
Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

τομεας τεχνολογιας πληροφορικης και υπολογιστων


# Μεταγραφή και Ευθυγράμμιση Στίχων με Σύγχρονες Τεχνικές Βαθιάς Μάθησης

## ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ


της

**Χριστίνας Κυπραίου**

**Επιβλέπων:** Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.



ΕΡΓΑΣΤΗΡΙΟ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΜΑΘΗΣΗΣ

Αθήνα, Ιούλιος 2023

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης

# Μεταγραφή και Ευθυγράμμιση Στίχων με Σύγχρονες Τεχνικές Βαθιάς Μάθησης

# ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

**Χριστίνας Κυπραίου**

**Επιβλέπων:** Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 6η Ιουλίου, 2023.

| ..... | ..... | ..... |
|---|---|---|
| Γεώργιος Στάμου | Αθανάσιος Βουλόδημος | Στέφανος Κόλλιας |
| Καθηγητής Ε.Μ.Π. | Επίκουρος Καθηγητής Ε.Μ.Π. | Καθηγητής Ε.Μ.Π. |

Αθήνα, Ιούλιος 2023

.....

**Χριστινα Κυπραιου**
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

*Στους γονείς μου*

# Περίληψη

Η παρούσα διπλωματική εργασία αποσκοπεί στην επίλυση του προβλήματος της μεταγραφής και ευθυγράμμισης στίχων με χρήση σύγχρονων μεθόδων βαθιάς μάθησης και στην σύγκριση των μοντέλων αυτών με κλασικά στατιστικά μοντέλα. Όσον αφορά τον τομέα Ανάκτησης Μουσικής Πληροφορίας (Music Information Retrieval) οι περισσότερες υπάρχουσες εφαρμογές επικεντρώνονται στη μεταγραφή του τόνου της φωνής του τραγουδιού, ελάχιστη έρευνα έχει γίνει για τη μεταγραφή των στίχων και την χρονική ευθυγράμμισή τους με το ηχητικό σήμα. Η αυτόματη ανάκτηση των στίχων τραγουδιών μπορεί να έχει σημαντικό αντίκτυπο στα εργαλεία σύνθεσης τραγουδιών, στις λεζάντες ήχου/βίντεο, στις εφαρμογές καραόκε, στη δημιουργία μουσικών καταλόγων, στη σύνθεση μουσικής, στη δημιουργία λιστών αναπαραγωγής και στην εκτίμηση πνευματικών δικαιωμάτων.

Το πρόβλημα της αυτόματης μεταγραφής στίχων είναι αντίστοιχο με το πρόβλημα της αυτόματης αναγνώρισης ομιλίας (ASR). Οι είσοδοι και των δύο συστημάτων είναι η ανθρώπινη φωνή και η αναμενόμενη έξοδος είναι οι μεταγραφές τους, ωστόσο το τραγούδι έχει συγκεκριμένα χαρακτηριστικά σε σύγκριση με τη φυσική ομιλία, τα οποία εισάγουν διάφορες προκλήσεις. Σε σύγκριση με την ευθυγράμμιση από κείμενο σε ομιλία, η ευθυγράμμιση στίχων παραμένει εξαιρετικά δύσκολη, παρά τις πολλές προσπάθειες να συνδυαστούν πλήθος επιμέρους μοντέλων, συμπεριλαμβανομένου του διαχωρισμού και της ανίχνευσης φωνής. Επιπλέον, η εκπαίδευση απαιτεί τη διαθεσιμότητα λεπτομερών επισημάνσεων σε συγκεκριμένη μορφή. Η αυτόματη αναγνώριση ομιλίας έχει σημειώσει σημαντική πρόοδο τα τελευταία χρόνια, ωστόσο το αντίστοιχο πρόβλημα στον τομέα του τραγουδιού πάσχει από περιορισμένα δεδομένα και υποβαθμισμένη κατανοησιμότητα των τραγουδισμένων στίχων.

Στην παρούσα διπλωματική εργασία επιχειρούμε να εκμεταλλευτούμε τις ομοιότητες μεταξύ ομιλίας και τραγουδιού. Πειραματιζόμαστε αρχικά με στατιστικά μοντέλα όπως τα Κρυφά Μοντέλα Markov (HMM). Στην συνέχεια δοκιμάζονται αρχιτεκτονικές νευρωνικών δικτύων όπως το Transformer που συνδυάζουν τον μηχανισμό προσοχής και την μοντελοποίηση από άκρο σε άκρο και αποτελούν τεχνικές της σύγχρονης ερευνητικής στάθμης (SOTA) στην Αυτόματη Αναγνώριση Ομιλίας. Επίσης εξετάζεται η επιρροή που ασκεί το πλήθος και η προέλευση των δεδομένων εκπαίδευσης.

**Λέξεις Κλειδιά** — Αυτόματη μεταγραφή στίχων, Ευθυγράμμιση ήχου με στίχους, Αυτόματη αναγνώριση ομιλίας, Ακουστική μοντελοποίηση, Γλωσσική μοντελοποίηση, Κρυφά μοντέλα Markov, Μηχανική μάθηση, Βαθιά νευρωνικά δίκτυα, Ακολουθία με ακολουθία, Μετασχηματιστές, Μεταφορά μάθησης

# Abstract

This thesis aims to solve the problem of transcription and alignment of lyrics using modern deep learning methods and to compare these models with classical statistical models. As far as the Music Information Retrieval domain is concerned most existing applications focus on transcribing the tone of the voice of the song, little research has been done on transcribing lyrics and their time alignment with the audio signal. Automatic retrieval of song lyrics can have a significant impact on song composition tools, audio/video captions, karaoke applications, music catalog creation, music composition, playlist creation, and copyright estimation.

The problem of automatic lyrics transcription is analogous to the problem of automatic speech recognition (ASR). The inputs of both systems are the human voice and the expected output is their transcriptions, however, singing has specific characteristics compared to natural speech, which introduce several challenges. Compared to speech-to-text alignment, lyric alignment remains extremely difficult, despite many attempts to combine numerous component models, including source separation and voice detection. In addition, training requires the availability of detailed annotations in a specific format. Automatic speech recognition has made significant progress in recent years, however the corresponding problem in the field of singing suffers from limited data and degraded intelligibility of sung lyrics.

In this thesis we attempt to exploit the similarities between speech and singing. We first experiment with statistical models such as Hidden Markov Models (HMMs). Then we test neural network architectures such as Transformer that combine the attention mechanism and end-to-end modeling and are state-of-the-art (SOTA) techniques in Automatic Speech Recognition (ASR). The impact of the quantity and origin domain of the training data is also studied.

**Keywords** — Automatic lyrics transcription, Audio-to-lyrics alignment, Automatic Speech Recognition, Acoustic Modeling, Language Modeling, Hidden Markov Models, Machine Learning, Deep Neural Networks, Sequence-to-sequence, Transformers, Transfer Learning

# Ευχαριστίες

Θα ήθελα καταρχάς να ευχαριστήσω τον καθηγητή κ. Γ. Στάμου για την επίβλεψη αυτής της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης. Επίσης ευχαριστώ ιδιαίτερα τον Ε. Δερβάκο για την καθοδήγησή του και την εξαιρετική συνεργασία που είχαμε. Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου και τους φίλους μου για την υποστήριξη που μου παρείχαν καθ όλη τη διάρκεια των σπουδών μου.

Χριστίνα Κυπραίου
Ιούλιος 2023

# Περιεχόμενα

# Λίστα Σχημάτων

# Κατάλογος Πινάκων

# Κεφάλαιο 1

# Εισαγωγή

## 1.1 Περιγραφή Εργασίας και Κίνητρα

Η μουσική και οι στίχοι είναι άρρηκτα συνδεδεμένοι στους περισσότερους ανθρώπινους πληθυσμούς παγκοσμίως, σχηματίζοντας ένα μέσο επικοινωνίας διαμέσου του τραγουδιού. Το τραγούδι με στίχους περιλαμβάνει τη συνύπαρξη δύο πολύπλοκων ακουστικών γνωστικών ικανοτήτων του ανθρώπου: της ομιλίας και της μουσικής. Σε αντίθεση με την αυθόρμητη ομιλία, το τραγούδι με στίχους μεταφέρει την επιδιωκόμενη πληροφορία μέσω της περίπλοκης αλληλεπίδρασης μεταξύ μουσικών και στιχουργικών δομών, εκτός από το γλωσσικό περιεχόμενο. Μεταγραφή είναι η διαδικασία μετατροπής προφορικού λόγου ή ηχογραφήσεων σε κείμενο. Ευθυγράμμιση είναι η διαδικασία συγχρονισμού του κειμένου με το αντίστοιχο ηχητικό. Η αυτόματη μεταγραφή στίχων (ALT) είναι μια υπολογιστική εργασία που βρίσκεται στη συμβολή της μουσικής, της ομιλίας και της επεξεργασίας της γλώσσας. Η παρούσα διπλωματική εργασία αποσκοπεί στην αντιμετώπιση αυτής της πρόκλησης με τη χρήση βαθιών νευρωνικών δικτύων (DNN) και τη διερεύνηση διαφόρων μεθόδων για τη βελτίωση και τη γενίκευση των επιδόσεων μεταγραφής στίχων.

Οι στίχοι μπορούν να χρησιμοποιηθούν για να εκφράσουν, να εξερευνήσουν και να συζητήσουν συναισθήματα, προβλήματα, προσωπικές ιδέες και πολιτικές απόψεις, να βοηθήσουν στην αντιμετώπιση καθημερινών προβλημάτων και να επηρεάσουν και να αντανακλούν τις καθημερινές ενέργειες και αποφάσεις του ακροατή, καθώς και τις συμπεριφορά τους ως προς την ακρόαση μουσικής. Οι ακροατές τείνουν να προτιμούν πλουσιότερους, πιο στοχαστικούς, πειστικούς, και συναισθηματικούς στίχους καθώς η προσθήκη στίχων σε ορχηστρικά τραγούδια προκαλεί ισχυρότερα συναισθήματα στους ακροατές. Όταν η μουσική και οι στίχοι έχουν αντιφατικές πληροφορίες για τη συναισθηματική φόρτιση, οι στίχοι έχουν μεγαλύτερη επιρροή στην προκαλούμενη διάθεση. Οι στίχοι των τραγουδιών μπορούν επίσης να δώσουν ενδείξεις για την κοινωνικοοικονομική κατάσταση των ακροατών. Από την άποψη αυτή, οι στίχοι αποτελούν ένα από τα βασικά στοιχεία της μουσικής που καθιερώνουν ένα έδαφος επικοινωνίας μεταξύ του τραγουδιστή/συνθέτη και του ακροατηρίου.

Με την επανάσταση των ψηφιακών μέσων, οι στίχοι των τραγουδιών έχουν επίσης αξιοποιηθεί εκτενώς για να για την οργάνωση και την ταξινόμηση μουσικών συλλογών. Ενώ οι στίχοι τραγουδιών μπορούν να επεξεργαστούν απευθείας από κείμενο, αυτοί συχνά δεν είναι διαθέσιμοι ή δεν αντιστοιχίζονται σωστά με τις μουσικές ηχογραφήσεις τους. Από αυτό το σημείο άποψη, η αυτόματη μεταγραφή των στιχουργικών δεδομένων έχει σπουδαίες προοπτικές για τη βελτίωση της δημιουργίας, ακρόασης και διανομής της μουσικής. Τέτοια εργαλεία αυτόματης μεταγραφής είναι γενικά αναφέρονται ως συστήματα αυτόματης μεταγραφής στίχων (ALT), τα οποία είναι ειδικά σχεδιασμένα για να μεταγράφουν τους τραγουδισμένους στίχους από τον ήχο σε κείμενο.

Εκ πρώτης όψεως, τα προαναφερθέντα συστήματα μπορούν να θεωρηθούν παρόμοια με τα συστήματα ομιλίας-σε κείμενο (ASR). Οι είσοδοι και των δύο συστημάτων είναι η ανθρώπινη φωνή και η αναμενόμενη έξοδος είναι οι ορθογραφικές μεταγραφές τους, ωστόσο το τραγούδι έχει συγκεκριμένα χαρακτηριστικά σε σύγκριση με τη φυσική ομιλία, τα οποία αποτελούν πρόκληση για τις τυπικές μηχανές ομιλίας-προς-κείμενο στην ακριβή μεταγραφή των τραγουδισμένων στίχων. Παρόμοια με αυτές τις μηχανές, η ευκρίνεια των λέξεων είναι χαμηλότερη για το τραγούδι από ό,τι για την ομιλία από τους ανθρώπινους ακροατές. Ένας από τους πιο

προφανείς λόγους γι' αυτό είναι ο τρόπος με τον οποίο προφέρονται τα φωνήεντα στο τραγούδι. Οι συγχύσεις που οφείλονται στις ηχητικές και χρονικές αλλαγές των φωνηέντων (οι οποίες οδηγούν επίσης σε σύγχυση στην αντίληψη των γύρω συμφώνων) μπορεί να είναι μια υποψήφια αιτία για τη χαμηλότερη ευκρίνεια των λέξεων.

Τα επιχειρήματα αυτά αφορούν κυρίως τους παράγοντες που σχετίζονται με τον ερμηνευτή (τραγουδιστή) και επηρεάζουν την κατανοησιμότητα των λέξεων, αν και δεν πρέπει να παραλείπονται οι παράγοντες που σχετίζονται με τον ακροατή και το περιβάλλον. Λόγω των ακουστικών φαινομένων (όπως η επικάλυψη), η κατανόηση των τραγουδισμένων λέξεων από τους ακροατές μπορεί να επηρεαστεί από το παρασκήνιο, ιδίως όταν υπάρχουν όργανα που συνοδεύουν τα φωνητικά. Η ακουστική και η αντήχηση του ακουστικού περιβάλλοντος και τα φωνητικά εφέ (όπως η τεχνητή αντήχηση, η χορωδία κ.λπ.) είναι μερικοί άλλοι σημαντικοί παράγοντες που επηρεάζουν την κατανοητότητα των λέξεων.

## 1.2 Ερευνητικός Στόχος και Συνεισφορά

Ορισμένες από τις άμεσες εφαρμογές του ALT περιλαμβάνουν τη λεζάντα/υποτιτλισμό βίντεο (π.χ. για καραόκε), την αναγνώριση μουσικής και την αναζήτηση μέσω τραγουδιού. Για το έργο της ευθυγράμμισης ήχου με στίχους - την αυτόματη ανάκτηση του χρονισμού των λέξεων σε μουσικά σήματα - οι προσεγγίσεις με τις καλύτερες επιδόσεις περιλαμβάνουν στον πυρήνα τους ένα προ-εκπαιδευμένο ακουστικό μοντέλο ALT. Από την άλλη πλευρά, το γλωσσικό μοντέλο εντός των μονάδων ALT μπορεί να χρησιμοποιηθεί για την παραγωγή στίχων. Τα μοντέλα ALT είναι επίσης χρήσιμα στη χρήση στίχων για την ανίχνευση διάθεσης/συναισθήματος τραγουδιού ή για τη βελτίωση του διαχωρισμού φωνητικών πηγών και την αναγνώριση τραγουδιών διασκευών. Επιπλέον, τα συστήματα αυτά μπορούν να χρησιμοποιηθούν σε μεθόδους μουσικοθεραπείας που αναπτύσσονται για την αντιμετώπιση γλωσσικών ελλειμμάτων. Ορισμένες άλλες πιθανές εφαρμογές του ALT περιλαμβάνουν τη σύνθεση μουσικής, τη δημιουργία λίστας αναπαραγωγής, τη σύσταση μουσικής και την πρόβλεψη δικαιωμάτων. Λαμβάνοντας υπόψη τις προαναφερθείσες εφαρμογές που θα μπορέσουν να υλοποιηθούν και να εμπλουτιστούν μέσω του ALT, η παρούσα διπλωματική έχει ως κίνητρο τη βελτίωση της δυνατότητας εφαρμογής αυτής της τεχνολογίας.

## 1.3 Δομή Διπλωματικής Εργασίας

Το περιεχόμενο της παρούσας διπλωματικής εργασίας είναι οργανωμένο σύμφωνα με την ακόλουθη δομή:

- Στο Κεφάλαιο 1 - Εισαγωγή, όπου περιγράφονται τα κίνητρα, οι στόχοι και η δομή της εργασίας.
- Στο Κεφάλαιο 2 - Θεωρητικό Υπόβαθρο, όπου εισάγονται βασικές έννοιες που αφορούν την αναπαράσταση, την επεξεργασία και την αναγνώριση φωνής. Παρουσιάζονται επίσης κλασικοί αλγόριθμοι αυτόματης αναγνώρισης ομιλίας καθώς και ευθυγράμμισης ομιλίας με κείμενο.
- Στο Κεφάλαιο 3 - Τρέχουσα τεχνολογική στάθμη (State-of-the-art), όπου περιγράφονται σύγχρονες τεχνικές μεταγραφής και ευθυγράμμισης με χρήση μηχανικής μάθησης και νευρωνικών δικτύων.
- Στο Κεφάλαιο 4 - Σύνολα Δεδομένων, όπου παρουσιάζεται το σύνολο δεδομένων και η διαδικασία προ-επεξεργασίας του.
- Στο Κεφάλαιο 5 - Αυτόματη Μεταγραφή Στίχων, στο οποίο περιγράφονται τα πειράματα και σχολιάζονται τα αποτελέσματα αναγνώρισης στίχων.
- Στο Κεφάλαιο 6 - Χρονική Ευθυγράμμιση Στίχων, όπου περιγράφεται η αρχιτεκτονική του συστήματος ευθυγράμμισης που χρησιμοποιήθηκε.
- Στο Κεφάλαιο 7 - Επίλογος, όπου συνοψίζονται τα ευρήματα της εργασίας και προτείνονται μελλοντικές επεκτάσεις.

# Κεφάλαιο 2

# Θεωρητικό Υπόβαθρο

## 2.1 Χαρακτηριστικά Φωνής

Θεμελιώδης σκοπός της φωνής είναι η ανθρώπινη επικοινωνία, δηλαδή η μετάδοση ενός μηνύματος από τον ομιλητή στον συνομιλητή του. Η φωνή είναι μια επίκτητη ικανότητα του ανθρώπου. Αναπτύσσεται, ελέγχεται και συντηρείται από την συνεχή ανατροφοδότηση του μηχανισμού ακοής, καθώς και των μυών υπεύθυνων για την παραγωγή της. Η πληροφορία που συγκροτείται από αυτά τα τμήματα του ανθρώπινου σώματος διαχειρίζεται από συγκεκριμένα τμήματα του εγκεφάλου τα οποία συντονίζουν ολόκληρη την λειτουργία της φωνής [1]. Η φωνή παρουσιάζει ιδιαίτερο ενδιαφέρον, τόσο ως προς τον μηχανισμό παραγωγής της όσο και στα χαρακτηριστικά της. Ο μηχανισμός παραγωγής της φωνής περιλαμβάνει την συνδυασμένη χρήση του μυαλού, της μύτης του στόματος, των πνευμόνων καθώς και κοιλοτήτων του ανθρώπινου σώματος. Τα χαρακτηριστικά της φωνής διαφέρουν από άνθρωπο σε άνθρωπο γεγονός που δυσχεραίνει την κατασκευή υπολογιστικών συστημάτων αναγνώρισης και επεξεργασίας.

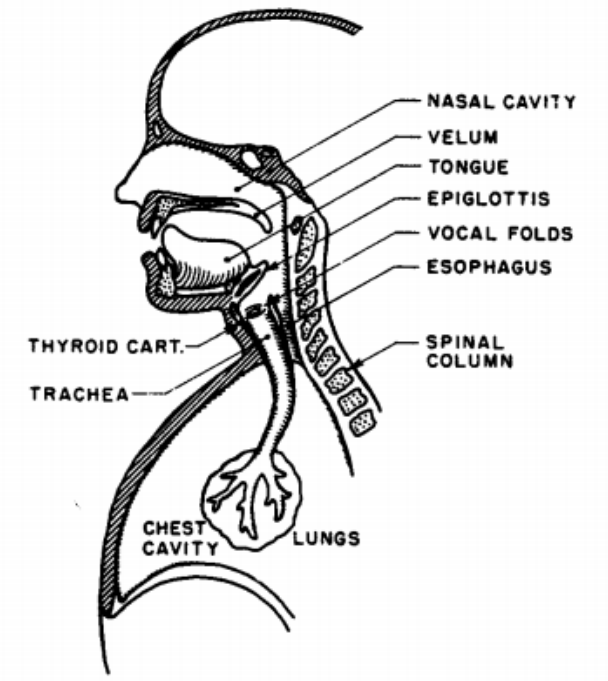
### 2.1.1 Παραγωγή Φωνής

Στο Σχήμα 2.1.1 αναπαριστάται η ανθρώπινη φωνητική οδός καθώς και τα μέλη του ανθρώπινου σώματος υπεύθυνα για την παραγωγή της φωνής. Η φωνητική οδός αποτελείται από τον φάρυγγα και την στοματική κοιλότητα. Ξεκινά από το άνοιγμα ανάμεσα στις φωνητικές χορδές, την γλωττίδα, και καταλήγει στα χείλη. Σε συγκεκριμένες περιπτώσεις παραγωγής φωνής χρησιμοποιείται και η ρινική οδός η οποία ξεκινά από τον ουρανίσκο και καταλήγει στα ρουθούνια.

Η διαδικασία παραγωγή φωνής διαφέρει στην περίπτωση έμφωνων και άφωνων ήχων. Τα μέλη της φωνητικής ή της ρινικής οδού αναγκάζονται να μετατοπιστούν προκειμένου να παραχθεί ο ζητούμενος ήχος. Στην γενική περίπτωση παραγωγής φωνής εισέρχεται αέρας μέσω της αναπνευστικής οδού. Καθώς ο αέρας εξέρχεται από τους πνεύμονες μέσω της τραχείας, οι φωνητικές χορδές πάλλονται περιοδικά. Στη συνέχεια οι περιοδικοί παλμοί διαμορφώνονται ως προς την συχνότητα τους και τέλος διοχετεύονται μέσω του φάρυγγα στην στοματική και πιθανώς στην ρινική κοιλότητα. Στην περίπτωση έμφωνων ήχων, όπως φωνηέντων, ο ήχος που παράγεται εξαρτάται άμεσα από την θέση της γλώσσας, του σαγονιού, και των χειλών. Οι άφωνοι ήχοι παράγονται μέσω ροής εξερχόμενου αέρα με σταθερό ρυθμό και στένωσης της φωνητικής οδού σε κάποιο σημείο της. Ανάλογα με το σημείο σύσφιξης της φωνητικής οδού παράγεται ο αντίστοιχος άφωνος ήχος. Ενδεικτικά για την παραγωγή του /$\varphi$/ η σύσφιξη γίνεται κοντά στα χείλη, ενώ για την παραγωγή του /$\theta$/ η σύσφιξη γίνεται στα δόντια.

#### Αλυσίδα Ομιλίας - *Speech Chain*

Η διαδικασία παραγωγής φωνής που αναφέρθηκε παραπάνω προϋποθέτει πως ο ομιλητής έχει κωδικοποιήσει το μήνυμα που επιθυμεί να μεταδώσει στον συνομιλητή του. Η κωδικοποίηση αυτή αντιστοιχεί σε μια ακολουθία συμβόλων στο λεξιλόγιο της γλώσσας. Από την πλευρά του ο συνομιλητής αποκωδικοποιεί την ακολουθία




Σχήμα 2.1.1: Αναπαράσταση παραγωγής φωνής.

ερμηνεύοντας το περιεχόμενο της. Ύστερα κωδικοποιεί το δικό του μήνυμα σε μια νέα ακολουθία και η διαδικασία επαναλαμβάνεται. Μια πλήρης αναπαράσταση της παραγωγής και λήψης μηνυμάτων προτάθηκε από τους *Denes* και *Pinson* [2].

Η αναπαράσταση της *speech chain* παρουσιάζεται στο Σχήμα 2.1.2. Σύμφωνα με την αναπαράσταση της αυτή τα μεταδιδόμενα μηνύματα ακολουθούν συγκεκριμένη διαδρομή ανάμεσα σε 3 διακριτά επίπεδα. Η διαδρομή είναι η εξής: γλωσσικό επίπεδο (*linguistic level*) στο οποίο επιλέγονται οι βασικοί ήχοι επικοινωνίας με σκοπό την έκφραση ενός μηνύματος, φυσικό επίπεδο (*physiological level*) στο οποίο οι συνιστώσες της φωνητικής οδού παράγουν τους αντίστοιχους ήχους, ακουστικό επίπεδο (*acoustic level*) στο οποίο το κωδικοποιημένο μήνυμα εξέρχεται μέσω της φωνής και λαμβάνεται τόσο από τον ομιλητή όσο και από τον συνομιλητή. Από την πλευρά του συνομιλητή, υπάρχει ομοίως το physiological level στο οποίο η φωνή αναλύεται μέσω του ακουστικού συστήματος του συνομιλητή και *linguistic level* στο οποίο τελικά η φωνή αποκωδικοποιείται στο λεξιλόγιο της γλώσσας.

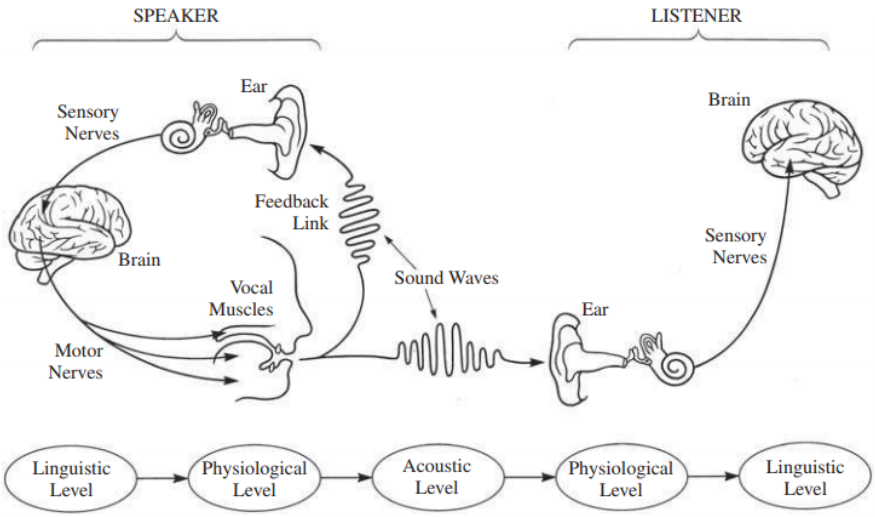
### 2.1.2 Εξαγωγή Χαρακτηριστικών Φωνής

#### Cepstrum χαρακτηριστικά

Το λογαριθμικό φάσμα είναι ένα συνεχές σήμα, κάτι που οφείλεται στην εξομάλυνση των παραθύρων. Επίσης έχει και περιοδικό χαρακτήρα, καθώς οι βασικές του κορυφές βρίσκονται στα ακέραια πολλαπλάσια της θεμελιώδης συχνότητας. Τα πιο σημαντικά χαρακτηριστικά του όμως, τα οποία είναι και υπεύθυνα π.χ. για την διαφοροποίηση των φωνημάτων, είναι στην μακροσκοπική δομή της συνάρτησης. Για αυτό το λόγο, πολλές φορές προσπαθούμε να προσεγγίσουμε την φασματική περιβάλλουσα αντί να δουλεύουμε με τα mel spectrograms.

Γενικά, η περιβάλλουσα ενός σήματος ταλάντωσης ορίζεται ως η ομαλή καμπύλη που αποτελεί το περίγραμμα των ακραίων τιμών της. Αποτελεί μια γενίκευση της έννοιας του πλάτους, καθώς σε κάθε χρονική στιγμή η περιβάλλουσα δείχνει το στιγμιαίο πλάτος. Η φασματική περιβάλλουσα είναι η περιβάλλουσα του λογαριθμικού φάσματος του σήματος και περιέχει την πληροφορία των μακροσκοπικών χαρακτηριστικών της.

Ο τρόπος που μπορούμε να πάρουμε πληροφορία για την φασματική περιβάλλουσα είναι χρησιμοποιώντας τον




Σχήμα 2.1.2: Αναπαράσταση αλυσίδας φωνής.

αντίστροφο μετασχηματισμό Fourier παίρνοντας τα cepstrum χαρακτηριστικά:

$$\mathcal{C}\{x\} = \left|\mathcal{F}^{-1}\left\{\log\left(|\mathcal{F}\{x(t)\}|^2\right)\right\}\right|^2, \tag{2.1.1}$$

όπου $\mathcal{C}\{x\}$ δηλώνει το *cepstrum* του σήματος, $\mathcal{F}\{x\}$ είναι ο μετασχηματισμός Fourier και $\mathcal{F}^{-1}\{x\}$ είναι ο αντίστροφος μετασχηματισμός Fourier.

Το όνομα *cepstrum* προκύπτει από την λέξη *spectrum* για να αντικατοπρίζεται το γεγονός ότι αποτελεί μια περίπλοκη αναδιάταξη μετασχηματισμών. Επειδή προκύπτει από αντίστροφο μετασχηματισμό από το πεδίο της συχνότητας, μετριέται στον άξονα του χρόνου, όπως φαίνεται και στο σχήμα 2.1.3.

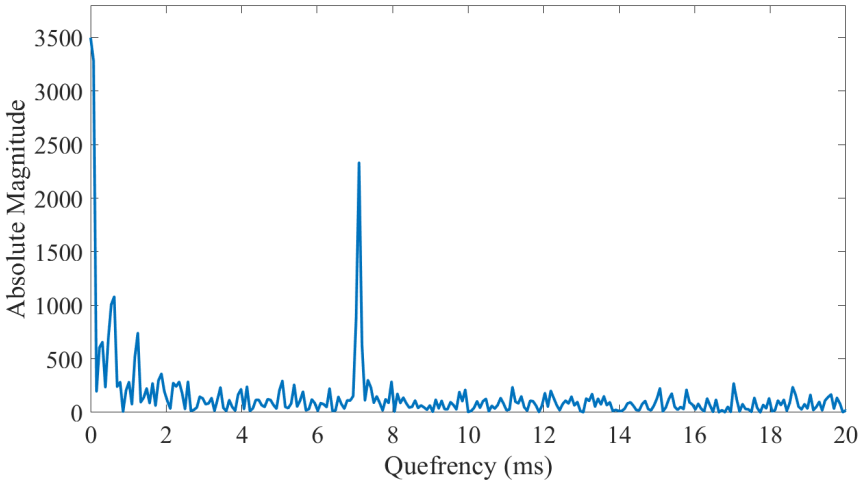
Οι τιμές του *cepstrum* που βρίσκονται στους χαμηλούς χρόνους δίνουν πληροφορία για τα χαρακτηριστικά του λογαριθμικού φάσματος που αλλάζουν με αργό ρυθμό, δηλαδή τα χαρακτηριστικά της φασματικής περιβάλλουσας που θέλουμε να μοντελοποιήσουμε. Επιπλέον, σε πιο μεγάλους χρόνους μπορούμε να δούμε χαρακτηριστικά του συχνοτικού περιεχομένου του φάσματος, με πιο ξεκάθαρη την θεμελιώδη συχνότητα η οποία στο σχήμα 2.1.3 αντιστοιχεί στην μεγάλη τιμή στα περίπου 7 ms. Τα 7 ms αντιστοιχούν σε συχνότητα περίπου 143 Hz οπότε έχουμε μία εκτίμηση της F0 του σήματος.

### Mel - Frequency Cepstral Coefficients (MFCCs)

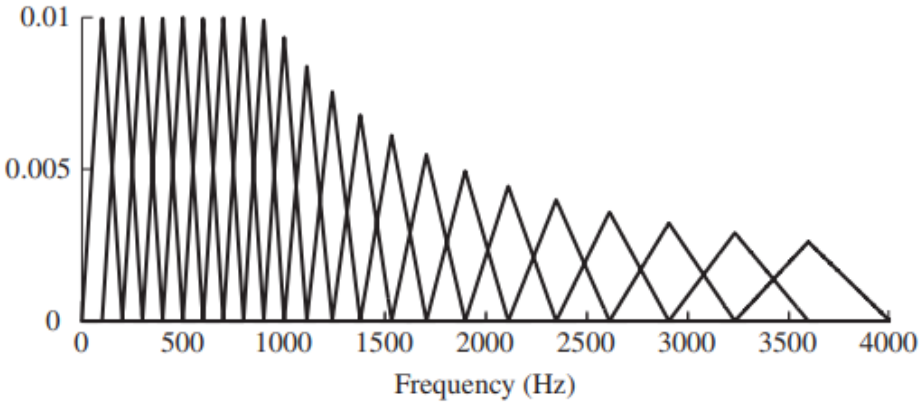
Τα *MFCCs* παρέχουν μια αναπαράσταση του σήματος προσομοιώνοντας τα εύρη ακοής του ανθρώπου. Οι συχνότητες του σήματος αναλύονται με βάση μια τριγωνική συστοιχία φίλτρων στην κλίμακα mel. Στην κλίμακα *Hz* τα φίλτρα διαθέτουν σταθερό εύρος ζώνης για χαμηλές συχνότητες (συνήθως $1Khz$) το οποίο αυξάνεται εκθετικά μέχρι την συχνότητα δειγματοληψίας $F_s$ του σήματος. Μια τριγωνική συστοιχία φίλτρων δίνεται στο Σχήμα 2.1.4.

Η κλίμακα *mel* (Σχήμα 2.1.5) αποσκοπεί στον μετασχηματισμό των συχνοτήτων από την κλίμακα *Hz* έτσι ώστε οι ίδιες να βρίσκονται ισοκατανεμημένες μεταξύ τους στο εύρος ακοής του ανθρώπου [3]. Επειδή ο άνθρωπος διακρίνει ευκολότερα μεταβολές σε μικρότερες συχνότητες από ότι σε μεγαλύτερες η κλίμακα *mel* είναι λογαριθμική. Παρ' όλα αυτά δεν υπάρχει μονοσήμαντος ορισμός της κλίμακας, διότι η αντίληψη συχνοτήτων του ανθρώπου είναι υποκειμενική. Μια φόρμουλα για την κλίμακα *mel* δίνεται ως:

$$m(f) = 1127 log_{10}(1 + \frac{f}{700}) \tag{2.1.2}$$




Σχήμα 2.1.3: Αναπαράσταση Cepstrum ενός ηχητικού σήματος.




Σχήμα 2.1.4: Τριγωνική συστοιχία φίλτρων.

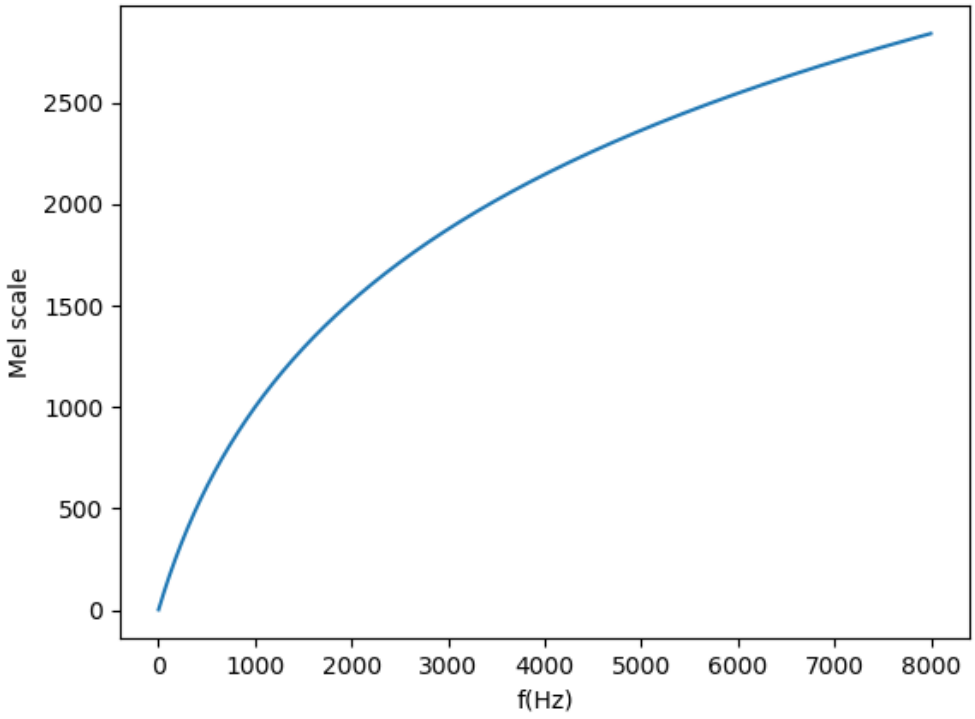
Ένα σήμα φωνής δίνεται στο Σχήμα 2.1.6. Αρχικά το σήμα χωρίζεται σε πλαίσια και παραθυρώνεται. Δημιουργούνται $l$ επικαλυπτόμενα πλαίσια μήκους $N$ δειγμάτων το καθένα. Το πρώτο βήμα για την εξαγωγή των *MFCCs* είναι ο Διακριτός Μετασχηματισμός Fourier (*Discrete Fourier Transform - DFT*) των πλαισίων από το πεδίο του χρόνου στο πεδίο της συχνότητας. Ο μετασχηματισμός ενός πλαισίου $x_m(n)$ για ένα τυχαίο πλαίσιο ορίζεται ως :

$$X_m[k] = \sum_{n=0}^{N-1} x_m[n] e^{(-j2\pi k/N)n} \tag{2.1.3}$$

Στη συνέχεια υπολογίζεται η ενέργεια εξόδου $P_m[r]$ κάθε φίλτρου $r = 1, 2 \dots R$ με συνάρτηση $V_r[k]$ και εύρος $[L_r, R_r]$ της συστοιχίας στην κλίμακα *mel* :

$$P_m[r] = \frac{\sum_{k=L_r}^{R_r} |V_r[k] X_m[k]|^2}{\sum_{k=L_r}^{R_r} |V_r[k]|^2} \tag{2.1.4}$$

Σε αυτό το σημείο το σήμα μπορεί να αναπαρασταθεί στο δισδιάστατο επίπεδο που ορίζουν οι άξονες χρό-




Σχήμα 2.1.5: Αναπαράσταση της λογαριθμικής κλίμακας *mel*.

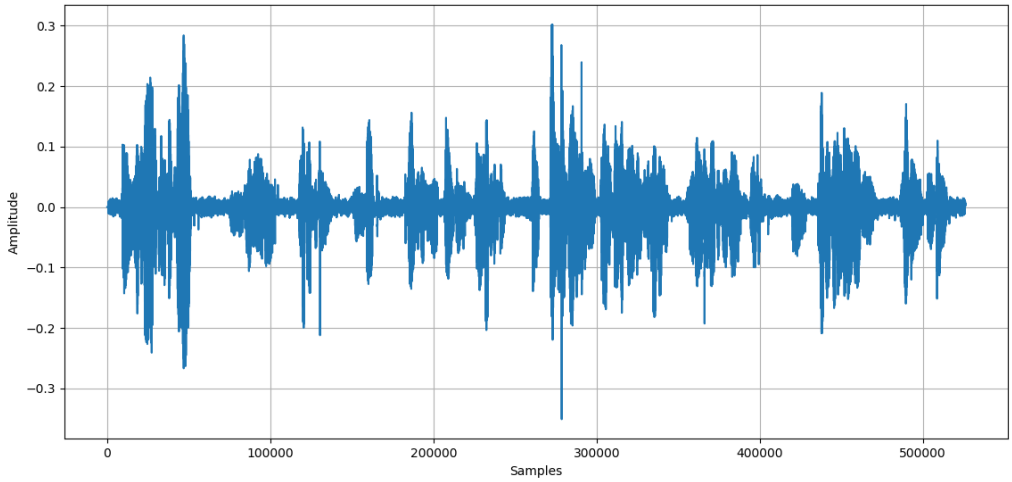
νου και συχνότητας. Ο άξονας χρόνου περιλαμβάνει όλα τα παράθυρα στα οποία εφαρμόζεται η τριγωνική συστοιχία ενώ ο άξονας συχνότητας αντιστοιχεί στις επιμέρους ενέργειες των εξόδων των φίλτρων σε κάθε εύρος συχνότητας ξεχωριστά. Η αναπαράσταση αυτή αναφέρεται ως *spectgram* ή *mel spectogram* αν ο άξονας της συχνότητας είναι στην κλίμακα *mel*, και φαίνεται στο Σχήμα 2.1.7. Στη συνέχεια για την εξαγωγή των *MFCC* συνιστωσών υπολογίζεται ο διακριτός μετασχηματισμός συνημιτόνου του λογαρίθμου της ενέργειας της εξόδου των φίλτρων :

$$MFCC_m[n] = \frac{1}{R}\sum_{r=1}^{R} log(P_m[r])cos\left(\frac{2n\pi}{R}(r+\frac{1}{2})\right) \tag{2.1.5}$$

### Θεμελιώδης Συχνότητα ($F_0$)

Η θεμελιώδης συχνότητα ενός σήματος φωνής αντιστοιχεί στην συχνότητα με την οποία ανοίγουν και κλείνουν οι φωνητικές χορδές. Συμβολίζεται με $F_0$ διότι αποτελεί την χαμηλότερη συχνότητα του σήματος, ακολουθούμενη από τις συχνότητες $F_1, F2, \ldots$ οι οποίες συμβολίζουν τα formants. Για τον υπολογισμό της έχουν προταθεί ποικίλοι αλγόριθμοι, από τους οποίους ξεχωρίζει ο αλγόριθμος με την χρήση αυτοσυσχέτισης [4].

Αρχικά το σήμα φωνής $s(n)$ φιλτράρεται μέσω ενός βαθυπερατού φίλτρου και στη συνέχεια παραθυρώνεται σε επικαλυπτόμενα πλαίσια $f_s(n,m)$, σύμφωνα με την εξίσωση 2.1.6a όπου $w(m-n)$ αντιστοιχεί στο παράθυρο μήκους $N_w$. Στα πλαίσια εφαρμόζεται μια συνθήκη κατωφλίου, 2.1.6b τιμής $C_{thr}$. Στη συνέχεια υπολογίζεται η αυτοσυσχέτιση σύμφωνα με την εξίσωση 2.1.6c όπου $\eta$ αντιστοιχεί στην καθυστέρηση κατά τον υπολογισμό της αυτοσυσχέτισης. Τέλος η $F_0$ προσεγγίζεται σύμφωνα με την εξίσωση 2.1.6d, όπου $F_s$ είναι η συχνότητα δειγματοληψίας, και $F_h, F_l$ είναι η μέγιστη και ελάχιστη αντιληπτή συχνότητα του ανθρώπου. Με τον υπολογισμό της $F_0$ έμμεσα υπολογίζεται και η Πιθανότητα Έμφωνου Ήχου (*Voice Probability*).




Σχήμα 2.1.6: Αναπαράσταση σήματος φωνής στο πεδίο του χρόνου.

$$f_s(n,m) = s(n)w(m-n) \tag{2.1.6a}$$

$$\hat{f}_s(n,m) = \left\{ \begin{array}{cc} f_s(n,m) - C_{thr} & |f_s(n,m)| \geq C_{thr}, \\ 0 & |f_s(n,m)| < C_{thr} \end{array} \right\} \tag{2.1.6b}$$

$$r_s(\eta,m) = \frac{1}{N} \sum_{n=m-N_w+1}^{m} \hat{f}_s(n,m) f_s(n-\eta,m) \tag{2.1.6c}$$

$$\hat{F}_0 = \frac{F_s}{N_w} \mathbf{argmax}_\eta \left\{ |r_s(\eta,m)| \right\} \begin{array}{c} n = N_w(F_h/F_s), \\ n = N_w(F_l/F_s) \end{array} \tag{2.1.6d}$$

### 2.1.3 Διαφορές μεταξύ ομιλίας και φωνής τραγουδιού

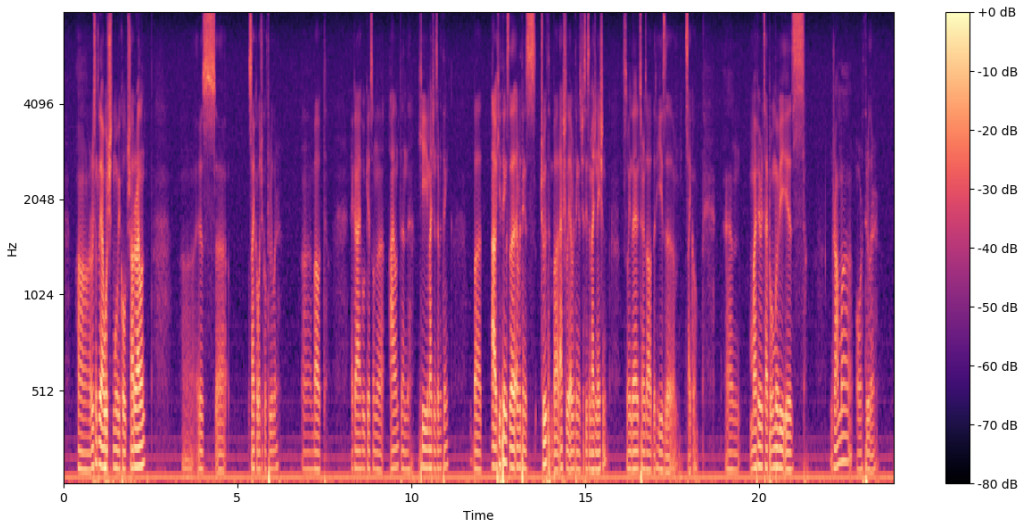
Κατά μία έννοια, η φωνή τραγουδιού μπορεί να θεωρηθεί ως μια ειδική μορφή ομιλίας, αλλά εξακολουθούν να υπάρχουν πολλές διαφορές μεταξύ τους[5]. Αυτές οι αποκλίσεις καθιστούν ακατάλληλη την άμεση μεταγραφή των στίχων από τη φωνή τραγουδιού με τη χρήση ενός μοντέλου αναγνώρισης ομιλίας που έχει εκπαιδευτεί σε δεδομένα ASR. Για να καταδείξουμε τις ασυμφωνίες, επιλέγουμε τυχαία 10K προτάσεις από το LibriSpeech [6] για το σώμα δεδομένων ομιλίας και το Dali [7] για το σύνολο δεδομένων φωνής τραγουδιού, και κάνουμε κάποια στατιστικά στοιχεία για αυτές. Η φυσική ομιλία και η φωνή τραγουδιού διαφέρουν κυρίως στις ακόλουθες πτυχές και τα αποτελέσματα της ανάλυσης παρατίθενται στον πίνακα 2.1.

#### Τόνος

Εξάγουμε τα περιγράμματα του τόνου (pitch) από τη φωνή τραγουδιού και την ομιλία και συγκρίνουμε το εύρος και την ομαλότητα του τόνου. Εδώ χρησιμοποιούμε το ημιτόνιο ως μονάδα του τόνου.

#### Εύρος τόνου

Σε γενικές γραμμές, το εύρος του τόνου (pitch range) στη φωνή τραγουδιού είναι μεγαλύτερο από αυτό της ομιλίας. Οι Loscos, Cano και Bonada [8] έχουν επισημάνει ότι το εύρος συχνοτήτων στην τραγουδιστική φωνή μπορεί να είναι πολύ μεγαλύτερο σε σύγκριση με αυτό της ομιλίας. Για κάθε πρόταση, υπολογίζουμε το εύρος του τόνου (η μέγιστη τιμή του τόνου μείον την ελάχιστη τιμή του τόνου σε αυτή την πρόταση).




Σχήμα 2.1.7: Αναπαράσταση σήματος φωνής μέσω *spectogram*.

Αφού υπολογίσαμε τον μέσο όρο του εύρους του ύψους των συνολικών 10K προτάσεων στο σώμα, οι μέσες τιμές παρατίθενται στο πεδίο Pitch Range στον πίνακα 2.1.

### Ομαλότητα του τόνου

Το ύψος κάθε πλαισίου σε μια συγκεκριμένη συλλαβή (όταν αντιστοιχεί σε νότα) στην τραγουδιστή φωνή παραμένει σχεδόν σταθερό, ενώ στην ομιλία το ύψος αλλάζει ελεύθερα μαζί με τα ηχητικά πλαίσια σε μια συλλαβή. Το χαρακτηριστικό της διατήρησης της τοπικής σταθερότητας εντός μιας συλλαβής το ονομάζουμε ομαλότητα του τόνου (Pitch Smoothness). Συγκεκριμένα, υπολογίζουμε τη διαφορά ύψους μεταξύ κάθε δύο γειτονικών πλαισίων σε μια πρόταση και τη μέση τιμή της σε ολόκληρο το σώμα των 10K προτάσεων. Όσο μικρότερη είναι η τιμή της Pitch Smoothness, τόσο πιο ομαλό είναι το περίγραμμα του τόνου.

### Διάρκεια

Αναλύουμε επίσης και συγκρίνουμε το εύρος και τη σταθερότητα της διάρκειας της συλλαβής στην τραγουδιστή φωνή και στην ομιλία. Η διάρκεια κάθε συλλαβής ποικίλλει πολύ μαζί με τη μελωδία στην τραγουδιστική φωνή. Ενώ στην ομιλία, εξαρτάται από τις συνήθειες προφοράς του συγκεκριμένου ομιλητή.

### Εύρος διάρκειας

Για κάθε πρόταση, υπολογίζουμε τη διαφορά μεταξύ της διάρκειας της μεγαλύτερης συλλαβής και της μικρότερης συλλαβής ως εύρος διάρκειας. Οι μέσες τιμές των εύρους διάρκειας σε ολόκληρο το σώμα κειμένων παρουσιάζονται ως Εύρος διάρκειας στον πίνακα 2.1.

### Διακύμανση διάρκειας

Υπολογίζουμε τη διακύμανση της διάρκειας των συλλαβών σε κάθε πρόταση και υπολογίζουμε το μέσο όρο των διαφορών όλων των προτάσεων σε ολόκληρο το σώμα κειμένων. Τα αποτελέσματα παρατίθενται ως Duration Variance στον Πίνακα 2.1 για να αντικατοπτρίζουν την ευελιξία της διάρκειας στην τραγουδιστική φωνή.

Εκτός από τις διαφορές στα χαρακτηριστικά που αναφέραμε παραπάνω, μερικές φορές οι τραγουδιστές μπορεί να προσθέσουν βιμπράτο σε ορισμένα μακρά φωνήεντα ή να κάνουν καλλιτεχνικές τροποποιήσεις στην προφορά ορισμένων λέξεων για να τις κάνουν να ακούγονται πιο μελωδικές, αν και αυτό θα έχει ως αποτέλεσμα την απώλεια της καταληπτότητας.

| Property | Speech | Singing Voice |
|---|---|---|
| Pitch Range (semitone) | 12.71 | 14.61 |
| Pitch Smoothness | 0.93 | 0.84 |
| Duration Range (s) | 0.44 | 2.40 |
| Duration Variance | 0.01 | 0.11 |

Πίνακας 2.1: Οι διαφορές των ακουστικών ιδιοτήτων μεταξύ ομιλίας και φωνής τραγουδιού.

| | speech | singing |
|---|---|---|
| men | 100-120Hz | 75-500Hz |
| women | 165-200Hz | 135-1100Hz |

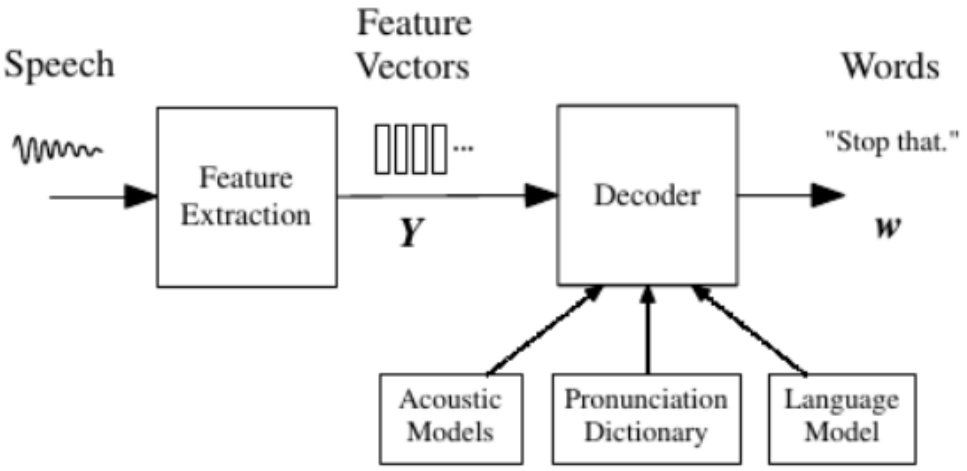
Πίνακας 2.2: Εύρος συχνοτήτων ομιλίας και τραγουδιού για άνδρες και γυναίκες.

## 2.2 Μέθοδοι Και Αλγόριθμοι Αναγνώρισης Ομιλίας

Σε αυτό το κεφάλαιο παρουσιάζουμε συνοπτικά τις κλασσικές μεθόδους και εργαλεία αναγνώρισης φωνής (ASR) παρότι πλέον δεν χρησιμοποιούνται ευρέως αλλά έχουν αντικατασταθεί από αντίστοιχα βαθιά νευρωνικά δίκτυα όπως θα δούμε στα επόμενα κεφάλαια. Αυτό γίνεται κατά κύριο λόγο να περιγράψουμε το πρόβλημα καθώς επίσης να αποκτήσουμε κάποια διαίσθηση τις επιμέρους προκλήσεις.

### 2.2.1 Διαδικασία Αυτόματης Αναγνώρισης Ομιλίας

Ο στόχος της αυτόματης αναγνώρισης ομιλίας στην απλούστερη του μορφή είναι η εξαγωγή γλωσσικών πληροφοριών από ένα ψηφιακό αρχείο ήχου. Ένα τυπικό τέτοιο αρχείο δεν περιέχει εν γένει κάποια πληροφορία που είναι άμεσα χρήσιμη για την διάκριση λέξεων ή φράσεων. Μπορεί να γίνει κάποια επιφανειακή ανάλυση της έντασης ή του συχνοτικού περιεχομένου όμως αυτά, δεν αρκούν για την αναγνώριση των ζητούμενων πληροφοριών. Για να επιτευχθεί ο στόχος, απαιτούνται πολλά ενδιάμεσα βήματα.




Σχήμα 2.2.1: Αρχιτεκτονική ενός Συστήματος ASR Βασισμένο σε HMM.

Τα βασικά συστατικά ενός συστήματος αναγνώρισης ομιλίας απεικονίζονται στο Σχήμα 2.2.1. Η κυματομορφή εισόδου από ένα μικρόφωνο μετατρέπεται σε μια σειρά από ακουστικά διανύσματα $Y = y_1, \ldots, y_T$ σε μια διαδικασία που ονομάζεται εξαγωγή χαρακτηριστικών. Ο αποκωδικοποιητής στη συνέχεια επιχειρεί να βρει την ακολουθία από λέξεις $W = w_1, \ldots, w_K$ που είναι πιθανότερο να έχει παράξει το $Y$, δηλαδή το:

$$\hat{w} = \arg max_w \{P(w|Y)\} \tag{2.2.1}$$

Ωστόσο επειδή το $P(w|Y)$είναι δύσκολο να μοντελοποιηθεί, χρησιμοποιείται ο νόμος του Bayes για να μετασχηματίσει την παραπάνω εξίσωση στο ισοδύναμο πρόβλημα:

$$\hat{w} = \arg max_w\{P(Y|w)P(w)\} \tag{2.2.2}$$

Η πιθανότητα $P(w)$ υπολογίζεται από το γλωσσικό μοντέλο, ενώ για τον υπολογισμό του $P(Y|w)$ απαιτείται ένα ακουστικό μοντέλο. Η βασική μονάδα ήχου που αναπαρίσταται από το ακουστικό μοντέλο ονομάζεται φώνημα. Η αγγλική γλώσσα απαιτεί περίπου 40 τέτοια φωνήματα, ενώ η ελληνική περίπου 30. Για οποιαδήποτε λέξη $w$, το αντίστοιχο ακουστικό μοντέλο δημιουργείται συνθέτοντας μοντέλα μεμονωμένων φωνημάτων για να σχηματίσουν λέξεις όπως αυτές ορίζονται στο λεξικό. προφορών. Οι παράμετροι των μοντέλων φωνημάτων εκτιμώνται από το σύνολο δεδομένων εκπαίδευσης, το οποίο αποτελείται από κυματομορφές ομιλίας και τις αντίστοιχες μεταγραφές τους (transcription). Το γλωσσικό μοντέλο είναι τυπικά της μορφής N-gram, στο οποίο η πιθανότητα μιας λέξης εξαρτάται από τις $N-1$ προηγούμενες.

Επομένως, συνοπτικά τα βήματα είναι:

- Χωρισμός του αρχείου ήχου σε πολλά μικρότερα κομμάτια (frames).
- Εξαγωγή χαρακτηριστικών από το αρχείο ήχου (MFCC είναι το πιο κοινό).
- Αναγνώριση φωνημάτων χρησιμοποιώντας το ακουστικό μοντέλο.
- Εύρεση των πιθανών λέξεων που σχηματίζονται από τα φωνήματα που αναγνωρίστηκαν (πχ όμως, ώμος ακούγονται το ίδιο. Και οι δυο θα ήταν υποψήφιες για την ίδια σειρά φωνημάτων).
- Επιλέγεται η πιο πιθανή λέξη με βάση τα συμφραζόμενα (context).
- Ενώνονται όλα μαζί και λαμβάνεται το τελικό αποτέλεσμα.

Στις ακόλουθες ενότητες θα παρουσιαστούν αναλυτικά όλα οι επιμέρους μαθηματικές έννοιες που απαιτούνται για την κατανόηση της διαδικασίας αυτής.

### 2.2.2 Κρυφά Μοντέλα Μαρκόβ (HMM)

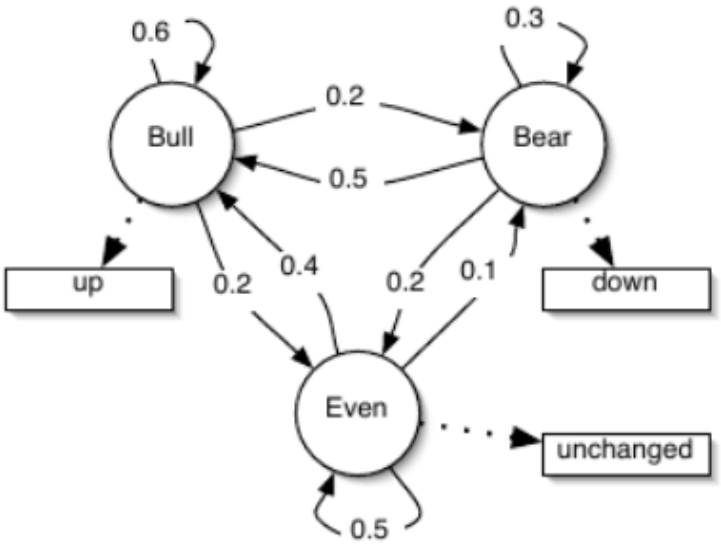
Τα Κρυφά Μοντέλα Μαρκόβ (Hidden Markov Models - HMM) [9] είναι ένα ισχυρό στατιστικό εργαλείο που χρησιμεύει στην μοντελοποίηση συστημάτων τα οποία χαρακτηρίζονται από μία διαδικασία που παράγει μία αισθητή ακολουθία. Τα Κρυφά Μοντέλα Μαρκόβ εφαρμόζονται σε πολλά πεδία ενδιαφέροντος στην επεξεργασία σήματος και ειδικά στην ανάλυση ομιλίας. Πιο συγκεκριμένα, έχουν επιτυχή εφαρμογή σε χαμηλού επιπέδου διεργασίες της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing-NLP), όπως στην επισήμανση μέρος-του-λόγου (part-of-speech tagging), στην κατάτμηση φράσης (phrase chunking) και στην εξαγωγή πληροφοριών από έγγραφα. Η Μαρκοβιανή θεωρία πήρε το όνομά της από τον Andrei Markov στις αρχές του 20ου αιώνα, αλλά η θεωρία των Κρυφών Μαρκοβιανών Μοντέλων στην πραγματικότητα αναπτύχθηκε από τον Baum και τους συνεργάτες του στην δεκαετία του 1960. Μαρκοβιανές Αλυσίδες: Στο Σχήμα 2.2.2 παρουσιάζεται μια Αλυσίδα Μαρκόβ. Το συγκεκριμένο σχήμα περιγράφει ένα απλό μοντέλο για ένα χρηματιστηριακό δείκτη. Έχει 3 καταστάσεις (Bull, Bear, Even) και 3 δείκτες παρατήρησης (πάνω, κάτω, αμετάβλητο). Το μοντέλο είναι μια μηχανή περιορισμένων καταστάσεων (finite state automaton) με πιθανοτικές μεταβάσεις ανάμεσα σε καταστάσεις (states). Δεδομένης μιας ακολουθίας παρατηρήσεων (observations), για παράδειγμα πάνω-κάτω-κάτω, μπορούμε εύκολα να επαληθεύσουμε ότι η σειρά καταστάσεων που παρήγαγε αυτές τις παρατηρήσεις ήταν η: $Bull - Bear - Bear$ και η πιθανότητα αυτής της ακολουθίας παρατηρήσεων είναι απλά το γινόμενο των βαρών των μεταβάσεων, δηλαδή $0.2 \times 0.3 \times 0.3$ .

Ο τυπικός ορισμός του HMM έχει ως εξής:

$$\lambda = (A, B, \pi) \tag{2.2.3}$$

S είναι το σύνολο των καταστάσεων, και V το σύνολο των παρατηρήσεων, δηλαδή:

$$S = (s_1, s_2, \ldots, s_N) \tag{2.2.4}$$




Σχήμα 2.2.2: Παράδειγμα Μαρκοβιανής αλυσίδας.

$$V = (v_1, v_2, \ldots, v_N) \tag{2.2.5}$$

Ορίζουμε $Q$ ως μια σταθερή ακολουθία καταστάσεων μήκους $T$ , και μια αντίστοιχη ακολουθία παρατηρήσεων $O$:

$$Q = (q_1, q_2, \ldots, q_T) \tag{2.2.6}$$

$$O = (o_1, o_2, ..., o_T) \tag{2.2.7}$$

είναι ένας πίνακας μετάβασης, περιέχει την πιθανότητα της κατάστασης $i$, που ακολουθείται από την κατάσταση $i$. Τονίζεται ότι οι πιθανότητες μετάβασης είναι ανεξάρτητες του χρόνου $t$:

$$A = [a_{ij}],\ a_{ij} = P(q_t = s_j | q_t) \tag{2.2.8}$$

$B$ είναι ο πίνακας παρατηρήσεων, περιέχει την πιθανότητα της παρατήρησης $k$ να έχει παραχθεί από την κατάσταση $j$, ανεξαρτήτως του $t$:

$$B = [b_i(K)],\ b_i(k) = P(x_t = v_k | q_t = s_i) \tag{2.2.9}$$

π είναι ο αρχικός πίνακας πιθανοτήτων:

$$\pi = [\pi_i],\ \pi_i = P(q_1 = s_i) \tag{2.2.10}$$

Το μοντέλο αυτό κάνει 2 υποθέσεις. Η πρώτη, που ονομάζεται Μαρκοβιανή υπόθεση (Markov assumption), δηλώνει ότι η παρούσα κατάσταση εξαρτάται μόνο από την αμέσως προηγούμενη. Αυτό εκπροσωπεί την μνήμη του μοντέλου:

$$P(q_1 | q_1^{t-1}) = P(q_t | q_{t-1}) \tag{2.2.11}$$

Η υπόθεση ανεξαρτησίας δηλώνει ότι η παρατήρηση την στιγμή t εξαρτάται μόνο από την παρούσα κατάσταση, είναι ανεξάρτητη από όλες τις προηγούμενες καταστάσεις και παρατηρήσεις:

$$P(o_t | o_1^{t-1}, q_1^{t_1}) = P(o_t | q_t) \tag{2.2.12}$$

Δεδομένου ενός HMM και μιας ακολουθίας παρατηρήσεων, επιθυμούμε τον υπολογισμό του P (O|λ), την πιθανότητα της ακολουθίας σύμφωνα με αυτό το δοσμένο μοντέλο. Το πρόβλημα μπορεί επίσης να χαρακτηριστεί και ως αξιολόγηση της ικανότητας ενός μοντέλου να προβλέψει μια ακολουθία παρατηρήσεων. Η πιθανότητα των παρατηρήσεων O για μια συγκεκριμένη ακολουθία Q είναι:

$$P(O|Q,\lambda) = \prod_{t=1}^{T} P(o_t|q_t,\lambda) = b_{q_1}(o_1) \times b_{q_2}(o_2) \dots b_{q_T}(o_2) \tag{2.2.13}$$

και η πιθανότητα της ακολουθίας καταστάσεων είναι:

$$P(Q|\lambda) = \pi_{q1} a_{q1q2} a_{q2q3} ... a_{qT-1} a_{qt} \tag{2.2.14}$$

ώστε να μπορούμε να υπολογίσουμε την πιθανότητα των παρατηρήσεων ως εξής:

$$P(O|\lambda) = \sum_{Q} P(O|Q,\lambda)P(Q|\lambda) = \sum_{q_1 \dots q_T} \pi_{q1} b_{q1}(o_1) a_{q_1 q_2} b_{q2}(o_2) ... a_{q_{T-1} q_T} b_{qT}(o_T) \tag{2.2.15}$$

Αυτό το αποτέλεσμα επιτρέπει την αξιολόγηση της πιθανότητας του O, όμως το να γίνει αυτό άμεσα θα ήταν δύσκολο, καθώς η πολυπλοκότητα αυξάνεται εκθετικά με το T. Μια καλύτερη προσέγγιση είναι να αναγνωριστεί ότι πολλοί περιττοί (επαναλαμβανόμενοι) υπολογισμοί θα πραγματοποιούνται με την παραπάνω εξίσωση και επομένως αποθηκεύοντας τα ενδιάμεσα αποτελέσματα σε cache οδηγούν σε μείωση πολυπλοκότητας.

Στο στάδιο της αναγνώρισης, υποθέτουμε ότι έχουμε στην διάθεση μας περισσότερα του ενός HMM, κάθε ένα εκ των οποίων περιγράφεται από ένα διαφορετικό σύνολο παραμέτρων. Προφανώς, κάθε HMM μοντελοποιεί μία διαφορετική στάσιμη κατά τμήματα διεργασία. Για παράδειγμα, ένα HMM μπορεί να μοντελοποιεί μία διεργασία δύο πηγών εκπομπής παρατηρήσεων, μία Gaussian και μία εκθετική $x^2$ (δύο καταστάσεις), ενώ ένα άλλο HMM αντιστοιχεί σε μία διεργασία τριών πηγών, πχ. Τριών Gaussians πηγών με διαφορετικές μέσες τιμές και μητρώα συνδιασποράς.

Κατά την φάση της αναγνώρισης, ο στόχος είναι ο ακόλουθος: με δεδομένη μία ακολουθία παρατηρήσεων και ένα πλήθος M από HMM (κάθε ένα εκ των οποίων μοντελοποιεί διαφορετική διεργασία), να αποφασιστεί ποιο από τα HMM είναι περισσότερο πιθανό να εκπέμψει την συγκεκριμένη ακολουθία παρατηρήσεων. Δύο μέθοδοι που δίνουν λύση σε αυτό το πρόβλημα είναι η μέθοδος Baum-Welch (γνωστή και ως μέθοδος οποιασδήποτε διαδρομής – any path method) και η μέθοδος Viterbi (ή μέθοδος καλύτερης διαδρομής – best path method). Οι δύο αυτές μέθοδοι υπολογίζουν, για κάθε HMM, ένα αποτέλεσμα (score) που βασίζεται σε πιθανότητες. Το HMM που γεννά το μέγιστο σκορ θεωρείται ως το πιο πιθανό να έχει εκπέμψει τη συγκεκριμένη ακολουθία παρατηρήσεων. Η φάση αναγνώρισης προϋποθέτει ότι όλες οι παράμετροι που προσδιορίζουν τα HMM έχουν προηγουμένως εκτιμηθεί και είναι επομένως γνωστές. Στην φάση της εκπαίδευσης γίνετε εκτίμηση των παραμέτρων του εκάστοτε HMM. Προς την κατεύθυνση αυτή, χρησιμοποιείται μία ακολουθία παρατηρήσεων ικανού μήκους (ή και περισσότερες), που έχει γεννηθεί από την αντίστοιχη στοχαστική διεργασία, προκειμένου να γίνει εκτίμηση των άγνωστων παραμέτρων (πχ. Χρησιμοποιώντας τεχνικές που βασίζονται στην λογική της μέγιστης πιθανοφάνειας για την εκτίμηση των παραμέτρων).

### 2.2.3 Μοντέλα Μείξης Γκαουσιανών (GMM)

Η βάση των Μοντέλων Μείξης Γκαουσιανών (Gaussian Mixture Models) [10] είναι η διαδεδομένη γκαουσιανή ή κανονική κατανομή, που δίνεται από τον τύπο:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{2.2.16}$$

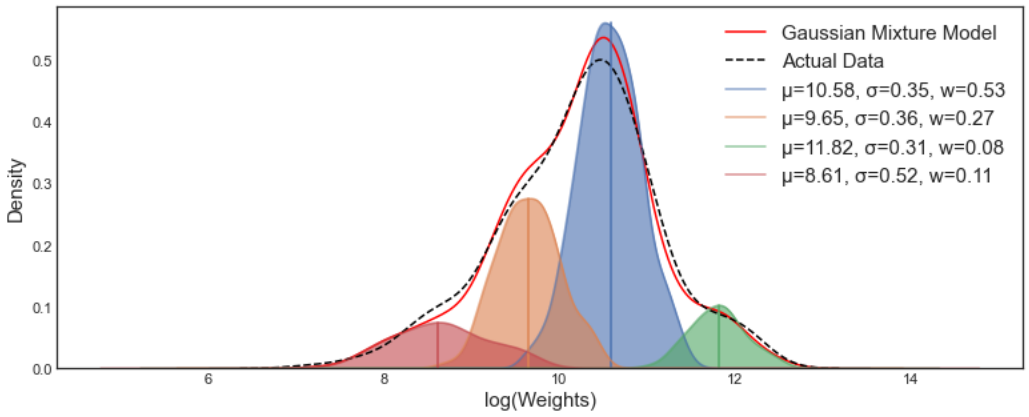
όπου $\mu$ η μέση τιμή και $\sigma$ η διασπορά.

Η κατανομή αυτή μπορεί να επεκταθεί και για πολλές μεταβλητές. Έστω μια τυχαία μεταβλητή n-διαστάσεων $X = [X_1, X_2, ..., X_n]$ με γκαουσιανή κατανομή $X \sim N(\mu_m, C_m)$. Η συνάρτηση πυκνότητας πιθανότητας για

αυτή τη τυχαία μεταβλητή είναι:

$$p(x) = \frac{1}{\sqrt{2\pi}^n \sqrt{\det C}} exp(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)) \tag{2.2.17}$$

όπου $\mu = [E[X_1], E[X_2], ..., E[X_n]]$ είναι το διάνυσμα μέσων τιμών και $C = [Cov[X_i, X_j]]; i, j = 1, 2, ..., n$ ο πίνακας συνδιασποράς.




Σχήμα 2.2.3: Παράδειγμα μοντέλου μίξης γκαουσιανών (GMM).

Παρ' όλα αυτά, η κατανομή γκαουσιανών πολλών μεταβλητών εξακολουθεί να έχει ξεκάθαρους περιορισμούς όσο αφορά την μοντελοποίηση πραγματικών δεδομένων, τα οποία δεν ακολουθούν κατ' ανάγκη την καμπύλη Gauss. Μια πολύ πιο ευέλικτη προσέγγιση είναι η πρόσθεση πολλών Γκαουσιανών κατανομών, η κάθε μια σταθμισμένη με ένα συντελεστή $c_m$. Αυτό ονομάζεται ένα μοντέλο μείξης γκαουσιανών και οι επιμέρους γκαουσιανές κατανομές $(\mu_m, C_m)$ ονομάζονται στοιχεία (components). Κάθε στοιχείο έχει δική του μέση τιμή $\mu_m$ και συνδιασπορά $C_m$. Η συνάρτηση πυκνότητας πιθανότητας για M στοιχεία δίνεται από τον τύπο:

$$p(x) = \sum_{m=1}^{M} c_m(\mu_m, C_m) = \sum_{m=1}^{M} \frac{c_m}{\sqrt{2\pi}^n \sqrt{\det C_m}} exp(-\frac{1}{2}(x-\mu_m)^T C_m^{-1}(x-\mu_m)) \tag{2.2.18}$$

Τονίζεται επίσης ότι το άθροισμα των συντελεστών ισούται με ένα.

$$\sum_{m=1}^{M} c_m = 1 \tag{2.2.19}$$

Αυξάνοντας τον αριθμό των στοιχείων και τροποποιώντας τις παραμέτρους του κάθε στοιχείου, το γκαουσιανό μίγμα μπορεί να αναπαραστήσει σχεδόν οποιοδήποτε σύνολο δεδομένων με υψηλή ακρίβεια.

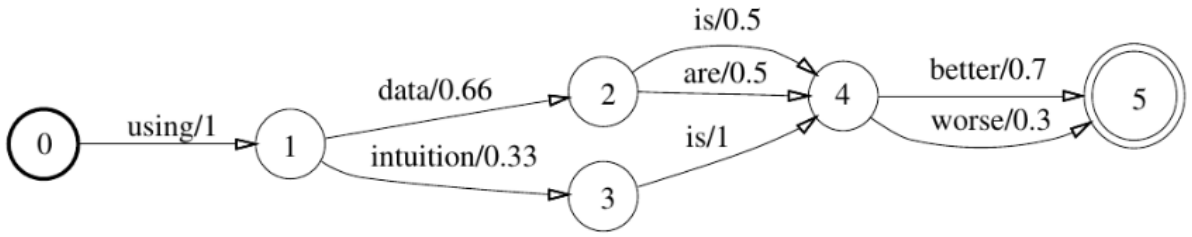
### 2.2.4 Σταθμισμένοι μετατροπείς πεπερασμένων καταστάσεων (WFST)

Οι σταθμισμένοι μετατροπείς πεπερασμένων καταστάσεων (weighted finite state transducers - WFST) [11] είναι μαθηματικές κατασκευές που ανήκουν στην ευρύτερη θεωρία αυτομάτων. Στο κεφάλαιο αυτό θα δοθεί μια γενική περιγραφή αυτών των μοντέλων και πως αυτά χρησιμοποιούνται με μεγάλη επιτυχία στην κατασκευή ολοκληρωμένων συστημάτων αυτόματης αναγνώρισης ομιλίας. Οι WFST προσφέρουν μια κοινή και φυσική αναπαράσταση πολλών δομών κρίσιμης σημασίας για τα συστήματα αυτόματης αναγνώρισης ομιλίας, όπως Κρυφά Μοντέλα Μαρκόβ (HMM), μοντέλα αναπαράστασης εξάρτησης από συμφραζόμενα, λεξικά προφορών, στατιστικών γραμματικών, και πλέγματα λέξεων ή φωνημάτων. Ελεύθερες βιβλιοθήκες γενικής χρήσης, όπως η OpenFST [12], που προσφέρουν αναλυτική υλοποίηση των WFST σε κάποια γνωστή γλώσσα προγραμματισμού (C++ στην προκειμένη περίπτωση), τους κάνουν ένα πολύ ισχυρό εργαλείο στα χέρια έμπειρου

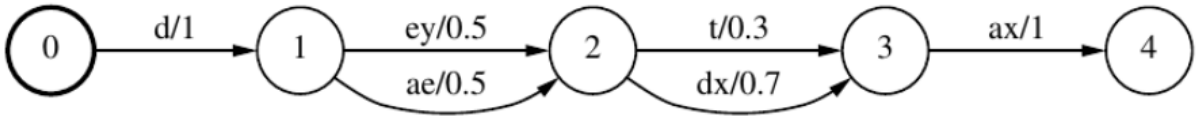
προγραμματιστή. Μεγάλο μέρος της αναγνώρισης ομιλίας σήμερα είναι βασισμένη σε μοντέλα όπως τα Κρυφά Μοντέλα Μαρκόβ, λεξικά ή n-gram γλωσσικά μοντέλα τα οποία μπορούν να αναπαρασταθούν με WFST.

Ένας σταθμισμένος μετατροπέας πεπερασμένων καταστάσεων είναι αυτόματο του οποίου οι μεταβάσεις καταστάσεων έχουν σύμβολα εισόδου και εξόδου. Κατά συνέπεια, ένα μονοπάτι μέσα από τον μετατροπέα κωδικοποιεί μια συμβολοσειρά στην είσοδο σε μια καινούρια στην έξοδο. Ένας σταθμισμένος μετατροπέας αναθέτει και βάρη στις μεταβάσεις, πέρα από την κωδικοποίηση συμβολοσειρών (strings) εισόδου εξόδου. Τα βάρη αυτά μπορεί να αναπαριστούν πιθανότητες, χρονικές διάρκειες, ποινές ή οποιαδήποτε άλλη ποσότητα που συσσωρεύεται κατά μήκος ενός μονοπατιού ώστε να υπολογιστεί το συνολικό "κόστος" μετατροπής μιας συμβολοσειράς εισόδου σε μια διαφορετική στην έξοδο. Επομένως, οι σταθμισμένοι μετατροπείς είναι μια φυσική επιλογή για την αναπαράσταση των πιθανοτικών μοντέλων πεπερασμένων καταστάσεων που συναντώνται στην επεξεργασία ομιλίας.

Σταθμισμένα αυτόματα πεπερασμένων καταστάσεων, ή σταθμισμένοι δέκτες (weighted acceptors), χρησιμοποιούνται ευρέως στην αυτόματη αναγνώριση ομιλίας. Τα παρακάτω σχήματα περιέχουν μερικά παραδείγματα. Το σχήμα 2.2.4a είναι ένα πολύ απλό γλωσσικό μοντέλο. Οι έγκυρες σειρές λέξεων προσδιορίζονται από τις λέξεις σε κάθε ολοκληρωμένη διαδρομή και οι πιθανότητα τους είναι το γινόμενο όλων των πιθανοτήτων μετάβασης, των βαρών. Το αυτόματο στο σχήμα 2.2.4b δίνει όλες τις πιθανές προφορές της λέξης *data*, που χρησιμοποιείται στο γλωσσικό μοντέλο. Κάθε έγκυρη προφορά είναι η σειρά φωνημάτων σε κάθε ολοκληρωμένη διαδρομή και η πιθανότητα του υπολογίζεται ακριβώς όπως και στην προηγούμενη περίπτωση. Συμβατικά, οι καταστάσεις αναπαρίστανται με κύκλους και σημειώνονται με έναν μοναδικό αριθμό που αντιστοιχεί σε αυτές. Η αρχική κατάσταση αναπαρίσταται από κύκλο με έντονο περίγραμμα, οι τελικές καταστάσεις με διπλό κύκλο. Η ετικέτα $l$ και το βάρος $w$ μιας μετάβασης σημειώνονται πάνω από τα βέλη που τους αντιστοιχούν με την μορφή $l/w$.




(a) Αυτόματο με καταστάσεις σε επίπεδο λέξης.




(b) Αυτόματο με καταστάσεις σε επίπεδο φωνήματος.

Σχήμα 2.2.4: Παραδείγματα σταθμισμένων δεκτών πεπερασμένων καταστάσεων.

Τα αυτόματα αυτά αποτελούνται από ένα σύνολο καταστάσεων: μια αρχική και μια ομάδα από τελικές καταστάσεις (με αντίστοιχα βάρη) και ένα σύνολο από μεταβάσεις ανάμεσα σε καταστάσεις. Κάθε μετάβαση έχει μια κατάσταση εκκίνησης και κατάσταση προορισμού, μια ετικέτα και ένα βάρος. Τέτοιες δομές ονομάζονται, σταθμισμένοι δέκτες πεπερασμένων καταστάσεων (weighted finite state acceptors - WFSA), καθώς δέχονται ή αναγνωρίζουν κάθε συμβολοσειρά που μπορεί να διαβαστεί πάνω σε ένα μονοπάτι από την αρχική προς κάποια τελική κατάσταση. Σε κάθε έγκυρη συμβολοσειρά ανατίθεται ένα βάρος, συγκεκριμένα το άθροισμα των βαρών των μεταβάσεων από τις οποίες πέρασε, μαζί με το βάρος της τελικής κατάστασης. Ουσιαστικά, ένας τέτοιος δέκτης αναπαριστά ένα σύνολο από συμβολοσειρές, αυτές που γίνονται δεκτές. Εάν είναι σταθμισμένος, τότε αναθέτει σε κάθε συμβολοσειρά και ένα βάρος.

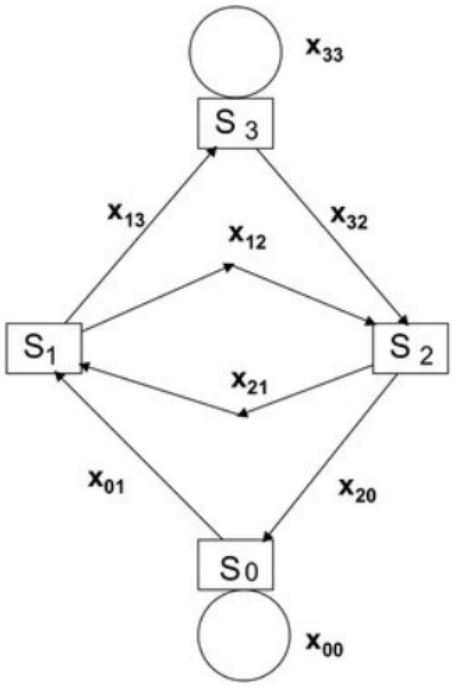
Τυπικά, συστήματα αναγνώρισης ομιλίας επιδιώκουν, με τον αποκωδικοποιητή τους, να συνδυάσουν και να

βελτιστοποιήσουν αυτόματα όπως αυτά στο σχήμα 2.2.4. Βρίσκει προφορές λέξεων από το λεξικό και τις αντικαθιστά στην γραμματική του. Αναπαραστάσεις φωνητικών δέντρων ενδεχομένως να χρησιμοποιούνται για να εντοπίσουν και να αφαιρέσουν πλεονασμούς στις διαδρομές, και κατά συνέπεια να βελτιώσουν την αποδοτικότητα των αναζητήσεων.

### 2.2.5 Αλγόριθμος Viterbi

Ο αλγόριθμος Viterbi [13] παράγει τις εκτιμήσεις μέγιστης πιθανότητας των διαδοχικών καταστάσεων μιας μηχανής πεπερασμένων καταστάσεων (finite-state machine) από την ακολουθία των αποτελεσμάτων που έχουν φθαρεί από διαδοχικούς ανεξάρτητους όρους παρεμβολών.

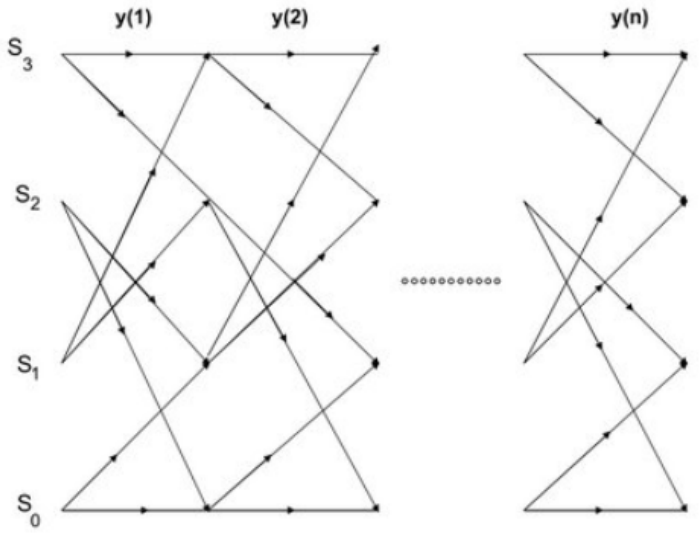
Αντίστοιχα, σε κάθε κατάσταση η γεννήτρια σήματος παράγει μια έξοδο x, η οποία είναι γενικά ένα πραγματικό διάνυσμα. Το φαινόμενο φθοράς, που ονομάζεται κανάλι στις εφαρμογές επικοινωνίας, μετατρέπει το x σε y, τα παρατηρήσιμα, τα οποία από τη φύση του σχηματισμού τους αποτελούν μια τυχαία Markov ακολουθία. Σημειώνεται ότι οι όροι του x επίσης συνιστούν μια Markov ακολουθία όποτε η ακολουθία εισόδου u στη FSM είναι τυχαία, δημιουργώντας έτσι μια κατανομή πιθανότητας στις μεταβάσεις της κατάστασης. Το σχήμα 2.2.5 είναι το διάγραμμα καταστάσεων για μια FSM με παραμέτρους Q=2 και L=2. Οι καταστάσεις τότε αντιστοιχούν στα περιεχόμενα των καταχωρητών: S0 = 00, S1 = 01, S2 = 10, S3 = 11




Σχήμα 2.2.5: Μαρκοβιανός γράφος τεσσάρων καταστάσεων.

Προφανώς, μόνο ένα υποσύνολο των μεταβάσεων είναι εφικτό. Από την ακολουθία των παρατηρήσεων y, αναζητούμε την πιο πιθανή πορεία μεταβάσεων μέσω των καταστάσεων του διαγράμματος. Στο Σχήμα 2.2.5, ορίζουμε επίσης τις εξόδους FSM σε κάθε κλάδο. Έτσι, ο κλάδος S0 ως το S1 φέρει την ένδειξη x01. Είναι επίσης βολικό για την περιγραφή του αλγορίθμου μια διαδρομή πολλών βημάτων κατά μήκος του γράφου να παρομοιαστεί ως μια πολλαπλών επιπέδων αναπαραγωγή του διαγράμματος καταστάσεων, γνωστό και ως trellis diagram. Το σχήμα 2.2.6 είναι το trellis diagram που αντιστοιχεί στο παράδειγμα του σχήματος 2.2.5.

Στην κορυφή του Σχήματος 2.2.6 φαίνονται οι διαδοχικοί παρατηρήσιμοι κλάδοι y(1), y(2), ...y(k), ... . Από εδώ και πέρα, χρησιμοποιούμε τον συμβολισμό y(k) για να δηλώσουμε το παρατηρήσιμο(α) του k-οστού συνεχόμενου κλάδου. Παρομοίως, με S(k) δηλώνουμε οποιαδήποτε κατάσταση στο k-οστό επίπεδο κόμβου.




Σχήμα 2.2.6: Το διάγραμμα trellis για τον Μαρκοβιανό Γράφο του Σχήματος 2.2.5.

Ο στόχος είναι να βρεθεί η πιο πιθανή διαδρομή εντός του διαγράμματος trellis. Δεδομένου ότι οι διαδοχικοί όροι u της ακολουθίας εισόδου είναι ανεξάρτητοι μεταξύ τους, οι πιθανότητες μετάβασης καταστάσεων $Pr[S(k-1) \rightarrow S(k)]$ είναι ανεξάρτητες μεταξύ τους για κάθε k όπως επίσης είναι και οι υπό συνθήκη πιθανότητες $p[y(k)|S(k-1) \rightarrow S(k)]$. Για οποιοδήποτε μονοπάτι από την αρχή (k=0) σε έναν τυχαίο κόμβο n, S(0), S(1), ..., S(n), η σχετική πιθανότητα μονοπατιού (likelihood function) δίνεται από τον τύπο:

$$L = \prod_{k=1}^{n} Pr[S(k-1) \rightarrow S(k)] p[y(k)|S(k-1) \rightarrow S(k)] \tag{2.2.20}$$

Για υπολογιστικούς σκοπούς είναι πιο βολικό να χρησιμοποιηθεί ο λογάριθμος της παραπάνω ποσότητας, το οποίο δίνεται από τη σχέση:

$$ln(L) = \sum_{k=1}^{n} m[y(k); S(k-1), S(k)] \tag{2.2.21}$$

όπου

$$m[y(k); S(k-1), S(k)] = lnPr[S(k-1) \rightarrow S(k)] + lnp[y(k)]|S(k-1) \rightarrow S(k) \tag{2.2.22}$$

το οποίο ονομάζεται Branch Metric ανάμεσα σε οποιεσδήποτε δυο καταστάσεις στα (k-1)- οστά και k-οστά επίπεδα κόμβων (σημειώνεται ότι οι μη-επιτρεπτές μεταβάσεις έχουν μηδενική πιθανότητα και κατά συνέπεια οι λογάριθμοι θα είναι -άπειρο, άρα εκτός συναγωνισμού). Στη συνέχεια ορίζουμε το State Metric, MK $(S_i)$ της κατάστασης $S_i(K)$ να είναι το μέγιστο από όλα τα μονοπάτια από την αρχή έως την i-οστή κατάσταση στον K-οστό κόμβο.

$$M_K(S_i) = \max\{\sum_{k=0}^{K-1} m[y(k); S(k-1), S(k)] + m[y(K); S(K-1), S_i(K)]\} \tag{2.2.23}$$

Στη συνέχεια, για να μεγιστοποιηθεί το παραπάνω άθροισμα για K όρους, αρκεί να μεγιστοποιηθεί το άθροισμα για τους πρώτους K-1 όρους για κάθε κατάσταση $S_j(K-1)$ στον (K-1)-οστό κόμβο και στη συνέχεια να μεγιστοποιηθεί το άθροισμα αυτού με τον K-οστό όρο επάνω σε όλες τις καταστάσεις $S(K-1)$. Επομένως:

$$M_K(S_i) = \max M_{K-1}(S_j) + m[y(K); S(K-1), S_i(K)] \tag{2.2.24}$$

Η αναδρομή αυτή είναι γνωστή και ως αλγόριθμος Viterbi. Η ευκολότερη περιγραφή του γίνεται σε συνδυασμό με το trellis diagram. Εάν επισημάνουμε κάθε κλάδο (επιτρεπτή μετάβαση μεταξύ καταστάσεων) με το Branch Metric του m[] και κάθε κατάσταση σε κάθε επίπεδο κόμβου με το State Metric του M[], τότε τα State Metrics στο κομβικό επίπεδο K λαμβάνονται από τα State Metrics στο κομβικό επίπεδο K-1 προσθέτοντας στο κάθε ένα τα Branch Metrics που το συνδέουν με τις καταστάσεις στο K-οστό επίπεδο και για κάθε κατάσταση στο επίπεδο K κρατάμε μόνο το μεγαλύτερο άθροισμα που φτάνει σε αυτό. Εάν επιπλέον σε κάθε επίπεδο διαγράψουμε όλους τους κλάδους εκτός από αυτόν που παράγει το μέγιστο, τότε θα παραμείνει μόνο ένα μονοπάτι που διασχίζει το trellis ξεκινώντας από την πηγή σε κάθε κατάσταση στο K-οστό επίπεδο, το οποίο θα είναι το πιο πιθανό μονοπάτι από την αρχή. Σε τυπικές (όχι σε όλες) εφαρμογές, τόσο η αρχική κατάσταση (origin) και η τελική κατάσταση ορίζονται να είναι $S_0$ και κατά συνέπεια ο αλγόριθμος παράγει το πιθανότερο μονοπάτι που διασχίζει το trellis με προέλευση και τέλος το $S_0$.

### 2.2.6 Γλωσσική Μοντελοποίηση

Ένα γλωσσικό μοντέλο (language model), είναι ένα μαθηματικό μοντέλο που αναθέτει πιθανότητες σε ακολουθίες από λέξεις [14]. Το πιο απλό τέτοιο μοντέλο ονομάζεται N-gram language model και είναι από τα σημαντικότερα εργαλεία σε οποιαδήποτε δραστηριότητα που σχετίζεται με επεξεργασία γλώσσας. N-gram ονομάζεται μια ακολουθία από N λέξεις: ένα 2-gram (bi-gram) είναι μια ακολουθία 2 λέξεων όπως "καλός καιρός", ένα 3-gram (tri-gram) είναι μια ακολουθία 3 λέξεων όπως "κυρίες και κύριοι". Στη συνέχεια, θα δούμε πως μπορούμε να αξιοποιήσουμε τα N-gram μοντέλα για να εκτιμήσουμε την πιθανότητα της τελευταίας λέξης μιας ακολουθίας, αλλά και για να αναθέσουμε πιθανότητες σε ολόκληρες προτάσεις.

#### N-grams

Επιθυμούμε να υπολογίσουμε το P (w|h), την πιθανότητα της λέξης w, δεδομένης κάποιας ιστορίας h. Έστω η ιστορία h είναι "its water is so transparent that" και ζητάμε την πιθανότητα $P$ η επόμενη λέξη να είναι "the".

Ένας τρόπος να εκτιμηθεί η πιθανότητα αυτή είναι από τις σχετικές συχνότητες εμφάνισης: παίρνοντας ένα μεγάλο corpus (συλλογή από προτάσεις), μετράμε πόσες φορές εμφανίζεται η πρόταση its water is so transparent that, και μετράμε τις φορές που αυτή ακολουθείται από the. Με αυτό το τρόπο, λαμβάνουμε απάντηση στην ερώτηση "Από όλες τις φορές που είδαμε την ιστορία h, πόσες φορές εμφανίστηκε η λέξη w αμέσως μετά;" ως εξής:

$$P(the|its\,water\,is\,so\,transparent\,that) = \frac{C(its\,water\,is\,so\,transparent\,that\,the)}{C(its\,water\,is\,so\,transparent\,that)} \tag{2.2.25}$$

Με ένα ικανοποιητικά μεγάλο corpus, όπως το διαδίκτυο, μπορούμε να υπολογίσουμε την ζητούμενη πιθανότητα χρησιμοποιώντας την εξίσωση 2.2.25. Ο τρόπος αυτός δουλεύει ικανοποιητικά καλά σε μερικές περιπτώσεις, και έχουμε δει ότι και το διαδίκτυο είναι αρκετά μεγάλο για να μας δώσει εκτιμήσεις με καλή ακρίβεια στις περισσότερες περιπτώσεις. Ωστόσο αυτό μπορεί και να μην συμβαίνει επειδή η γλώσσα είναι κάτι δημιουργικό: νέες προτάσεις δημιουργούνται διαρκώς, και δεν γίνεται πάντα να μετρήσουμε ολόκληρες προτάσεις. Ακόμα και απλές επεκτάσεις της παραπάνω πρότασης μπορεί να μην εμφανίζονται πουθενά στο internet. Παρομοίως, αν ζητάμε την πιθανότητα ολόκληρης ακολουθίας λέξεως όπως "its water is so transparent", θα μπορούσαμε να το κάνουμε απαντώντας την ερώτηση: από όλους τους συνδυασμούς 5 λέξεων, πόσοι από αυτούς είναι "its water is so transparent" ; Απαιτείται ο εντοπισμός όλων των προτάσεων με 5 λέξεις, πρακτικά ανέφικτο! Χρειαζόμαστε εξυπνότερο τρόπο υπολογισμού των πιθανοτήτων αυτών.

Για την αναπαράσταση της πιθανότητας μιας τυχαίας μεταβλητής $X_i$ να λαμβάνει την τιμή "the", δηλαδή $P(X_i = \text{"the"})$, θα χρησιμοποιείται $P(the)$. Μια ακολουθία N λέξεων θα συμβολίζεται είτε με$w_1...w_n$ είτε $w_1^N$ . Για την πιθανότητα κάθε λέξης μιας ακολουθίας να έχει συγκεκριμένη τιμή, δηλαδή $P(X = w_1, Y = w_2, Z = w_3, ..., W = w_n)$ θα χρησιμοποιείται $P(w_1, w_2, ..., w_n)$. Πώς υπολογίζονται πιθανότητες ολόκληρων προτάσεων όπως $P(w_1, w_2, ..., w_n)$; Σε πρώτη φάση, σπάμε τους υπολογισμούς με βάση τον κανόνα της αλυσίδας:

$$P(X_1...X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1^2)...P(X_n|X_1^n - 1) = \prod_{k=1}^{N} P(X_k|X_1^{k-1}) \tag{2.2.26}$$

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)...P(w_n|w_1^P n - 1) = \prod_{k=1}^{N} P(w_k|w_1^P k - 1) \tag{2.2.27}$$

Ο κανόνας της αλυσίδας δείχνει την σχέση ανάμεσα στον υπολογισμό πιθανότητας ολόκληρης της ακολουθίας και υπολογίζοντας την υπό συνθήκη πιθανότητα μιας λέξης δεδομένων των προηγούμενων. Η τελευταία εξίσωση υποδεικνύει ότι μπορούμε να εκτιμήσουμε την πιθανότητα ολόκληρης πρότασης πολλαπλασιάζοντας πολλές επιμέρους υπό συνθήκη πιθανότητες. Αλλά αυτό δεν βοηθάει ιδιαίτερα, είναι πρακτικά αδύνατο να υπολογιστεί με ακρίβεια η πιθανότητα μιας λέξης δεδομένης μεγάλης ακολουθίας προηγούμενων λέξεων, το $P(w_n|w_1^{n-1})$ . Είναι και υπολογιστικά δύσκολο, όμως και λογικά, καθώς όπως αναφέρθηκε προηγουμένως, η γλώσσα είναι μια δημιουργική διαδικασία και οποιαδήποτε "σωστή" σειρά λέξεων μπορεί να μην έχει ξαναεμφανιστεί ποτέ.

Η διαίσθηση των μοντέλων N-gram είναι ότι αντί να χρησιμοποιούμε ολόκληρη την ιστορία για να υπολογίσουμε με ακρίβεια την πιθανότητα μιας λέξης, προσεγγίζουμε την επίδραση της ιστορίας χρησιμοποιώντας μόνο μερικές από τις πιο πρόσφατες λέξεις. Σε ένα bigram υπολογίζει την πιθανότητα μιας λέξης δεδομένων όλων των προηγούμενων, χρησιμοποιώντας μόνο την υπό συνθήκη πιθανότητα της αμέσως προηγούμενης λέξης.

Η υπόθεση ότι η πιθανότητα μιας λέξης εξαρτάται μόνο από την αμέσως προηγούμενη ονομάζεται υπόθεση Markov [15]. Μαρκοβιανά μοντέλα είναι η κατηγορία πιθανοτικών μοντέλων με τα οποία μπορούμε να προβλέψουμε την πιθανότητα κάποιου μελλοντικού στοιχείου χωρίς να χρειάζεται να κοιτάξουμε πολύ μακρυά στο παρελθόν. Το bigram (που κοιτάει 1 λέξη προς τα πίσω) γενικεύεται στο trigram (που κοιτάει 2 λέξεις προς τα πίσω) και ακόμα πιο γενικά στο N-gram (που κοιτάει N-1 λέξεις στο παρελθόν).

### 2.2.7 Ποσοστό Σφαλμάτων Λέξης (WER)

Το ποσοστό σφαλμάτων λέξης (word error rate - WER) [16] είναι μια κοινή μετρική για την αξιολόγηση της απόδοσης ενός συστήματος αναγνώρισης ομιλίας ή μηχανικής μετάφρασης.

Η γενική δυσκολία μέτρησης της απόδοσης έγκειται στο γεγονός ότι η αναγνωρισμένη ακολουθία λέξεων μπορεί να έχει διαφορετικό μήκος από την ακολουθία λέξεων αναφοράς (υποτίθεται ότι είναι η σωστή). Το WER προέρχεται από την απόσταση Levenshtein, λειτουργώντας σε επίπεδο λέξης αντί σε επίπεδο φωνήματος. Το WER είναι ένα πολύτιμο εργαλείο για τη σύγκριση διαφορετικών συστημάτων καθώς και για την αξιολόγηση βελτιώσεων εντός ενός συστήματος. Αυτό το είδος μέτρησης, ωστόσο, δεν παρέχει λεπτομέρειες σχετικά με τη φύση των μεταφραστικών σφαλμάτων και, επομένως, απαιτείται περαιτέρω εργασία για τον εντοπισμό της κύριας πηγής (ή των πηγών) σφαλμάτων και για την εστίαση κάθε ερευνητικής προσπάθειας.

Το πρόβλημα αυτό επιλύεται με την πρώτη ευθυγράμμιση της αναγνωρισμένης ακολουθίας λέξεων με την ακολουθία λέξεων αναφοράς (προφορικού λόγου) με τη χρήση δυναμικής ευθυγράμμισης συμβολοσειρών. Η εξέταση αυτού του ζητήματος γίνεται αντιληπτή μέσω μιας θεωρίας που ονομάζεται νόμος της δύναμης και δηλώνει τη συσχέτιση μεταξύ της περιπλοκότητας και του ποσοστού σφάλματος λέξης[1].

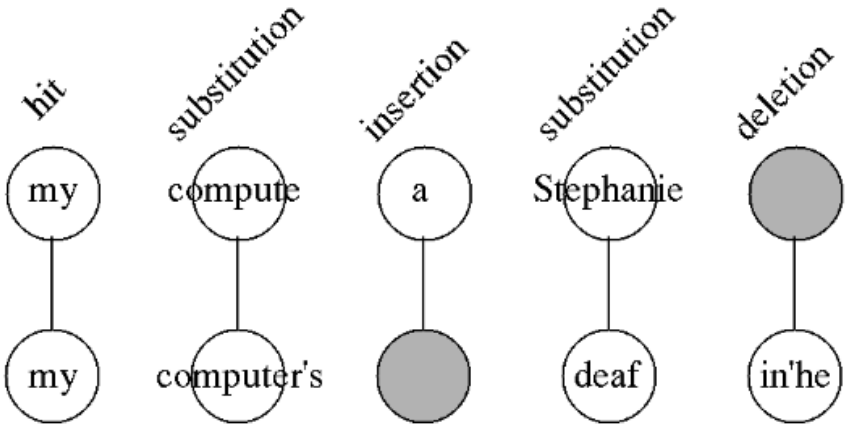
Το ποσοστό σφάλματος λέξης μπορεί στη συνέχεια να υπολογιστεί ως εξής:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{2.2.28}$$

όπου:
S είναι ο αριθμός των αντικαταστάσεων,
D είναι ο αριθμός των διαγραφών,
I είναι ο αριθμός των εισαγωγών,

C είναι ο αριθμός των σωστών λέξεων,
N είναι ο αριθμός των λέξεων της αναφοράς (N=S+D+C)

Στο Σχήμα 2.2.7 φαίνονται τα διάφορα σφάλματα αναγνώρισης όπου στην πρώτη γραμμή είναι η έξοδος του συστήματος αναγνώρισης, δηλαδή η μεταγραφή του ηχητικού, ενώ στην δεύτερη γραμμή είναι το κείμενο που εκφωνείται πραγματικά από τον ομιλητή.




Σχήμα 2.2.7: Παραδείγματα σφαλμάτων αναγνώρισης.

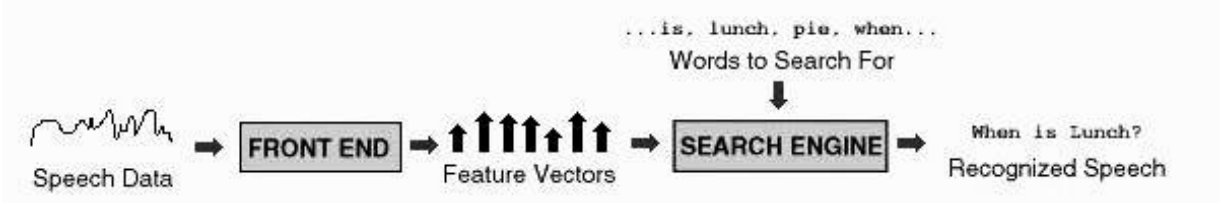
## 2.3 Ευθυγράμμιση

Το πρόβλημα της αυτόματης αναγνώρισης στίχων συνολικά δεν αφορά μόνο την ανάκτηση των λέξεων που εκφωνούνται με μια συγκεκριμένη σειρά, αλλά και τους χρόνους τους και τον τρόπο με τον οποίο ευθυγραμμίζονται με τη συνολική μουσική δομή, καθώς ο στόχος των μεταγραφών είναι να καθοδηγήσουν τους εκτελεστές και τους ακροατές σχετικά με το μουσικό κομμάτι. Από την άποψη της ανάκτησης πληροφοριών, το έργο της ανάκτησης των χρονισμών των λέξεων μπορεί να συνδεθεί με το πρόβλημα της ευθυγράμμισης (alignment).

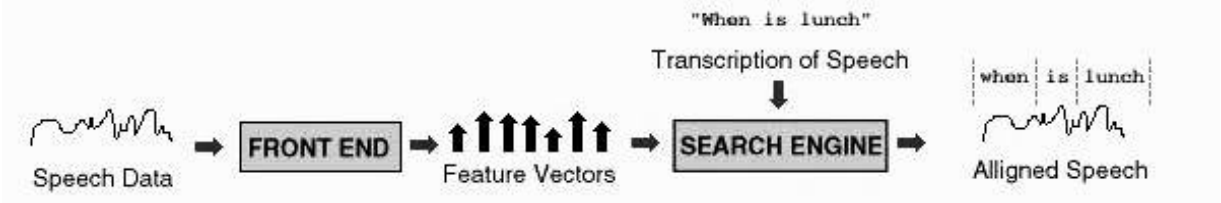
### 2.3.1 Forced Alignment

Η αναγκαστική ευθυγράμμιση αναφέρεται στη διαδικασία με την οποία οι ορθογραφικές μεταγραφές ευθυγραμμίζονται με τις ηχογραφήσεις για την αυτόματη δημιουργία τμηματοποίησης σε επίπεδο φωνήματος. Ενώ η αυτόματη ευθυγράμμιση δεν ανταγωνίζεται ακόμη τη χειροκίνητη ευθυγράμμιση, ο χρόνος που κερδίζεται μέσω της αναγκαστικής ευθυγράμμισης συχνά αξίζει τη μικρή μείωση της ακρίβειας για πολλά έργα.

Η αναγκαστική ευθυγράμμιση λειτουργεί καλύτερα σε ηχογραφήσεις που έχουν έναν ομιλητή να μιλάει κάθε φορά και έχουν ελάχιστο περιβαλλοντικό θόρυβο, αλλά και άλλοι τύποι ηχογραφήσεων μπορούν επίσης να υποστούν καλή επεξεργασία.




(a)




(b)

Σχήμα 2.3.1: Αυτόματη αναγνώριση ομιλίας με ευθυγράμμιση.

Ένα σύστημα αναγνώρισης ομιλίας χρησιμοποιεί μια μηχανή αναζήτησης μαζί με ένα ακουστικό και γλωσσικό μοντέλο που περιέχει ένα σύνολο πιθανών λέξεων, φωνημάτων ή κάποιο άλλο σύνολο δεδομένων για να αντιστοιχίσει τα δεδομένα ομιλίας με το σωστό προφορικό φώνημα. Η μηχανή αναζήτησης επεξεργάζεται τα χαρακτηριστικά που εξάγονται από τα δεδομένα ομιλίας για να εντοπίσει τις εμφανίσεις των λέξεων, των φωνημάτων ή οποιουδήποτε άλλου συνόλου δεδομένων είναι εξοπλισμένη για να αναζητήσει και επιστρέφει τα αποτελέσματα.

Η αναγκαστική ευθυγράμμιση είναι παρόμοια με αυτή τη διαδικασία, αλλά διαφέρει σε ένα σημαντικό σημείο. Αντί να δίνεται ένα σύνολο πιθανών λέξεων για αναζήτηση, δίνεται στη μηχανή αναζήτησης μια ακριβής μεταγραφή του τι λέγεται στα δεδομένα ομιλίας. Στη συνέχεια, το σύστημα ευθυγραμμίζει τα μεταγραμμένα δεδομένα με τα δεδομένα ομιλίας, προσδιορίζοντας ποια χρονικά τμήματα στα δεδομένα ομιλίας αντιστοιχούν σε συγκεκριμένες λέξεις στα δεδομένα μεταγραφής.

Η αναγκαστική ευθυγράμμιση μπορεί επίσης να χρησιμοποιηθεί για την ευθυγράμμιση των φωνημάτων των δεδομένων μεταγραφής με τα δεδομένα ομιλίας που δίνονται, παρόμοια με την παρακάτω εικόνα, αν και με πιο σαφώς καθορισμένα όρια για το πού αρχίζει και πού τελειώνει κάθε φώνημα.

### 2.3.2 Μετρικές Αξιολόγησης

Η ακρίβεια της ευθυγράμμισης μπορεί να αξιολογηθεί σε διάφορα επίπεδα λεπτομέρειας, ανάλογα με την εφαρμογή. Σε αυτό το πλαίσιο, η ακρίβεια μετράται με τη χρήση διαφόρων οντοτήτων, οι οποίες αναφέρονται ως "λυρικές μονάδες" στη συζήτηση που ακολουθεί. Αυτές οι μονάδες μπορεί να είναι φωνήματα, συλλαβές, λέξεις, γραμμές/φράσεις στίχων ή πλήρεις παράγραφοι/τμήματα στίχων. Για παράδειγμα, κατά τη δημιουργία υποτίτλων για μουσικά βίντεο, η ευθυγράμμιση σε επίπεδο γραμμής ή φράσης μπορεί να είναι επαρκής. Ωστόσο, για ακριβείς απαιτήσεις ευθυγράμμισης, όπως η αυτόματη παραγωγή στιγμιότυπων για καραόκε, είναι απαραίτητη η ευθυγράμμιση σε επίπεδο συλλαβών ή ακόμη και φωνημάτων.

Δεδομένου ότι το πρόβλημα αυτό είναι σχετικά ανεξερεύνητο, δεν υπάρχει καθιερωμένη τυποποιημένη μετρική αξιολόγησης. Έχουν προταθεί διάφορες μετρικές, αλλά καθεμία έχει χρησιμοποιηθεί μόνο σε μία ή δύο μελέτες.

#### Μέσο απόλυτο σφάλμα/απόκλιση

Η μετρική του μέσου απόλυτου σφάλματος/απόκλισης εισήχθη αρχικά από τους Mesaros και Virtanen [5]. Μετρά τη χρονική μετατόπιση μεταξύ της πραγματικής χρονοσήμανσης $t_i$ και της εκτιμώμενης αντίστοιχης $\hat{t_i}$ στην αρχή και στο τέλος κάθε λυρικής ενότητας.

$$e = \frac{1}{N_k} \sum_{word\, i} (\hat{t_i} - t_i) \tag{2.3.1}$$

Στη συνέχεια, το σφάλμα υπολογίζεται κατά μέσο όρο για όλες τις $N_k$ λέξεις του συνόλου δεδομένων. Η αξιολόγηση πραγματοποιήθηκε σε χρονοσφραγίδες στα όρια των στιχουργικών γραμμών. Οι συγγραφείς αναγνωρίζουν ότι η χρήση ενός σφάλματος σε απόλυτους όρους έχει ένα μειονέκτημα: η αντίληψη ενός σφάλματος με την ίδια διάρκεια μπορεί να διαφέρει ανάλογα με το ρυθμό του τραγουδιού.

#### Ποσοστό σωστών τμημάτων

Για να μετριαστεί η αντιληπτική εξάρτηση από τον ρυθμό, προτείνεται η μέτρηση του ποσοστού k του συνολικού μήκους των σωστά επισημασμένων ηχητικών τμημάτων σε ένα τραγούδι k, σε σχέση με το συνολικό μήκος του τραγουδιού. Το επίπεδο λεπτομέρειας στο οποίο αξιολογεί είναι οι στιχουργικές γραμμές. Αυτή η μετρική μπορεί να θεωρηθεί ως ειδική περίπτωση της μετρικής ομαδοποίησης πλαισίων που χρησιμοποιείται για την αξιολόγηση της δομικής κατάτμησης.

$$\rho^k = \frac{length\, of\, correct\, segments}{total\, length\, of\, the\, song} \times 100 \tag{2.3.2}$$

Ουσιαστικά, αυτή η μετρική είναι ισοδύναμη με το ποσοστό των σωστών τμημάτων αν θεωρήσουμε μια λυρική μονάδα ως "τμήμα". Ωστόσο, παρά το γεγονός ότι επηρεάζεται λιγότερο από τον ρυθμό και είναι πιο αυστηρό, το ποσοστό των πλαισίων δεν παρέχει μια διαισθητική εκτίμηση από αντιληπτική άποψη, καθώς η συσχέτισή του με την έκταση του απόλυτου σφάλματος δεν είναι άμεσα εμφανής.

#### Ποσοστό ορθών εκτιμήσεων σύμφωνα με ένα παράθυρο ανοχής

Μια μετρική που λαμβάνει υπόψη ότι οι μετατοπίσεις από τη βασική αλήθεια κάτω από ένα ορισμένο όριο θα μπορούσαν να γίνουν ανεκτές από τους ανθρώπινους ακροατές. Οι συγγραφείς αξιολογούν το μέσο ποσοστό των εκτιμήσεων του χρόνου έναρξης $\hat{t_i}$ που εμπίπτουν εντός δευτερολέπτων από τον χρόνο έναρξης $t_i$ της αντίστοιχης μονάδας στίχων βασικής αλήθειας.

$${\rho_t}^k = \frac{1}{N_k} \sum_{word\, i} 1_{|\hat{t_i} - t_i| < r} \times 100 \tag{2.3.3}$$

όπου N k είναι ο αριθμός των λέξεων σε ένα δεδομένο τραγούδι k. Η τελική μετρική υπολογίζεται υπολογίζοντας το μέσο όρο k σε όλα τα τραγούδια.

# Κεφάλαιο 3

# Τρέχουσα τεχνολογική στάθμη (State-of-the-art)

Στο παρόν Κεφάλαιο θα θέσουμε τα θεμέλια πάνω στα οποία βασίζεται η εργασία μας. Παρουσιάζουμε το τεχνικό μας υπόβαθρο, όσον αφορά τη Μηχανική Μάθηση και ειδικότερα το υποπεδίο της Βαθιάς Μάθησης. Αναλύουμε προσεγγίσεις της βαθιάς μάθησης για το πρόβλημα της αυτόματης αναγνώρισης ομιλίας και εστιάζουμε στη βάση των μοντέλων μας, την αρχιτεκτονική Transformer, και τη χρήση της στη μοντελοποίηση ακολουθίας προς ακολουθία. Τέλος, παρουσιάζουμε βασικές τεχνικές μάθησης όπως η μεταφορά μάθησης και η λεπτομερής ρύθμιση.

## 3.1 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη (ΤΝ) γνωστή και ως Artificial Intelligence (AI) μπορεί να οριστεί ως η νοημοσύνη που επιδεικνύουν οι μηχανές, ένας τύπος νοημοσύνης που μπορεί να οριστεί ή να υπολογιστεί, οπότε διαφέρει από τη φυσική νοημοσύνη που συναντάται στις μορφές ζωής, η οποία περιλαμβάνει συνείδηση και συναισθηματισμό. Μια πρώτη διάκριση γίνεται στην ισχυρή ΤΝ (Τεχνητή Γενική Νοημοσύνη), η οποία σχετίζεται με την υποθετική ικανότητα μιας μηχανής να κατανοεί και να μαθαίνει οποιαδήποτε διανοητική εργασία που θεωρείται χαρακτηριστική της ανθρώπινης νοημοσύνης, και στην ασθενή τεχνητή νοημοσύνη, η οποία επικεντρώνεται στην επίλυση μιας συγκεκριμένης εργασίας και είναι πιο κοντά στο επίπεδο που έχει επιτευχθεί μέχρι σήμερα. Ως αφετηρία της ιστορίας της Τεχνητής Νοημοσύνης αναγνωρίζεται γενικά η εργασία των McCulloch και Pitts για τους τεχνητούς νευρώνες. Στην εργασία αυτή, οι συγγραφείς περιγράφουν ένα απλό υπολογιστικό μοντέλο λογικής συνάρτησης για ένα κύτταρο, που ονομάζεται νευρώνας, σε αυτό που πιστεύεται ότι είναι η πρώτη περιγραφή των νευρωνικών δικτύων. Η ανάπτυξη αυτού του μοντέλου, καθώς και οι συνδέσεις με τη βιολογία που γίνονται, είναι φυσικά προϊόν της εποχής του, δεδομένου ότι η περίοδος αυτή σημαδεύτηκε από τις ανακαλύψεις στη νευροβιολογία, τη θεωρία της πληροφορίας και την κυβερνητική, καθώς και από τη διορατικότητα που έδωσε η θεωρία του υπολογισμού του Alan Turing. Η τεχνητή νοημοσύνη άρχισε να αποτελεί ξεχωριστό πεδίο το 1956, αφού διαχωρίστηκε από την κυβερνητική. Κατά τη διάρκεια των επόμενων 50 ετών, νέες παραλλαγές αυτής της νέας τεχνολογίας χρησιμοποιήθηκαν για την επίλυση προβλημάτων που ήταν εύκολο να διατυπωθούν και δύσκολο να λυθούν από τον άνθρωπο, όπως για παράδειγμα το παιχνίδι της ντάμας (1954) ή η απόδειξη λογικών θεωρημάτων (1956). Η πιο επιτυχημένη και διάσημη εφαρμογή ήταν ο Deep Blue3, ένα μοντέλο που παίζει σκάκι, το οποίο το 1997 κατάφερε να νικήσει τον παγκόσμιο πρωταθλητή, Garry Kasparov.

Κατά τη διάρκεια αυτής της περιόδου, υπήρξαν πολλές διαφορετικές προσεγγίσεις στην τεχνητή νοημοσύνη, καθώς και φάσεις ανάπτυξης και παρακμής, αλλά το αναλλοίωτο ήταν το εμπόδιο που παρουσιάστηκε κατά την δοκιμή της με δύσκολα τυποποιήσιμες εργασίες, τις οποίες οι άνθρωποι μπορούν να εκτελέσουν εύκολα, σχεδόν χωρίς να σκέφτονται, αλλά οι μηχανές δεν μπορούν. Μια προσέγγιση για την επίλυση αυτών των καθηκόντων του πραγματικού κόσμου, ήταν η εισαγωγή σκληρά κωδικοποιημένης γνώσης του κόσμου σε ένα

μοντέλο, χρησιμοποιώντας ειδικές τυπικές γλώσσες. Αυτό περιελάμβανε μια μεθοδολογία που ονομάζεται τεχνητή νοημοσύνη βασισμένη στη γνώση. Λόγω της αυξανόμενης δυσκολίας και της ανθρώπινης προσπάθειας που απαιτείται για την κωδικοποίηση αυτής της γνώσης, το ερευνητικό ενδιαφέρον απομακρύνθηκε. Μια διαφορετική σχολή σκέψης υποστήριξε ότι τα μοντέλα που αναπτύσσονται πρέπει να είναι σε θέση να αποκτούν τα ίδια τη γνώση, χωρίς να εξαρτώνται από ανθρώπινους εμπειρογνώμονες και ρητές αναπαραστάσεις. Αντίστοιχα με την ανθρώπινη διαδικασία μάθησης, θα μπορούσαν να βελτιώνονται μέσω της εμπειρίας και των δεδομένων. Αυτή η προσέγγιση ονομάζεται μηχανική μάθηση (ΜΜ) και μπορεί να θεωρηθεί ως μέρος ή υποπεδίο της ΤΝ. Και τα δύο πεδία έχουν κοινές ρίζες και στόχους, με την ΜΜ να εστιάζει μακριά από τις προαναφερθείσες συμβολικές αναπαραστάσεις, ενώ δανείζεται μεθόδους και μοντέλα από τη στατιστική και τη θεωρία πιθανοτήτων.

## 3.2 Μηχανική Μάθηση

Η μηχανική μάθηση (Machine Learning - ML) είναι ένα υποπεδίο της Τεχνητής Νοημοσύνης, το οποίο μελετά αλγορίθμους υπολογιστών οι οποίοι βελτιώνονται αυτόματα μέσω δεδομένων και εμπειρίας. Οι αλγόριθμοι μηχανικής μάθησης δημιουργούν ένα μοντέλο με βάση δειγματικά δεδομένα, προκειμένου να κάνουν προβλέψεις ή να λαμβάνουν αποφάσεις χωρίς να απαιτείται ρητός προγραμματισμός συγκεκριμένης εργασίας. Οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται με μεγάλη επιτυχία για διάφορες εφαρμογές, ιδίως σε σενάρια στα οποία είναι δύσκολο για έναν άνθρωπο να σχεδιάσει χειροκίνητα αλγόριθμους για την επίλυσή τους, όπως στην στην Επεξεργασία Ομιλίας και στην Όραση Υπολογιστών. Ο όρος Μηχανική Μάθηση επινοήθηκε το 1959 από τον Arthur Samuel. Ένας συχνά αναφερόμενος, επίσημος ορισμός της ML δίνεται από τον Tom Mitchel: " Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία E σε σχέση με κάποια κατηγορία εργασιών T και μέτρο απόδοσης P αν η απόδοσή του σε εργασίες στο T, όπως μετράται από το P, βελτιώνεται με την εμπειρία E. " Η πιο συνηθισμένη διάκριση των προσεγγίσεων Μηχανικής Μάθησης εξαρτάται από την ανατροφοδότηση που είναι διαθέσιμη στον αλγόριθμο μάθησης και ο διαχωρισμός είναι τριπλός. Οι τρεις κατηγορίες αναλύονται εν συντομία παρακάτω, αλλά θα πρέπει να σημειωθεί ότι υπάρχουν προσεγγίσεις που δεν εντάσσονται σε αυτή την κατηγοριοποίηση ή συστήματα που χρησιμοποιούν περισσότερες από μία προσεγγίσεις, για παράδειγμα μάθηση με ημιεπίβλεψη.

#### Μάθηση με επίβλεψη

Η μάθηση με επίβλεψη μπορεί να θεωρηθεί ως μια οικογένεια αλγορίθμων που μαθαίνουν μια συνάρτηση από τα παραδείγματα εισόδου σε τιμές-στόχους δεδομένου ενός συνόλου δεδομένων, για τα οποία οι τιμές-στόχοι είναι γνωστές. Αυτά τα ζεύγη εισόδου-εξόδου ονομάζονται δεδομένα εκπαίδευσης, ενώ η έξοδος αναφέρεται ως βασική αλήθεια. Τα μοντέλα με επίβλεψη σχεδιάζονται για να εξάγουν μια συνάρτηση αντιστοίχησης $g$ που προσεγγίζει τις αφανείς σχέσεις του συνόλου δεδομένων εκπαίδευσης, με στόχο να κάνει προβλέψεις για προηγουμένως αθέατες εισόδους.

Έτσι, δεδομένου ενός συνόλου $N$ παραδειγμάτων εκπαίδευσης της μορφής$\{(x_1, y_1), ..., (x_N,\ y_N)\}$ έτσι ώστε $x_i$ είναι το διάνυσμα χαρακτηριστικών του i-οστού παραδείγματος και $y_i$ είναι η ετικέτα του (δηλαδή η κλάση ή η τιμή), ο αλγόριθμος μάθησης αναζητά μια συνάρτηση $g : X \to Y$ , όπου $X$ είναι ο χώρος εισόδου και $Y$ είναι ο χώρος εξόδου. Μπορούμε να συμβολίσουμε μια συνάρτηση βαθμολόγησης $f : X \times Y \to \mathbb{R}$, έτσι ώστε η $g$ να επιστρέφει την τιμή $y$ με την υψηλότερη βαθμολογία: $g(x) = \arg\max_y f(x, y)$. Για να μετρήσουμε πόσο καλά μια συνάρτηση ταιριάζει στα δεδομένα εκπαίδευσης, ορίζεται μια συνάρτηση σφάλματος $\mathcal{L} : Y \times Y \to \mathbb{R}^{\geq 0}$. Θέλουμε να ελαχιστοποιήσουμε το άθροισμα της συνάρτησης σφάλματος για όλα τα παραδείγματα εκπαίδευσης, με το σφάλμα πρόβλεψης κάθε τιμής εξόδου να είναι $\mathcal{L}(y_i, \hat{y})$, για ένα παράδειγμα εκπαίδευσης $(x_i,\ y_i)$ και πρόβλεψη $\hat{y}$.

Οι αλγόριθμοι με επίβλεψη χωρίζονται σε δύο κύριες κατηγορίες, με βάση την επιθυμητή έξοδο: (α) ταξινόμηση και (β) παλινδρόμηση. Η πρώτη αναφέρεται στην πρόβλεψη εξόδων που είναι περιορίζονται σε ένα διακριτό σύνολο τιμών, που ονομάζεται κλάση, ενώ η δεύτερη αναφέρεται στο πρόβλημα της εκτίμησης εξόδων πραγματικών τιμών εντός ενός προκαθορισμένου εύρους.

### Μάθηση χωρίς επίβλεψη

Η μάθηση χωρίς επίβλεψη είναι μια οικογένεια αλγορίθμων που μαθαίνουν να συμπεραίνουν μοτίβα, χωρίς να δίνονται τιμές-στόχοι για κάθε παράδειγμα μάθησης. Ο αλγόριθμος ανακαλύπτει την υποκείμενη δομή ή κατανομή των δεδομένων, μέσω της μίμησης. Ένα πολύ συνηθισμένο πρόβλημα μη επιβλεπόμενης μάθησης είναι η ομαδοποίηση, κατά την οποία ένα μοντέλο ομαδοποιεί ή χωρίζει τα σημεία των δεδομένων σε κατηγορίες με βάση ένα μέτρο ομοιότητας, έτσι ώστε τα σημεία που ανήκουν στην ίδια συστάδα να είναι περισσότερο όμοια με τα σημεία που ανήκουν σε άλλες συστάδες. Ένα άλλο σύνολο τεχνικών και εργασιών είναι η μάθηση αναπαράστασης, όπου το μοντέλο ανακαλύπτει σημαντικές αναπαραστάσεις που μπορούν αργότερα να χρησιμοποιηθούν για να βοηθήσουν σε άλλες εργασίες, αντικαθιστώντας τη χειροκίνητη μηχανική χαρακτηριστικών. Τέλος, πολλά παραγωγικά μοντέλα μπορούν να δημιουργήσουν νέα δεδομένα που προέρχονται από μια κατανομή πραγματικών δεδομένων που δίνονται ως είσοδος.

### Ενισχυτική μάθηση

Η ενισχυτική μάθηση είναι ένας τομέας της Μηχανικής Μάθησης που ερευνά τον τρόπο με τον οποίο ευφυείς παράγοντες αναλαμβάνουν δράσεις σε ένα δυναμικό περιβάλλον, με στόχο τη μεγιστοποίηση μιας σωρευτικής επιβράβευσης. Όπως και στη μάθηση χωρίς επίβλεψη, δεν υπάρχει ανάγκη για επισημασμένα ζεύγη εισόδου-εξόδου. Το επίκεντρο αυτής της προσέγγισης είναι η εξεύρεση ισορροπίας μεταξύ της εξερεύνησης (αχαρτογράφητου εδάφους) και της εκμετάλλευσης (της τρέχουσας γνώσης), ενώ οι μη βέλτιστες ενέργειες δεν χρειάζεται να διορθώνονται ρητά. Το πρόβλημα μπορεί να διατυπωθεί με τη μορφή μιας διαδικασίας απόφασης Markov. Για να δώσουμε έναν πιο επίσημο ορισμό, ας συμβολίσουμε:

- ένα σύνολο καταστάσεων του περιβάλλοντος και του παράγοντα, $S$
- ένα σύνολο ενεργειών του παράγοντα, $A$
- την πιθανότητα μετάβασης (τη χρονική στιγμή $t$) από την κατάσταση $s$ στην κατάσταση $s'$ υπό την ενέργεια $\alpha$ $P_\alpha(s, s') = \Pr(s_{t+1} = s' \mid s_t = s, \alpha_t = \alpha)$
- η άμεση επιβράβευση μετά τη μετάβαση από την κατάσταση $s$ στην κατάσταση $s'$ με τη δράση $\alpha$ $R_\alpha(s, s')$

Σε κάθε διακριτό χρονικό βήμα $t$, ο παράγοντας λαμβάνει την τρέχουσα κατάσταση $s_t$ και την ανταμοιβή $r_t$ και στη συνέχεια επιλέγει μια ενέργεια $\alpha_t$ από ένα σύνολο διαθέσιμων ενεργειών. Το περιβάλλον, δεδομένης της $\alpha_t$ , μετακινείται σε μια νέα κατάσταση $s_{t+1}$ και ο πράκτορας λαμβάνει τη νέα ανταμοιβή $r_{t+1}$ που σχετίζεται με τη μετάβαση $(s_t, a_t, s_{t+1})$. Ο στόχος του πράκτορα είναι να μάθει μια πολιτική:$\pi : A \times S \to [0, 1]$, $\pi(a, s) = \Pr(a_t = a \mid s_t = s)$ που μεγιστοποιεί την αναμενόμενη αθροιστική ανταμοιβή.

## 3.2.1 Συνάρτηση Σφάλματος

Ο στόχος των αλγορίθμων επιβλεπόμενης μάθησης είναι η προσέγγιση μίας συνάρτησης από τα δεδομένα σε κάποια δοσμένη τιμή. Ο τρόπος που επιτυγχάνεται αυτός ο σκοπός είναι πολύ συχνά, όπως στην περίπτωση των νευρωνικών δικτύων, χρησιμοποιώντας μια επιπλέον συνάρτηση που αντιπροσωπεύει την απόδοση του μοντέλου. Η συνάρτηση αυτή είναι συνάρτηση των δεδομένων του προβλήματος, αλλά και των παραμέτρων του μοντέλου που χρησιμοποιείται. Στη συνέχεια, με μια διαδικασία που θα εξηγηθεί αργότερα, βρίσκουμε ένα σύνολο από παραμέτρους στις οποίες η συνάρτηση παίρνει μια επιθυμητή τιμή, συνήθως ελάχιστο ή μέγιστο. Όταν η συνάρτηση ποσοτικοποιεί το σφάλμα του μοντέλου στα δεδομένα, λέγεται συνάρτηση σφάλματος και ο σκοπός της εκπαίδευσης είναι η ελαχιστοποίηση της.

Δεδομένου ενός συνόλου εκπαίδευσης με δεδομένα $x_i,\ i = 1, \ldots, n$, *labels* $y_i,\ i = 1, \ldots, n$, ενός μοντέλου που υλοποιεί την συνάρτηση $f$ η οποία με παραμέτρους $\theta$ προσπαθεί να προβλέψει τα *labels* από τα δεδομένα, αρχικά ορίζουμε μια συνάρτηση κόστους ανά δεδομένο $L(y, \hat{y})$ όπου $\hat{y}_i = f(x_i, \theta)$. Στη συνέχεια, ορίζουμε τη συνολική απώλεια του μοντέλου με παραμέτρους $\theta$ πάνω σε όλο το σύνολο των δεδομένων ως τη μέση απώλεια πάνω σε όλα τα δεδομένα εκπαίδευσης.

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{y_i}) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i, \theta)) \tag{3.2.1}$$

Ο στόχος τώρα της εκπαίδευσης είναι να υπολογίσουμε την τιμή του $\theta$ για την οποία έχουμε το ελάχιστο σφάλμα, δηλαδή:

$$\hat{\theta} = \arg\min_{\theta} L(\theta) = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i, \theta)) \tag{3.2.2}$$

Παρατηρούμε ότι η συνάρτηση σφάλματος είναι συνάρτηση μόνο των παραμέτρων του μοντέλου αφού η εξάρτηση από κάθε δεδομένο εξαφανίστηκε αθροίζοντας. Φυσικά η εξάρτηση από τα δεδομένα παραμένει έμμεση από τον ορισμό της συνάρτησης. Ελπίζουμε ότι με κατάλληλη επιλογή των δεδομένων, η συνάρτηση είναι αρκετά γενική ώστε να αποτελεί καλή μετρική για το πρόβλημα, και ότι ένα μοντέλο που πετυχαίνει ικανοποιητική τιμή σε αυτή τη συνάρτηση θα λύνει ικανοποιητικά το πρόβλημα και για εισόδους που δεν έχει ξαναδεί.

Η επιλογή της συνάρτησης σφάλματος είναι ένα πολύ σημαντικό κομμάτι της μηχανικής μάθησης. Συνήθως επιλέγεται ώστε να αντιπροσωπεύει ένα μέτρο απόστασης μεταξύ των δύο τιμών, που στην γενική περίπτωση είναι διανύσματα, με τέτοιο τρόπο ώστε όταν το διάνυσμα που βγάζει το μοντέλο είναι ίδιο με το ζητούμενο, το σφάλμα είναι 0. Κοινές επιλογές για συνάρτηση σφάλματος είναι οι $p$-νόρμες οι οποίες ορίζονται:

$$||x - y||_p = \left( \sum_{i=1}^{\dim(x)} |x_i - y_i|^p \right)^{1/p} \tag{3.2.3}$$

Ιδιαίτερα συνηθισμένη είναι η 1-νόρμα που λέγεται και απόλυτο σφάλμα, καθώς και το τετράγωνο της 2-νόρμας που αντιστοιχεί στην Ευκλείδια απόσταση και ονομάζεται τετραγωνικό σφάλμα. Άλλες συναρτήσεις σφάλματος θα αναλυθούν αργότερα.

Οι συναρτήσεις σφάλματος, χρησιμοποιούνται και σε μη επιβλεπόμενη μάθηση, απλά σε αυτή τη περίπτωση δεν αντιστοιχούν απλά σε μια μετρική απόσταση από τις εξόδους του μοντέλου με τις επιθυμητές εξόδους, και είναι πολύ διαφορετικές μεταξύ τους ανάλογα με το πρόβλημα. Η βασική ιδέα πάλι παραμένει ότι η συνάρτηση αντιπροσωπεύει το πόσο καλά το μοντέλο λύνει το ζητούμενο πρόβλημα.

Στα περισσότερα προβλήματα, ο τρόπος που βρίσκονται οι ζητούμενες τιμές των παραμέτρων $\theta$ είναι με την μέθοδο *Gradient Descent*. Ο τρόπος που επιλέγουμε την μοντελοποίηση της συνάρτησης είναι έτσι ώστε η έξοδος του μοντέλο να είναι παραγωγίσιμη. Επίσης, η συνάρτηση σφάλματος επιλέγεται να είναι και αυτή παραγωγίσιμη και σαν αποτέλεσμα, μπορούμε να βρούμε την παράγωγο της συνάρτησης σφάλματος ως προς την κάθε παράμετρο του μοντέλου. Επειδή οι παράμετροι είναι περισσότεροι από μία, ουσιαστικά βρίσκουμε το διάνυσμα που αποτελεί την κλίση της συνάρατησης σφάλματος, το οποίο μας δείχνει την κατεύθυνση στην οποία η συνάρτηση σφάλματος μεγαλώνει με τον μεγαλύτερο ρυθμό. Στη συνέχεια αλλάζουμε τις τιμές των παραμέτρων ώστε να κινηθούν προς στην αντίθετη κατεύθυνση από την κλίση, κάτι που θα μειώσει την τιμή της συνάρτησης σφάλματος αν το βήμα που κάναμε είναι αρκετά μικρό. Στο τέλος της εκπαίδευσης, ελπίζουμε ότι θα έχουμε καταλήξει σε ένα τοπικό ελάχιστο της συνάρτησης και ιδανικά σε ένα τοπικό ελάχιστο με όσο το δυνατόν μικρότερη τιμή σφάλματος. Αυτό βέβαια δεν εξασφαλίζεται μαθηματικά εκτός από σε πολύ περιορισμένες περιπτώσεις.

Συγκεκριμένα, οι τιμές των παραμέτρων στην επανάληψη $n$ ανανεώνονται ως εξής:

$$\theta_{n+1} = \theta_n - \gamma \nabla L(\theta_n) \tag{3.2.4}$$

όπου το $\gamma$ ονομάζεται ρυθμός μάθησης.

Επειδή η συνάρτηση κόστους εξαρτάται από όλα τα δεδομένα και το κόστος υπολογισμού της σε χρόνο και μνήμη είναι μεγάλο, συνήθως η βελτιστοποίηση της συνάρτησης γίνεται μεσω του *stochastic gradient descent* που προσεγγίζει την κλίση της συνάρτησης χρησιμοποιώντας μόνο μερικά δείγματα από τα δεδομένα:

$$\theta_{n+1} = \theta_n - \gamma \nabla \frac{1}{b} \sum_{i=1}^{b} L(y_i, f(x_i, \theta_n)) \tag{3.2.5}$$

όπου το $b$ ονομάζεται *batch size*.

Παράμετροι όπως ο ρυθμός μάθησης και το *batch size*, που καθορίζονται πριν την εκπαίδευση και δεν αλλάζουν ανάλογα με την απόδοση του μοντέλου ονομάζονται υπερπαράμετροι.

Στην πράξη, χρησιμοποιούνται πολλοί διαφορετικοί αλγόριθμοι βελτιστοποίησης, οι οποίοι αποτελούν παραλλαγές του *stochastic gradient descent* και εξασφαλίζουνε καλύτερη ταχύτητα σύγκλισης, μεγαλύτερη πιθανότητα σύγκλισης σε επιθυμητό σημείο ή άλλα επιθυμητά χαρακτηριστικά. Αυτό συχνά γίνεται αλλάζοντας το ρυθμό μάθησης κατά την διάρκεια της εκπαίδευσης με κάποιο προκαθορισμένο τρόπο. Για παράδειγμα, καθώς το μοντέλο γίνεται καλύτερο, ο ρυθμός μάθησης πολύ συχνά μειώνεται έτσι ώστε το μοντέλο να μάθει πιο πολλές λεπτομέρειες καθώς δεν χρειάζονται μεγάλες αλλαγές στα βάρη. Ένας πολύ συνηθισμένος αλγόριθμος βελτιστοποίησης είναι ο *Adam* [17].

## 3.3 Μεταφορά Μάθησης

Η μάθηση μεταφοράς (Transfer Learning) είναι ένα ερευνητικό πρόβλημα στη μηχανική μάθηση (Machine Learning) που επικεντρώνεται στην εφαρμογή της γνώσης που αποκτήθηκε κατά την επίλυση μιας εργασίας σε μια συναφή εργασία. Για παράδειγμα, η γνώση που αποκτήθηκε κατά την εκμάθηση της αναγνώρισης αυτοκινήτων θα μπορούσε να εφαρμοστεί κατά την προσπάθεια αναγνώρισης φορτηγών. Το θέμα αυτό σχετίζεται με την ψυχολογική βιβλιογραφία για τη μεταφορά της μάθησης, αν και οι πρακτικοί δεσμοί μεταξύ των δύο πεδίων είναι περιορισμένοι. Η επαναχρησιμοποίηση/μεταφορά πληροφοριών από εργασίες που έχουν διδαχθεί προηγουμένως σε νέες εργασίες έχει τη δυνατότητα να βελτιώσει σημαντικά την αποτελεσματικότητα της μάθησης.

Στη μάθηση μεταφοράς(Transfer Learning) ή προσαρμογή τομέα (Domain Adaptation), εκπαιδεύουμε το μοντέλο με ένα σύνολο δεδομένων. Στη συνέχεια, εκπαιδεύουμε το ίδιο μοντέλο με ένα άλλο σύνολο δεδομένων που έχει διαφορετική κατανομή των κλάσεων, ή ακόμη και με άλλες κλάσεις από ό,τι στο πρώτο σύνολο δεδομένων εκπαίδευσης).

Στο *Fine-tuning*, μια προσέγγιση της Μάθησης Μεταφοράς, έχουμε ένα σύνολο δεδομένων και χρησιμοποιούμε ας πούμε το 90% αυτού στην εκπαίδευση. Στη συνέχεια, εκπαιδεύουμε το ίδιο μοντέλο με το υπόλοιπο 10%. Συνήθως, αλλάζουμε τον ρυθμό μάθησης σε μικρότερο, ώστε να μην έχει σημαντικό αντίκτυπο στα ήδη προσαρμοσμένα βάρη. Υπάρχει ένα βασικό μοντέλο που λειτουργεί για μια παρόμοια εργασία και στη συνέχεια να "παγώνουν" ορισμένα από τα επίπεδα για να διατηρηθεί η παλιά γνώση όταν εκτελείται η νέα συνεδρία εκπαίδευσης με τα νέα δεδομένα. Το στρώμα εξόδου μπορεί επίσης να είναι διαφορετικό και να έχει παγώσει κάποιο από αυτό όσον αφορά την εκπαίδευση.

Χρησιμοποιώντας Transfer Learning χρειάζεται να παγώσουν κάποια στρώματα, κυρίως τα προ-εκπαιδευμένα και να εκπαιδευθούν μόνο στα προστιθέμενα, και να μειωθεί ο ρυθμός μάθησης για να προσαρμοστούν τα βάρη χωρίς να αναμειχθεί η σημασία τους για το δίκτυο. Αν αυξάναμε το ρυθμό μάθησης, μπορεί να προκύψουν φτωχά αποτελέσματα λόγω των μεγάλων βημάτων στη βελτιστοποίηση της κλίσης καθόδου. Αυτό μπορεί να οδηγήσει σε μια κατάσταση όπου το νευρωνικό δίκτυο δεν μπορεί να βρει το παγκόσμιο ελάχιστο αλλά μόνο ένα τοπικό.

Χρησιμοποιώντας ένα προ-εκπαιδευμένο μοντέλο σε μια παρόμοια εργασία, συνήθως έχουμε εξαιρετικά αποτελέσματα όταν χρησιμοποιούμε το Fine-tuning. Ωστόσο, εάν δεν υπάρχουν επαρκή δεδομένα στο νέο σύνολο δεδομένων ή ακόμη και οι υπερπαράμετροι δεν είναι οι καλύτερες, μπορεί να έχετε μη ικανοποιητικά αποτελέσματα. Η μηχανική μάθηση εξαρτάται πάντα από το σύνολο δεδομένων και τις παραμέτρους του δικτύου.

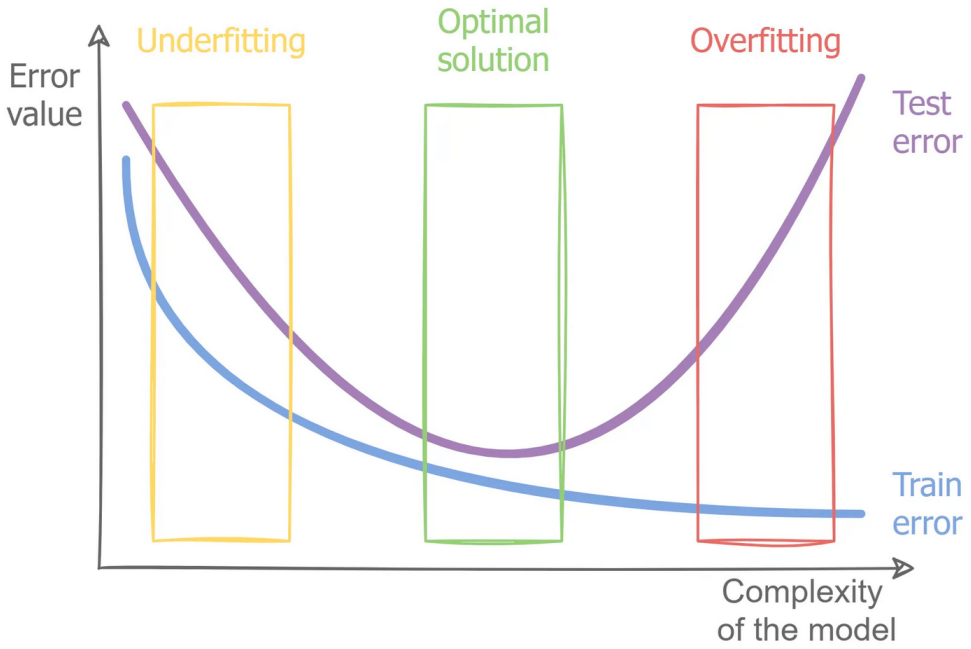
## 3.4 Υπερπροσαρμογή και υποπροσαρμογή

### Υπερπροσαρμογή

Ένα στατιστικό μοντέλο λέγεται ότι είναι υπερβολικά προσαρμοσμένο όταν το μοντέλο δεν κάνει ακριβείς προβλέψεις σε δεδομένα δοκιμών. Όταν ένα μοντέλο εκπαιδεύεται με τόσα πολλά δεδομένα, αρχίζει να μαθαίνει από το θόρυβο και τις ανακριβείς καταχωρήσεις δεδομένων στο σύνολο δεδομένων μας. Και όταν η δοκιμή με δεδομένα δοκιμής οδηγεί σε Υψηλή διακύμανση. Τότε το μοντέλο δεν κατηγοριοποιεί σωστά τα δεδομένα, λόγω των πολλών λεπτομερειών και του θορύβου. Οι αιτίες της υπερπροσαρμογής είναι οι μη παραμετρικές και μη γραμμικές μέθοδοι, επειδή αυτοί οι τύποι αλγορίθμων μηχανικής μάθησης έχουν μεγαλύτερη ελευθερία στην κατασκευή του μοντέλου με βάση το σύνολο δεδομένων και επομένως μπορούν πραγματικά να κατασκευάσουν μη ρεαλιστικά μοντέλα. Μια λύση για την αποφυγή της υπερπροσαρμογής είναι η χρήση ενός γραμμικού αλγορίθμου αν έχουμε γραμμικά δεδομένα ή η χρήση των παραμέτρων όπως το μέγιστο βάθος αν χρησιμοποιούμε δέντρα αποφάσεων.

Με λίγα λόγια, η υπερπροσαρμογή είναι ένα πρόβλημα όπου η αξιολόγηση των αλγορίθμων μηχανικής μάθησης σε δεδομένα εκπαίδευσης είναι διαφορετική από τα αθέατα δεδομένα. Ορισμένοι λόγοι για την υπερπροσαρμογή μπορεί να είναι η υψηλή διακύμανση και η χαμηλή μεροληψία η υψηλή πολυπλοκότητα του μοντέλου ή το μέγεθος των δεδομένων εκπαίδευσης.

Μπορούν να χρησιμοποιηθούν διάφορες τεχνικές για τη μείωση της υπερπροσαρμογής, όπως η αύξηση των δεδομένων εκπαίδευσης ή η μείωση της πολυπλοκότητας του μοντέλου. Μια άλλη τεχνική είναι η έγκαιρη διακοπή κατά τη φάση της εκπαίδευσης (παρακολούθηση των απωλειών κατά τη διάρκεια της περιόδου εκπαίδευσης, μόλις οι απώλειες αρχίσουν να αυξάνονται σταματά η εκπαίδευση). Μπορεί κανείς επίσης να χρησιμοποιήσει τη διακοπή (dropout) [18] για τα νευρωνικά δίκτυα για την αντιμετώπιση της υπερπροσαρμογής




Σχήμα 3.4.1: Σχέση σφάλματος και πολυπλοκότητας του μοντέλου.

### Υποπροσαρμογή

Ένα στατιστικό μοντέλο ή ένας αλγόριθμος μηχανικής μάθησης λέγεται ότι έχει υποπροσαρμογή όταν δεν μπορεί να συλλάβει την υποκείμενη τάση των δεδομένων, δηλαδή αποδίδει καλά μόνο σε δεδομένα εκπαίδευσης αλλά αποδίδει ελάχιστα σε δεδομένα δοκιμής. Η υποπροσαρμογή καταστρέφει την ακρίβεια του μοντέλου μηχανικής μάθησης. Η εμφάνισή της σημαίνει απλώς ότι το μοντέλο μας ή ο αλγόριθμος δεν προσαρμόζεται

αρκετά καλά στα δεδομένα. Συνήθως συμβαίνει όταν έχουμε λιγότερα δεδομένα για να φτιάξουμε ένα ακριβές μοντέλο και επίσης όταν προσπαθούμε να φτιάξουμε ένα γραμμικό μοντέλο με λιγότερα μη γραμμικά δεδομένα. Σε τέτοιες περιπτώσεις, οι κανόνες του μοντέλου μηχανικής μάθησης είναι πολύ εύκολοι και ευέλικτοι για να εφαρμοστούν σε τόσο ελάχιστα δεδομένα, και ως εκ τούτου το μοντέλο θα κάνει πιθανότατα πολλές λανθασμένες προβλέψεις. Η υποπροσαρμογή μπορεί να αποφευχθεί με τη χρήση περισσότερων δεδομένων και επίσης με τη μείωση των χαρακτηριστικών με την επιλογή χαρακτηριστικών.

Με λίγα λόγια, η υποπροσαρμογή αναφέρεται σε ένα μοντέλο που αποδίδει ελάχιστα τόσο στα δεδομένα εκπαίδευσης όσο και στα νέα, αθέατα δεδομένα. Διάφοροι παράγοντες συμβάλλουν στην υποπροσαρμογή, όπως η υψηλή μεροληψία και η χαμηλή διακύμανση, το ανεπαρκές μέγεθος του συνόλου δεδομένων εκπαίδευσης, η υπερβολικά απλή πολυπλοκότητα του μοντέλου και τα μη καθαρισμένα δεδομένα εκπαίδευσης που περιέχουν θόρυβο.

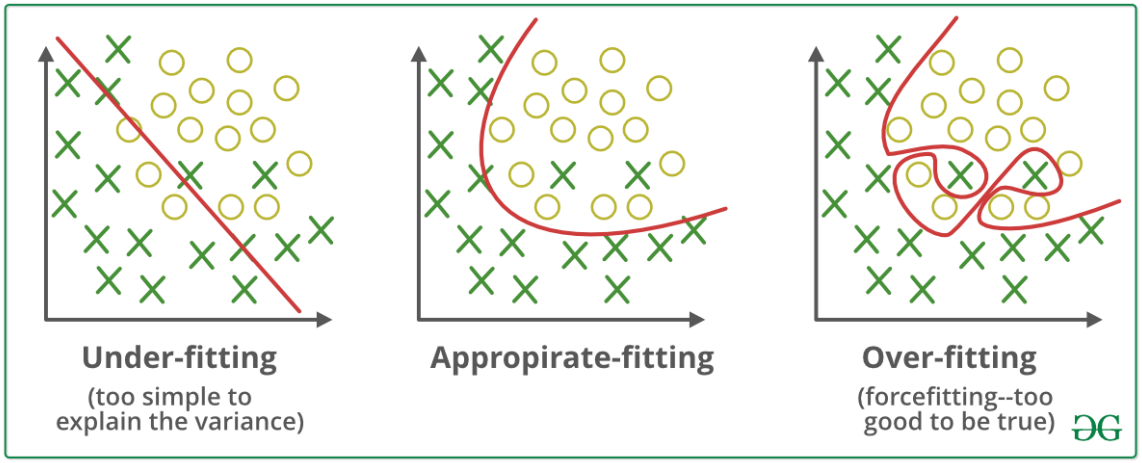
Για την αντιμετώπιση της υποπροσαρμογής, μπορούν να χρησιμοποιηθούν διάφορες τεχνικές:

Αύξηση της πολυπλοκότητας του μοντέλου, με την εισαγωγή πιο σύνθετων μοντέλων, όπως η χρήση πολυωνυμικών συναρτήσεων υψηλότερου βαθμού ή η αύξηση του αριθμού των επιπέδων σε ένα νευρωνικό δίκτυο, μπορούμε να ενισχύσουμε την ικανότητα του μοντέλου να συλλαμβάνει περίπλοκα μοτίβα στα δεδομένα.

Αύξηση του αριθμού των χαρακτηριστικών και εκτέλεση μηχανικής χαρακτηριστικών, επεκτείνοντας το σύνολο των χαρακτηριστικών με τη συμπερίληψη σχετικών μεταβλητών ή τη δημιουργία νέων χαρακτηριστικών μέσω της γνώσης του τομέα μπορεί να παρέχει στο μοντέλο περισσότερες πληροφορίες για μάθηση.

Η αφαίρεση του θορύβου από τα δεδομένα με τον καθαρισμό των δεδομένων εκπαίδευσης με την εξάλειψη των ακραίων τιμών, τη διόρθωση των σφαλμάτων ή την εξομάλυνση των θορυβωδών δεδομένων μπορεί να βοηθήσει το μοντέλο να επικεντρωθεί στις υποκείμενες τάσεις και τα μοτίβα, μειώνοντας την επιρροή των άσχετων ή λανθασμένων πληροφοριών.

Αύξηση του αριθμού των εποχών ή της διάρκειας της εκπαίδευσης με την επέκταση της διαδικασίας εκπαίδευσης είτε αυξάνοντας τον αριθμό των εποχών είτε αφήνοντας περισσότερο χρόνο στο μοντέλο να μάθει μπορεί να βελτιώσει την απόδοσή του προσαρμόζοντας επαναληπτικά τα βάρη και τις προκαταλήψεις ώστε να προσαρμόζονται με μεγαλύτερη ακρίβεια στα δεδομένα.




Σχήμα 3.4.2: Απεικόνιση υποπροσαρμογής και υπερπροσαρμογής σε μοντέλο ταξινόμησης.

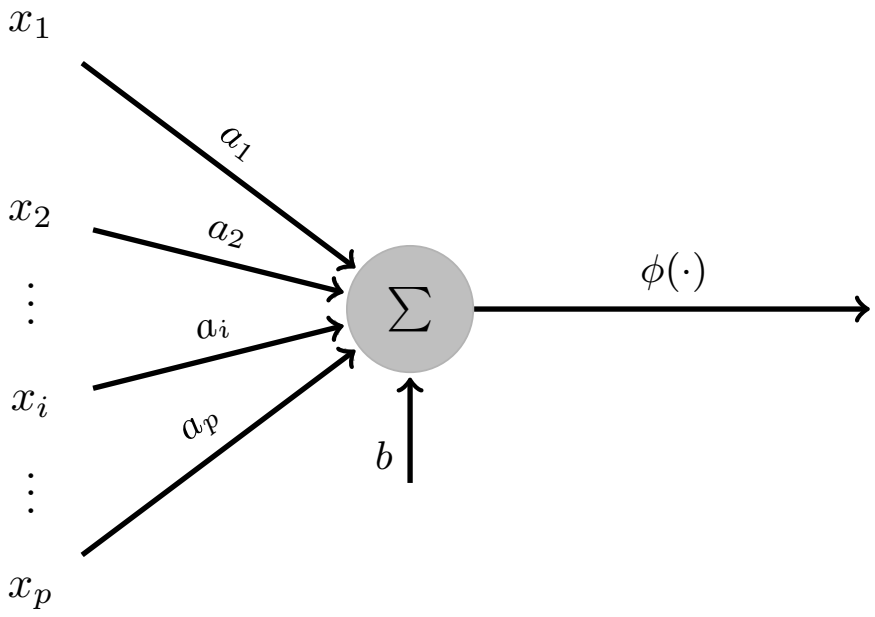
## 3.5 Βαθιά Μάθηση

Η διάκριση μεταξύ της βαθιάς μάθησης και της μηχανικής μάθησης είναι αναμφισβήτητα θολή. Τα μοντέλα βαθιάς μάθησης είναι γενικά πιο πολύπλοκα και περιλαμβάνουν μεγαλύτερη ποσότητα μαθησιακής συνάρτησης ή σύνθεσης εννοιών από ό,τι η παραδοσιακή Μηχανική Μάθηση. Το επίθετο "βαθύ" αναφέρεται στη χρήση

πολλαπλών επιπέδων σε ένα δίκτυο, αλλά δεν περιορίζεται σε αυτό. Επίσης, η βαθιά μάθηση βασίζεται στα νευρωνικά δίκτυα και τη μάθηση αναπαράστασης. Σε αυτό το υποκεφάλαιο θα συζητήσουμε βασικές έννοιες (που δεν περιορίζονται στη Βαθιά Μάθηση) και επιτυχημένες αρχιτεκτονικές που έχουν αναπτυχθεί.

### 3.5.1 Δίκτυο Single-layer Perceptron

Το *perceptron* αποτελεί την απλούστερή δυνατή μορφή νευρωνικού δικτύου, δέχεται ένα διάνυσμα χαρακτηριστικών $\mathbf{x} \in \mathbb{R}^n$. το οποίο πολλαπλασιάζεται με ένα διάνυσμα βαρών $w$. Οι πολλαπλασιασμένες συντεταγμένες αθροίζονται μαζί με μία σταθερά πόλωσης $b$. Η έξοδος της άθροισης, χρησιμοποιείται ως είσοδο στην συνάρτηση ενεργοποίησης (*activation function*) του perceptron, $\phi(.)$ παράγοντας την έξοδο του, $y$. Το *perceptron* δίνεται στο Σχήμα 3.5.1. Η εξίσωση 3.5.1 περιγράφει ένα υπερεπίπεδο διαχωρισμού. Ωστόσο για την εύρεση του υπερεπιπέδου δεν χρησιμοποιείται το περιθώριο. Αντίθετα, το υπερεπίπεδο προκύπτει από τις τελικές τιμές των παραμέτρων του *perceptron*, δηλαδή τα βάρη και την πόλωσή του, ύστερα από εφαρμογή ενός αλγορίθμου βελτιστοποίησης της *loss function*. Συνεπώς το perceptron λύνει το πρόβλημα ταξινόμησης προτύπων δύο κλάσεων, όταν τα πρότυπα είναι γραμμικά διαχωρίσιμα. Όσο αναφορά την *activation function* υπάρχουν πολλές δυνατές επιλογές.




Σχήμα 3.5.1: Σχηματική αναπαράσταση του *perceptron*

Αποτελείται από τις εισόδους, ένα n-διάστατο διάνυσμα $x$, το οποίο στη συνέχεια πολλαπλασιάζεται (εσωτερικό γινόμενο) με ένα n-διάστατο διάνυσμα $w$, το διάνυσμα των βαρών, ενώ προσθέτουμε στο άθροισμα αυτών την πόλωση (bias) $b$. Η τελική έξοδος του νευρώνα προκύπτει περνώντας το προηγούμενο αποτέλεσμα μέσα από μια συνάρτηση ενεργοποίησης $\phi$. Συνολικά, η έξοδος του νευρωνικού δίνεται από την σχέση

$$y = \phi\left(\sum_{i=1}^{n} \mathbf{x}_i \times \mathbf{w_i} + b\right) \tag{3.5.1}$$

### 3.5.2 Βαθιές Αρχιτεκτονικές

Όπως αναφέρθηκε, τα περισσότερα πραγματικά προβλήματα ταξινόμησης δεν έχουν γραμμικά διαχωρίσιμα πρότυπα ή διαθέτουν περισσότερες από δύο δυνατές κλάσεις ταξινόμησης. Συνεπώς το *perceptron* από

μόνο του δεν βρίσκει ιδιαίτερη εφαρμογή. Τέτοια προβλήματα μπορούν να προσεγγιστούν μέσω βαθιών αρχιτεκτονικών.

Οι βαθιές αρχιτεκτονικές χρησιμοποιούν ως στοιχειώδη μονάδα το *perceptron*. Συνήθως διαθέτουν ένα επίπεδο εισόδου το διανύσματος $\mathbf{x}$, κάποια κρυφά επίπεδα νευρώνων, και ένα επίπεδο εξόδου. Τα κρυφά επίπεδα αποτελούνται από perceptrons που επικοινωνούν μεταξύ τους μέσω συνάψεων - βάρη. Τόσο ο αριθμός των κρυφών επιπέδων όσο και οι συνάψεις μεταξύ νευρώνων μπορεί να διαφέρουν από αρχιτεκτονική σε αρχιτεκτονική. Το επίπεδο εξόδου επικοινωνεί με το τελευταίο κρυφό επίπεδο. Στη γενική περίπτωση αν ο αριθμός δυνατών κλάσεων του προβλήματος, $n > 2$ η αρχιτεκτονική διαθέτει $n$ εξόδους (στην περίπτωση των δύο κλάσεων αρκεί μία δυαδική έξοδος).

Το δίκτυο λαμβάνει ένα διάνυσμα εισόδου $\mathbf{x} \in \mathbb{R}^n$, το κρυφό επίπεδο μετασχηματίζει το διάνυσμα εισόδου στο διάνυσμα $\mathbf{h} \in \mathbb{R}^l$. Το επίπεδο εξόδου έχει $k$ εξόδους δηλαδή $\mathbf{y} \in \mathbb{R}^k$. Στο παράδειγμα αυτό το δίκτυο είναι πλήρως συνδεδεμένο, δηλαδή υπάρχει σύναψη μεταξύ κάθε νευρώνα του προηγούμενου και του επόμενου επιπέδου.

### Back Propagation

Η εκπαίδευση των νευρωνικών δικτύων γίνεται με τον αλγόριθμο gradient decent. Για να δουλέψει ο αλγόριθμος, χρειάζεται σε κάθε επανάληψη να υπολογίσουμε την μερική παράγωγο της συνάρτησης σφάλματος ως προς όλες τις παραμέτρους όλων των επιπέδων του δικτύου. Αυτό μπορεί να επιτευχθεί αποτελεσματικά με τον αλγόριθμο . Αρχικά βρίσκεται η παράγωγος της συνάρτησης σφάλματος ως προς τις παραμέτρους του τελευταίου επιπέδου με άμεσο τρόπο. Στη συνέχεια, χρησιμοποιώντας τον κανόνα της αλυσίδας $\frac{df}{dx} = \frac{df}{dy}\frac{dy}{dx}$ μπορούμε να βρούμε τις παραγώγους των παραμέτρων του αμέσως προηγούμενου επιπέδου επαναχρησιμοποιώντας τις μερικές παραγώγους που έχουμε ήδη υπολογίσει και τις γνωστές παραγώγους των συναρτήσεων ενεργοποίησης και των αφινικών μετασχηματισμών. Συνεχίζουμε αυτή την διαδικασία μέχρι το πρώτο επίπεδο και στο τέλος ανανεώνουμε τα βάρη με κατάλληλο τρόπο ώστε να μειωθεί το σφάλμα. Η όλη διαδικασία επαναλαμβάνεται μέχρι να βρεθούμε σε ένα ζητούμενο τοπικό ελάχιστο, ολοκληρώνοντας την εκπαίδευση.

### Κανονικοποίηση (Regularization)

Όπως αναφέρθηκε και προηγουμένως, η αφελής ελαχιστοποίησης την συνάρτησης σφάλματος, μπορεί να οδηγήσει στο φαινόμενο του  δηλαδή ένα δίκτυο που απλά απομνημονεύει τα δεδομένα εισόδου και έχει κακή απόδοση στο test set. Το ίδιο πρόβλημα υπάρχει όταν το μοντέλο μαθαίνει χαρακτηριστικά που υπάρχουν στο αλλά είναι προϊόν τυχαίου θορύβου των συγκεκριμένων δεδομένων και δεν γενικεύει σε δεδομένα που δεν έχουν συναντηθεί στην εκπαίδευση. το *(Regularization)* είναι η διαδικασία η οποία έχει στόχο την βελτίωση της ικανότητας γενίκευσης των μοντέλων και γίνεται με διάφορους τρόπους.

Ένας συχνός τρόπος *Regularization* είναι να προσθέσουμε στην συνάρτηση σφάλματος έναν επιπλέον όρο ως εξής:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i, \theta_n)) + \lambda R(\theta) \tag{3.5.2}$$

όπου το $\lambda$ είναι υπερπαράμετρος.

Οι όροι κανονικοποίσης αποσκοπούν στο να μειωθεί η περιπλοκότητα του μοντέλου και πολύ συχνά κωδικοποιούν μια εκ των προτέρων γνώση (*prior*) που έχουμε για το σύστημα. Συχνές επιλογές είναι οι:

- L2 norm:
$$R(\theta) = \frac{1}{2}\lambda||\theta||_2^2 = \frac{1}{2}\lambda w^T w \tag{3.5.3}$$
όπου $W$ το διάνυσμα βαρών του νευρωνικού δικτύου.

Η L2 νόρμα, περιορίζει τα βάρη του δικτύου ώστε να μην γίνουν πολύ μεγάλα με τέτοιο τρόπο ώστε τα μεγαλύτερα βάρη να τιμωρούνται περισσότερο από τα μικρά και είναι ισοδύναμη με το να μειώνουμε τα βάρη κατά ένα ποσοστό σε κάθε επανάληψη (*weight decay*).

- L1 norm:

$$R(\theta) = \lambda||\theta||_1 \tag{3.5.4}$$

Η L1 νόρμα, περιορίζει όλα τα βάρη σε μικρότερες τιμές ανεξαρτήτως από το μέγεθος τους και σαν αποτέλεσμα οδηγεί σε μικρότερη πυκνότητα στο δίκτυο, κάτι που αναφέρεται ως μεγαλύτερο *sparsity*. Αυτός ο τρόπος *(Regularization)* λέγεται και lasso**[lasso]**.

Μια διαφορετική προσέγγιση που είναι πολύ δημοφιλής είναι το *dropout*[18], το οποίο σε κάθε επανάληψη μηδενίζει κάθε βάρος ενός δικτύου με κάποια δεδομένη πιθανότητα. Όταν το μοντέλο αξιολογείται, χρησιμοποιούνται όλα τα βάρη κανονικά. Η βασική ιδέα είναι ότι εκπαιδεύουμε πολλά (συγκεκριμένα εκθετικά πολλά) μικρότερα νευρωνικά δίκτυα με κοινούς παραμέτρους (τα οποία προκύπτουν με την αφαίρεση συνδέσεων μεταξύ επιπέδων) και στο τέλος βρίσκουμε τον μέσο όρο όλων των προβλέψεων.

Για να αξιολογήσουμε την απόδοση ενός μοντέλου, με τρόπο που δεν είναι ευαίσθητος στο overfitting, διαλέγουμε ένα υποσύνολο των δεδομένων (σύνολο επαλήθευσης ή *Validation set*) τα οποία δεν χρησιμοποιούμε καθόλου κατά την διάρκεια της εκπαίδευσης. Το σύνολο επαλήθευσης χρησιμοποιείται επίσης για την επιλογή των υπερπαραμέτρων της εκπαίδευσης.

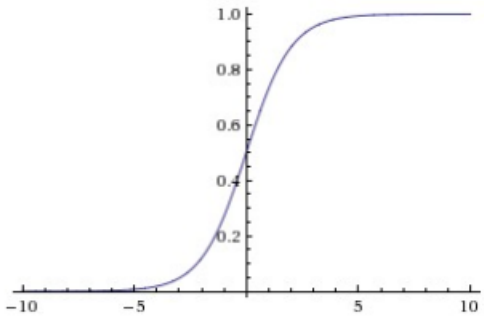
### 3.5.3 Συναρτήσεις Ενεργοποίησης

Ο ρόλος των συναρτήσεων ενεργοποίησης είναι να προσθέτουν μια μη-γραμμικότητα στο δίκτυο και είναι απαραίτητος για να μπορεί το νευρωνικό δίκτυο να προσεγγίσει περίπλοκες συναρτήσεις.

Οι συνηθέστερες επιλογές για την συνάρτηση ενεργοποίησης είναι:

- Σιγμοειδής συνάρτηση (sigmoid)

Η σιγμοειδής συνάρτηση στα μαθηματικά, είναι οποιαδήποτε συνάρτηση η οποία έχει σχήμα όπως το 3.5.2 αλλά στο πλαίσιο της μηχανικής μάθησης αναφέρεται στην λογιστική συνάρτηση η οποία ορίστηκε στην εξίσωση **??** και είναι ιστορικά από τις πρώτες συναρτήσεις που χρησιμοποιήθηκαν στην μηχανική μάθηση.




Σχήμα 3.5.2: Η σιγμοειδής συνάρτηση ενεργοποίησης.

Το πλεονέκτημα της συνάρτησης αυτής είναι ότι η παράγωγός της είναι $\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x)$ και σαν συνέπεια μπορεί να υπολογιστεί σχετικά γρήγορα κατά το *back propagation* χρησιμοποιώντας τις τιμές της συνάρτησης. Επειδή όμως η παράγωγός της είναι πάντα μικρότερη από το 1, όταν χρησιμοποιείται σε νευρωνικά δίκτυα με πολλά επίπεδα, οι παράγωγοι της συνάρτησης σφάλματος ως προς τις παραμέτρους των πρώτων επιπέδων έχει γίνει σχεδόν μηδέν με αποτέλεσμα να μην μπορούν να ανανεωθούν με τέτοιο τρόπο ώστε να μειωθεί το σφάλμα. Αυτό συμβαίνει επειδή σύμφωνα με τον κανόνα της αλυσίδας, όταν παραγωγίζουμε τις παραμέτρους ενός επιπέδου ως προς το επόμενο, κάθε παράγωγος πολλαπλασιάζεται

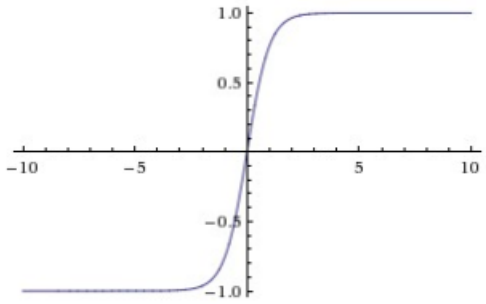
με έναν αριθμό μικρότερο του ενός και η παράγωγος των πρώτων επιπέδων μειώνεται εκθετικά ως προς τον αριθμό των επιπέδων. Το πρόβλημα αυτό ονομάζεται vanishing gradient problem. Για αυτό το λόγο, στα βαθιά νευρωνικά δίκτυα, η σιγμοειδής συνάρτηση χρησιμοποιείται σχεδόν αποκλειστικά όταν θέλουμε να περιορίσουμε έναν αριθμό στο διάστημα $[0, 1]$, όπως για παράδειγμα στην δυαδική ταξινόμηση.

- Υπερβολική εφαπτομένη (tanh)

  Αυτή η συνάρτηση φαίνεται στο σχήμα 3.5.3 και ορίζεται ως:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3.5.5}$$

  Σε αυτή την περίπτωση, η παράγωγος της συνάρτησης είναι $\frac{d}{dx}\tanh(x) = 1 - \tanh^2(x)$ οπότε έχουμε το ίδιο πλεονέκτημα και μειονέκτημα που είχαμε με την λογιστική συνάρτηση. Η διαφορά είναι ότι η έξοδος της υπερβολικής εφαπτομένης είναι στο $[-1, 1]$ αντί στο $[0, 1]$ και χρησιμοποιείται αναλόγως.

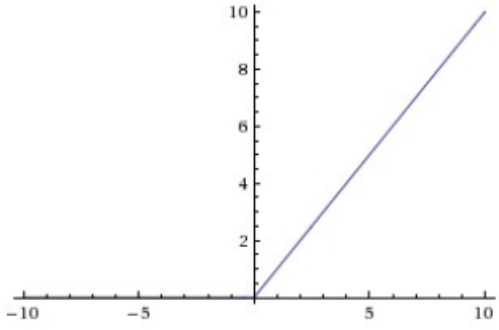



Σχήμα 3.5.3: Η συνάρτηση ενεργοποίησης υπερβολικής εφαπτομένης *tanh*.

- ReLU (Rectified Linear Unit)

  Αυτή η συνάρτηση φαίνεται στο σχήμα 3.5.4 και ορίζεται ως:

$$f(x) = \max(x, 0) \tag{3.5.6}$$




Σχήμα 3.5.4: Η συνάρτηση ενεργοποίησης ReLU.

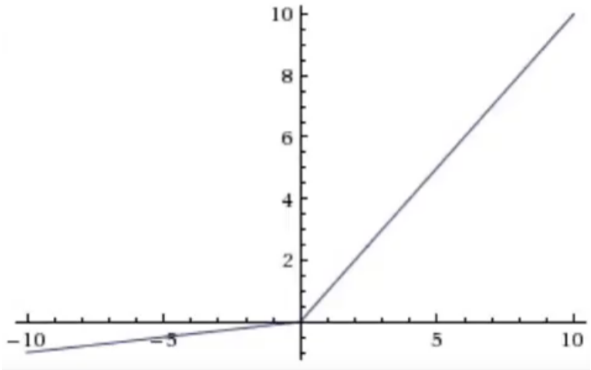
Τα πλεονέκτηματα της συνάρτησης *ReLU* είναι ότι είναι πάρα πολύ γρήγορη να υλοποιηθεί, αφού ουσιαστικά αποτελείται μόνο από μια σύγκριση, και οτι αποφεύγει το vanishing gradient πρόβλημα.

Αυτό εξηγείται από το ότι η παράγωγος της *ReLU* είναι ή 1, ή 0, ανάλογα με το πρόσημο της εισόδου και δεν μειώνεται η πληροφορία της παραγώγου κατά την διάρκεια της παραγώγισης ως προς τα πρώτα επίπεδα. Το μειονέκτημα που έχει η *ReLU* είναι ότι αν κάποιος νευρώνας δέχεται μόνο αρνητικές εισόδους, η έξοδος είναι πάντα μηδέν, και επειδή η παράγωγος είναι 0 στις αρνητικές τιμές, δεν υπάρχει τρόπος να αξιοποιηθεί η πληροφορία της εισόδου σε κανένα στάδιο της εκπαίδευσης. Σε αυτή την περίπτωση, η *ReLU* έχει κολλήσει και λέμε ότι είναι νεκρή.

- Leaky ReLU

  Αυτή η συνάρτηση φαίνεται στο σχήμα 3.5.5 και ορίζεται ως:

$$f(x) = \begin{cases} \alpha x, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{3.5.7}$$



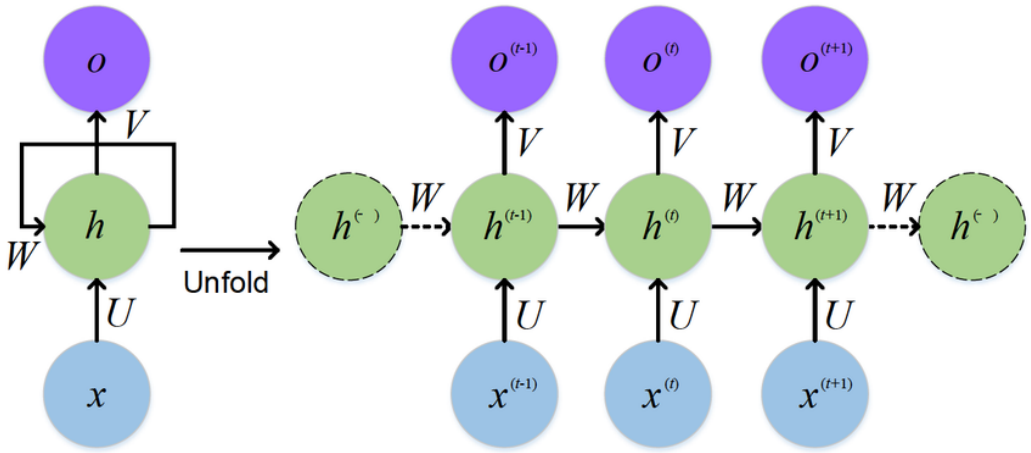

Σχήμα 3.5.5: Η συνάρτηση ενεργοποίησης leaky ReLU.

όπου το $\alpha$ είναι μια μικρή παράμετρος (π.χ. $\alpha = 0.01$). Η συμπεριφορά αυτής της συνάρτησης είναι πολύ παρόμοια με αυτήν της *ReLU*, με την διαφορά ότι αν η εκπαίδευση κολλήσει στην αρνητική περιοχή, υπάρχει πιθανότητα να ξεκολλήσει γιατί τώρα η παράγωγος στις αρνητικές τιμές δεν είναι 0.

### 3.5.4 Αναδρομικά Νευρωνικά Δίκτυα (RNN)

Τα Αναδρομικά Νευρωνικά Δίκτυα (*Reccurent Neural Networks - RNNs*) ανήκουν σε ένα κλάδο νευρωνικών δικτύων ικανά να διαχειρίζονται ακολουθιακά δεδομένα της μορφής $x^{(1)}, x^{(2)}, \ldots x^{(\tau)}$ . Στα κλασικά νευρωνικά δίκτυα γίνεται η υπόθεση πως τα δεδομένα εισόδου είναι ανεξάρτητα μεταξύ τους. Ωστόσο σε πολλά προβλήματα ταξινόμησης αυτή η υπόθεση είναι λάθος. Σε προβλήματα αναγνώρισης ή μετάφρασης προτάσεων, η ακριβής αντιστοιχία μιας συγκεκριμένης λέξης είναι ευκολότερη δεδομένου των προηγούμενων αποτελεσμάτων. Υπό αυτό το πρίσμα τα RNNs διαθέτουν ένα είδος "μνήμης", μπορούν να διατηρήσουν την πληροφορία πού έχει επεξεργαστεί μέσα στο δίκτυο μέχρι την τωρινή χρονική στιγμή και έτσι να διαχειριστούν δεδομένα εισόδου που διαφέρουν ως προς το μήκος τους. Με τον όρο χρονική στιγμή (*time step*) συνήθως αναφέρεται το δείγμα της ακολουθίας που εξετάζεται από το δίκτυο. Το δείγμα αυτό διαφέρει ανάλογα με το πρόβλημα ταξινόμησης. Σε προβλήματα Επεξεργασίας Φωνής το δείγμα μπορεί να περιέχει μια λέξη από μια πρόταση ενώ σε προβλήματα Όρασης κάποιο *pixel* της εικόνας. Στην αυτόματη αναγνώριση ομιλίας το δείγμα μπορεί να αφορά κάποιο τμήμα της ομιλίας, όπως ένα πλαίσιο ή μια ομάδα πλαισίων.

Το βασικό δομικό στοιχείο ενός αναδρομικού νευρωνικού δικτύου είναι οι νευρώνες (μονάδες ή κόμβοι του συστήματος). Οι νευρώνες συνδέονται μεταξύ τους με συναπτικούς δεσμούς οι οποίοι περιγράφονται από το βάρος τους. Διαχωρίζονται συνήθως σε κόμβους εισόδου, εσωτερικούς κόμβους και εξόδου. Κάθε χρονική στιγμή οι κόμβοι έχουν μια τιμή ενεργοποίησης, η οποία αναπαρίσταται με $u(n)$ για τους κόμβους εισόδου, $x(n)$ για τους εσωτερικούς και $y(n)$ για τους κόμβους εξόδου. Πολλές φορές αναφερόμαστε σε όλους με την ονομασία $x(n)$.



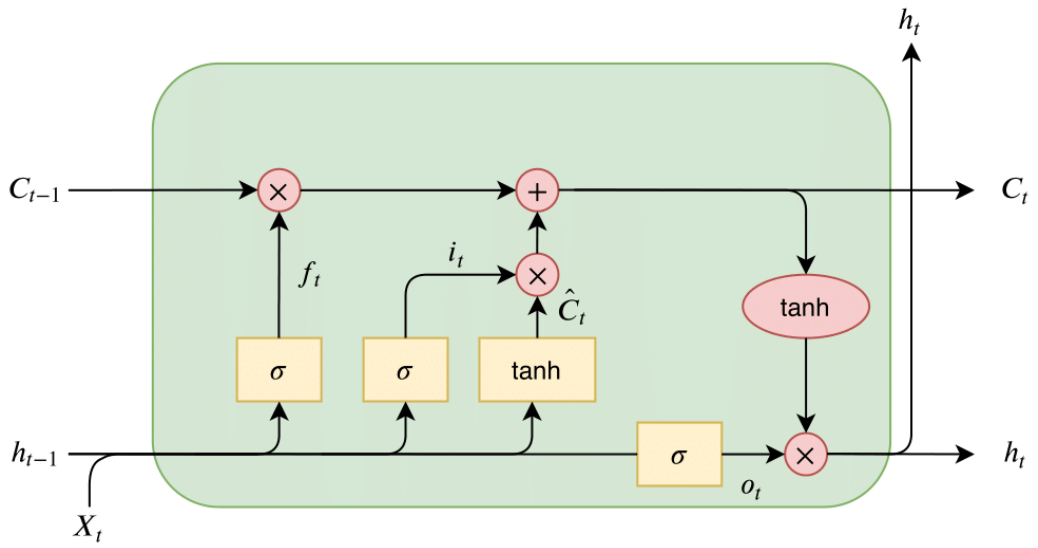

Σχήμα 3.5.6: Το τυπικό και το αναδιπλωμένο RNN.

Τα αναδρομικά νευρωνικά δίκτυα κατηγοριοποιούνται σε πλήρως αναδρομικά δίκτυα, στα οποία επιτρέπεται η σύνδεση όλων των νευρώνων μεταξύ τους και στα τοπικά αναδρομικά δίκτυα. Στα πλήρως αναδρομικά δίκτυα δεν υπάρχει διάκριση στους κόμβους εισόδου και κάθε κόμβος μπορεί να αποτελέσει είσοδο για οποιονδήποτε άλλο κόμβο, συμπεριλαμβανομένου και του εαυτού του. Τα δεύτερα περιέχουν, συνήθως, μόνο αναδρομικές συνδέσεις μεταξύ των νευρώνων που βρίσκονται ίδιο επίπεδο και η διαδικασία της διάδοσης προς τα εμπρός κατά την εκμάθηση μοιάζει πολύ με αυτή των feedforward δικτύων. Συγκριτικά με τα τοπικά αναδρομικά δίκτυα, τα πλήρως αναδρομικά πάσχουν από θέματα ευστάθειας κατά το training και απαιτούν τη χρήση περίπλοκων και χρονοβόρων αλγορίθμων εκμάθησης. Τα τοπικά αναδρομικά δίκτυα έχουν, αντιθέτως, πιο απλή δομή που τα κάνει πιο αποδοτικά κατά τη διαδικασία εκμάθησης και παρέχουν τη δυνατότητα για έλεγχο της ευστάθειας στους εσωτερικούς τους κόμβους. Μια ακόμα διάκριση των RNN που χρησιμοποιούνται σε εφαρμογές διακριτού χρόνου είναι σε αναδρομικά δίκτυα με χρονοκαθυστέρηση και σύγχρονα αναδρομικά δίκτυα. Τα δίκτυα που λειτουργούν με χρονοκαθυστέρηση εκπαιδεύονται με στόχο τη μείωση του σφάλματος πρόβλεψης, ενώ τα σύγχρονα δίκτυα δε στοχεύουν στην ιδιότητα του να έχουν μνήμη ή καλύτερη πρόβλεψη όσο εκπαιδεύεται το δίκτυο, αλλά κάνουν χρήση των αναδράσεων με στόχο να προσφέρουν καλύτερη δυνατότητα προσέγγισης συναρτήσεων σύμφωνα με τις αρχές της θεωρίας του Turing και της περιπλοκότητας. Σε αυτό τον τομέα έχει αποδειχθεί ότι έχουν πολύ ισχυρές ικανότητες και έχει φανεί πειραματικά ότι μπορούν να "μάθουν" οποιαδήποτε συνάρτηση πηγάζει από κάποιο MLP (Multi Layer Perceptron), χωρίς όμως να ισχύει το αντίστροφο.

### 3.5.5 Δίκτυα Μακράς Βραχύχρονης Μνήμης (LSTM)

Τα δίκτυα με ανατροφόδοτηση (recurrent networks) χρησιμοποιούν τις συνδέσεις ανάδρασης προκειμένου να αποθηκεύσουν πληροφορία των πρόσφατων γεγονότων εισόδου ως αποτέλεσμα της συνάρτησης ενεργοποίησης. Αυτό το χαρακτηριστικό τους δίνει τη δυνατότητα μιας προσωρινής μνήμης (short-term memory). Παρότι τα δίκτυα αυτού του τύπου είναι αποτελεσματικά για αρκετές εφαρμογές (πχ. αναγνώριση φωνής) παρουσιάζουν αδυναμίες σε περιπτώσεις που υπάρχει υπολογίσιμο χρονικό διάστημα μεταξύ της εισόδου και της αναμενόμενης εξόδου (time lag). Τα δίκτυα "Long Short-Term Memory" ή κοινώς LSTM είναι recurrent networks που έχουν σχεδιαστεί ώστε να ξεπεράσουν το πρόβλημα της γεφύρωσης των γεγονότων εισόδου χωρίς να έχουν απώλειες λόγω χρονικών κενών [19]. Αυτό επιτυγχάνεται μέσω ενός αποδοτικού αλγορίθμου που είναι δομημένος σε επίπεδα.

Αποτελούνται από κελιά μνήμης (cell), πύλες εισόδου, πύλες εξόδου και πύλες επιλεκτικής συγκράτησης (forget gate). Τα κελιά μνήμης επιλέγουν ποιες πληροφορίες θα αποθηκεύσουν με το να συνδυάζουν την προηγούμενη κατάσταση, την τρέχουσα μνήμη και την είσοδο.




Σχήμα 3.5.7: Η αρχιτεκτονική ενός κελιού LSTM.

Οι τρεις πύλες, forget, input και output, φαίνονται στο Σχήμα 3.5.7 ως $f_t$, $i_t$ και $o_t$, αντίστοιχα. Οι πύλες έχουν μια απλή διαίσθηση πίσω τους:

Η πύλη forget λέει στο κελί ποιες πληροφορίες πρέπει να "ξεχάσει" ή να απορρίψει από την εσωτερική κατάσταση του κελιού. Η πύλη εισόδου λέει στο κελί ποιες νέες πληροφορίες να αποθηκεύσει στην εσωτερική κατάσταση του κελιού. Η πύλη εξόδου είναι τότε αυτό που το κελί εξάγει, δηλαδή μια φιλτραρισμένη έκδοση της εσωτερικής κατάστασης του κελιού.

Όπως και σε κάθε νευρωνικό δίκτυο, τα βάρη και οι προκαταλήψεις συνδέονται με κάθε πύλη. Αυτοί οι πίνακες βαρών χρησιμοποιούνται σε συνδυασμό με βελτιστοποίηση βάσει κλίσης για να μάθει το κύτταρο LSTM. Στη συνέχεια, αυτά τα κύτταρα αλυσιδωτά συνδέονται μεταξύ τους- αυτό είναι που επιτρέπει στο δίκτυο RNN-LSTM να διατηρεί πληροφορίες από προηγούμενα χρονικά βήματα και να κάνει προβλέψεις χρονοσειρών. Με τη χρήση της αρχιτεκτονικής των κυττάρων LSTM, το δίκτυο έχει έναν τρόπο να εξαλείψει το πρόβλημα της εξαφανιζόμενης κλίσης. Αυτό το πρόβλημα εμπόδιζε τις παλαιότερες αρχιτεκτονικές RNN να επιτύχουν εξαιρετικές προβλέψεις χρονοσειρών.

Η παραπάνω αρχιτεκτονική είναι μια σχετικά τυπική εκδοχή μιας κυψέλης LSTM- υπάρχουν όμως πολλές παραλλαγές και οι ερευνητές βελτιώνουν και τροποποιούν συνεχώς την αρχιτεκτονική της κυψέλης για να κάνουν το δίκτυο LSTM να αποδίδει καλύτερα και πιο ισχυρά για διάφορες εργασίες. Μια άλλη διαδεδομένη τροποποίηση του LSTM είναι η λεγόμενη gated recurrent unit ή GRU [20]. Η κύρια διαφορά μεταξύ της GRU και του LSTM είναι ότι η GRU συγχωνεύει τις πύλες εισόδου και forget σε μια ενιαία πύλη ενημέρωσης. Επιπλέον, συνδυάζει την εσωτερική κατάσταση των κυττάρων και την κρυφή κατάσταση. Το προκύπτον κελί GRU είναι, επομένως, ελαφρώς πιο απλό από το παραδοσιακό LSTM. Στην βιβλιογραφία υπάρχουν ακόμη πολλές παραλλαγές του δικτύου LSTM [21].

## 3.6 Αρχιτεκτονική Ακολουθίας-προς-Ακολουθία

Η ανάλυση ακολουθίας προς ακολουθία Sequence-to-sequence (seq2seq) είναι μια οικογένεια προσεγγίσεων μηχανικής μάθησης που χρησιμοποιούνται στην επεξεργασία ομιλίας, γλώσσας και σε άλλους τομείς. Ασχολείται με το έργο της δημιουργίας μιας ακολουθίας εξόδου δεδομένης μιας ακολουθίας εισόδου, όταν οι δύο ακολουθίες έχουν διαφορετικά μήκη και όχι ρητή αντιστοιχία ένα προς ένα. Για παράδειγμα, μια πολύ τυπική εργασία που επιλύεται με τη χρήση μιας προσέγγισης seq2seq, είναι η νευρωνική μηχανική μετάφραση (NMT). Άλλες εργασίες που επιλύονται με παρόμοιο τρόπο είναι η περίληψη εγγράφων, τα μοντέλα συνομιλίας και η λεζάντα εικόνας, για να αναφέρουμε μερικές.

Δεδομένου ότι οι δύο ακολουθίες δεν έχουν την ίδια δομή, η χρήση απλώς μιας ακολουθιακής αρχιτεκτονικής, όπως το RNN που περιγράψαμε παραπάνω, δεν είναι επαρκής. Για να αντιμετωπιστεί αυτό, εισήχθη η αρχιτεκτονική seq2seq. Η seq2seq βασίζεται σε ένα πλαίσιο κωδικοποιητή-αποκωδικοποιητή. Ο κωδικοποιητής κωδικοποιεί την ακολουθία εισόδου σε μια κρυφή (πλαισιωμένη) αναπαράσταση, ενώ ο αποκωδικοποιητής λαμβάνει αυτή την αναπαράσταση ως είσοδο για να παράγει την τελική έξοδο. Στην αρχική υλοποίηση, ως κωδικοποιητής και αποκωδικοποιητής χρησιμοποιούνται RNNs (συγκεκριμένα βαθιά LSTMs), ενώ η κωδικοποιημένη αναπαράσταση είναι ένα διάνυσμα σταθερής διάστασης.

Αν συμβολίσουμε την ακολουθία εισόδου ως $x_1, \ldots, x_n$, την ακολουθία εξόδου ως $y_1, \ldots, y_m$ και το διάνυσμα σταθερού μεγέθους ως $c$ (context vector), μπορούμε να εκφράσουμε τυπικά το έργο της δημιουργίας της ακολουθίας εξόδου δεδομένης της εισόδου, με την υπό συνθήκη πιθανότητα:

$$P(y_1, \ldots, y_m \mid x_1, \ldots, x_n) = \prod_{i=1}^{m} P(y_i \mid c, y_1, \ldots, y_{i-1}) \tag{3.6.1}$$

Αναλυτικότερα, ο κωδικοποιητής RNN επεξεργάζεται την ακολουθία εισόδου ένα token κάθε φορά και η τελική κρυφή κατάσταση χρησιμοποιείται ως διάνυσμα πλαισίου $c$. Στη συνέχεια, η κρυφή κατάσταση του αποκωδικοποιητή RNN αρχικοποιείται με το $c$. Η πρώτη είσοδος στον αποκωδικοποιητή είναι ένα ειδικό token που ονομάζεται <bos> (αρχή της ακολουθίας) και οι επόμενες είσοδοι είναι οι έξοδοι του προηγούμενου χρονικού βήματος. Μια άλλη τεχνική που χρησιμοποιείται κατά τη διάρκεια της εκπαίδευσης, η οποία ονομάζεται **teacher forcing**, χρησιμοποιεί τα tokens της ακολουθίας-στόχου ως είσοδο αντί της προβλεπόμενης εξόδου και συνήθως οδηγεί σε ταχύτερη σύγκλιση. Κατά τη διάρκεια της εξαγωγής συμπερασμάτων (ή του χρόνου δοκιμής) χρησιμοποιείται η πρώτη τεχνική, καθώς δεν υπάρχει βασική αλήθεια.

Αν και η προσέγγιση αυτή είναι πολύ χρήσιμη κατά την επίλυση εργασιών χωρίς υποθέσεις σχετικά με τη δομή της ακολουθίας, έχει ένα σημαντικό μειονέκτημα. Η είσοδος μπορεί να προσπελαστεί μόνο μέσω του διανύσματος πλαισίου. Η κωδικοποίηση μιας ακολουθίας αυθαίρετου μήκους σε μια αναπαράσταση σταθερού μεγέθους δημιουργεί μια συμφόρηση πληροφοριών. Επίσης, όταν έχουμε να κάνουμε με μεγάλες ακολουθίες, οι πληροφορίες στην αρχή μπορεί να έχουν ξεχαστεί σημαντικά μέχρι τη στιγμή της επεξεργασίας του τελευταίου συμβόλου.

## 3.7 Ο Μηχανισμός Προσοχής

Ο άνθρωπος συνήθως σε προβλήματα επεξεργασίας ακολουθιακών δεδομένων δίνει περαιτέρω σημασία στο παρόν γεγονός, παρά σε προηγούμενα ή μελλοντικά. Ενδεικτικά κατά την μετάφραση ενός κειμένου δίνεται κυρίως έμφαση στην επί του παρόντος μεταφραζόμενη λέξη. Στα νευρωνικά δίκτυα η ικανότητα αυτή επιτυγχάνεται μέσω του μηχανισμού προσοχής *Attention*. Στην πράξη ο μηχανισμός αυτός αντιστοιχεί ως ένα επιπλέον επίπεδο νευρώνων στο δίκτυο και αποδίδει συντελεστές - βάρη στις εξόδους που παράγονται από ένα ακολουθιακό μοντέλο, εστιάζοντας σε συγκεκριμένες από αυτές.

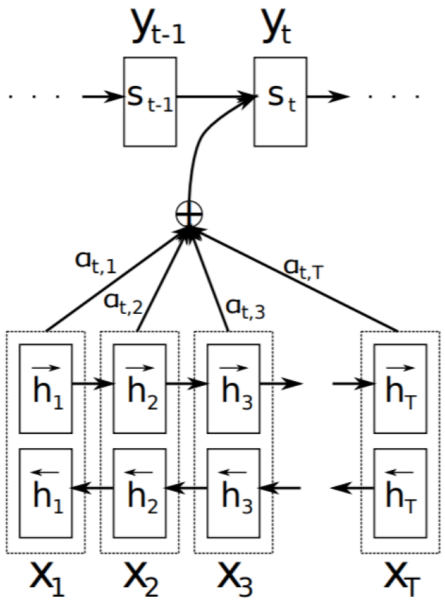
Στο Σχήμα 3.7.1 δίνεται η σχηματική αναπαράσταση του μηχανισμού. Αν $h_t$ είναι η έξοδος του μοντέλου σε κάθε *time step*, τότε υπολογίζεται ένα διάνυσμα συμφραζομένων (*"context" vector"*), $\mathbf{c}$. Το διάνυσμα αυτό αντιστοιχεί στον σταθμισμένο μέσο όρο των εξόδων $h_t$, $t = 1, 2, \ldots, T$ :

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t h_t \tag{3.7.1}$$

Τα βάρη $a_t$ υπολογίζονται μέσω των :

$$e_t = f(h_t) \tag{3.7.2a}$$

$$a_t = softmax(e_t) = \frac{exp(e_t)}{\sum_{t=1}^{T} exp(e_k)} \tag{3.7.2b}$$



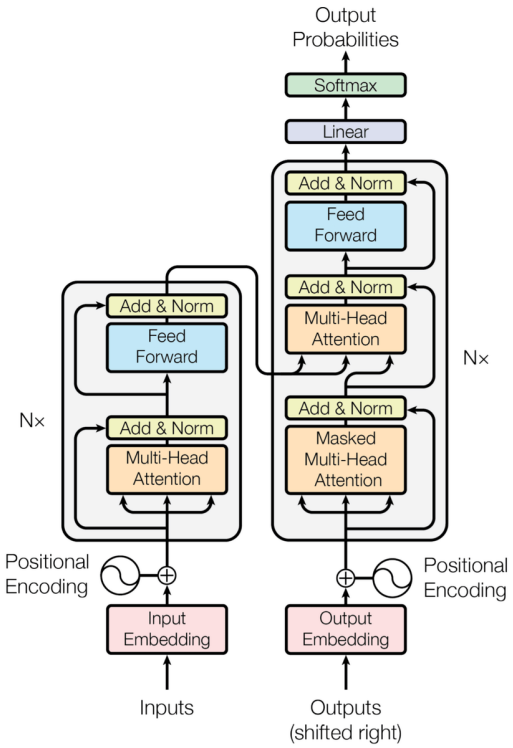

Σχήμα 3.7.1: Σχηματική αναπαράσταση του μηχανισμού Attention.

Όπως φαίνεται και στην εξίσωση 3.7.1 ο μηχανισμός *Attention* επιστρέφει ένα διάνυσμα διάστασης ίσης με το μήκος της ακολουθίας. Τα βάρη $a_t \in [0, 1]$ προσδιορίζονται με εφαρμογή της *softmax* σε κάθε συντεταγμένη $e_t$ της εξόδου της συνάρτησης ενεργοποίησης του επιπέδου $f$. Υψηλές τιμές στα βάρη $a_t$ αντιστοιχούν σε time steps μεγάλης σημασίας. Ουσιαστικά μέσω το μηχανισμού το δίκτυο δίνει έμφαση σε συγκεκριμένες περιοχές στον χώρο των εξόδων.

## 3.8 Αρχιτεχτονική Transformer

Ο Transformer είναι μια αρχιτεκτονική βαθιού νευρωνικού δικτύου που παρέχει μια seq2seq προσέγγιση στο έργο του NMT. Εμπνευσμένος από την επιτυχία του μηχανισμού προσοχής και προσπαθώντας να καταπολεμήσει τις αργές ταχύτητες εκπαίδευσης που επιβάλλει η διαδοχική φύση των αναδρομικών δικτύων, ο Transformer βασίζεται μόνο στην προσοχή και έτσι είναι σε θέση να επεξεργάζεται δεδομένα παράλληλα. Αυτό τον καθιστά πολύ ταχύτερο από οποιαδήποτε ακολουθιακή αρχιτεκτονική, ενώ έχει πολύ καλύτερες επιδόσεις.

Ο Transformer είναι μια αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή. Ο κωδικοποιητής και ο αποκωδικοποιητής αποτελούνται από $N$ στοιβαγμένα επίπεδα, όπου η έξοδος κάθε επιπέδου είναι η είσοδος του επόμενου. Τα στρώματα είναι πανομοιότυπα, με διαφορετικά βάρη. Κάθε στρώμα κωδικοποιητή αποτελείται από ένα υποστρώμα αυτοσυντήρησης και ένα πλήρως συνδεδεμένο υποστρώμα τροφοδότησης που παρεμβάλλονται με μια υπολειμματική σύνδεση και κανονικοποίηση του στρώματος. Ο αποκωδικοποιητής έχει την ίδια δομή, με την προσθήκη ενός υποστρώματος διασταυρούμενης προσοχής μεταξύ της αυτοπροσοχής και της τροφοδότησης. Το υποεπίπεδο αυτοπροσοχής είναι ικανό να δημιουργεί μια αναπαράσταση της ακολουθίας με βάση το πλαίσιο, αναλύοντας την εξάρτηση μεταξύ των tokens της ακολουθίας. Το υποεπίπεδο διασταυρούμενης προσοχής είναι υπεύθυνο για την ανάλυση της εξάρτησης μεταξύ των ακολουθιών εισόδου και εξόδου. Θα πρέπει να σημειωθεί ότι η αυτοπροσοχή στον κωδικοποιητή είναι "αμφίδρομη", ενώ η αυτοπροσοχή του αποκωδικοποιητή είναι "μονόδρομη". Αυτό υλοποιείται με τη χρήση τριγωνικής μάσκας και σκοπός της είναι να αποφεύγεται η προσοχή σε μελλοντικά tokens της ακολουθίας, η οποία έχει δυσμενείς επιπτώσεις κατά την πρόβλεψη του τρέχοντος token. Το υποστρώμα διασταυρούμενης προσοχής κάθε στρώματος αποκωδικοποιητή εξαρτάται από την έξοδο του τελευταίου στρώματος του κωδικοποιητή. Η έξοδος του τελευταίου στρώματος αποκωδικοποιητή μετατρέπεται τελικά σε token πιθανότητας, χρησιμοποιώντας γραμμικό μετασχηματισμό και softmax.




Σχήμα 3.8.1: Η αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή Transformer.

Ο Transformer στο στάδιο εκπαίδευσης επεξεργάζεται τα δεδομένα ως εξής: Η ακολουθία εισόδου μετατρέπεται σε embeddings (με κωδικοποίηση θέσης) και τροφοδοτείται στον κωδικοποιητή. Η στοίβα των κωδικοποιητών το επεξεργάζεται και παράγει μια κωδικοποιημένη αναπαράσταση της ακολουθίας εισόδου. Η ακολουθία-στόχος προστίθεται με ένα σύμβολο έναρξης της πρότασης, μετατρέπεται σε embeddings (με Position Encoding) και τροφοδοτείται στον Decoder. Η στοίβα των αποκωδικοποιητών το επεξεργάζεται αυτό μαζί με την κωδικοποιημένη αναπαράσταση της στοίβας κωδικοποιητών για να παράγει μια κωδικοποιημένη αναπαράσταση της ακολουθίας στόχου. Το στρώμα εξόδου τη μετατρέπει σε πιθανότητες λέξεων και στην τελική ακολουθία εξόδου. Η συνάρτηση απωλειών του μετασχηματιστή συγκρίνει αυτή την ακολουθία εξόδου με την ακολουθία-στόχο από τα δεδομένα εκπαίδευσης. Αυτή η απώλεια χρησιμοποιείται για τη δημιουργία κλίσεων για την εκπαίδευση του μετασχηματιστή κατά την αντίστροφη διάδοση.

Κατά τη διαδικασία inference ο Transformer εκτελεί τα ακόλουθα βήματα: Η ακολουθία εισόδου μετατρέπεται σε embeddings (με κωδικοποίηση Θέσης) και τροφοδοτείται στον κωδικοποιητή. Η στοίβα των κωδικοποιητών την επεξεργάζεται και παράγει μια κωδικοποιημένη αναπαράσταση της ακολουθίας εισόδου. Αντί για την ακολουθία-στόχο, χρησιμοποιούμε μια κενή ακολουθία με μόνο ένα σύμβολο αρχής πρότασης. Αυτό μετατρέπεται σε embeddings (με κωδικοποίηση θέσης) και τροφοδοτείται στον αποκωδικοποιητή. Η στοίβα των αποκωδικοποιητών την επεξεργάζεται μαζί με την κωδικοποιημένη αναπαράσταση της στοίβας των κωδικοποιητών για να παράγει μια κωδικοποιημένη αναπαράσταση της ακολουθίας-στόχου. Το στρώμα εξόδου τη μετατρέπει σε πιθανότητες λέξεων και παράγει μια ακολουθία εξόδου. Παίρνουμε την τελευταία λέξη της ακολουθίας εξόδου ως την προβλεπόμενη λέξη. Αυτή η λέξη συμπληρώνεται τώρα στη δεύτερη θέση της ακολουθίας εισόδου του αποκωδικοποιητή μας, η οποία περιέχει τώρα ένα σύμβολο αρχής πρότασης και την πρώτη λέξη. Όπως και πριν, τροφοδοτείται η νέα ακολουθία αποκωδικοποιητή στο μοντέλο. Στη συνέχεια, η δεύτερη λέξη της εξόδου και προστίθεται στην ακολουθία του αποκωδικοποιητή. Αυτό επαναλαμβάνεται μέχρι να προβλέψει ένα σύμβολο τέλους πρότασης. Σημειώνουμε ότι η ακολουθία του κωδικοποιητή δεν αλλάζει για κάθε επανάληψη.

Οι Transformers είναι πολύ ευέλικτοι και χρησιμοποιούνται για τις περισσότερες εργασίες NLP, όπως γλωσσικά μοντέλα και ταξινόμηση κειμένου. Χρησιμοποιούνται συχνά σε μοντέλα ακολουθίας προς ακολουθία

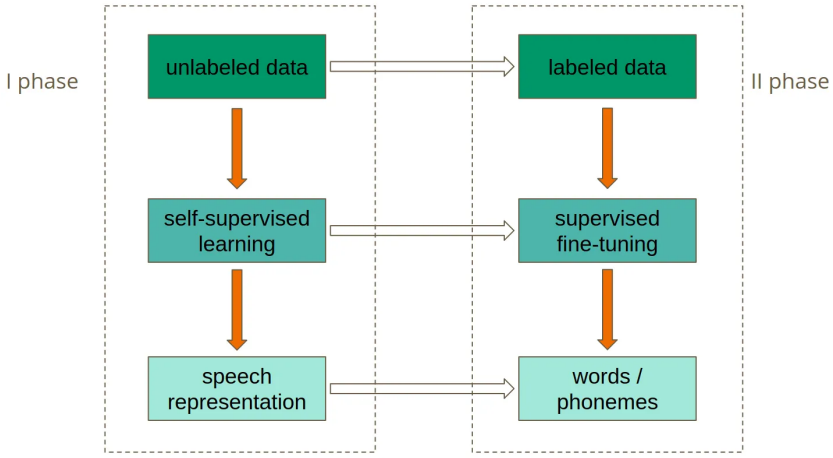
για εφαρμογές όπως η μηχανική μετάφραση, η περίληψη κειμένου, η απάντηση ερωτήσεων, η αναγνώριση ονομαστικών οντοτήτων και η αναγνώριση ομιλίας.

Υπάρχουν διαφορετικές εκδοχές της αρχιτεκτονικής Transformer για διαφορετικά προβλήματα. Το βασικό στρώμα κωδικοποιητή χρησιμοποιείται ως κοινό δομικό στοιχείο για αυτές τις αρχιτεκτονικές, με διαφορετικές "κεφαλές" για συγκεκριμένες εφαρμογές ανάλογα με το πρόβλημα που επιλύεται.

## 3.9 Ο αλγόριθμος wav2vec2.0

Το Wav2Vec 2.0 [22] είναι ένα από τα σύγχρονα μοντέλα αυτόματης αναγνώρισης ομιλίας χάρη στην αυτοεπιβλεπόμενη εκπαίδευση, η οποία είναι μια αρκετά νέα έννοια στον τομέα αυτό. Αυτός ο τρόπος εκπαίδευσης μας επιτρέπει να προ-εκπαιδεύσουμε ένα μοντέλο σε μη επισημασμένα δεδομένα, τα οποία είναι πάντα πιο προσιτά. Στη συνέχεια, το μοντέλο μπορεί να τελειοποιηθεί σε ένα συγκεκριμένο σύνολο δεδομένων για έναν συγκεκριμένο σκοπό. Όπως δείχνουν οι προηγούμενες εργασίες, αυτός ο τρόπος εκπαίδευσης είναι πολύ ισχυρός.

Το μοντέλο εκπαιδεύεται σε δύο φάσεις. Η πρώτη φάση είναι αυτοεπιβλεπόμενη, η οποία γίνεται με τη χρήση μη επισημασμένων δεδομένων και αποσκοπεί στην επίτευξη της καλύτερης δυνατής αναπαράστασης ομιλίας. Μπορείτε να το σκεφτείτε αυτό με παρόμοιο τρόπο όπως σκέφτεστε τις ενσωματώσεις λέξεων. Οι ενσωματώσεις λέξεων στοχεύουν επίσης στην επίτευξη της καλύτερης αναπαράστασης της φυσικής γλώσσας. Η κύρια διαφορά είναι ότι το Wav2Vec 2.0 επεξεργάζεται ήχο αντί για κείμενο. Η δεύτερη φάση της εκπαίδευσης είναι η επιτηρούμενη τελειοποίηση, κατά την οποία χρησιμοποιούνται επισημασμένα δεδομένα για να διδαχθεί το μοντέλο να προβλέπει συγκεκριμένες λέξεις ή φωνήματα. Η λέξη "φώνημα" ορίζετε ως η μικρότερη δυνατή μονάδα ήχου σε μια συγκεκριμένη γλώσσα, που συνήθως αντιπροσωπεύεται από ένα ή δύο γράμματα.



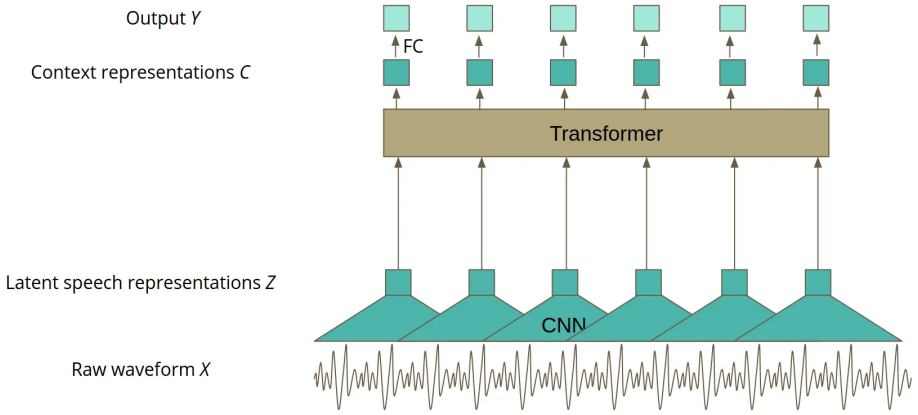

Σχήμα 3.9.1: Η διαδικασία εκπαίδευσης του wav2vec2.0 .

Η πρώτη φάση της εκπαίδευσης είναι το κύριο πλεονέκτημα αυτού του μοντέλου. Η εκμάθηση μιας πολύ καλής αναπαράστασης ομιλίας επιτρέπει την επίτευξη κορυφαίων αποτελεσμάτων σε μικρό όγκο δεδομένων με ετικέτες. Η διαδικασία εκπαίδευσης του μοντέλου φαίνεται στο Σχήμα 3.9.1.

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

- συνελικτικά στρώματα που επεξεργάζονται την ακατέργαστη είσοδο της κυματομορφής για να λάβουν λανθάνουσα αναπαράσταση - $Z$,
- στρώματα Transformer, που δημιουργούν πλαισιωμένη αναπαράσταση - $C$,
- γραμμική προβολή στην έξοδο - $Y$.

Το μοντέλο εκπαιδεύεται να αναγνωρίζει αν δύο μετασχηματισμοί της εισόδου εξακολουθούν να είναι το ίδιο αντικείμενο. Στο Wav2Vec 2.0, τα στρώματα Transformer είναι ο πρώτος τρόπος μετασχηματισμού, ο δεύτερος γίνεται με κβαντισμό. Πιο τυπικά, για μια καλυμμένη λανθάνουσα αναπαράσταση $z_t$, θα θέλαμε να πάρουμε μια τέτοια αναπαράσταση πλαισίου $c_t$ για να μπορέσουμε να μαντέψουμε τη σωστή κβαντισμένη αναπαράσταση $q_t$ μεταξύ άλλων κβαντισμένων αναπαραστάσεων.




Σχήμα 3.9.2: Η αρχιτεκτονική του μοντέλου Wav2Vec 2.0 για εκπαίδευση με αυτοεπίβλεψη.

| | Self-supervised Training | Supervised fine-tuning |
|---|---|---|
| Convolutional Layers | × | × |
| Quantization | × | |
| Transformer | × | × |
| Output Linear Projection | | × |

Πίνακας 3.1: Στάδια εκπαίδευσης και fine-tuning του wav2vec2.0 .

# Κεφάλαιο 4

# Σύνολα Δεδομένων

## 4.1 Το σύνολο δεδομένων DAMP Smule Sing!300x30x2

Η βάση δεδομένων DAMP αποτελείται από ηχογραφήσεις καραόκε που συλλέγονται μέσω μιας εμπορικής εφαρμογής για κινητά, το Smule [23]. Οι ερμηνευτές είναι οι χρήστες της εφαρμογής για κινητά και γενικά ερασιτέχνες τραγουδιστές.

Οι περισσότερες ηχογραφήσεις είναι μονοφωνικές και δεν υπάρχει κυρίαρχη μουσική υπόκρουση που να συνοδεύει το τραγούδι. Αξίζει να σημειωθεί ότι οι ακουστικές συνθήκες του περιβάλλοντος ηχογράφησης ποικίλλουν λόγω της διαδικασίας δημιουργίας των δεδομένων. Αυτό αντικατοπτρίζεται συνήθως ως αντήχηση, ηχώ ή παραμόρφωση των σημάτων φωνής τραγουδιού, εύρος γλωσσών των χρηστών και των θέσεων ηχογράφησης, καθιστώντας το παρόν σύνολο δεδομένων αρκετά αντιπροσωπευτικό των δεδομένων τραγουδιού σε πραγματικές συνθήκες.

Η παρούσα διπλωματική χρησιμοποιεί το σύνολο δεδομένων Sing! 300x30x2 του DAMP, το οποίο περιλαμβάνει χρονικές σημειώσεις σε επίπεδο γραμμής των αντίστοιχων στίχων. Επιπλέον, το σύνολο δεδομένων έχει ισορροπία μεταξύ των φύλων των τραγουδιστών και περιέχει δείγματα από 30 διαφορετικές χώρες. Εξαιτίας αυτού, το σύνολο δεδομένων παρέχει μια μεγάλη ποικιλία προφορών, προσωδίας και εύρους τόνων, καθιστώντας το ακόμη πιο κατάλληλο για το πρόβλημα της αυτόματης αναγνώρισης και ευθυγράμμισης στίχων.

Η οργάνωση και δομή του συνόλου δεδομένων είναι μέσα στο φάκελο sing_300x30x2 όπου υπάρχουν 30 φάκελοι, ένας για κάθε χώρα. Ο κατάλογος των 30 χωρών και οι κωδικοί τους φαίνονται στον Πίνακα 4.1. Μέσα σε κάθε φάκελο χώρας, υπάρχουν τρεις φάκελοι που περιέχουν metadata για κάθε μία από τις 300 διασκευές, χρονικά διαστήματα στίχων για κάθε διασκευή, και τις 600 εκτελέσεις τραγουδιών (1 άνδρας, 1 γυναίκα).

| | | | | | |
|---|---|---|---|---|---|
| AE | United Arab Emirates | HU | Hungary | NO | Norway |
| AR | Argentina | IN | India | PH | Philippines |
| AU | Australia | ID | Indonesia | PT | Portugal |
| BR | Brazil | IR | Iran | RU | Russia |
| CL | Chile | IQ | Iraq | SA | Saudi Arabia |
| CN | China | IT | Italy | SG | Singapore |
| DE | Germany | JP | Japan | TH | Thailand |
| ES | Spain | KR | South Korea | US | United States of America |
| FR | France | MX | Mexico | VN | Vietnam |
| GB | United Kingdom (Great Britain) | MY | Malaysia | ZA | South Africa |

Πίνακας 4.1: Κωδικοί των 30 χωρών στο σύνολο δεδομένων Sing!300x30x2.

Η τελευταία έκδοση δεδομένων της Smule, Sing!, περιέχει 18.676 ερμηνείες από 13.154 τραγουδιστές που

καλύπτουν 5.690 τραγούδια με ίσο αριθμό ερμηνειών ανά φύλο, χωρισμένες ανά χώρα των τραγουδιστών. Παρέχει επίσης τις στιχουργικές προτροπές που παρουσιάστηκαν στον χρήστη-ερμηνευτή μαζί με τους χρόνους των προτροπών, οπότε είναι πολύ πιο εύκολο να ευθυγραμμιστούν οι στίχοι των τραγουδιών με τα σήματα ήχου. Ως προτροπή αναφέρεται ο στίχος που εμφανίζεται (ή υπογραμμίζεται) κάθε φορά στην οθόνη στην εφαρμογή καραόκε που "προτρέπει" τον χρήστη να το τραγουδήσει. Τα δεδομένα συλλέχθηκαν και επεξεργάστηκαν από την Smule κατά το δεύτερο εξάμηνο του 2017 και κυκλοφόρησαν στις αρχές του 2018, επιλέγοντας τους δύο πιο δημοφιλείς τραγουδιστές (άντρες και γυναίκες), από τις 300 πιο δημοφιλείς διασκευές τραγουδιών, από 30 χώρες. Η δημοτικότητα των διασκευών τραγουδιών προσδιορίστηκε μετρώντας τον αριθμό των ερμηνειών, ενώ η δημοτικότητα των ερμηνειών προσδιορίστηκε μετρώντας τον αριθμό των ακροάσεων και των ψήφων που έδωσαν οι χρήστες της εφαρμογής Smule. Μπορούμε να υποθέσουμε ότι οι ερμηνείες που έχουν ψηφιστεί προς τα πάνω είναι καλά τραγουδισμένες, καλής ποιότητας ηχογραφήσεις.

Όταν χρησιμοποιούν την εφαρμογή Smule για κινητά, οι χρήστες τραγουδούν μαζί με ένα κομμάτι συνοδείας καραόκε που παίζει στη συσκευή τους. Οι χρήστες συνήθως χρησιμοποιούν τα ακουστικά τους, ώστε η φωνή τους να καταγράφεται απομονωμένη από τη συνοδεία. Αυτό σημαίνει ότι τα δεδομένα μπορούν να χρησιμοποιηθούν για τη μελέτη της αναγνώρισης τραγουδιστής ομιλίας χωρίς τις προκλήσεις του διαχωρισμού της μουσικής πηγής. Η έρευνα μπορεί, αντίθετα, να επικεντρωθεί στις προκλήσεις της ίδιας της αναγνώρισης τραγουδιστής ομιλίας. Σε ένα δείγμα 100 τυχαία επιλεγμένων ηχογραφήσεων διαπιστώθηκε ότι για το 88% αυτών των ηχογραφήσεων οι χρήστες φορούσαν ακουστικά, ενώ για το 12% δεν φορούσαν, όπως προκύπτει από την απουσία ή την παρουσία του συνοδευτικού ήχου. Από τα δεδομένα χωρίς συνοδεία, περίπου το 15% είχε αξιοσημείωτα επίπεδα θορύβου από το περιβάλλον, δηλαδή οι ερμηνευτές χρησιμοποιούσαν την εφαρμογή σε θορυβώδη χώρο.

Οι πιο δημοφιλείς διασκευές προσδιορίστηκαν μετρώντας τα τραγούδια τις ενάρξεις ή τις ενώσεις (ντουέτο/ομάδα) των ηχογραφήσεων για κάθε διασκευή, εντός χώρας ενδιαφέροντος. Ο όρος "διασκευή" χρησιμοποιείται, επειδή μπορεί να υπάρχουν πολλαπλές διασκευές του ίδιου τραγουδιού. Οι πιο δημοφιλείς εκτελέσεις και οι τραγουδιστές αυτών των διασκευών προσδιορίστηκαν μετρώντας τις ακροάσεις ή/και τα likes για όλες τις εκτελέσεις κάθε διασκευής.

## 4.2 Προ-επεξεργασία του συνόλου δεδομένων

Τα δεδομένα Sing! παρέχουν τις προτροπές κειμένου που εμφανίστηκαν στους ερμηνευτές μαζί με τον χρόνο που εμφανίστηκαν. Για τα περισσότερα τραγούδια, αυτά είναι σε βολική μορφή σε επίπεδο γραμμής-στίχου, δηλαδή, μία προτροπή και μία χρονοσφραγίδα ανά φράση που πρέπει να τραγουδηθεί (συνήθως μία μόνο γραμμή από τους στίχους του τραγουδιού). Για ορισμένα τραγούδια όμως, οι προτροπές παρουσιάζονται στα δεδομένα ως ακολουθία λέξεων ή/και συλλαβών με ξεχωριστές χρονοσφραγίδες ανά μονάδα και χωρίς δείκτη που να υποδεικνύει πού αρχίζουν και πού τελειώνουν οι γραμμές-στίχοι. Για να ανακτήσουμε τους χρόνους για την έναρξη κάθε γραμμής, ανακατασκευάζουμε αυτόματα τις προτροπές σε επίπεδο γραμμής από αυτές τις προτροπές σε επίπεδο λέξεων/συλλαβών, αντιστοιχίζοντας τις λέξεις και τις συλλαβές με τις γραμμές του τραγουδιού στίχων του τραγουδιού που μπορούν να ανακτηθούν από τον ιστότοπο της Smule.

Όπως αναφέρθηκε παραπάνω, το Sing! είναι μια συλλογή πολυγλωσσικών ηχογραφήσεων από φωνητικά κομμάτια καραόκε. Στην διπλωματική αυτή ενδιαφερόμαστε μόνο για τραγούδια που τραγουδιούνται στα αγγλικά. Τα μετα-δεδομένα δεν παρέχουν αναγνωριστικό γλώσσας, οπότε τα μη-αγγλικά τραγούδια αναγνωρίζονται με τη χρήση του CLD2 Naive Bayesian Classifier που έχει εκπαιδευτεί σε κείμενο από ιστοσελίδες. Συγκεκριμένα, κάθε τραγούδι για το οποίο λιγότερο από το 60% των προτάσεων ταξινομείται ως μη αγγλικό αφαιρέθηκε από το σύνολο δεδομένων. Η επιθεώρηση έδειξε ότι αυτή η διαδικασία είναι εύρωστη και μετά το φιλτράρισμα των 18.676 τραγουδιών της αρχικής έκδοσης, παρέμειναν 4.460 αγγλικά τραγούδια.

### 4.2.1 Ευθυγράμμιση και τμηματοποίηση ήχου

Το στάδιο ευθυγράμμισης αποσκοπεί στην παραγωγή μιας ακολουθίας τμηματοποιημένων εκφωνημάτων και της αντίστοιχης μεταγραφής τους, χρησιμοποιώντας τα δεδομένα προτροπής (λέξεις και χρονισμό προτροπής) και το μη τμηματοποιημένο ηχητικό σήμα ως είσοδο.

Υπάρχουν τρεις κύριες προκλήσεις για τη διαδικασία ευθυγράμμισης. Πρώτον, υπάρχει συχνά αναντιστοιχία

μεταξύ των στίχων που ζητήθηκαν και των λέξεων που τραγουδήθηκαν πραγματικά από τον ερμηνευτή. Αυτό συμβαίνει επειδή οι τραγουδιστές παραλείπουν, αλλάζουν ή εισάγουν ολόκληρες φράσεις, είτε κατά λάθος είτε για να δημιουργήσουν μια προσωπική ερμηνεία. Δεύτερον, μπορεί να υπάρχουν σημαντικές διαφορές μεταξύ των χρονικών στιγμών της προτροπής και των χρονικών στιγμών των αντίστοιχων εκφωνήσεων. Γενικά, οι προτροπές εμφανίζονται νωρίς για να δώσουν χρόνο στον τραγουδιστή να προετοιμαστεί, αλλά ο χρόνος προβολής δεν είναι πάντα ίσος. Επιπλέον, οι τραγουδιστές μπορεί να αρχίσουν τις εκφωνήσεις με σημαντική καθυστέρηση, αν δεν είναι εξοικειωμένοι με το τραγούδι. Τέλος, δεν υπάρχει αντιστοιχία ένα προς ένα μεταξύ των προτροπών σε επίπεδο εκφώνησης και των τραγουδισμένων εκφωνήσεων. Μία συνεχώς τραγουδισμένη εκφώνηση μπορεί να καλύπτει περισσότερες από μία προτροπές, δηλαδή δεν υπάρχει μια φυσική παύση στο τέλος κάθε γραμμής ενός τραγουδιού. Αυτό ισχύει ιδιαίτερα για έμπειρους τραγουδιστές που γνωρίζουν τους στίχους του τραγουδιού και δεν χρειάζεται να κάνουν παύσεις για να διαβάσουν ή να προετοιμαστούν.

Η διαδικασία ευθυγράμμισης επιχειρεί να αντιμετωπίσει τις παραπάνω προκλήσεις χρησιμοποιώντας έναν αλγόριθμο βασισμένο σε κανόνες. Ο αλγόριθμος αντιστοιχίζει μια ακολουθία προτροπών σε επίπεδο εκφώνησης με μια ακολουθία μη-σιωπηλών τμημάτων σήματος που εξάγονται από τις ηχογραφήσεις. Οι προτροπές σε επίπεδο εκφώνησης ανακτώνται χρησιμοποιώντας τη διαδικασία που περιγράφηκε στην προηγούμενη υποενότητα. Με κάθε προτροπή συσχετίζεται ένας χρόνος λήξης, ο οποίος λαμβάνεται ως ο χρόνος έναρξης της επόμενης προτροπής. Τα μη-σιωπηλά τμήματα σήματος εξάγονται από το σήμα με τη χρήση ενός απλού ανιχνευτή δραστηριότητας με βάση την ενέργεια, χρησιμοποιώντας μια περιβάλλουσα ενέργειας που παράγεται με την εφαρμογή Pydub. Ο αλγόριθμος χρησιμοποιεί ένα παράθυρο 20 ms και ένα βήμα πλαισίου 1 ms και ταξινομεί τα πλαίσια είτε ως σιωπή είτε ως μη σιωπή ανάλογα με το αν η rms ενέργεια του παραθύρου είναι χαμηλότερη ή υψηλότερη από -25 ντεσιμπέλ (dB) κάτω από το μέγιστο πλάτος του σήματος. Οι σιωπές μικρότερες των 20 ms (π.χ. εντός σιωπών λέξεων) μετατρέπονται σε μη σιωπή. Στη συνέχεια, εντοπίζονται όλα τα τμήματα μη σιωπής (δηλ. ακολουθία μη σιωπηλών πλαισίων που οριοθετούνται από σιωπή). Σημειώνεται ο χρόνος έναρξης και λήξης κάθε τμήματος.

Στον αλγόριθμο ευθυγράμμισης χρησιμοποιούμε τους χρόνους έναρξης και λήξης για να αντιστοιχίσουμε τις προτροπές εκφωνήσεων με τα αντίστοιχα τμήματα σήματος. Ωστόσο, σε ορισμένες περιπτώσεις είναι απαραίτητο να ενώσουμε δύο ή περισσότερα τμήματα που δεν είναι σιωπηλά για να ταυτιστούν με μία μόνο προτροπή. Αυτό συμβαίνει όταν μια εκφώνηση έχει χωριστεί λόγω της ύπαρξης μιας μικρής σιωπής. Σε άλλες περιπτώσεις είναι απαραίτητο να προτρέπονται κείμενα για να ταιριάξουν με μία μόνο εκφώνηση, δηλαδή όταν ο ερμηνευτής τραγουδάει περισσότερες από μία γραμμές χωρίς ενδιάμεση παύση. Για να επιτευχθεί αυτό, ο αλγόριθμος προχωρά ως εξής:

1. Οι προτροπές που δεν τέμνονται με κανένα υπάρχον μη-σιωπηλό τμήμα απορρίπτονται (ο τραγουδιστής απέτυχε να τραγουδήσει το στίχο).
2. Απορρίπτονται τα μη-σιωπηλά τμήματα που δεν τέμνονται με οποιαδήποτε υπάρχουσα προτροπή (δηλαδή, συνήθως ξένος θόρυβος, όπως βήχας).
3. Όπου περισσότερα από ένα μη-σιωπηλά τμήματα τέμνονται με την ίδια προτροπή, τα τμήματα ενώνονται.
4. Όπου περισσότερες από μία προτροπές τέμνονται με το ίδιο μη-σιωπηλό τμήμα, οι προτροπές ενώνονται.
5. Εάν κάθε τμήμα δεν τέμνει μόνο μία προτροπή, επιστρέψτε στο βήμα 4.
6. Τα μη-σιωπηλά τμήματα αντιστοιχίζονται τώρα με την προτροπή που τα τέμνει.

Μετά την εκτέλεση του αλγορίθμου, εξετάστηκε ένα δείγμα 100 τμημάτων για να αξιολογηθεί η ποιότητα της ευθυγράμμισης. Διαπιστώθηκε ότι για το 60% τα τμήματα ευθυγραμμίστηκαν σωστά με την προτροπή, δηλαδή με σωστούς χρόνους και με προτροπές που παρείχαν τη σωστή μεταγραφή. Ένα επιπλέον 32% ήταν μόνο εν μέρει σωστό. Σε αυτές τις περιπτώσεις, η αντιστοίχιση του τμήματος με την προτροπή ήταν σωστή, αλλά στον τραγουδιστή υπήρξε προσθήκη, αφαίρεση ή αντικατάσταση μιας ή περισσότερων λέξεων σε σχέση με τον προτρεπτικό στίχο. Στο 8% των τμημάτων η ευθυγράμμιση είχε αποτύχει εντελώς. Συνήθως σε αυτές τις περιπτώσεις οι προτροπές ευθυγραμμίζονταν με τμήματα που περιείχαν μόνο θόρυβο υποβάθρου που εισήχθη λόγω αποτυχίας του προηγούμενου σταδίου ανίχνευσης φωνητικής δραστηριότητας.

Λόγω αυτών των ατελειών, η βασική μας επεξεργασία ευθυγράμμισης χρησιμοποιείται μόνο για τη δημιουργία των δεδομένων εκπαίδευσης. Για να διασφαλιστεί ότι οι ακριβείς επιδόσεις αναγνώρισης μπορούν να

μετρηθούν, για τα δεδομένα δοκιμής (test), κατασκευάστηκε ένα πρότυπο σύνολο με τη χρήση ανθρώπινων σχολιαστών για τη διόρθωση των χρονισμών ευθυγράμμισης και την εκ νέου μεταγραφή της ομιλίας.

### 4.2.2 Ορισμός συνόλου εκπαίδευσης και δοκιμής

Πρώτον, αξιοποιώντας τα πλεονεκτήματα των πληροφοριών για τις χώρες των τραγουδιστών, τα δεδομένα χωρίζονται σε τρία σύνολα δεδομένων, DSing1, DSing3 και DSing30, τα οποία εισάγουν σταδιακά τις ηχογραφήσεις από ένα ευρύτερο σύνολο χωρών (Πίνακας 4.2). Το DSing1, κατασκευάζεται χρησιμοποιώντας το υποσύνολο των ηχογραφήσεων από τραγουδιστές που είναι εγγεγραμμένοι ως χρήστες στη Μεγάλη Βρετανία. Το DSing3 κατασκευάζεται από το υποσύνολο των ηχογραφήσεων από τραγουδιστές που είναι εγγεγραμμένοι σε μία από τις τρεις αγγλόφωνες χώρες, δηλαδή τη Μεγάλη Βρετανία, τις ΗΠΑ και την Αυστραλία. Τέλος, το μεγαλύτερο σύνολο δεδομένων, το DSing30, κατασκευάζεται χρησιμοποιώντας τραγουδιστές και από τις 30 χώρες που είναι διαθέσιμες στο σύνολο δεδομένων Sing!. Σημειώστε ότι σε όλες τις περιπτώσεις χρησιμοποιούνται μόνο τα αγγλικά τραγούδια, δηλαδή το DSing30 θα περιέχει πολλές ηχογραφήσεις που τραγουδήθηκαν στα αγγλικά από ομιλητές που η αγγλική γλώσσα δεν είναι η μητρική τους.

| | DSing1 | DSing3 | DSing30 |
|---|---|---|---|
| countries | GB | GB, US, AU | All 30 |
| singers | 352 | 1050 | 3205 |
| songs | 434 | 1343 | 4324 |
| utterances | 8794 | 25526 | 81092 |
| hours | 15.1 | 44.7 | 149.1 |

Πίνακας 4.2: Διαχωρισμός του συνόλου δεδομένων.

Τα δεδομένα στο DSing1 χωρίζονται περαιτέρω σε σύνολα εκπαίδευσης (train), δοκιμής (test) και ανάπτυξης (dev) που περιλαμβάνουν το 80%, το 10% και το 10% των δεδομένων αντίστοιχα. Έχει ληφθεί μέριμνα ώστε τα σύνολα να είναι ασύνδετα όσον αφορά τόσο τους τραγουδιστές όσο και τις διασκευές, δηλαδή κανένας τραγουδιστής ή διασκευή που εμφανίζεται σε ένα σύνολο δεν εμφανίζεται σε οποιοδήποτε άλλο σύνολο. Αυτό περιπλέκεται λόγω της συσχέτισης πολλών προς πολλούς μεταξύ τραγουδιστών και διασκευών και πρέπει να χαθούν κάποια δεδομένα για να ικανοποιηθεί αυτός ο περιορισμός. Τυχόν διασκευές που εμφανίζονται στα σύνολα ανάπτυξης και δοκιμής DSing1 αφαιρούνται από τα σύνολα DSing3 και DSing30, έτσι ώστε αυτά τα σύνολα δεδομένων να μπορούν να χρησιμοποιηθούν για εκπαίδευση.

Για την κατασκευή των προτύπων δεδομένων αξιολόγησης, επιλέχθηκαν τυχαία περίπου 600 εκφωνήσεις από τα δεδομένα που αντιστοιχούσαν στα σύνολα ανάπτυξης και δοκιμής. Για αυτές τις εκφωνήσεις οι ευθυγραμμίσεις διορθώθηκαν από ανθρώπους και οι προτροπές αντικαταστάθηκαν με ανθρώπινες μεταγραφές των λέξεων που πραγματικά τραγουδήθηκαν. Απορρίφθηκαν ηχογραφήσεις στις οποίες διαπιστώθηκε ότι οι χρήστες δεν φορούσαν ακουστικά με συνέπεια να ακούγεται σημαντικά το τραγούδι στο υπόβαθρο, ενώ λήφθηκε μέριμνα ώστε να διατηρούνται το πολύ 20 ηχογραφήσεις ανά ομιλητή. Η διαδικασία αυτή οδήγησε σε 482 ηχογραφήσεις που κάλυπταν 40 ομιλητές (27 γυναίκες και 13 άνδρες) για το σύνολο ανάπτυξης(dev) και 480 ηχογραφήσεις που κάλυπταν 43 ομιλητές (30 γυναίκες και 13 άνδρες) για το σύνολο δοκιμής(test). Τα παραπάνω παρουσιάζονται ποσοτικά στον Πίνακα 4.3

| Gold Standard Evaluation Dataset | | | | |
|---|---|---|---|---|
| | utterances | speakers | songs | hours |
| dev | 482 | 40 | 66 | 0.7 |
| test | 480 | 43 | 70 | 0.8 |

Πίνακας 4.3: Σύνολα ανάπτυξης και δοκιμής.

# Κεφάλαιο 5

# Αυτόματη Μεταγραφή Στίχων

## 5.1 Hidden Markov Models

### 5.1.1 Το εργαλείο Kaldi

Το Kaldi [24] είναι μια εργαλειοθήκη αναγνώρισης ομιλίας ανοικτού κώδικα γραμμένη σε C++ για αναγνώριση ομιλίας και επεξεργασία σήματος. Το Kaldi παρέχει ένα σύστημα αναγνώρισης ομιλίας που βασίζεται σε μετατροπείς πεπερασμένης κατάστασης (χρησιμοποιώντας το ελεύθερα διαθέσιμο OpenFst [12]), μαζί με λεπτομερή έγγραφα και scripts για την κατασκευή ολοκληρωμένων συστημάτων αναγνώρισης. Το Kaldi είναι γραμμένο σε C++, και η βασική βιβλιοθήκη υποστηρίζει μοντελοποίηση αυθαίρετων μεγεθών φωνητικού πλαισίου, ακουστική μοντελοποίηση με υποχωρικά μοντέλα μίξης Gauss (SGMM) καθώς και τυπικά μοντέλα μίξης Gauss, μαζί με όλους τους ευρέως χρησιμοποιούμενους γραμμικούς και αφινικούς μετασχηματισμούς. Το Kaldi κυκλοφορεί υπό την Άδεια Apache License v2.0, η οποία είναι εξαιρετικά μη περιοριστική, καθιστώντας το κατάλληλο για μια ευρεία κοινότητα χρηστών.

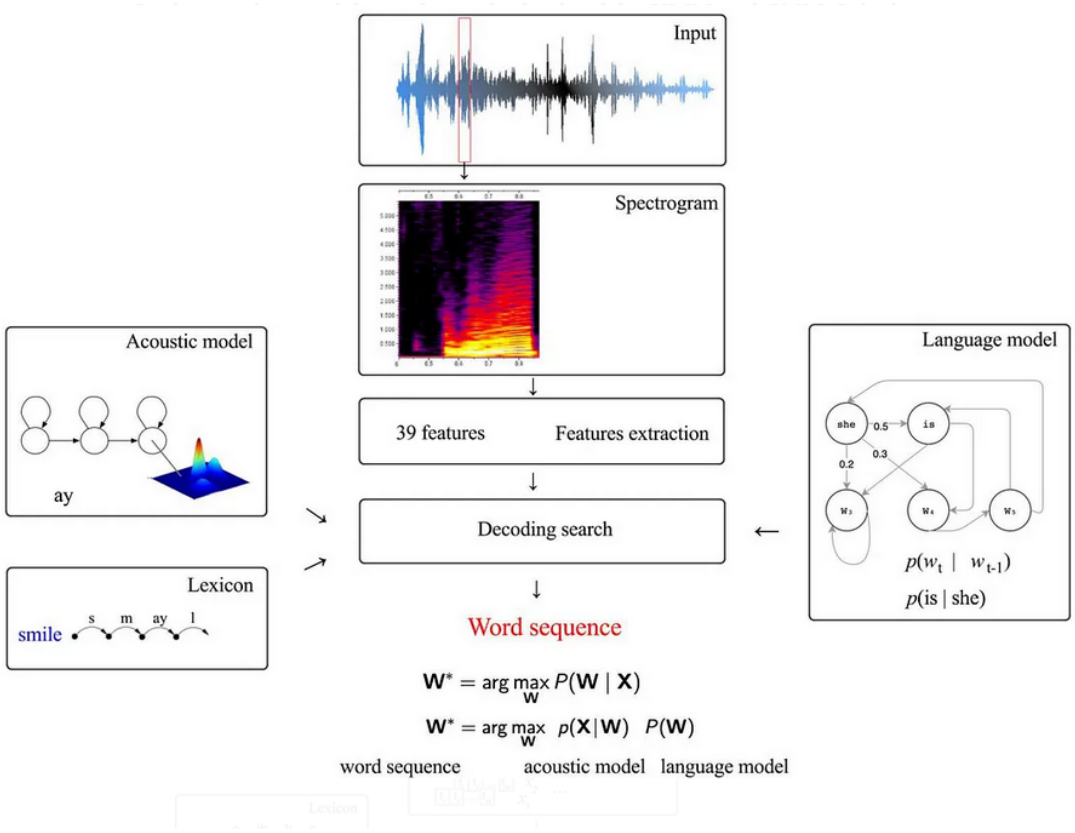
Σε ένα μεγάλο σύνολο αρχείων ήχου (audio data) που περιέχουν ομιλία, για την ανάκτηση του γραπτού λόγου που προφέρουν οι ομιλητές (transcription) μέσω του Kaldi, η διαδικασία είναι η εξής:

- Χωρισμός των αρχείων ήχου σε 2 σετ, εκπαίδευσης (train) και αξιολόγησης (test). Η αναλογία πρέπει να είναι περίπου 90%/10%, δηλαδή 90% των αρχείων ήχου να μπουν στην κατηγορία train και το υπόλοιπο 10% να μπει στην κατηγορία test.
- Προετοιμασία αρχείων κειμένου που θα περιέχουν πληροφορίες για το περιεχόμενο τον αρχείων ήχου, όπως λεξικό (όλες οι λέξεις που εμφανίζονται), λίστα φωνημάτων, πληροφορίες για τους ομιλητές και σε ποια αρχεία ήχου εμφανίζονται κλπ.
- Προετοιμασία μερικών ακόμα αρχείων κειμένου όπως το γλωσσικό μοντέλο, ρυθμίσεις για τα script του Kaldi κ.α. Από τη στιγμή που έχουν δημιουργηθεί σωστά τα αρχεία στο προηγούμενο βήμα, τα αρχεία αυτά θα δημιουργηθούν αυτόματα καλώντας κατάλληλα script του Kaldi.
- Δημιουργία γλωσσικού μοντέλου. Θεωρητικά μπορεί να γίνει με οποιοδήποτε πρόγραμμα, αρκεί να παράγει γλωσσικά μοντέλο σύμφωνα με το ARPA πρότυπο.
- Εξαγωγή διανυσμάτων χαρακτηριστικών (feature vectors) των αρχείων ήχου. Στο Kaldi χρησιμοποιούνται οι MFCC κατά κύριο λόγο. Αφορά και το train και το test.
- Εκπαίδευση (training) ακουστικού μοντέλου. Τυπικά χρησιμοποιούνται πολλά και συγκρίνονται τα αποτελέσματα που δίνει το καθένα και έπειτα γίνεται επιλογή αυτού που δίνει τα καλύτερα αποτελέσματα. Αυτό γίνεται χρησιμοποιώντας μόνο τα αρχεία του train set.
- Δημιουργία γράφων HCLG.

- Ανάλυση των αρχείων ήχου (decoding) και αξιολόγηση των αποτελεσμάτων. Αυτό γίνεται πάνω στα αρχεία του test set.

### 5.1.2 Ακουστικό μοντέλο GMM-HMM

Εκπαιδεύσαμε έναν αυτόματο σύστημα αναγνώρισης ομιλίας ακολουθώντας την παραδοσιακή προσέγγιση της δημιουργίας ακουστικών μοντέλων φωνημάτων με επίγνωση του συμφραζομένου με βάση την αρχιτεκτονική GMM-HMM. Η θεμελιώδης ιδέα πίσω από ένα σύστημα αυτόματης αναγνώρισης ομιλίας είναι η δημιουργία ενός μοντέλου που προβλέπει με επιτυχία την ακολουθία των λέξεων δεδομένης της ηχητικής καταγραφής μιας εκφώνησης.




Σχήμα 5.1.1: Αναγνώριση στίχων με το ακουστικό μοντέλο HMM.

Ξεκινάμε με την εκπαίδευση ενός μονόφωνου ακουστικού μοντέλου GMM-HMM χρησιμοποιώντας 13 χαρακτηριστικά MFCC με μέγεθος παραθύρου 25 χιλιοστά του δευτερολέπτου και μέγεθος άλματος 10 χιλιοστά του δευτερολέπτου, με ρυθμίσεις που φαίνονται στον Πίνακα 5.1. Στη συνέχεια εφαρμόζουμε κανονικοποίηση του μέσου όρου και της διακύμανσης (CMVN). Για να συμπεριλάβουμε την εξάρτηση από τα συμφραζόμενα στα ακουστικά μοντέλα, επανεκπαιδεύουμε το μοντέλο χρησιμοποιώντας τα χαρακτηριστικά delta και delta-delta, τα οποία μοντελοποιούν τα τηλέφωνα ως "τρίφωνα". Στη συνέχεια, εφαρμόζεται μείωση των διαστάσεων στα διανύσματα χαρακτηριστικών με τη χρήση γραμμικής ανάλυσης διάκρισης (LDA). Στη συνέχεια, εφαρμόζουμε μετασχηματισμό "Μέγιστης γραμμικής πιθανοφάνειας χώρου χαρακτηριστικών" (fMLLR) στα χαρακτηριστικά εισόδου, προσαρμόζοντας τις παραμέτρους του GMM για την απόκτηση μιας αναπαράστασης του χώρου χαρακτηριστικών ανεξάρτητης από τους τραγουδιστές.

Τα δεδομένα εκπαίδευσης που χρησιμοποιήθηκαν ήταν ένα μικρό μέρος του συνόλου των δεδομένων, το οποίο αποτελείται από τις ηχογραφήσεις που προέρχονται μόνο από τις χώρες που έχουν ως μητρική γλώσσα την αγγλική γλώσσα στο σύνολο δεδομένων (Μεγάλη Βρετανία, Ηνωμένες Πολιτείες και Αυστραλία).

| Configuration | Parameters |
|---|---|
| use energy | False |
| sample frequency | 16kHz |
| allow downsample | True |
| number of cepstrals | 23 |
| low cut-off frequency | 20Hz |
| high cut-off frequency | 7600Hz |
| frame length | 20ms |

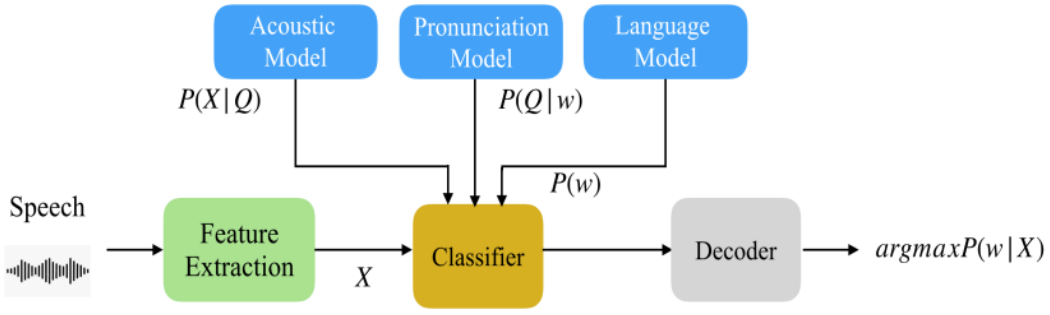
Πίνακας 5.1: Ρυθμίσεις για την εξαγωγή των χαρακτηριστικών MFCC.

### Γλωσσικό μοντέλο

Το γλωσσικό μοντέλο (LM) βασίζεται στο σύνολο κειμένων που αποτελείται από τους στίχους των τραγουδιών που περιέχονται στο σύνολο δεδομένων Dsing!. Ορισμένες από αυτές τις στρατηγικές περιλαμβάνουν την αφαίρεση των μη ASCII χαρακτήρων και των μη λυρικών λέξεων (όπως "verse" ή "chorus" κ.λπ.). Διορθώσαμε την ορθογραφία ορισμένων λέξεων στα ακατέργαστα δεδομένα των στίχων που περιλαμβάνουν επαναλαμβανόμενα φωνήεντα που υποδηλώνουν διατήρηση στην αντίστοιχη συλλαβή (π.χ. "YEEEEAAAAAH" σε "YEAH"). Οι αριθμοί απορρίπτονται εάν αναπαρίστανται ως ψηφία. Ως αποτέλεσμα, λαμβάνουμε 1.747.287 γραμμές στίχων και 91.654 μοναδικές λέξεις ως δεδομένα κειμένου. Κατασκευάσαμε ένα γλωσσικό μοντέλο μέγιστης εντροπίας 3gram (MaxEnt LM) για το σύστημα αναγνώρισης ομιλίας μας. Σε γενικές γραμμές, τα μοντέλα 4gram υπερτερούν των μοντέλων 3gram ωστόσο παρατηρήσαμε το αντίθετο στα αποτελέσματα του τριφωνικού μοντέλου GMM-HMM του ποσοστού σφάλματος λέξης (WER). Κατασκευάσαμε το LM μέγιστης εντροπίας χρησιμοποιώντας την εργαλειοθήκη SRILM [25].

### Λεξικό

Στο σύστημα ευθυγράμμισης που διαθέτουμε, χρησιμοποιούμε ένα προ-εκπαιδευμένο τριφωνικό ακουστικό μοντέλο για να προβλέψουμε την ακολουθία των λέξεων σε επίπεδο πλαισίου και, κατά συνέπεια, να λάβουμε χρονικές ευθυγραμμίσεις. Δεδομένου ότι ο τελικός στόχος είναι η ευθυγράμμιση σε επίπεδο λέξης, αντί σε επίπεδο φωνήματος, απαιτείται μια γλωσσολογικά τεκμηριωμένη αντιστοίχιση των λέξεων με τη φωνημική τους αναπαράσταση. Η εν λόγω αντιστοίχιση αναφέρεται συνήθως ως λεξικό (προφοράς). Στο σύστημά μας, χρησιμοποιήσαμε το CMU Sphinx English Pronunciation Dictionary [26] ως λεξικό για την μετατροπή των λέξεων σε φωνήματα.




Σχήμα 5.1.2: Διαδικασία αναγνώρισης στίχων με το ακουστικό μοντέλο HMM.

Μετά τη δημιουργία τριφωνικών μοντέλων GMM-HMM με επίγνωση των συμφραζομένων, λαμβάνουμε ευθυγραμμίσεις για ολόκληρο το σύνολο δεδομένων, οι οποίες απαιτούνται για την εκπαίδευση GMM-HMM.

#### Ευθυγράμμιση

Η ευθυγράμμιση των φωνημάτων με το ηχητικό σήμα πραγματοποιείται μέσω μιας μεθόδου "αναγκαστικής ευθυγράμμισης" (forced alignment). Η αναγκαστική ευθυγράμμιση είναι η διαδικασία εύρεσης της καλύτερης διαδρομής από μια ακολουθία γεγονότων-στόχων που ελαχιστοποιεί το συνολικό κόστος. Στο σύστημά μας, οι πιθανότητες μετάβασης των φωνημάτων εκτιμώνται μετά την εκπαίδευση των ακουστικών καταστάσεων HMM στο σύστημα αναγνώρισης ομιλίας που περιγράφεται παραπάνω. Εμείς, στη συνέχεια, χρησιμοποιούμε τον αλγόριθμο αποκωδικοποίησης Viterbi [13] για την απόκτηση της πιο πιθανής αλυσίδας φωνημάτων με την αντιστοίχιση των πιθανοτήτων μετάβασης με τα ακουστικά χαρακτηριστικά που εξάγονται από τα ακουστικά πλαίσια. Ο αλγόριθμος Viterbi χρησιμοποιεί τον αλγόριθμο αναζήτησης δέσμης (beam search) για την εύρεση της καλύτερης διαδρομής, όπου οι εμφανίσεις φωνημάτων χαμηλής πιθανότητας περικόπτονται για την αποφυγή συσσωρευμένων σφαλμάτων ευθυγράμμισης και για λόγους αποδοτικότητας της μνήμης.

### 5.1.3 Αποτελέσματα

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα που προέκυψαν με τη χρήση ενός GMM-HMM. Τα ακουστικά χαρακτηριστικά που χρησιμοποιήθηκαν είναι 13 MFCC συν delta, delta-delta και ενέργεια, με μήκος πλαισίου 25 χιλιοστά του δευτερολέπτου και επικάλυψη 15 χιλιοστών του δευτερολέπτου. Αρχικά εκπαιδεύσαμε ένα GMM-HMM τριφωνικό GMM προσαρμοσμένο στον ομιλητή πάνω στο fMLLR. Αυτό το μοντέλο χρησιμοποιήθηκε για την εφαρμογή μιας διαδικασίας καθαρισμού (τυπική στο Kaldi) στο σύνολο ακουστικής εκπαίδευσης για την αφαίρεση των κακών εκφωνήσεων από τα δεδομένα εκπαίδευσης, (π.χ. φιλτράροντας εκείνες με λανθασμένη μεταγραφή). Αυτή η διαδικασία αφαίρεσε περίπου το 10% των εκφωνημάτων εκπαίδευσης. Για όλα τα πειράματα, το LM που χρησιμοποιείται είναι ένα μοντέλο MaxEnt 3gram που εκπαιδεύτηκε στα δεδομένα LMSmule+ χρησιμοποιώντας ένα λεξιλόγιο 28K λέξεων. Παρουσιάζουμε τη βασική μας γραμμή με τα αποτελέσματα που επιτεύχθηκαν κατά την εκπαίδευση στα τρία σύνολα δεδομένων εκπαίδευσης DSing καραόκε Smule Sing! που περιγράφηκαν προηγουμένως, δηλαδή τα DSing1, DSing3 και DSing30. Για κάθε σύστημα, η απόδοση μετράται με τη χρήση των χρυσών προτύπων συνόλων ανάπτυξης και δοκιμής που περιγράφονται στο Κεφάλαιο 4.

| train data | %WER | insertions | deletions | substitutions |
|---|---|---|---|---|
| Dsing1 | 61.69 | 270 | 711 | 1873 |
| Dsing3 | 54.39 | 263 | 485 | 1768 |
| DSing30 | 51.41 | 183 | 566 | 1629 |

Πίνακας 5.2: Αποτελέσματα σφαλμάτων αναγνώρισης με το ακουστικό μοντέλο GMM-HMM.

Από τα πειράματα προκύπτει ότι όσο αυξάνουμε το μέγεθος του συνόλου εκπαίδευσης, τόσο μειώνεται το σφάλμα αναγνώρισης WER. Τα περισσότερα δεδομένα επιτρέπουν στο σύστημα να καταγράφει και να ερμηνεύει με ακρίβεια τους στίχους καθώς μαθαίνει καλύτερα τις αντιστοιχίες ήχων-φωνημάτων σε συνδυασμό με τις συνθήκες ηχογράφησης και τη μουσική υποβάθρου. Πιο αναλυτικά, μερικοί λόγοι για τους οποίους η προσθήκη περισσότερων δεδομένων εκπαίδευσης είναι επωφελής είναι οι εξής:

Μεγαλύτερη κάλυψη των στυλ τραγουδιού και της διακύμανσης: Το τραγούδι περιλαμβάνει ένα ευρύ φάσμα στυλ, ειδών και φωνητικών τεχνικών. Με περισσότερα δεδομένα εκπαίδευσης, καταγράφεται μια πιο ολοκληρωμένη αντιπροσώπευση των φωνών τραγουδιού. Αυτό περιλαμβάνει διαφορετικές φωνητικές καταγραφές, μοτίβα βιμπράτο, μελωδικές διαμορφώσεις, παραλλαγές άρθρωσης και πολλά άλλα.

Βελτιωμένη γενίκευση: Εκθέτοντας το μοντέλο σε περισσότερα παραδείγματα από διάφορους τραγουδιστές, τραγούδια και παραστάσεις, μαθαίνει να εξάγει κοινά μοτίβα και χαρακτηριστικά που είναι συνεπή σε διαφορετικές περιπτώσεις. Αυτό επιτρέπει στο σύστημα να χειρίζεται πιο αποτελεσματικά άγνωστες για το μοντέλο φωνές τραγουδιού, μειώνοντας την πιθανότητα σφαλμάτων και βελτιώνοντας τη συνολική ακρίβεια αναγνώρισης.

Μειωμένη μεροληψία και ανθεκτικότητα: Τα συστήματα αναγνώρισης που εκπαιδεύονται σε περιορισμένα δεδομένα ενδέχεται να πάσχουν από προκαταλήψεις και περιορισμούς στις δυνατότητες αναγνώρισής τους. Αυτές οι μεροληψίες μπορεί να προκύψουν από ανισορροπίες στα δεδομένα εκπαίδευσης, όπως η υπερεκπροσώπηση συγκεκριμένων τραγουδιστών ή ειδών.

Καταγραφή σπάνιων ή δύσκολων φωνητικών περιστατικών: Το τραγούδι περιλαμβάνει όχι μόνο κοινά και ευρέως εκτελούμενα τραγούδια αλλά και μοναδικές ή αντισυμβατικές φωνητικές εκτελέσεις. Συμπεριλαμβάνοντας περισσότερα δεδομένα εκπαίδευσης, αυξάνεται η πιθανότητα καταγραφής απαιτητικών περιπτώσεων που μπορεί να εμφανίζονται λιγότερο συχνά. Αυτές οι περιπτώσεις θα μπορούσαν να περιλαμβάνουν σύνθετες φωνητικές τεχνικές, υψηλές ή χαμηλές νότες, περάσματα με γρήγορο ρυθμό ή αντισυμβατικά φωνητικά εφέ.

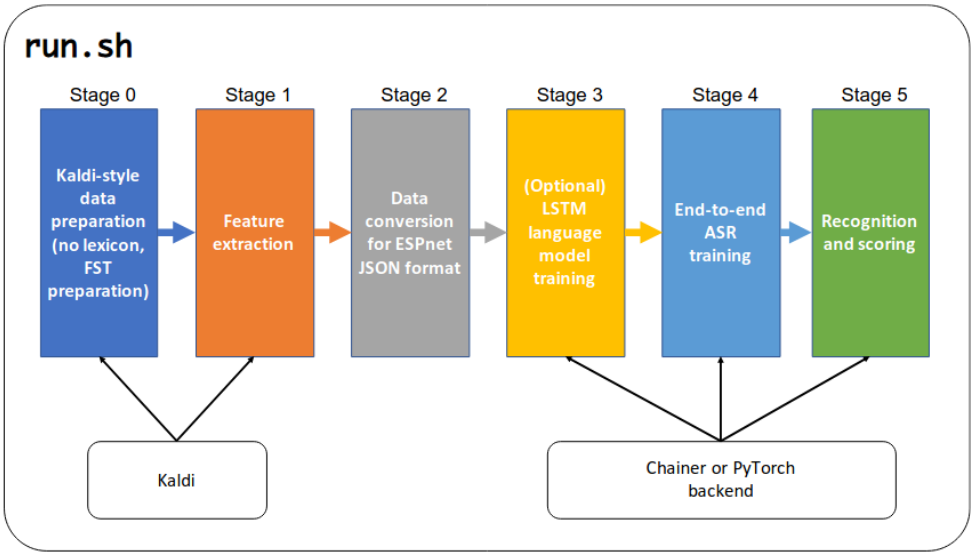
Βελτιωμένη ακουστική μοντελοποίηση: Το μοντέλο μπορεί να μάθει καλύτερες αναπαραστάσεις διαφόρων ακουστικών μοτίβων, όπως τα περιγράμματα του τόνου, τις διακυμάνσεις του ηχοχρώματος, τη δυναμική και τις αρθρωτικές αποχρώσεις που αφορούν ειδικά την τραγουδιστή φωνή.

Είναι σημαντικό να σημειωθεί ότι η προσθήκη περισσότερων δεδομένων εκπαίδευσης από μόνη της μπορεί να μην εγγυάται βέλτιστα αποτελέσματα. Η κατάλληλη προεπεξεργασία, η μηχανική των χαρακτηριστικών, η αρχιτεκτονική του μοντέλου είναι επίσης κρίσιμοι παράγοντες για τον σχεδιασμό ενός αποτελεσματικού συστήματος αναγνώρισης στίχων. Παρ' όλα αυτά, η ενσωμάτωση ενός μεγαλύτερου και ποικίλου συνόλου δεδομένων εκπαίδευσης παρέχει μια ισχυρή βάση για τη βελτίωση της απόδοσης αναγνώρισης και την μείωση του σφάλματος αναγνώρισης WER.

## 5.2 Transformers

### 5.2.1 Το εργαλείο ESPNet

Το ESPnet [27] είναι μια ολοκληρωμένη εργαλειοθήκη επεξεργασίας ομιλίας που καλύπτει την ολοκληρωμένη αναγνώριση ομιλίας, τη μετατροπή κειμένου σε ομιλία, τη μετάφραση ομιλίας, την ενίσχυση ομιλίας, την καταγραφή ομιλητή, την κατανόηση προφορικού λόγου και άλλα. Το ESPnet χρησιμοποιεί το pytorch ως μια μηχανή βαθιάς μάθησης και ακολουθεί επίσης την επεξεργασία δεδομένων στο στυλ του Kaldi, την εξαγωγή/μορφοποίηση χαρακτηριστικών και τις "συνταγές" για να παρέχει μια πλήρη διάταξη για διάφορα πειράματα επεξεργασίας ομιλίας. Τα στάδια για την εκπαίδευση και την αναγνώριση στίχων μέσω του εργαλείου ESPNet απεικονίζονται στο Σχήμα 5.2.1.




Σχήμα 5.2.1: Τα στάδια εκπαίδευσης του ESPNet.

### 5.2.2 Σχεδιασμός Πειράματος με Speech Transformer

Ο Speech Transformer [28] είναι ένα μοντέλο μεταγωγής που βασίζεται εξ ολοκλήρου σε προσοχή, αντικαθιστώντας τα επαναλαμβανόμενα στρώματα που χρησιμοποιούνται συνήθως στις αρχιτεκτονικές κωδικοποιητή-αποκωδικοποιητή με την αυτο-προσοχή με πολλαπλές κεφαλές.

Συνδυάζουμε δύο σύγχρονες τεχνικές ASR: κοινή αποκωδικοποίηση της συνδεσμολογικής χρονικής ταξινόμησης (CTC) και του γλωσσικού μοντέλου (LM) και του Transformer, για να επιτύχουμε καλύτερες επιδόσεις αναγνώρισης στίχων. [29]

Σε αυτό το πείραμα, ο Transformer προβλέπει μια ακολουθία εξόδου αναγνωριστικών χαρακτήρων $Y$ από μια ακολουθία εισόδου με χαρακτηριστικά ομιλίας τράπεζας φίλτρων log-Mel $X^{fbank}$. Μπορούμε να χωρίσουμε αυτή την αρχιτεκτονική Transformer σε δύο μέρη. Το ένα είναι ένα δίκτυο κωδικοποιητή που μετατρέπει το $X^{fbank}$ σε μια ενδιάμεση ακολουθία κωδικοποιημένων χαρακτηριστικών $X_e$. Το άλλο είναι ένα δίκτυο αποκωδικοποιητή που προβλέπει το νέο χαρακτήρα $Y[u+1]$ δεδομένου του $X_e$ και των χαρακτήρων προθέματος $Y[1], \ldots, Y[u]$. Και τα δύο δίκτυα αποτελούνται από μονάδες δικτύου προσοχής και τροφοδότησης. Μια σημαντική διαφορά μεταξύ του μετασχηματιστή και του RNN είναι ένας μηχανισμός αυτοπροσοχής που μετασχηματίζει την ακολουθία χρησιμοποιώντας πίνακες προσοχής στα πλαίσια εισόδου αντί της αναδρομικής σύνδεσης.

**Multi-head attention**

Ο Transformer περιέχει ένα επίπεδο dot-προσοχής με κλιμάκωση:

$$\text{att}(X_q, X_k, X_v) = \text{softmax}\left(\frac{X_q X_k^\top}{\sqrt{d_{\text{att}}}}\right) X_v$$

όπου τα $X_k, X_v \in \mathbb{R}^{n_k \times d_{\text{att}}}$ και το $X_q \in \mathbb{R}^{n_q \times d_{\text{att}}}$ είναι ακολουθίες εισόδου για αυτό το επίπεδο προσοχής, όπου το $d_{\text{att}}$ είναι ο αριθμός των χαρακτηριστικών διαστάσεων, το $n_q$ είναι το μήκος του $X_q$ και το $n_k$ είναι το μήκος των $X_k$ και $X_v$. Αναφερόμαστε στο $X_q X_k^\top$ ως "πίνακα προσοχής". Αυτές οι είσοδοι $X_q, X_k$ και $X_v$ θεωρούνται ως ένα ερώτημα και ένα σύνολο ζευγαριού κλειδιού-τιμής, αντίστοιχα.

Επιπλέον, για να επιτρέψουμε στο μοντέλο να δίνει πολλαπλές προσοχές παράλληλα, επεκτείνουμε αυτό το επίπεδο προσοχής στη μορφή πολυκεφαλικής προσοχής (MHA): $MHA(Q, K, V) = [H_1, H_2, \ldots, H_{d_{head}}]W^{head}$, όπου τα $K, V \in \mathbb{R}^{n_k \times d_{\text{att}}}$

### Αρχιτεκτονική κωδικοποιητή ομιλίας

Η ομιλία εισόδου αναπαρίσταται ως μια ακολουθία χαρακτηριστικών φίλτρου log-Mel $X_{\text{fbank}} \in \mathbb{R}^{n_{\text{fbank}} \times 80}$, όπου το $n_{\text{fbank}}$ είναι το μήκος της εισόδου. Αρχικά, υποδειγματοληπτούμε το $X_{\text{fbank}}$ σε $X_{\text{sub}} \in \mathbb{R}^{n_{\text{sub}} \times d_{\text{att}}}$ χρησιμοποιώντας ένα δίκτυο CNN με δύο επίπεδα στον άξονα του χρόνου, ενεργοποίηση ReLU, $d_{\text{att}}$ κανάλια, μέγεθος βήματος 2 και μέγεθος πυρήνα 3, όπου το $n_{\text{sub}}$ είναι το μήκος της εξόδου της ακολουθίας CNN. Έπειτα, το δίκτυο κωδικοποιητή Transformer μετατρέπει αυτήν την είσοδο σε μια ακολουθία κωδικοποιημένων χαρακτηριστικών $X_e \in \mathbb{R}^{n_{\text{sub}} \times d_{\text{att}}}$ για το CTC και το δίκτυο αποκωδικοποίησης, ως εξής:

$$X_0 = [[X^{\text{sub}}[1], \text{PE}[1]]^T >, \ldots, [X^{\text{sub}}[n_{\text{sub}}], \text{PE}[n_{\text{sub}}]]^T],$$

$$X'_i = X_i + \text{MHA}_i(X_i, X_i, X_i),$$

$$X_{i+1} = X'_i + \text{FF}_i(X'_i),$$

όπου $i = 0, \ldots, e$ είναι ο δείκτης των επιπέδων κωδικοποιητή, $e$ είναι ο αριθμός των επιπέδων κωδικοποιητή, PE είναι η συνημιτονική θέσαριστική κωδικοποίηση:

$$\text{PE}[t] = \begin{cases} \sin\left(\frac{t}{10000^{\frac{t}{d_{\text{att}}}}}\right) & \text{αν } t \text{ είναι άρτιος,} \\ \cos\left(\frac{t}{10000^{\frac{t}{d_{\text{att}}}}}\right) & \text{αν } t \text{ είναι περιττός,} \end{cases}$$

$\text{FF}_i$ είναι το δίκτυο τροφοδοσίας με δύο επίπεδα:

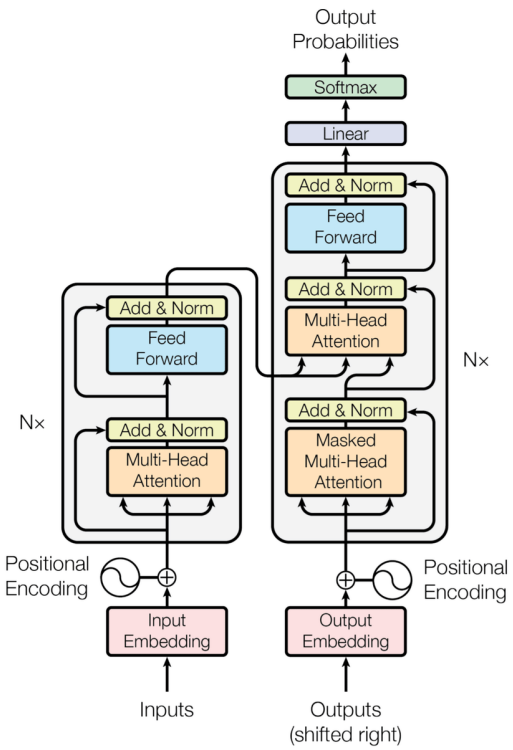
$$\text{FF}(X[t]) = \text{ReLU}(X[t]W_{\text{ff1}} + b_{\text{ff1}})W_{\text{ff2}} + b_{\text{ff2}},$$

όπου το $X[t] \in \mathbb{R}^{d_{\text{att}}}$ είναι το $t$-οστό πλαίσιο της ακολουθίας εισόδου $X$, τα $W_{\text{ff1}} \in \mathbb{R}^{d_{\text{att}} \times d_{\text{ff}}}$, $W_{\text{ff2}} \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{att}}}$ είναι μάθησιμοι πίνακες βαρών, και τα $b_{\text{ff1}} \in \mathbb{R}^{d_{\text{ff}}}$, $b_{\text{ff2}} \in \mathbb{R}^{d_{\text{att}}}$ είναι μάθησιμοι διανύσματα παραμέτρων. Αναφερόμαστε στο $MHA(X_i, X_i, X_i)$ ως "προσοχή εαυτού" (self-attention).

### Αρχιτεκτονική αποκωδικοποιητή χαρακτήρων

Ο αποκωδικοποιητής Transformer λαμβάνει την κωδικοποιημένη ακολουθία $X_e$ από την εξίσωση (5) και την προθέματα ακολουθία χαρακτήρων (π.χ., αλφάβητο, ιαπωνικοί χαρακτήρες κλπ) με αναγνωριστικά Y[1:u] = Y[1], ..., Y[u]. Στη συνέχεια, προβλέπει τα ακολουθούντα αναγνωριστικά Y[2:u+1] ως εξής:

$$\begin{aligned} E &= Embed(Y[1:u]), \\ Z_0 &= [[E[1], PE[1]]^T, ..., [E[u], PE[u]]^T], \\ Z'_j &= Z_j + MHA_{self\,j}(Z_j, Z_j, Z_j), \\ Z''_j &= Z_j + MHA_{src\,j}(Z'_j, X_e, X_e), \\ Z_{j+1} &= Z''_j + FF_j(Z''_j) \end{aligned} \tag{5.2.1}$$




Σχήμα 5.2.2: Η αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή Transformer.

όπου η $Embed(\cdot)$ είναι μια ενσωμάτωση που μετατρέπει την ακολουθία των αναγνωριστικών χαρακτήρων Y σε μια ακολουθία μάθησης διανυσμάτων που ευρίσκονται στα αναγνωριστικά$E \in \mathbb{R}^{u \times d_{att}}, j = 0, \ldots, d$ είναι ο δείκτης των επιπέδων αποκωδικοποιητή και d είναι ο αριθμός των επιπέδων αποκωδικοποιητή. Τέλος, ο αποκωδικοποιητής εκπέμπει τις υπόθετες πιθανότητες του επόμενου αναγνωριστικού χαρακτήρα $Y[u+1]$ δεδομένων των προθέματος αναγνωριστικών $Y[1], \ldots, Y[u]$ και του κωδικοποιημένου ήχου $X_e$.

### Εκπαίδευση και αποκωδικοποίηση

Το στάδιο της εκπαίδευσης, ο αποκωδικοποιητής του Transformer προβλέπει όλα τα πλαίσια χαρακτήρων ως $p_{s2s}(Y|Xe)$, όπου το $Y$ είναι μια ακολουθία πραγματικών χαρακτήρων, και υπολογίζει το σφάλμα εκπαίδευσης ως $L_{s2s} = -\log p_{s2s}(Y|Xe)$ μαζί όλα ταυτόχρονα και παράλληλα, αντίθετα από τα μοντέλα S2S που βασίζονται σε αναδρομικά νευρωνικά δίκτυα, καθώς δεν υπάρχει σειριακή λειτουργία. Η συνάρτηση $MHA(\cdot)$ αγνοεί τον κάτω τριγωνικό μέρος στον πίνακα προσοχής $X_q X_k^\top$ στην Εξίσωση (1) για να αποτρέψει την έξοδο από να επιδράσει σε μεταγενέστερες θέσεις στην ακολουθία ερωτήματος $X_q$.

Κατά τη φάση αποκωδικοποίησης, ο αποκωδικοποιητής του Transformer προβλέπει ατομικούς χαρακτήρες σειριακά χρησιμοποιώντας αναζήτηση με προθέματα παρόμοια με τα μοντέλα S2S που βασίζονται σε αναδρομικά νευρωνικά δίκτυα, για να εντοπίσει την πιο πιθανή υπόθεση: $\hat{Y} = \mathbf{argmax}_{Y \in Y^*} \log p_{s2s}(Y|Xe)$ όπου $Y^*$ είναι η έξοδος - υπόθεση.

### Συνδετική Χρονική Ταξινόμηση (Connectionist Temporal Classification - CTC)

Το CTC μαθαίνει ρητά τη μονοτονική αλληλουχία μεταξύ των χαρακτηριστικών ομιλίας και της απόδοσης των χαρακτήρων. Η κοινή εκπαίδευση με το CTC βοηθά την προσοχή του μοντέλου S2S να είναι μονοτονική, πράγμα που είναι λογικό για μια εργασία ASR. Επομένως, οδηγεί σε ταχύτερη σύγκλιση. Παρόμοια, εισάγουμε ένα νέο κλάδο CTC από την έξοδο του αποκωδικοποιητή του Transformer προς την κωδικοποιητική διαδικασία.

Πρώτον, το στρώμα CTC λαμβάνει την ακολουθία εξόδου του κωδικοποιητή, $Xe$. Στη συνέχεια, υπολογίζει

την πιθανότητα $p_{ctc}(Y|Xe)$ για μια αυθαίρετη αλληλουχία.

Η έξοδος του CTC και το $C[t, \pi[t]]$ είναι η πιθανότητα της αλληλουχίας ανάμεσα στον χαρακτήρα εξόδου $\pi[t]$ και το t-οστό καρέ στο $Xe$. Η πολλαπλή-προς-μία απεικόνιση $B(\pi)$ αφαιρεί τυχόν περιττά σύμβολα από την αλληλουχία.

Κατά την εκπαιδευτική διαδικασία, χρησιμοποιούμε πολυστρατική απώλεια (multi-task loss) [4], η οποία συνδυάζει τον αρνητικό λογαριθμικό όρο πιθανότητας από το S2S μοντέλο και τη CTC:

$$L_{mtl} = -\alpha L_{s2s} - (1-\alpha)\log p_{ctc}(Y|Xe), \quad (14)$$

όπου $\alpha$ είναι ένα υπερπαράμετρος. Χρησιμοποιούμε το δυναμικό προγραμματισμό που περιγράφεται στο [10] για τον υπολογισμό του όρου CTC, $p_{ctc}(Y|Xe)$.

#### Κοινή αποκωδικοποίηση με CTC και γλωσσικό μοντέλο (Language Model - LM)

Η CTC έχει δυνατότητα ASR όπως φαίνεται στην Εξίσωση (13), και γι' αυτό μπορούμε να εκμεταλλευτούμε τις προβλέψεις της κατά τη διάρκεια της αποκωδικοποίησης. Ακολουθούμε την κοινή προσέγγιση κοινής αποκωδικοποίησης [15], [16], που απλά παίρνει το άθροισμα των λογαριθμικών πιθανοτήτων από τα μοντέλα ως εξής:

$$\hat{Y} = \arg\max_{Y \in Y^*} \{\lambda \log p_{s2s}(Y|Xe) + (1-\lambda)\log p_{ctc}(Y|Xe) + \gamma \log p_{lm}(Y)\}, \quad (15)$$

όπου $p_{lm}(Y)$ είναι η πιθανότητα του γλωσσικού μοντέλου για την ακολουθία $Y$, και $\lambda$ και $\gamma$ είναι υπερ-παράμετροι με τα ονόματα "CTC weight" και "LM weight", αντίστοιχα. Η υλοποίησή μας βασίζεται στο [16], όπου υπολογίζεται το $\arg\max_{Y \in Y^*}$ μέσω ενός δέντρου προθέματος με αξιολόγηση κατά τη διάρκεια της εκτέλεσης.

### 5.2.3 Αποτελέσματα

Με την αρχιτεκτονική αυτή εκτελέσθηκαν 2 πειράματα. Στο πρώτο πείραμα το μοντέλο έχει εκπαιδευτεί στο σύνολο δεδομένων LibriSpeech [6]. Το LibriSpeech προέρχεται από ηχητικά βιβλία (audiobooks) που αποτελούν μέρος του προγράμματος LibriVox και περιέχει 1000 ώρες ομιλίας με δειγματοληψία στα 16 kHz. Τα audiobooks από το Project Gutenberg αποτελούν το μεγαλύτερο μέρος της συλλογής. Για την παραγωγή ενός σώματος αγγλικής αναγνωσμένης ομιλίας κατάλληλου για την εκπαίδευση συστημάτων αναγνώρισης ομιλίας, το LibriSpeech ευθυγραμμίζει και τμηματοποιεί αυτόματα την αναγνωσμένη ομιλία των audiobooks με το αντίστοιχο κείμενο του βιβλίου, φιλτράρει τα τμήματα με θορυβώδεις μεταγραφές και παράγει το σώμα. Το σώμα είναι ελεύθερα διαθέσιμο για λήψη, μαζί με ξεχωριστά προετοιμασμένα δεδομένα εκπαίδευσης γλωσσικών μοντέλων και προκατασκευασμένα γλωσσικά μοντέλα.

| **Configuration** | **Parameters** |
|---|---|
| lm_conf | nlayers: 2 |
| | unit: 650 |
| optim | sgd |
| batch_type | folded |
| batch_size | 64 |
| max_epoch | 20 |
| patience | 3 |
| best_model_criterion | - valid |
| | - loss |
| | - min |
| keep_nbest_models | 1 |

Πίνακας 5.3: Ρυθμίσεις για την εκπαίδευση του γλωσσικού μοντέλου.

Στο δεύτερο το μοντέλο εκπαιδεύεται στο σύνολο δεδομένων Dsing όπως και στα προηγούμενα πειράματα. Τα αποτελέσματα αξιολόγησης της αναγνώρισης παρουσιάζονται στον Πίνακα 5.6.

| Configuration | Parameters |
|---|---|
| encoder | transformer |
| encoder_conf | input_layer: "conv2d" |
| | num_blocks: 12 |
| | linear_units: 2048 |
| | dropout_rate: 0.1 |
| | output_size: 256 |
| | attention_heads: 4 |
| | attention_dropout_rate: 0.0 |
| decoder | transformer |
| decoder_conf | input_layer: "embed" |
| | num_blocks: 6 |
| | linear_units: 2048 |
| | dropout_rate: 0.1 |
| model_conf | ctc_weight: 0.3 |
| | lsm_weight: 0.1 |
| | length_normalized_loss: false |
| batch_type | folded |
| batch_size | 32 |
| optim | adam |
| accum_grad | 2 |
| grad_clip | 5 |
| patience | 15 |
| max_epoch | 100 |
| optim_conf | lr: 1.0 |
| scheduler | noamlr |
| scheduler_conf | warmup_steps: 25000 |
| best_model_criterion | - valid |
| | - acc |
| | - max |
| keep_nbest_models | 10 |
| init | xavier_uniform |

Πίνακας 5.4: Ρυθμίσεις για την εκπαίδευση του μοντέλου αναγνώρισης.

Βλέπουμε ότι με την βελτίωση της αρχιτεκτονικής του μοντέλου που χρησιμοποιούμε μειώνεται το σφάλμα αναγνώρισης WER.

Οι μετασχηματιστές ενσωματώνουν μηχανισμούς προσοχής που τους επιτρέπουν να εξετάζουν ταυτόχρονα ολόκληρη την ακολουθία εισόδου. Αυτή η σφαιρική κατανόηση του πλαισίου βοηθά στη σύλληψη εξαρτήσεων μεγάλης εμβέλειας και στη μοντελοποίηση πολύπλοκων σχέσεων μεταξύ διαφορετικών τμημάτων του ήχου. Αντίθετα, τα μοντέλα HMM βασίζονται συνήθως σε τοπικά παράθυρα συμφραζομένων, περιορίζοντας την ικανότητά τους να συλλαμβάνουν εκτεταμένες πληροφορίες συμφραζομένων. Ο μηχανισμός προσοχής στους μετασχηματιστές τους επιτρέπει να κατανοούν καλύτερα το πλαίσιο και να κάνουν πιο τεκμηριωμένες προβλέψεις, οδηγώντας σε βελτιωμένη ακρίβεια μεταγραφής.

Οι μετασχηματιστές προσφέρουν μια end-to-end προσέγγιση στην ASR, πράγμα που σημαίνει ότι αντιστοιχίζουν απευθείας τον ήχο εισόδου στις μεταγραφές εξόδου χωρίς να βασίζονται σε ρητά μοντέλα ευθυγράμμισης ή πολύπλοκα χειροποίητα χαρακτηριστικά. Αυτή η από άκρο σε άκρο μοντελοποίηση απλοποιεί το σύστημα ASR και αποφεύγει τη διάδοση σφαλμάτων που μπορεί να συμβεί σε συστήματα πολλαπλών σταδίων, όπως τα μοντέλα που βασίζονται σε HMM. Οι μετασχηματιστές μαθαίνουν απευθείας από το ηχητικό σήμα, επιτρέποντάς τους να εκμεταλλεύονται τα δεδομένα πιο αποτελεσματικά και να συλλαμβάνουν λεπτομερή χαρακτηριστικά, με αποτέλεσμα υψηλότερη ακρίβεια μεταγραφής.

Οι μετασχηματιστές είναι ικανοί να μαθαίνουν ισχυρές σύνθετες αναπαραστάσεις των δεδομένων ήχου και

| Configuration | Parameters |
|---|---|
| batch_size | 1 |
| beam_size | 10 |
| penalty | 0.0 |
| maxlenratio | 0.0 |
| minlenratio | 0.0 |
| ctc_weight | 0.5 |
| lm_weight | 0.3 |

Πίνακας 5.5: Ρυθμίσεις για την αποκωδικοποίηση της αναγνώρισης.

| train data | %WER | insertions | deletions | substitutions |
|---|---|---|---|---|
| librispeech | 59.27 | 460 | 446 | 1836 |
| dsing | 35.21 | 248 | 453 | 928 |

Πίνακας 5.6: Αποτελέσματα σφαλμάτων αναγνώρισης με το μοντέλο Transformer.

κειμένου μέσω πολλαπλών επιπέδων self-attention και νευρωνικών δικτύων που μπορούν να συλλάβουν τόσο τοπικά ακουστικά χαρακτηριστικά όσο και γλωσσικές πληροφορίες υψηλότερου επιπέδου. Αντίθετα, τα μοντέλα HMM συχνά βασίζονται σε τυποποιημένα χαρακτηριστικά και υποθέτουν απλουστευμένες στατιστικές υποθέσεις σχετικά με την κατανομή των δεδομένων.

Οι μετασχηματιστές έχουν επιδείξει σημαντικά οφέλη από τη μεταφορά μάθησης και την προ-εκπαίδευση μεγάλης κλίμακας σε μεγάλα σώματα δεδομένων, όπως πολύγλωσσα ή σύνολα δεδομένων γενικού τομέα. Με την αξιοποίηση της προ-εκπαίδευσης, οι μετασχηματιστές μπορούν να μάθουν χρήσιμες αναπαραστάσεις της ομιλίας και της γλώσσας από τεράστιες ποσότητες δεδομένων, οι οποίες μπορούν στη συνέχεια να ρυθμιστούν λεπτομερώς σε προβλήματα ASR συγκεκριμένων τομέων. Αυτή η εκμάθηση μεταφοράς επιτρέπει στους μετασχηματιστές να αξιοποιούν τη γνώση που αποκτάται από την προ-εκπαίδευση για τη βελτίωση της ακρίβειας μεταγραφής, ακόμη και όταν εκπαιδεύονται σε σχετικά μικρότερα σύνολα δεδομένων. Τα μοντέλα HMM, από την άλλη πλευρά, συνήθως απαιτούν περισσότερη χειροκίνητη μηχανική των χαρακτηριστικών και δεν επωφελούνται τόσο πολύ από την προεκπαίδευση μεγάλης κλίμακας.

Συνοπτικά, οι μετασχηματιστές προσφέρουν πλεονεκτήματα έναντι των μοντέλων HMM όσον αφορά τον μηχανισμό προσοχής, τη μοντελοποίηση από άκρο σε άκρο, την ικανότητα εκμάθησης αναπαραστάσεων, τις δυνατότητες εκμάθησης μεταφοράς και τη χωρητικότητα του μοντέλου. Αυτοί οι παράγοντες συμβάλλουν στην ανώτερη ακρίβεια μεταγραφής τους σε σύγκριση με τα μοντέλα HMM σε πολλές εφαρμογές ASR.

Η εκπαίδευση ενός μοντέλου ASR ειδικά σε δεδομένα τραγουδιού, αντί για δεδομένα γενικής ομιλίας, μπορεί οδηγεί σε βελτιωμένη αναγνώριση φωνών τραγουδιού, όπως φαίνεται και στον Πίνακα 5.6. Ακολουθούν ορισμένοι λόγοι που εξηγούν γιατί η εκπαίδευση του μοντέλου σε δεδομένα τραγουδιού είναι επωφελής:

Χαρακτηριστικά τραγουδιστής φωνής: Το τραγούδι περιλαμβάνει ένα ξεχωριστό σύνολο χαρακτηριστικών που το διαφοροποιούν από την κανονική ομιλία. Αυτά τα χαρακτηριστικά περιλαμβάνουν διακυμάνσεις του τόνου, μελωδική φρασεολογία, βιμπράτο, παρατεταμένες νότες, φωνητικές διακοσμήσεις και πολλά άλλα. Με την εκπαίδευση του μοντέλου σε δεδομένα τραγουδιού, μπορεί να μάθει ειδικά να αναγνωρίζει και να ερμηνεύει αυτά τα μοναδικά χαρακτηριστικά, τα οποία μπορεί να μην είναι τόσο εμφανή ή σημαντικά στη γενική ομιλία.

Ειδικό λεξιλόγιο τομέα: Το τραγούδι συχνά ενσωματώνει ένα ειδικό λεξιλόγιο, όπως στίχους, τίτλους τραγουδιών, μουσική ορολογία και ονόματα καλλιτεχνών. Αυτή η στοχευμένη εκπαίδευση βελτιώνει την ακρίβεια και την ακρίβεια της αναγνώρισης λέξεων που σχετίζονται με το τραγούδι, οδηγώντας σε πιο αξιόπιστες μεταγραφές και αποτελέσματα υψηλότερης ποιότητας.

Διαφορετικά μοτίβα προφοράς: Το τραγούδι μπορεί να εισάγει διαφοροποιήσεις στην προφορά σε σύγκριση με την κανονική ομιλία. Αυτές οι διαφοροποιήσεις μπορεί να προκύψουν λόγω της ανάγκης να ταιριάξουν με τη μελωδία, να ακολουθήσουν το ρυθμό ή να τονίσουν ορισμένες συλλαβές στους στίχους. Η εκπαίδευση του μοντέλου σε δεδομένα τραγουδιού το εκθέτει σε ένα ευρύτερο φάσμα μοτίβων προφοράς ειδικά για το

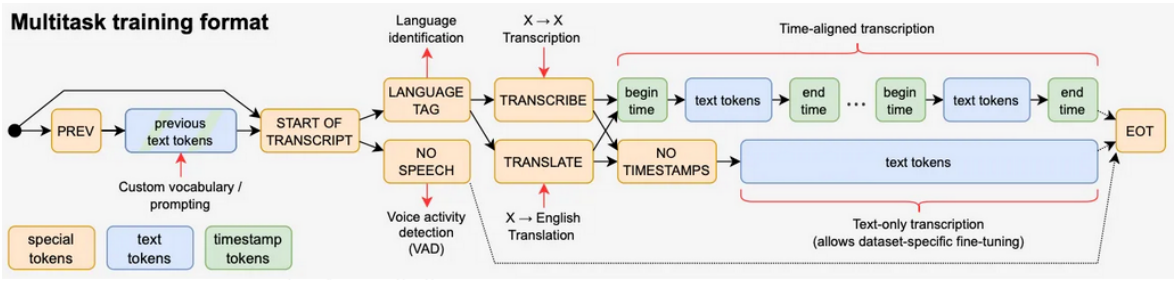
τραγούδι, βοηθώντας το να μάθει και να προσαρμοστεί σε αυτές τις παραλλαγές.

Μειωμένη παρεμβολή από τη μουσική υπόκρουση: Σε πολλά σενάρια τραγουδιού, υπάρχει συχνά μουσική υπόκρουση που συνοδεύει τα φωνητικά. Με την εκπαίδευση του μοντέλου σε δεδομένα τραγουδιού, γίνεται πιο ανθεκτικό στην παρουσία μουσικής υποβάθρου και μαθαίνει να εστιάζει στα φωνητικά στοιχεία, διαχωρίζοντάς τα αποτελεσματικά από τη μουσική συνοδεία, βελτιώνει την ικανότητα του μοντέλου να διακρίνει και να αναγνωρίζει τη φωνή του τραγουδιού μέσα στο μουσικό πλαίσιο, με αποτέλεσμα πιο ακριβείς μεταγραφές.

## 5.3 Το μοντέλο Whisper

Το Whisper [30] είναι ένα μοντέλο τελευταίας τεχνολογίας (SOTA) αυτόματης αναγνώρισης ομιλίας (ASR) που εκπαιδεύτηκε σε 680.000 ώρες πολύγλωσσων και πολλαπλών εργασιών δεδομένων που συλλέχθηκαν από το διαδίκτυο. Η χρήση ενός τόσο μεγάλου και ποικίλου συνόλου δεδομένων οδηγεί σε βελτιωμένη ανθεκτικότητα σε προφορές, θόρυβο υποβάθρου και τεχνική γλώσσα. [31] Επιπλέον, επιτρέπει τη μεταγραφή σε πολλές γλώσσες, καθώς και τη μετάφραση από τις γλώσσες αυτές στα αγγλικά. Τα μοντέλα και ο κώδικας εξαγωγής συμπερασμάτων είναι ανοικτά για να χρησιμεύσουν ως βάση για περαιτέρω έρευνα σχετικά με την εύρωστη επεξεργασία ομιλίας. Το μοντέλο whisper αν και πολύ πρόσφατο, έχει χρησιμοποιηθεί ήδη από την βιβλιογραφία για την επίλυση διαφόρων παρεμφερών προβλημάτων όπως στα [32], [33].

Αν και η πρόβλεψη των λέξεων που εκφωνήθηκαν σε ένα δεδομένο ηχητικό απόσπασμα αποτελεί βασικό μέρος του πλήρους προβλήματος της αναγνώρισης ομιλίας και έχει μελετηθεί εκτενώς στην έρευνα, δεν είναι το μοναδικό μέρος του Whisper. Το συγκεκριμένο μοντέλο αποτελεί ένα πλήρως εξοπλισμένο σύστημα αναγνώρισης ομιλίας αφού περιλαμβάνει πολλά πρόσθετα στοιχεία, όπως η ανίχνευση φωνητικής δραστηριότητας, η καταγραφή ομιλητή και την κανονικοποίηση κειμένου. Αυτά τα στοιχεία συχνά αντιμετωπίζονται χωριστά, με αποτέλεσμα ένα σχετικά πολύπλοκο σύστημα γύρω από το κύριο μοντέλο αναγνώρισης ομιλίας. Για να μειωθεί αυτή η πολυπλοκότητα, στόχος του Whisper είναι να έχει ένα μόνο μοντέλο που να εκτελεί ολόκληρη την επεξεργασίας ομιλίας, όχι μόνο το βασικό τμήμα αναγνώρισης. Ένα σημαντικό στοιχείο εδώ είναι η διεπαφή για το μοντέλο. Υπάρχουν πολλές διαφορετικές εργασίες που μπορούν να εκτελεστούν στο το ίδιο ηχητικό σήμα εισόδου: μεταγραφή, μετάφραση, φωνητική ανίχνευση δραστηριότητας, ευθυγράμμιση και αναγνώριση γλώσσας είναι μερικά παραδείγματα.




Σχήμα 5.3.1: Εκπαίδευση του multitask μοντέλου Whisper.

Η προσέγγιση για το whisper ήταν να εκπαιδευτεί σε όσο το δυνατόν περισσότερες ώρες ομιλίας από το διαδίκτυο, όσο το δυνατόν πιο διαφορετικές. Το Whisper εκτελεί συμπερασμό εκτός κατανομής όταν δοκιμάζεται στο σύνολο δοκιμών Librispeech [6]. Η ικανότητα του Whisper να γενικεύει σε διαφορετικά σύνολα δεδομένων και τομείς οφείλεται στον τεράστιο όγκο δεδομένων με διαφορετική κατανομή. Είναι ίσο με 77 χρόνια ακρόασης.

Το Whisper διαφέρει από τα προηγούμενα μοντέλα με διάφορους τρόπους:

- Τεράστια ποσότητα δεδομένων εκπαίδευσης: Το Whisper εκπαιδεύτηκε σε μεγάλο όγκο δεδομένων, τα οποία θα συζητηθούν λεπτομερέστερα στη συνέχεια.
- Πλήρως εποπτευόμενη εκπαίδευση: Σε αντίθεση με ορισμένα προηγούμενα μοντέλα SOTA, όπως το Wav2Vec2 [22], τα οποία χρησιμοποιούσαν αυτοεπίβλεψη για την εκπαίδευση, το Whisper εκπαιδεύτηκε με πλήρη επίβλεψη.
- Πλήρης υποδομή ASR: Το Whisper εκπαιδεύτηκε για να χειρίζεται πολλαπλές εργασίες που απαιτούνται για ένα σύστημα αυτόματης αναγνώρισης ομιλίας (ASR) μέσα σε ένα μόνο μοντέλο. Περιλαμβάνει ανίχνευση φωνητικής δραστηριότητας, ανίχνευση γλώσσας, μετατροπή ομιλίας σε κείμενο, μετάφραση και μερική ευθυγράμμιση για 96 διαφορετικές γλώσσες.

Το σύνολο δεδομένων εκπαίδευσης για το Whisper κατασκευάστηκε με τη χρήση ήχου σε συνδυασμό με μεταγραφές που προέρχονται από το Διαδίκτυο. Αυτή η προσέγγιση οδήγησε σε ένα ποικιλόμορφο σύνολο

δεδομένων που περιλαμβάνει ένα ευρύ φάσμα δειγμάτων ήχου από διάφορα περιβάλλοντα, ρυθμίσεις ηχογράφησης, ομιλητές και γλώσσες. Ενώ η ποικιλομορφία του ήχου βοηθάει στην εκπαίδευση ενός ισχυρού μοντέλου, η ποιότητα των μεταγραφών αποδείχθηκε κρίσιμη. Η αρχική ανάλυση αποκάλυψε έναν σημαντικό αριθμό υποβαθμισμένων μεταγραφών στο ακατέργαστο σύνολο δεδομένων. Για την αντιμετώπιση αυτού του ζητήματος, αναπτύχθηκαν διάφορες αυτοματοποιημένες μέθοδοι φιλτραρίσματος για τη βελτίωση της ποιότητας των μεταγραφών.

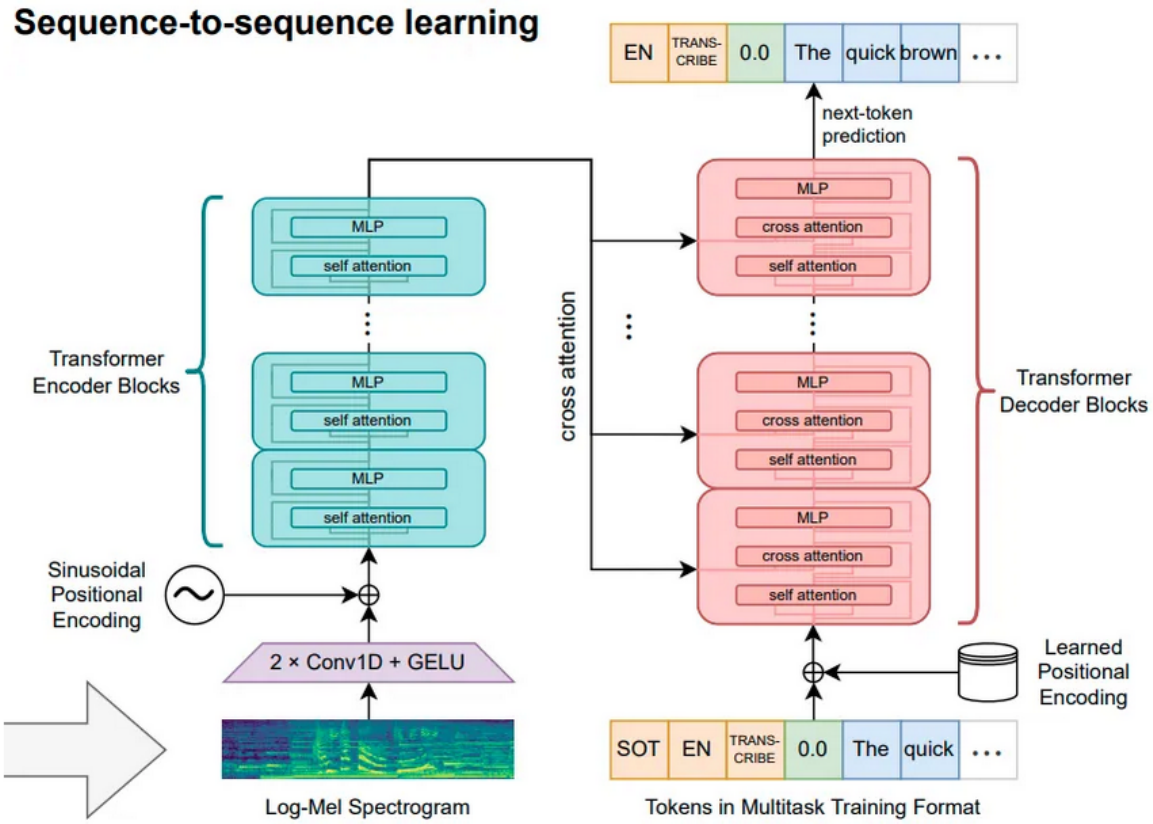
Αξίζει να σημειωθεί ότι πολλά κείμενα που διατίθενται στο διαδίκτυο δεν παράγονται από τον άνθρωπο, αλλά παράγονται από υπάρχοντα συστήματα ASR. Πρόσφατες έρευνες έχουν δείξει ότι η εκπαίδευση σε μικτά σύνολα δεδομένων που περιλαμβάνουν τόσο δεδομένα που παράγονται από τον άνθρωπο όσο και δεδομένα που παράγονται από μηχανήματα μπορεί να επηρεάσει αρνητικά την απόδοση των συστημάτων μετάφρασης. Ως εκ τούτου, καταβλήθηκαν προσπάθειες για τον εντοπισμό και την εξάλειψη των μεταγραφών που παράγονται από μηχανήματα από το σύνολο δεδομένων εκπαίδευσης. Χρησιμοποιήθηκαν διάφορες ευρετικές μέθοδοι για τον εντοπισμό χαρακτηριστικών που είναι ενδεικτικά των μεταγραφών που παράγονται από μηχανήματα. Για παράδειγμα, τα μεταγράμματα που παράγονται από μηχανήματα συχνά στερούνται σύνθετων σημείων στίξης, κενών διαστημάτων μορφοποίησης ή υφολογικών πτυχών και μπορεί να είναι σταθερά γραμμένα με κεφαλαία ή πεζά γράμματα.

Επιπλέον, χρησιμοποιήθηκε ένας ανιχνευτής γλώσσας ήχου για να διασφαλιστεί ότι η προφορική γλώσσα ταιριάζει με τη γλώσσα του αντίστοιχου κειμένου. Με τη λεπτομερή ρύθμιση ενός πρωτότυπου μοντέλου σε ένα σύνολο δεδομένων που ονομάζεται VoxLingua107 , ο ανιχνευτής γλώσσας ήταν σε θέση να εντοπίσει τα αναντίστοιχα ζεύγη γλωσσών. Εάν η ομιλούμενη γλώσσα και η γλώσσα του μεταγράφου δεν ταίριαζαν σύμφωνα με τον CLD2 (Compact Language Detector 2), το ζεύγος (ήχου, κειμένου) δεν συμπεριλαμβανόταν ως παράδειγμα εκπαίδευσης για την αναγνώριση ομιλίας, εκτός από τις περιπτώσεις όπου η γλώσσα του κειμένου ήταν η αγγλική. Σε αυτές τις περιπτώσεις, τα ζεύγη προστέθηκαν στο σύνολο δεδομένων ως παραδείγματα εκπαίδευσης μετάφρασης ομιλίας από τη γλώσσα X στα αγγλικά. Εφαρμόστηκαν τεχνικές ασαφούς αντιγραφής για τη μείωση της επανάληψης και του αυτόματα παραγόμενου περιεχομένου στο σύνολο δεδομένων εκπαίδευσης.

Μετά την εκπαίδευση ενός αρχικού μοντέλου, πραγματοποιήθηκε ένα πρόσθετο πέρασμα φιλτραρίσματος. Συγκεντρώθηκαν πληροφορίες σχετικά με το ποσοστό σφάλματος του μοντέλου σε πηγές δεδομένων εκπαίδευσης και πραγματοποιήθηκε χειροκίνητη επιθεώρηση σε αυτές τις πηγές. Η διαδικασία επιθεώρησης έδωσε προτεραιότητα στις πηγές δεδομένων με υψηλά ποσοστά σφάλματος και μεγαλύτερα μεγέθη για τον αποτελεσματικό εντοπισμό και την αφαίρεση πηγών χαμηλής ποιότητας. Η επιθεώρηση αποκάλυψε μερικώς απομαγνητοφωνημένα ή ανεπαρκώς ευθυγραμμισμένα/απομαγνητοφωνημένα αντίγραφα, καθώς και εναπομείνασες χαμηλής ποιότητας μηχανικά παραγόμενες λεζάντες που δεν εντοπίστηκαν από τις ευρετικές λειτουργίες φιλτραρίσματος.

### 5.3.1 Αρχιτεκτονική του μοντέλου Whisper

Η αρχιτεκτονική του μοντέλου Whisper αποτελείται από έναν Transformer κωδικοποιητή-αποκωδικοποιητή, καθώς αυτή η αρχιτεκτονική έχει επικυρωθεί για την αξιόπιστη κλιμάκωσή της. Όλος ο ήχος αναδειγματοληπτείται σε 16.000 Hz και υπολογίζεται μια αναπαράσταση log Mel φασματογραφήματος 80 καναλιών σε παράθυρα 25 χιλιοστών του δευτερολέπτου με βήμα 10 χιλιοστών του δευτερολέπτου. Για την κανονικοποίηση των χαρακτηριστικών, αναπροσαρμόζουμε συνολικά την είσοδο ώστε να είναι μεταξύ -1 και 1 με περίπου μηδενική μέση τιμή στο σύνολο δεδομένων προ-εκπαίδευσης. Ο κωδικοποιητής επεξεργάζεται αυτή την αναπαράσταση εισόδου με ένα μικρό στέλεχος που αποτελείται από δύο στρώματα συνέλιξης με πλάτος φίλτρου 3 και τη συνάρτηση ενεργοποίησης GELU, όπου το δεύτερο στρώμα συνέλιξης έχει βήμα δύο. Στη συνέχεια, προστίθενται ημιτονοειδή embeddings θέσης στην έξοδο του στελέχους και στη συνέχεια εφαρμόζονται τα μπλοκ Transformer του κωδικοποιητή. Ο Transformer χρησιμοποιεί residual blocks πριν από την ενεργοποίηση , και μια τελική κανονικοποίηση στρώματος εφαρμόζεται στην έξοδο του κωδικοποιητή. Ο αποκωδικοποιητής χρησιμοποιεί εκπαιδευμένα position embeddings και συνδεδεμένες αναπαραστάσεις συμβόλων εισόδου-εξόδου. Ο κωδικοποιητής και ο αποκωδικοποιητής έχουν το ίδιο πλάτος και τον ίδιο αριθμό μπλοκ Transformer. Η αρχιτεκτονική του μοντέλου απεικονίζεται στο Σχήμα 5.3.2




Σχήμα 5.3.2: Η αρχιτεκτονική του μοντέλου Whisper.

### 5.3.2 Σχεδιασμός πειραμάτων και Αποτελέσματα

**Το Hugging Face**

Το Hugging Face Hub [34] είναι μια πλατφόρμα με πάνω από 120 χιλιάδες μοντέλα, 20 χιλιάδες σύνολα δεδομένων και 50 χιλιάδες εφαρμογές επίδειξης (Spaces), όλα ανοιχτού κώδικα και δημόσια διαθέσιμα, σε μια διαδικτυακή πλατφόρμα όπου οι άνθρωποι μπορούν εύκολα να συνεργαστούν και να δημιουργήσουν μαζί εφαρμογές Μηχανικής Μάθησης. Είναι πιο γνωστή για τη βιβλιοθήκη μετασχηματιστών (Transformers) που έχει κατασκευαστεί για εφαρμογές επεξεργασίας φυσικής γλώσσας και την πλατφόρμα που επιτρέπει στους χρήστες να μοιράζονται μοντέλα μηχανικής μάθησης και σύνολα δεδομένων. Το Hub λειτουργεί ως ένα κεντρικό σημείο όπου ο καθένας μπορεί να εξερευνήσει, να πειραματιστεί, να συνεργαστεί και να δημιουργήσει τεχνολογία με τη Μηχανική Μάθηση.

Το Hugging Face Hub ως πλατφόρμα (κεντρική διαδικτυακή υπηρεσία) φιλοξενεί:

- Αποθετήρια κώδικα με βάση το Git, με χαρακτηριστικά παρόμοια με το GitHub, συμπεριλαμβανομένων των συζητήσεων και των αιτημάτων pull για έργα.
- Μοντέλα, επίσης με έλεγχο έκδοσης βασισμένο στο Git,
- Σύνολα δεδομένων, κυρίως σε κείμενο, εικόνες και ήχο,
- Διαδικτυακές εφαρμογές ("spaces" και "widgets"), που προορίζονται για μικρής κλίμακας επιδείξεις εφαρμογών μηχανικής μάθησης.

Εκτός από τους Transformers και το Hugging Face Hub, το οικοσύστημα Hugging Face περιέχει βιβλιοθήκες για άλλες εργασίες, όπως επεξεργασία συνόλων δεδομένων ("Datasets"), αξιολόγηση μοντέλων

| Model | Layers | Width | Heads | Parameters |
|---|---|---|---|---|
| Tiny | 4 | 384 | 6 | 39M |
| Base | 6 | 512 | 8 | 74M |
| Small | 12 | 768 | 12 | 244M |
| Medium | 24 | 1024 | 16 | 769M |
| Large | 32 | 1280 | 20 | 1550M |

Πίνακας 5.7: Στοιχεία αρχιτεκτονικής της σειράς μοντέλων Whisper.

| Hyperparameter | Value |
|---|---|
| Updates | 1048576 |
| Batch Size | 256 |
| Warmup Updates | 2048 |
| Max grad norm | 1.0 |
| Optimizer | AdamW |
| $\beta_1$ | 0.9 |
| $\beta_2$ | 0.98 |
| $\epsilon$ | $10^{-6}$ |
| Weight Decay | 0.1 |
| Weight Init | Gaussian Fan-In |
| Learning Rate Schedule | Linear Decay |
| Speechless audio subsample factor | 10× |
| Condition on prior text rate | 50% |

Πίνακας 5.8: Υπερπαράμετροι εκπαίδευσης του Whisper.

("Evaluate"), προσομοίωση ("Simulate"), επιδείξεις μηχανικής μάθησης ("Gradio").

### Εκπαίδευση του μοντέλου Whisper

Μια σειρά μοντέλων διαφόρων μεγεθών εκπαιδεύτηκαν προκειμένου να μελετηθούν οι ιδιότητες κλιμάκωσης του Whisper, τα μεγέθη των οποίν παρουσιάζονται στον Πίνακα 5.7. Τα μοντέλα εκπαιδεύτηκαν με παραλληλία δεδομένων σε όλους τους επιταχυντές χρησιμοποιώντας FP16 με dynamic loss scaling και activation checkpointing. Η εκπαίδευση έγινε με το AdamW και gradient norm clipping με γραμμική μείωση του ρυθμού μάθησης στο μηδέν μετά από μια περίοδο προθέρμανσης κατά τις πρώτες 2048 ενημερώσεις. Χρησιμοποιήθηκε ένα μέγεθος batch 256 τμημάτων και τα μοντέλα εκπαιδεύονται για 220 ενημερώσεις, δηλαδή για δύο έως τρία περάσματα πάνω από το σύνολο δεδομένων. Λόγω της εκπαίδευσης μόνο για λίγες εποχές, το over-fitting δεν αποτελεί ανησυχία και δεν χρησιμοποιήθηκε καμία επαύξηση δεδομένων ή κανονικοποίηση, αντ' αυτού βασιζόμαστε στην ποικιλομορφία που περιέχεται σε ένα τόσο μεγάλο σύνολο δεδομένων για να ενθαρρύνουμε τη γενίκευση και την ευρωστία.

### Κλιμάκωση μοντέλου

Ένα μεγάλο μέρος της υπόσχεσης των προσεγγίσεων εκπαίδευσης με ασθενή επίβλεψη είναι η δυνατότητά τους να χρησιμοποιούν σύνολα δεδομένων πολύ μεγαλύτερα από εκείνα της παραδοσιακής μάθησης με επίβλεψη. Ωστόσο, αυτό συνεπάγεται το κόστος της χρήσης δεδομένων που είναι πιθανώς πολύ πιο θορυβώδη και χαμηλότερης ποιότητας από τα δεδομένα που χρησιμοποιούν τα χρυσά πρότυπα επίβλεψης. Μια ανησυχία με αυτή την προσέγγιση είναι ότι, αν και μπορεί να φαίνεται αρχικά πολλά υποσχόμενη, η απόδοση των μοντέλων που εκπαιδεύονται σε αυτού του είδους τα δεδομένα μπορεί να κορεστεί στο εγγενές επίπεδο ποιότητας του συνόλου δεδομένων, το οποίο μπορεί να είναι πολύ κάτω από το ανθρώπινο επίπεδο. Μια συναφής ανησυχία είναι ότι, καθώς αυξάνεται η χωρητικότητα και ο υπολογισμός που δαπανάται για την εκπαίδευση στο σύνολο δεδομένων, τα μοντέλα μπορεί να μάθουν να εκμεταλλεύονται τις ιδιαιτερότητες του συνόλου δεδομένων και η ικανότητά τους να γενικεύουν ισχυρά σε δεδομένα εκτός διανομής μπορεί ακόμη και να υποβαθμιστεί.

| Model | Max Learning Rate |
|---|---|
| Tiny | $1.5 \times 10^{-3}$ |
| Base | $1 \times 10^{-3}$ |
| Small | $5 \times 10^{-4}$ |
| Medium | $2.5 \times 10^{-4}$ |
| Large-v2 | $2 \times 10^{-4}$ |

Πίνακας 5.9: Ρυθμοί εκμάθησης των μοντέλων Whisper.

| Model | parameters ($\times 10^6$) | multilingual(%WER) | english(%WER) |
|---|---|---|---|
| Tiny | 37 | 45.05 | 34.34 |
| Base | 71 | 35.43 | 30.29 |
| Small | 240 | 16.52 | 15.34 |
| Medium | 762 | 14.14 | 17.01 |
| Large | 1,541 | 10.80 | - |

Πίνακας 5.10: Αποτελέσματα word error rate (WER) των μοντέλων Whisper.

## 5.4 Συγκεντρωτική παρουσίαση των αποτελεσμάτων και Σχόλια

Στον Πίνακα 5.11 παρουσιάζονται συγκεντρωτικά τα ποσοστά σφαλμάτων λέξης (WER) που παρουσίασαν τα μοντέλα που εξετάσαμε σε δεδομένο σύνολο δοκιμής.

Το baseline μοντέλο αποτελεί ένα μοντέλο HMM-GMM που υλοποιήθηκε με τη χρήση της εργαλειοθήκης Kaldi [24]. Το μοντέλο χρησιμοποιούσε προκαθορισμένα βασικά χαρακτηριστικά ως είσοδο. Τα προκαταρκτικά πειράματα με αυτό το μοντέλο αποκάλυψαν ότι η αύξηση του όγκου των δεδομένων εκπαίδευσης βελτίωσε την ποιότητα της μεταγραφής και μείωσε τα σφάλματα αναγνώρισης λέξεων (WER). Η βελτίωση αυτή προήλθε από την ικανότητα του μοντέλου να μαθαίνει καλύτερα τις αντιστοιχίες ήχου-φωνής αναλύοντας μια πιο ολοκληρωμένη αναπαράσταση των φωνών των τραγουδιών.

Η αρχιτεκτονική Transformer ξεπέρασε τα μοντέλα HMM-GMM για τη μεταγραφή στίχων, ακόμη και όταν εκπαιδεύτηκε αποκλειστικά σε δεδομένα ομιλίας. Οι αρχιτεκτονικές Transformer προσέφεραν διακριτά πλεονεκτήματα έναντι των μοντέλων HMM, συμπεριλαμβανομένων των μηχανισμών προσοχής, της από άκρη σε άκρη μοντελοποίησης, της ικανότητας μάθησης αναπαράστασης και της αυξημένης χωρητικότητας του μοντέλου. Με τη χρήση της αρχιτεκτονικής κωδικοποιητή-αποκωδικοποιητή και του μηχανισμού προσοχής, το μοντέλο απέκτησε σημαντικά πιο ισχυρές αναπαραστάσεις που είχαν επίγνωση του πλαισίου όταν του δόθηκε ως είσοδος το ακατέργαστο αρχείο ήχου.

Επιπλέον, εστιάζοντας ειδικά την εκπαίδευση του μοντέλου Transformer σε δεδομένα τραγουδιών, οι επιδόσεις αναγνώρισης μπορούσαν να αξιοποιήσουν τα μοναδικά χαρακτηριστικά, το λεξιλόγιο, τα μοτίβα εκφώνησης και τις προκλήσεις που ενυπάρχουν στις φωνές τραγουδιού. Αυτή η εξειδικευμένη εκπαίδευση διευκόλυνε την καλύτερη κατανόηση και ερμηνεία από το μοντέλο των χαρακτηριστικών που σχετίζονται με το τραγούδι.

Σύμφωνα με την έννοια της μεταφοράς μάθησης, χρησιμοποιήθηκε επίσης το μοντέλο αναγνώρισης Whisper τελευταίας τεχνολογίας, το οποίο βασίζεται στο Transformers. Το Whisper, το μοντέλο ASR της OpenAI, ξεπέρασε τα προηγούμενα μοντέλα στην αυτόματη αναγνώριση ομιλίας. Εκπαιδευμένο σε 680K ώρες πολύγλωσσων δεδομένων, πέτυχε υψηλή ακρίβεια και χαμηλό ποσοστό σφάλματος λέξης (WER). Παρατηρήθηκε ότι η κλιμάκωση των παραμέτρων της μεγαλύτερης έκδοσης του μοντέλου Whisper απέδωσε βελτιωμένη απόδοση σε σύγκριση με το βασικό μοντέλο. Ενώ αυτά τα μεγαλύτερα μοντέλα προσέφεραν αυξημένη ακρίβεια, απαιτούσαν περισσότερους υπολογιστικούς πόρους κατά τη διάρκεια της εξαγωγής συμπερασμάτων.

| **model** | **trainset** | **%WER** |
|---|---|---|
| HMM-GMM | DSing1 | 61.69 |
| HMM-GMM | DSing3 | 54.39 |
| HMM-GMM | DSing30 | 51.41 |
| Transformers | LibriSpeech | 59.27 |
| Transformers | DSing30 | 35.21 |
| Whisper tiny | 680k hours | 45.05 |
| Whisper tiny.en | 680k hours | 34.34 |
| Whisper base | 680k hours | 35.43 |
| Whisper base.en | 680k hours | 30.29 |
| Whisper small | 680k hours | 16.52 |
| Whisper small.en | 680k hours | 15.34 |
| Whisper medium | 680k hours | 14.14 |
| Whisper medium.en | 680k hours | 17.01 |
| Whisper large | 680k hours | 10.80 |

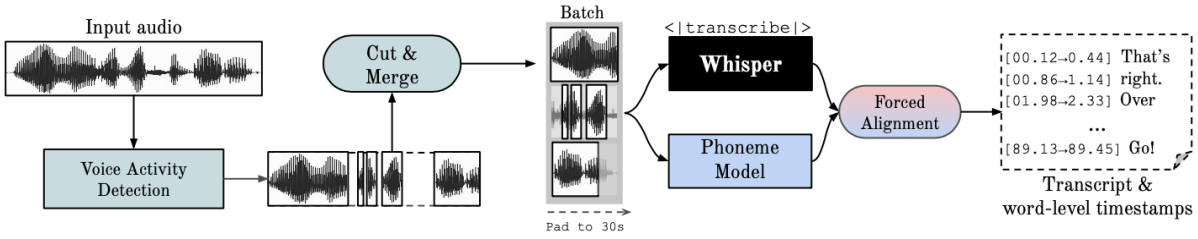
Πίνακας 5.11: Συγκεντρωτική παρουσίαση των αποτελεσμάτων για το πρόβλημα της μεταγραφής.

# Κεφάλαιο 6

# Χρονική Ευθυγράμμιση Στίχων

## 6.1 WhisperX

Το WhisperX [35] είναι ένα σύστημα για την αποτελεσματική μεταγραφή ομιλίας ήχου μεγάλης διάρκειας με ακριβείς χρονοσφραγίδες σε επίπεδο λέξης. Αποτελείται από τρία πρόσθετα στάδια στη διαδικασία της μεταγραφής του Whisper: (i) προ-τμηματοποίηση του ήχου εισόδου με ένα εξωτερικό μοντέλο ανίχνευσης φωνητικής δραστηριότητας (VAD) (ii) αποκοπή και συγχώνευση των τμημάτων VAD που προκύπτουν σε κομμάτια εισόδου διάρκειας περίπου 30 δευτερολέπτων με όρια που βρίσκονται σε ελάχιστα ενεργές περιοχές ομιλίας, επιτρέποντας την ομαδοποιημένη μεταγραφή του Whisper [30] - και τέλος (iii) αναγκαστική ευθυγράμμιση με ένα εξωτερικό μοντέλο φωνημάτων για την παροχή ακριβών χρονοσφραγίδων σε επίπεδο λέξης. Η αρχιτεκτονική του φαίνεται στο Σχήμα 6.1.1




Σχήμα 6.1.1: Η αρχιτεκτνική του μοντέλου WhisperX.

### Ανίχνευση φωνητικής δραστηριότητας (VAD)

Πριν από τη μεταγραφή και την ευθυγράμμιση, ο ήχος τμηματοποιείται εκ των προτέρων με τη χρήση VAD. Αυτό έχει δύο πλεονεκτήματα. Πρώτον, επιτρέπει τον διαχωρισμό του ήχου σε κομμάτια που δεν περιέχουν ενεργές περιοχές ομιλίας, μειώνοντας τα σφάλματα που προκαλούνται από φαινόμενα ορίων και επιτρέποντας την τμηματική μεταγραφή. Δεύτερον, τα χρονικά όρια κάθε τμήματος μπορούν να χρησιμοποιηθούν για τον περιορισμό της ευθυγράμμισης σε τοπικά τμήματα, εξαλείφοντας την εξάρτηση από αναξιόπιστες χρονοσφραγίδες από το μοντέλο Whisper. Δεδομένης μιας ηχητικής κυματομορφής $A$, το VAD παράγει έναν κατάλογο από $N$ μη επικαλυπτόμενα τμήματα $S$, όπου κάθε τμήμα αντιστοιχεί στους χρόνους έναρξης και λήξης των ενεργών περιοχών ομιλίας $S_i = (t_{i0}, t_{i1})$.

### Αποκοπή και συγχώνευση VAD

Τα τμήματα VAD $S$ μπορεί να έχουν διαφορετικό μήκος, μικρότερο ή μεγαλύτερο από τη διάρκεια εισόδου στην οποία εκπαιδεύτηκε το μοντέλο ASR (Whisper) $|A_{train}| = 30$ δευτερόλεπτα). Ενώ η αρχιτεκτονική

του Transformer μπορεί να χειριστεί ακολουθίες αυθαίρετου μήκους, η λειτουργία προσοχής attention κλιμακώνεται με το τετράγωνο του μήκους εισόδου, με αποτέλεσμα υψηλή κατανάλωση μνήμης για μεγάλα τμήματα χωρίς ανώτατο όριο διάρκειας. Για να αντιμετωπιστεί αυτό, προτείνεται μια λειτουργία min-cut. Τμήματα μεγαλύτερα από το $|A_{train}|$ διαιρούνται στο σημείο με τη χαμηλότερη βαθμολογία ενεργοποίησης φωνής από το μοντέλο VAD, διασφαλίζοντας ότι τα νέα διαιρεμένα τμήματα δεν υπάρχουν σε όρια λέξεων και ελαχιστοποιώντας τα σφάλματα ορίων. Επιπλέον, τα πολύ μικρά τμήματα δημιουργούν τις δικές τους προκλήσεις, καθώς δεν διαθέτουν επαρκές πλαίσιο για την ακριβή μοντελοποίηση της ομιλίας και η μεταγραφή πολλών μικρότερων τμημάτων αυξάνει το συνολικό χρόνο μεταγραφής. Για τον μετριασμό αυτών των ζητημάτων, προτείνεται μια λειτουργία συγχώνευσης. Γειτονικά σύντομα τμήματα ομιλίας συγχωνεύονται έως ότου η συνολική τους διάρκεια δεν είναι μεγαλύτερη από το μέγεθος εισόδου του μοντέλου μεταγραφής $\tau = |A_{train}|$. Αυτό παρέχει το μεγαλύτερο δυνατό πλαίσιο κατά τη διάρκεια της μεταγραφής και διατηρεί μια παρόμοια κατανομή δεδομένων όπως αυτή που παρατηρήθηκε κατά την εκπαίδευση.

### Μεταγραφή με το Whisper

Τα προκύπτοντα τμήματα ομιλίας $S_i$, τα οποία έχουν τώρα διάρκεια περίπου ίση με το μέγεθος εισόδου του μοντέλου $|S_i| \approx |A_{train}|$ για όλα τα $i \in N$ και όρια που δεν συμπίπτουν με ενεργό ομιλία, μπορούν να μεταγραφούν αποτελεσματικά χρησιμοποιώντας το μοντέλο Whisper ASR. Κάθε τμήμα $S_i$ μεταγράφεται ανεξάρτητα, χωρίς να εξαρτάται από προηγούμενο κείμενο, καθώς αυτό θα παραβίαζε την υπόθεση ανεξαρτησίας κάθε δείγματος στη δέσμη. Η μεταγραφή χωρίς κλιμάκωση σε προηγούμενο κείμενο συμβάλλει στη βελτίωση της ευρωστίας μειώνοντας την πιθανότητα ψευδαισθήσεων.

### Αναγκαστική ευθυγράμμιση φωνημάτων

Ο στόχος σε αυτό το βήμα είναι να εκτιμηθούν οι χρόνοι έναρξης και λήξης κάθε λέξης σε ένα τμήμα $S_i$ και η αντίστοιχη μεταγραφή κειμένου $T_i$, η οποία αποτελείται από μια ακολουθία λέξεων $T_i = [w_0, w_1, ..., w_m]$. Για να επιτευχθεί αυτό, χρησιμοποιείται ένα μοντέλο αναγνώρισης φωνημάτων. Αυτό το μοντέλο εκπαιδεύεται για να ταξινομεί τη μικρότερη μονάδα ομιλίας που διακρίνει μια λέξη από μια άλλη, όπως το φώνημα "p" στη λέξη "tap". Ο ταξινομητής φωνημάτων δέχεται ως είσοδο ένα τμήμα ήχου $S$ και παράγει έναν πίνακα logits $L$ μεγέθους $RK \times T$, όπου το $T$ ποικίλλει με βάση τη χρονική ανάλυση του μοντέλου φωνημάτων. Η διαδικασία ευθυγράμμισης περιλαμβάνει τα ακόλουθα βήματα: 1) εκτέλεση ταξινόμησης φωνημάτων στο τμήμα εισόδου $S_i$ χρησιμοποιώντας ένα περιορισμένο σύνολο κλάσεων φωνημάτων $C'$ που αντιστοιχούν στα φωνήματα στη μεταγραφή του τμήματος $T_i$, 2) εφαρμογή δυναμικής χρονικής στρέβλωσης (DTW) στον προκύπτοντα πίνακα logits $L_i$ για την εύρεση της βέλτιστης χρονικής διαδρομής των φωνημάτων στο $T_i$, και 3) λήψη των χρόνων έναρξης και λήξης για κάθε λέξη $w_i$ στο $T_i$ λαμβάνοντας τους χρόνους έναρξης και λήξης του πρώτου και του τελευταίου φωνήματος, αντίστοιχα. Εάν ένας χαρακτήρας στο κείμενο δεν υπάρχει στο λεξικό φωνημάτων $C$, αποδίδεται η χρονική στιγμή από το αμέσως επόμενο πλησιέστερο φώνημα στο κείμενο. Αυτή η διαδικασία μπορεί να παραλληλιστεί για αποτελεσματική μεταγραφή και ευθυγράμμιση λέξεων σε ήχο μεγάλης διάρκειας.

### Πολύγλωσση μεταγραφή και ευθυγράμμιση

Το WhisperX μπορεί επίσης να χρησιμοποιηθεί για πολύγλωσση μεταγραφή. Για να επιτευχθεί αυτό, το μοντέλο VAD θα πρέπει να είναι ανθεκτικό σε διαφορετικές γλώσσες και το μοντέλο φωνημάτων ευθυγράμμισης θα πρέπει να εκπαιδευτεί στη γλώσσα (στις γλώσσες) ενδιαφέροντος. Εναλλακτικά, μπορεί να χρησιμοποιηθεί ένα πολύγλωσσο μοντέλο αναγνώρισης φωνημάτων, το οποίο μπορεί να γενικευτεί σε γλώσσες που δεν έχουν παρατηρηθεί κατά την εκπαίδευση, αντιστοιχίζοντας τα φωνήματα που δεν εξαρτώνται από τη γλώσσα στα φωνήματα της γλώσσας (των γλωσσών) στόχου.

### Χρονοσφραγίδες σε επίπεδο λέξης χωρίς αναγνώριση φωνημάτων

Υπάρχει η δυνατότητα εξαγωγής χρονοσφραγίδων (timestamps) σε επίπεδο λέξης απευθείας από το μοντέλο Whisper χωρίς τη χρήση εξωτερικού μοντέλου φωνημάτων. Αυτή η μέθοδος θα εξαλείψει το πρόσθετο κόστος εξαγωγής συμπερασμάτων και την ανάγκη αντιστοίχισης μεταξύ των λεξικών του Whisper και των φωνημάτων. Ωστόσο, έχει διαπιστωθεί ότι αυτή η προσέγγιση υπολείπεται σε απόδοση σε σύγκριση με την εξωτερική προσέγγιση φωνημάτων και είναι επιρρεπής σε ανακρίβειες χρονοσφραγίδων.

Στην υλοποίηση του WhisperX, η προεπιλεγμένη διαμόρφωση που καθορίζεται στον Πίνακα 6.1 χρησιμοποιείται για όλα τα πειράματα, εκτός αν αναφέρεται διαφορετικά. Ακολουθούν οι λεπτομέρειες για τα άλλα μοντέλα που χρησιμοποιούνται:

- Whisper [30]: Για μεταγραφή και ευθυγράμμιση λέξεων μόνο με το Whisper, χρησιμοποιείται η προεπιλεγμένη διαμόρφωση από τον Πίνακα 6.1. Οι χρονοσφραγίδες σε επίπεδο λέξης εξάγονται από τις κορυφές διασταυρούμενης προσοχής στα αποκωδικοποιημένα tokens. Ωστόσο, πρέπει να εφαρμοστούν ευρετικές μέθοδοι χρονοσφραγίδων, συμπεριλαμβανομένης της σύσφιξης αρνητικών χρονοσφραγίδων διάρκειας, για την αποφυγή αποτυχιών ευθυγράμμισης.
- Wav2vec2.0 [22]: Για τη μεταγραφή wav2vec2.0 και την ευθυγράμμιση λέξεων, χρησιμοποιούνται οι προεπιλεγμένες ρυθμίσεις του Πίνακα 6.1, εκτός αν ορίζεται διαφορετικά. Οι εκδόσεις των μοντέλων που χρησιμοποιούνται προέρχονται από το επίσημο αποθετήριο torchaudio. Τα μοντέλα Base 960h, Large 960h και HuBERT τελειοποιήθηκαν σε δεδομένα Librispeech [6].

Για τη συγκριτική αξιολόγηση της ταχύτητας εξαγωγής συμπερασμάτων των μοντέλων, όλες οι μετρήσεις πραγματοποιούνται σε μια GPU NVIDIA A40.

| Type | Hyperparameter | Default Value |
|---|---|---|
| VAD | Model | pyannote |
| | Onset threshold | 0.767 |
| | Offset threshold | 0.377 |
| | Min. duration on | 0.136 |
| | Min. duration off | 0.067 |
| Whisper | Model version | large-v2 |
| | Decoding strategy | greedy |
| | Condition on previous text | False |
| Phoneme Recognition | Architecture | Architecture |
| | Model version | BASE 960H |
| | Decoding strategy | greedy |

Πίνακας 6.1: Ρυθμίσεις των μοντέλων του WhisperX.

## 6.2 Αποτελέσματα

Σε συνέχεια των προηγούμενων πειραμάτων, για την ευθυγράμμιση των στίχων επιλέγουμε να χρησιμοποιήσουμε την μεταγραφή των μοντέλων Whisper [30] το οποίο αναγνωρίζει τους στίχους που εκφωνούνται με μεγαλύτερη ακρίβεια.

Χρησιμοποιώντας το wav2vec2.0, το οποίο είναι ένα ισχυρό μοντέλο μάθησης αναπαράστασης ομιλίας με αυτοεπίβλεψη, το σύστημα μπορεί να μάθει ισχυρές αναπαραστάσεις των δεδομένων ήχου [36].

Η διαδικασία της αναγκαστικής ευθυγράμμισης περιλαμβάνει την ευθυγράμμιση των μεταγραφών των στίχων με τα αντίστοιχα ηχητικά τμήματα. Με τη βοήθεια του wav2vec2.0, το σύστημα μπορεί να αντλήσει ακουστικές αναπαραστάσεις υψηλής ποιότητας από τις ηχογραφήσεις. Αυτές οι αναπαραστάσεις μπορούν στη συνέχεια να χρησιμοποιηθούν σε συνδυασμό με τις πληροφορίες κειμένου για να πραγματοποιηθεί ευθυγράμμιση σε πιο ακριβές και λεπτομερές επίπεδο.

Χρησιμοποιώντας αναγκαστική ευθυγράμμιση με το wav2vec2.0, το σύστημα αυτόματης μεταγραφής στίχων μπορεί να επιτύχει βελτιωμένο συγχρονισμό μεταξύ του ήχου και των στίχων. Αυτή η ευθυγράμμιση μπορεί να βελτιώσει σημαντικά την ακρίβεια και τη χρονική αντιστοιχία μεταξύ των μεταγραμμένων στίχων και των ηχητικών τμημάτων στα οποία αντιστοιχούν. Αυτή η προσέγγιση έχει τη δυνατότητα να ξεπεράσει τις προκλήσεις που δημιουργούν οι διακυμάνσεις στη φωνητική απόδοση, οι διακυμάνσεις του ρυθμού και άλλες ηχητικές ιδιαιτερότητες, οδηγώντας σε πιο αξιόπιστα και ακριβή αποτελέσματα μεταγραφής.

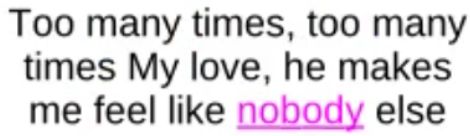
Η έξοδος του μοντέλου είναι στην μορφή:

| | | |
|---|---|---|
| 60.702005730659025 | 60.82234957020057 | Cause |
| 62.166189111747855 | 62.40687679083094 | all |
| 62.44699140401146 | 62.54727793696275 | of |
| 62.6676217765043 | 63.12893982808023 | me |
| 63.83094555873925 | 64.4727793696275 | loves |
| 65.9570200573066 | 66.13753581661891 | all |
| 66.35816618911174 | 66.45845272206304 | of |
| 66.53868194842407 | 66.69914040114614 | you |

όπου στην πρώτη και στην δεύτερη στήλη σημειώνονται η χρονική στιγμή έναρξης και λήξης της εκφώνησης κάθε λέξης που προέβλεψε το Whisper, όπως προκύπτει από τον αλγόριθμο ευθυγράμμισης.

Είδαμε πριν ότι τα μικρότερα μοντέλα του Whisper εισάγουν μεγαλύτερο σφάλμα στην μεταγραφή των στίχων. Αυτό έχει ως συνέπεια να δυσκολεύει τη διαδικασία του alignment επηρεάζοντας αρνητικά την ακρίβεια του συγχρονισμού στα σημεία που έχει αναγνωρισθεί εσφαλμένα κάποια λέξη αλλά και γύρω από αυτά.

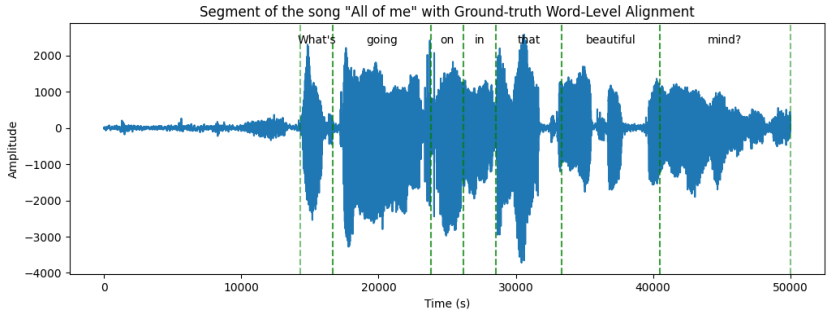
Για την οπτικοποίηση της ευθυγράμμισης, δημιουργήσαμε ένα βίντεο-demo για κάθε ηχητική καταγραφή, όπου προβάλλουμε συγχρονισμένα το ηχητικό σήμα και τους στίχους που μεταγράφηκαν και ευθυγραμμίστηκαν. Σε κάθε εκφώνηση μιας λέξης, η συγκεκριμένη λέξη εμφανίζεται με διαφορετικό χρώμα από τον υπόλοιπο στίχο, όπως και σε μια εφαρμογή καραόκε. Στο Σχήμα 6.2.1 φαίνεται η χρονική στιγμή που εκφωνείται η λέξη "nobody" από τον χρήστη,στο μουσικό κομμάτι "New Rules". Η μεταγραφή στο συγκεκριμένο παράδειγμα έγινε με το μοντέλο Whisper medium. Η δημιουργία του βίντεο έγινε με χρήση του εργαλείου επεξεργασίας βίντεο *ffmpeg* [37].



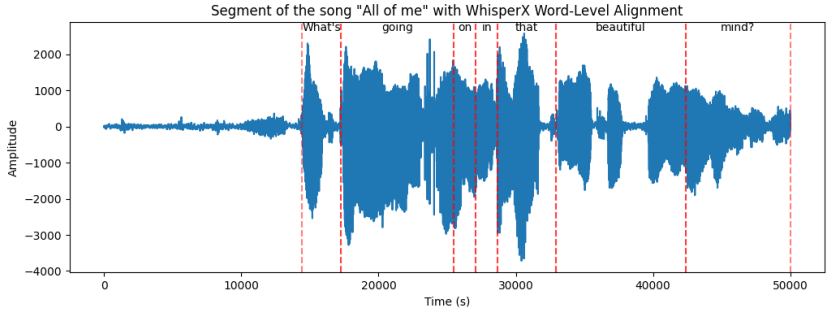

Σχήμα 6.2.1: Οπτικοποίηση της χρονικής στιγμής της εκφώνησης της λέξης "nobody".

Για την ποσοτική αξιολόγηση του αλγορίθμου ευθυγράμμισης, και την σύγκριση της απόδοσης δύο ή περισσότερων μοντέλων απαιτούνται αποτελέσματα ευθυγράμμισης όπου τα ζεύγη ηχητικού-κειμένου να είναι τα ίδια. Πρακτικά, αυτό σημαίνει ότι σε αποτελέσματα πειραμάτων αναγνώρισης, όπου κάθε μοντέλο δίνει διαφορετική μεταγραφή δεν μπορούμε να εκφράσουμε αριθμητικά την ποιότητα της ευθυγράμμισης. Έτσι, και οι μετρικές αξιολόγησης που περιγράφονται στο Κεφάλαιο 2 δεν μπορούν να εφαρμοστούν στο παρόν πείραμα.

Μια ακόμη παράμετρος που καθιστά δύσκολη την επαλήθευση της ευθυγράμμισης είναι το γεγονός ότι στο dataset οι χρονοσφραγίδες στα περισσότερα αρχεία είναι σε επίπεδο γραμμής-στίχου και όχι σε επίπεδο λέξης, όπως περιγράψαμε και στο Κεφάλαιο 4.




(a) Πραγματική χρονική ευθυγράμμιση.




(b) Χρονική ευθυγράμμιση με το μοντέλο WhisperX.

Σχήμα 6.2.2: Συγχρονισμός στίχου και ηχητικού στο μουσικό κομμάτι *All of Me* .

# Κεφάλαιο 7

# Επίλογος

Στο κεφάλαιο αυτό γίνεται σύνοψη της διπλωματικής εργασίας και παρουσιάζονται τα συμπεράσματα που προκύπτουν από τα πειράματα που διεξήχθησαν. Τέλος, παρουσιάζονται κάποιες μελλοντικές επεκτάσεις που μπορούν να εφαρμοστούν.

## 7.1 Σύνοψη και Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήθηκε το πρόβλημα της αυτόματης μεταγραφής και ευθυγράμμισης στίχων τραγουδιών. Για τη σύνοψη και την ανάδειξη των ερευνητικών αποτελεσμάτων αυτής της εργασίας, παρατίθενται στη συνέχεια ορισμένα σημαντικά ευρήματα.

Αρχικά, υλοποιήθηκε ένα μοντέλο HMM-GMM σαν βάση (baseline) με χρήση του Kaldi toolkit [24]. Ως είσοδος στο μοντέλο χρησιμοποιήθηκαν προκαθορισμένα βασικά χαρακτηριστικά. Στα πρώτα πειράματα με το μοντέλο αυτό βλέπουμε με ποιόν τρόπο η αύξηση του όγκου των δεδομένων εκπαίδευσης ενισχύει την ποιότητα της μεταγραφής και μειώνει το σφάλμα αναγνώρισης λέξης (WER). Αυτό συμβαίνει επειδή έτσι δίνουμε στο μοντέλο τη δυνατότητα να μάθει καλύτερα τις αντιστοιχίες ήχων-φωνημάτων, κοιτάζοντας μια πιο ολοκληρωμένη αντιπροσώπευση των φωνών τραγουδιού. Αυτό περιλαμβάνει διαφορετικές φωνητικές καταγραφές, φωνητικές τεχνικές, διακυμάνσεις του ηχοχρώματος, παραλλαγές άρθρωσης και προφοράς, τις συνθήκες ηχογράφηση τη μουσική υποβάθρου και πολλά άλλα.

Σύμφωνα με τους πίνακες αποτελεσμάτων στο Κεφάλαιο 5, η αρχιτεκτονική των Transformer [29] για τη μεταγραφή στίχων είναι ανώτερη από τα μοντέλα HMM-GMM, ακόμα και με εκπαίδευση μόνο σε δεδομένα ομιλίας. Οι αρχιτεκτονικές Transformer προσφέρουν πλεονεκτήματα έναντι των μοντέλων HMM όσον αφορά τον μηχανισμό προσοχής, τη μοντελοποίηση από άκρο σε άκρο, την ικανότητα εκμάθησης αναπαραστάσεων, και τη χωρητικότητα του μοντέλο. Μέσω της αρχιτεκτονικής κωδικοποιητή – αποκωδικοποιητή και του μηχανισμού προσοχής το μοντέλο μαθαίνει πολύ πιο ισχυρές αναπαραστάσεις, με επίγνωση των συμφραζομένων με είσοδο το ανεπεξέργαστο αρχείο ήχου.

Στη συνέχεια, επικεντρώνοντας την εκπαίδευση του ίδιου μοντέλου Transformer σε δεδομένα τραγουδιού, η απόδοση της αναγνώρισης μπορεί να επωφεληθεί από τα μοναδικά χαρακτηριστικά, το λεξιλόγιο, τα μοτίβα εκφώνησης και τις προκλήσεις που χαρακτηρίζουν τις τραγουδιστές φωνές. Αυτή η εξειδικευμένη εκπαίδευση επιτρέπει στο μοντέλο να κατανοεί και να ερμηνεύει καλύτερα τα χαρακτηριστικά που σχετίζονται με το τραγούδι, με αποτέλεσμα βελτιωμένη ακρίβεια αναγνώρισης και πιο αξιόπιστες μεταγραφές για ερμηνείες τραγουδιού.

Στην ίδια λογική της μεταφοράς μάθησης, χρησιμοποιήθηκε το σύγχρονο state-of-the-art μοντέλο αναγνώρισης Whisper [30], η αρχιτεκτονική του οποίου είναι βασισμένη σε Transformers. Το Whisper, το μοντέλο ASR της OpenAI, υπερέχει στην αυτόματη αναγνώριση ομιλίας, ξεπερνώντας τα προηγούμενα μοντέλα. Εκπαιδευμένο σε 680K ώρες πολύγλωσσων δεδομένων με ασθενή επίβλεψη, επιτυγχάνει υψηλή ακρίβεια και χαμηλό ποσοστό σφάλματος λέξης (WER). Είδαμε ότι κλιμακώνοντας τις παραμέτρους, μεγαλύτερη έκδοση

του μοντέλου Whisper προσφέρει βελτιωμένες επιδόσεις σε σύγκριση με το βασικό μοντέλο. Μπορεί να επιτύχει καλύτερη ακρίβεια, αλλά απαιτεί περισσότερους υπολογιστικούς πόρους για την εξαγωγή συμπερασμάτων. Η μικρότερη παραλλαγή του μοντέλου Whisper σχεδιάστηκε για να είναι ελαφρύ και αποδοτικό, καθιστώντας το κατάλληλο για ανάπτυξη σε συσκευές με περιορισμένες υπολογιστικές δυνατότητες, όπως κινητά τηλέφωνα ή ενσωματωμένα συστήματα.

## 7.2 Μελλοντικές Επεκτάσεις

Οι μελλοντικές επεκτάσεις της παρούσας εργασίας θα μπορούσαν να αφορούν διάφορες κατευθύνσεις. Μια κύρια διάσταση είναι η έμφαση στη βελτίωση των επιδόσεων μέσω της χρήσης καλύτερων μοντέλων και πλουσιότερων συνόλων δεδομένων, με την αξιοποίηση μοντέλων τελευταίας τεχνολογίας, όπως τα self-supervised συστήματα αυτόματης αναγνώρισης.

Η παρούσα εργασία επικεντρώνεται στη μεταγραφή στίχων μόνο για την αγγλική γλώσσα και η έρευνα για άλλες γλώσσες αποτελεί μια μελλοντική διάσταση λαμβάνοντας υπόψη τις πιθανές εφαρμογές της αυτόματης μεταγραφής στίχων. Εκτός από την απαίτηση δεδομένων για την εκάστοτε γλώσσα, μια βασική πρόκληση για την πολύγλωσση μεταγραφή στίχων είναι η εξάρτηση από γλωσσολογική γνώση ειδικών για την κατασκευή του μοντέλου προφοράς από λέξη σε φώνημα.

Ένας άλλος σημαντικός τομέας για μελλοντική έρευνα είναι η αναγνώριση και η ευθυγράμμιση των στίχων σε τραγούδια που δεν είναι a cappella. Αυτό μπορεί να επιτευχθεί μέσω μιας προσέγγισης δύο φάσεων. Αρχικά, τα υπάρχοντα συστήματα μπορούν να χρησιμοποιηθούν για τον διαχωρισμό της τραγουδιστής φωνής από τη μουσική υπόκρουση. Στη συνέχεια, οι μεταγραφές μπορούν να αναγνωριστούν και να ευθυγραμμιστούν άμεσα. Αυτό θα επιτρέψει στο σύστημα να καταγράψει με ακρίβεια τους στίχους σε τραγούδια με πολύπλοκες συνθέσεις και ενορχηστρώσεις.

Επιπλέον, επέκταση της εργασίας θα μπορούσε να περιλαμβάνει η ανάπτυξη συστημάτων ικανών να αναγνωρίζουν και να ευθυγραμμίζουν τους στίχους σε ένα ολόκληρο τραγούδι χωρίς την ανάγκη για ένα ξεχωριστό στάδιο διαχωρισμού με προτροπή. Αυτό θα απλοποιούσε τη διαδικασία μεταγραφής και θα την καθιστούσε πιο αποτελεσματική, καθώς θα εξαλείφονταν οι ανάγκες για στάδια προεπεξεργασίας.

Μια άλλη κατεύθυνση στην οποία μπορούμε να κινηθούμε είναι η αναγνώριση και ευθυγράμμιση της μελωδίας παράλληλα με τους στίχους. Αυτό θα διευκόλυνε την αυτόματη δημιουργία παρτιτούρας, επιτρέποντας στο σύστημα όχι μόνο να μεταγράφει τους στίχους αλλά και να εξάγει τα μελωδικά στοιχεία του τραγουδιού όπως νότες ή συγχορδίες. Με την ενσωμάτωση αυτής της πτυχής, το σύστημα θα μπορούσε να συμβάλει στη δημιουργία μουσικών παρτιτούρων για τραγούδια, γεγονός που μπορεί να έχει μεγάλη αξία για τους μουσικούς, τους συνθέτες και τους λάτρεις της μουσικής.

# Βιβλιογραφία


[1] Flanagan, J. *Speech Analysis Synthesis and Perception:* Communication and Cybernetics. Springer Berlin Heidelberg, 1972. ISBN: 9780387055619. URL:

[2] Whiteside, S. P. "Peter B. Denes and Elliot N. Pinson The Speech Chain: The Physics and Biology of Spoken Language, 2nd edition. Oxford: W.H. Freeman and Company, 1993. Pp. 246 Pb. US$14.95. ISBN 0-7167-2344-1." In: *Journal of the International Phonetic Association* 23.2 (1993), pp. 98–101. DOI: 10.1017/S0025100300004904.

[3] Stevens, S. S., Volkmann, J. E., and Newman, E. B. "A Scale for the Measurement of the Psychological Magnitude Pitch". In: *Journal of the Acoustical Society of America* 8 (1937), pp. 185–190.

[4] Nakatani, T. et al. "A Method for Fundamental Frequency Estimation and Voicing Decision: Application to Infant Utterances Recorded in Real Acoustical Environments". In: *Speech Commun.* 50.3 (Mar. 2008), pp. 203–214. ISSN: 0167-6393. DOI: 10.1016/j.specom.2007.09.003. URL:

[5] Mesaros, A. and Virtanen, T. "Automatic Recognition of Lyrics in Singing". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2010 (2010), pp. 1–11.

[6] Panayotov, V. et al. "Librispeech: An ASR corpus based on public domain audio books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[7] Meseguer-Brocal, G., Cohen-Hadria, A., and Peeters, G. "DALI: A Large Dataset of Synchronized Audio, Lyrics and notes, Automatically Created using Teacher-student Machine Learning Paradigm." In: (2018). DOI: 10.5281/ZENODO.1492443. URL:

[8] Loscos, A., Cano, P., and Bonada, J. "Low-Delay Singing Voice Alignment to Text". In: *International Conference on Mathematics and Computing*. 1999.

[9] Gales, M. J. F. and Young, S. J. "The Application of Hidden Markov Models in Speech Recognition". In: *Found. Trends Signal Process.* 1 (2007), pp. 195–304.

[10] Reynolds, D. A. "Gaussian Mixture Models". In: *Encyclopedia of Biometrics*. 2009.

[11] Mohri, M., Pereira, F., and Riley, M. "Speech Recognition with Weighted Finite-State Transducers". In: *Springer Handbook of Speech Processing*. Ed. by J. Benesty, M. M. Sondhi, and Y. A. Huang. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 559–584. ISBN: 978-3-540-49127-9. DOI: 10.1007/978-3-540-49127-9_28. URL:

[12] Riley, M., Allauzen, C., and Jansche, M. "OpenFst: An Open-Source, Weighted Finite-State Transducer Library and its Applications to Speech and Language". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*. Boulder, Colorado: Association for Computational Linguistics, May 2009, pp. 9–10. URL:

[13] Wang, Q., Wei, L., and Kennedy, R. "Iterative Viterbi decoding, trellis shaping, and multilevel structure for high-rate parity-concatenated TCM". In: *IEEE Transactions on Communications* 50.1 (2002), pp. 48–55. DOI: 10.1109/26.975743.

[14] Jurafsky, D. and Martin, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Vol. 2. Feb. 2008.

[15] Shoenfield, J. R. "Markov A. A.. Theory of Algorithms. English Translation by Schorr-Kon Jacques J., and Program for Scientific Translations Staff. Published for the National Science Foundation, Washington, D.C., and the Department of Commerce by the Israel Program for Scientific Translations, Jerusalem 1961, 444 Pp". In: *Journal of Symbolic Logic* 27.2 (1962), pp. 244–244. DOI: 10.2307/2964158.

[16] Woodard, J. and Nelson, J. "An information theoretic measure of speech recognition performance". In: 1982.

[17] Kingma, D. P. and Ba, J. *Adam: A Method for Stochastic Optimization*. 2014. arXiv: 1412.6980 [cs.LG].

[18] Srivastava, N. et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL:

[19] Hochreiter, S. and Schmidhuber, J. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: URL:

[20] Cho, K. et al. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. arXiv: 1406.1078 [cs.CL].

[21] Greff, K. et al. "LSTM: A Search Space Odyssey". In: *IEEE Transactions on Neural Networks and Learning Systems* 28.10 (Oct. 2017), pp. 2222–2232. DOI: 10.1109/tnnls.2016.2582924. URL:

[22] Kim, J. and Kang, P. *K-Wav2vec 2.0: Automatic Speech Recognition based on Joint Decoding of Graphemes and Syllables*. 2021. DOI: 10.48550/ARXIV.2110.05172. URL:

[23] Smule, I. *DAMP-MVP: Digital Archive of Mobile Performances - Smule Multilingual Vocal Performance 300x30x2*. Version 1.0.0. Zenodo, Apr. 2018. DOI: 10.5281/zenodo.2747436. URL:

[24] Povey, D. et al. "The Kaldi speech recognition toolkit". In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (Jan. 2011).

[25] Stolcke, A. "SRILM - an extensible language modeling toolkit." In: *INTERSPEECH*. Ed. by J. H. L. Hansen and B. L. Pellom. ISCA, 2002. URL:

[26] *Sphinx knowledge base tools*. URL:

[27] Watanabe, S. et al. *ESPnet: End-to-End Speech Processing Toolkit*. 2018. DOI: 10.48550/ARXIV.1804.00015. URL:

[28] Dong, L., Xu, S., and Xu, B. "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 5884–5888. DOI: 10.1109/ICASSP.2018.8462506.

[29] Karita, S. et al. "Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration". In: Sept. 2019, pp. 1408–1412. DOI: 10.21437/Interspeech.2019-1938.

[30] Radford, A. et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. arXiv: 2212.04356 [eess.AS].

[31] Chan, W. et al. *SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network*. 2021. arXiv: 2104.02133 [cs.CL].

[32] Zhuo, L. et al. *LyricWhiz: Robust Multilingual Zero-shot Lyrics Transcription by Whispering to ChatGPT*. 2023. arXiv: 2306.17103 [cs.CL].

[33] Kadlčík, M. et al. *A Whisper transformer for audio captioning trained with synthetic captions and transfer learning*. 2023. arXiv: 2305.09690 [cs.SD].

[34] *Hugging Face – The AI community building the future*. URL:

[35] Bain, M. et al. *WhisperX: Time-Accurate Speech Transcription of Long-Form Audio*. 2023. arXiv: 2303.00747 [cs.SD].

[36] Zhu, J., Zhang, C., and Jurgens, D. *Phone-to-audio alignment without text: A Semi-supervised Approach*. 2022. arXiv: 2110.03876 [cs.CL].

[37] Tomar, S. "Converting video formats with FFmpeg". In: *Linux Journal* 2006.146 (2006), p. 10.