμενη επανάληψη των <b>, για την κατασκευή ”διαδρομών” που έχουν το ίδιο μήκος με την ακολουθία ομιλίας εισόδου. Στο Σχήμα 3.2 φαίνεται ένα παράδειγμα τριών διαδρομών CTC για τη λέξη ”team”. Δηλώνουμε την ακολουθία ομιλίας εισόδου ως x, την αρχική ετικέτα εξό-

δου ακολουθία ως y, και όλες οι διαδρομές CTC που αντιστοιχίζονται από y ως $B^{-1}(y)$. Το δίκτυο κωδικοποιητή χρησιμοποιείται για τη μετατροπή του ακουστικού χαρακτηριστικού x_t σε αναπαράσταση υψηλότερου επιπέδου h_t^{enc}. Η συνάρτηση απώλειας CTC ορίζεται ως ο αρνητικός λογάριθμος πιθανοτήτων των σωστών ετικετών δεδομένης της ομιλίας εισόδου:

$$L_{CTC} = -lnP(y|x) \tag{3.1}$$

με

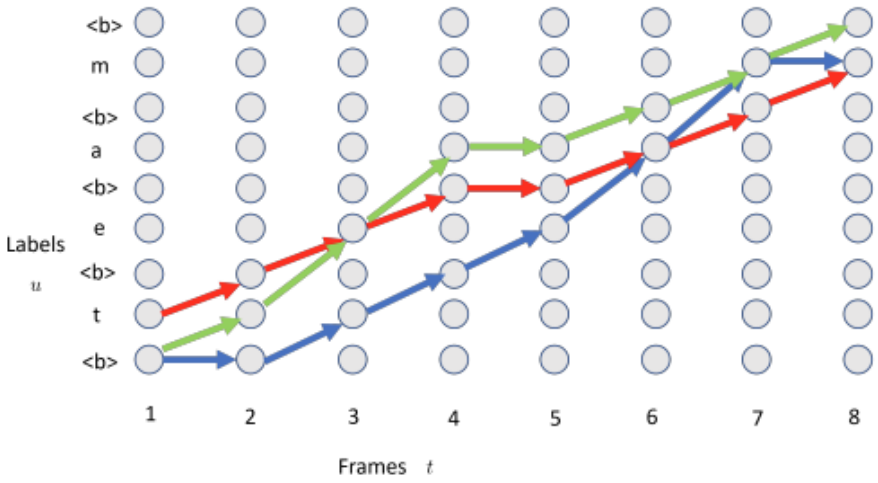
$$P(y|x) = \sum_{q \in B^{-1}(y)} P(q|x) \tag{3.2}$$

όπου q είναι μια διαδρομή (path). Με την υπόθεση για ανεξαρτησία υπό όρους το $P(q|x)$ μπορεί να αποσυντεθεί σε πολλαπλάσια των posterior πιθανοτήτων:

$$P(q|x) = \prod_{t=1}^{T} P(q_t|x) \tag{3.3}$$

όπου T είναι το μήκος της ακολουθίας την ομιλίας.

Η CTC είναι η πιο ευρέως χρησιμοποιούμενη τεχνολογία E2E σε συστήματα ASR, ωστόσο, η υπόθεση για ανεξαρτησία υπό όρους επικρίνεται αρκετά. Ένας τρόπος για να αρθεί αυτή η υπόθεση είναι να χρησιμοποιηθεί ο μηχανισμός προσοχής που εισάγει μηχανισμό μοντελοποίησης γλώσσας στα πλαίσια της ομιλίας. Τέτοια μοντέλα CTC που βασίζονται στην προσοχή χαλαρώνουν την υπόθεση της υπό όρους ανεξαρτησίας και βελτιώνουν την απόδοση του κωδικοποιητή χωρίς αλλαγή του κριτηρίου βελτιστοποίησης, επομένως βοηθούν στην απλότητα στη μοντελοποίησης. Αντικαθιστώντας την μακροπρόθεσμη μνήμη (LSTM) με μετασχηματιστές (transformers) στον κωδικοποιητή, ο οποίος επιτρέπει τη χρήση ενός πιο ισχυρού μηχανισμού προσοχής, ενισχύεται η επίδοση ακόμα περισσότερο όπως αποδεικνύεται πειραματικά. Η αυτοεποπτευόμενη μάθηση (Self-Supervised Learning - SSL) βοηθά στην εκμάθηση μιας πολύ καλής αναπαράστασης που μεταφέρει σημασιολογικές πληροφορίες.



Σχήμα 3.2: Παράδειγμα διαδρομών CTC για τη λέξη ”team”

3.1.2 Κωδικοποιητής-Αποκωδικοποιητής βασισμένος στην προσοχή (AED)

Το μοντέλο κωδικοποιητή-αποκωδικοποιητή (AED) που βασίζεται στην προσοχή είναι άλλος τύπος μοντέλου E2E. Οπως φαίνεται στο Σχήμα 3.1 ο AED έχει ένα δίκτυο κωδικοποιητή, μια μονάδα προσοχής και ένα δίκτυο αποκωδικοποιητή. Το μοντέλο υπολογίζει την πιθανότητα ως:

$$P(y|x) = \prod_{u} P(y_u|x, y_{1:u-1}) \tag{3.4}$$

όπου u είναι ο δείκτης της ετικέτας εξόδου. Η στόχος της εκπαίδευσης είναι, όπως και στο CTC, η ελαχιστοποίηση του $-lnP(y|x)$.

Το δίκτυο κωδικοποιητών εκτελεί την ίδια λειτουργία όπως του κωδικοποιητή στο CTC με μετατροπή της ακολουθίας εισόδου σε ακολουθίες κρυφών χαρακτηριστικών υψηλού επιπέδου. Η μονάδα προσοχής υπολογίζει τα βάρη προσοχής μεταξύ της προηγούμενης εξόδου του αποκωδικοποιητή και την έξοδο κωδικοποιητή καθενός πλαισίου χρησιμοποιώντας τη λειτουργία προσοχής, όπως προσθετική προσοχή (additive attention) ή προσοχή εσωτερικού γινομένου (dot-product attention) [7] και, στη συνέχεια, δημιουργεί ένα διάνυσμα συμφραζομένων, ως το σταθμισμένο άθροισμα των εξόδων του κωδικοποιητή. Το δίκτυο αποκωδικοποιητή παίρνει μαζί την προηγούμενη ετικέτα εξόδου με το παραπάνω διάνυσμα για τη δημιουργία της εξόδου του και υπολογίζει τον όρο $P(y_u|x, y_{1:u-1}$ και λειτουργεί με αυτοπαλινδρομικό (autoregressive) τρόπο ως συνάρτηση των προηγούμενων εξόδων ετικέτας χωρίς την υπόθεση ανεξαρτησίας υπό όρους.

Ενώ η προσοχή στην πλήρη ακολουθία στη μέθοδο AED είναι η λογική λύση στην αυτόματη μετάφραση όπου η σειρά των λέξεων εναλλάσσεται μεταξύ των γλωσσών πηγής και προορισμού, μπορεί να μην είναι ιδανική για ASR συστήματα επειδή το σήμα ομιλίας και η ακολουθία ετικετών εξόδου είναι μονότονες. Για να έχουμε καλύτερη ευθυγράμμιση μεταξύ του σήματος ομιλίας και της ακολουθίας ετικέτας, το μοντέλο AED έχει βελτιστοποιηθεί μαζί με ένα μοντέλο CTC σε ένα πλαίσιο εκμάθησης πολλαπλών εργασιών (multitask learning) με κοινή χρήση του κωδικοποιητή. Μια τέτοια στρατηγική εκπαίδευσης βελτιώνει σημαντικά τη σύγκλιση του μοντέλου που βασίζεται στην προσοχή και μετριάζει το θέμα ευθυγράμμισης, συνεπώς γίνεται η πλέον διαδεδομένη συνταγή εκπαίδευσης για τα περισσότερα μοντέλα AED. Μια περαιτέρω βελτίωση προτάθηκε συνδυάζοντας τις βαθμολογίες (scoring) και από τα δύο μοντέλα, και το AED και το CTC κατά την αποκωδικοποίηση.

Οι περισσότερες εμπορικές εφαρμογές με ροή πραγματικού χρόνου έχουν ανάγκη τα συστήματα ASR με χαμηλή καθυστέρηση, που σημαίνει ότι τα αποτελέσματα αναγνώρισης πρέπει να παράγονται ταυτόχρονα με τον χρήστη που μιλάει. Στα μοντέλα AED η προσοχή εφαρμόζεται σε ολόκληρη την εκφορά/άρθρωση (utterance) για να επιτευχθεί καλή απόδοση. Η καθυστέρηση μπορεί να είναι σημαντική επειδή το μοντέλο πρέπει να λάβει την πλήρη έκφραση πριν την αποκωδικοποίηση, και δεν είναι πρακτικό όταν υπάρχει ροή στο σήμα ομιλίας, δηλαδή όταν έρχεται σε συνεχή λειτουργία χωρίς τμηματοποίηση. Γίνονται πολλές προσπάθειες δημιουργίας κατάλληλων μοντέλων AED σε σενάρια συνεχόμενης ροής δεδομένων. Η βασική ιδέα αυτών των μεθόδων είναι η εφαρμογή προσοχής στα κομμάτια της ομιλίας εισόδου. Η διαφορά μεταξύ αυτών των προσπαθειών είναι ο τρόπος με τον οποίο καθορίζεται και χρησιμοποιείται η προσοχή στα κομμάτια. Ενώ όλες αυτές οι μέθοδοι μπορούν να είναι επιτυχημένες σε κάποιο βαθμό, σε σενάρια ροής, συνήθως δεν πετυχαίνουν χαμηλή καθυστέρηση, που είναι ένας άλλος σημαντικός παράγοντας για εμπορικά συστήματα ASR. Αυτή η πρόκληση αντιμετωπίστηκε με την εκπαίδευση μοντέλων AED ροής χαμηλής καθυστέρησης αξιοποιώντας την εξωτερική σκληρή ευθυγράμμιση. Το προσκοπικό δίκτυο χρησιμοποιήθηκε για την πρόβλεψη του ορίου των λέξεων, και η πρόβλεψη στη συνέχεια χρησιμοποιείται από το δίκτυο ASR για να προβλέψει την επόμενη υπο-λέξη αξιοποιώντας τις πληροφορίες από όλα τα πλαίσια ομιλίας που προηγούνται από το τρέχον πλαίσιο. Εν τέλει τα μοντέλα AED δεν μπορούν να έχουν καλή απόδοση σε μεγάλες εκφωνήσεις λόγου, επομένως όπου υπάρχει η ανάγκη για αναγνώριση σε πραγματικό χρόνο και για χαμηλή καθυστέρηση η βιομηχανία τείνει να επιλέγει μοντέλα RNN Transducer, που παρουσιάζεται στη συνέχεια ως το κυρίαρχο streaming E2E μοντέλο, ενώ το AED εξασφαλίζει τη θέση του σε κάποια non-streaming σενάρια.

3.1.3 Επαναλαμβανόμενος Μετατροπέας Νευρωνικού Δικτύου (RNN-Transducer)

Ο μετατροπέας RNN (RNN-T) όπως ειπώθηκε είναι ιδανικό μοντέλο για ροή εισόδου, ενώ επίσης την υπό όρους ανεξαρτησία υπόθεση του CTC. Όπως απεικονίζεται στο Σχήμα 3.1 αποτελείται από έναν κωδικοποιητή, ένα δίκτυο πρόβλεψης και ένα κοινό δίκτυο. Το δίκτυο κωδικοποιητών είναι

το ίδιο με αυτό στο CTC και το AED και δημιουργεί την αναπαράσταση χαρακτηριστικών υψηλού επιπέδου h_t^{enc}. Το δίκτυο πρόβλεψης παράγει μια αναπαράσταση υψηλού επιπέδου h_u^{pre} με βάση την προηγούμενη ετικέτα εξόδου y_{u-1} του RNN-T. Το κοινό δίκτυο είναι ένα εμπρόσθιο δίκτυο τροφοδοσίας (feed forward neural network - FFNN) που συνδυάζει τα h_t^{enc} και h_u^{pre} ως:

$$z_{t,u} = \psi(Qh_t^{enc} + Vh_u^{pre} + b_z) \tag{3.5}$$

όπου Q και V είναι πίνακες βάρους, το b_z είναι ένα διάνυσμα πόλωσης (bias vector), και η ψ είναι μια μη γραμμική συνάρτηση (π.χ. RELU ή Tanh). Το $z_{t,u}$ συνδέεται με το στρώμα εξόδου με τον γραμμικό μετασχηματισμό:

$$h_{t,u} = W_y z_{t,u} + b_y \tag{3.6}$$

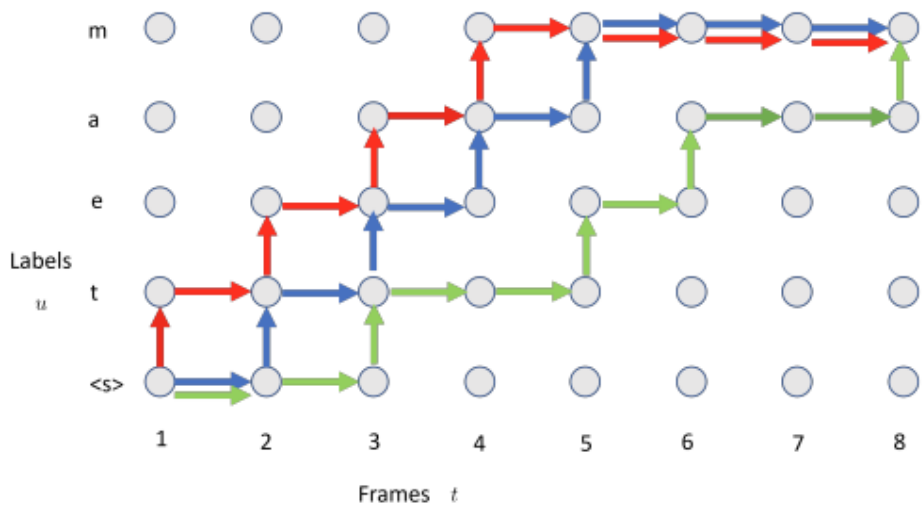
όπου W_y και b_y υποδηλώνουν έναν πίνακα βαρών και ένα διάνυσμα πόλωσης αντίστοιχα. Η πιθανότητα για κάθε διακριτικό (token) εξόδου k είναι:

$$P(y_u = k|x_{1:t}, y_{1:u-1}) = softmax(h_{t,u}^k) \tag{3.7}$$

Η συνάρτηση απώλειας του RNN-T είναι κι εδώ ίση με $-lnP(y|x)$ με:

$$P(y|x) = \sum_{a \in A^{-1}(y)} P(a|x) \tag{3.8}$$

ως το άθροισμα όλων των πιθανών διαδρομών ευθυγράμμισης που αντιστοιχούν στην ακολουθία ετικέτας y. Η αντιστοίχηση από την ευθυγράμμιση του μονοπατιού a στην ακολουθία ετικέτας y ορίζεται ως $A(a) = y$.



Σχήμα 3.3: Παράδειγμα ευθυγράμμισης διαδρομών RNN-T για τη λέξη "team"

Στο Σχήμα 3.3, απεικονίζονται τρία παραδείγματα μονοπατιών ευθυγράμμισης για την ακολουθία ομιλίας $x = (x_1, x_2, ..., x_8)$ και την ακολουθία ετικετών $y = (< s >, t, e, a, m)$, όπου το $< s >$ είναι ένα διακριτικό για την έναρξη της πρότασης. Όλες οι έγκυρες διαδρομές ευθυγράμμισης πηγαίνουν από την κάτω αριστερή γωνία στην επάνω δεξιά γωνία του πλέγματος TxU, εξ ου και το μήκος κάθε διαδρομής ευθυγράμμισης είναι $T + U$. Σε μια διαδρομή ευθυγράμμισης, το οριζόντιο βέλος προχωρά ένα βήμα με α κενή ετικέτα διατηρώντας την κατάσταση δικτύου πρόβλεψης ενώ το κατακόρυφο βέλος εκπέμπει μια μη κενή ετικέτα εξόδου. Οι posteriors πιθανότητες της ευθυγράμμισης του πλέγματος που συντίθενται από τον κωδικοποιητή και την πρόβλεψη πρέπει να υπολογιστούν σε κάθε σημείο του πλέγματος. Αυτός είναι ένας τρισδιάστατος τανυστής (tensor) που

απαιτεί πολύ περισσότερη μνήμη από αυτή που χρειάζεται σthn εκπαίδευση άλλων μοντέλων E2E όπως to CTC και to AED. Η εξίσωση (8) υπολογίζεται με βάση τον αλγόριθμο εμπρός-πίσω (forward-backward) ενώ για τη βελτίωση της αποτελεσματικότητας της εκπαίδευσης, οι πιθανότητες προς τα εμπρός και προς τα πίσω μπορούν να διανυσματοποιηθούν με έναν λοξό μετασχηματισμό βρόχου (loop skewing transformation) και οι αναδρομές μπορούν να υπολογιστούν σε έναν μόνο βρόχο αντί για δύο εμφωλευμένους βρόχους. Η συγχώνευση συναρτήσεων οδήγησε σε σημαντική μείωση του κόστους της μνήμης εκπαίδευσης, έτσι ώστε να γίνουν μεγαλύτερες παρτίδες (mini-batches) που χρησιμοποιήθηκαν για τη βελτίωση της αποτελεσματικότητας της εκπαίδευσης. Αξίζει να σημειωθεί ότι η καθυστέρηση ASR είναι μια πολύ σημαντική μέτρηση που επηρεάζει την εμπειρία του χρήστη. Εάν η καθυστέρηση είναι μεγάλη, οι χρήστες θα αισθάνονται ότι το σύστημα ASR δεν ανταποκρίνεται. Επομένως, τα συστήματα του πραγματικού κόσμου πρέπει να έχουν μικρή καθυστέρηση προκειμένου να προσφέρουν στους χρήστες μια καλή εμπειρία.

3.2 Αυτο-εποπτευόμενη μάθηση (Self-Supervised Learning - SSL)

Η αυτο-εποπτευόμενη μάθηση είναι μια υποκατηγορία της μη εποπτευόμενης μάθηση επειδή αξιοποιεί τα δεδομένα χωρίς ετικέτα. Η βασική ιδέα είναι ότι επιτραπεί στο μοντέλο να μάθει την αναπαράσταση δεδομένων χωρίς οι ετικέτες να είναι υποσημειωμένες. Μόλις το μοντέλο μάθει πώς να αναπαριστά δεδομένα στη συνέχεια μπορεί να χρησιμοποιηθεί για μεταγενέστερες εργασίες με μικρότερο αριθμό δεδομένων με ετικέτα για να επιτύχει παρόμοια ή καλύτερη απόδοση από τα μοντέλα χωρίς αυτοεποπτευόμενη μάθηση. Αποτελείται από τρία βήματα:

1. Δημιουργία των δεδομένων εισόδου και των ετικετών από τα δεδομένα χωρίς ετικέτα μέσω προγραμματισμού με βάση την κατανόηση των δεδομένων.

2. Προεκπαίδευση (pre-training): εκπαιδεύουμε το μοντέλο με δεδομένα/ετικέτες από το προηγούμενο βήμα

3. Βελτιστοποίηση (fine-tuning): χρησιμοποιούμε το προεκπαιδευμένο μοντέλο και τα αρχικά του βάρη για εκπαίδευση πάνω σε εργασίες που μας ενδιαφέρουν.

Εάν χρησιμοποιήσουμε τα δεδομένα με μη αυτόματες ετικέτες αντί για ετικέτες που δημιουργούνται αυτόματα στο δεύτερο βήμα, θα ήταν εποπτευόμενη η προεκπαίδευση, που είναι γνωστή ως (ένα βήμα) μεταφορά μάθησης (transfer learning).

Η αυτο-εποπτευόμενη μάθηση ήταν επιτυχής σε πολλά πεδία, π.χ. κείμενο, εικόνα/βίντεο, ομιλία και γράφους. Ουσιαστικά, η αυτοεποπτευόμενη μάθηση κάνει εξόρυξη των δεδομένων χωρίς ετικέτα ενισχύοντας τελικά την απόδοση. Η αυτοεποπτευόμενη μάθηση μπορεί να αποκτήσει περισσότερες χρήσιμες πληροφορίες από κάθε δείγμα από την εποπτευόμενη μάθηση. Οι ετικέτες που δημιουργούνται από τον άνθρωπο συνήθως εστιάζουν σε μια συγκεκριμένη προβολή των δεδομένων. Για παράδειγμα, μπορούμε να περιγράψουμε μια εικόνα ενός αλόγου στο γρασίδι με έναν μόνο όρο ”άλογο” για την αναγνώριση εικόνας και να παρέχουμε τις συντεταγμένες των pixel για τη σημασιολογική τμηματοποίηση. Ωστόσο, υπάρχουν πολύ περισσότερες πληροφορίες στα δεδομένα, π.χ. το κεφάλι και η ουρά του αλόγου βρίσκονται στην αντίθετη πλευρά του σώματος ή το άλογο είναι συνήθως πάνω από το γρασίδι (όχι από κάτω). Τα μοντέλα μπορούν ενδεχομένως να μάθουν καλύτερες και πιο σύνθετες αναπαραστάσεις από τα δεδομένα απευθείας αντί από χειροκίνητες ετικέτες. Οι χειροκίνητες ετικέτες μπορεί να είναι λάθος μερικές φορές, κάτι που είναι επιβλαβές για τα μοντέλα, διότι μπορεί να οδηγήσουν σε χειρότερη απόδοση.

Η επισήμανση δεδομένων είναι δαπανηρή, χρονοβόρα και αποτελεί μια εντατική εργασία. Επιπλέον, οι εποπτευόμενες προσεγγίσεις μάθησης θα χρειάζονταν διαφορετικές ετικέτες για νέα δεδομένα/ετικέτες και νέες εργασίες. Το πιο σημαντικό, έχει αποδειχθεί ότι η αυτο-εποπτευόμενη προεκπαίδευση ξεπέρασε ακόμη και την εποπτευόμενη προ-εκπαίδευση για εργασίες που βασίζονται σε εικόνες όπως για παράδειγμα αναγνώριση εικόνας, ανίχνευση αντικειμένων, σημασιολογική τμηματοποίηση. Με άλλα λόγια, η εξαγωγή πληροφοριών απευθείας από δεδομένα είναι πιο χρήσιμη από

τις χειροκίνητες ετικέτες. Τότε, ίσως να μην χρειαζόμαστε πολλές ετικέτες και δαπανηρή προσπάθεια για την υποσημείωσή τους, με πιο προηγμένη αυτοεποπτευόμενη μάθηση είτε τώρα είτε στο εγγύς μέλλον, αναλόγως φυσικά με τις εργασίες. Η ανωτερότητα της αυτο-εποπτευόμενης μάθησης έχει επικυρωθεί σε εργασίες που βασίζονται σε εικόνες χάρη στα μεγάλης κλίμακας σύνολα δεδομένων με ετικέτα, τα οποία έχουν μεγαλύτερη ιστορία από άλλους τομείς στις πρόσφατες τάσεις της βαθιάς μάθησης. Πιστεύεται ότι παρόμοια υπεροχή θα αποδειχθεί και σε άλλους τομείς στο μέλλον και ως εκ τούτου, η αυτοεποπτευόμενη μάθηση θα παίξει σημαντικό ρόλο στην εξέλιξη του τομέα μηχανικής μάθησης. Συνήθως, όταν κυκλοφορεί ένα αυτο-εποπτευόμενο μοντέλο, μπορούμε να κατεβάσουμε το προεκπαιδευμένο και στη συνέχεια μπορούμε να ρυθμίσουμε το προεκπαιδευμένο μοντέλο και να χρησιμοποιήσουμε το βελτιστοποιημένο για μια συγκεκριμένη εργασία (downstream task). Το πιο γνωστό παράδειγμα αυτοεποπτευόμενης μάθησης είναι πιθανώς το BERT, το οποίο είχε προεκπαιδευτεί σε 3,3 δισεκατομμύρια λέξεις με τη μέθοδο αυτή. Μπορούμε να βελτιστοποιήσουμε το BERT για μια εργασία που σχετίζεται με κείμενο, όπως η ταξινόμηση προτάσεων, με πολύ λιγότερη προσπάθεια και δεδομένα από την εκπαίδευση ενός μοντέλου από την αρχή. Οι κατηγορίες της αυτοεποπτευόμενης μάθησης επί γραμματικά είναι:

1. Γενετικές (Generative) προσεγγίσεις (Σχήμα 3.4): ανάκτηση των αρχικών πληροφοριών α: μη αυτο-παλινδρομικό (Non-autoregressive), όπου γίνεται κάλυψη ενός διακριτικού/εικονοστοιχείου και πρόβλεψή του (π.χ., μοντελοποίηση γλωσσικού μοντέλου με χρήση μάσακας (MLM)) β: αυτο-παλινδρομικό (Autoregressive): πρόβλεψη του επόμενου διακριτικού/pixel

2. Προσεγγίσεις πρόβλεψης (Predictive) (Σχήμα 3.5): σχεδιασμός ετικετών με βάση την κατανόηση, τη ομαδοποίηση ή την επαύξηση των δεδομένων α: πρόβλεψη το περιβάλλοντος/context (π.χ., πρόβλεψη της σχετικής θέσης κάποιων σημείων σε εικόνα, πρόβλεψη εάν το επόμενο τμήμα είναι η επόμενη πρόταση) β: πρόβλεψη το αναγνωριστικού ομαδοποίησης (cluster id) κάθε δείγματος γ: πρόβλεψη της γωνίας περιστροφής μιας εικόνας

3. Αντιθετική (Contrastive) μάθηση (γνωστή και ως διάκριση αντίθεσης περίπτωσης) (Σχήμα 3.6): ορίζουμε ένα πρόβλημα δυαδικής ταξινόμησης με βάση θετικά και αρνητικά ζεύγη δειγμάτων που δημιουργούνται με επαύξηση δεδομένων

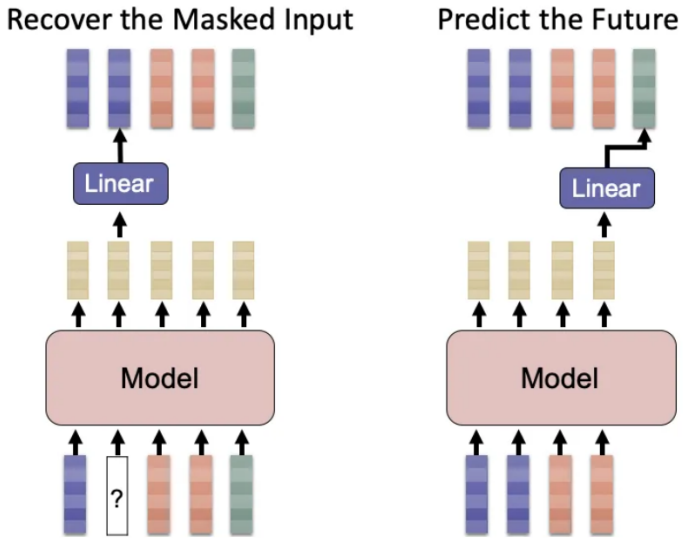
4. Προσεγγίσεις εκκίνησης (Bootstrapping) (Σχήμα 3.7): χρησιμοποιούμε δύο παρόμοια αλλά διαφορετικά δίκτυα για να μάθουν την ίδια αναπαράσταση από τα επαυξημένα ζεύγη του ίδιου δείγματος

5. Κανονικοποίηση (Regularization) (Σχήμα 3.8): άθροισμα των όρων απώλειας (loss) και κανονικοποίηση με βάση τις υποθέσεις/διαισθήσεις: α: τα θετικά ζεύγη πρέπει να είναι παρόμοια β: τα αποτελέσματα από διαφορετικά δείγματα στην ίδια παρτίδα πρέπει να είναι διαφορετικά

3.2.1 Γενετικές (Generative) προσεγγίσεις

Η πρόβλεψη καλυμμένων εισόδων από δεδομένα περιβάλλοντος είναι η παλαιότερη κατηγορία μεθόδων αυτο-εποπτευόμενης μάθησης. Η ιδέα στην πραγματικότητα μπορεί να ανιχνευθεί στο απόσπασμα, «Θα γνωρίζουμε μια λέξη από την παρέα που διατηρεί.». Αυτή η σειρά αλγορίθμων ξεκίνησε από το μοντέλο word2vec πάνω σε κειμένου το 2013. Η έννοια του συνεχούς σάκου λέξεων (CBOW) του word2vec προβλέπει μια κεντρική λέξη από τους γείτονές της, η οποία μοιάζει πολύ με την ELMo και την μοντελοποίηση μάσκας γλώσσας (MLM) του BERT. Όλα αυτά τα μοντέλα κατηγοριοποιήθηκαν ως μη-αυτοπαλινδρομικές γενετικές προσεγγίσεις. Οι κύριες διαφορές ήταν ότι τα μεταγενέστερα μοντέλα χρησιμοποιούσαν πιο προηγμένες δομές όπως αμφίδρομο LSTM (για ELMo) και μετασχηματιστή (για BERT), και τα πρόσφατα μοντέλα δημιούργησαν ενσωματωμένα διανύσματα (embeddings) με βάση τα συμφραζόμενα. Στο πεδίο ομιλίας, το Mockingjay κάλυψε όλες τις διαστάσεις των διαδοχικών χαρακτηριστικών και το TERA κάλυψε το συγκεκριμένο υποσύνολο διαστάσεων χαρακτηριστικών. Στο πεδίο εικόνας, το OpenAI εφάρμοσε κάτι ανάλογο του BERT ενώ στο πεδίο των γράφων,

το GPT-GNN κάλυπτε επίσης χαρακτηριστικά και ακμές. Όλες αυτές οι μέθοδοι κάλυπταν μερικώς τα δεδομένα εισαγωγής και προσπάθησαν να τα προβλέψουν ξανά. Από την άλλη πλευρά, μια άλλη γενετική προσέγγιση είναι η πρόβλεψη του επόμενου διακριτικού/pixel/ακουστικού χαρακτηριστικού. Στο πεδίο κειμένου, τα μοντέλα της σειράς GPT είναι οι πρωτοπόροι σε αυτήν την κατηγορία ενώ το APC και το ImageGPT εφάρμοσαν την ίδια ιδέα στα πεδία ομιλίας και εικόνας αντίστοιχα. Είναι ενδιαφέρον ότι επειδή τα γειτονικά ακουστικά χαρακτηριστικά είναι τόσο εύκολο να προβλεφθούν, το μοντέλο συνήθως καλείται να προβλέψει το διακριτικό (token) στην επόμενη ακολουθία (τουλάχιστον 3 token μακριά). Οι μεγάλες επιτυχίες της αυτοεποπτευόμενης μάθησης (ειδικά το BERT/GPT) παρακίνησαν τους ερευνητές να εφαρμόσουν παρόμοιες γενετικές προσεγγίσεις σε άλλα πεδία όπως η εικόνα και ο λόγος. Ωστόσο, για δεδομένα εικόνας και ομιλίας, είναι πιο δύσκολο να δημιουργηθούν είσοδοι με κάλυψη, καθώς η επιλογή περιορισμένου αριθμού διακριτικών κειμένου είναι ευκολότερη από την επιλογή απεριόριστου αριθμού εικονοστοιχείων εικόνας / ακουστικών χαρακτηριστικών. Οι βελτιώσεις απόδοσης δεν ήταν τόσο μεγάλες όσο στο πεδίο του κειμένου, συνεπώς οι ερευνητές ανέπτυξαν αργότερα πολλές άλλες μη γενετικές προσεγγίσεις.

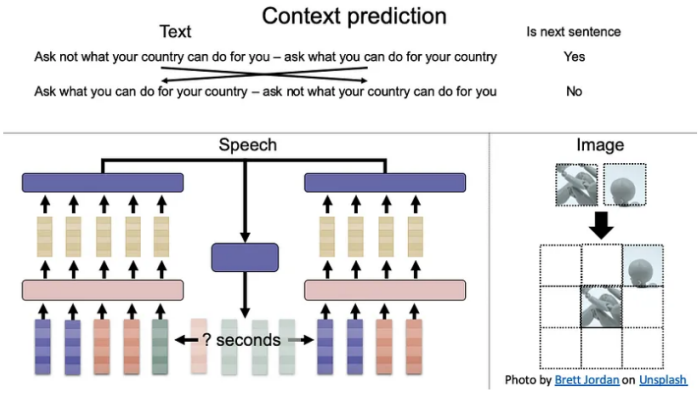


Σχήμα 3.4: Γενετικές (Generative) προσεγγίσεις

3.2.2 Προσεγγίσεις πρόβλεψης (Predictive)

Η κύρια ιδέα της προσέγγισης είναι ο σχεδιασμός πιο απλουστευμένων στόχων για να αποφευχθεί η δημιουργία νέων δεδομένων. Το πιο κρίσιμο και δύσκολο σημείο είναι ότι η εργασία πρέπει να είναι στο κατάλληλο επίπεδο δυσκολίας για να μάθει το μοντέλο. Για παράδειγμα, στην πρόβλεψη του context στο πεδίο κειμένου, τόσο το BERT όσο και το ALBERT προέβλεψαν εάν το επόμενο τμήμα αντιστοιχεί στην επόμενη πρόταση. Στο BERT δόθηκαν αρνητικά δείγματα εκπαίδευσης ανταλλάσσοντας τυχαία το επόμενο τμήμα με ένα άλλο τμήμα (πρόβλεψη επόμενης πρότασης, Next Sentence Prediction - NSP) ενώ στο ALBERT δόθηκαν αρνητικά δείγματα εκπαίδευσης ανταλλάσσοντας το προηγούμενο και το επόμενο τμήμα (πρόβλεψη σειράς πρότασης, Sentence Order Prediction - SOP). Το SOP έχει αποδειχθεί ότι υπερέχει του NSP, και μια εξήγηση είναι ότι είναι τόσο εύκολο να διακρίνουμε τυχαία ζεύγη προτάσεων από την πρόβλεψη του θέματος που το μοντέλο δεν έμαθε πολλά από την εργασία NSP, ενώ η εργασία SOP επιτρέπει στο μοντέλο να μάθει τη σχέση συνοχής. Συμπερασματικά λοιπόν, χρειάζεται γνώση τομέα για να σχεδιάσει καλές εργασίες και πειράματα για να επικυρώσει την αποτελεσματικότητα μιας εργασίας. Η ιδέα της πρόβλεψης του πλαισίου (context) εφαρμόστηκε επίσης σε εικόνες (ως πρόβλεψη της σχετικής θέσης των κενών εικόνας) και στο πεδίο ομιλίας (ως πρόβλεψη του χρονικού διαστήματος μεταξύ δύο ομάδων ακουστικών χαρακτηριστικών). Μια άλλη προσέγγιση είναι η δημιουργία των ετικετών με ομαδοποίηση. Στο πεδίο εικόνας, το

DeepCluster εφάρμοσε την ομαδοποίηση k-means, στο πεδίο ομιλίας το HuBERT εφάρμοσε ομαδοποίηση k-means και το BEST-RQ χρησιμοποίησε έναν κβαντιστή τυχαίας προβολής. Άλλες εργασίες στο πεδίο της εικόνας είναι η πρόβλεψη του καναλιού κλίμακας του γκρι από τα κανάλια χρώματος των εικόνων και αντίστροφα, η ανασύνθεση της τυχαίας περικοπής των εικόνων, η ανακατασκευή των εικόνων αρχικής ανάλυσης, η πρόβλεψη της γωνίας περιστροφής των εικόνων και η πρόβλεψη των χρωμάτων των εικόνων.

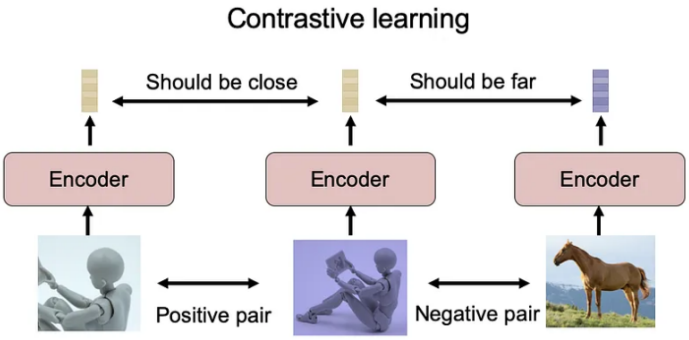


Σχήμα 3.5: Προσεγγίσεις πρόβλεψης (Predictive)

3.2.3 Αντιθετική (Contrastive) μάθηση

Η βασική ιδέα της αντιθετικής μάθησης είναι να δημιουργήσει τα θετικά και αρνητικά ζεύγη δειγμάτων εκπαίδευσης με βάση την κατανόηση των δεδομένων. Το μοντέλο πρέπει να μάθει μια συνάρτηση έτσι ώστε τα δύο θετικά δείγματα να έχουν υψηλές βαθμολογίες ομοιότητας και δύο αρνητικά δείγματα χαμηλές βαθμολογίες ομοιότητας. Ως αποτέλεσμα, η κατάλληλη δημιουργία δείγματος είναι απαραίτητη για να διασφαλιστεί ότι το μοντέλο μαθαίνει τα υποκείμενα χαρακτηριστικά/δομές των δεδομένων. Η αντιθετική μάθηση στο πεδίο εικόνας εφαρμόζει δύο διαφορετικές επαυξήσεις δεδομένων από την ίδια αρχική εικόνα για να δημιουργήσει θετικά ζεύγη δειγμάτων και να χρησιμοποιήσει δύο διαφορετικές εικόνες ως ζεύγη αρνητικών δειγμάτων. Τα δύο πιο κρίσιμα και γεμάτα προκλήσεις μέρη είναι η διαδικασία επαύξησης και η επιλογή αρνητικών ζευγών δειγμάτων. Εάν η επαύξηση είναι τέτοια ώστε να μην υπάρχει σχέση μεταξύ των δύο επαυξημένων δειγμάτων από το ίδιο δείγμα, το μοντέλο δεν μπορεί να μάθει. Ομοίως, εάν η αύξηση είναι τόσο μικρή που το μοντέλο μπορεί να λύσει εύκολα το πρόβλημα, τότε το μοντέλο δεν μπορεί επίσης να μάθει χρήσιμες πληροφορίες για το κυρίως task (το downstream task όπως λέγεται). Όσον αφορά την επιλογή ζευγών αρνητικών δειγμάτων, εάν αντιστοιχίσουμε τυχαία δύο εικόνες ως αρνητικό ζεύγος, μπορεί να είναι της ίδιας κατηγορίας (π.χ. δύο εικόνες γατών), γεγονός που εισάγει αντικρουόμενο θόρυβο στο μοντέλο. Εάν τα αρνητικά ζεύγη είναι πολύ εύκολο να διακριθούν, τότε το μοντέλο δεν μπορεί να μάθει τα υποκείμενα χαρακτηριστικά των δεδομένων. Όσον αφορά το πεδίο ομιλίας, μια προσέγγιση είναι η εφαρμογή επαύξησης όπως το SimCLR (Speech SimCLR) ενώ μια δεύτερη είναι να χρησιμοποιηθούν γειτονικά χαρακτηριστικά ως θετικά ζεύγη και χαρακτηριστικά από διαφορετικά δείγματα ως αρνητικά ζεύγη (όπως στο Wav2vec (v1, v2.0) και στο Discret BERT). Μια ενδιαφέρουσα οπτική είναι ότι η ταξινόμηση στην αυτο-εποπτευόμενη μάθηση στο πεδίο κειμένου είναι στην πραγματικότητα παρόμοια με την αντιθετική μάθηση εννοιολογικά. Η ταξινόμηση μεγιστοποιεί την έξοδο της θετικής τάξης και ελαχιστοποιεί τα αποτελέσματα των αρνητικών κλάσεων. Ομοίως, η αντιθετική μάθηση μεγιστοποιεί επίσης την απόδοση των θετικών ζευγών και ελαχιστοποιεί τα αποτελέσματα των αρνητικών ζευγών. Η βασική διάκριση είναι ότι η ταξινόμηση έχει πεπερασμένο αριθμό αρνητικών κατηγοριών (στην περίπτωση των tokens κειμένου) ενώ η αντιθετική μάθηση έχει άπειρο αριθμό αρνητικών κατηγοριών (στην περίπτωση εικόνων και ακουστικών χαρακτηριστικών). Θεωρητικά, μπορούμε να σχεδιάσουμε έναν ταξινομητή για εικόνες/ομιλίες με δεδομένο μικρό αριθμό κλάσεων. Μία κλάση είναι μία πρωτότυπη εικόνα και η είσοδος είναι οι επαυξημένες εικόνες. Ωστόσο, αυτό δεν θα ήταν πρακτικό, καθώς

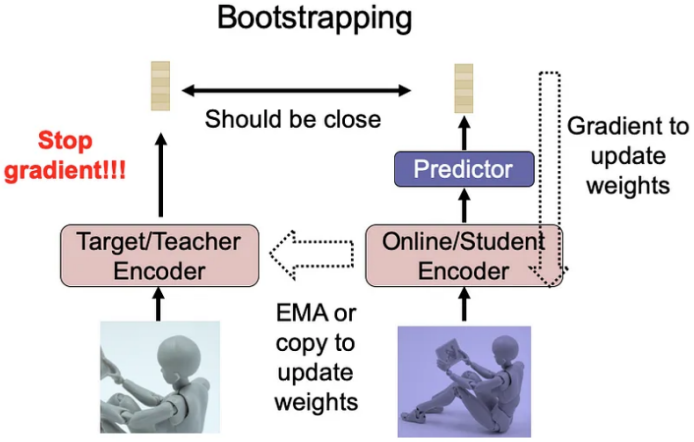
θα μπορούσε να εφαρμοστεί μόνο σε περιορισμένο αριθμό εικόνων/κλάσεων.



Σχήμα 3.6: Αντιθετική (Contrastive) μάθηση

3.2.4 Προσεγγίσεις εκκίνησης (Bootstrapping)

Περαιτέρω προσεγγίσεις αναπτύχθηκαν με σκοπό να αποφευχθεί η χρήση αρνητικών παραδειγμάτων, καθώς είναι υπολογιστικά κοστοβόρο για την εκπαίδευση και δεν είναι εύκολο να επιλεγούν καλά αρνητικά παραδείγματα. Οι βασικές ιδέες των προσεγγίσεων bootstrapping είναι 1) η δημιουργία ενός θετικού ζεύγους δειγμάτων από δύο επαυξήσεις του ίδιου αρχικού δείγματος (ακριβώς όπως στην αντιθετική μάθηση). 2) η ρύθμιση ενός δικτύου προορισμού/στόχου (ονομάζεται επίσης δίκτυο δασκάλου) και ένα άλλο δίκτυο ως online δίκτυο (ονομάζεται επίσης δίκτυο μαθητή), το οποίο είναι η ίδια αρχιτεκτονική με το δίκτυο προορισμού συν ένα πρόσθετο πρόσθιο (feed-forward) επίπεδο (που ονομάζεται πρόβλεψη) 3) ο καθορισμός των βαρών του δικτύου δασκάλου/μαθητή και η ενημέρωση μόνο του δικτύου μαθητή. 4) ενημέρωση των βαρών του δικτύου δασκάλου με βάση τα βάρη του δικτύου μαθητή. Τα πιο σημαντικά σημεία είναι ότι: 1) το δίκτυο μαθητή πρέπει να έχει το προγνωστικό επίπεδο (predictor), το οποίο είναι ένα επιπλέον επίπεδο. 2) μόνο τα βάρη του δικτύου μαθητή μπορούν να ενημερωθούν. Το Data2Vec, που αναπτύχθηκε από την εταιρεία Meta, είναι ένα ενοποιημένο framework που χρησιμοποείται για εικόνες, ομιλία και κείμενο. Τροφοδοτεί το δίκτυο προορισμού-/δασκάλου με τα αρχικά δεδομένα και το δίκτυο μαθητή με τα καλυμμένα δεδομένα. Ένας σημαντικός σχεδιασμός είναι ότι στόχος του είναι η πρόβλεψη των ενσωματωμένων διανυσμάτων (embeddings) των καλυμμένων περιοχών εισόδου/token των κορυφαίων επιπέδων στο δίκτυο στόχου/δασκάλου.

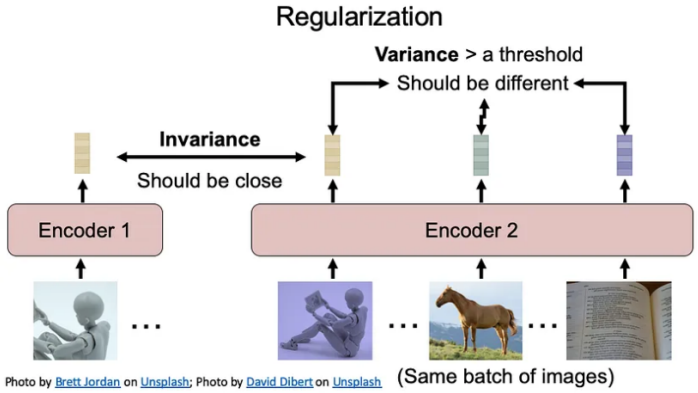


Σχήμα 3.7: Προσεγγίσεις εκκίνησης (Bootstrapping)

3.2.5 Κανονικοποίηση (Regularization)

Αυτή είναι μια άλλη προσέγγιση που χρειάζεται θετικά ζεύγη, χωρίς αρνητικά παραδείγματα. Παραδόξως, αυτές οι μέθοδοι μπορούν να χρησιμοποιήσουν τις ίδιες αρχιτεκτονικές για τα δύο δίκτυα και επίσης δεν χρειάζονται τον μηχανισμό «διακοπής κλίσης» (gradient stop) για να ενημερώσουν μόνο ένα από τα δίκτυα κατά τη διάρκεια της εκπαίδευσης. Με την προσθήκη επιπλέον όρων κανονικοποίησης, το μοντέλο επίσης δεν καταρρέει. Οι όροι της αντικειμενικής συνάρτησης περιλαμβάνουν:

Αμεταβλητότητα (invariance): ο όρος απώλειας διατηρεί τα δύο ενσωματωμένα διανύσματα (embeddings) από το ίδιο θετικό ζεύγος όσο το δυνατόν πιο όμοια. Διακύμανση (variance): ο όρος κανονικοποίησης κρατά τα δείγματα στην ίδια παρτίδα αρκετά διαφορετικά αφού δεν είναι το ίδιο δείγμα. Αυτός ο όρος μπορεί να ενισχύσει σημαντικά την απόδοση και μεγιστοποιεί την αποτελεσματικότητα της χρήσης όλων των διαστάσεων της ενσωματωμένων διανυσμάτων.



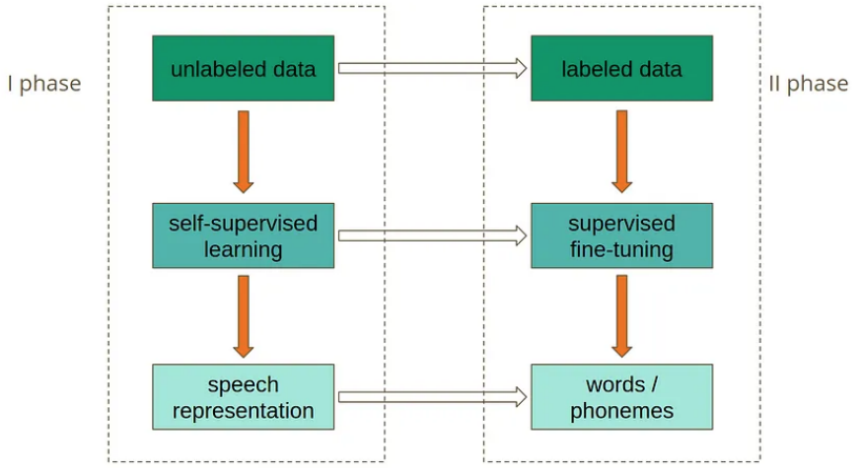
Σχήμα 3.8: Κανονικοποίηση (Regularization)

Η τεχνική SSL εξελίσσεται από τη πρόβλεψη με κάλυψη (masked prediction), την πρόβλεψη του επόμενου όρου, την αντιθετική μάθηση έως το bootstrapping και την κανονικοποίηση πάνω σε πολλαπλές μορφές δεδομένων όπως κείμενο, εικόνα, ήχου/ομιλίας και γράφων. Όπως αναφέρθηκε αρχικά, το μοντέλο χρησιμοποιεί δεδομένα για εκμάθηση και δεν έχει ανάγκη την παρουσία ετικετών. Στη συνέχεια, το μοντέλο μπορεί να μάθει από τις εργασίες που έχουν σχεδιαστεί με βάση την κατανόηση των δεδομένων, ενώ η αύξηση δεδομένων, τα παραδείγματα θετικών και αρνητικών ζευγών επιτρέπουν την αντιθετική μάθηση. Το πιο εκπληκτικό είναι ότι με τις τεχνικές εκκίνησης ή τους όρους κανονικοποίησης, το μοντέλο μπορεί ακόμη και να μάθει χωρίς αρνητικά παραδείγματα. Με καλύτερη κατανόηση του SSL στο άμεσο μέλλον, μπορούμε να αναπτύξουμε πιο στιβαρά μοντέλα με λιγότερα δεδομένα, χρόνο και προσπάθεια.

3.3 Αρχιτεκτονική Wav2Vec 2.0

Το Wav2Vec 2.0 είναι ένα από τα τρέχοντα μοντέλα τελευταίας τεχνολογίας για Αυτόματη Αναγνώριση Ομιλίας, που βασίζεται στην ιδέα της αυτο-εποπτευόμενης μάθησης. Αυτός ο τρόπος εκπαίδευσης μας επιτρέπει να εκπαιδεύσουμε εκ των προτέρων ένα μοντέλο σε δεδομένα χωρίς ετικέτα, το οποίο είναι πάντα πιο προσιτό. Στη συνέχεια, το μοντέλο μπορεί να προσαρμοστεί με ακρίβεια σε ένα συγκεκριμένο σύνολο δεδομένων για έναν συγκεκριμένο σκοπό. Όπως φαίνεται από διάφορα πειράματα για ποικίλλες εργασίες, αυτός ο τρόπος εκπαίδευσης είναι πολύ αποδοτικός. Όπως παρουσιάζεται στο Σχήμα 3.9 , το μοντέλο εκπαιδεύεται σε δύο φάσεις. Η πρώτη φάση είναι σε λειτουργία αυτοεποπτευόμενης, η οποία γίνεται με χρήση δεδομένων χωρίς ετικέτα και στοχεύει στην επίτευξη της καλύτερης δυνατής αναπαράστασης της ομιλίας. Μπορούμε να το σκεφτούμε με παρόμοιο τρόπο όπως στα ενσωματωμένα διανύσματα λέξεων (word embeddings), αφού εκείνα αναλόγως στοχεύουν στην καλύτερη αναπαράσταση της φυσικής γλώσσας. Η κύρια διαφορά είναι ότι το Wav2Vec 2.0 επεξεργάζεται ήχο αντί για κείμενο. Η δεύτερη φάση της εκπαίδευσης είναι το εποπτευόμενο

”κούρδισμα” (fine-tuning), κατά την οποία χρησιμοποιούνται δεδομένα με ετικέτα για να διδάξουν το μοντέλο να προβλέπει συγκεκριμένες λέξεις ή φωνήματα. Όπως έχουμε προαναφέρει «φώνημα» είναι η μικρότερη δυνατή μονάδα ήχου σε μια συγκεκριμένη γλώσσα, που συνήθως αντιπροσωπεύεται από ένα ή δύο γράμματα.



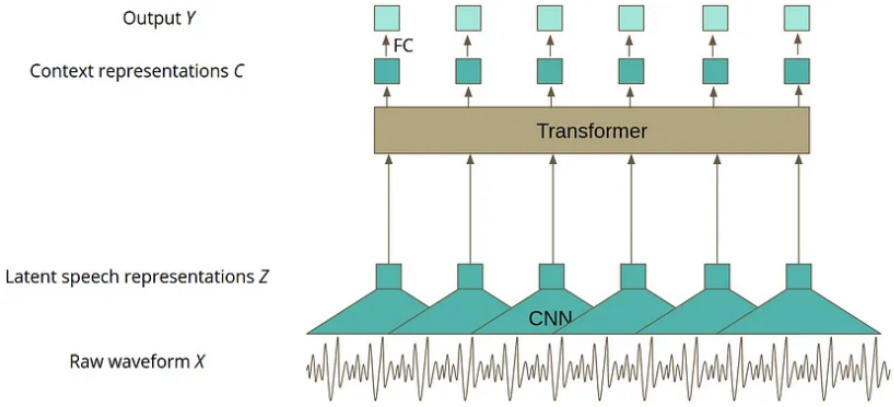
Σχήμα 3.9: Φάσεις εκπαίδευσης του Wav2Vec 2.0

Η πρώτη φάση της εκπαίδευσης είναι το κύριο πλεονέκτημα αυτού του μοντέλου. Η εκμάθηση μιας πολύ καλής αναπαράστασης ομιλίας επιτρέπει την επίτευξη αποτελεσμάτων αιχμής σε μια μικρή ποσότητα δεδομένων με ετικέτα. Για παράδειγμα, οι συγγραφείς του δημοσιευμένου paper έχουν εκπαιδεύσει το μοντέλο σε ένα τεράστιο σύνολο δεδομένων το LibriVox. Στη συνέχεια, χρησιμοποίησαν ολόκληρο το σύνολο δεδομένων του Libri Speech για fine-tuning που είχε ως αποτέλεσμα 1,8% ποσοστό λάθους λέξης (WER) στο υποσύνολο test-clean και 3,3% WER στο test-other. Τα αποτελέσματα αναλογικά με την ποσότητα των δεδομένων φαίνονται στο Σχήμα 3.10:

Σχήμα 3.10: Χρόνος εκπαίδευσης έναντι WER του Wav2Vec 2.0

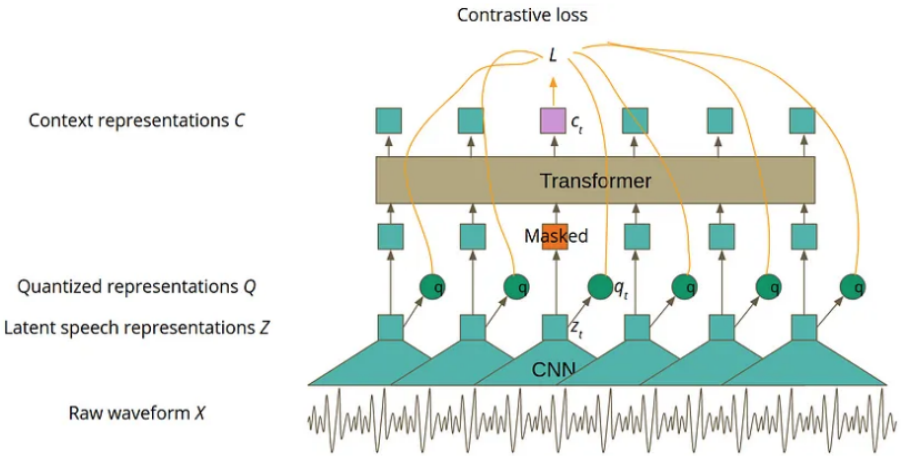
Η αρχιτεκτονική του τελικού μοντέλου που χρησιμοποιείται για την πρόβλεψη αποτελείται από τρία κύρια μέρη:

- συνελικτικά στρώματα που επεξεργάζονται την ακατέργαστη είσοδο κυματομορφής για να λάβουν την κρυφή αναπαράσταση Z,
- στρώματα μετασχηματιστή, που δημιουργούν αναπαράσταση με βάση τα συμφραζόμενα, C,
- μια γραμμική προβολή στην έξοδο, Y.



Σχήμα 3.11: Fine-tuned μοντέλο Wav2Vec 2.0 που χρησιμοποιείται για πρόβλεψη

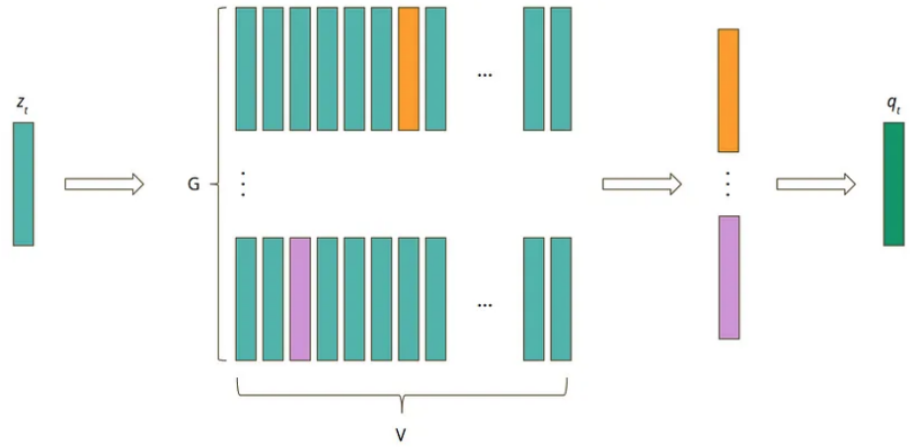
Στο Σχήμα 3.11 φαίνεται το μοντέλο μετά το τελικό fine-tuning, έτοιμο να τεθεί σε λειτουργία σε περιβάλλον παραγωγής. Όλη η μαγεία συμβαίνει κατά την πρώτη φάση της εκπαίδευσης, στην αυτο-εποπτευόμενη λειτουργία, όπου και είναι λίγο διαφορετικό. Το μοντέλο εκπαιδεύεται χωρίς τη γραμμική προβολή που δημιουργεί την πρόβλεψη εξόδου. Η κύρια ιδέα της προεκπαίδευσης στην αναπαράσταση ομιλίας είναι παρόμοια με το BERT: μέρος της εισόδου του μετασχηματιστή είναι καλυμμένο και ο στόχος είναι να μαντέψουμε την αναπαράσταση του κρυφού διανύσματος της κάλυψης z_t. Σε αυτή την ιδέα έρχεται να βοηθήσει η αντιθετική μάθηση, που αναφερθήκαμε νωρίτερα και η οποία είναι μια έννοια στην οποία τα δεδομένα μετασχηματίζονται με δύο διαφορετικούς τρόπους. Στη συνέχεια, το μοντέλο εκπαιδεύεται να αναγνωρίζει εάν δύο μετασχηματισμοί της εισόδου εξακολουθούν να είναι το ίδιο αντικείμενο. Στο Wav2Vec 2.0, τα στρώματα μετασχηματιστή είναι ο πρώτος τρόπος μετασχηματισμού, ο δεύτερος γίνεται με κβαντισμό, ο οποίος θα εξηγηθεί στη συνέχεια. Ουσιαστικά, για την καλυμμένη κρυφή αναπαράσταση z_t, θα θέλαμε να λάβουμε μια τέτοια αναπαράσταση του context c_t για να μπορούμε να μαντέψουμε τη σωστή κβαντισμένη αναπαράσταση q_t μεταξύ άλλων κβαντισμένων αναπαραστάσεων. Η έκδοση του Wav2Vec 2.0 που χρησιμοποιείται για αυτο-εποπτευόμενη εκπαίδευση παρουσιάζεται στο Σχήμα 3.12.



Σχήμα 3.12: Self-Supervised εκπαίδευση του Wav2Vec 2.0

Η κβαντοποίηση (ή κβαντισμός - quantization) είναι μια διαδικασία μετατροπής τιμών από έναν συνεχή χώρο σε ένα πεπερασμένο σύνολο τιμών σε ένα διακριτό χώρο. Ας υποθέσουμε ότι το κρυφό διάνυσμα αναπαράστασης ομιλίας z_t καλύπτει δύο φωνήματα. Ο αριθμός των φωνημάτων σε μια γλώσσα είναι πεπερασμένος κι επιπλέον ο αριθμός όλων των πιθανών ζευγών φωνημάτων είναι πε-

περασμένος, που σημαίνει ότι μπορούν να αναπαρασταθούν τέλεια από την ίδια κρυφή αναπαράσταση ομιλίας. Επιπλέον, ο αριθμός τους είναι πεπερασμένος, οπότε μπορούμε να δημιουργήσουμε ένα βιβλίο κωδικών (codebook) που περιέχει όλα τα πιθανά ζεύγη φωνημάτων. Οπότε, η κβαντοποίηση καταλήγει στην κατάλληλη επιλογή της σωστής κωδικής λέξης από το βιβλίο κωδικών. Ωστόσο, μπορείτε να φανταστούμε ότι ο αριθμός όλων των πιθανών ήχων είναι τεράστιος, άρα για να διευκολυνθεί η εκπαίδευση και η χρήση δημιουργήθηκαν βιβλία κωδικών G, το καθένα αποτελούμενο από λέξεις κωδικών V. Για να δημιουργηθεί μια κβαντισμένη αναπαράσταση, θα πρέπει να επιλεγεί η καλύτερη λέξη από κάθε βιβλίο κωδικών. Στη συνέχεια, τα επιλεγμένα διανύσματα ενώνονται και υποβάλλονται σε επεξεργασία με έναν γραμμικό μετασχηματισμό για να ληφθεί μια κβαντισμένη αναπαράσταση. Η διαδικασία του κβαντισμού παρουσιάζεται στο Σχήμα 3.13:



Σχήμα 3.13: Κβαντισμός στο Wav2Vec 2.0

Δεδομένου ότι είναι μια εργασία ταξινόμησης, η συνάρτηση softmax φαίνεται να είναι μια λογική επιλογή για την επιλογή της καλύτερης λέξης κώδικα σε κάθε βιβλίο κωδικών. Στην περίπτωσή μας, η Gumbel softmax είναι καλύτερη από την γνωστή softmax, διότι έρχεται με δύο βελτιώσεις: τυχαιότητα (randomization) και θερμοκρασία τ. Λόγω της τυχαιότητας, το μοντέλο είναι πιο πρόθυμο να επιλέξει διαφορετικές κωδικές λέξεις κατά τη διάρκεια της εκπαίδευσης και στη συνέχεια να ενημερώσει τα βάρη τους. Είναι σημαντικό, ειδικά στην αρχή της εκπαίδευσης, να αποτρέπεται η χρήση μόνο ενός υποσυνόλου βιβλίων κωδικών. Η θερμοκρασία μειώνεται με την πάροδο του χρόνου από 2 σε 0,5, οι τιμές προέρχονται από πειραματισμό, επομένως ο αντίκτυπος της τυχαιότητας γίνεται μικρότερος με την πάροδο του χρόνου. Η Gumbel softmax, με βάση την οποία επιλέγεται η καλύτερη κωδική λέξη από κάθε codebook υπολογίζεται ως:

$$p_{g,u} = \frac{\exp\left(sim(l_{g,u} + n_u)/\tau\right)}{\sum_{k=1}^{V} \exp\left((l_{g,k} + n_k)/\tau\right)} \tag{3.9}$$

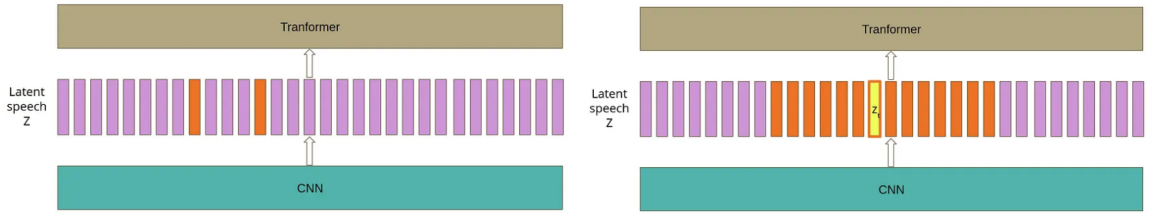
όπου:

- sim - η ομοιότητα συνημιτόνου (cosine similarity),
- $l \in R^{G*V}$ - τα υπολογισμένα logits από το z,
- $n_k = -log(-log(u_k))$,
- u_k - δειγματοληπτείται από την ομοιόμορφη κατανομή $U(0, 1)$,
- τ - η θερμοκρασία.

Στη διαδικασία της κάλυψης (masking) ορίζονται δύο υπερπαράμετροι: p = 0,065 και M = 10 και γίνονται τα ακόλουθα βήματα:

1. παίρνουμε όλα τα χρονικά βήματα από το χώρο της κρυφής αναπαράστασης ομιλίας Z.
2. γίνεται δειγματοληψία χωρίς αντικατάσταση ποσοστό p των διανυσμάτων από το προηγούμενο βήμα.
3. τα επιλεγμένα χρονικά βήματα είναι οι δείκτες έναρξης
4. για κάθε δείκτη i, διαδοχικά M χρονικά βήματα καλύπτονται.

Όπως παρουσιάζεται στο Σχήμα 3.14, επιλέξαμε τυχαία δύο διανύσματα ως αρχικούς δείκτες, όπου φαίνονται με πορτοκαλί χρώμα:



(a) Επιλογή αρχικών δεικτών για masking **(b)** Κάλυψη M συνεχόμενων χρονικών βημάτων

Σχήμα 3.14: Δειγματοληψία διανυσμάτων για κάλυψη (masking)

Στη συνέχεια, ξεκινώντας από κάθε επιλεγμένο διάνυσμα, καλύπτονται M = 10 διαδοχικά χρονικά βήματα. Τα διαστήματα μπορεί να επικαλύπτονται και επειδή το κενό μεταξύ τους ισούται με 3 χρονικά βήματα, καλύπτουμε 14 διαδοχικά χρονικά βήματα. Τέλος, η απώλεια αντίθεσης (contrastive loss) υπολογίζεται μόνο για το κεντρικό χρονικό βήμα της μάσκας.

Ο αντικειμενικός στόχος της εκπαίδευσης είναι ένα άθροισμα δύο συναρτήσεων απώλειας: απώλεια αντίθεσης (contrastive loss) και απώλεια ποικιλομορφίας (diversity loss).

$$L = L_m + \alpha L_d \tag{3.10}$$

Η απώλεια αντίθεσης είναι υπεύθυνη για την εκπαίδευση του μοντέλου ώστε να προβλέπει τη σωστή κβαντισμένη αναπαράσταση q_t μεταξύ $K + 1$ κβαντισμένων υποψήφιων αναπαραστάσεων $q' \in Q_t$. Το σετ Q_t αποτελείται από στόχου; q_t και K διαχωριστές (distractors) που έχουν ληφθεί ομοιόμορφα από άλλα χρονικά βήματα της κάλυψης/μάσκας. Η απώλεια της αντίθεσης δίνεται από τη φόρμουλα:

$$L_m = -log \frac{\exp\left(sim(c_t, q_t)/k\right)}{\sum\limits_{\hat{q} \in Q_t} \exp\left(sim(c_t, \hat{q}_t)/k\right)} \tag{3.11}$$

Το είναι μια τιμή θερμοκρασίας που είναι σταθερή κατά τη διάρκεια της εκπαίδευσης. Ο όρος sim δίνει την ομοιότητα συνημίτονου και το κύριο μέρος της συνάρτησης L_m είναι παρόμοιο με τη softmax, αλλά αντί για βαθμολογίες παίρνουμε ομοιότητες συνημιτόνου μεταξύ της αναπαράστασης του context c_t και των κβαντισμένων αναπαραστάσεων q_t. Για ευκολότερη βελτιστοποίηση βάζουμε επίσης $-log$ σε αυτό το κλάσμα.

Η απώλεια ποικιλομορφίας είναι ένα είδος τεχνικής κανονικοποίησης. Έχει οριστεί $G = 2$ βιβλία κωδικών με $V = 320$ κωδικές λέξεις σε κάθε βιβλίο κωδικών. Θεωρητικά δίνει 320*320=102400 αριθμό πιθανών κβαντισμένων αναπαραστάσεων. Ωστόσο, δεν γνωρίζουμε αν το μοντέλο θα χρησιμοποιήσει πραγματικά όλες αυτές τις πιαθανότητες. Διαφορετικά, θα μάθει μόνο να χρησιμοποιεί, για παράδειγμα, 100 κωδικές λέξεις από κάθε βιβλίο κωδικών και θα σπαταλήσει όλες τις άλλες του

codebook. Γι' αυτό η απώλεια ποικιλομορφίας να είναι χρήσιμη και βασίζεται στην εντροπία, η οποία μπορεί να υπολογιστεί με τον ακόλουθο τύπο:

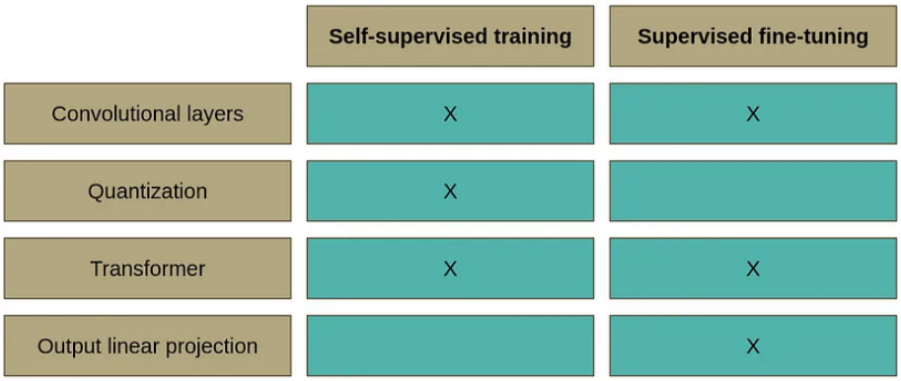
$$H(X) = \sum_{x} P(x) log(P(x)) \tag{3.12}$$

όπου x είναι ένα πιθανό αποτέλεσμα της διακριτής μεταβλητής χ, και $P(x)$ η πιθανότητα του γεγονότος x.

Η εντροπία λαμβάνει τη μέγιστη τιμή όταν η κατανομή των δεδομένων είναι ομοιόμορφη. Στην περίπτωσή μας, σημαίνει ότι όλες οι κωδικές λέξεις χρησιμοποιούνται με την ίδια συχνότητα. Με αυτό, μπορούμε να υπολογίσουμε την εντροπία κάθε βιβλίου κωδίκων σε ολόκληρη την παρτίδα των παραδειγμάτων εκπαίδευσης για να ελέγξουμε αν οι κωδικές λέξεις χρησιμοποιούνται με την ίδια συχνότητα. Η μεγιστοποίηση αυτής της εντροπίας θα ενθαρρύνει το μοντέλο να εκμεταλλευτεί όλες τις κωδικές λέξεις. Η μεγιστοποίηση ισοδυναμεί με ελαχιστοποίηση της αρνητικής εντροπίας που ισούται με την απώλεια ποικιλομορφίας L_d, που υπολογίζεται ως:

$$L_d = \frac{1}{GV} * (-H(\hat{p}_g)) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \hat{p}_{g,v} log(\hat{p}_{g,v}) \tag{3.13}$$

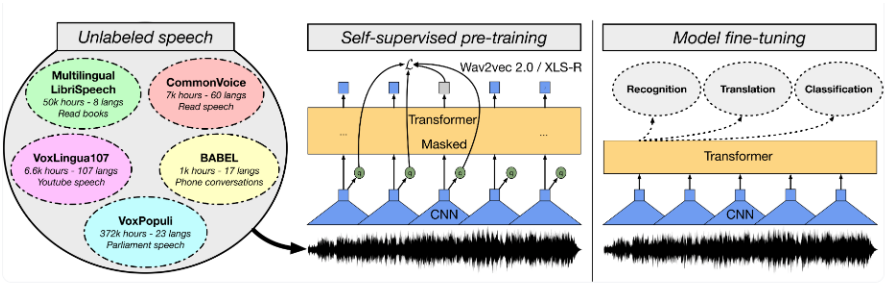
Στη φάση του fine-tuning του Wav2Vec 2.0 η κβαντοποίηση δεν χρησιμοποιείται, όπως είπαμε. Αντί αυτού, προστίθεται ένα τυχαία αρχικοποιημένο επίπεδο γραμμικής προβολής πάνω από την αναπαράσταση του context C. Στη συνέχεια, το μοντέλο ρυθμίζεται με χρήση της απώλειας Connectionist Temporal Classification (CTC) και μια τροποποιημένη έκδοση του SpecAugment ενώ η χρήση μάσκας κάλυψης γίνεται επειδή μπορεί να χρησιμεύσει ως τεχνική κανονικοποίησης. Εν τέλει στο Σχήμα 3.15 φαίνονται συνοπτικά ποια από τα κομμάτια του μοντέλου Wav2Vec 2.0 συμμετέχουν είτε στο fine-tuning είτε στην πρώτη φάση της αυτο-εποπτευόμενη εκπαίδευσης:



Σχήμα 3.15: Fine-Tuning & Self-Supervised εκπαίδευση του Wav2Vec 2.0

3.3.1 Μοντέλο XLS-R (Cross-lingual Representation Learning for Speech Recognition)

Λίγο μετά την επιτυχία του Wav2Vec 2.0 σε ένα από τα πιο δημοφιλή σύνολα δεδομένων της αγγλικής γλώσσας για ASR, που ονομάζεται LibriSpeech, το Facebook AI παρουσίασε μια πολύγλωσση έκδοση του Wav2Vec2, που ονομάζεται XLSR. Το XLSR αποτελείται από διαγλωσσικές αναπαραστάσεις ομιλίας και αναφέρεται στην ικανότητα του μοντέλου να μαθαίνει αναπαραστάσεις που είναι χρήσιμες σε πολλές γλώσσες.



Σχήμα 3.16: Πολυγλωσσικό μοντέλο XLS-R

Ο διάδοχος του XLSR, που ονομάζεται απλά XLS-R (αναφερόμενος στο XLM-R for Speech), κυκλοφόρησε τον Νοέμβριο του 2021 από τους Arun Babu, Changhan Wang, Andros Tjandra, et al. Το XLS-R χρησιμοποίησε σχεδόν μισό εκατομμύριο ώρες δεδομένων ήχου σε 128 γλώσσες για αυτοεπιβλεπόμενη προ-εκπαίδευση και διατίθεται σε μεγέθη που κυμαίνονται από 300 εκατομμύρια έως δύο δισεκατομμύρια παραμέτρους. Μπορούμε να βρούμε τα προεκπαιδευμένα σημεία (checkpoints) στο HuggingFace Hub ως: Wav2Vec2-XLS-R-300M, Wav2Vec2-XLS-R-1B, Wav2Vec2-XLS-R-2B. Παρόμοια με τον στόχο μοντελοποίησης καλυμμένης γλώσσας (masked language modeling) του BERT, το XLS-R μαθαίνει αναπαραστάσεις ομιλίας με βάση τα συμφραζόμενα καλύπτοντας τυχαία διανύσματα χαρακτηριστικών προτού τα διαβιβάσει σε ένα δίκτυο μετασχηματιστών (transformer network) κατά τη διάρκεια της αυτοεποπτευόμενης εκπαίδευσης, όπως φαίνεται στο Σχήμα 3.16. Για το ”κούρδισμα” των παραμέτρων (fine-tuning), προστίθεται ένα ενιαίο γραμμικό στρώμα στην κορυφή του προεκπαιδευμένου δικτύου για να εκπαιδεύσει το μοντέλο σε δεδομένα ήχου για τις περαιτέρω εργασίες (downstream task), όπως η αναγνώριση ομιλίας, η μετάφραση ομιλίας και η ταξινόμηση ήχου, όπως φαίνεται στο Σχήμα 3.16 στα δεξιά.

3.4 Αλγόριθμος Συνδετικής Χρονικής Ταξινόμησης (CTC)

Στην αναγνώριση ομιλίας έχουμε ένα σύνολο δεδομένων από ηχητικά κλιπ και τις αντίστοιχες μεταγραφές. Το πρόβλημα δυστυχώς είναι ότι δεν γνωρίζουμε πώς οι χαρακτήρες στη μεταγραφή ευθυγραμμίζονται με τον ήχο. Αυτό κάνει την εκπαίδευση ενός συστήματος αναγνώρισης ομιλίας πιο δύσκολη από ό,τι φαίνεται στην αρχή. Χωρίς αυτήν την ευθυγράμμιση, οι απλές προσεγγίσεις δεν είναι διαθέσιμες. Θα μπορούσαμε να επινοήσουμε έναν κανόνα όπως «ένας χαρακτήρας αντιστοιχεί σε δέκα εισόδους». Αλλά οι ρυθμοί ομιλίας των ανθρώπων ποικίλλουν, επομένως αυτός ο τύπος κανόνα μπορεί πάντα να παραβιάζεται. Μια άλλη εναλλακτική είναι να ευθυγραμμίσουμε με το χέρι κάθε χαρακτήρα στη θέση του στον ήχο. Από την άποψη της μοντελοποίησης, αυτό λειτουργεί καλά, θα γνωρίζαμε τη βασική αλήθεια για κάθε χρονικό βήμα εισόδου. Ωστόσο, για οποιοδήποτε σύνολο δεδομένων λογικού μεγέθους, αυτό είναι απαγορευτικά χρονοβόρο. Αυτό το πρόβλημα δεν εμφανίζεται μόνο στην αναγνώριση ομιλίας. Το βλέπουμε και σε πολλές άλλες εφαρμογές. Η αναγνώριση χειρόγραφου από εικόνες ή ακολουθίες πινελιών με στυλό είναι ένα παράδειγμα, ενώ η επισήμανση ενεργειών (action labelling) στα βίντεο είναι άλλη. Η Συνδετική Χρονική Ταξινόμηση (CTC) είναι ένας τρόπος να λυθεί το πρόβλημα χωρίς να γνωρίζουμε την ευθυγράμμιση μεταξύ της εισόδου και της εξόδου και είναι ιδιαίτερα κατάλληλο για εφαρμογές όπως η αναγνώριση ομιλίας και γραφής, όπως φαίνεται π.χ. στο Σχήμα 3.17.

Σχήμα 3.17: Αναγνώριση χειρόγραφου από εικόνες (αριστερά), Αναγνώριση ομιλίας από φασματογράφημα (δεξιά)

Χρειαζόμαστε να βρούμε μια ακριβή αντιστοιχία της ακολουθίας εισόδου $X = [x_1, x_2, ..., x_T]$, όπως ήχος, σε αντίστοιχες ακολουθίες εξόδου $Y = [y_1, y_2, ..., y_U]$, όπως μεταγραφές. Υπάρχουν προκλήσεις που μας εμποδίζουν να χρησιμοποιούμε απλούστερους αλγόριθμους εποπτευόμενης μάθησης. Συγκεκριμένα:

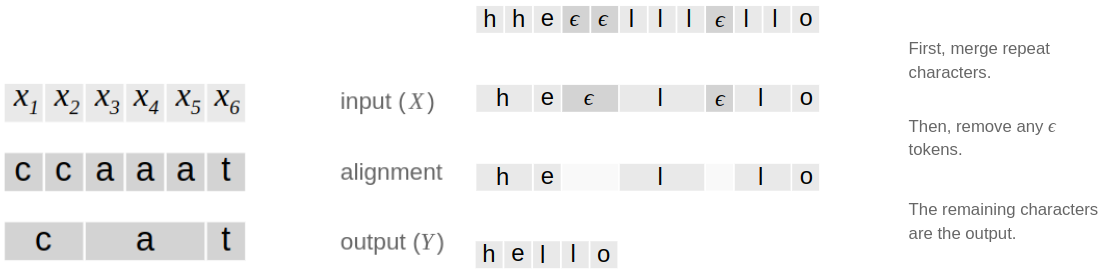
- Τα X και Y μπορεί να ποικίλλουν σε μήκος.
- Ο λόγος των μηκών των X και Y μπορεί να ποικίλλει.
- Δεν έχουμε ακριβή αντιστοιχία των στοιχείων των X και Y.

Ο αλγόριθμος CTC ξεπερνά αυτές τις προκλήσεις. Για ένα δεδομένο X μας δίνει μια κατανομή εξόδου σε όλα τα δυνατά Y. Μπορούμε να χρησιμοποιήσουμε αυτήν την κατανομή είτε για να συμπεράνουμε μια πιθανή έξοδο είτε για να εκτιμήσουμε την πιθανότητα μιας δεδομένης εξόδου. Για μια δεδομένη είσοδο, θα θέλαμε να εκπαιδεύσουμε το μοντέλο μας ώστε να μεγιστοποιήσει την πιθανότητα που εκχωρεί στη σωστή απάντηση. Για να γίνει αυτό κατά την εκπαίδευση, θα χρειαστεί να υπολογίσουμε αποτελεσματικά την υπό όρους πιθανότητα $p(Y \square X)$, ενώ η συνάρτηση αυτή θα πρέπει επίσης να είναι διαφοροποιήσιμη, ώστε να μπορούμε να χρησιμοποιήσουμε τη μέθοδο καθόδου κλίσης (gradient descent). Για την πρόβλεψη (inference) φυσικά, αφού έχουμε εκπαιδεύσει το μοντέλο, θέλουμε να το χρησιμοποιήσουμε για να συμπεράνουμε ένα πιθανό Y με δεδομένο X. Την επίλυση δηλαδή του:

$$Y^* = \underset{Y}{\operatorname{argmax}}\, p(Y|X), \tag{3.14}$$

που ιδανικά μπορεί να βρεθεί αποτελεσματικά, όμως με το CTC θα αρκεστούμε σε μια κατά προσέγγιση λύση που δεν είναι πολύ ακριβή.

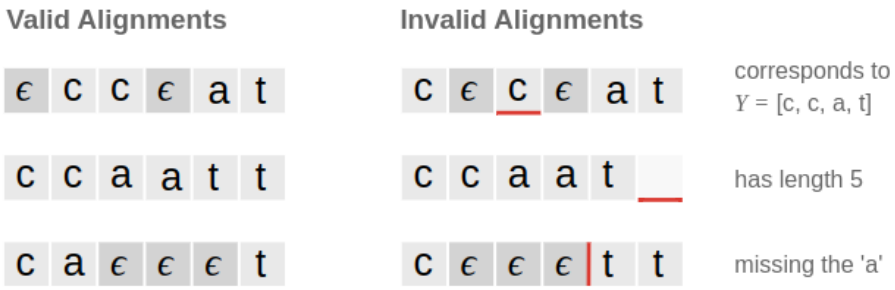
Ο αλγόριθμος CTC είναι χωρίς ευθυγράμμιση (alignment-free), δεν απαιτεί ευθυγράμμιση μεταξύ της εισόδου και της εξόδου. Ωστόσο, για να ληφθεί η πιθανότητα μιας εξόδου δεδομένης μιας εισόδου, το CTC λειτουργεί αθροίζοντας την πιθανότητα όλων των πιθανών ευθυγραμμίσεων μεταξύ των δύο. Πρέπει να καταλάβουμε ποιες είναι αυτές οι ευθυγραμμίσεις για να κατανοήσουμε πώς υπολογίζεται τελικά η συνάρτηση απώλειας. Ας υποθέσουμε ότι η είσοδος έχει μήκος έξι και ότι $Y = [c, a, t]$. Ένας τρόπος ευθυγράμμισης X και Y θα ήταν να να εκχωρήσουμε έναν χαρακτήρα εξόδου σε κάθε βήμα εισόδου και να συμπτύξουμε τις επαναλήψεις, όπως στο Σχήμα 3.18a. Αυτή η προσέγγιση όμως έχει 2 προβλήματα: α) Συχνά, δεν έχει νόημα να αναγκάζουμε κάθε βήμα εισόδου να ευθυγραμμιστεί με κάποια έξοδο. Στην αναγνώριση ομιλίας, για παράδειγμα, η είσοδος μπορεί να έχει τμήματα σιωπής χωρίς αντίστοιχη έξοδο. β) Δεν έχουμε τρόπο να παράγουμε εξόδους με πολλούς χαρακτήρες στη σειρά. Αν θεωρήσουμε τη στοίχιση [h, h, e, l, l, l, o] η σύμπτυξη των επαναλήψεων θα παράγει ”helo” αντί για ”hello”. Για να ξεπεραστούν αυτά τα προβλήματα, ο CTC εισάγει ένα νέο διακριτικό στο σύνολο των επιτρεπόμενων εξόδων. Αυτό το νέο διακριτικό ονομάζεται μερικές φορές το κενό διακριτικό και θα το αναφέρουμε εδώ ως ϵ ή ως ”$-$”. Το διακριτικό αυτό δεν αντιστοιχεί σε τίποτα και απλώς αφαιρείται από την έξοδο. Οι ευθυγραμμίσεις που επιτρέπονται από το CTC έχουν το ίδιο μήκος με την είσοδο. Επιτρέπουμε οποιαδήποτε στοίχιση στην οποία αντιστοιχίζεται στο Y μετά τη συγχώνευση των επαναλήψεων και την αφαίρεση των όρων ϵ, όπως φαίνεται στο Σχήμα 3.18b.



(a) Σύμπτυξη χαρακτήρων σε επανάληψη **(b)** Εισαγωγή του κενού διακριτικού

Σχήμα 3.18: Επίλυση ευθυγράμμισης στον CTC

Εάν το Y έχει δύο ίδιους χαρακτήρες διαδοχικά τότε μια έγκυρη ευθυγράμμιση θα πρέπει να έχει το ϵ ανάμεσά τους. Στο παρακάτω Σχήμα 3.19 φαίνονται μερικές έγκυρες και μη ευθυγραμμίσεις:



Σχήμα 3.19: Έγκυρες και μη ευθυγραμμίσεις

Οι ευθυγραμμίσεις CTC έχουν μερικές αξιοσημείωτες ιδιότητες. Πρώτον, οι επιτρεπόμενες ευθυγραμμίσεις μεταξύ X και Y είναι μονοτονικές. Εάν προχωρήσουμε στην επόμενη είσοδο, μπορούμε να διατηρήσουμε την αντίστοιχη έξοδο ίδια ή να προχωρήσουμε στην επόμενη. Μια δεύτερη ιδιότητα είναι ότι η ευθυγράμμιση του X στο Y ακολουθεί σχέση πολλά-προς-ένα (many-to-one), δηλαδή ένα ή περισσότερα στοιχεία εισόδου μπορούν να ευθυγραμμιστούν με ένα μόνο στοιχείο εξόδου αλλά όχι το αντίστροφο. Αυτό συνεπάγεται μια τρίτη ιδιότητα ότι το μήκος του Y δεν μπορεί να είναι μεγαλύτερο από το μήκος του X.

Πολλαπλές διαδρομές/ευθυγραμμίσεις μπορούν να δώσουν μια σωστή λύση, και επομένως, όλες οι σωστές λύσεις πρέπει να εξεταστούν. Ο ίδιος ο αλγόριθμος CTC όπως ειπώθηκε είναι ”χωρίς ευθυγράμμιση”, ωστόσο, αυτές οι «ψευδοευθυγραμμίσεις» χρησιμοποιούνται για τον υπολογισμό της μεγαλύτερης πιθανότητας. Έτσι παράγει μια κατανομή εξόδου σε όλα τα πιθανά Y, τα οποία μπορούν να χρησιμοποιηθούν για να συμπεράνουμε την πιθανότητα μιας συγκεκριμένης εξόδου, Y . Η υπό όρους πιθανότητα, $P(Y|X)$, υπολογίζεται αθροίζοντας όλες τις πιθανές ευθυγραμμίσεις μεταξύ της εισόδου και της εξόδου, όπως φαίνεται στο Σχήμα 3.20. Μαθηματικά μπορούμε να ορίσουμε την υπό συνθήκη πιθανότητα μιας αντιστοίχησης a_t, ως το γινόμενο κάθε κατάστασης στην ακολουθία:

$$P(\alpha|X) = \prod_{t=1}^{T} P(\alpha_t|X) \tag{3.15}$$

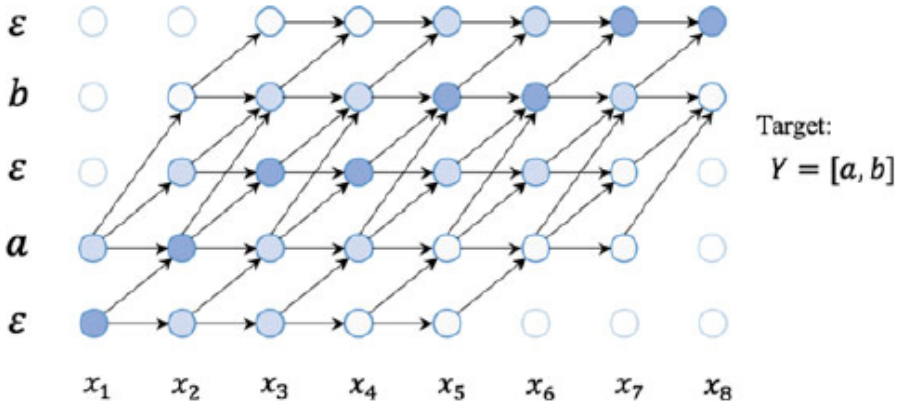
Όλες οι διαδρομές θεωρούνται αμοιβαία αποκλειόμενες, επομένως αθροίζουμε την πιθανότητα όλων των ευθυγραμμίσεων, δίνοντας την υπό όρους πιθανότητα για μία μόνο έκφραση/utterance (X, Y):

$$P(Y|X) = \sum_{A \in A_{x,y}} \prod_{t=1}^{T} P(\alpha_t|X) \tag{3.16}$$

όπου $A_{X,Y}$ είναι το σύνολο των έγκυρων ευθυγραμμίσεων. Ο δυναμικός προγραμματισμός χρησιμοποιείται για βελτίωση του υπολογισμού της συνάρτησης απώλειας CTC. Παρέχοντας τους κενούς όρους γύρω από την κάθε ετικέτα στην ακολουθία, τα μονοπάτια μπορούν εύκολα να συγκριθούν και να συγχωνευθούν όταν αυτά φτάνουν στην ίδια έξοδο στο ίδιο χρονικό βήμα. Ο συνδυασμός όλων δίνει την συνάρτηση απώλειας και αντικειμενικό στόχο βελτιστοποίησης για το CTC:

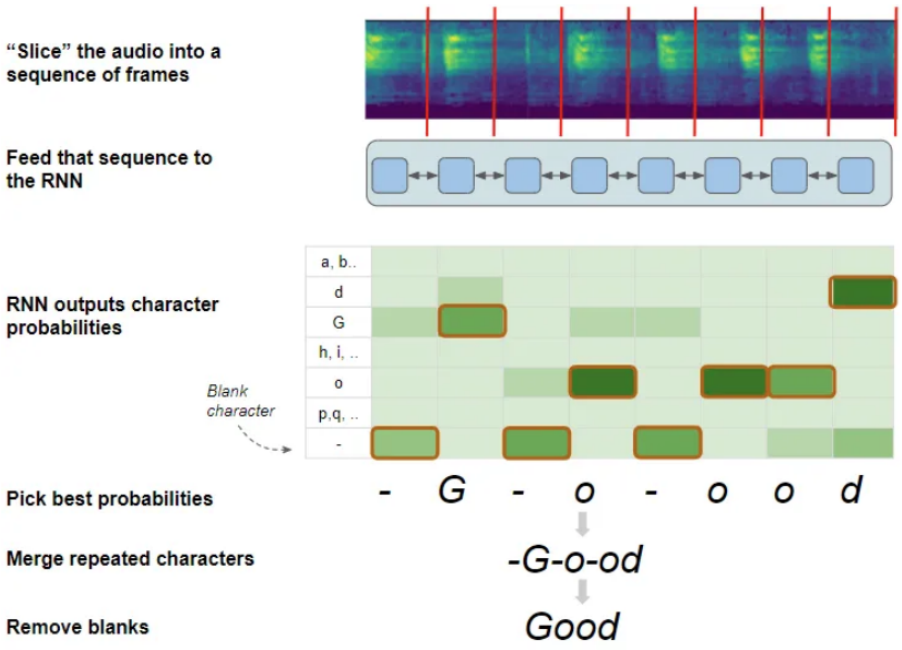
$$L_{ctc}(X,Y) = -log \sum_{A \in A_{x,y}} \prod_{t=1}^{T} P(\alpha_t|X) \tag{3.17}$$

Η κλίση για την αντίστροφη διάδοση μπορεί να υπολογιστεί για κάθε χρονικό βήμα από τις πιθανότητες του κάθε πλαισίου. Ο CTC υποθέτει υπό όρους ανεξαρτησία σε κάθε χρονικό βήμα, δηλαδή ότι ισχύει πως η έξοδος σε κάθε χρονικό βήμα είναι ανεξάρτητη από εκείνη στα προηγούμενα χρονικά βήματα. Αν και αυτή η υπόθεση επιτρέπει τη διάδοση κλίσης κατά πλαίσιο (frame-wise gradient propagation), περιορίζει την ικανότητα μάθησης διαδοχικών εξαρτήσεων. Η χρήση ενός γλωσσικού μοντέλου μετριάζει ορισμένα από τα ζητήματα, παρέχοντας ένα context λέξης ή n-gram context.



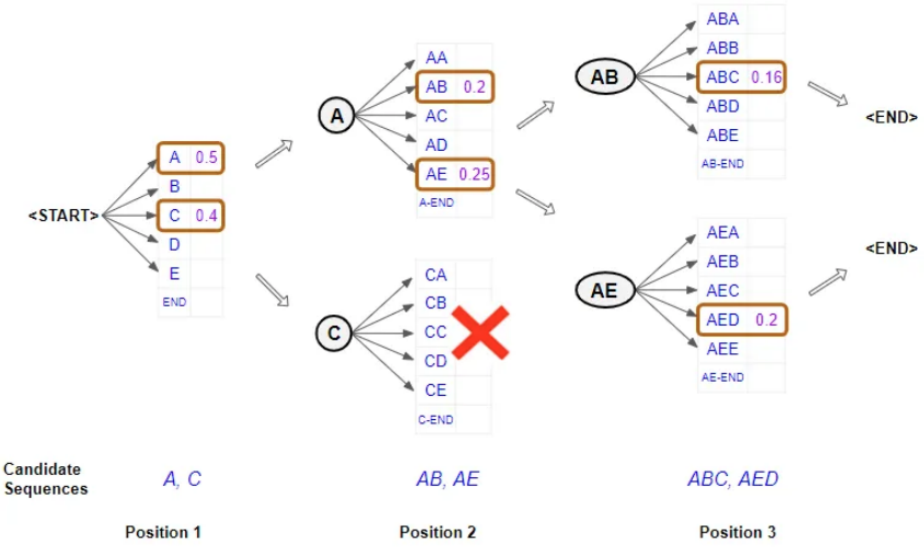
Σχήμα 3.20: Έγκυρα CTC μονοπάτια

Για την πρόβλεψη απλά χρησιμοποιούμε τις πιθανότητες χαρακτήρων για να επιλέξετε τον πιο πιθανό χαρακτήρα για κάθε καρέ, συμπεριλαμβανομένων των κενών. π.χ. ”-G-o-ood” και συγχονεύουμε τυχόν χαρακτήρες που επαναλαμβάνονται και δεν χωρίζονται με κενό. Για παράδειγμα, μπορούμε να συγχωνεύσουμε το ”oo” σε ένα μόνο ”o”, αλλά δεν μπορούμε να συγχωνεύσουμε το ”o-oo”. Αυτός είναι ο τρόπος με τον οποίο το CTC μπορεί να διακρίνει ότι υπάρχουν δύο ξεχωριστά ”o” και να παράγει λέξεις που γράφονται με επαναλαμβανόμενους χαρακτήρες. π.χ. ”-G-o-od”. Τέλος, εφόσον τα κενά έχουν εξυπηρετήσει τον σκοπό τους, αφαιρούνται όλοι οι κενοί χαρακτήρες κι έχουμε τη λέξη. π.χ. ”Good”, στο παράδειγμα που φαίνεται στο Σχήμα 3.21.



Σχήμα 3.21: Αποκωδικοποίηση CTC για πρόβλεψη/έξοδο

Κατά την αποκωδικοποίηση CTC που γίνεται στο inference, υποθέσαμε σιωπηρά ότι ο αλγόριθμος επιλέγει πάντα έναν χαρακτήρα, αυτόν με την υψηλότερη πιθανότητα σε κάθε χρονικό βήμα. Αυτό είναι γνωστό ως Άπληστη Αναζήτηση (Greedy Search). Ωστόσο, γνωρίζουμε ότι μπορούμε να έχουμε καλύτερα αποτελέσματα χρησιμοποιώντας μια εναλλακτική μέθοδο που ονομάζεται Αναζήτηση Δέσμης (Beam Search) και χρησιμοποιείται συχνά σε προβλήματα NLP, δεν είναι συγκεκριμένη για στα ASR συστήματα. Η Αναζήτηση Δέσμης κάνει δύο βελτιώσεις σε σχέση με την Άπληστη Αναζήτηση. Η Άπληστη Αναζήτηση εξετάζει κάθε θέση μεμονωμένα, όπως είπαμε και μόλις εντοπίσει την καλύτερη λέξη για αυτή τη θέση, με τη μεγαλύτερη δηλαδή πιθανότητα, δεν εξετάζει τι προηγήθηκε, δηλαδή στην προηγούμενη θέση ή μετά από αυτήν. Αντίθετα, η Αναζήτηση Δέσμης επιλέγει τις καλύτερες N-ακολουθίες (Σχήμα 3.22) ως τη δεδομένη στιγμή και εξετάζει τις πιθανότητες του συνδυασμού όλων των προηγούμενων λέξεων μαζί με τη λέξη στην τρέχουσα θέση.



Σχήμα 3.22: Αναζήτηση Δέσμης (Beam Search) με N=2

Με άλλα λόγια, ρίχνει τη «δέσμη φωτός της αναζήτησής του» λίγο πιο ευρεία από την Άπληστη Αναζήτηση, και αυτό είναι που του δίνει το όνομά του. Η υπερπαράμετρος «N» είναι γνωστή ως πλάτος δέσμης. Διαισθητικά είναι λογικό ότι αυτό μας δίνει καλύτερα αποτελέσματα σε σχέση με την Άπληστη Αναζήτηση, διότι, αυτό που πραγματικά μας ενδιαφέρει είναι η καλύτερη πλήρης πρόταση και μπορεί να μας διαφεύγει αν επιλέγαμε μόνο την καλύτερη μεμονωμένη λέξη σε κάθε θέση.

3.5 Γλωσσικό Μοντέλο KenLM

Ένα γλωσσικό μοντέλο, καθορίζει πόσο πιθανή είναι η εμφάνιση της πρότασης σε μια γλώσσα. Μακράν το πιο ευρέως χρησιμοποιούμενο γλωσσικό μοντέλο είναι το μοντέλο γλώσσας n-gram, το οποίο χωρίζει μια πρόταση σε μικρότερες ακολουθίες λέξεων και υπολογίζει την πιθανότητα με βάση μεμονωμένες πιθανότητες n-gram. Με δεδομένο ένα μεγάλο σώμα απλού κειμένου, θα θέλαμε να εκπαιδεύσουμε ένα μοντέλο γλώσσας n-gram και να υπολογίσουμε την πιθανότητα για μια αυθαίρετη πρόταση. Ένα απλό πιθανολογικό γλωσσικό μοντέλο, λοιπόν όπως έχει ήδη αναφερθεί, κατασκευάζεται υπολογίζοντας τις πιθανότητες n-gram, όπου n-gram είναι μια ακολουθία n λέξεων, το n είναι ένας ακέραιος αριθμός μεγαλύτερος από 0. Η πιθανότητα ενός n-gram είναι η υπό όρους πιθανότητα η τελευταία λέξη του n-gram να ακολουθεί τις συγκεκριμένες n-1 λέξεις (χωρίς την τελευταία λέξη). Πρακτικά, είναι το ποσοστό των εμφανίσεων της τελευταίας λέξης μετά τις n-1 λέξεις (grams) αφήνοντας την τελευταία λέξη εκτός. Αυτή η έννοια είναι μια υπόθεση Markov, όπου δεδομένου του n-1 gram στο παρόν, οι πιθανότητες n-gram στο μέλλον δεν εξαρτώνται από τα n-2, n-3, grams του παρελθόντος.

Υπάρχουν προφανή μειονεκτήματα αυτής της προσέγγισης. Το πιο σημαντικό, είναι ότι οι προηγούμενες n λέξεις επηρεάζουν την κατανομή πιθανοτήτων της επόμενης λέξης. Τα περίπλοκα κείμενα έχουν βαθύ πλαίσιο που μπορεί να έχει καθοριστική επίδραση στην επιλογή της επόμενης λέξης. Έτσι, το ποια είναι η επόμενη λέξη μπορεί να μην είναι εμφανές από τις προηγούμενες n-λέξεις, ούτε καν αν το n ισούται με 20 ή 50. Ένας όρος επηρεάζει μια προηγούμενη επιλογή λέξης, π.χ. η λέξη 'United' είναι πολύ πιο πιθανή αν ακολουθείται από τις λέξεις 'States of America'. Επιπλέον, είναι προφανές ότι αυτή η προσέγγιση δεν επεκτείνεται εύκολα: καθώς αυξάνεται το μέγεθος n, ο αριθμός των πιθανών μεταθέσεων εκτοξεύεται στα ύψη, παρόλο που οι περισσότερες από τις μεταθέσεις δεν εμφανίζονται ποτέ στο κείμενο. Όλες οι πιθανότητες (ή όλες οι μετρήσεις n-gram) πρέπει να υπολογιστούν και να αποθηκευτούν. Τα μη εμφανιζόμενα n-grams δημιουργούν πρόβλημα αραιότητας (sparsity όπως λέγεται), καθώς η κατανομή πιθανοτήτων μπορεί να είναι αρκετά χαμηλή (οι πιθανότητες λέξεων έχουν λίγες διαφορετικές τιμές, επομένως οι περισσότερες λέξεις έχουν την ίδια πιθανότητα).

Το KenLM Language Model Toolkit χρησιμοποιείται για να δημιουργήσουμε ένα μοντέλο γλώσσας n-gram, ενώ τελικά θα δημιουργήσει ένα αρχείο σε μορφή ARPA που θα περιέχει όλα τα μοντέλα N-gram και υπιοχώρησης (back-off), οπότε μπορείτε να υπολογίσουμε την πιθανότητα που δίνει το γλωσσικό μοντέλο για οποιεσδήποτε προτάσεις. Επιπλέον το KenLM χρησιμοποιεί μια μέθοδο εξομάλυνσης που ονομάζεται τροποποιημένη Kneser-Ney. Η εξομάλυνση είναι μια τεχνική για την προσαρμογή της κατανομής πιθανοτήτων σε n-grams για να γίνουν καλύτερες εκτιμήσεις των πιθανοτήτων προτάσεων. Για παράδειγμα, σε οποιαδήποτε n-gram σε μια πρόταση ερωτήματος που δεν εμφανίστηκε στο σώμα εκπαίδευσης θα εκχωρηθεί μια πιθανότητα μηδέν, αλλά αυτό είναι προφανώς λάθος. Δεν μπορούμε να καλύψουμε όλα τα πιθανά n-gram που θα μπορούσαν να εμφανιστούν σε μια γλώσσα ανεξάρτητα από το πόσο μεγάλο είναι το σώμα, και μόνο και μόνο επειδή το n-gram δεν εμφανίστηκε σε ένα σώμα δεν σημαίνει ότι δεν θα εμφανιζόταν ποτέ σε κανένα κείμενο. Το KenLM είναι πολύ αποδοτικό framework σε μνήμη και χρόνο. Η μορφή ARPA περιέχει τις λογαριθμικές πιθανότητες και τις τιμές των βαρών υποχώρησης για κάθε n-gram. Για να υπολογίσουμε την πιθανότητα μιας πρότασης, εξετάζουμε κάθε n-gram της πρότασης από την αρχή. Εάν το n-gram βρίσκεται στον πίνακα, απλώς διαβάζουμε την λογαριθμική πιθανότητα και την προσθέτουμε (καθώς είναι ο λογάριθμος, μπορούμε να χρησιμοποιήσουμε πρόσθεση αντί για γινόμενο μεμονωμένων πιθανοτήτων). Εάν το n-gram δεν βρίσκεται στον πίνακα, κάνουμε πίσω στο n-gram σε χαμηλότερη τάξη του και χρησιμοποιούμε την πιθανότητα του, προσθέτοντας τα βάρη back-off (και πάλι τα προσθέτουμε αφού

πρόκειται για λογαρίθμους). Ένα απόσπασμα του αρχείου μοντέλου γλώσσας ARPA είναι το εξής:

```
2-grams
-1.7037368      <s> I   -0.35425213
-3.1241505      a boy   -0.19261438
-1.9892355      am a    -0.08787394
-1.0562452      boy .   -0.19261438
.....

3-grams
-1.4910358      <s> I am
-1.1888235      I am a
-0.6548149      a boy .
-1.1425415      . </s>  0
.....
```

Κεφάλαιο 4

ASR Συγκεκριμένου Τομέα και Πειραματική Διαδικασία

4.1 ASR Συγκεκριμένου Τομέα

Παρόλο που η αυτόματη αναγνώριση ομιλίας σε διεργασίες γενικού σκοπού δουλεύει πολύ καλά, συχνά με ακρίβεια πάνω από 90%, σε κανονικές συνθήκες που περιλαμβάνουν ειδικού σκοπού λεξιλόγιο, όπως ορολογίες για παράδειγμα η απόδοση των συστημάτων μειώνεται δραστικά. Το πρόβλημα αυτό γίνεται περισσότερο ορατό όταν θέτουμε σε εφαρμογή ένα ASR σύστημα που έχει εκπαιδευτεί σε ηχητικά δεδομένα γενικού σκοπού και αναλαμβάνει να αναγνωρίσει ηχογραφήσεις που περιέχουν ορολογίες. Κατα συνέπεια λοιπόν προκαλείται μεγάλη αύξηση στην μετρική απόδοσης, που συνήθως είναι το Word Error Ratio (WER), της τάξης του 50%, που εξηγείται αφ' ενός από την παρουσία των πολλών "άγνωστων" ετικετών (<UNK>) που αντικαθιστούν τις λέξεις που βρίσκονται εκτός λεξιλογίου (out of vocabulary - OOV) και αφ' ετέρου από τις λέξεις γενικού σκοπού, πάνω στις οποίες έχει εκπαιδευτεί το μοντέλο, που τις χρησιμοποιεί για να αντικαταστήσει κάποιους από τους ειδικούς όρους, όπως για παράδειγμα η λέξη "Audi" μπορεί να αντικατασταθεί από τη λέξη "howdy".

Έτσι, ενώ λοιπόν οι εικονικοί βοηθοί, όπως το Siri και η Alexa μπορούν να λειτουργήσουν καλά χρησιμοποιώντας μοντέλα εκπαιδευμένα σε δεδομένα γενικού σκοπού, σπάνια μπορούν να ικανοποιήσουν εμπορικές ανάγκες όπως φωνητικά chatbots για μια εφαρμογή ή για ένα τηλεφωνικό κέντρο. Οι εφαρμογές αυτές απαιτούν αναγνώριση των ονομάτων των προϊόντων ή των υπηρεσιών που σχεδόν πάντα είναι εκτός λεξιλογίου και διαφορετικές για την κάθε εταιρεία, αφού συνήθως αναφέρονται σε ονόματα εξοπλισμών παραγωγής ή πιο πολύπλοκα όταν υπάρχουν διαφορετικά ονόματα για την ίδια συσκευή. Εξαιτίας όλων των παραπάνω απαιτείται η εκπαίδευση του μοντέλου σε δεδομένα που ανήκουν στο συγκεκριμένο πεδίο εφαρμογής, όμως στις περισσότερες περιπτώσεις αυτό δεν είναι δυνατό είτε γιατί οι εταιρείες δεν διαθέτουν αρκετό όγκο δεδομένων είτε δεν μπορούν να τα μοιραστούν λόγω νομικών κολλημάτων. Παρόλα αυτά ακόμα κι αν τα δεδομένα είναι λιγοστά είναι δυνατόν να προσαρμόσουμε το μοντέλο και να εμπλουτίσουμε την ικανότητα αναγνώρισης που διαθέτει. Αυτό μπορεί να επιτευχθεί με τις λεγόμενες τεχνικές fine-tuning, που περιλαμβάνουν τη χρήση ενός ευρέως προ-εκπαιδευμένου μοντέλου που έχει ήδη αποκτήσει τη γνώση πάνω στη γλώσσα γενικά και απαιτείται να μάθει να αναγνωρίζει συγκεκριμένες λέξεις και συνδυασμούς λέξεων. Αυτή η μέθοδος επιτρέπει να ξεπεραστεί το πρόβλημα των λέξεων εκτός λεξιλογίου (OOV) και να μειωθεί το WER του μοντέλου.

4.2 Πειραματική Διαδικασία

Η διαδικασία που ακολουθήσαμε είναι η εξής:

1. Κατεβάσαμε και συλλέξαμε τα βίντεο αγώνων ποδοσφαίρου με περιγραφή στα ελληνικά και στη συνέχεια έγινε εξαγωγή του ηχητικού σήματος και διαχωρισμός σε αρθρώσεις (utterances) διάρκειας 30 - 60 δευτερολέπτων.

2. Έπειτα στείλαμε τις ηχητικές αυτές προτάσεις στο API του Google ASR μέσω κώδικα και έτσι παρήχθησαν τα λεκτικά utterances στα ελληνικά, ως αρχικό κείμενο.

3. Χρησιμοποιήσαμε το Label Studio ώστε να γίνει η διόρθωση και η υποσημείωση(annotation) των δεδομένων, από όπου και εξαγάγαμε την τελική μορφή των δεδομένων εκπαίδευσης του μοντέλου μας.

4. Τέλος τροφοδοτήσαμε αυτά τα δεδομένα στο δικό μας ASR μοντέλο, εφαρμόσαμε κάποια πειράματα και αξιολογήσαμε την απόδοση του συστήματός μας κάνοντας συγκρίσεις με βάση το Word Error Rate (WER).

4.3 Σύνολο δεδομένων και προετοιμασία

Γενικά για την τροφοδότηση και εκπαίδευση του μοντέλου που κατασκευάστηκε χρησιμοποιήσαμε ένα υβριδικό σύνολο δεδομένων που αποτελείται από 3 διαφορετικά σετ, συγκεκριμένα:

- Το Common Voice (CV) [4], [27] είναι ένα σύνολο δεδομένων ομιλιών αποτελούμενο από μοναδικά .mp3 ηχητικά και τα αντίστοιχα αρχεία κειμένου. Υπάρχουν 26.119 καταγεγραμμένες ώρες, ενώ περιλαμβάνει επίσης δημογραφικά μεταδεδομένα όπως ηλικία, φύλο και προφορά. Το σύνολο δεδομένων αποτελείται από 17.127 επικυρωμένες ώρες σε 104 γλώσσες. Εμείς χρησιμοποιήσαμε τα ηχητικά που περιλαμβάνουν ελληνικά και ενέρχονται σε 10 επικυρωμένες (από τις 30 καταγεγραμμένες) ώρες. Σημειωτέον ότι στο συγκεκριμένο σετ ο οποιοσδήποτε μπορεί να συμβάλλει χρησιμοποιώντας απλά το μικρόφωνό του, θα πρέπει όμως το ηχητικό και το αντίστοιχο κείμενο να περάσουν πρώτα απο επικύρωση πριν ενσωματωθούν στο κυρίως τμήμα (corpus).

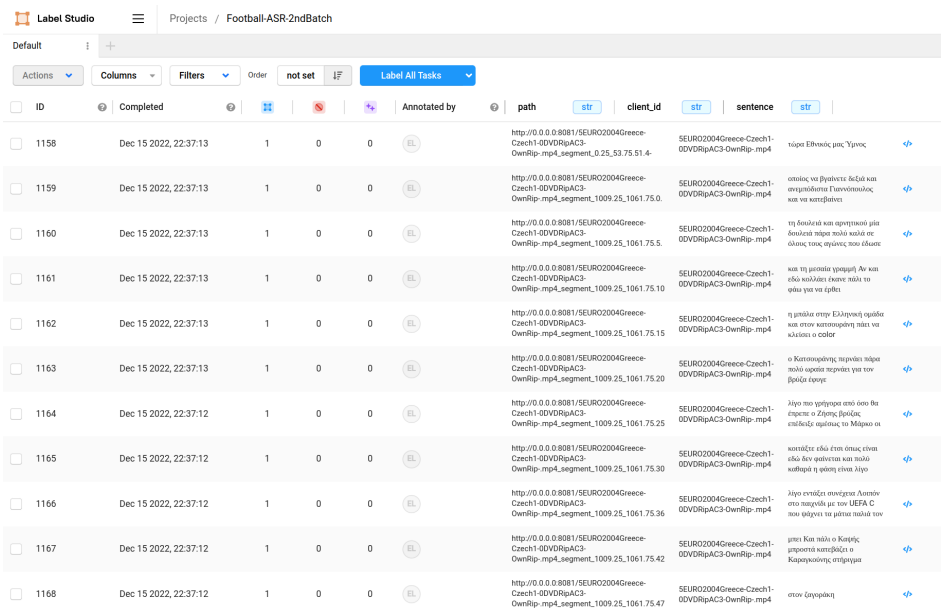
- Το δεύτερο σετ αποτελείται από βίντεο που περιλαμβάνουν συζητήσεις, παρουσιάσεις, εκπομπές, συνεδριάσεις της ελληνικής βουλής και αντλήθηκαν από το διαδίκτυο (προερχόμενα από το youtube) μαζί με τις αντίστοιχες μεταγραφές κειμένου. Όπως και στο ελληνικό Common Voice δεν χρειάστηκε να παρέμβουμε με διορθώσεις ή και υποσημείωση του κειμένου για την παρούσα δουλειά, καθώς οι μεταγραφές ήταν ήδη διαθέσιμες.

- Στο τρίτο σετ τα πρωτόγονα δεδομένα είναι σε μορφή βίντεο .mp4, συνολικής διάρκειας 12 ωρών και περιλαμβάνουν τις περιγραφές 5 ποδοσφαιρικών αγώνων. Η εξαγωγή της ηχητικής μορφής (.wav) αλλά και η τμηματοποίηση σε προτάσεις έγινε με κώδικα που χρησιμοποιεί το πακέτο $pyAudioAnalysis$ της $python$. Συνολικά παρήχθησαν 6.700 utterances που χρησιμοποιήθηκαν σε 2 ”δόσεις” για να τροφοδοτήσουν το μοντέλο μας. Τα utterances αφού εστάλησαν στο Google ASR για μια πρώτη αρχική μεταγραφή στα ελληνικά, φορτώθηκαν στο Label Studio όπου εκεί έγινε χειροκίνητα η διόρθωση και η τελική μεταγραφή του κειμένου. Παρακάτω φαίνεται ένα στιγμιότυπο των δεδομένων πριν γίνουν εξαγωγή από το Label Studio [18], όπως φαίνεται και στο Σχήμα 4.1.

Συνοπτικά λοιπόν είχαμε:

Σετ Δεδομένων	Διάρκεια (σε ώρες)	# train	# test	# validation
Common Voice (CV)	10	2721	410	0
Ελληνική Τηλεόραση (YouTube)	6	4186	1173	727
Ποδοσφαιρικοί αγώνες	12	4492	1951	0

Πίνακας 4.1: Σύνολο δεδομένων για εκπαίδευση του μοντέλου



Σχήμα 4.1: Διόρθωση και υποσημείωση δεδομένων στο Label Studio [18]

Κατά την προεπεξεργασία κειμένου και στα 3 σετ το γράμμα ’ς’ κανονικοποιείται σε ’σ’, ο λόγος είναι ότι και τα δύο γράμματα ακούγονται το ίδιο με το ’ς’ να χρησιμοποιείται μόνο ως χαρακτήρας λήξης των λέξεων. Έτσι, η αλλαγή μπορεί να αντιστοιχιστεί εύκολα στη σωστή υπαγόρευση. Επίσης αφαιρέθηκαν όλοι οι τόνοι από τα γράμματα και τα σημεία στίξης, όπως δείχνει και το Σχήμα 4.2 και χρησιμοποιήθηκαν 24 γράμματα και ο κενός χαρακτήρας για το ακουστικό μοντέλο, πράγμα που βοηθά στην απλούστευση του προβλήματος και βελτιώνει σημαντικά το WER. Η έξοδος του ακουστικού μοντέλου διορθώνεται με τη χρήση ενός γλωσσικού μοντέλου, το οποίο επιβάλλει τη σωστή ορθογραφία και τους συντακτικούς κανόνες στην έξοδο. Ένα άλλο πράγμα που θα μπορούσαμε να δοκιμάσουμε θα ήταν να αλλάξουμε όλα τα ι, η ... κτλ σε έναν μόνο χαρακτήρα αφού όλα ακούγονται το ίδιο, ενώ παρόμοια για το ο και το ω, και ίσως να βοηθούσαν σημαντικά το ακουστικό μοντέλο αφού όλοι αυτοί οι χαρακτήρες χαρτογραφούνται στον ίδιο ήχο.

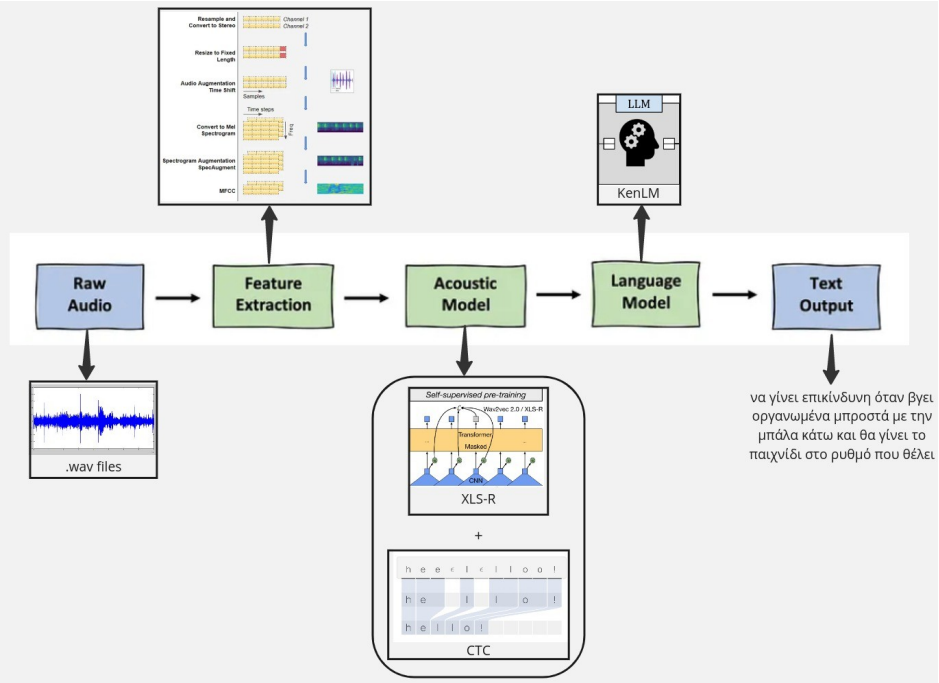
```
Target text: ευρωπαικο πρωταθλημα με το οποιο ητανε καθαρο δεν εγιναν σκοπιμα φαουλ
Input array shape: 69632
Sampling rate: 16000
Target text: το πουλμαν λεσ και το ξεραν οι πορτογαλοι εγραψαν στο πουλμαν ειχε δωδεκα θεουσ
Input array shape: 86016
Sampling rate: 16000
Target text: απο τον κατσουρανη το πλαγιο για τουσ τσεχουσ πλανα απο την εξεδρα τησ τσεχιασ κι εδω
Input array shape: 81920
Sampling rate: 16000
Target text: του ονομα ειναι μιγκελ σουαρεσ περειρα ριμπειρο γκομεζ το διαλεξε ετσι επειδη του
Input array shape: 83968
Sampling rate: 16000
Target text: μπεραρντι στο οριο ο βερατι δεν το δοκιμασε ο μπεραρντι ειχε κανει καλη κινηση παντωσ κιεζα
Input array shape: 106496
Sampling rate: 16000
Target text: ξανα ροζιτσκι ξανα κολερ σουτ πανω στην κινηση χαθηκε η μπαλα σαυτη τη φαση ουτε να τη δουμε δεν
Input array shape: 86016
Sampling rate: 16000
Target text: γιαννακοπουλοσ παιρνει την κεφαλια να γυρισει για το βρυζα και το χαριστεα ειναι και αδυναμη ομωσ εκει η κεφαλια
Input array shape: 79872
Sampling rate: 16000
Target text: εξουδετερωσει τελειωσ τον πανυψηλο κολερ τον πιο επικινδυνο παικτη τησ τσεχιασ
Input array shape: 120832
Sampling rate: 16000
Target text: αυτοσ που μπαινει δεν ειναι καλοσ βεβαια προσ θεου ο βλαντιμιρ σμιτσερ μεγαλοσ παικτησ και αυτοσ τριανταενα ετων στη λιβερπουλ
Input array shape: 120832
Sampling rate: 16000
Target text: διδυμο στην ντορτμουντ ο ψηλοσ με τον κοντο συνδυαζονται με κλειστα και βρισκονται με κλειστα ματια
Input array shape: 81920
Sampling rate: 16000
```

Σχήμα 4.2: Δείγμα των δεδομένων κατά την εκπαίδευση

4.4 Ακουστικό Μοντέλο και Μοντέλο Γλώσσας

Τα κομμάτια του E2E μοντέλου που παρουσιάζουμε στο Σχήμα 4.3 έχουν περιγραφεί αναλυτικά στο προηγούμενο κεφάλαιο, οπότε επιγραμματικά: το Ακουστικό μοντέλο είναι το πολυγλωσσικό XLS-R της Meta, που βασίζεται στη αρχιτεκτονική wav2vec 2.0 και είναι Tranformer-based και εκ-

παιδεύεται (ή γίνεται fine-tune, αναλόγως την περίπτωση) με τον αλγόριθμο CTC. Για το γλωσσικό μοντέλο χρησιμοποιήθηκε το KenLM Toolkit, όπως έχει αναφερθεί ανήκει στην κατηγορία των n-gram μοντέλων και έχει εκπαιδευτεί στην ελληνική γλώσσα για τη δική μας χρήση.



Σχήμα 4.3: End-to-End ASR μοντέλο που κατασκευάσαμε

4.5 Πειράματα και Συγκριτικά Αποτελέσματα

Μετά την εκπαίδευση του δικτύου μας, πρέπει να αξιολογήσουμε πόσο καλά αποδίδει. Η μέτρηση που θα χρησιμοποιήσουμε είναι το ποσοστό σφάλματος λέξης, η οποία συγκρίνει την προβλεπόμενη έξοδο και τη μεταγραφή στόχου, λέξη προς λέξη για να υπολογίσει τον αριθμό των διαφορών μεταξύ τους. Μια διαφορά θα μπορούσε να είναι μια λέξη που υπάρχει στη μεταγραφή αλλά λείπει από την πρόβλεψη (υπολογίζεται ως Διαγραφή), μια λέξη που δεν είναι στη μεταγραφή αλλά έχει προστεθεί στην πρόβλεψη (μια Εισαγωγή) ή μια λέξη που έχει τροποποιηθεί μεταξύ της πρόβλεψης και της μεταγραφής (μια Αντικατάσταση).

Πραγματοποιήσαμε τα παρακάτω πειράματα, με τη μετρική απόδοσης WER να φαίνεται στον Πίνακα 4.2 για το καθένα από αυτά:

- Ως πρώτο πείραμα βάλαμε το ASR μοντέλο μας να αξιολογηθεί ενώ έχει εκπαιδευτεί στα δεδομένα του Common Voice και της Ελληνικής Τηλεόρασης, δηλαδή σε δεδομένα γενικής κι όχι στοχευμένης περιοχής.

- Σαν δεύτερο πείραμα πήραμε το εκπαιδευμένο μοντέλο του πρώτου πειράματος και κάναμε fine-tuning στα δεδομένα του σετ δεδομένων ποδοσφαίρου. Όπως φαίνεται και απο τον Πίνακα τα αποτελέσματα είναι πολύ καλά.

- Τέλος εκπαιδεύσαμε ξανά το μοντέλο αυτή τη φορά μόνο στα δεδομένα του ποδοσφαίρου. Η απόδοσή του είναι όπως αναμενόταν καλύτερη από του πρώτου πειράματος και εφάμιλλη του δεύτερου. Παρόλα αυτά η εκπαίδευση έρχεται με επιπλέον κόστος, που είναι αρκετά πιο πάνω από το fine-tuning που πραγματοποιήσαμε στο δεύτερο πείραμα.

Σετ Δεδομένων	# test set CV	# test set YT	# **test set Football**
CV + YT (training)	11.6	29.7	39.7
CV + YT (training) + Football (fine-tune)	-	35.6	**20.5**
Football (training)	-	46.2	21.7

Πίνακας 4.2: Αποτελέσματα Word Error Rate (WER) ανά πείραμα

4.6 Συμπεράσματα και Επεκτάσεις

4.6.1 Συμπεράσματα

Ερευνήθηκε και αξιολογήθηκε πειραματικά η κατασκευή ενός σύγχρονου συστήματος αυτόματης αναγνώρισης ομιλίας από άκρο σε άκρο και χρησιμοποιήθηκαν σύγχρονα μοντέλα βαθιάς μάθησης τόσο στην διεργασία μετατροπής από ομιλία σε κείμενο όσο και στο γλωσσικό μοντέλο. Αντιμετωπίστηκαν αρκετές δυσκολίες, όπως φυσικά συμβαίνει και με κάθε σετ δεδομένων πραγματικού κόσμου, όπως εξωτερικός θόρυβος ή ταυτόχρονη εκφορά λόγου από 2 ομιλητές, όμως τα πειράματα έδειξαν ξεκάθαρα μια συμφέρουσα εναλλακτική λύση στην κατασκευή ενός ASR συστήματος.

Έγινε μια επισκόπηση και δόθηκε μια σφαιρική εικόνα για την έννοια της Αυτο-Εποπτευόμενης Μάθησης (SSL), η οποία ναι μεν ήταν γνωστή στα πρώιμα στάδιά της, τώρα βλέπουμε ότι μπορεί να εφαρμοστεί και στο πεδίο που ασχολούμαστε, της αναγνώρισης ομιλίας. Δόθηκε έμφαση στις τρεις επικρατέστερες οικογένειες μοντέλων και τεχνικών E2E και συγκεκριμένα στη Συνδετική Χρονική Ταξινόμηση (Connectionist Temporal Classification - CTC), στον Κωδικοποιητή-Αποκωδικοποιητή που βασίζεται στην προσοχή (Attention-based Encoder-Decoder - AED) και στον Επαναλαμβανόμενο Μετατροπέα Νευρωνικού Δικτύου (Recurrent Neural Network-Transducer - RNN-T). Επιπλέον, περιγράφηκαν τα διαφορετικά σενάρια χρήσης του καθενός, εντοπίζοντας πλεονεκτήματα και μειονεκτήματα για το καθένα.

Παρόλο που τα γενικού σκοπού ASR συστήματα εμφανίζουν πειστική μεν αλλά όχι ιδιαίτερα καλή απόδοση όταν εφαρμόστηκαν σε συγκεκριμένο τομέα και πρόβλημα πρόβλεψης συμπεραίνουμε ότι η λιγότερο δαπανηρή εναλλακτική λύση που παρέχεται από ήδη εκπαιδευμένο μοντέλο και από την πλευρά μας εφαρμογή ”κουρδίσματος”/fine-tuning στον αντίστοιχο τομέα. Η λύση αυτή είναι η πιο αποτελεσματική, καθώς επιτρέπει τη μείωση του χρόνου που καταναλώνουμε για την κατασκευή του μοντέλου με αξιοσημείωτο αντίκτυπο στην ακρίβεια των προβλέψεων, οι οποίες γίνονται πολύ πιο εύστοχες και με χρήση πολύ λίγων δεδομένων, βλέπουμε δηλαδή μεγάλη βελτίωση ακόμα και με 10 ώρες annotated δεδομένων. Τέλος αξίζει τον κόπο να διερευνηθούν περαιτέρω οι δυνατότητες των σύγχρονων γλωσσικών μοντέλων για να υπάρξει ακόμα μεγαλύτερη βελτίωση, όπως συμβαίνει και σε άλλους τομείς όπως η επεξεργασία φυσικής γλώσσας όπου και χρησιμοποιούνται σαν κύριο εργαλείο.

4.6.2 Προτάσεις για επέκταση

Μια απλή επέκταση της μελέτης θα ήταν να χρησιμοποιηθούν και τα επιπλέον δεδομένα που μαζέψαμε, που αποτελούνται από ακόμη 10 ώρες περιγραφής αγώνων, να ακολουθηθεί η ίδια διαδικασία υποσημείωσης κι έτσι να αξιολογηθεί περαιτέρω η προβλεπτική ικανότητα του μοντέλου. Μια ενδιαφέρουσα κι αναμενόμενη παρατήρηση που προέκυψε κατά την ακρόαση και υποσημείωση των ηχητικών αποσπασμάτων ήταν η μεγαλύτερη ευκολία υποσημείωσης όσων αγώνων είναι πιο πρόσφατοι καθώς η ποιότητα και η καθαρότητα του ήχου βοηθά και το πρόβλημα του θορύβου είναι μειωμένο σημαντικό βαθμό.

Επιπλέον θα μπορούσαμε να χρησιμοποιήσουμε διαφορετικό γλωσσικό μοντέλο, είτε κάποια αρχιτεκτονική Recurrent Neural Network είτε βασισμένο σε Transformers, αντί για το ”old-school” στατιστικό που χρησιμοποιήσαμε. Η ιδέα αυτή όμως έρχεται και με έξτρα υπολογιστικό κόστος αλλά και πολυπλοκότητα, ενώ είναι άγνωστο το κατά πόσο θα μπορούσε να ανταποκριθεί σε συνθήκες πραγματικού χρόνου (ή σχεδόν πραγματικού χρόνου), π.χ. σε real-time μετάδοση/ηχογράφηση.

Μια ακόμη ιδέα είναι να δοκιμάσουμε το speech-to-text μοντέλο Whisper [15], [45] της OpenAI, για μεταγραφή, που είναι ένα μοντέλο κωδικοποιητή-αποκωδικοποιητή (encoder-decoder) βασιμένο σε Transformers. Προσφέρεται τόσο για inference, μέσω API calls στα προκαθορισμένα endpoints όσο και για fine-tuning στα δικά μας δεδομένα [20].

Τέλος θα είχε ενδιαφέρον να δοκιμάσουμε το μοντέλο μας σε μεταγραφή ομιλίας περιγραφής κάποιου άλλου αθλήματος, π.χ. καλαθοσφαίρισης ή αντισφαίρισης, ή ακόμη και σε εκπομπή που ασχολείται με αθλητικά, ή σε συνεντεύξεις τύπου προπονητών ή αθλητών.

Βιβλιογραφία

[1] *Asr inference with ctc decoder*. https://pytorch.org/audio/stable/tutorials/asr_inference_with_ctc_decoder_tutorial.html.

[2] *Asr language modeling*. https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/asr/asr_language_modeling.html.

[3] *Automatic speech recognition for specific domains*. https://medium.com/product-ai/automatic-speech-recognition-for-specific-domains-28f400fb9000.

[4] *Common voice open-source dataset by mozilla*. https://commonvoice.mozilla.org/en/datasets.

[5] *Essential guide to automatic speech recognition technology*. https://developer.nvidia.com/blog/essential-guide-to-automatic-speech-recognition-technology/.

[6] *Evaluating an automatic speech recognition service*. https://aws.amazon.com/blogs/machine-learning/evaluating-an-automatic-speech-recognition-service/.

[7] *Foundations of nlp explained — bleu score and wer metrics*. https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b.

[8] *Global conversational al market size*. https://www.globalmarketestimates.com/market-report/conversational-ai-market-3805.

[9] *Global speech-to-text market size*. https://www.globalmarketestimates.com/market-report/speech-to-text-market-3839.

[10] *A history of voice technology*. https://info.keylimeinteractive.com/history-of-voice-technology.

[11] *How rev.com harnesses human-in-the-loop and deep learning to build the world's best english speech recognition engine*. https://www.youtube.com/watch?v=sR6_bZ6VkAg.

[12] *How to build domain specific automatic speech recognition models on gpus*. https://developer.nvidia.com/blog/how-to-build-domain-specific-automatic-speech-recognition-models-on-gpus/.

[13] *How to create wav2vec2 with language model*. https://discuss.huggingface.co/t/how-to-create-wav2vec2-with-language-model/12703.

[14] *How to train and run a simple language model*. https://lukesalamone.github.io/posts/running-simple-language-model/.

[15] *Introducing whisper*. https://openai.com/research/whisper.

[16] *Introduction to automatic speech recognition (asr)*. https://maelfabien.github.io/machinelearning/speech_reco/#.

[17] *Kenlm model*. https://github.com/kmario23/KenLM-training.

[18] *Label studio*. https://labelstud.io/.

[19] *Language modeling*. https://www.techtarget.com/searchenterpriseai/definition/language-modeling.

[20] *Managed transcription with openai whisper and hugging face inference endpoints*. https://www.philschmid.de/whisper-inference-endpoints.

[21] *A python library for audio feature extraction, classification, segmentation and applications*. https://github.com/tyiannak/pyAudioAnalysis.

[22] *Speech recognition — review state-of-the-art papers — part 1*. https://medium.com/aiguys/speech-recognition-review-state-of-the-art-papers-part-1-ce37f6bc45e1.

[23] *Using the speech-to-text api with python*. https://codelabs.developers.google.com/codelabs/cloud-speech-text-python3#0.

[24] *What is word error rate (wer)?* https://www.trywingman.com/blog-posts/what-is-word-error-rate-in-automatic-speech-recognition.

[25] *Xls-r model*. https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec/xlsr.

[26] *Νέα Ελληνική Γλώσσα: Φθόγγοι, φωνήεντα και σύμφωνα*. https://users.sch.gr/ipap/NEGlossa/fon-sim_NE-Gl.htm.

[27] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, *Common voice: A massively-multilingual speech corpus*, in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 4211–4215.

[28] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, *Xls-r: Self-supervised cross-lingual speech representation learning at scale*, 2021.

[29] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, *wav2vec 2.0: A framework for self-supervised learning of speech representations*, 2020.

[30] K. Doshi, *Audio deep learning made simple: State-of-the-art techniques*, https://towardsdatascience.com/, (2021).

[31] S. Gautam, C. Midoglu, S. Shafiee Sabet, D. B. Kshatri, and P. Halvorsen, *Soccer game summarization using audio commentary, metadata, and captions*, in Proceedings of the 1st Workshop on User-Centric Narrative Summarization of Long Videos, NarSUM '22, New York, NY, USA, 2022, Association for Computing Machinery, p. 13–22.

[32] Georgian, *How to make an end to end automatic speech recognition system with wav2vec 2.0*, https://towardsdatascience.com, (2021).

[33] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, *Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks*, in Proceedings of the 23rd International Conference on Machine Learning, ICML '06, New York, NY, USA, 2006, Association for Computing Machinery, p. 369–376.

[34] M. Hagiwara, *Training an n-gram language model and estimating sentence probability*, https://masatohagiwara.net, (2021).

[35] A. Hannun, *Sequence modeling with ctc*, Distill, (2017). https://distill.pub/2017/ctc.

[36] K. He, *Automatic speech recognition: Breaking down components of speech*, https://towardsdatascience.com/, (2021).

[37] U. Kamath, J. Liu, and J. Whitaker, *Automatic Speech Recognition*, Springer International Publishing, Cham, 2019, pp. 369–404, 537–574.

[38] M. Kapronczay, *A beginner's guide to language models*, https://towardsdatascience.com, (2021).

[39] J. Li, *Advancing end-to-end automatic speech recognition*, Tech. Rep. MSR-TR-2021-32, Microsoft, October 2021. Keynote talk at the conference on Computational Linguistics and Speech Processing, 2021.

[40] J. Li, *Recent advances in end-to-end automatic speech recognition*, APSIPA Transactions on Signal and Information Processing, (2022).

[41] J. C.-H. Lin, *Self-supervised learning (ssl) overview*, https://towardsdatascience.com/, (2022).

[42] A. Mukhamadiyev, I. Khujayarov, O. Djuraev, and J. Cho, *Automatic speech recognition method based on deep learning approaches for uzbek language*, Sensors, 22 (2022).

[43] I. Papastratis, *Speech recognition: a review of the different deep learning approaches*, https://theaisummer.com/, (2021).

[44] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, *End-to-end speech recognition: A survey*, 2023.

[45] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022.

[46] A. Sable, *End to end automatic speech recognition: State of the art*, https://www.paperspace.com, (2022).

[47] H. Scheidl, *An intuitive explanation of connectionist temporal classification*, https://towardsdatascience.com/, (2018).

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, 2017.

[49] P. von Platen, *Fine-tuning xls-r for multi-lingual asr with hugging face transformers*, huggingface.co, (2021). https://huggingface.co/blog/fine-tune-xlsr-wav2vec2.

[50] I. M. Yann LeCun, *Self-supervised learning: The dark matter of intelligence*, https://ai.facebook.com/blog/, (2021).

[51] Łukasz Sus, *Wav2vec 2.0: A framework for self-supervised learning of speech representations*, https://towardsdatascience.com, (2021).