Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

ΕΡΓΑΣΙΑ ΣΤΗΝ ΤΕΧΝΟΛΟΓΙΑ ΗΧΟΥ ΚΑΙ ΕΙΚΟΝΑΣ

Αυτόματος Συγχρονισμός Υποτίτλων

Λέτρος Κωνσταντίνος (8851)

Μανουσαρίδης Ιωάννης (8855)

Μουράτης Ιωάννης (8859)

Χατζηαντωνίου Κωνσταντίνος (8941)

Περίληψη

Η εργασία αυτή μελετάει το θέμα του *Αυτόματου Συγχρονισμού των Υποτίτλων μίας Ταινίας.* Τα ασυγχρόνιστα αρχεία βίντεο και υποτίτλων είναι συχνά φαινόμενα στα μέσα ροής. Δεδομένου ότι οι υπότιτλοι είναι σε μεγάλο βαθμό ένα ουσιαστικό μέρος της εμπειρίας της προβολής ενός βίντεο, αυτό μπορεί να έχει σημαντικές συνέπειες τελικά, ενδεχομένως καθιστώντας το περιεχόμενο απρόσιτο. Η ανίχνευση του ασυγχρόνιστου υποτίτλου και ο συγχρονισμός του είναι απαραίτητοι σε αυτή την περίπτωση.

Στο πλαίσιο της εργασίας αυτής, διερευνώνται τρόποι με τους οποίους μπορεί να επιλυθεί το προαναφερθέν ζήτημα. Παρουσιάζονται αναλυτικά πέντε υπάρχουσες υλοποιήσεις καθώς και δύο νέες προσεγγίσεις, τροποποιήσεις αυτών που υλοποιήθηκαν, από τους συγγραφείς της εργασίας. Οι υλοποιήσεις αυτές κάνουν χρήση τεχνικών μηχανικής μάθησης καθώς και επεξεργασίας ήχου για την επίτευξη του σκοπού αυτού .

Η αναφορά αυτή περιλαμβάνει πρωτίστως την θεωρητική επεξήγηση των υλοποιήσεων σχετικά με την διαδικασία που ακολουθούν και έπειτα την αντίστοιχη πειραματική τους εφαρμογή σε ένα σύνολο δεδομένων δεκαέξι ταινιών, παραθέτοντας τα αντίστοιχα αποτελέσματα και συμπεράσματα.

Καταλήγοντας, η εργασία παρουσιάζει μία σύνοψη με τις επιδόσεις των επτά υλοποιήσεων. Η αποδοτικότερη υλοποίηση προκειμένου να συγχρονίσει τον υπότιτλο μίας δίωρης ταινίας απαιτεί κατά μέσο όρο 60 δευτερόλεπτα και παρουσιάζει ακρίβεια περίπου 70%.

Λέτρος Κωνσταντίνος, konsletr@ece.auth.gr
Μανουσαρίδης Ιωάννης, imanousar@ece.auth.gr
Μουράτης Ιωάννης, mouratiis@ece.auth.gr
Χατζηαντωνίου Κωνσταντίνος, konstantic@ece.auth.gr

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών,
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Ελλάδα
Νοέμβριος 2019

Επίβλεψη

Δημούλας Α. Χαράλαμπος, Αναπληρωτής Καθηγητής
Σεβαστιάδης Χρήστος, Εργαστηριακό Διδακτικό Προσωπικό
Θωίδης Ιορδάνης, Υποψήφιος Διδάκτορας

Abstract

This paper approaches the topic of the *Automatic Synchronization of a Movie's Subtitle*. Unsynchronized audio and subtitle files are common within streaming media. As subtitles often are an essential part of the viewing experience, this can have large consequences, possibly making the content inaccessible. Detection of the asynchronous subtitle and its synchronization are necessary in this case.

In the context of this work, ways in which this issue can be resolved are explored. Five existing implementations are presented in detail as well as a sixth one based on an existing implementation and modified by the authors of the work. These implementations, which try to achieve this goal, rely on artificial intelligence such as machine learning techniques and speech recognition.

This report firstly includes a theoretical explanation of the embodiments of the procedure followed and then their corresponding experimental application to a dataset of sixteen movies, listing the respective results and conclusions.

In conclusion, the work presents a summary of the performance of the seven implementations. The most efficient implementation to synchronize the subtitles of a two-hour movie requires an average of 60 seconds and has an accuracy of about 70%.

Letros Konstantinos, konsletr@ece.auth.gr
Manousaridis Ioannis, imanousar@ece.auth.gr
Mouratis Ioannis, mouratiis@ece.auth.gr
Chatziantoniou Konstantinos, konstantic@ece.auth.gr

Department of Electrical and Computer Engineering,
Aristotle University of Thessaloniki, Greece
November 2019

Supervision

Dimoulas A. Charalambos, Associate Professor
Sevastiadis Christos, Laboratory Teaching Staff
Thoidis Iordanis, PhD Student

Περιεχόμενα

1. Εισαγωγή

Το υλικό και οι κώδικες της παρούσας εργασίας μπορούν να βρεθούν στο παρακάτω repo στο Github: Αυτόματος Συγχρονισμός Υποτίτλων.

1.1 Γενικά

Οι υπότιτλοι έχουν κυρίαρχο ρόλο στα σύγχρονα βίντεο. Η χρήση τους καθιστά δυνατή την παρακολούθηση βίντεο από άτομα με ακουστικά προβλήματα, λειτουργώντας ως υποκατάστατο του ήχου, αλλά και από άτομα που μιλούν διαφορετική γλώσσα. Συγχρόνως, η οπτικοποίηση της ομιλίας μέσω κειμένου διευκολύνει την κατανόησή του, ειδικά σε περιπτώσεις όπου οι διάλογοι είναι πολύ γρήγοροι ή ο ήχος χαμηλής έντασης. Περαιτέρω, έρευνες έχουν δείξει ότι τα βίντεο με ενσωματωμένους υπότιτλους έχουν μεγαλύτερη απήχηση στα μέσα κοινωνικής δικτύωσης και υψηλότερη αξιολόγηση από τις μηχανές αναζήτησης .

1.2 Παρουσίαση Προβλήματος της Εργασίας

Το πρόβλημα που εξετάζεται στην παρούσα εργασία έγκειται στον αυτόματο συγχρονισμό υποτίτλων σε μια ταινία βάση του ήχου. Συγκεκριμένα, η διαδικασία εκκινεί με τη φόρτωση του αρχείου υποτίτλων και του αντίστοιχου βίντεο αρχείου της ταινίας στο προς κατασκευή σύστημα και περατώνεται με την παραγωγή ενός νέου συγχρονισμένου αρχείου υποτίτλων από αυτό.

Υπάρχουν αρκετοί λόγοι που μπορούν να δημιουργήσουν πρόβλημα στο συγχρονισμό [4] των υποτίτλων με την ταινία όπως :

- **Αρχική Χρονοσφραγίδα:** Πολλές εκδόσεις της ίδιας ταινίας ή βίντεο ενδέχεται να διαφέρουν από μερικά χιλιοστά του δευτερολέπτου έως μερικά δευτερόλεπτα. Αυτό έχει ως αποτέλεσμα ένας υπότιτλος που είναι κατάλληλα συγχρονισμένος για μία έκδοση Α μιας ταινίας, να μην είναι κατάλληλος για μία άλλη έκδοση, Β.
- **Κομμένες Σκηνές:** Ταινίες που περιέχουν κομμένες σκηνές διαφέρουν μερικά δευτερόλεπτα σε σύγκριση με την πρώτη έκδοση της ταινίας.
- **Διαφορές Περιεχομένου:** Οι υπότιτλοι μπορεί να περιέχουν (ή όχι) μεταφράσεις για οπτικές πληροφορίες (πχ. πινακίδες), ή ήχους ζώων (πχ. σκυλιά που γαβγίζουν). Περαιτέρω, ενδέχεται η ύπαρξη υποτίτλων σε σημεία χωρίς φωνή, όπως για παράδειγμα όταν σε μία ταινία γίνεται χρονική μετάβαση και αναφέρεται ως μετάφραση το εξής «Τέσσερις μήνες μετά ...» .

- **Στίχοι Τραγουδιών:** Διαφορετικοί υπότιτλοι ενδέχεται να περιλαμβάνουν (ή όχι) τους στίχους από τα τραγούδια που ακούγονται κατά την αναπαραγωγή της ταινίας ή του βίντεο.

1.3 Τυπογραφικές Παραδοχές του Εγγράφου

Για το συγκεκριμένο έγγραφο επιλέχθηκε η γραμματοσειρά «Arial», σε μέγεθος 11pt για το βασικό κείμενο, 10pt για τις περιγραφές των σχημάτων και 15pt, 13.5pt και 12pt για τους τίτλους κεφαλαίων, ενοτήτων και υποενοτήτων αντίστοιχα. Όλοι οι τίτλοι είναι σε έντονη γραφή (Bold), οι υποενότητες είναι και υπογραμμισμένες, ενώ οι περιγραφές των σχημάτων είναι σε πλάγια γραφή. Σε όλη την έκταση του εγγράφου χρησιμοποιούνται παρενθέσεις () για διευκρινίσεις, παρατηρήσεις και μεταφράσεις όρων. Τα σχήματα, οι εικόνες και οι πίνακες περιέχουν αριθμημένες λεζάντες σε όλη την έκταση του εγγράφου. Βιβλιογραφία και επιπλέον αναφορές βρίσκονται στο τέλος του εγγράφου. Υπάρχει πλήρης στοίχιση στο κείμενο κορμού.

1.4 Αναγνωστικό κοινό και τρόπος ανάγνωσης

Η ανάγνωση μπορεί να γίνει είτε σειριακά με την συνήθη διάταξη γραφής, είτε τμηματικά, δίνοντας βάση σε συγκεκριμένα κεφάλαια, ανάλογα με τις ανάγκες του αναγνώστη. Ωστόσο οι αναγνώστες καλούνται, προκειμένου να διευκολυνθούν αλλά και για την πλήρη και σωστή κατανόηση του εγγράφου, να ακολουθήσουν την ροή του.

Το έγγραφο αυτό απευθύνεται στις εξής ομάδες ατόμων: Τα μέλη της επιτροπής αξιολόγησης αυτής της εργασίας, τους μηχανικούς που θα αναλάβουν τη συγγραφή, δοκιμή και αποσφαλμάτωση του συγκεκριμένου έργου, τους μηχανικούς που θα αναλάβουν τη συντήρηση και την επέκτασή του και τέλος οποιοδήποτε άλλον ενδιαφερόμενο. Οι παράγραφοι της αναφοράς αυτής είναι γραμμένες σε όσο είναι εφικτό, απλή γλώσσα, κρίνοντας όμως αναγκαίες για την κατανόησή τους κάποιες βασικές γνώσεις που αφορούν μερικούς κλάδους, όπως των μαθηματικών, του προγραμματισμού, της επεξεργασίας σήματος και ήχου και της αναγνώρισης προτύπων.

1.5 Δομή του Εγγράφου

Η υπάρχουσα αναφορά χωρίζεται σε οκτώ κεφάλαια. Ολοκληρώνοντας αυτή τη στιγμή το πρώτο, εισαγωγικό, κεφάλαιο η περαιτέρω ανάγνωση περιλαμβάνει κατά σειρά:

- Εισαγωγικές έννοιες σχετικές με το αντικείμενο της εργασίας στις οποίες γίνεται επεξήγηση ορολογιών και τεχνικών που θα αναφερθούν μετέπειτα.
- Στρατηγικές που έχουν ακολουθηθεί στο παρελθόν αναφέροντας πρώτα συνοπτικά τη διαδικασία εκτέλεσής τους ακολουθιακά με βήματα (στρατηγικές) και έπειτα την αναλυτική εφαρμογή τους με λεπτομέρειες της εκάστοτε υλοποίησης.
- Εργαλεία και δεδομένα που χρησιμοποιήθηκαν.

- Επεξήγηση των αλγορίθμων με τους οποίους πραγματοποιήθηκε ο απο-συγχρονισμός των υποτίτλων για την κατασκευή του συνόλου δεδομένων και η αξιολόγηση των συγχρονισμένων υποτίτλων.
- Αξιολόγηση των υλοποιήσεων που εντοπίστηκαν στη βιβλιογραφία με την παράθεση των αποτελεσμάτων τους μέσω διαγραμμάτων και πινάκων καθώς και το σχολιασμό αυτών.
- Αξιολόγηση δύο νέων υλοποιήσεων βασισμένων στις υλοποιήσεις αυτές, οι οποίες δημιουργήθηκαν στο πλαίσιο της εργασίας.
- Σύγκριση μεταξύ των υλοποιήσεων με την παράθεση και τη σύγκριση των αποτελεσμάτων τους.
- Βιβλιογραφικές αναφορές.

2. Εισαγωγικές Έννοιες – Αντικείμενο της Εργασίας

Σε αυτό το κεφάλαιο θα γίνει μια σύντομη ανάλυση του θεωρητικού υπόβαθρου που απαιτείται για την πλήρη κατανόηση της εργασίας. Η κατανόηση και η γνώση των εννοιών που παρουσιάζονται εδώ, αποσκοπούν στη θεωρητική προετοιμασία του αναγνώστη για την ευκολότερη και καλύτερη παρακολούθηση των επόμενων κεφαλαίων. Η ανάγνωση του κεφαλαίου αυτού από ένα έμπειρο αναγνώστη είναι προαιρετική, ωστόσο προτείνεται.

2.1 Ήχος και Πρότυπα (Audio and Patterns)

«Ο ήχος είναι η αίσθηση που προκαλείται λόγω της διέγερσης των αισθητηρίων οργάνων της ακοής από μεταβολές πίεσης του ατμοσφαιρικού αέρα. Αυτές οι μεταβολές διαδίδονται με τη μορφή ηχητικών κυμάτων» [14]. Οι συχνότητες που αντιλαμβάνεται ο άνθρωπος κυμαίνονται προσεγγιστικά από 20 Hz έως 20 kHz.

Η επεξεργασία του ήχου στον υπολογιστή γίνεται με τη χρήση ειδικού τύπου αρχείων όπως τα WAV (Waveform Audio), τα AAC (Advanced Audio Coding), τα PCM (Pulse-Code Modulation), τα MP3 (MPEG-1 Audio Layer 3) και τα FLAC (Free Lossless Audio Codec).

Πιο συγκεκριμένα, στα πλαίσια της εργασίας θα χρησιμοποιηθούν τα αρχεία WAV [16], τα οποια είναι μία μορφή αρχείων που χρησιμοποιούνται για την αποθήκευση του ψηφιακού ήχου (digital audio) χωρίς συμπίεση και υποστηρίζουν μεγάλη ποικιλία όσον αφορά τη διφιακή ανάλυση (bit resolution), τους ρυθμούς δειγματοληψίας (sample rates) και τα κανάλια (channels) του ήχου.

2.2 Υπότιτλοι και Πρότυπα (Subtitles and Patterns)

Οι υπότιτλοι [8] είναι μια μορφή γραπτής μετάφρασης του διαλόγου ή της ομιλίας ενός βίντεο ή ταινίας σε μια ξένη γλώσσα, ή της απλής γραπτής απόδοσής τους στην ίδια γλώσσα, με ή χωρίς επιπρόσθετες πληροφορίες. Ένα παράδειγμα χρήσης τους είναι η διευκόλυνση των θεατών που πάσχουν από κώφωση ή βαρηκοΐα και αδυνατούν να κατανοήσουν πλήρως τον προφορικό λόγο ή βρίσκονται σε θορυβώδες περιβάλλον.

Λόγω της μεγάλης πρακτικής σημασίας τους, υπάρχουν συγκεκριμένα πρότυπα τα οποία ένας υπότιτλος πρέπει να ακολουθεί ώστε να θεωρείται καλός. Σύμφωνα με τα πρότυπα του BCC [2] και του Netflix [1] ορισμένα χαρακτηριστικά είναι :

- Η προτίμηση της αυτολεξεί (verbatim) μετάφρασης.
- Η διάρκεια του υποτίτλου πρέπει να κυμαίνεται από 5/6 του δευτερολέπτου έως 7 δευτερόλεπτα.

- Χρήση δύο γραμμών ώς μέγιστο. Η εναλλαγή των γραμμών επίσης πρέπει να πληροί κάποιες αρχές, όπως να μην διαχωρίζει το ουσιαστικό από το άρθρο και να γίνεται μετά από σημεία στίξης.
- Η τοποθέτηση των υποτίτλων πρέπει να είναι στο κέντρο της οθόνης, είτε στο κάτω είτε στο πάνω μέρος της.

Υπάρχουν διάφορες μορφές-πρότυπα αρχείων υποτίτλων, τα πιο διαδεδομένα εκ των οποίων είναι τα εξής: α) Time Text Markup Language (TTML) , β) Stereolithography (STL) , γ) SubRip Subtitle file (SRT) και δ) Web Video Text Tracks (VTT).

Το πλέον ευρέως χρησιμοποιούμενο είναι το πρότυπο SRT, το οποίο αποτελείται από τέσσερα τμήματα [5] :

1. Έναν αριθμό που αποτελεί την αρίθμηση του υποτίτλου κατά σειρά στο αρχείο.
2. Το χρόνο εμφάνισης και εξαφάνισης του υποτίτλου στην οθόνη (με ακρίβεια χιλιοστού).
3. Το κείμενο του υποτίτλου.
4. Μια κενή γραμμή που υποδεικνύει την έναρξη του επόμενου υποτίτλου.

Σχήμα 2-1 : Η ταινία «V for Vendetta» σε μορφή .mp4 και το αντίστοιχο αρχείο υποτίτλων σε μορφή .srt

2.3 Εντοπισμός Φωνητικής Δραστηριότητας (Voice Activity Detection - VAD)

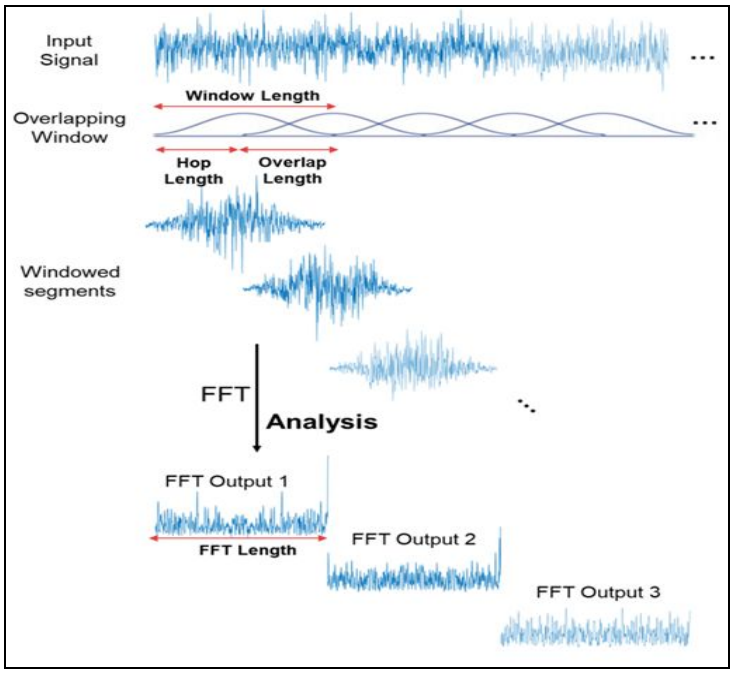
Ο εντοπισμός φωνητικής δραστηριότητας [9], γνωστός και ως εντοπισμός ομιλίας (Speech Detection), είναι μία τεχνική που αναφέρεται στην ανίχνευση παρουσίας ή απουσίας της ανθρώπινης φωνής. Η σημασία του εντοπισμού φωνητικής δραστηριότητας είναι ύψιστη στις σύγχρονες εφαρμογές που βασίζονται στην ομιλία. Ως αποτέλεσμα αυτού, έχουν αναπτυχθεί πολλοί τέτοιοι αλγόριθμοι που διαφοροποιούνται κυρίως ως προς την αδράνεια (Latency), την ευαισθησία (Sensitivity), την ακρίβεια (Accuracy) και το υπολογιστικό κόστος (Computational Cost).

Η τυπική σχεδίαση ενός αλγορίθμου εντοπισμού φωνής περιλαμβάνει τα εξής στάδια:

1. Ένα στάδιο μείωσης θορύβου.
2. Την εξόρυξη κάποιων χαρακτηριστικών από το σήμα εισόδου.
3. Την εφαρμογή ενός κανόνα ταξινόμησης, με βάση τα χαρακτηριστικά αυτά, σε τμήματα του σήματος εισόδου ώστε να ταξινομηθούν ως τμήματα που περιέχουν ομιλία (Speech) ή όχι (non-Speech).

2.4 Φασματογράφημα (Spectrogram)

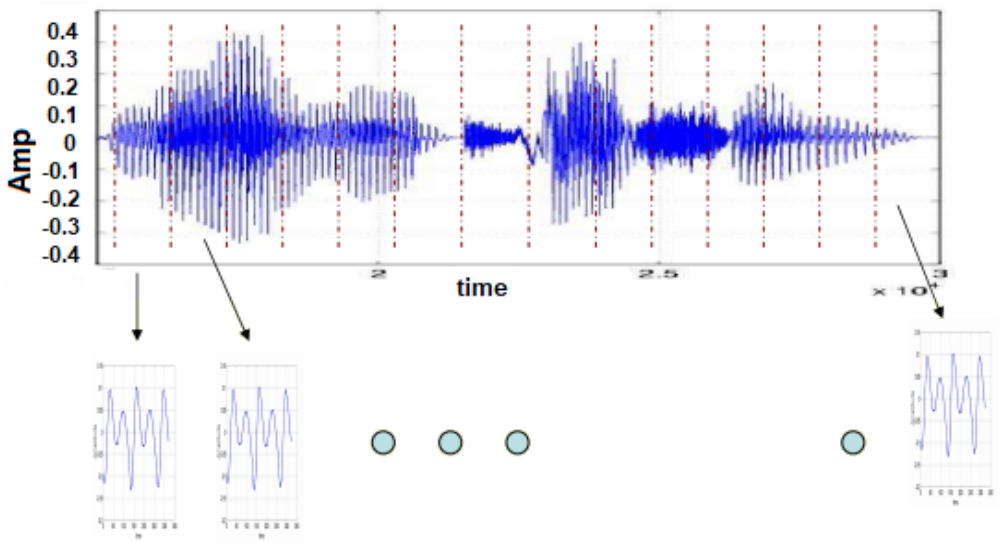
Το φασματογράφημα είναι ένα βασικό εργαλείο στη φασματική ανάλυση του ήχου ενώ έχει εφαρμοστεί εκτενώς στην ανάλυση ομιλίας αλλά και σε άλλους τομείς. Μπορεί να οριστεί ως μια γραφική και ταυτόχρονα χρωματική αναπαράσταση της έντασης (συνήθως σε λογαριθμική κλίμακα, όπως dB) του μέτρου του βραχυχρόνιου μετασχηματισμού Fourier (Short Time Fourier Transform - STFT). Ο STFT είναι μια ακολουθία DFT (Discrete Fourier Transform) - φασμάτων που έχουν προκύψει ο καθένας από το γινόμενο μετατοπισμένων συναρτήσεων-παράθυρο με το αρχικό σήμα ήχου. Τα παράθυρα αυτά επιτρέπεται να επικαλύπτονται στο χρόνο, συνήθως κατά 25-50%. Αποτελεί μια σημαντική αντιπροσώπευση των ηχητικών δεδομένων, καθώς η ανθρώπινη ακοή βασίζεται σε ένα είδος φασματογράφων πραγματικού χρόνου που κωδικοποιείται από τον κοχλία στο εσωτερικό του αυτιού.



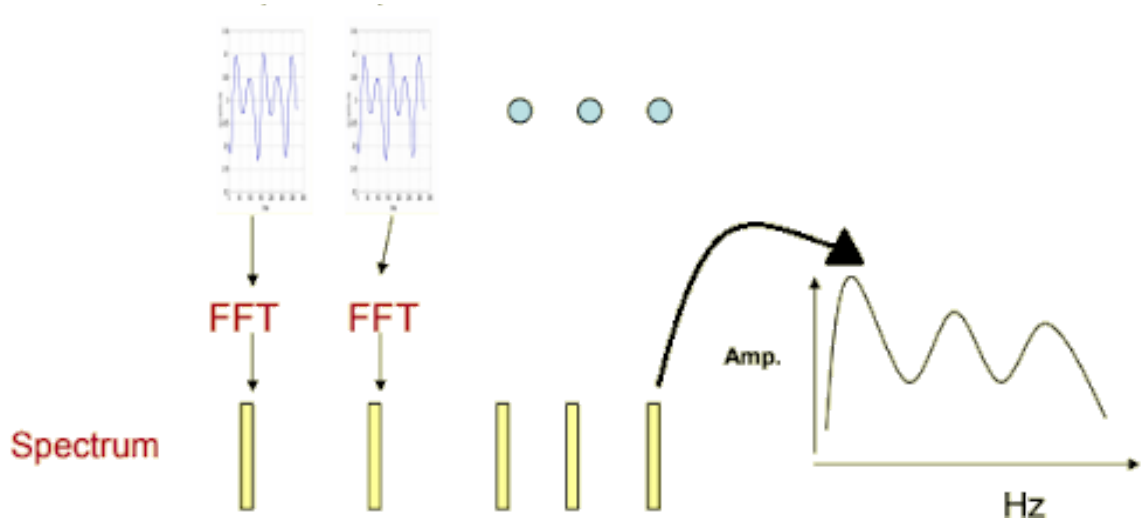
Σχήμα 2-2 : Παραγωγή ακολουθίας φασμάτων για την κατασκευή του STFT του σήματος στο χρόνο.
Πηγή: https://www.mathworks.com/help/dsp/ref/dsp.stft.html

Συγκεκριμένα, η διαδικασία μετατροπής του ηχητικού σήματος και απεικόνισής του σε φασματογράφημα περιλαμβάνει τα εξής βήματα:

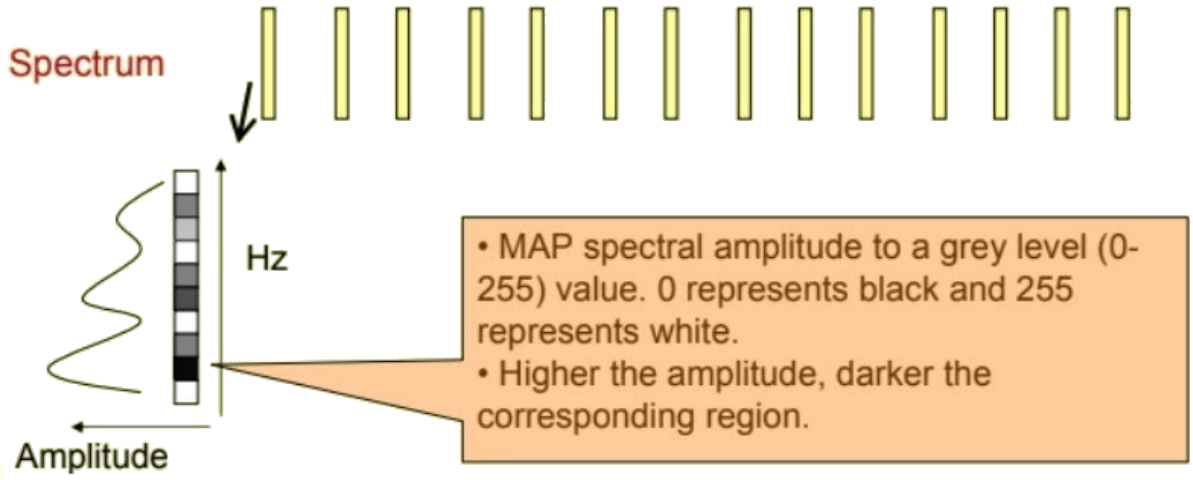
1. Αρχικά το σήμα χωρίζεται σε επιμέρους επικαλυπτόμενα τμήματα. Τα τμήματα αυτά προκύπτουν πολλαπλασιάζοντας το σήμα κάθε φορά με κάποια μετατοπισμένη συνάρτηση παράθυρο (Window).
2. Σε κάθε ένα από αυτά τα τμήματα εφαρμόζεται ο Διακριτός Μετασχηματισμός Fourier (Discrete Fourier Transform) με τη βοήθεια του αλγορίθμου FFT (Fast Fourier Transform), δημιουργώντας ένα φάσμα (Spectrum) για κάθε τμήμα.
3. Το πλάτος του φάσματος αντιστοιχίζεται (mapping) στη κλίμακα του γκρι από το 0 (μαύρο) έως το 255 (λευκό). Έτσι μπορεί πλέον να σχηματιστεί μια τρισδιάστατη απεικόνιση του αρχικού σήματος όπου οι δύο διαστάσεις, του χρόνου και της συχνότητας, αναπαρίστανται από τους άξονες του καρτεσιανού επιπέδου ενώ η τρίτη διάσταση, του πλάτους του φάσματος, αναπαρίσταται από χρώμα.
4. Από την ένωση όλων των απεικονίσεων για κάθε τμήμα προκύπτει το φασματογράφημα.



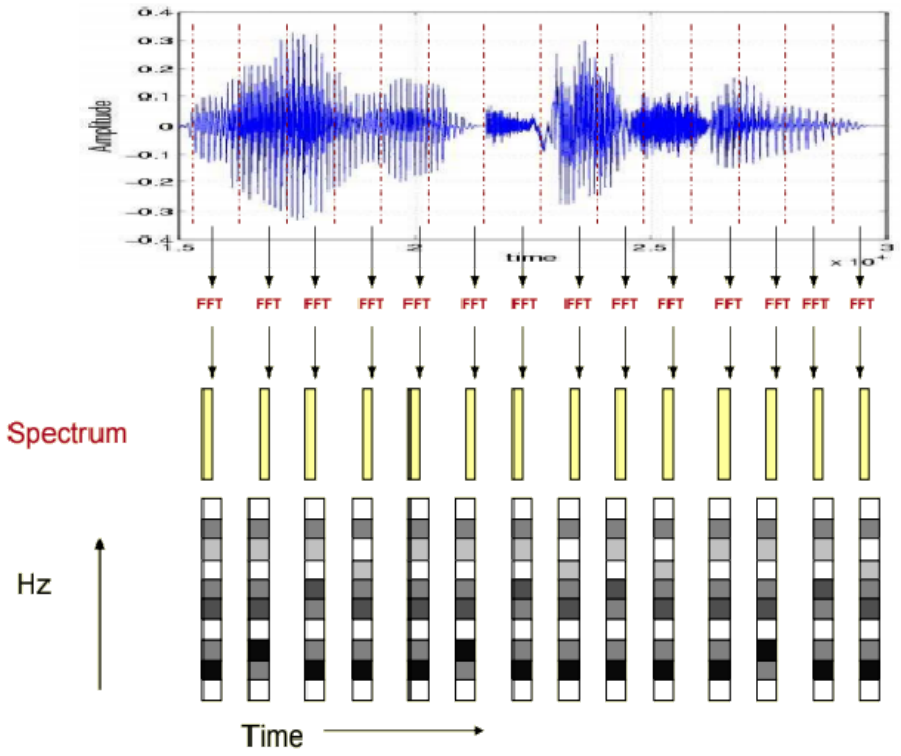
Σχήμα 2-3 : Βήμα 1 - Χωρισμός σε επιμέρους τμήματα του ηχητικού σήματος



Σχήμα 2-4 : Βήμα 2 - Εφαρμογή το FFT στα επιμέρους τμήματα του ηχητικού σήματος



Σχήμα 2-5 : Βήμα 3 - Αντιστοίχηση του πλάτους στη κλίμακα του γκρι



Σχήμα 2-6 : Βήμα 4 - Ένωση των χρωματισμένων φασμάτων και δημιουργία του φασματογραφήματος
Πηγή των Σχήμα 2-3 έως 2-6: http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

2.5 Κλίμακα Mel (Mel Scale)

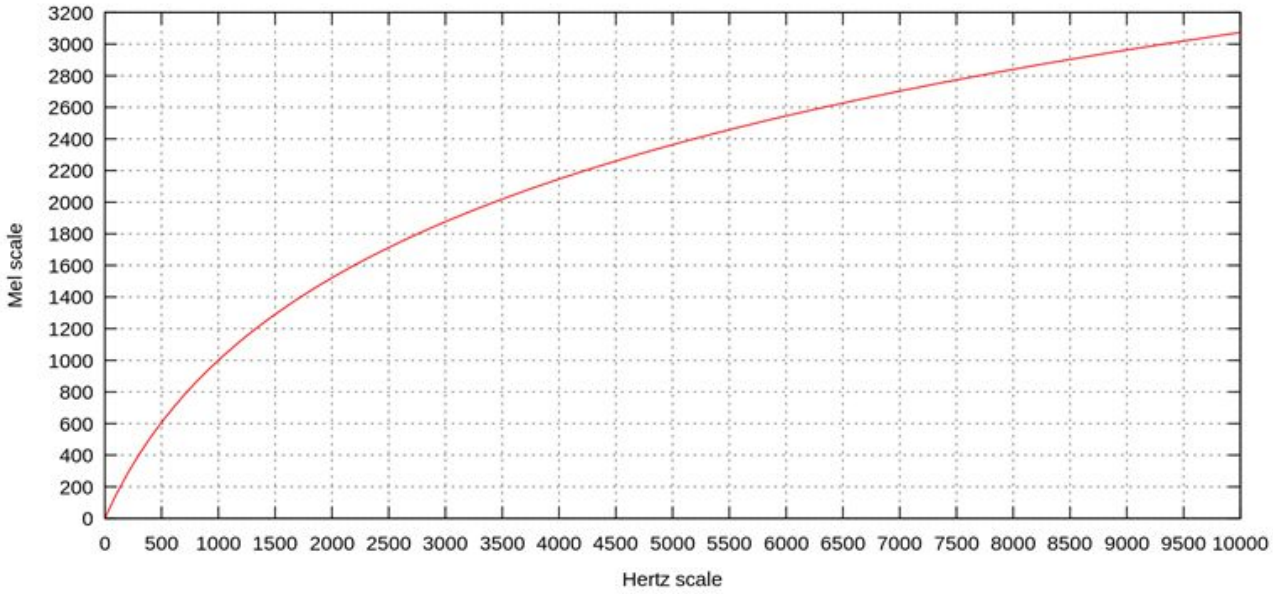
Η κλίμακα Mel σχετίζεται με τις συχνότητες που αντιλαμβάνεται το ανθρώπινο αυτί. Ψυχοφυσικές έρευνες έχουν δείξει ότι ο άνθρωπος αντιλαμβάνεται καλύτερα τις μεταβολές στις χαμηλές συχνότητες παρά στις υψηλές. Η κλίμακα Mel χρησιμοποιείται προκειμένου να προσεγγιστεί καλύτερα η ανθρώπινη ακοή. Συγκεκριμένα, στις χαμηλές συχνότητες παρουσιάζει γραμμική συμπεριφορά ενώ στις υψηλές λογαριθμική.

Η μετατροπή των συχνοτήτων στην κλίμακα Mel γίνεται με τον ακόλουθο τύπο:

$$m = 2595\, log_{10}(1 + \frac{f}{700}),\ (mels) \qquad (2.1)$$

Ενώ η επιστροφή στον χώρο των συχνοτήτων γίνεται από τον τύπο:

$$f = 700\left(10^{\frac{m}{2595}} - 1\right),\ (Hertz) \qquad (2.2)$$



Σχήμα 2-7 : Γραφική παράσταση που απεικονίζει τη συνάρτηση της κλίμακας των συχνοτήτων με την κλίμακα Mel. Πηγή: https://en.wikipedia.org/wiki/Mel_scale

2.6 Mel Frequency Cepstral Coefficient - MFCC

Τα MFCC [3][12] αποτελούν μία τεχνική επεξεργασίας του ήχου που χρησιμοποιείται για την εξαγωγή χρήσιμων χαρακτηριστικών (Features) και την αναγνώριση φωνητικών γνωρισμάτων σε αυτόν, ενώ συγχρόνως επιτυγχάνεται απόρριψη ανεπιθύμητου τύπου ηχητικών χαρακτηριστικών όπως ο παρασκηνιακός θόρυβος (Background Noise). Παρουσιάστηκαν για πρώτη φορά από τους Davis και Mermelstein το 1980. Έκτοτε αποτελούν τεχνολογία αιχμής (State-of-the-Art) [13] στο πεδίο τους και χρησιμοποιούνται κατά κόρον στην αυτόματη ομιλία (Automatic Speech) και την αναγνώριση ομιλητή (Speaker Recognition).

Η προσέγγιση για να εξαχθούν τα MFCC περιγράφεται από τα ακόλουθα βήματα:

1. Ένα στάδιο προ-έμφασης (pre-emphasis) για την ενίσχυση των υψηλών συχνοτήτων.

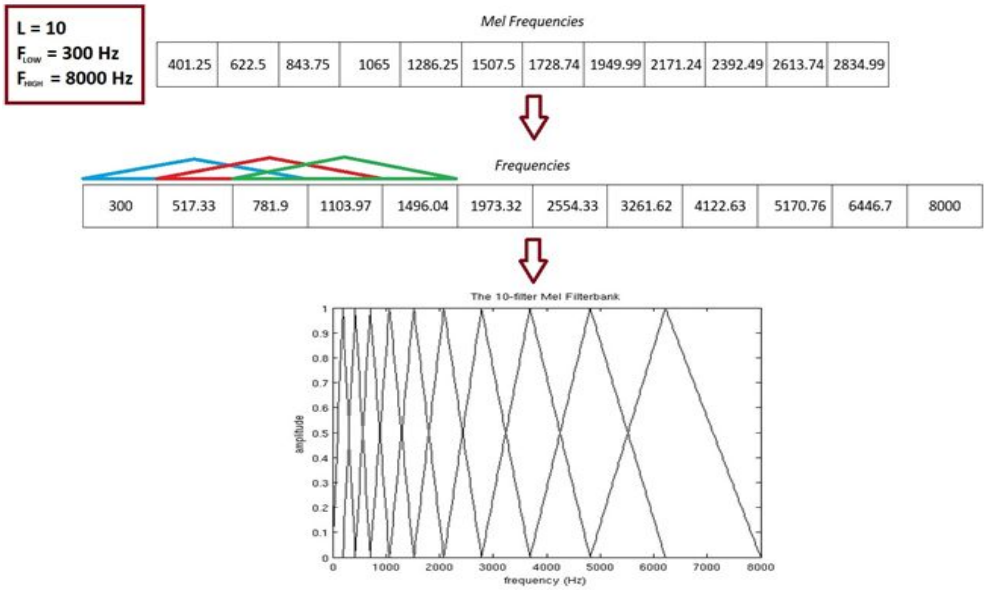
2. Δημιουργία μίας σειράς από επικαλυπτόμενα πλαίσια (frames) του ηχητικού σήματος.

3. Εκτίμηση περιοδικού φάσματος του φάσματος ισχύος για καθένα από τα πλαίσια με χρήση του αλγορίθμου Fast Fourier Transform (FFT).

4. Εφαρμογή μίας Mel ομάδας τριγωνικών φίλτρων (Mel Filterbank) στα φάσματα ισχύος και άθροιση της ενέργειας του κάθε φίλτρου.

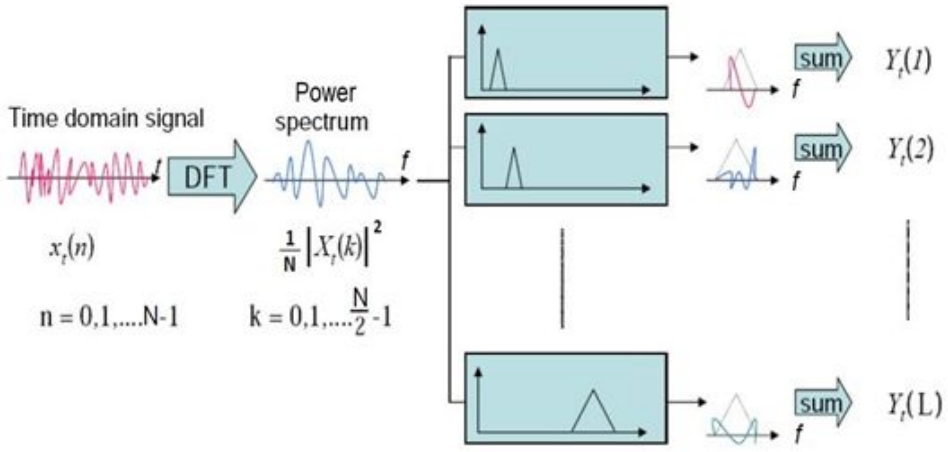
   Μια ομάδα L τριγωνικών φίλτρων (L τυπικά από 20 έως 40) προκύπτει μετά από μετασχηματισμό της χαμηλότερης και υψηλότερης συχνότητας του φάσματος ισχύος στην κλίμακα Mel (Mel Scale) και τη δημιουργία ενός διανύσματος L+2 Mel συχνοτήτων

(Mel Frequencies) που κάθε μία ισαπέχει από την επόμενη. Το διάνυσμα αυτό μετασχηματίζεται στη συνέχεια σε ένα νέο διάνυσμα με συμβατικές συχνότητες το οποίο χρησιμοποιείται για την κατασκευή μιας σειράς από τριγωνικά φίλτρα.

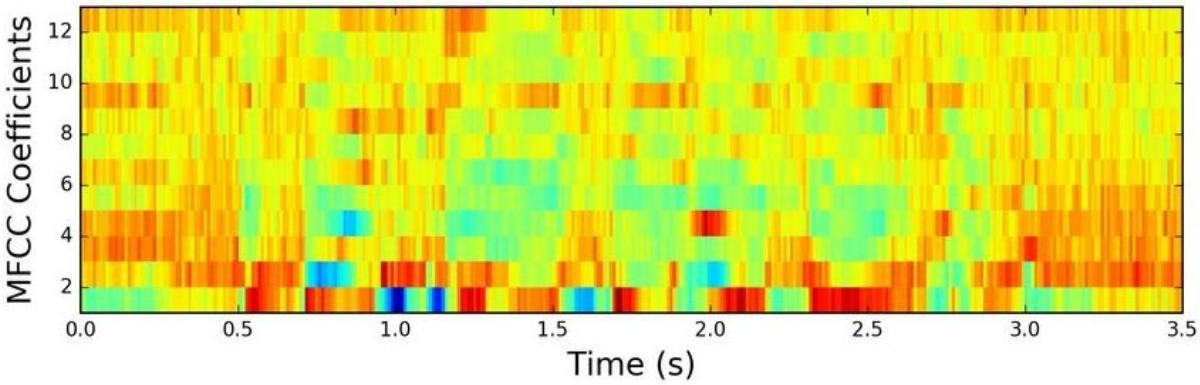


Σχήμα 2-8 : Παράδειγμα κατασκευής Mel ομάδας τριγωνικών φίλτρων (L=10) για φάσμα με συχνοτικό περιεχόμενο από 300 Hz έως 8000 Hz.

5. Υπολογισμός του λογαρίθμου (log) κάθε στοιχείου και εφαρμογή του διακριτού μετασχηματισμού συνημιτόνου (Discrete Cosine Transform - DCT) για κάθε ένα από τα πλαίσια.
6. Οι 12 συντελεστές που προκύπτουν από το 2 έως το 13 του διακριτού μετασχηματισμού συνημιτόνου αποτελούν τους MFCC του αντίστοιχου πλαισίου.



Σχήμα 2-9 : Σχηματική Αναπαράσταση των βημάτων 3 - 5
Πηγή:https://docplayer.net/30202985-Feature-extraction-mel-frequency-cepstral-coefficients-mfcc-mustafa-yankayis.html



Σχήμα 2-10 : Σχηματική απεικόνιση των MFCC χαρακτηριστικών με βάση το χρόνο.
Πηγή: https://machinelearnings.co/automatic-subtitle-synchronization-e188a9275617

Για να είναι αποδεκτή η προσέγγιση αυτή γίνεται η παραδοχή ότι το ηχητικό σήμα δεν μεταβάλλεται στοχαστικά σε μικρές χρονικές κλίμακες (στάσιμο). Προκειμένου να ισχύει αυτή η παραδοχή το μήκος των πλαισίων θα πρέπει να κυμαίνεται μεταξύ 20-40 χιλιοστών του δευτερολέπτου. Περαιτέρω, ο λόγος που υπολογίζεται η φασματική ισχύς, αντί απλά του φάσματος, οφείλεται στο γεγονός ότι ο ανθρώπινος κοχλίας, στο αυτί, προβαίνει σε μία παρόμοια διαδικασία φιλτραρίσματος των εισερχόμενων συχνοτήτων. Ανάλογα με τις εισερχόμενες συχνότητες, δονούνται διαφορετικά μέρη του κοχλία και ενεργοποιούνται διαφορετικοί νευρώνες για να αποστείλουν την πληροφορία στον εγκέφαλο.

2.7 Μη Κυρτό Πρόβλημα Βελτιστοποίησης (Non Convex Optimization Problem)

Πρόβλημα βελτιστοποίησης [15] ονομάζεται ένα πρόβλημα εύρεσης της βέλτιστης λύσης από ένα σύνολο εφικτές λύσεις (λύσεις που πληρούν τους περιορισμούς του προβλήματος). Η συνήθης μορφή ενός συνεχούς προβλήματος βελτιστοποίησης είναι:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_j(x) = 0, \quad j = 1, \ldots, p \end{aligned}$$

Σχήμα 2-11 : Πρόβλημα Ελαχιστοποίησης της συνάρτησης κόστους f ως προς το διάνυσμα x δεδομένων περιορισμών.
Πηγή: https://en.wikipedia.org/wiki/Optimization_problem

- $f : \mathbb{R}^n \rightarrow \mathbb{R}$ είναι η συνάρτηση κόστους προς ελαχιστοποίηση ως προς το διάνυσμα $x \, \varepsilon \, \mathbb{R}^n$,
- $g_i(x) \leq 0$ είναι οι ανισοτικοί περιορισμοί του προβλήματος,

- $h_j(x) = 0$ είναι οι ισοτικοί περιορισμοί του προβλήματος και
- $m \geq 0$, $p \geq 0$.

Αν $m = p = 0$, το πρόβλημα δεν έχει περιορισμούς. Συμβατικά η τυπική μορφή ενός προβλήματος βελτιστοποίησης ορίζεται ως πρόβλημα ελαχιστοποίησης (και όχι μεγιστοποίησης). Το πρόβλημα βελτιστοποίησης, όταν το $x \varepsilon \mathbb{Z}^n$, λέγεται διακριτό.

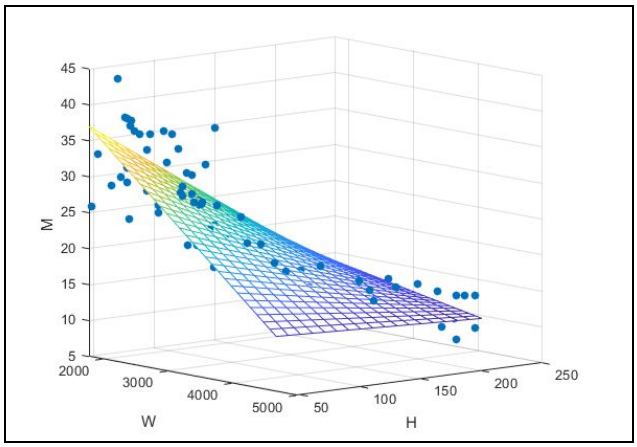
Τα προβλήματα βελτιστοποίησης που περιέχουν μόνο κυρτές συναρτήσεις f , g_i, h_j για κάθε i, j ονομάζονται Κυρτά Προβλήματα Βελτιστοποίησης και παρουσιάζουν δύο πολύ χρήσιμες ιδιότητες:

- Κάθε τοπικό ελάχιστο της συνάρτησης κόστους f είναι και ολικό της ελάχιστο, με αποτέλεσμα η εύρεση ενός μονάχα τοπικού ελαχίστου στο πρόβλημα αυτό, να σηματοδοτεί άμεσα και την επίλυση του.
- Αν η συνάρτηση κόστους f είναι αυστηρά κυρτή, τότε το πρόβλημα έχει μοναδική θέση ελαχίστου.

Διαφορετικά, το πρόβλημα βελτιστοποίησης ονομάζεται μη κυρτό και η δυσκολία επίλυσής του είναι σε άγνωστο βαθμό αυξημένη καθώς οι παραπάνω ιδιότητες δεν ισχύουν αναγκαία.

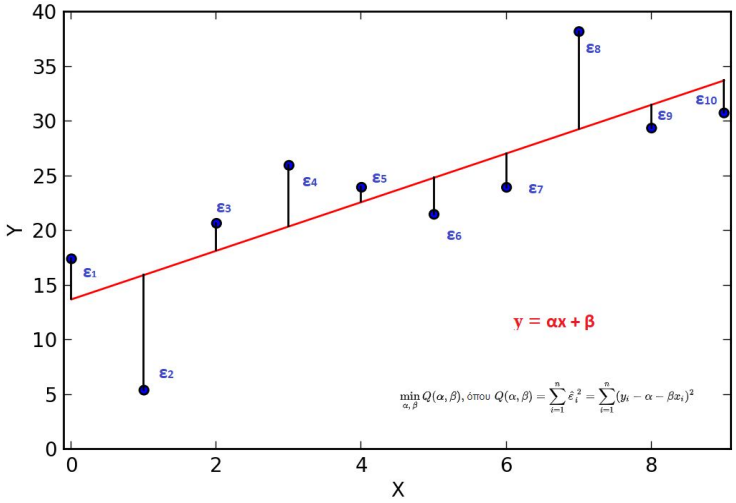
2.8 Γραμμική Παλινδρόμηση (Linear Regression)

Στο χώρο της στατιστικής, η γραμμική παλινδρόμηση αναφέρεται ως μία προσεγγιστική μοντελοποίηση της σχέσης ανάμεσα σε μία απλή εξαρτημένη (dependent) μεταβλητή *y* με μία ή περισσότερες ανεξάρτητες (independent) μεταβλητές { x_1, x_2, …, x_n }[24]. Η μεταβλητή *y* θεωρείται τυχαία μεταβλητή ενώ η x_i όχι. Στην περίπτωση που υπάρχει μόνο μία ανεξάρτητη μεταβλητή x τότε η μοντελοποίηση ονομάζεται απλή γραμμική παλινδρόμηση (simple linear regression).



Σχήμα 2-12 : Παράδειγμα γραμμικής παλινδρόμησης.
Πηγή: https://medium.com/datadriveninvestor/basics-of-linear-regression-9b5

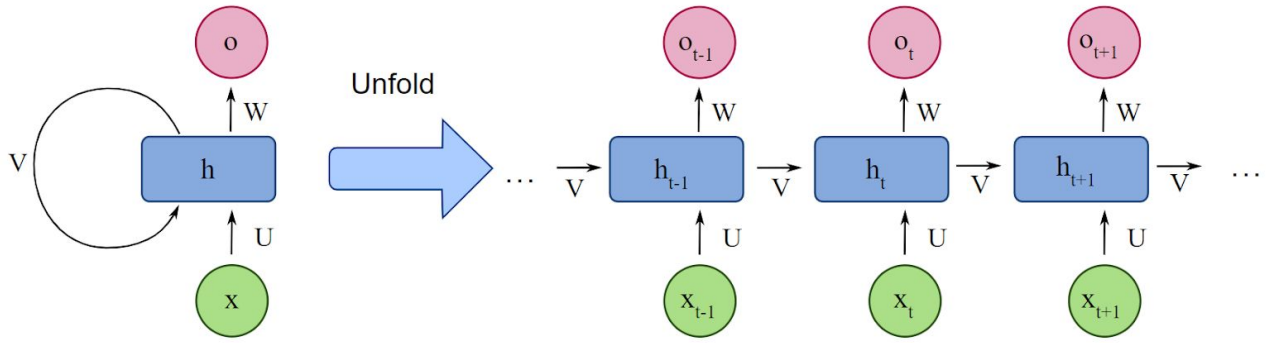
Στην απλή γραμμική παλινδρόμηση, αναζητείται μία ευθεία $f(x) = y = ax + \beta$, για ένα σύνολο δειγμάτων με τιμές $\{x_i, y_i\}$. Σκοπός είναι η εύρεση μίας ευθείας η οποία προσεγγίζει βέλτιστα το σύνολο των δειγμάτων. Η διαδικασία εύρεσης της ευθείας αυτής βασίζεται συνήθως στη *Μέθοδο των Ελαχίστων Τετραγώνων.* Στη διαδικασία αυτή η βέλτιστη ευθεία θεωρείται εκείνη η οποία το άθροισμα των τετραγώνων των αποστάσεων όλων των δειγμάτων είναι το ελάχιστο σε σύγκριση με όλες τις άλλες ευθείες.



Σχήμα 2-13 : Παράδειγμα απλής γραμμικής παλινδρόμησης με τη Μέθοδο των Ελαχίστων Τετραγώνων.
Πηγή: https://cdn-images-1.medium.com/max/1200/0*FjKhbw6Va8O8bCkF.png

2.9 Ανατροφοδοτούμενα Νευρωνικά Δίκτυα (Recurrent Neural Networks)

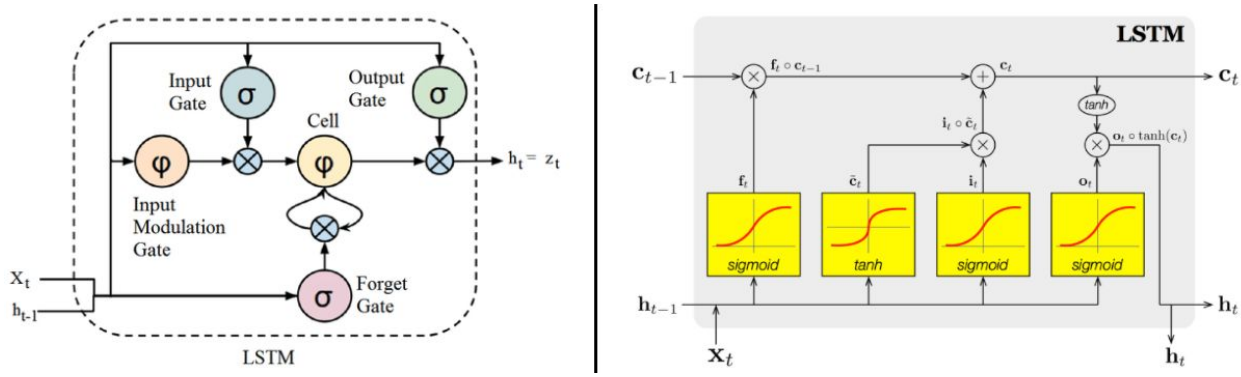
Τα ανατροφοδοτούμενα νευρωνικά δίκτυα (Recurrent Neural Networks), είναι μία κατηγορία νευρωνικού δικτύου η οποία χρησιμοποιεί δεδομένα χωρισμένα σε διακριτά χρονικά βήματα όπως οι χρονοσειρές [25]. Η διαφορά τους με τα απλά νευρωνικά δίκτυα έγκειται στο ότι εκτός από την κανονική είσοδο στην συνάρτηση ενεργοποίησης τους, παίρνουν και την ενεργοποίηση που είχαν στο προηγούμενο χρονικό βήμα τους. Χρησιμοποιώντας αυτήν την ενεργοποίηση που έχουν ως μνήμη μπορούν να έχουν μια διαμοιρασμένη αναπαράσταση των δεδομένων που έχουν ήδη επεξεργαστεί και να τη χρησιμοποιήσουν συνδυαστικά με τις καινούριες εισόδους. Αυτό επιτρέπει στα ανατροφοδοτούμενα δίκτυα να μπορούν να επεξεργαστούν εισόδους απροσδιόριστου μεγέθους όπως ομιλίες ή κείμενα



Σχήμα 2-14 : Παράδειγμα της κανονικής και της ξετυλιγμένης μορφής ενός ανατροφοδοτούμενου νευρωνικού δικτύου.
Πηγή: https://en.wikipedia.org/wiki/Recurrent_neural_network

2.10 Μονάδες Χρόνιας Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory)

Οι μονάδες χρόνιας βραχυπρόθεσμης μνήμης (Long Short-Term Memory - LSTM) αποτελούν μια αρχιτεκτονική τεχνητών ανατροφοδοτούμενων νευρωνικών δικτύων η οποία χρησιμοποιείται συχνά στον τομέα της βαθιάς μάθησης (Deep Learning). Σε αντίθεση με τα συμβατικά εμπρόσθια-τροφοδοτούμενα δίκτυα (feedforward networks), τα LSTM διαθέτουν συνδέσεις ανατροφοδότησης. Επιπλέον, έχουν τη δυνατότητα επεξεργασίας τόσο για μονάδες (πχ. εικόνες), όσο και για ολόκληρες ακολουθίες δεδομένων (πχ. ομιλία ή βίντεο). Συνήθεις εφαρμογές των LSTM είναι για παράδειγμα η αναγνώριση ομιλίας και η ανίχνευση ανωμαλιών σε κυκλοφοριακά δίκτυα ή συστήματα ανίχνευσης εισβολών (IDS).



Σχήμα 2-15: Μονάδα LSTM.

Μια κοινή μονάδα LSTM αποτελείται από ένα κύτταρο, μια πύλη εισόδου, μια πύλη εξόδου και μια πύλη απώλειας μνήμης. Το κύτταρο έχει τη δυνατότητα να απομνημονεύει τιμές της

εισόδου για αυθαίρετα χρονικά διαστήματα προγενέστερα της παρούσα τιμής ενώ οι τρεις πύλες ρυθμίζουν τη ροή πληροφοριών μέσα και έξω από το κύτταρο.

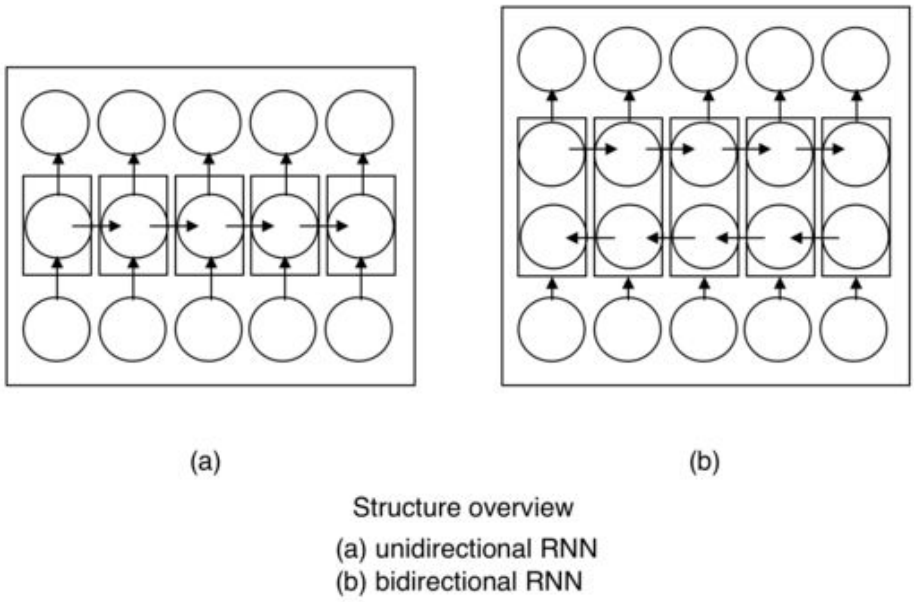
Οι μονάδες LSTM είναι κατάλληλες για την ταξινόμηση, επεξεργασία και πρόβλεψη που βασίζονται σε δεδομένα χρονοσειρών.

2.11 Αμφίδρομα Ανατροφοδοτούμενα Νευρωνικά Δίκτυα (Bidirectional Recurrent Neural Networks)

Τα αμφίδρομα ανατροφοδοτούμενα νευρωνικά δίκτυα (Bidirectional Recurrent Neural Networks - BRNN) είναι νευρωνικά δίκτυα τα οποία διασυνδέουν νευρώνες από δύο κρυφά επίπεδα αντίθετων κατευθύνσεων στην ίδια έξοδο. Έτσι, το επίπεδο εξόδου μπορεί να λαμβάνει ταυτόχρονα πληροφορία από προηγούμενες (προς τα πίσω) και ταυτόχρονα μελλοντικές (προς τα εμπρός) καταστάσεις. Τα BRNN εισήχθησαν το 1997 με σκοπό την αύξηση της πληροφορίας εισόδου που είναι διαθέσιμη στο δίκτυο συγκριτικά με άλλα δίκτυα.

Για παράδειγμα τα πολυστρωματικά Perceptron (Multilayer Perceptron - MLP) αφενός έχουν περιορισμούς στην ευελιξία των δεδομένων εισόδου, και αφετέρου απαιτούν τον εκ των προτέρων καθορισμό των δεδομένων εισόδου. Τα συμβατικά ανατροφοδοτούμενα νευρωνικά δίκτυα (RNN) έχουν επίσης περιορισμούς καθώς δεν προσφέρουν τη δυνατότητα εκ των προτέρων γνώσης για μελλοντικές πληροφορίες εισόδου από την τρέχουσα κατάσταση. Αντίθετα, τα BRNN δεν απαιτούν τον πλήρη καθορισμό των δεδομένων εισόδου εξ αρχής και επιπλέον οι μελλοντικές πληροφορίες-δεδομένα εισόδου είναι προσβάσιμες από την εκάστοτε τρέχουσα κατάσταση.

Η βασική αρχή των BRNN είναι ο διαχωρισμός των νευρώνων ενός συμβατικού RNN σε δύο κατευθύνσεις, μία για θετική χρονική κατεύθυνση (προς τα εμπρός) και άλλη για αρνητική κατεύθυνση χρόνου. Η έξοδος των δύο αυτών νευρώνων συνδέεται με εισόδους των καταστάσεων ίδιας κατεύθυνσης. Η γενική δομή των RNN και BRNN φαίνεται στη συνέχεια.



Σχήμα 2-16: BRNN vs RNN
Πηγή: https://en.wikipedia.org/wiki/Bidirectional_recurrent_neural_networks

Χρησιμοποιώντας δύο χρονικές κατευθύνσεις, η πληροφορία εισόδου από το παρελθόν και το μέλλον του τρέχοντος χρονικού πλαισίου μπορούν να χρησιμοποιηθούν, σε αντίθεση με τα συμβατικά RNN.

Τα BRNN μπορούν να εκπαιδευτούν χρησιμοποιώντας παρόμοιους αλγορίθμους με τα RNNs, επειδή οι δύο κατευθυντικοί νευρώνες δεν έχουν καμία αλληλεπίδραση μεταξύ τους. Ωστόσο, όταν εφαρμόζεται η μέθοδος οπισθοδιάδοσης (back-propagation), για τον υπολογισμό των βαρών του δικτύου απαιτούνται επιπρόσθετες διαδικασίες καθώς η ενημέρωση των επιπέδων εισόδου και εξόδου δεν μπορεί να γίνει ταυτόχρονα.

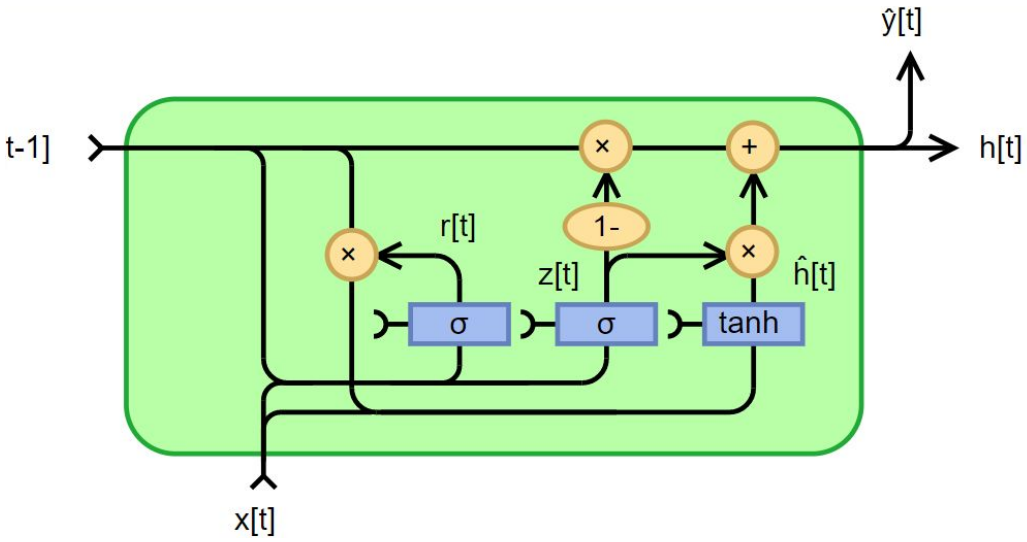
Οι πλέον συνήθεις εφαρμογές των BRNN είναι οι εξής:

- Αναγνώριση ομιλίας (σε συνδυασμό με χρόνια βραχυπρόθεσμη μνήμη)
- Μετάφραση
- Χειρόγραφη αναγνώριση
- Πρόβλεψη δομής πρωτεϊνών
- Κατηγοριοποίηση τμήματος ομιλίας

2.12 Ανατροφοδοτούμενες Μονάδες με Πύλες (Gated Recurrent Unit)

Οι Ανατροφοδοτούμενες Μονάδες με Πύλες είναι βελτιωμένη έκδοση του τυπικού ανατροφοδοτούμενου νευρωνικού δικτύου [26]. Οι ανατροφοδοτούμενες μονάδες με πύλες

(GRUs) χρησιμοποιούνται για την επίλυση του προβλήματος της εκλειπόμενης παραγώγου που υπήρχε στα απλά ανατροφοδοτούμενα δίκτυα. Ένα GRU είναι τυπικά ένα LSTM χωρίς την πύλη εξόδου. Το GRU χρησιμοποιεί δύο πύλες, μία πύλη ανανέωσης (update) και μία επαναφοράς (reset). Οι δύο πύλες αυτές καθορίζουν ποιες πληροφορίες θα δοθούν στην έξοδο. Ιδιαίτερα σημαντικό είναι ότι μπορούν να εκπαιδευτούν ώστε να αποθηκεύουν πολύ παλιές πληροφορίες.



Σχήμα 2-17 : Παράδειγμα μιας Ανατροφοδοτούμενης Μονάδας με Πύλες (GRU).
Πηγή: https://en.wikipedia.org/wiki/Gated_recurrent_unit

3. Βιβλιογραφική Επισκόπηση και Αποτελέσματα

Ο αυτόματος συγχρονισμός των υποτίτλων σε μία ταινία πρόκειται για ένα πρόβλημα που φαίνεται να έχει προσεγγιστεί στο κοντινό παρελθόν και έχουν δοθεί ικανοποιητικές λύσεις χωρίς, ωστόσο, να μπορεί να θεωρηθεί λυμένο. Οι προσεγγίσεις που παρουσιάζονται στη συνέχεια διακρίνονται με βάση τη στρατηγική που ακολουθούν. Κάθε στρατηγική αποτελεί μια μέθοδο με συγκεκριμένη σειρά βημάτων που επιλύει το πρόβλημα. Ωστόσο η υλοποίηση κάθε βήματος μπορεί να επιδέχεται διαφορετική προσέγγιση και να αντιμετωπίζεται διαφορετικά.

3.1 Στρατηγική 1 - Συγχρονισμός Υποτίτλων με Εντοπισμό Φωνής

3.1.1 Στρατηγική 1 - Συνοπτική Ανάλυση

Η στρατηγική αυτή στηρίζεται στη δημιουργία ενός συστήματος εντοπισμού φωνητικής δραστηριότητας (VAD - Voice Activity Detector) το οποίο θα χρησιμοποιεί στη συνέχεια το εξαγόμενο αρχείο ήχου της ταινίας προς συγχρονισμό [6]. Η μέθοδος αυτή προϋποθέτει ότι ο υπότιτλος προς συγχρονισμό είναι εξ ολοκλήρου μετατοπισμένος κατά ένα χρονικό παράθυρο και αρκεί να υπολογιστεί το μέγεθος της μετατόπισης αυτής. Με τη μετατόπιση κάθε υπότιτλου-διαλόγου ανεξάρτητα, κατά το ίδιο χρονικό διάστημα που υπολογίστηκε στο προηγούμενο βήμα, ολοκληρώνεται η διαδικασία συγχρονισμού.

Συγκεκριμένα τα βήματα που ακολουθούνται είναι τα εξής:

Βήμα 1

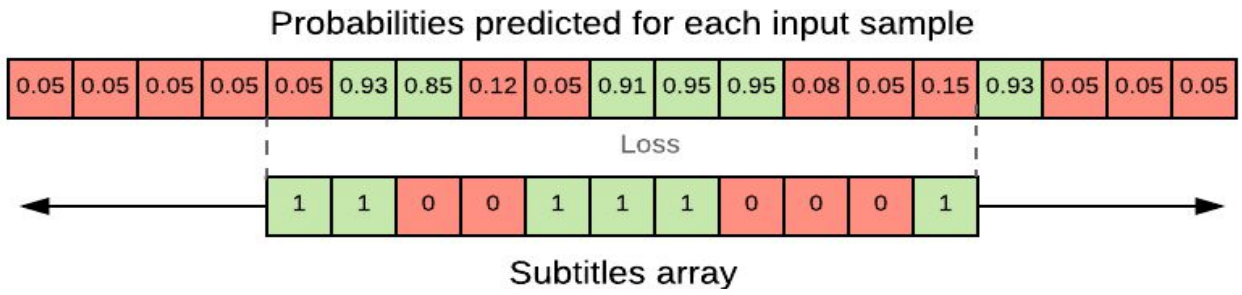
Αρχικά διαχωρίζεται και δειγματοληπτείται κατάλληλα το αρχείο της ταινίας ώστε να εξαχθεί το κομμάτι του ήχου, σε ασυμπίεστη μορφή wav.

Βήμα 2

Το σύστημα εντοπισμού φωνητικής δραστηριότητας (VAD) δέχεται ως είσοδο το αρχείο ήχου, το οποίο χωρίζεται στη συνέχεια σε ισομήκη τμήματα μικρής διάρκειας, ώστε καθένα από αυτά να κατηγοριοποιηθεί ως τμήμα που περιέχει ανθρώπινη φωνή ή όχι (με τη μορφή πιθανότητας που ξεπερνάει ένα κατώφλι ή όχι). Με τον τρόπο αυτό δημιουργείται ένα διάνυσμα, με στοιχεία την πιθανότητα το εκάστοτε τμήμα του ηχητικού αρχείου να περιέχει ομιλία.

Βήμα 3

Στη συνέχεια το αρχείο υπότιτλου αναλύεται και αυτό με τη σειρά του σε ισομήκη τμήματα (ίσης διάρκειας με αυτά που αναφέρθηκαν προηγουμένως), με βάση τους χρόνους που εμφανίζονται οι υπότιτλοι, και δημιουργείται ένα διάνυσμα (μικρότερου ή ίσου μήκους σε σχέση με το προηγούμενο). Το διάνυσμα αυτό περιέχει στοιχεία με τιμή 0 αν δεν αντιστοιχεί υπότιτλος σε εκείνο το τμήμα ή με τιμή 1 αν αντιστοιχεί.

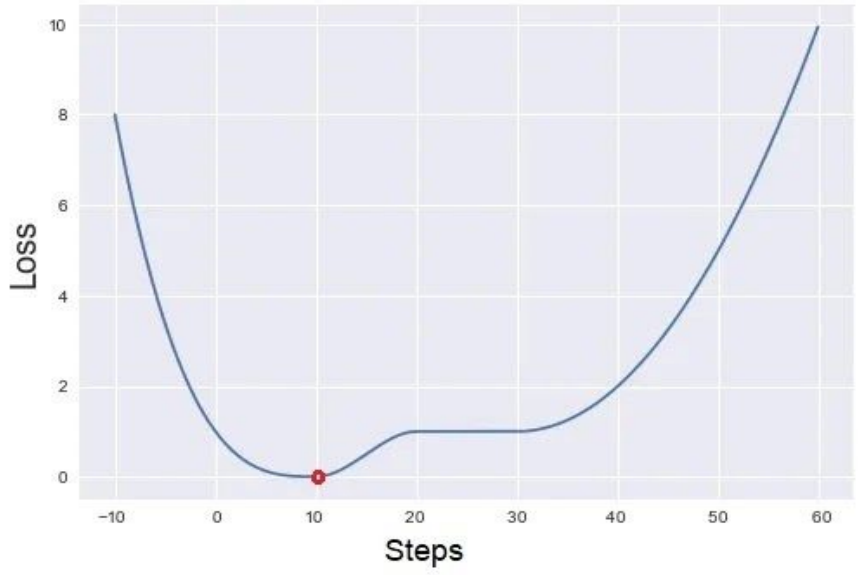


Σχήμα 3-1 : Τα δύο διανύσματα (ήχου και υπότιτλου) που πρέπει να αντιστοιχηθούν για να επιτευχθεί ο συγχρονισμός. Το πράσινο χρώμα υποδεικνύει ότι η πιθανότητα ξεπερνάει το κατώφλι που έχει οριστεί ώστε να θεωρείται τμήμα ομιλίας ενώ το κόκκινο όχι.

Πηγή: https://machinelearnings.co/automatic-subtitle-synchronization-e188a9275617

Βήμα 4

Έτσι πλέον το αρχικό πρόβλημα του συγχρονισμού ανάγεται σε ένα μη κυρτό διακριτό πρόβλημα βελτιστοποίησης - ελαχιστοποίησης κάποιας μετρικής ανομοιότητας ως προς τη χρονική μετατόπιση. Η επίλυση αυτού συνεπάγεται τη βέλτιστη αντιστοίχιση των δύο διανυσμάτων και κατ' επέκταση το συγχρονισμό. Συγκεκριμένα, ολισθαίνοντας επαναληπτικά το διάνυσμα του υποτίτλου (το μικρότερο διάνυσμα) κατά μία θέση κάθε φορά εξάγεται μια τιμή αξιολόγησης, μετά από κάθε ολίσθηση, μέσω της μετρικής ανομοιότητας. Το σύνολο των τιμών αυτών σε συνάρτηση με τα βήματα που μετατοπίστηκε το αρχικό διάνυσμα υποτίτλων σχηματίζει μια καμπύλη στο επίπεδο, της οποίας αναζητείται το ολικό ελάχιστο.



Σχήμα 3-2 : Ελαχιστοποίηση της μετρικής ανομοιότητας ως προς τον αριθμό βημάτων μετατόπισης του υπότιτλου για τη βέλτιστη αντιστοίχιση των διανυσμάτων ήχου και υποτίτλου.

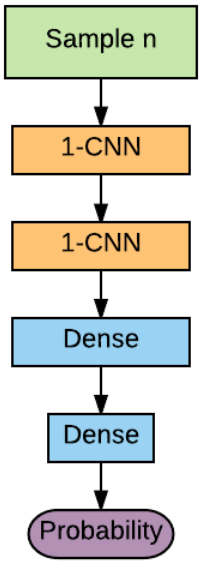
Βήμα 5

Ο αριθμός βημάτων που ελαχιστοποιεί τη συνάρτηση ανομοιότητας είναι αυτός με τον οποίο πρέπει να μετατοπιστεί ο αρχικός υπότιτλος, αφού μετατραπεί κατάλληλα σε χιλιοστά του δευτερολέπτου, ώστε να εξαχθεί ο συγχρονισμένος.

3.1.2 Στρατηγική 1 - Αναλυτική Προσέγγιση

Η προσέγγιση αυτή βασίζεται στη χρήση τεχνικών βαθιάς μάθησης (Deep Learning) αλλά και συνελικτικών νευρωνικών δικτύων (Convolutional Neural Networks) για την εκπαίδευση ενός ταξινομητή (Classifier) εντοπισμού δραστηριότητας φωνής [6].

Συγκεκριμένα, η εκπαίδευση γίνεται με τη χρήση ενός βαθύ νευρωνικού δικτύου (Deep Neural Network) που περιλαμβάνει αρχικά δύο μονοδιάστατα συνελικτικά επίπεδα (1D Convolutional Layers) και στη συνέχεια δύο έντονα συνδεδεμένα επίπεδα (Dense Layers).



Σχήμα 3-3: Αλληλουχία Επιπέδων Νευρωνικού Δικτύου για την Εκπαίδευση του Συστήματος Εντοπισμού Φωνητικής Δραστηριότητας.
Πηγή: https://machinelearnings.co/automatic-subtitle-synchronization-e188a9275617

Ως είσοδος του παραπάνω νευρωνικού δικτύου επιλέγεται μια σειρά χαρακτηριστικών (Features/Attributes), τα οποία εξάγονται από τα ηχητικά αρχεία των βίντεο που ανήκουν στο

σύνολο δεδομένων (Dataset). Τα χαρακτηριστικά αυτά είναι γνωστά ως Mel Frequency Cepstral Coefficient (MFCC).

Η διαδικασία που ακολουθεί προϋποθέτει ότι το ηχητικό σήμα δεν μεταβάλλεται στοχαστικά (στάσιμο) σε χρονική περίοδο 20 ms έως 40 ms. Το σήμα διακριτοποιείται σε χρονικά διαστήματα διάρκειας αυτής της τάξης, τα οποία αποκαλούνται πλαίσια (frames). Για κάθε πλαίσιο υπολογίζονται 12 σταθερές-χαρακτηριστικά οι οποίες στη συνέχεια τροφοδοτούν το δίκτυο-ταξινομητή κατά τη διαδικασία εκπαίδευσης.

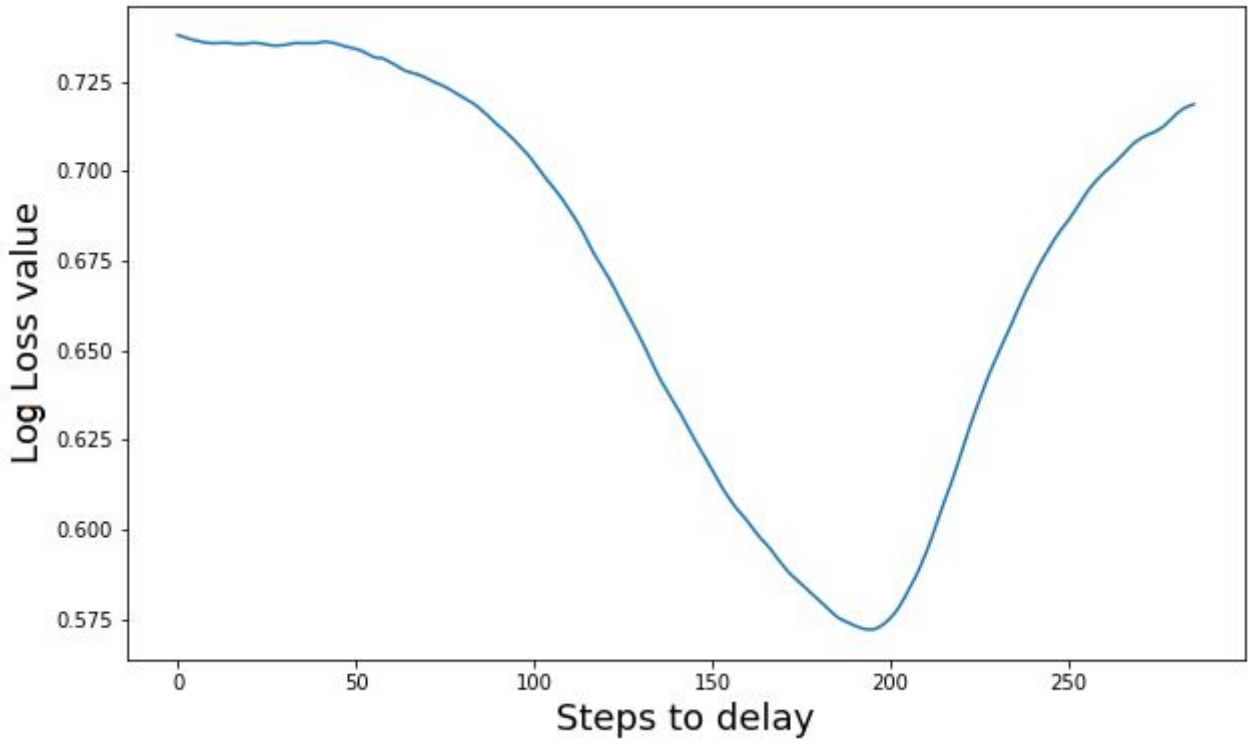
Αναφορικά με το χαρακτηριστικό κλάση (Class Attribute), που περιλαμβάνει τη θεωρητικά βέλτιστη τιμή εξόδου για την εκπαίδευση και στη συνέχεια την αξιολόγηση του εκπαιδευόμενου ταξινομητή, γίνεται για κάθε βίντεο, χρήση των αντίστοιχων συγχρονισμένων υποτίτλων, οι οποίοι μετατρέπονται σε μορφή διανύσματος όπως γίνεται και με τα ασυγχρόνιστα αρχεία.

Στη συγκεκριμένη περίπτωση έγινε χρήση ενός συνόλου δεδομένων που αποτελούνταν από δύο κεφάλαια διαφορετικών τηλεοπτικών σειρών και δύο ταινίες σε διαφορετικές γλώσσες, συνολικής διάρκειας πέντε ωρών. Η συχνότητα δειγματοληψίας για την παραγωγή των ηχητικών αρχείων που εξάγονται από τα αρχεία βίντεο είναι τα 16 KHz μονοφωνικού ήχου, ενώ για την εξαγωγή των χαρακτηριστικών MFCC από κάθε πλαίσιο χρησιμοποιούνται 512 δείγματα ήχου (ή 0,032 δευτερόλεπτα ήχου ανά πλαίσιο). Ως μετρική ανομοιότητας των διανυσμάτων πιθανοτήτων και υπότιτλου χρησιμοποιήθηκε η Λογαριθμική Συνάρτηση Κόστους (Log Loss / Cross Entropy Loss Function).

$$LogLoss(p,y) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot log_2(p_i) + (1-y_i)\cdot log_2(1-p_i)\right] \qquad (3.1)$$

Όπου p το διάνυσμα πιθανοτήτων του ηχητικού αρχείου, μεγέθους N και y το διάνυσμα υποτίτλων συμπληρωμένο με μηδενικά κατάλληλα, ώστε να έχει και αυτό μέγεθος N.

Τέλος, αξίζει να αναφερθεί η διάρκεια συγχρονισμού ενός υπότιτλου μετά το πέρας της εκπαίδευσης με σωστά ρυθμισμένες υπερ-παραμέτρους (hyperparameters). Παρατηρείται ότι η διάρκεια αυτή εξαρτάται σημαντικά από τη συνολική διάρκεια της ταινίας και το μήκος του υποτίτλου με αποτέλεσμα να εμφανίζονται προβλήματα μεγάλης καθυστέρησης στο συγχρονισμό ταινιών μεγάλου μήκους (πχ. 13 λεπτά για μια ταινία 100 λεπτών). Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί με τη χρήση ενός αναδρομικού αλγορίθμου «Διαίρει και Βασίλευε» (Divide and Conquer Algorithm) ώστε να μειωθεί σημαντικά ο χρόνος εύρεσης ελαχίστου του προβλήματος βελτιστοποίησης που συζητήθηκε.



Σχήμα 3-4 : Υπολογισμός της Μετρικής Ανομοιότητας Log Loss σε σχέση με τα τμήματα χρόνου που απαιτείται να μετατοπιστεί ο ασυγχρόνιστος υπότιτλο.
Πηγή: https://machinelearnings.co/automatic-subtitle-synchronization-e188a9275617

3.2 Στρατηγική 2 - Συγχρονισμός Υποτίτλων με Αναγνώριση Ομιλίας

3.2.1 Στρατηγική 2 - Συνοπτική Ανάλυση

Η στρατηγική αυτή στηρίζεται στη δημιουργία ενός συστήματος αναγνώρισης ομιλίας (Speech Recognition), σε συνδυασμό με ένα μεταφραστικό σύστημα, το μεταφραστή (Translator), όταν η γλώσσα του υπότιτλου διαφέρει από αυτή της ταινίας και ένα σύστημα συσχέτισης (Correlator) για τη συσχέτιση των υποτίτλων με τον ήχο της ταινίας [17], [18], [19], [20]. Η μέθοδος αυτή δεν προϋποθέτει ότι ο υπότιτλος προς συγχρονισμό είναι εξ ολοκλήρου μετατοπισμένος κατά ένα χρονικό παράθυρο, σε αντίθεση με τη στρατηγική 1, αλλά συγχρονίζει συνεχόμενα τμήματα μικρότερης διάρκειας διαδοχικά. Συγκεκριμένα, τα βήματα που ακολουθούνται είναι τα εξής:

Βήμα 1

Αρχικά δειγματοληπτείται κατάλληλα το αρχείο της ταινίας ώστε να εξαχθεί μια σειρά από μη επικαλυπτόμενα τμήματα ήχου, καθένα από τα οποία εισέρχεται σε ένα σύστημα αναγνώρισης ομιλίας.

Βήμα 2

Το σύστημα αναγνώρισης ομιλίας παράγει, για κάθε τμήμα ήχου, μια λίστα από λέξεις οι οποίες αναγνωρίστηκαν σε αυτό, καθώς επίσης και τις χρονοσφραγίδες (timestamps) που αντιστοιχούν στις λέξεις αυτές. Επίσης, για κάθε λέξη παράγεται και μία τιμή αξιολόγησης (Score) που αντικατοπτρίζει τη βεβαιότητα του συστήματος αναγνώρισης ομιλίας για τη συγκεκριμένη λέξη που αναγνωρίστηκε. Για κάθε θέση της λίστας που περιέχει τα τρία παραπάνω χαρακτηριστικά (word-timestamp-score) θα χρησιμοποιείται ο όρος «Λέξη» (Word).

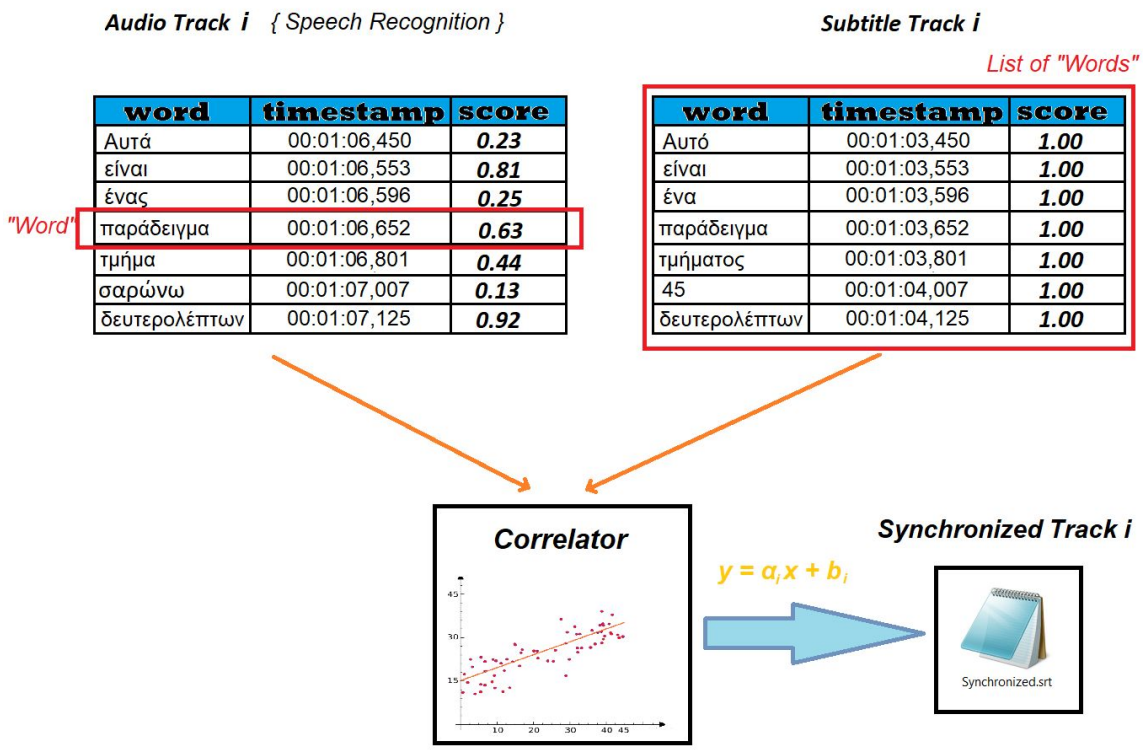
Η διαδικασία αναγνώρισης της ομιλίας πραγματοποιείται σε διαφορετικά νήματα (threads) ώστε να μειωθεί σημαντικά ο συνολικός χρόνος εκτέλεσης.

Βήμα 3

Όμοια με την παραπάνω διαδικασία το αρχείο του υποτίτλου αναλύεται και αυτό σε «Λέξεις» με τη μόνη διαφορά ότι η τιμή αξιολόγησης για κάθε «Λέξη» είναι πάντοτε 1. Αυτό συμβαίνει καθώς δεν εισάγεται καμία αβεβαιότητα για την αναγνώριση των «Λέξεων» αυτών, αφού όλες λαμβάνονται από το κείμενο του υποτίτλου.

Βήμα 4

Σε περίπτωση που η γλώσσα των υποτίτλων διαφέρει από τη γλώσσα της ταινίας υπεισέρχεται η ανάγκη για μετάφραση των διαλόγων της ταινίας στη γλώσσα του υποτίτλου. Για το σκοπό αυτό χρησιμοποιείται ο μεταφραστής (Translator), ο οποίος αναλαμβάνει να μεταφράσει τη λίστα «Λέξεων» της ταινίας στη γλώσσα των υποτίτλων.



Σχήμα 3-5 : Σχηματική Αναπαράσταση των Βημάτων 4 και 5 για το τμήμα ήχου-υπότιτλου i. Η έξοδος του συσχετιστή είναι μια ευθεία. Η τιμή x αντικαθίσταται από κάθε χρονοσφραγίδα του ασυγχρόνιστου τμήματος υπότιτλου i και παράγονται οι χρονοσφραγίδες y του συγχρονισμένου για εκείνο το τμήμα.

Βήμα 5

Στη συνέχεια λαμβάνει χώρα ο συσχετιστής (Correlator). Ο συσχετιστής δέχεται ως είσοδο δύο λίστες από «Λέξεις». Η μία λίστα περιλαμβάνει τις «Λέξεις» του ήχου της ταινίας και η άλλη τις «Λέξεις» των υποτίτλων. Χρησιμοποιώντας αυτές τις δύο λίστες και με βάση την ομοιότητα ανάμεσα σε κάθε πιθανό ζεύγος λέξεων γίνεται μια προσεγγιστική αντιστοίχηση των δύο λιστών και παράγεται μια νέα λίστα με τις αντιστοιχίσεις. Με βάση την τελευταία λίστα καθορίζεται το σφάλμα συγχρονισμού και συνεπώς η απόκλιση των υποτίτλων από τον ήχο.

Βήμα 6

Η παραπάνω διαδικασία επαναλαμβάνεται για όλα τα τμήματα ήχου-υποτίτλων και όλες οι έξοδοι του συσχετιστή μαζί με το αρχικό αρχείο των υποτίτλων εισέρχονται στο Συλλέκτη Υποτίτλων (Subtitles Collector), ο οποίος πραγματοποιεί τις μετατοπίσεις και τις διορθώσεις σε αυτό, δημιουργώντας το νέο συγχρονισμένο αρχείο υποτίτλων.

3.2.2 Στρατηγική 2 - Αναλυτική Προσέγγιση

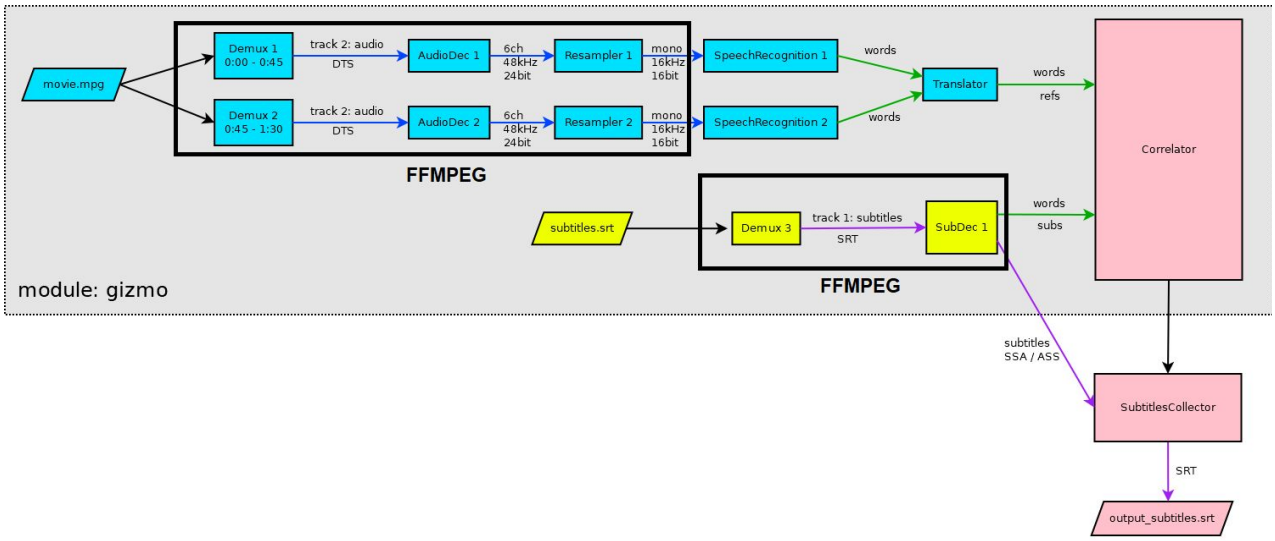
Η επίλυση αυτή [17], [18], [19], [20] χρησιμοποιεί σε πρώτο στάδιο το εργαλείο FFmpeg [21]. Αρχικά, το FFmpeg δέχεται ως είσοδο την ταινία και με τη χρήση πολυπλεκτών

(Demultiplexers [22]) την διαχωρίζει σε μικρότερα μη επικαλυπτόμενα τμήματα διάρκειας 45 δευτερολέπτων τα οποία περιλαμβάνουν μόνο το ηχητικό της μέρος. Στη συνέχεια, το κάθε τμήμα αποκωδικοποιείται (από τον AudioDec) και δειγματοληπτείται εκ νέου (Resampler) ώστε να αποκτήσει κατάλληλη μορφή για το επόμενο στάδιο, την αναγνώριση της ομιλίας.

Η αναγνώριση της ομιλίας (Speech Recognition) πραγματοποιείται με τη χρήση του εργαλείου CMUSphinx [23]. Το εργαλείο αυτό προτιμήθηκε καθώς άλλα εργαλεία που θεωρούνται αιχμής (State-of-the-Art), όπως το σύστημα αναγνώρισης ομιλίας που χρησιμοποιεί το YouTube, δεν έχουν αναρτήσει κώδικα διαθέσιμο στο κοινό (Open Source). Το CMUSphinx παρέχει αποτελέσματα με ακρίβεια της τάξης του 10% στις λέξεις που δημιουργεί, ποσοστό που αν και μικρό είναι αρκετό για τη σωστή λειτουργία της υλοποίησης.

Το εργαλείο αυτό θα δεχθεί ως είσοδο το κάθε τμήμα ήχου 45 δευτερολέπτων και θα παράξει μία λίστα από «Λέξεις». Στη συνέχεια περιγράφεται η διαδικασία για ένα τμήμα ήχου, η οποία ωστόσο, είναι ίδια για όλα τα τμήματα.

Η κάθε «Λέξη», όπως αναφέρθηκε και προηγουμένως, είναι μια δομή που περιέχει τη λέξη που εντόπισε το σύστημα αναγνώρισης ομιλίας, τη χρονοσφραγίδα της (timestamp) και μία τιμή αξιολόγησης (Score) που περιγράφει τη βεβαιότητα του συστήματος αναγνώρισης ομιλίας για τη συγκεκριμένη λέξη. Αυτή η λίστα «Λέξεων» που δημιουργήθηκε θα δοθεί ως είσοδος στο μεταφραστή.



Σχήμα 3-6 : Συνοπτική αρχιτεκτονική απεικόνιση της εφαρμογής SubSync.
Πηγή: http://sc0ty.pl/2019/04/subsync-architecture-overview/

Ο μεταφραστής (Translator) περιέχει μία σειρά από λεξικά (Dictionaries) διαφόρων γλωσσών τα οποία λαμβάνονται από το διαδίκτυο πριν ξεκινήσει η διαδικασία συγχρονισμού ανάλογα με τη γλώσσα της ταινίας και του υπότιτλου. Δέχεται ως είσοδο τη λίστα από «Λέξεις» του ηχητικού τμήματος και στη συνέχεια, για κάθε «Λέξη» της λίστας, προβαίνει σε μία διαδικασία

αναζήτησης της μετάφρασης στο κατάλληλο λεξικό. Έτσι ο μεταφραστής παράγει ως έξοδο μία νέα λίστα από μεταφρασμένες «Λέξεις» στη γλώσσα του υπότιτλου.

Ωστόσο, μια λέξη μπορεί να έχει πολλαπλές μεταφράσεις σε μια άλλη γλώσσα. Έτσι για κάθε πιθανή μετάφραση δημιουργείται μια νέα «Λέξη» στη μεταφρασμένη λίστα «Λέξεων» ώστε να ελεγχθούν όλες οι πιθανές ερμηνείες-μεταφράσεις που μπορεί να υπάρχουν. Οι διάφορες αυτές «Λέξεις» - μεταφράσεις έχουν, προφανώς, την ίδια χρονοσφραγίδα. Τελικά, η έξοδος του μεταφραστή είναι η (μεταφρασμένη) λίστα «Λέξεων» του ήχου.

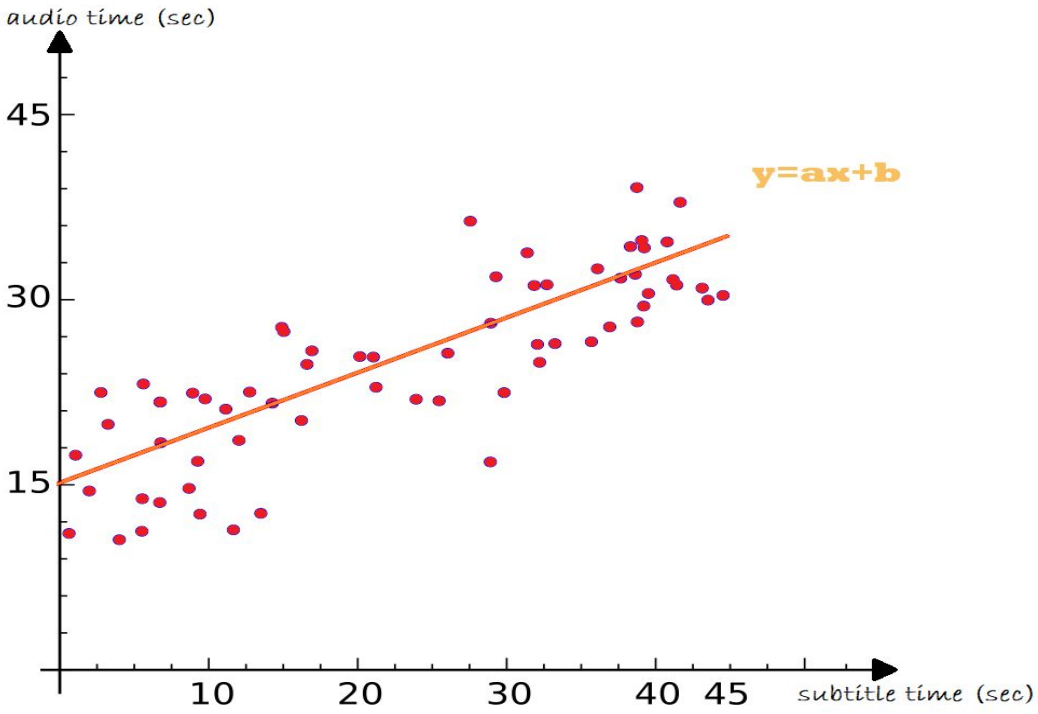
Όσον αφορά το αρχείο των υποτίτλων, όμοια διαχωρίζεται σε τμήματα των 45 δευτερολέπτων και ακολουθείται και πάλι μια διαδικασία παραγωγής μιας λίστας «Λέξεων», όπως αναφέρεται και στη συνοπτική ανάλυση. Ωστόσο, το βήμα της μετάφρασης για τη λίστα αυτή παραλείπεται. Έτσι τελικά είναι διαθέσιμες δύο λίστες «Λέξεων», αυτή με τις μεταφράσεις του ηχητικού τμήματος και αυτή του τμήματος των υποτίτλων, που είναι πλέον στην ίδια γλώσσα.

Οι λίστες αυτές δίνονται ως είσοδοι στο συσχετιστή (Correlator). Ο ρόλος του συσχετιστή είναι να επιλύσει το πρόβλημα συσχέτισης δύο λιστών.

Αρχικά με τη χρήση μιας συνάρτησης ομοιότητας συγκρίνεται κάθε λέξη της λίστας υπότιτλου με κάθε λέξη της λίστας ήχου. Συγκεκριμένα, σε κάθε σύγκριση ελέγχεται πόσα γράμματα της μίας λέξης είναι ίδια με τα αντίστοιχα γράμματα της άλλης λέξης και για κάθε συνδυασμό λέξεων αντιστοιχίζεται μια τιμή ομοιότητας (similarity score).

Στη συνέχεια οι συνδυασμοί λέξεων με τιμή ομοιότητας που ξεπερνάει ένα καθορισμένο κατώφλι, διατηρούνται ενώ οι υπόλοιποι συνδυασμοί απορρίπτονται. Έτσι δημιουργείται μια νέα, μικρότερη, κατά κανόνα, λίστα καθώς η χαμηλή ακρίβεια του συστήματος αναγνώρισης ομιλίας οδηγεί στην ύπαρξη συνδυασμών με χαμηλές τιμές ομοιότητας οι οποίες απορρίπτονται. Η νέα αυτή λίστα περιέχει σε κάθε θέση τις δύο χρονικές στιγμές που εμφανίζονται αντίστοιχα, οι δύο παρόμοιες λέξεις του τμήματος υπότιτλου και ήχου.

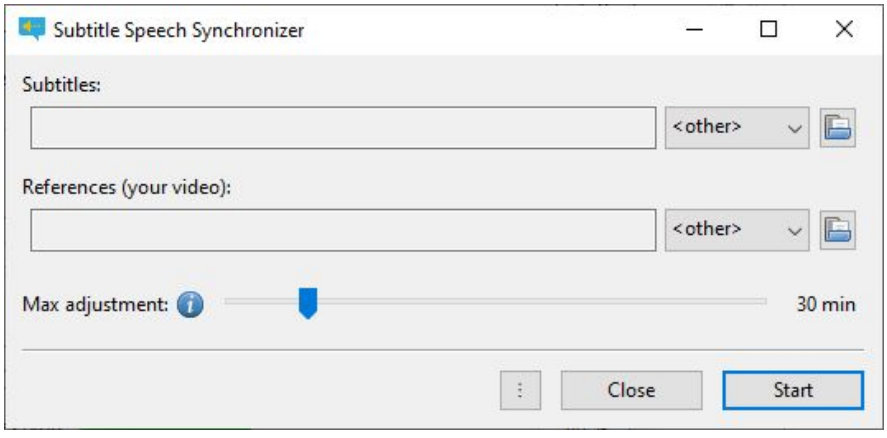
Έτσι πλέον το αρχικό πρόβλημα ανάγεται σε ένα πρόβλημα γραμμικής παλινδρόμησης όπου αναζητείται η βέλτιστη ευθεία που προσεγγίζει καλύτερα τα σημεία της νέας λίστας. Η κλίση της ευθείας αυτής συμβολίζει τη μεταβολή της ταχύτητας ενώ η μετατόπισή της τη χρονική καθυστέρηση που πρέπει να εφαρμοστεί στο αντίστοιχο τμήμα υπότιτλου.



Σχήμα 3-7: Γραφική απεικόνιση του προβλήματος γραμμικής παλινδρόμησης και της βέλτιστης ευθείας.

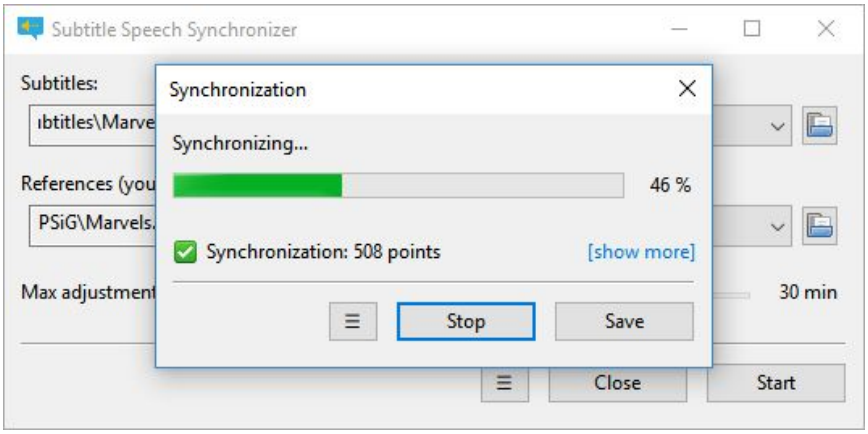
Η παραπάνω διαδικασία επαναλαμβάνεται για όλα τα τμήματα ήχου - υποτίτλων και παράγεται μια σειρά από βέλτιστες ευθείες (μία για κάθε τμήμα).

Οι ευθείες αυτές μαζί με το αρχικό αρχείο του υπότιτλου δίνονται ως είσοδος στο Συλλέκτη Υποτίτλων (Subtitles Collector). Εκεί, εφαρμόζονται όλες οι μετατοπίσεις-μετατροπές των χρονοσφραγίδων και δημιουργείται το νέο αρχείο συγχρονισμένων υποτίτλων.



Σχήμα 3-8 : Εφαρμογή SubSync - Αρχική Σελίδα
Πηγή : https://sc0ty.github.io/subsync/en/screens.html

Σημειώνεται ότι υπάρχει η δυνατότητα πρόωρης εξαγωγής του συγχρονισμένου υπότιτλου πριν από την ολοκλήρωση της εκτέλεσης του SubSync, όπου ο αλγόριθμος έχει συγχρονιστεί μόνο με βάση τα πρώτα λεπτά της ταινίας. Η δυνατότητα αυτή μπορεί να φανεί χρήσιμη σε περιπτώσεις όπου το αρχείο του υποτίτλου είναι εξ ολοκλήρου μετατοπισμένο κατά μια σταθερή χρονική τιμή και αρκεί να επεξεργαστούν τα πρώτα λεπτά της ταινίας ώστε να γίνει ο συγχρονισμός. Ωστόσο, το βέλτιστο αποτέλεσμα λαμβάνεται θεωρητικά όταν ολοκληρωθεί η διαδικασία.



Σχήμα 3-9 : Εφαρμογή SubSync - Σελίδα Συγχρονισμού
Πηγή : https://sc0ty.github.io/subsync/en/screens.html

4. Βασικά Εργαλεία και Δεδομένα

4.1 Εργαλεία

Τα εργαλεία που χρησιμοποιήθηκαν κατά την εκτέλεση της εργασίας παρουσιάζονται στη συνέχεια:

- FFmpeg (Εξαγωγή του ήχου από το αρχείο βίντεο)
- Microsoft Excel (Γραφική αναπαράσταση των αποτελεσμάτων)
- Keras (Βιβλιοθήκη της Python για Μηχανική Μάθηση)
- Tensorflow (Βιβλιοθήκη της Python για Μηχανική Μάθηση)
- CMUSphinx (Εργαλείο της Python για Συστήματα Αναγνώρισης Ομιλίας)
- WebRTC (Εργαλείο της Python για Συστήματα Εντοπισμού Φωνής)

Καθώς και οι απαραίτητες βιβλιοθήκες της Python. Οι γλώσσες προγραμματισμού που χρησιμοποιήθηκαν είναι:

- Python 3

4.2 Σύνολο Δεδομένων

Ως σύνολο δεδομένων για την εκπαίδευση των απαραίτητων ταξινομητών της εργασίας θα χρησιμοποιηθούν κατά βάση μια σειρά από 16 ταινίες διάρκειας 100’-150’ κατά μέσο όρο. Ως βάση αλήθειας (Ground Truth) θα χρησιμοποιηθούν οι συγχρονισμένοι αγγλικοί υπότιτλοι που αντιστοιχούν στην κάθε ταινία. Πριν τη διαδικασία της εκπαίδευσης θα πραγματοποιηθεί απο-συγχρονισμός των υποτίτλων οι οποίοι θα διαχωριστούν μετέπειτα σε σύνολο εκπαίδευσης (Training Set - 10 ταινίες) και σύνολο ελέγχου (Test Set - 6 ταινίες). Το σύνολο δεδομένων θα περιέχει διαφορετικών ειδών ταινίες σε διαφορετικές γλώσσες μαζί με τους αντίστοιχους υπότιτλους σε ελληνικά, αγγλικά (και ισπανικά σε ορισμένες περιπτώσεις). Συγκεκριμένα η λίστα με τις ταινίες και τους υπότιτλους φαίνεται στη συνέχεια:

Πίνακας 4.1

#	Τίτλος Ταινίας (Έτος Κυκλοφορίας)	Είδος Ταινίας	Γλώσσα Ταινίας	Γλώσσες Υπότιτλων
1	A Star Is Born (2018)	Μιούζικαλ, Ρομαντική, Δράμα	Αγγλικά	Αγγλικά, Ελληνικά, Ισπανικά
2	Barbara (2012)	Δράμα	Γερμανικά	Αγγλικά, Ελληνικά

3	Champions (2018)	Κωμωδία, Αθλητική	Ισπανικά	Αγγλικά, Ελληνικά
4	Coco (2017)	Κινουμένων Σχεδίων	Αγγλικά	Αγγλικά, Ελληνικά, Ισπανικά
5	Deadpool 2 (2018)	Δράσης, Κωμωδία	Αγγλικά	Αγγλικά, Ελληνικά
6	How To Train Your Dragon (2010)	Κινουμένων Σχεδίων	Αγγλικά	Αγγλικά, Ελληνικά, Ισπανικά
7	In The Aisles (2018)	Δράμα	Γερμανικά	Αγγλικά, Ελληνικά, Ισπανικά
8	La La Land (2016)	Μιούζικαλ, Κωμωδία, Δράμα	Αγγλικά	Αγγλικά, Ελληνικά
9	Mamma Mia (2008)	Μιούζικαλ, Ρομαντική	Αγγλικά, Ελληνικά	Αγγλικά, Ελληνικά, Ισπανικά
10	Mirage (2018)	Δράμα, Μυστηρίου	Ισπανικά	Αγγλικά, Ελληνικά
11	Mission Impossible - Fallout (2018)	Δράσης, Θρίλερ	Αγγλικά	Αγγλικά, Ελληνικά
12	Never Look Away (2018)	Ρομαντική, Δράμα	Γερμανικά	Αγγλικά, Ελληνικά
13	Pan's Labyrinth (2006)	Φαντασίας, Πολεμική	Ισπανικά	Αγγλικά, Ελληνικά
14	Skyfall (2012)	Δράσης	Αγγλικά	Αγγλικά, Ελληνικά, Ισπανικά
15	The Invisible Guest (2016)	Έγκλημα, Μυστηρίου, Θρίλερ	Ισπανικά	Αγγλικά, Ελληνικά
16	The Lives Of Others (2006)	Μυστηρίου, Θρίλερ	Γερμανικά	Αγγλικά, Ελληνικά, Ισπανικά

Το σύνολο αρχείων των ταινιών μπορεί να βρεθεί στον παρακάτω σύνδεσμο: **ταινίες**.

Το σύνολο των υποτίτλων των ταινιών μπορεί να βρεθεί στον παρακάτω σύνδεσμο: **υπότιτλοι**.

4.3 Αποσυγχρονισμός Υποτίτλων

Ο αποσυγχρονισμός των υποτίτλων των ταινιών γίνεται αυτόματα με μία ρουτίνα γραμμένη σε γλώσσα Python 3 (srt_creator.py). Ο αλγόριθμος αυτός δέχεται ως είσοδο ένα αρχείο .srt και παράγει τέσσερα αρχεία .srt. Το πρώτο περιλαμβάνει τους αρχικούς υπότιτλους όπου όλοι είναι μετατοπισμένοι κατά μία σταθερή καθυστέρηση μείων τριών δευτερολέπτων (-3000 ms). Τα υπόλοιπα τρία αρχεία περιλαμβάνουν τους αρχικούς υποτίτλους οι οποίοι έχουν υποστεί τυχαία-μεταβλητή μετατόπιση χωρίς, ωστόσο, να μεταβάλλεται η σειρά εμφάνισής τους. Η μετατόπιση αυτή κυμαίνεται μεταξύ μείων ενός έως μείων έξι δευτερολέπτων.

Αλγόριθμος 4.1 srt_creator(sub)

```
1:
2:  sub1 = copyfile(sub)
3:
4:  for line in sub1{
5:        line -= 3000   // (msec)
6:  }
7:
8:  for i in range(1,3){
9:       subs = copyfile(sub)
10:      limit = 0
11:      offset = 0
12:      for line in subs{
13:           offset = (line.start - limit) * random.number(0.3, 0.7) + 1000
14:           if(offset > 6000){
15:                offset = (line.start - limit) * random.number(0.01, 0.2) + 1000
16:           }else if(offset < 1250){
17:                offset = (line.start - limit) * random.number(0.8, 1) + 1000
18:                line -= offset
19:                limit = line.end
20:           }
21:      }
22:  }
```

4.4 Αξιολόγηση Συντονισμού δύο Υποτίτλων

Για την αξιολόγηση του συστήματος συγχρονισμού και συγκεκριμένα τη μέτρηση της ακρίβειας του δημιουργείται η ανάγκη για σύγκριση του παραγόμενου-συγχρονισμένου υπότιτλου με κάποιον υπότιτλο-αναφορά. Για το σκοπό αυτό κατασκευάζεται ένας αλγόριθμος ο οποίος στηρίζεται στη σύγκριση μεταξύ της εκκίνησης κάθε διαλόγου ανάμεσα στους δύο υπότιτλους, subs1.start[i] και subs2.start[j] όπως επίσης και τον τερματισμό του ίδιου διαλόγου subs1.end[i] και subs2.end[k].

Η διαδικασία αυτή εξηγείται στη συνέχεια για την αξιολόγηση ενός διαλόγου, ενώ αθροίζοντας τις τιμές αξιολόγησης όλων των διαλόγων λαμβάνεται η συνολική αξιολόγηση. Συγκεκριμένα, η διαδικασία που πραγματοποιείται για το διάλογο i του υπότιτλου subs1 είναι η εξής:

1. Ελέγχεται πότε ξεκινάει και πότε τελειώνει να εμφανίζεται. Έστω τη χρονική στιγμή subs1[i].start = τ και subs1[i].end = π αντίστοιχα.

2. Έπειτα ο αλγόριθμος αναζητά στο χρονικό διάστημα [τ-ε , τ+ε] αν ξεκινάει (και όχι αν απλά προβάλλεται) κάποιος διάλογος στο αρχείο subs2. Αν αυτό συμβαίνει τότε μια λογική μεταβλητή, startFlag παίρνει τιμή «Αληθής» (True).

3. Όμοια, ο αλγόριθμος αναζητά στο χρονικό διάστημα [π-ε , π+ε] αν εξαφανίζεται (και όχι αν γενικά δεν προβάλλεται) κάποιος διάλογος στο αρχείο subs2. Αν αυτό συμβαίνει τότε μια λογική μεταβλητή, endFlag παίρνει τιμή «Αληθής» (True).

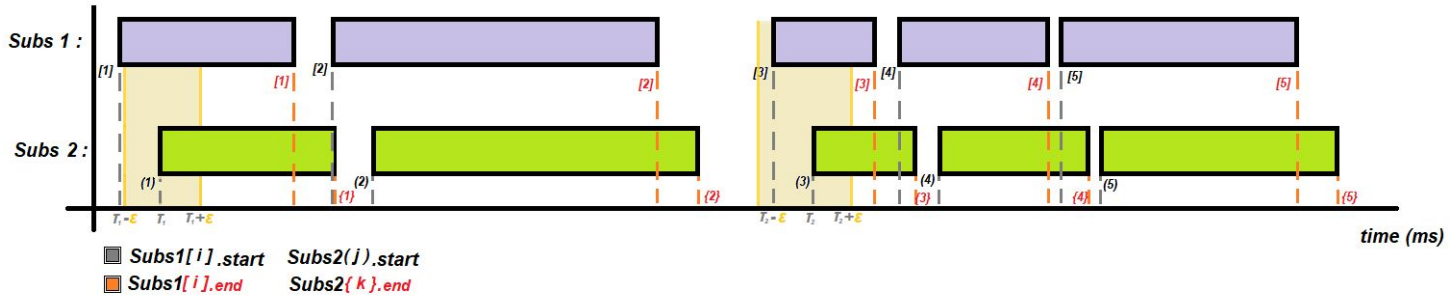
4. Αν και οι δύο λογικές μεταβλητές startFlag και endFlag είναι αληθείς, τότε το συνολικό score αυξάνεται γραμμικά σε σχέση με την απόσταση των διαλόγων κατά ένα παράγοντα:

$$score = score + maxReward \cdot \left(1 - \frac{d}{\varepsilon}\right) \quad (4.1)$$

   όπου maxReward η μέγιστη τιμή αξιολόγησης που μπορεί να λάβει ένας διάλογος (όταν τα σημεία εκκίνησης και τερματισμού ενός διαλόγου συμπίπτουν στα δύο αρχεία) και d η χρονική απόσταση των διαλόγων ανάμεσα στα δύο αρχεία. Παρατηρείται ότι για d = 0 (μέγιστος συγχρονισμός διαλόγου) προστίθεται η μέγιστη τιμή αξιολόγησης, maxReward, ενώ για d = ε η συνολική τιμή αξιολόγησης δεν αλλάζει.

   Διαφορετικά αν τουλάχιστον μία λογική μεταβλητή είναι ψευδής (False), τότε το συνολικό score μειώνεται ως εξής:

$$score = score - 1 \quad (4.2)$$



Σχήμα 4-1 : Γραφικό παράδειγμα των χρονικών διαστημάτων που παίζονται οι διάλογοι των αρχείων υποτίτλων subs1.srt και subs2.srt. Ο πρώτος διάλογος του subs1, (1), φαίνεται να ξεκινάει τη στιγμή τ_1 . Ωστόσο, στο διάστημα $[\tau_1-\varepsilon , \tau_1+\varepsilon]$ δεν ξεκινάει κάποιος διάλογος στο subs2 (κίτρινη περιοχή 1), επομένως τίθεται startFlag = False. Αντίθετα για τον τρίτο διάλογο του subs1, (3), φαίνεται ότι στο αντίστοιχο διάστημα στο subs2 ξεκινάει να εμφανίζεται ο υπότιτλος [3]. Επομένως τίθεται startFlag = True.

Ο αλγόριθμος που περιγράφηκε προηγουμένως παρουσιάζεται στη συνέχεια συνοπτικά, με τη μορφή ψευδοκώδικα.

Αλγόριθμος 4.2 srt_evaulator (subs1, subs2)

```
1:  score = 0
2:  threshold = 1000   // epsilon (msec)
3:  maxReward = 1.5
4:  j, k = 0
5:
6:  // subs2: subtitle with the earliest timestamp
7:   for (i = 0 : length(subs1) - 1){
8:      d = 0 // distance between corresponding dialogs
9:      // Check start times
10:    while (subs2[ j ].start < subs1[ i ].start - threshold){
11:          j = j + 1
12:    }
13:    if (subs2[ j ].start > subs1[ i ].start + threshold){
14:         startFlag = False
15:    }else{
16:         d = d + abs(subs2[ j ].start - subs1[ i ].start)/2
17:   }
18:     // Check end times
19:     while (subs2[ k ].end < subs1[ i ].end - threshold){
20:          k = k + 1
21:    }
```

```
22:    if (subs2[ k ].end > subs1[ i ].end + threshold){
23:          endFlag = False
24:    }else{
25:          d = d + abs(subs2[ k ].end - subs1[ i ].end)/2
26:    }
27:   // Calculate Score
28:    if (startFlag == True and endFlag == True){
29:          score += maxReward * ( 1 - d / threshold )
30:    }else{
31:          score -= 1
32:    }
```

Σημείωση: Ιδανικά, οι υπότιτλοι πρέπει να ξεκινούν με την έναρξη της ομιλίας και να σταματούν με το τέλος του τμήματος ομιλίας. Ωστόσο, υπότιτλοι που ξεκινούν μέχρι 1.5 δευτερόλεπτα πριν από την ομιλία και διαρκούν μέχρι 1.5 δευτερόλεπτα μετά την ολοκλήρωση της ομιλίας, θεωρούνται επίσης οριακά εντός του ιδανικού εύρους [2]. Η επιλογή του αυστηρότερου κατωφλίου, threshold, στο 1 δευτερόλεπτο έγινε με βάση αυτό το κριτήριο.

5. Αξιολόγηση Υλοποιήσεων και Σχολιασμός Αποτελεσμάτων

Στην ενότητα αυτή πραγματοποιείται αξιολόγηση πέντε υλοποιήσεων που εντοπίστηκαν στη βιβλιογραφία και η διανομή του κώδικα που χρησιμοποιήθηκε είναι ελεύθερη. Οι υλοποιήσεις αυτές ακολουθούν τη λογική των στρατηγικών που περιγράφονται στην ενότητα 3. Συμπτωματικά οι αλγόριθμοι-υλοποιήσεις έχουν την ίδια ονομασία, SubSync (Subtitle Synchronizer), επομένως διακρίνονται με βάση τα ψευδώνυμα των δημιουργών τους, όπως φαίνεται στον παρακάτω πίνακα.

Πίνακας 5.1

#	Ψευδώνυμο Δημιουργού SubSync	Αρχή Λειτουργίας	Γλώσσα Προγραμματισμού
1	Tympanix	Στρατηγική 1	Python 3
2	Smacke	Στρατηγική 1	Python 3
3	Sc0ty	Στρατηγική 2	C#
4	Kaegi	Άλλη Στρατηγική (με VAD)	Rust, Python 3
5	Koenkk (SubSync - PyAMC)	Στρατηγική 2	Python 3

Σημείωση:

Οδηγίες εγκατάστασης για την κάθε υλοποίηση μπορούν να βρεθούν στον παρακάτω σύνδεσμο: https://github.com/imanousar/Automatic-Subtitles-Synchronization/tree/master/Implemantations

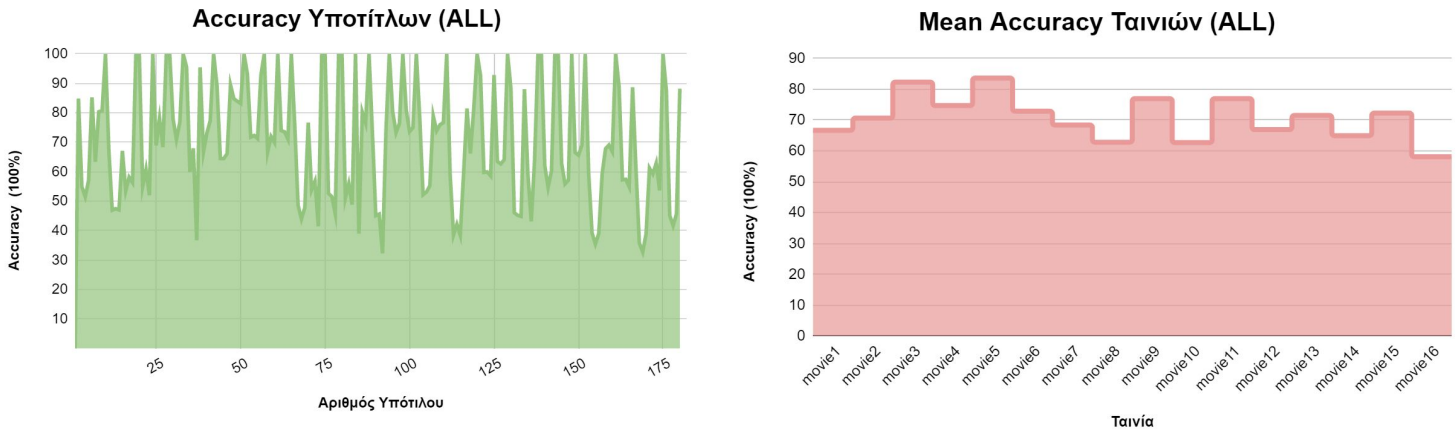
Οι υλοποιήσεις αυτές αξιολογούνται πάνω στο σύνολο δεδομένων που δημιουργείται με χρήση της ρουτίνας απο-συγχρονισμού όπως περιγράφεται στην ενότητα 4.1. Στη συνέχεια για κάθε ταινία και υπότιτλο, εκτελείται ο κώδικας κάθε δημιουργού και τέλος πραγματοποιείται αξιολόγηση της εκάστοτε υλοποίησης με χρήση του αλγόριθμου που περιγράφεται στην ενότητα 4.2.

Ως αποτέλεσμα της παραπάνω διαδικασίας, παρουσιάζονται στη συνέχεια στατιστικά δεδομένα αναφορικά με την ακρίβεια της κάθε υλοποίησης. Συγκεκριμένα, αναλύεται η ακρίβεια όλων των αρχείων υποτίτλων προς συγχρονισμό ξεχωριστά αλλά και σε ομάδες (ανά ταινία). Στη συνέχεια, επιπλέον, οι υπότιτλοι διακρίνονται σε δύο κατηγορίες με βάση τη μορφή απο-συγχρονισμού που έχουν υποστεί. Συγκεκριμένα, διακρίνονται σε υπότιτλους με (τυχαία) μεταβλητή μετατόπιση διαλόγων και σε υπότιτλους με σταθερή μετατόπιση (3 δευτερολέπτων). Τέλος, για τις υλοποιήσεις που περιλαμβάνουν αναγνώριση ομιλίας οι υπότιτλοι διακρίνονται επιπλέον και σε αγγλικούς/άλλους υπότιτλους για λόγους αξιολόγησης της μετάφρασης μετά

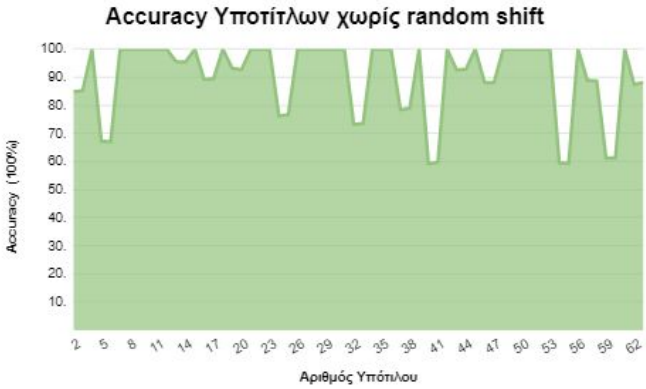
την αναγνώριση ομιλίας. Τέλος, αναφέρεται ο μέσος χρόνος εκτέλεσης συγχρονισμού. Πρέπει να σημειωθεί ότι η τιμή αυτή είναι ενδεικτική καθώς εξαρτάται τόσο από την υπολογιστική ισχύ που διαθέτει το εκάστοτε σύστημα όσο και από το μέγεθος της κάθε ταινίας.

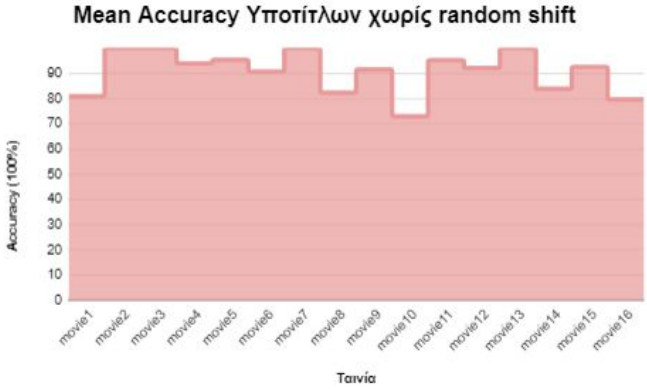
5.1 Υλοποίηση 1 - SubSync Tympanix

5.1.1 Αποτελέσματα-Drive



Σχήμα 5-1: Ακρίβεια μεταβλητά και σταθερά μετατοπισμένων υποτίτλων αντίστοιχα. Μέση Ακρίβεια για τους μεταβλητά αποσυγχρονισμένους καθώς και για τους σταθερά αποσυγχρονισμένους υπότιτλους κάθε ταινίας.





Σχήμα 5-2 : Ακρίβεια μεταβλητά και σταθερά μετατοπισμένων υποτίτλων αντίστοιχα. Μέση Ακρίβεια για τους μεταβλητά αποσυγχρονισμένους καθώς και για τους σταθερά αποσυγχρονισμένους υπότιτλους κάθε ταινίας.

Πίνακας 5.2

Μέσος Χρόνος Συγχρονισμού για έναν υπότιτλο.	**61.02 sec**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με μεταβλητή μετατόπιση. (%)**	***60.39***
Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο. (%)	**70.80**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με σταθερή μετατόπιση. (%)**	***90.65***
Ταινίες με Μέση Ακρίβεια > 60%	**15/16**		

5.1.2 Σχολιασμός Αποτελεσμάτων

Από τα παραπάνω διαγράμματα είναι εμφανές ότι η υλοποίηση αυτή είναι αρκετά αποδοτική. Συγκεκριμένα, φαίνεται ότι η μετρική ακρίβειας ξεπερνάει το ποσοστό των 60% επιτυχημένων συγχρονισμένων διαλόγων ανά ταινία, για 15/16 ταινίες, ενώ έχει μέση συνολική ακρίβεια 70,8%. Όπως ήταν αναμενόμενο, ότι η ακρίβεια συγχρονισμού είναι μειωμένη για υπότιτλους που έχουν υποστεί τυχαίο-μεταβλητό αποσυγχρονισμό σε σχέση με τους σταθερά μετατοπισμένους.

Μεταβλητή Μετατόπιση

Μειωμένη απόδοση παρουσιάζεται για τις ταινίες 7,8,16 δηλαδή το In the Aisles, το La La Land και το The Live of others. Οι ταινίες αυτές ανήκουν αντίστοιχα στις κατηγορίες Δράμα, Μιούζικαλ και Μυστηρίου, Θρίλερ. Το μιούζικαλ δικαιολογεί τη δυσκολία συγχρονισμού των υποτίτλων τους δεδομένου του θορύβου οποιασδήποτε μορφής (ηχητικά εφέ, τραγούδια κτλ.) που υπάρχουν σε αυτά τα είδη ταινιών. Όσον αφορά τις άλλες δύο ταινίες το πρόβλημα συγχρονισμού ενδεχομένως να οφείλεται στη χρήση της Γερμανικής αντί της αγγλικής

γλώσσας ή πιο πιθανό σε διάφορα άλλα εφέ των ταινιών (background μουσική, φωνές τρόμου κτλ.).Επιπλέον, ένα ακόμα αίτιο αδυναμίας συγχρονισμού υψηλής ακρίβειας για την ταινία La La Land είναι ότι πολύ διάλογοι είναι στίχοι τραγουδιών. Σε κάποια αρχεία υποτίτλων οι δημιουργοί αυτών επέλεξαν να μεταφράσουν τους στίχους των τραγουδιών ενώ σε άλλα όχι με αποτέλεσμα η ακρίβεια τόσο της διαδικασίας συγχρονισμού όσο και της διαδικασία αξιολόγησης αυτού να ελαττώνεται, καθώς η αξιολόγηση γίνεται με βάση τη σύγκριση δύο διαφορετικών αρχείων υποτίτλων.

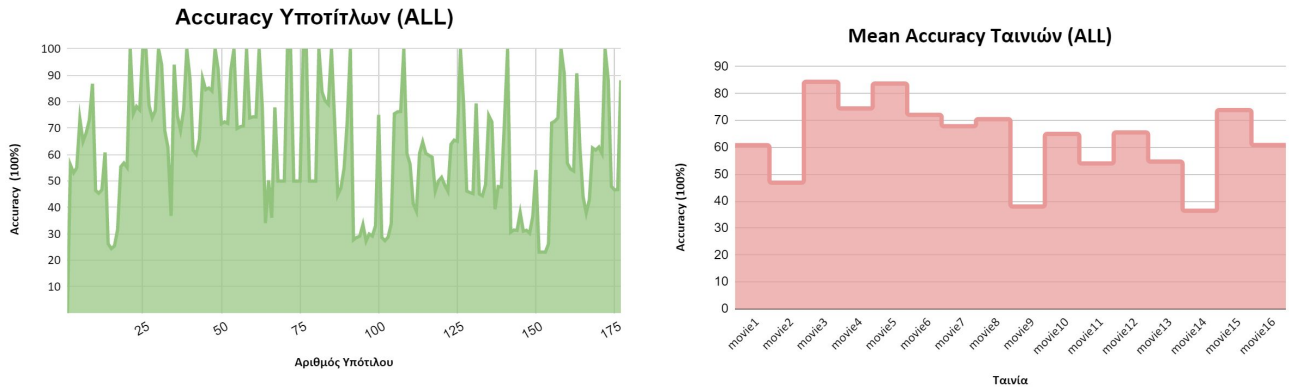
Σταθερή Μετατόπιση

Σχετικά με τους υπότιτλους που έχουν υποστεί σταθερή μετατόπιση τα αποτελέσματα είναι καλύτερα. Μειωμένη είναι η απόδοση στη ταινία 10, Mirage. Παρατηρώντας τις αριθμητικές τιμές της μετρικής αξιολόγησης φαίνεται ότι για την ταινία Mirage η ακρίβεια συγχρονισμού αυξήθηκε σε σύγκριση με τους υπότιτλους με μεταβλητή μετατόπιση, φτάνοντας το ποσοστό ακριβείας 73%.

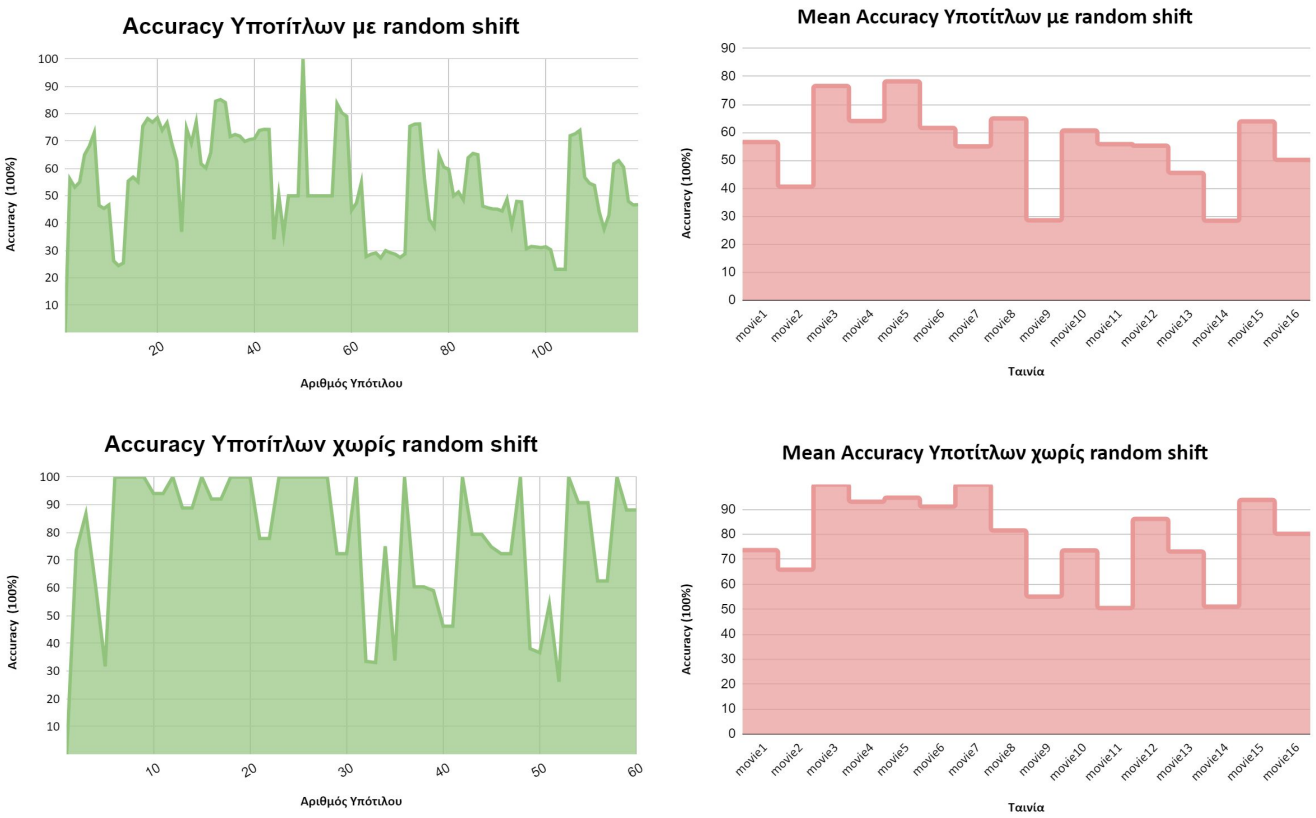
Τέλος, σημειώνεται ο μέσος χρόνος εκτέλεσης για κάθε υπότιτλο, ο οποίος είναι της τάξης των 61 δευτερολέπτων. Το παραπάνω χρονικό διάστημα είναι αρκετά σύντομο δεδομένης της όλης διαδικασίας που λαμβάνει χώρα μέχρι το συντονισμό.

5.2 Υλοποίηση 2 - SubSync Smacke

5.2.1 Αποτελέσματα-Drive



Σχήμα 5-3 : Ακρίβεια μεταβλητά και σταθερά μετατοπισμένων υποτίτλων αντίστοιχα. Μέση Ακρίβεια για τους μεταβλητά αποσυγχρονισμένους καθώς και για τους σταθερά αποσυγχρονισμένους υπότιτλους κάθε ταινίας.



Σχήμα 5-4 : Ακρίβεια μεταβλητά και σταθερά μετατοπισμένων υποτίτλων αντίστοιχα. Μέση Ακρίβεια για τους μεταβλητά αποσυγχρονισμένους καθώς και για τους σταθερά αποσυγχρονισμένους υπότιτλους κάθε ταινίας.

Πίνακας 5.3

Μέσος Χρόνος Συγχρονισμού για έναν υπότιτλο.	**40.35 sec**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με μεταβλητή μετατόπιση. (%)**	***55.36***
Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο. (%)	**63.08**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με σταθερή μετατόπιση. (%)**	***78.97***
Ταινίες με Μέση Ακρίβεια > 60%	**11/16**		

5.2.2 Σχολιασμός Αποτελεσμάτων

Από τα παραπάνω διαγράμματα είναι εμφανές ότι η υλοποίηση αυτή είναι μεν αποδοτική, αλλά όχι όσο η προηγούμενη. Συγκεκριμένα, φαίνεται ότι η μετρική ακρίβειας ξεπερνάει το ποσοστό των 60% επιτυχημένων συγχρονισμένων διαλόγων ανά ταινία, για 11/16 ταινίες, ενώ έχει μέση συνολική ακρίβεια 63,1%. Όπως ήταν αναμενόμενο, η ακρίβεια συγχρονισμού είναι μειωμένη για υπότιτλους που έχουν υποστεί τυχαίο-μεταβλητό αποσυγχρονισμό σε σχέση με τους σταθερά μετατοπισμένους.

Μεταβλητή Μετατόπιση

Μειωμένη απόδοση παρουσιάζεται για τις ταινίες 2, 9 και 14 δηλαδή τις Barbara, Mamma Mia και The Invisible Guest. Οι ταινίες αυτές ανήκουν αντίστοιχα στις κατηγορίες Δράμα, Μιούζικαλ και Έγκλημα-Θρίλερ, πράγμα που δικαιολογεί τη δυσκολία συγχρονισμού των υποτίτλων τους δεδομένου του θορύβου οποιασδήποτε μορφής (ηχητικά εφέ, τραγούδια κτλ.) που υπάρχουν σε αυτά τα είδη ταινιών. Παρατηρείται και πάλι η δυσκολία συγχρονισμού της ταινίας Mamma Mia, ωστόσο είναι αξιοσημείωτο ότι τα αποτελέσματα αυτής της μεθόδου είναι καλύτερα για τις ταινίες 4,6 και 11 σε σχέση με την προηγούμενη, στην περίπτωση της μεταβλητής μετατόπισης.

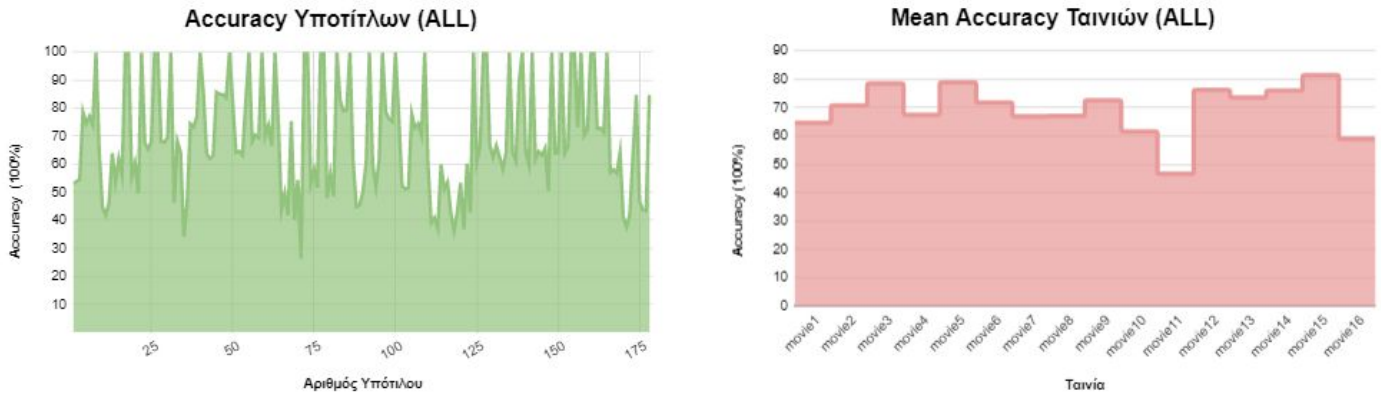
Σταθερή Μετατόπιση

Σχετικά με τους υπότιτλους που έχουν υποστεί σταθερή μετατόπιση τα αποτελέσματα είναι καλύτερα. Μειωμένη είναι η απόδοση στις ταινίες 9, 11 και 14 δηλαδή Mamma Mia, Mission Impossible Fallout και The Invisible Guest αντίστοιχα. Παρατηρώντας τις αριθμητικές τιμές της μετρικής αξιολόγησης φαίνεται ότι η ακρίβεια συγχρονισμού για τους υπότιτλους της ταινίας Mission Impossible Fallout είναι μειωμένη στην περίπτωση της σταθερής μετατόπισης. Αυτό μπορεί να οφείλεται στην ύπαρξη μη αγγλικών διαλόγων κατά την εκτέλεση της ταινίας οι οποίες μπορεί να μεταφράζονται ή όχι από τους δημιουργούς των υποτίτλων. Επιπλέον τα επίπεδα θορύβου είναι αρκετά υψηλά, όπως έχει αναφερθεί, ως ταινία δράσης.

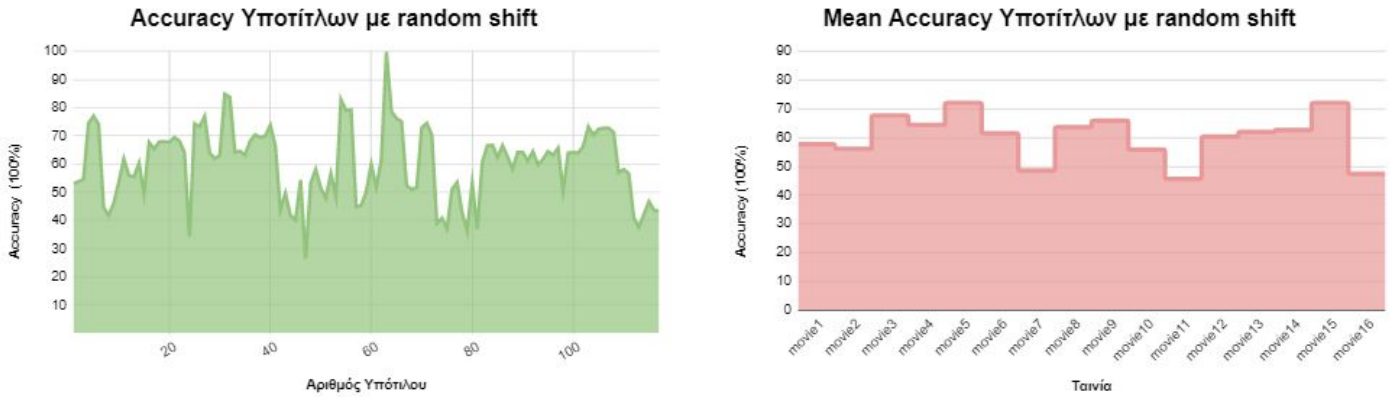
Γενικά, οι ταινίες 2,9,14 είχαν συνολικά κακό συγχρονισμό με τη ταινία 2 να έχει ποσοστό κάτω από 50% και τις άλλες δύο κάτω από 40%. Τέλος, σημειώνεται ο μέσος χρόνος εκτέλεσης για κάθε υπότιτλο, ο οποίος είναι της τάξης των 63 δευτερολέπτων. Το παραπάνω χρονικό διάστημα είναι αρκετά σύντομο δεδομένης της όλης διαδικασίας που λαμβάνει χώρα μέχρι το συντονισμό.

5.3 Υλοποίηση 3 - SubSync Sc0ty

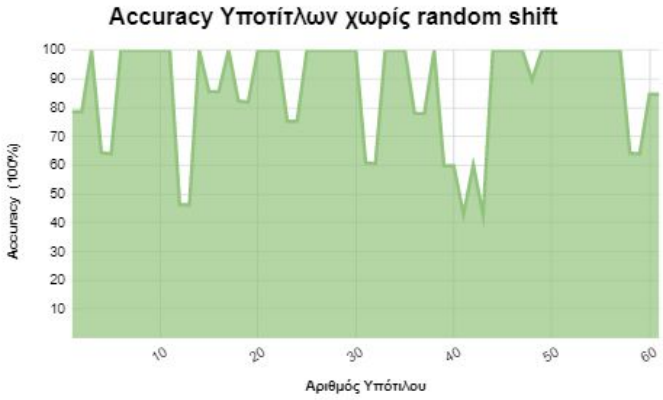
5.3.1 Αποτελέσματα-Drive

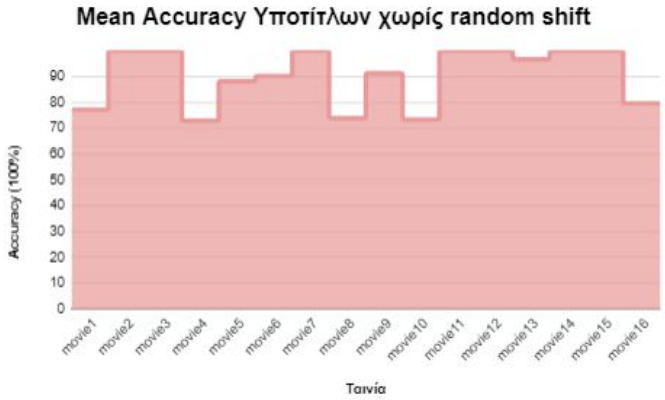


Σχήμα 5-5 : Ακρίβεια κάθε Υποτίτλου που ανήκει στο σύνολο δεδομένων και Μέση Ακρίβεια για τους υπότιτλους κάθε ταινίας.



Σχήμα 5-6: Ακρίβεια μεταβλητά μετατοπισμένων υποτίτλων. Μέση Ακρίβεια για τους μεταβλητά αποσυγχρονισμένους υπότιτλους κάθε ταινίας.





Σχήμα 5-7: Ακρίβεια σταθερά μετατοπισμένων υποτίτλων. Μέση Ακρίβεια για τους σταθερά αποσυγχρονισμένους υπότιτλους κάθε ταινίας.

Πίνακας 5.4

Μέσος Χρόνος Συγχρονισμού για έναν υπότιτλο.	**273.87 sec**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με μεταβλητή μετατόπιση. (%)**	***60.21***
Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο. (%)	**69.52**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με σταθερή μετατόπιση. (%)**	***90.14***
Ταινίες με Μέση Ακρίβεια > 60% (Καλός Συγχρονισμός)	**14/16**		

5.3.2 Σχολιασμός Αποτελεσμάτων

Από τα παραπάνω διαγράμματα είναι εμφανές ότι η υλοποίηση αυτή είναι αρκετά αποδοτική. Συγκεκριμένα, φαίνεται ότι η μετρική ακρίβειας ξεπερνάει το ποσοστό των 60% επιτυχημένων συγχρονισμένων διαλόγων ανά ταινία, για 14/16 ταινίες, ενώ έχει μέση συνολική ακρίβεια 69,53%. Όπως ήταν αναμενόμενο, η ακρίβεια συγχρονισμού είναι μειωμένη για υπότιτλους που έχουν υποστεί τυχαίο-μεταβλητό αποσυγχρονισμό σε σχέση με τους σταθερά μετατοπισμένους. Μάλιστα, η διαφορά είναι πολύ μεγάλη, περίπου 30%, με τους υπότιτλους με σταθερή μετατόπιση να συγχρονίζονται με ακρίβεια 90%, ενώ αυτοί με τυχαία-μεταβλητή μετατόπιση να συγχρονίζονται με ακρίβεια 60%.

Μεταβλητή Μετατόπιση

Μειωμένη απόδοση παρουσιάζεται για τις ταινίες 1, 2, 6, 12, 13, 14 που είναι κοντά στο 60%, ενώ αρκετά χαμηλή απόδοση έχουν οι ταινίες 7,11,16 με ποσοστό κάτω από το 50%.

Αναλυτικά, οι ταινίες με μέτρια απόδοση, περίπου 60% είναι οι A Star is Born (Αγγλική - Μιούζικαλ), Barbara (Γερμανική - Δράμα), How to train your dragon (Αγγλική - Κινουμένων Σχεδίων), Never Look Away (Γερμανική - Ρομαντική) , Pan’s Labyrinth (Ισπανική - Φαντασίας) και Skyfall (Αγγλική - Δράσης), ενώ οι ταινίες με χαμηλή απόδοση (< 50%) είναι οι εξής: In the Aisles (Γερμανική - Δράμα), Mission Impossible - Fallout (Αγγλική - Δράσης) και The lives of others (Γερμανική - Μυστηρίου, Θρίλερ).

Οι ταινίες με μέτρια απόδοση είναι τρεις αγγλικές ταινίες που είναι μιούζικαλ, κινουμένων σχεδίων και δράσης, και τρεις ξενόγλωσσες. Οι ταινίες με χαμηλή απόδοση είναι τρεις και εξ αυτών είναι ξενόγλωσσες ενώ η τρίτη είναι αγγλική και δράσης. Οι ξενόγλωσσες ταινίες και οι ταινίες δράσης φαίνεται πως δυσκολεύουν τον συγχρονισμό. Αυτό είναι και λογικό καθώς η μέθοδος αυτή βασίζεται στην αναγνώριση ομιλίας, επομένως είναι λογικό η απόδοση των μη αγγλικών ταινιών να είναι μειωμένη. Περαιτέρω, οι ταινίες δράσης λόγω θορύβου οποιασδήποτε μορφής (ηχητικά εφέ, εκρήξεις, τραγούδια κτλ.) δυσκολεύουν τον συγχρονισμό.

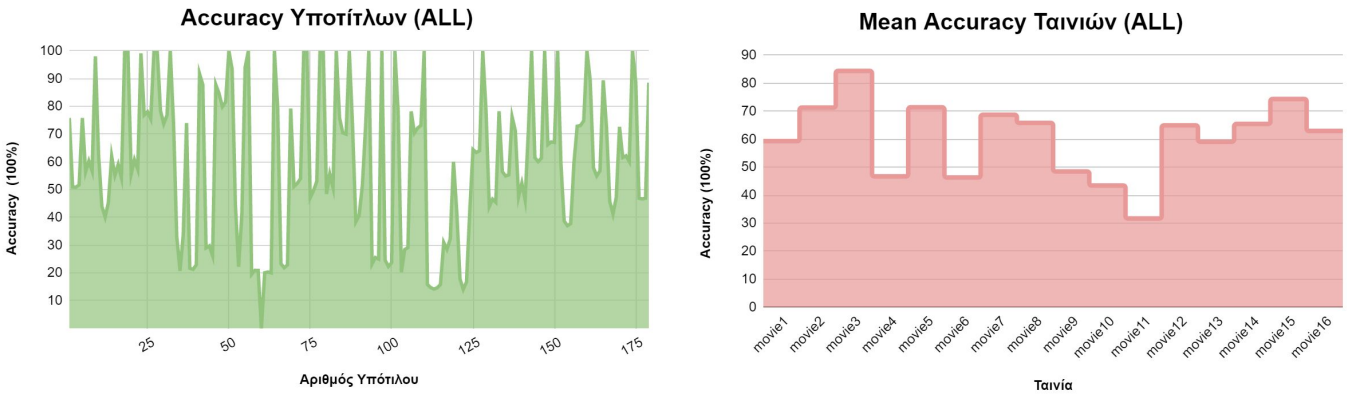
Σταθερή Μετατόπιση

Σχετικά με τους υπότιτλους που έχουν υποστεί σταθερή μετατόπιση τα αποτελέσματα είναι βελτιωμένα κατά πολύ. Η χειρότερα συγχρονισμένα ταινία είναι η ταινία 4 - Coco (κινουμένων Σχεδίων), με ποσοστό κοντά στο 72,5%. Επιπλέον, αξίζει να σημειωθεί ότι 7 ταινίες έχουν μέση ακρίβεια 100%.

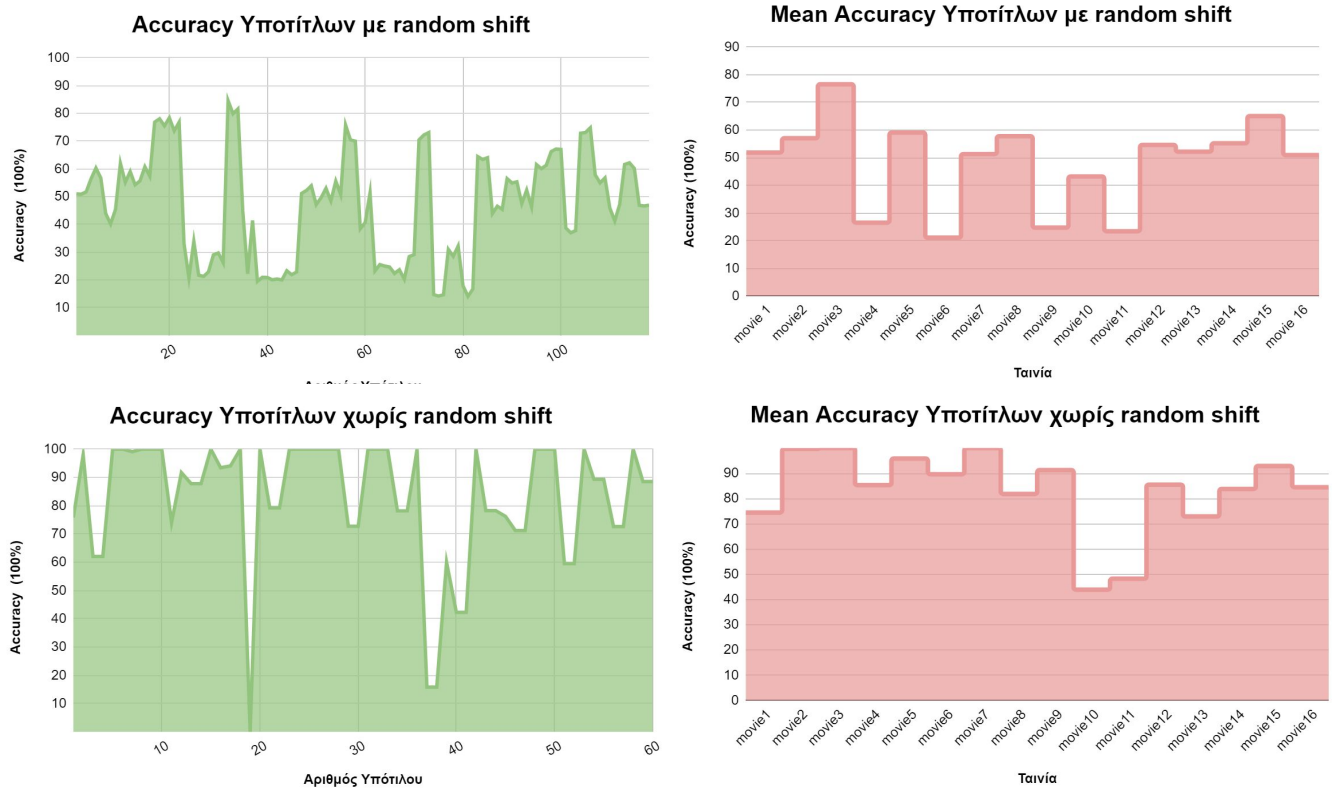
Τέλος, σημειώνεται ο μέσος χρόνος εκτέλεσης για κάθε υπότιτλο, ο οποίος είναι της τάξης των 4,5 λεπτών. Το παραπάνω χρονικό διάστημα δεν είναι τόσο σύντομο σε σύγκριση με άλλες μεθόδους που αναφέρθηκαν, ωστόσο είναι ικανοποιητικό.

5.4 Υλοποίηση 4 - SubSync Kaegi

5.4.1 Αποτελέσματα-Drive



Σχήμα 5-8: Ακρίβεια κάθε Υποτίτλου που ανήκει στο σύνολο δεδομένων και Μέση Ακρίβεια για τους υπότιτλους κάθε ταινίας.



Σχήμα 5-9: Ακρίβεια μεταβλητά και σταθερά μετατοπισμένων υποτίτλων αντίστοιχα. Μέση Ακρίβεια για τους μεταβλητά αποσυγχρονισμένους καθώς και για τους σταθερά αποσυγχρονισμένους υπότιτλους κάθε ταινίας.

Πίνακας 5.5

Μέσος Χρόνος Συγχρονισμού για έναν υπότιτλο.	**252.93 sec**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με μεταβλητή μετατόπιση. (%)**	***48.10***
Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο. (%)	**60.26**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με σταθερή μετατόπιση. (%)**	***83.10***
Ταινίες με Μέση Ακρίβεια > 60% (Καλός Συγχρονισμός)	**11/16**		

Από τα παραπάνω διαγράμματα είναι εμφανές ότι η υλοποίηση αυτή είναι σχετικά υψηλής απόδοσης. Συγκεκριμένα, φαίνεται ότι η μετρική ακρίβειας ξεπερνάει το ποσοστό των 60% επιτυχημένων συγχρονισμένων διαλόγων ανά ταινία, για 11/16 ταινίες, ενώ έχει μέση συνολική ακρίβεια 60,26%. Όπως ήταν αναμενόμενο, η ακρίβεια συγχρονισμού είναι μειωμένη για υπότιτλους που έχουν υποστεί τυχαίο-μεταβλητό αποσυγχρονισμό σε σχέση με τους σταθερά μετατοπισμένους.

Μεταβλητή Μετατόπιση

Μειωμένη απόδοση παρουσιάζεται για τις ταινίες 4, 8, 9 και 11 δηλαδή τις Coco, La La Land, Mamma Mia και Mission Impossible - Fallout. Οι ταινίες αυτές ανήκουν αντίστοιχα στις κατηγορίες Κινουμένων Σχεδίων, Μιούζικαλ, Μιούζικαλ, Δράσης-Θρίλερ πράγμα που δικαιολογεί τη δυσκολία συγχρονισμού των υποτίτλων τους δεδομένου του θορύβου οποιασδήποτε μορφής (ηχητικά εφέ, μουσική κτλ.) που υπάρχουν σε αυτά τα είδη ταινιών.

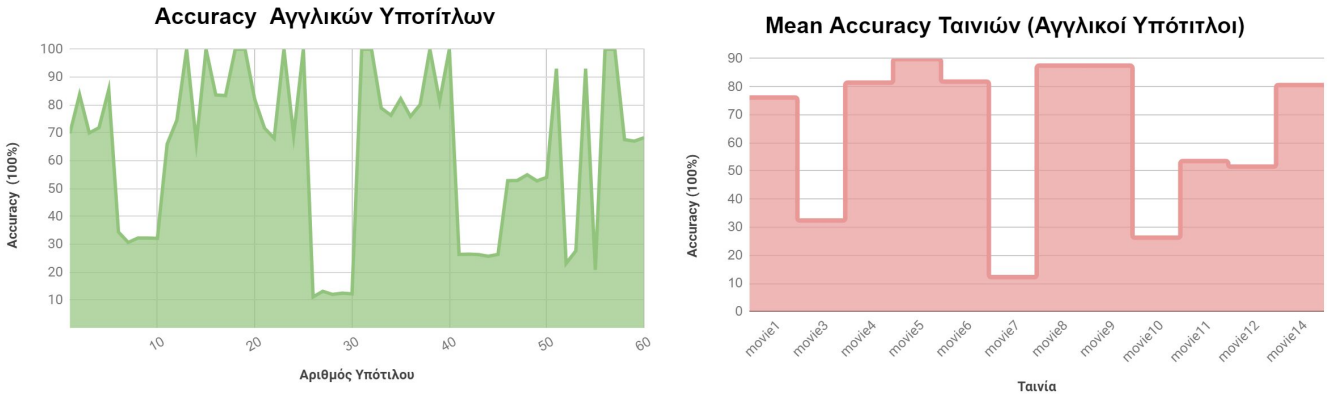
Σταθερή Μετατόπιση

Σχετικά με τους υπότιτλους που έχουν υποστεί σταθερή μετατόπιση τα αποτελέσματα είναι καλύτερα. Μειωμένη είναι η απόδοση στις ταινίες 10 και 11 δηλαδή τις Mirage και Mission Impossible - Fallout αντίστοιχα. Παρατηρώντας τις αριθμητικές τιμές της μετρικής αξιολόγησης φαίνεται ότι η ακρίβεια συγχρονισμού για τους υπότιτλους της ταινίας Mission Impossible Fallout είναι μειωμένη στην περίπτωση της σταθερής μετατόπισης. Αυτό μπορεί να οφείλεται στην ύπαρξη μη αγγλικών διαλόγων κατά την εκτέλεση της ταινίας οι οποίες μπορεί να μεταφράζονται ή όχι από τους δημιουργούς των υποτίτλων. Επιπλέον τα επίπεδα θορύβου είναι αρκετά υψηλά, όπως έχει αναφερθεί, ως ταινία δράσης.

Τέλος, σημειώνεται ο μέσος χρόνος εκτέλεσης για κάθε υπότιτλο, ο οποίος είναι της τάξης των 4 λεπτών. Το παραπάνω χρονικό διάστημα δεν είναι τόσο σύντομο σε σχέση με άλλες μεθόδους που έχουν αναφερθεί προηγουμένως.

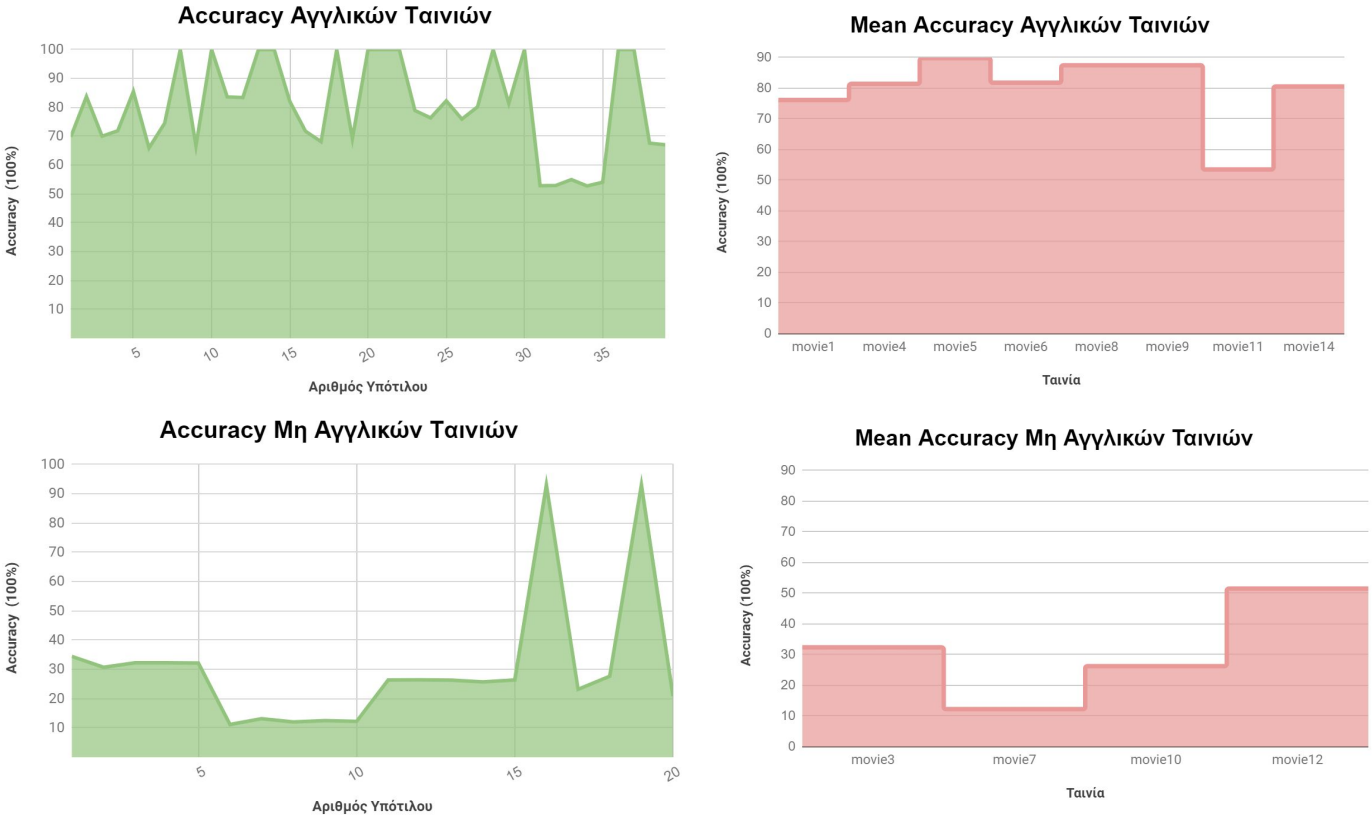
5.5 Υλοποίηση 5 - SubSync Koenkk

5.5.1 Αποτελέσματα-Drive



Σχήμα 5-10: Ακρίβεια κάθε Υποτίτλου που ανήκει στο σύνολο δεδομένων και Μέση Ακρίβεια για τους υπότιτλους κάθε ταινίας.

Σχήμα 5-11: Ακρίβεια μεταβλητά και σταθερά μετατοπισμένων υποτίτλων αντίστοιχα. Μέση Ακρίβεια για τους μεταβλητά αποσυγχρονισμένους καθώς και για τους σταθερά αποσυγχρονισμένους υπότιτλους κάθε ταινίας.



Σχήμα 5-12: Ακρίβεια Αγγλικών υποτίτλων και όλων των υπολοίπων ξεχωριστά. Μέση Ακρίβεια για τους αγγλικούς υπότιτλους κάθε ταινίας καθώς και για τους υπόλοιπους υπότιτλους.

Πίνακας 5.6

Μέσος Χρόνος Συγχρονισμού για έναν υπότιτλο.	**675.14 sec**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με μεταβλητή μετατόπιση. (%)**	***55.53***
Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο. (%)	**63.38**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με σταθερή μετατόπιση. (%)**	***75.16***
Ταινίες με Μέση Ακρίβεια > 60% (Καλός Συγχρονισμός)	**7/12**		

5.5.2 Σχολιασμός Αποτελεσμάτων

Αρχικά, όπως φαίνεται από τα αποτέλεσματα η υλοποίηση αυτή είχε συγκριτικά πιο χαμηλές αποδόσεις με κυριότερο μειονέκτημα ότι 4 από τις 16 ταινίες, δηλαδη 25% απέτυχαν να συγχρονιστούν. Οι ταινίες 2, 13, 15, 16 είναι οι ταινίες που απέτυχαν να συγχρονιστούν και έχουν ως κοινό χαρακτηριστικό ότι όλες είναι μη αγγλόφωνες ταινίες. Περαιτέρω, οι ταινίες 3, 7

και 10, μη αγγλόφωνες και αυτές, παρουσίασαν πολύ κακή ακρίβεια συγχρονισμού, με ακρίβεια για όλες κάτω από 33%, ενώ επιπλέον η τελευταία μη αγγλόφωνη ταινία, η 12, παρουσίασε περίπου 52% ακρίβεια.

Όσον αφορά τις αγγλόφωνες ταινίες τα ποσοστά είναι αισθητά βελτιωμένα με ποσοστά κοντά στο 80%. Εξαίρεση αποτελεί η ταινία 11, Mission Impossible - Fallout. Αυτό ενδεχομένως να οφείλεται στο ότι είναι ταινία δράσης και τα σημεία στην ταινία με θορύβους και μουσική είναι συχνά.

Τέλος, σχετικά με την τυχαία-μεταβλητή και σταθερή μετατόπιση τα αποτελέσματα είναι ανάμεικτα και στις δύο περιπτώσεις Ωστόσο, συγκρίνοντας τους δύο πίνακες φαίνεται ότι τα χαμηλά ποσοστά και στις δύο περιπτώσεις (με ή χωρίς random shift) αντιστοιχούν στις μη αγγλόφωνες ταινίες ενώ οι αγγλόφωνες ταινίες παρουσιάζουν υψηλά ποσοστά ανεξαρτήτως αν έχουν τυχαία-μεταβλητή ή σταθερή μετατόπιση.

5.6 Συμπεράσματα

Από τα παραπάνω διαγράμματα εξάγεται το συμπέρασμα ότι η πρώτη υλοποίηση είναι η πλέον αποτελεσματική, καθώς η συνολική μέση ακρίβεια είναι η μέγιστη ενώ συγχρόνως και ο χρόνος συγχρονισμού είναι ο μικρότερος όλων των υλοποιήσεων. Για το λόγο αυτό στη συνέχεια επιλέγεται ο κώδικας που αντιστοιχεί σε αυτή την υλοποίηση ώστε να πραγματοποιηθεί μια απόπειρα περαιτέρω βελτίωσης μέσω ορισμένων τροποποιήσεων στο μοντέλο ταξινόμησης που χρησιμοποιείται.

6. Υλοποίηση SubSync - Auth

Στη συνέχεια παρουσιάζονται οι δύο νέες υλοποίησεις οι οποίες υλοποιήθηκαν από τους συγγραφείς της εργασίας. Στηρίζονται στην πρώτη στρατηγική και αποτελούν τροποποιημένη μορφή της πρώτης υλοποίησης (SubSync - Tympanix). Οι τροποποιήσεις αφορούν τη δομή του μοντέλου ταξινόμησης για τη διαδικασία εντοπισμού φωνητικής δραστηριότητας που λαμβάνει χώρα πριν το συγχρονισμό.

Τα νευρωνικά δίκτυα που υλοποιήθηκαν περιέχουν μονοδιάστατα συνελικτικά επίπεδα (1D Convolutional Layer), έντονα συνδεδεμένα επίπεδα και μία μονάδα-επίπεδο που ανήκει στην κατηγορία των ανατροφοδοτούμενων νευρωνικών δικτύων (Recurrent Neural Networks). Τα δίκτυα αυτά παρουσιάζουν υψηλή απόδοση σε εφαρμογές εντοπισμού χαρακτηριστικών που σχετίζονται χρονικά [25].

6.1 Ανάλυση Υλοποίησης SubSync - Auth

6.1.1 Διαδικασία εκπαίδευσης

Συνολικά, δημιουργήθηκαν 8 μοντέλα τα οποία προέκυψαν ως αποτέλεσμα συνδυασμού των εξής επιλογών: α) ένα ή δύο μονοδιάστατα συνελικτικά επίπεδα β) αμφίδρομο ή όχι, ανατροφοδοτούμενο νευρωνικό δίκτυο γ) ανατροφοδοτούμενο νευρωνικό δίκτυο με αναδρομικές μονάδες πύλης (Gated Recurrent Unit - GRU) ή αμφίδρομο ανατροφοδοτούμενο δίκτυο χρόνιας βραχυπρόθεσμης-μνήμης (Long Short-Term Memory - LSTM). Σε όλα τα μοντέλα, το τελευταίο στάδιο αποτελείται από δύο έντονα συνδεδεμένα επίπεδα (Dense Layers) και έναν νευρώνα με σιγμοειδή ενεργοποίηση. Επομένως τα μοντέλα πραγματοποιούν δυαδική ταξινόμηση (Binary Classification) για την ύπαρξη ή όχι φωνητικής δραστηριότητας σε συγκεκριμένο τμήμα ήχου.

Η δημιουργία των μοντέλων έγινε σε Python 3 με χρήση της βιβλιοθήκης Keras. Οι υπόλοιπες παράμετροι και τεχνικές είναι ίδιες και για τα δύο μοντέλα.

Ως δεδομένα εισόδου χρησιμοποιήθηκαν τα χαρακτηριστικά MFCC των τμημάτων ήχου των ταινιών, ενώ ως βάση αλήθειας (Ground Truth) χρησιμοποιήθηκαν τα αρχεία συγχρονισμένων αγγλικών υποτίτλων - το τμήμα ήχου ανήκει στο χαρακτηριστικό 1 αν υπάρχει ενεργός διάλογος ή στο 0 αν δεν υπάρχει.

Το σύνολο των δεδομένων εκπαίδευσης εξισορροπήθηκε, ώστε κάθε κλάση να έχει τον ίδιο αριθμό δειγμάτων.

Η εκπαίδευση γίνεται σε εποχές. Ο μέγιστος αριθμός τίθεται στις 50 εποχές, ενώ γίνεται πρόωρος τερματισμός (Early Stopping) σε περίπτωση που η τιμή της συνάρτησης κόστους γίνει μικρότερη από 0.001 και παραμείνει μικρή για 10 διαδοχικές εποχές. Κατά μέσο όρο τα μοντέλα εκπαιδεύτηκαν για 40 εποχές.

Ως συνάρτηση κόστους/απώλειας (Cost/Loss Function) χρησιμοποιείται η συνάρτηση μέσου τετραγωνικού σφάλματος (Mean Squared Error Loss Function) ενώ ως «βελτιστοποιητής» (Optimizer) ο αλγόριθμος Adam (Adaptive Moment Estimation).

Ως συναρτήσεις ενεργοποίησης (Activation Functions) επιλέγονται οι σιγμοειδής (Sigmoid) και ενεργοποίηση ανόρθωσης (ReLU). Η σιγμοειδής συνάρτηση ενεργοποίησης χρησιμοποιείται μόνο στο τελευταίο επίπεδο (επίπεδο εξόδου) ώστε να απεικονίσει την έξοδο στο εύρος $(0, 1)$. Σε όλα τα υπόλοιπα επίπεδα γίνεται χρήση της ενεργοποίησης ανόρθωσης, καθώς έχει την ικανότητα να αποτρέπει το φαινόμενο εκλειπόμενης παραγώγου (Vanishing Gradients) κατά το οποίο η τιμή των βαρών παύει να μεταβάλλεται για μεγάλο πλήθος επιπέδων - νευρώνων ενώ ταυτόχρονα είναι υπολογιστικά φθηνή.

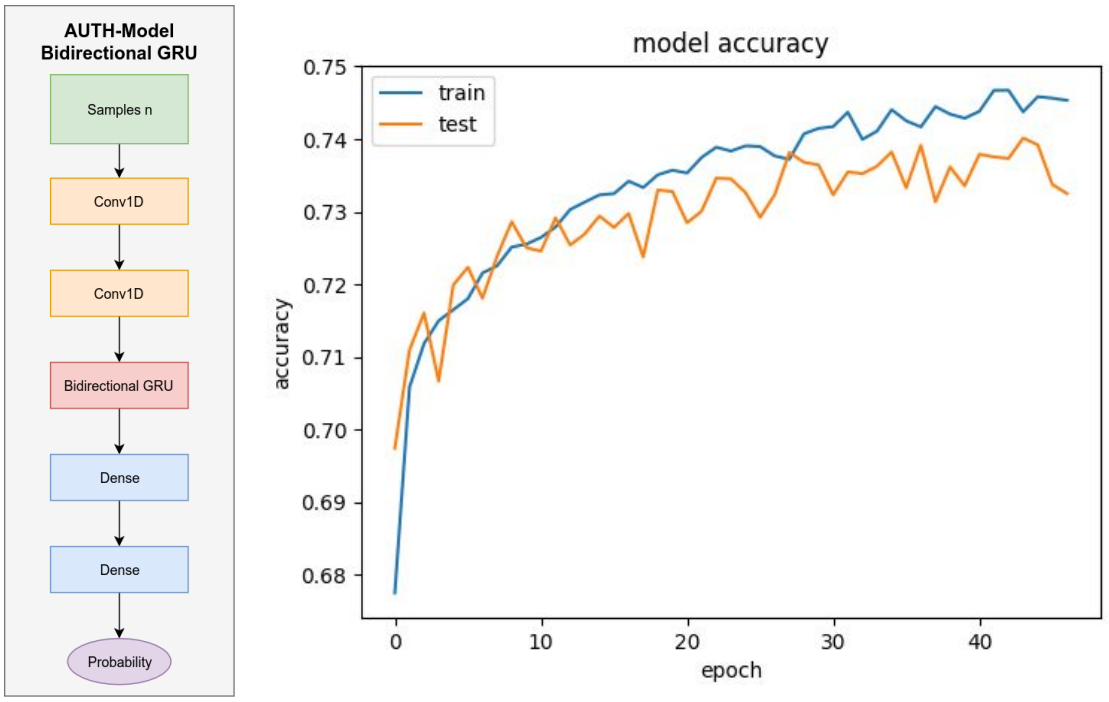
Επίσης χρησιμοποιούνται και δύο τεχνικές βελτιστοποίησης πριν από κάθε έντονα συνδεδεμένο επίπεδο. Γίνεται χρήση της τεχνικής-λειτουργίας ‘Dropout’, η οποία κατά την εκπαίδευση απενεργοποιεί έναν αριθμό (στην προκειμένη περίπτωση 20%) τυχαία επιλεγμένων νευρώνων του δικτύου, προκειμένου αφενός να αποφευχθούν φαινόμενα υπερ-εκπαίδευσης και αφετέρου το μοντέλο να αποκτήσει ανοχή στο θόρυβο. Επιπλέον, γίνεται χρήση και της τεχνικής κανονικοποίησης παρτίδας (Batch Normalization) προκειμένου να αυξηθεί η ευστάθεια του δικτύου κανονικοποιώντας (standardization) την έξοδο κάθε νευρώνα με κατάλληλο τρόπο.

Τέλος, κατά τη διαδικασία της εκπαίδευσης, ένα τμήμα των δεδομένων εκπαίδευσης (30%) χρησιμοποιείται για επικύρωση (Validation) μετά από κάθε εποχή, ως ένα επιπρόσθετο μέτρο αποφυγής της υπερεκπαίδευσης.

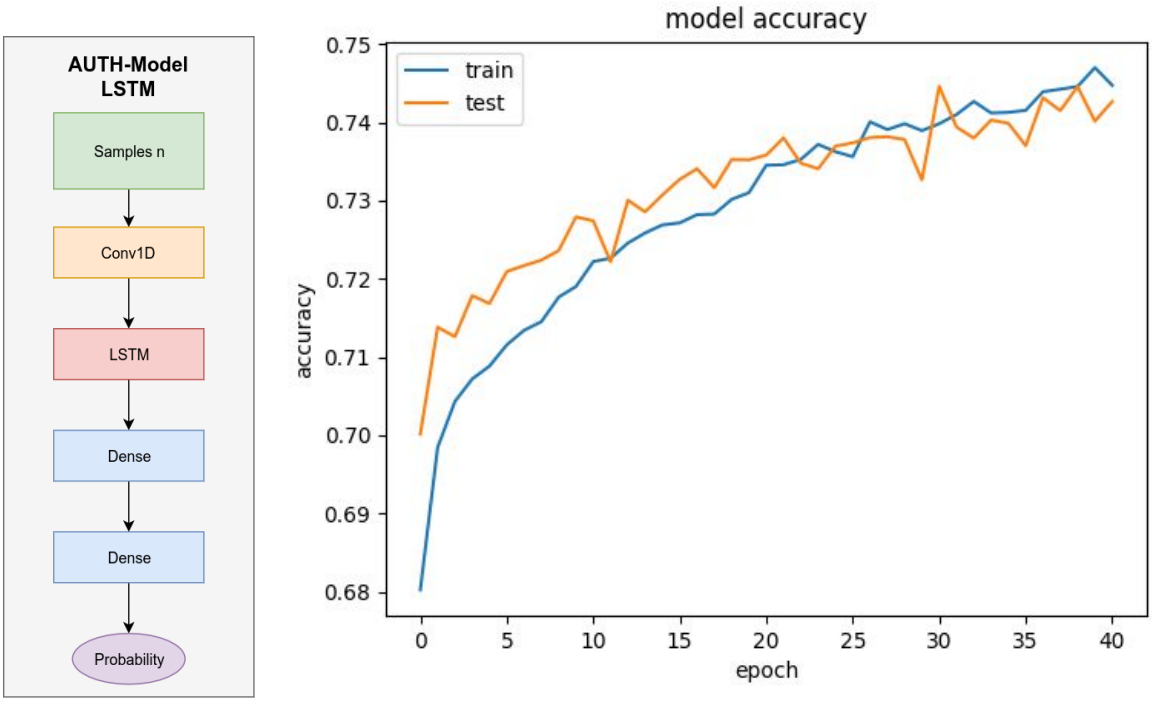
6.1.2 Επιλογή μοντέλων

Αφού τα μοντέλα εκπαιδεύτηκαν, ενσωματώθηκαν στην εφαρμογή συγχρονισμού υποτίτλων. Πραγματοποιήθηκε έλεγχος ως προς το σύνολο εκπαίδευσης και ελέγχου και αξιολογήθηκαν οι συγχρονισμένοι υπότιτλοι που προέκυψαν. Η επιλογή των μοντέλων έγινε βάση των αξιολογήσεων για το σύνολο ελέγχου.

Αποφασίστηκε να επιλεχθούν τα δύο μοντέλα με την υψηλότερη μέση ακρίβεια. Παρουσιάζεται η αλληλουχία επιπέδων του κάθε μοντέλου μαζί με την εξέλιξη της μετρικής ακρίβειας κατά την εκπαίδευση:



Σχήμα 6-1: α) Αλληλουχία επιπέδων νευρωνικού δικτύου του πρώτου μοντέλου για εντοπισμό φωνητικής δραστηριότητας. β) Η καμπύλη ακρίβειας του πρώτου μοντέλου κατά την εκπαίδευση.



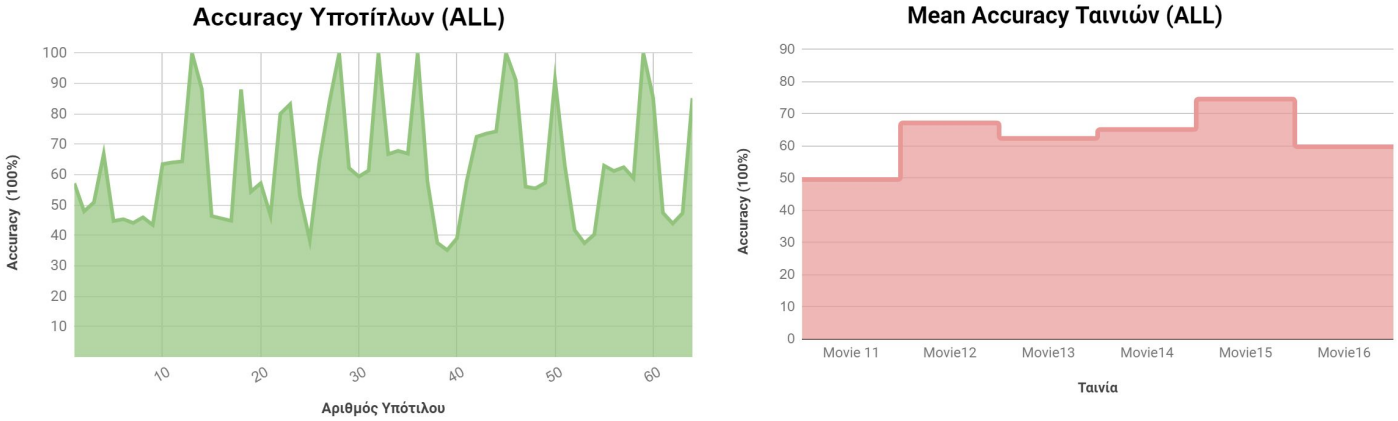
Σχήμα 6-2: α) Αλληλουχία επιπέδων νευρωνικού δικτύου του δεύτερου μοντέλου για εντοπισμό φωνητικής δραστηριότητας. β) Η καμπύλη ακρίβειας του δεύτερου μοντέλου κατά την εκπαίδευση.

Για την εκπαίδευση των νευρωνικών δικτύων χρησιμοποιήθηκαν οι 10 πρώτες ταινίες του συνόλου δεδομένων, ενώ για την αξιολόγηση οι υπόλοιπες 6 όπως αναφέρθηκε και στην ενότητα 4.2. Η διαδικασία της εκπαίδευσης για καθένα από τα δύο νευρωνικά δίκτυα (σε επεξεργαστή) διήρκησε δύο ώρες, συμπεριλαμβανομένων της επεξεργασίας του ήχου και εξαγωγής χαρακτηριστικών.

Στη συνέχεια παρουσιάζεται αναλυτικά η αξιολόγηση των μοντέλων ως συνολική εφαρμογή συγχρονισμού υποτίτλων στην οποία ενσωματώθηκαν.

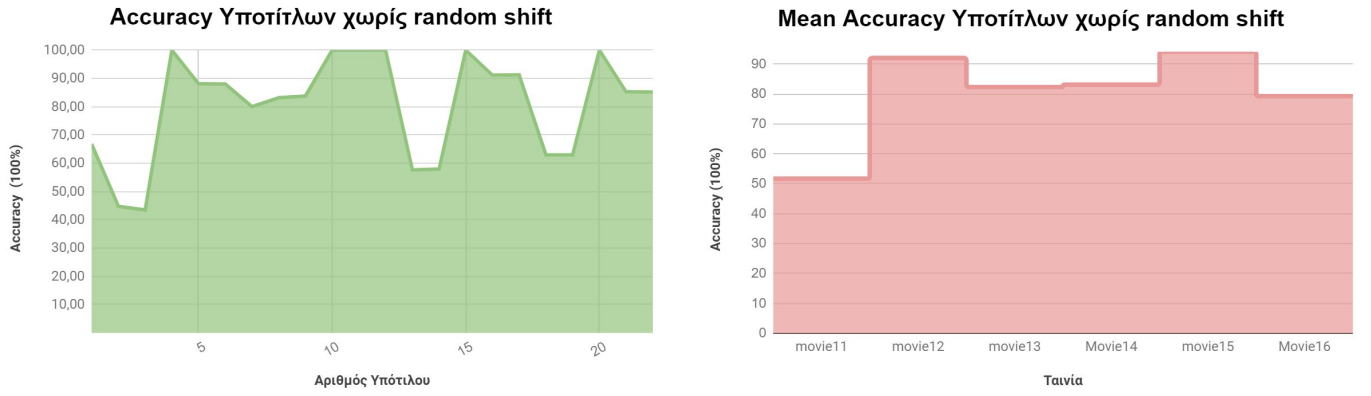
6.2 Αποτελέσματα SubSync - Auth

6.2.1 Neural Network - Gated Recurrent Unit - GRU



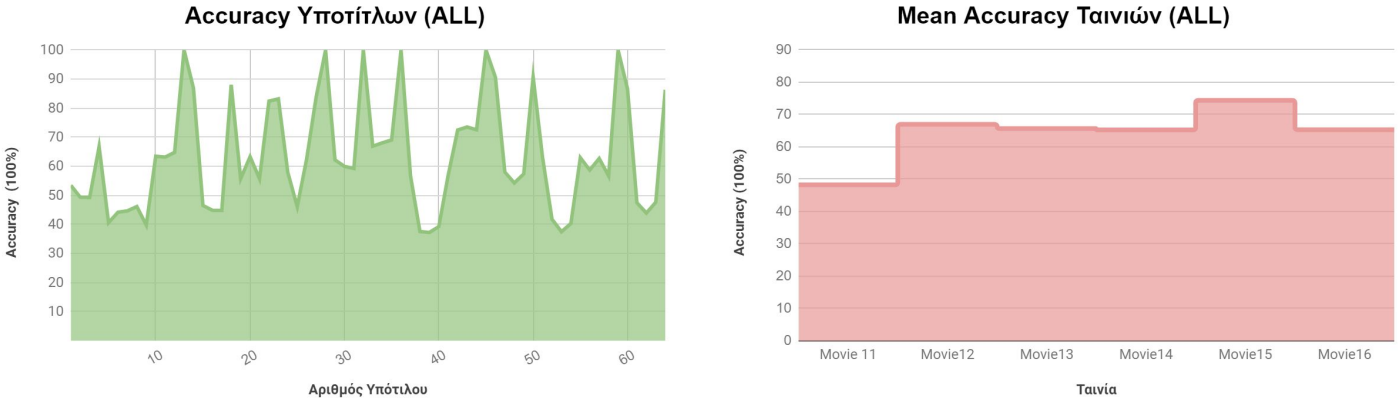
Σχήμα 6-3: Η ακρίβεια και η μέση ακριβεία συγχρονισμού όλων των υποτίτλων για το πρώτο μοντέλο.

Σχήμα 6-4: Η ακρίβεια και η μέση ακριβεία συγχρονισμού υποτίτλων με τυχαία μετατόπιση για το πρώτο μοντέλο.

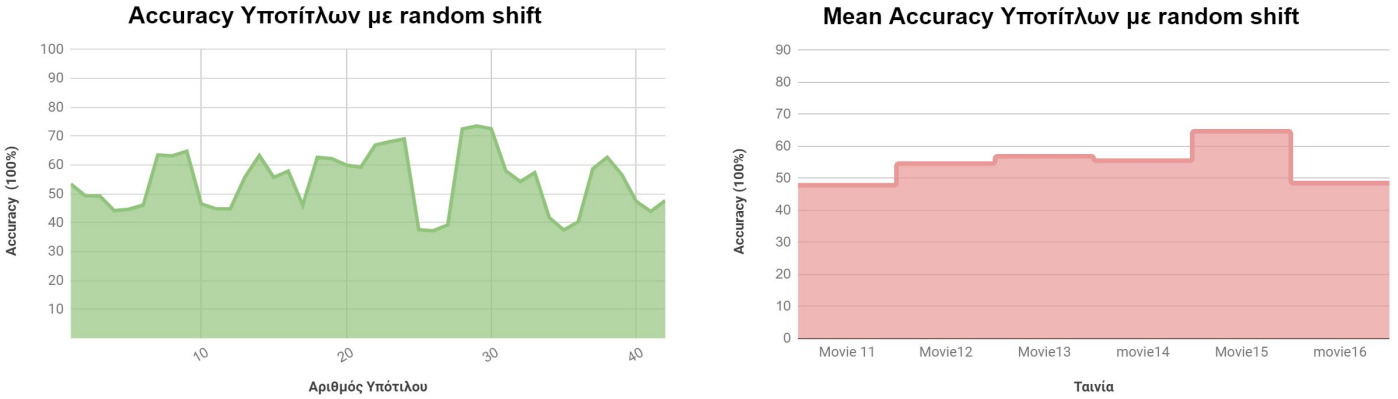


Σχήμα 6-5: Η ακρίβεια και η μέση ακριβεία συγχρονισμού υποτίτλων με σταθερή μετατόπιση για το πρώτο μοντέλο.

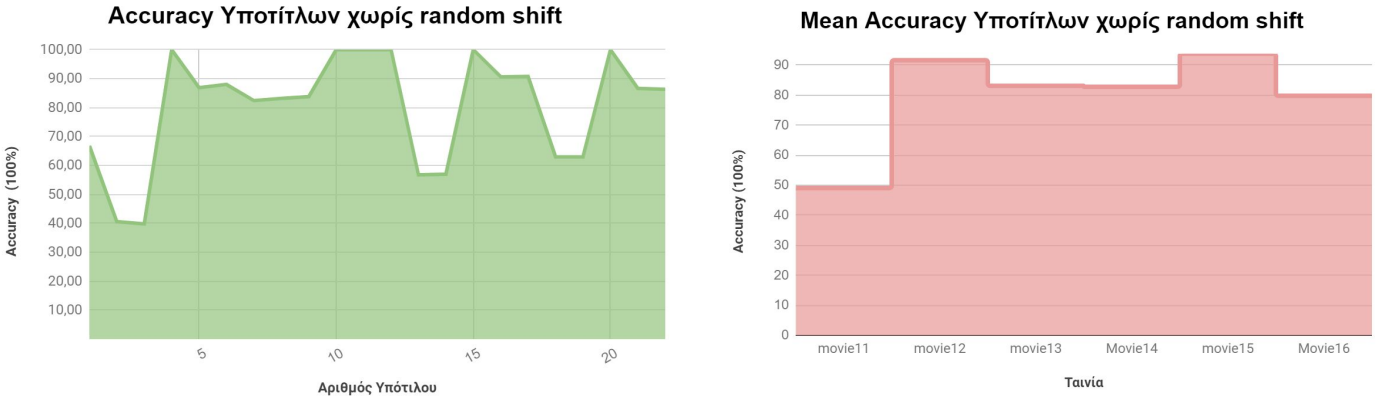
6.2.2 Neural Network - Bidirectional Long Short-Term Memory - B. LSTM



Σχήμα 6-6: Η ακρίβεια και η μέση ακριβεία συγχρονισμού όλων των υποτίτλων για το δεύτερο μοντέλο.



Σχήμα 6-7: Η ακρίβεια και η μέση ακριβεία συγχρονισμού υποτίτλων με τυχαία μετατόπιση για το δεύτερο μοντέλο.



Σχήμα 6-8: Η ακρίβεια και η μέση ακριβεία συγχρονισμού υποτίτλων με σταθερή μετατόπιση για το δεύτερο μοντέλο.

Στον πίνακα 6.1 γίνεται μία σύγκριση της ακρίβειας συγχρονισμού των δύο νέων-τροποποιημένων μοντέλων με το αρχικό ως προς το σύνολο εκπαίδευσης και στον πίνακα 6.2 ως προς το σύνολο ελέγχου.

Πίνακας 6.1

Movie / Implementation	Tympanix	B. GRU - AUTH	LSTM - AUTH
movie1	66.69	57.71	59.75
movie2	70.65	69.87	69.93
movie3	82.33	84.20	82.99
movie4	74.74	72.87	73.89
movie5	83.66	82.63	82.95
movie6	72.89	68.50	68.47
movie7	68.41	69.35	69.03
movie8	62.84	67.56	65.38
movie9	76.98	73.89	74.14
movie10	62.72	61.78	61.77
Total means	72.19	70.84	70.83
# Movies > 60 Accuracy	15/16	13/16	14/16

Πίνακας 6.2

Movie / Implementation	Tympanix	B. GRU - AUTH	LSTM - AUTH
movie11	76.97	49.63	48.26
movie12	66.94	67.21	66.95
movie13	71.53	62.38	65.63
movie14	64.97	65.14	65.24
movie15	72.30	74.62	74.40
movie16	58.13	59.81	65.30
Total means	68.48	63.13	64.30
# Movies > 60 Accuracy	5/6	4/6	5/6

6.3 Σχολιασμός Αποτελεσμάτων SubSync - Auth

Στον πίνακα 6.3 φαίνονται συνοπτικά κάποια αποτελέσματα για το B. GRU μοντέλο και στον 6.4 τα αντίστοιχα αποτελέσματα για το LSTM μοντέλο.

Πίνακας 6.3.

Μέσος Χρόνος Συγχρονισμού για έναν υπότιτλο.	**17,07**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με μεταβλητή μετατόπιση. (%)**	**54,13**
Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο. (%)	**63,13**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με σταθερή μετατόπιση. (%)**	**80,43**
Ταινίες με Μέση Ακρίβεια > 60% (Καλός Συγχρονισμός)	**13/16**		

Πίνακας 6.4

Μέσος Χρόνος Συγχρονισμού για έναν υπότιτλο.	**16,35**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με μεταβλητή μετατόπιση. (%)**	**54,69**
Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο. (%)	**64,30**	**Μέση Ακρίβεια Συγχρονισμού για έναν υπότιτλο με σταθερή μετατόπιση. (%)**	**80,24**
Ταινίες με Μέση Ακρίβεια > 60% (Καλός Συγχρονισμός)	**14/16**		

Παρατηρώντας την ακρίβεια της εφαρμογής συγχρονισμού υποτίτλων, διακρίνουμε τα ίδια προβλήματα με την αρχική υλοποίηση. Η εφαρμογή υστερεί σε υπότιτλους που έχουν υποστεί μεταβαλλόμενο συγχρονισμό.

Το LSTM μοντέλο πετυχαίνει ακρίβεια μεγαλύτερη από 60% σε 5 από τις 6 ταινίες, όπως και του Tympanix, όμως διαφέρει σε ποια ταινία αποτυγχάνουν. Ομοίως, το μοντέλο B.GRU πετυχαίνει ακρίβεια μεγαλύτερη από 60% σε 4 από τις 6 ταινίες, με μια ταινία να είναι μόλις 0.2% κάτω από το όριο του 60%. Στις ταινίες 12,14,15,16 τα μοντέλα πετυχαίνουν οριακά καλύτερη ακρίβεια, με τη μεγαλύτερη να διαφορά να βρίσκεται στην ταινία 16 όπου το LSTM μοντέλο πετυχαίνει 65.3% ενώ του Tympanix 58.13% ακρίβεια.

Σημαντικό «πλήγμα» και για τα δύο νέα μοντέλα είναι το αποτέλεσμα της αξιολόγησης της ταινίας 11, στην οποία πετυχαίνουν ακρίβεια 49.63% και 48.26% το καθένα αντίστοιχα, και μειώνει αρκετά τη μέση ακρίβεια των δύο μοντέλων. Η ταινία 11 ανήκει στο είδος Δράσης - Θρίλερ. Επομένως, είναι πολύ πιθανό το μοντέλο να αποτυγχάνει να αναγνωρίσει-διαχωρίσει τα ηχητικά εφέ της ταινίας.

Σε γενικές γραμμές, και τα δύο μοντέλα βρίσκονται αρκετά κοντά στη μέση ακρίβεια των καλύτερων υλοποιήσεων χωρίς όμως να τα ξεπερνούν.

Αξιοσημείωτο είναι, επίσης, πως τα δύο καλύτερα μοντέλα έχουν αρκετά διαφορετική δομή επιπέδων - το πρώτο έχει δύο ενώ το δεύτερο ένα συνελικτικό επίπεδο, το πρώτο έχει αμφίδρομο RNN ενώ το δεύτερο όχι αμφίδρομο και τέλος, το πρώτο έχει GRU ενώ το δεύτερο LSTM.

7. Σύνοψη Αποτελεσμάτων

7.1 Αποτελέσματα - Σύγκριση Υλοποιήσεων

Στον πίνακα 7.1 παρουσιάζονται συνοπτικά τα αποτελέσματα (ακρίβεια συγχρονισμού) όλων των υλοποιήσεων για τις ταινίες 1 έως 16 του συνόλου των δεδομένων. Είναι φανερό ότι η υλοποίηση του δημιουργού Tympanix (καθώς και τα δύο μοντέλα που εκπαιδεύτηκαν BGRU-LSTM) και του Sc0ty παρουσίασαν τα καλύτερα αποτελέσματα. Το μοντέλο με την υψηλότερη αξιολόγηση είναι με ισχνή διαφορά αυτό της αρχικής υλοποίησης του Tympanix.

Πίνακας 7.1

Movie / Implementation	Tympanix	Smacke	Sc0ty	Kaegi	Koenkk	B. GRU - AUTH	LSTM - AUTH
movie1	66.69	61.08	64.65	59.30	74.27	57.71	59.75
movie2	70.65	52.79	70.74	71.27	-	69.87	69.93
movie3	82.33	84.40	78.47	84.37	32.42	84.20	82.99
movie4	74.74	74.45	67.41	46.71	81.42	72.87	73.89
movie5	83.66	83.74	78.87	71.35	87.24	82.63	82.95
movie6	72.89	72.12	71.70	46.31	77.24	68.50	68.47
movie7	68.41	67.86	66.91	68.71	12.00	69.35	69.03
movie8	62.84	70.50	66.98	65.80	84.36	67.56	65.38
movie9	76.98	38.02	72.46	48.48	84.34	73.89	74.14
movie10	62.72	64.99	61.59	43.45	26.23	61.78	61.77
movie11	76.97	54.02	46.62	31.68	53.70	49.63	48.26
movie12	66.94	65.56	76.18	64.92	41.18	67.21	66.95
movie13	71.53	54.73	73.52	59.07	-	62.38	65.63
movie14	64.97	36.46	76.00	65.42	75.72	65.14	65.24
movie15	72.30	73.88	81.41	74.34	-	74.62	74.40
movie16	58.13	60.88	58.83	62.93	-	59.81	65.30
Total means	70.80	63.47	69.52	60.26	60.84	67.95	68.38
# Movies > 60 Accuracy	15/16	11/16	14/16	9/16	7/12	13/16	14/16

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα που αφορούν μόνο το σύνολο ελέγχου, δηλαδή αφορούν τις τελευταίες 6 ταινίες.

Πίνακας 7.2

Movie / Implementation	Tympanix	Smacke	Sc0ty	Kaegi	Koenkk	B. GRU - AUTH	LSTM - AUTH
movie11	76.97	54.02	46.62	31.68	53.70	49.63	48.26
movie12	66.94	65.56	76.18	64.92	41.18	67.21	66.95
movie13	71.53	54.73	73.52	59.07	-	62.38	65.63
movie14	64.97	36.46	76.00	65.42	75.72	65.14	65.24
movie15	72.30	73.88	81.41	74.34	-	74.62	74.40
movie16	58.13	60.88	58.83	62.93	-	59.81	65.30
Total means	68.48	57.59	68.76	59.73	56.86	63.13	64.30
# Movies > 60 Accuracy	5/6	3/6	4/6	4/6	1/3	4/6	5/6

7.2 Συμπεράσματα

Συνολικά, θα μπορούσε να εξαχθεί το συμπέρασμα ότι σε περιπτώσεις σταθερά απο-συγχρονισμένων υποτίτλων προτιμότερη είναι η χρήση της υλοποίησης του Tympanix, τόσο για την υψηλή της ακρίβεια όσο και για την ταχύτητα της διαδικασίας. Αντίθετα στην περίπτωση των μεταβαλλόμενα απο-συγχρονισμένων υποτίτλων ως βέλτιστη υλοποίηση θα μπορούσε να θεωρηθεί αυτή του Sc0ty αν και η διαδικασία συγχρονισμού μπορεί να διαρκεί περισσότερο.

Εν κατακλείδι, οι λύσεις για το πρόβλημα του συγχρονισμού υποτίτλων που παρουσιάστηκαν, αν και παρουσιάζουν υψηλή ακρίβεια, δεν μπορούν να εγγυηθούν το συγχρονισμό σε όλες τις περιπτώσεις. Επιπλέον, φαίνεται ότι η κάθε μέθοδος συγχρονισμού έχει πλεονεκτήματα και μειονεκτήματα επομένως μπορεί να υπερτερεί σε κάποιες περιπτώσεις έναντι των υπολοίπων αλλά και το αντίθετο.

8. Βιβλιογραφικές Αναφορές

[1] Netflix, “Timed Text Style Guide: General Requirements”, Jul 20, 2019

[2] BBC, “Subtitle Guidelines” , Apr 2019

[3] Olofsson, O.” Detecting Unsynchronized Audio and Subtitles using Machine Learning”, Master thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2019.

[4] Kaegi, “Language Agnostic Subtitle Synchronization”, Bachelor Thesis, 2019

[5] Matroska Media Container, “SRT Subtitles”

[6] A. Sabater, “Automatic Subtitle Synchronization through Machine Learning”, Sep 14, 2017 [Online]. Available: Medium, https://machinelearnings.co/automatic-subtitle-synchronization-e188a9275617

[7] A. Koretzky, “ Audio AI: isolating vocals from stereo music using Convolutional Neural Networks” 4 Feb 2017 [Online], Available: Medium, https://towardsdatascience.com/audio-ai-isolating-vocals-from-stereo-music-using-convolutional-neural-networks-210532383785

[8] Wikipedia, last edited Nov 24, 2019, “Subtitles”

[9] Ramírez, J.; J. M. Górriz; J. C. Segura (2007). "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness" (PDF). In M. Grimm and K. Kroschel (ed.). Robust Speech Recognition and Understanding. pp. 1–22. ISBN:978-3-902613-08-0.

[10] Pacific Northwest Seismic Network, “What is a Spectrogram”

[11] K. Prahallad, “Speech Technology: A Practical Introduction Topic: Spectrogram, Cepstrum and Analysis “, Carnegie Mellon University & International Institute of Information Technology Hyderabad, 2019

[12] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/, Mel Frequency Cepstral Coefficient (MFCC) Tutorial

[13] Fawaz S. Al-Anzi, Dia AbuZeina The Capacity of Mel Frequency Cepstral Coefficients for Speech Recognition, International Journal of Computer and Information

Engineering, Vol:11, No:10, 2017
https://pdfs.semanticscholar.org/d3a9/d1a6addf05f48293a968c434f45837b68a75.pdf

[14] *Bachus, John (1977).* The Acoustical Foundations of Music *(Second Edition), W. W. Norton & Company, Inc., New York. ISBN 0-393-09096-5.*
https://el.wikipedia.org/wiki/%CE%89%CF%87%CE%BF%CF%82

[15] Boyd, Stephen P.; Vandenberghe, Lieven (2004). *Convex Optimization* (pdf). Cambridge University Press. p. 129. ISBN 978-0-521-83378-3
https://en.wikipedia.org/wiki/Optimization_problem

[16] Waveform Audio File Format - WAV, http://midi.teragonaudio.com/tech/wave.htm

[17] Sc0ty, SubSync, 2019, https://github.com/sc0ty/subsync

[18] Sc0ty, SubSync – architecture overview, 2019,
http://sc0ty.pl/2019/04/subsync-architecture-overview/

[19] http://sc0ty.pl/2019/04/subsync-synchronize-movie-subtitles-with-audio-track/, Sc0ty, SubSync – synchronize movie subtitles with audio track, 2019

[20] https://www.linuxuprising.com/2019/01/subsync-auto-subtitle-synchronization.html, SubSync: Auto Subtitle Synchronization Tool Based On Audio Track, 2019

[21] https://ffmpeg.org/, FFmpeg

[22] https://www.afterdawn.com/glossary/term.cfm/demux, Demux
https://en.wkipedia.org/wiki/Demultiplexer_(media_file)

[23] CMUSphinx – Github, https://cmusphinx.github.io/

[24] Μπούτσικας Μιχαήλ. «Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression)» (PDF). Σημειώσεις μαθήματος "Στατιστικά Προγράμματα". Πανεπιστήμιο Πειραιώς. Ανακτήθηκε στις 30 Απριλίου 2013.

[25] Τσαντεκίδης, Α. (2016). Αρχιτεκτονικές και Εκπαίδευση Βαθιών Νευρωνικών Δικτύων (No. GRI-2016-17594). Aristotle University of Thessaloniki.

[26] Γεώργιος, Β. (2017). Τεχνικές Βαθιάς Μηχανικής Μάθησης Για Την Αυτόματη Δημιουργία Περιγραφών Εικόνων.