ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Τμήμα Πληροφορικής

Εργασία Μαθήματος ***ΕΠΕΞΕΡΓΑΣΙΑ ΣΗΜΑΤΩΝ ΦΩΝΗΣ ΚΑΙ ΗΧΟΥ***

Αριθμός εργασίας – Τίτλος εργασίας	***Εργασία***
Όνομα φοιτητή	Καλλίγερος Αναστάσιος
Αρ. Μητρώου	Π19253
Ημερομηνία παράδοσης	30/05/2023

Εκφώνηση εργασίας

Εργασία Εαρινού Εξαμήνου 2022-2023:

- *Ημερομηνία παράδοσης: Ημερομηνία εξέτασης του μαθήματος, ώρα 23:59μμ. Ομάδες **14** φοιτητών. Η παράδοση της εργασίας γίνεται μέσω της πλατφόρμας e-class χωρίς παραπομπές σε εξωτερικούς συνδέσμους. Παραδίδονται:*

α) *η τεκμηρίωση της εργασίας σε ένα αρχείο pdf, στην πρώτη σελίδα της οποίας αναγράφονται τα ονοματεπώνυμα των φοιτητών και οι ΑΜ. Δεν θα βαθμολογηθούν εργασίες που δεν περιέχουν τεκμηρίωση ή που δεν αναφέρουν τα ονόματα των μελών της*

*ομάδας στην τεκμηρίωση**. β)** τα αρχεία source code σε ένα συμπιεσμένο αρχείο με όνομα source2023.zip (ή .rar ή άλλη σχετική κατάληξη).*

***γ)** οποιαδήποτε άλλα συνοδευτικά αρχεία η ομάδα κρίνει απαραίτητα σε ένα συμπιεσμένο αρχείο με το όνομα auxiliary2023.zip (ή .rar ή άλλη σχετική κατάληξη).*

- Η εργασία ισχύει και τον Σεπτέμβριο.

- Αποδεκτές γλώσσες είναι οι Octave/Matlab και Python.

- Η αντιγραφή ή η χρήση generative bots οδηγεί σε μηδενισμό στον τελικό βαθμό.

***Προσοχή!!!**: **Δεν** μπορείτε να χρησιμοποιήσετε **συνελικτικά** νευρωνικά δίκτυα **ή Recurrent Neural Networks** οποιουδήποτε τύπου. **Δεν** είναι αποδεκτή η χρήση έτοιμων web services ή APIs για speech recognition. Δεν μπορείτε να χρησιμοποιήσετε **transfer learning** από ήδη εκπαιδευμένα δίκτυα. Οι αντίστοιχες λύσεις μηδενίζονται.*

***Θέμα (4 βαθμοί):** Καλείστε να υλοποιήσετε ένα ASR σύστημα, που δέχεται είσοδο μία ηχογράφηση κάθε φορά, η οποία συνιστά πρόταση αποτελούμενη από 5-10 ψηφία της Αγγλικής γλώσσας που έχουν ειπωθεί με αρκούντως μεγάλα διαστήματα παύσης.*

***1)** Το σύστημα προχωρά στην κατάτμηση της πρότασης σε λέξεις χρησιμοποιώντας υποχρεωτικά έναν ταξινομητή background vs foreground της επιλογής σας. Από τις λέξεις που προκύπτουν, υπολογίστε τη θεμελιώδη συχνότητα του ομιλητή.*

***2)** Στη συνέχεια, το σύστημα αναγνωρίζει κάθε λέξη χρησιμοποιώντας ως φασματική αναπαράσταση μόνο το απλό φασματογράφημα (επιλέξτε φασματική περιοχή αν θέλετε). Αν χρειαστείτε δεδομένα εκπαίδευσης, χρησιμοποιήστε **μόνο** δημόσια σύνολο(α) δεδομένων από το Internet. Δεν χρειάζεται να τα συμπεριλάβετε στα παραδοτέα αλλά μόνο να περιγράψετε πώς ελήφθησαν.*

***3)** Στην έξοδο παράγεται κείμενο με τα ψηφία που αναγνωρίστηκαν.*

- *Δώστε έμφαση στην επεξεργασία του σήματος, προτού αρχίσουν τα στάδια κατάτμησης/αναγνώρισης (π.χ., με κατάλληλα φίλτρα, αλλαγή ρυθμού δειγματοληψίας, κ.λ.π).*
- *Είναι σημαντικό να **περιγράψετε** το σύστημα αλγοριθμικά (εξαγωγή χαρακτηριστικών, αλγόριθμος αναγνώρισης) και να εξηγήσετε τις επιδόσεις του χρησιμοποιώντας τις κατάλληλες μετρικές.*
- *Πρέπει να εξηγήσετε ποια δεδομένα χρησιμοποιήσατε κατά τον έλεγχο και την εκπαίδευση του συστήματος. Αν είναι δικά σας, πώς τα δημιουργήσατε.*

- *Προσπαθήστε να μην εξαρτάται το σύστημα από τα χαρακτηριστικά της φωνής του ομιλητή, αλλά να είναι όσο το δυνατόν ανεξάρτητο ομιλητή.*

1 Επίδειξη της λύσης

Ερώτημα 1 και βήματα προεπεξεργασίας

Οι δύο ηχογραφήσεις που χρησιμοποιήθηκαν ως είσοδος στο πρόγραμμα βρίσκονται στον φάκελο auxiliary και το όνομα τους αντιστοιχεί στα ψηφία που τα αποτελούν. Όλα τα δείγματα είναι μονοκαναλικές ηχογραφήσεις και έχουν περάσει από ανωπερατό φίλτρο, ώστε να αποκλειστούν όλες οι συχνότητες κάτω από 40Hz και στην συνέχεια έγινε υποδειγματοληψία με τον νέο ρυθμό δειγματοληψίας να είναι τα 6000 δείγματα ανά δευτερόλεπτο.

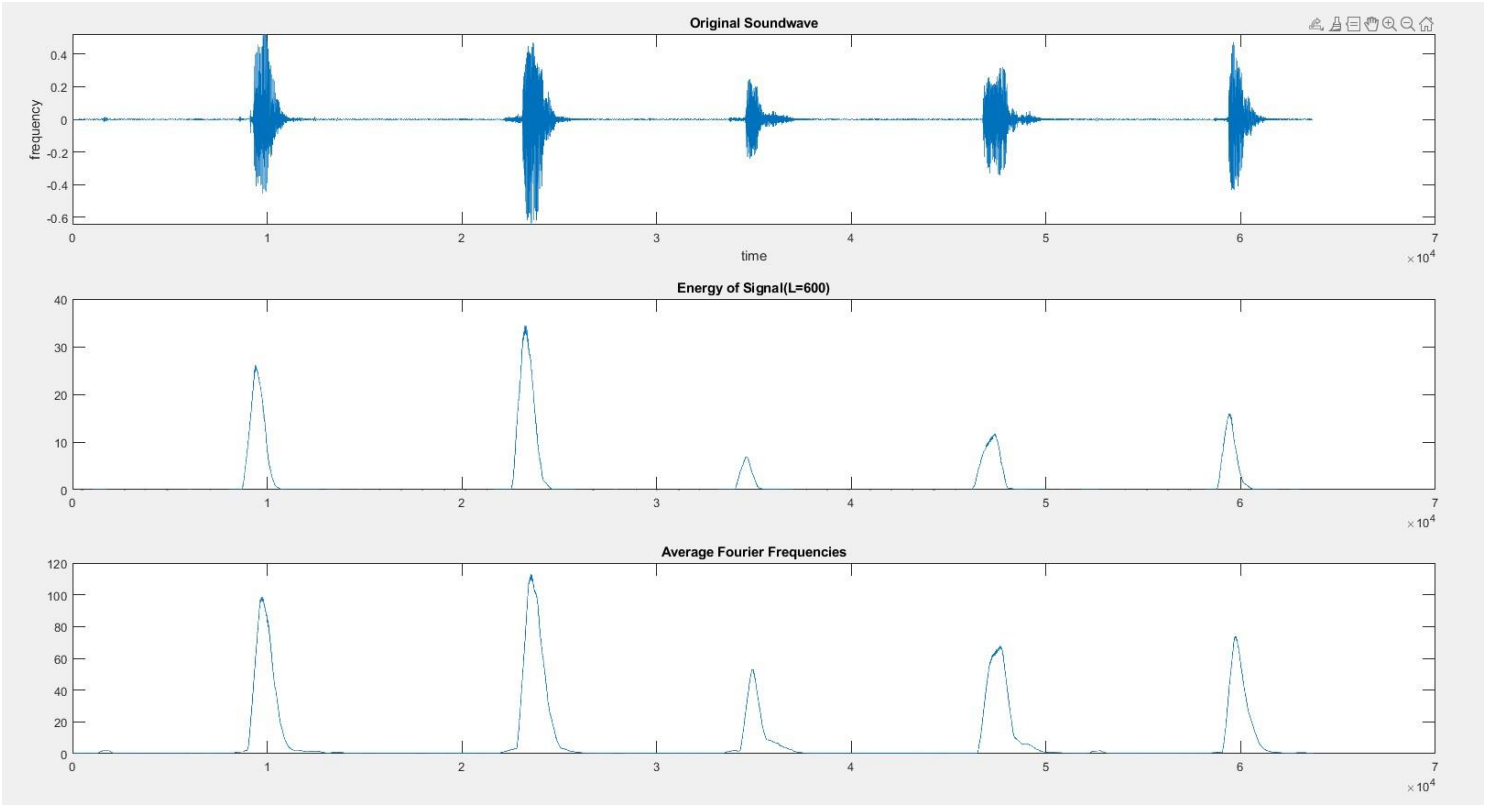
Ο ταξινομητής background vs foreground που επιλέξαμε λειτουργεί με την τεχνική της κατωφλίωσης των δειγμάτων. Συγκεκριμένα με την απόλυτη τιμή

των συντελεστών του μετασχηματισμού fourier και την ενέργεια του σήματος. Τα διαστήματα έχουν μήκος 600 δείγματα κάθε φορά ή αλλιώς περίπου το 1% της συνολικής ηχογράφησης. Η ενέργεια υπολογίστηκε με βάση τον παρακάτω τύπο.

$$e(n) = \sum_{n}^{n+L-1} |x(n)^2|$$

Όπου n το πρώτο δείγμα του παραθύρου και L το μήκος του.

Η εκτέλεση του ερωτήματος αυτού, όπως και η τελική κατηγοριοποίηση γίνονται στο main_program.m και παρακάτω παρατίθεται ένα γράφημα που παρουσιάζει την διαδικασία που περιγράψαμε ως εδώ.



Όπως μπορούμε να παρατηρήσουμε στην παραπάνω γραφική παράσταση, η ενέργεια του σήματος και οι συντελεστές Fourier συμπίπτουν σχεδόν ακριβώς με τα σημεία της ηχογράφησης που ακούγεται κάποιο ψηφίο, επομένως με τα εμπειρικά όρια που θέσαμε στην κατωφλίωση μπορούμε να διαχωρίσουμε την ηχογράφηση σε επιμέρους μέρη που αποτελούν ένα ψηφίο το κάθε ένα.

Ερώτημα 2

Σε αυτό το ερώτημα εκπαιδεύσαμε τον ταξινομητή που αναθέτει το την ηχογράφηση στο εκάστοτε ψηφίο.Το μοντέλο που εκπαιδεύσαμε είναι 10 μηχανές διανυσματικής στήριξης (SVM,support vector machines),όπου η έξοδος είναι το 0 η το 1.Έχουμε ακολουθήσει μία 1 vs all προσέγγιση, δηλαδή το κάθε ένα svm αντιστοιχεί σέ ένα ψηφίο και όσες είσοδοι αντιστοιχούν σε αυτό το ψηφίο έχουν έξοδο 1 και όλες οι άλλες έξοδο 0.

Η είσοδος είναι οι απόλυτες τιμές του φασματογραφήματος,με την χρήση της συνάρτησης imresize,η οποία κανονικά χρησιμοποιείται για εικόνες, αλλά στο παράδειγμα μας χρησιμοποιήθηκε για να φέρει όλα τα φασματογραφήματα στο ίδιο μέγεθος,καθώς το φασματογράφημα είναι η οπτικοποίηση των συντελεστών Fourier στην μονάδα του χρόνου,επομένως κανένα δείγμα δεν έχει ίδιο μήκος φασματογραφήματος με κάποιο άλλο.Η εκπαίδευση έγινε με το σύνολο δεδομένων του kaggle Free Spoken Digits Dataset.Μετά την εκπαίδευση τα svm αποθηκεύονται στο αρχείο classifiers.matΛόγω του πώς ταξινομεί το svm έχουμε δύο παραμέτρους όταν εξετάζουμε εάν ένα δείγμα ανήκει στην κλάση αυτή το score,όπου όσο μικρότερο τόσο καλύτερα και το

εάν ανήκει που είναι είτε 0 είτε 1.Η ταξινόμηση γίνεται στο μικρότερο score (μικρότερη απόσταση) που το δείγμα ανήκει στην κλάση αυτή.Η έξοδος είναι τα ψηφία που προβλέπει το σύστημα.

```
Predicted Output:
4    4    4    4    5
```

two_four_six_eight_five.wav

```
Predicted Output:
6    9    4    9    4
```

eight_seven_four_nine_one.wav

ΥΠΟΣΗΜΕΙΩΣΗ

Λόγω μεγάλου μεγέθους λείπει από τον φάκελο auxiliary.zip το αρχείο classifiers.mat, το οποίο είναι μοντέλο φτιαγμένο σε svm και λόγω του μικρού μεγέθους του συνόλου δεδομένων μπορεί να εκπαιδευτεί πιο αποδοτικά από άλλα μοντέλα. Παρόλα αυτά μέσω των ενδεικτικών screenshots που παρέχουμε παραπάνω βλέπετε τα αποτελέσματα.