ΕΠΕΞΕΡΓΑΣΙΑ ΣΗΜΑΤΩΝ ΦΩΝΗΣ ΚΑΙ ΗΧΟΥ

ΑΛΕΞΙΟΣ ΛΥΜΠΕΡΗΣ Π15074

ΠΕΤΡΟ ΠΡΙΦΤΗ Π15121

ΕΚΦΩΝΗΣΗ ΕΡΓΑΣΙΑΣ

Καλείστε να υλοποιήσετε ένα ASR σύστημα, που δέχεται είσοδο μία ηχογράφηση κάθε φορά, η οποία συνιστά πρόταση αποτελούμενη από 4-10 ψηφία που έχουν ειπωθεί με αρκούντως μεγάλα διαστήματα παύσης.

Το σύστημα προχωρά στην κατάτμηση της πρότασης και στη συνέχεια αναγνωρίζει κάθε λέξη. Στην έξοδο παράγεται κείμενο με τα ψηφία που αναγνωρίστηκαν.

Προσπαθήστε να μην εξαρτάται το σύστημα από τα χαρακτηριστικά της φωνής του ομιλητή, αλλά να είναι όσο το δυνατόν ανεξάρτητο ομιλητή (2 βαθμοί)

Δώστε έμφαση στην επεξεργασία του σήματος, προτού αρχίσει το στάδιο της 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.

Η εργασία παραδίδεται με email. Χρησιμοποιήστε dropbox links ή σχετικές δυνατότητες, μόνο όταν ο όγκος των αρχείων είναι απαγορευτικός για την αποστολή με email.

Η εργασία εκπονείται σε ομάδες 1-3 φοιτητών. Για την εκπόνηση της εργασίας αποδεκτές γλώσσες είναι οι Octave/Matlab, Python και C/C++.

Τελική ημερομηνία παράδοσης είναι η ημερομηνία εξέτασης του μαθήματος στο εαρινό εξάμηνο, ώρα 23:59. Η εργασία είναι απαλλακτική και δεν παραδίδεται τον Σεπτέμβριο ή σε τυχόν εμβόλιμες εξεταστικές.

Αν δεν παραδοθεί εργασία ή αν η εργασία δεν λάβει βαθμό >=5, τότε ο φοιτητής/φοιτήτρια δίνει γραπτές εξετάσεις τον Σεπτέμβριο με άριστα το 10.

Θα γίνει αυστηρός έλεγχος ως προς την αντιγραφή.

ΦΑΚΕΛΟΣ ΕΡΓΑΣΙΑΣ

Ο φάκελος του project είναι ο ASR. Μέσα περιέχει το εκτελέσιμο αρχείο asrsytem.py καθώς και το φάκελο audio με όλα τα .wav αρχεία του project π.χ. λεξιλόγιο μοντέλου και προτάσεις εισόδου[test0.wav] .Τέλος ο κώδικας της εργασίας αναπτύχθηκε σε γλώσσα python.

ΒΙΒΛΙΟΘΗΚΕΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ

Χρησιμοποιήθηκαν οι παρακάτω βιβλιοθήκες:

-Librosa: είναι ενα python package για ανάλυση μουσικής και ήχου. Με τη βοήθεια αυτής της βιβλιοθήκης φορτώνονται οι ήχοι στο πρόγραμμα και μέσω αυτής έγιναν extract τα κατάλληλα Features που χρειαζόμασταν.

-Sklearn: περιέχει εργαλεία για data mining και data analysis (Στη δική μας περίπτωση svm SVC).

-**Numpy**

-**Scipy**

-Matplotlib

ΕΠΕΞΗΓΗΣΗ ΠΡΟΓΡΑΜΜΑΤΟΣ

ΣΥΝΟΠΤΙΚΗ ΠΕΡΙΓΡΑΦΗ

Έχουμε υλοποιήσει ένα ASR σύστημα. Έχουμε φτιάξει συγκεκριμένο λεξιλόγιο με αρχεία(λέξεις) που βρίσκονται στο φάκελο audio του project.Ως είσοδο κάθε φορά δέχεται μια ηχογράφηση . Ο αλγόριθμός μας «σπάει» την ηχογράφηση(πρόταση) και τη χωρίζει σε λέξεις. Μετά αναγνωρίζει κάθε μία ξεχωριστά και στο τέλος εμφανίζει όλες τις αναγνωρισμένες λέξεις.

ΔΗΜΙΟΥΡΓΙΑ ΛΕΞΙΛΟΓΙΟΥ

Το λεξιλόγιο βρίσκεται μέσα στον φάκελο audio. Κάθε ηχητικό αρχείο που έχουμε βάλει στο λεξιλόγιο με βάση το path του πηγαίνει στη συνάρτηση wav2mfcc .Εκεί το αρχείο φορτώνεται από τη librosa και αποθηκελυεται ένας πίνακας x_one(numpy array) και το sampling rate sr_one.

Σε αυτο εδώ το σημείο γίνονται extract τα **MFCC features** από το ηχητικό. Είναι σημαντικό να αναφερθεί ότι χρησιμοποιούνται τα συγκεκριμένα features ,διότι έχει ζητηθεί το σύστημα να είναι ανεξάρτητο από τα χαρακτηριστικά φωνής του ομιλητή ,και αυτά μας δίνουν τη δυνατότητα αυτή .

Χρησιμοποιούμε έναν Scaler, ο οποίος κάνει scale τα features ώστε να έχουν μηδενική μέση τιμή και κοινή διακυμανση (μόνο ένας scaler χρησιμοποιείται σε όλο τον κώδικα) .

Αφού έχει τελιώσει η παραπάνω διαδικασία για κάθε λέξη του λεξιλογίου ,βάζουμε τα feature όλων των λέξεων σε έναν πίνακα (features) και φτιαχνουμε έναν άλλο πίνακα(labels) που περιέχει target/reference labels π.χ. το 0 αντιστοιχεί στην 1^{η} λέξη , το 1 στη δευτερη κ.ο.κ.

Μετά δημιουργούμε έναν classifier μοντέλο με τη βοήθεια της sklearn[model=sklearn.svm.SVC()]. Κάνουμε train(προπονούμε) τον classifier με του δυο πίνακες που φτιάξαμε παραπάνω(features,labels).

Επίσης να σημιωθεί ότι στη μεταβλητή n_mfcc[int] βάλαμε αριθμό που εμείς θεωρούσαμε κατάλληλο.

Με αυτόν τον τρόπο φτιάξαμε το λεξιλόγιο του συστήματός μας.

ΕΠΕΞΕΡΓΑΣΙΑ ΠΡΟΤΑΣΗΣ ΕΙΣΟΔΟΥ

Τα αρχεία που «μπαίνουν» ως προτάσεις-εισόδοι στο σύστημα βρίσκονται στο φάκελο audio και έχουν ονομασίες test0.wav, test1.wav, test2.wav.

Ο αλγόριθμος αναγνώρισής τους ξεκινάει ως εξής :

Παίρνουμε το path του ηχητικού και μεταφερόμαστε στη συνάρτηση silence_audio. Αυτή παίρνει το ηχητικό-πρόταση και το φορτώνει από τη librosa.

Threshold: Του έχουμε βάλει προκαθορισμένη τιμή, η οποία βασίζεται στο μικρόφωνο και στην ένταση στην οποία μιλάει ο χρήστης.

'Οποτε ο αλγόριθμος συναντήσει στοιχείο μέσα στον πίνακα με απόλυτη τιμή που είναι μικρότερη από το threshold ,και δεξιά του υπάρχουν τουλάχιστον άλλα 5000 στοχεία που έχουν και αυτά μικρότερη τιμή από το threshold τότε αντικαθιστά την τιμή του με 0,ενώ τα υπόλοιπα τα αφήνει όπως ήταν.Στην περίπτωση που υπάρχουν 2000 στοιχεία του πίνακα με απόλυτη τιμή μικρότερη από το threshold τότε μπαίνει πάλι η αρχική τους τιμή.Αφού τελειώσει ο έλεγχος του καθενός περνάνε όλα σε μία λίστα .

Το **count** έχει καθοριστεί απαραίτητα από εμάς .Αν αλλαξεί πιθανώς να δώσει χειρότερα αποτελέσματα.

Με την παραπάνω διαδικασία απαλείφεται η φασαρία ενδιάμεσα των λέξεων.Οπότε τώρα είναι πιο έυκολη η διαδικασία για να κόψουμε την πρόταση σε λέξεις.

Μόλις τελειώσει η παραπάνω διαδικασία ,καλείται η συνάρτηση split_audio. Χωρίζει τη λίστα που έχουν δώσει με βάση τα μηδενικά και γυρνάει ένα πίνακα με n υποπίνακες αν οι λέξεις τις πρότασης είναι n.

Τέλος καλείται η συνάρτηση print_words .Σε αυτήν ξανακαλείται η συνάρτηση wav2mfcc, αλλά αυτή τη φορά με όρισμα τον πίνακα με τους υποπίνακες των λέξεων για να πάρουμε ένα prediction από το προπονημένο μοντέλο μας.Μετά μετράμε τα predicted labels για κάθε υποπίνακα ,και αναλόγως αντιστοιχίζουμε τη κατάλληλη λέξη από το λεξιλόγιο μας.

ΑΠΟΤΕΛΕΣΜΑΤΑ

Τα αποτελέσματα βγαίνουν ¾ σωστά(και οι 2 φωνές στα αρχεία testN.wav). Στο αρχείο test3.wav όμως 2 στις 4 λέξεις βγαίνουν σωστά. Για να βελτιωθεί αυτό πρέπει να πάρουμε περισσότερα δείγματα για κάθε λέξη του λεξιλογίου.Επίσης στο μετασχημάτισμο mfcc ίσως υπάρχει καλύτερο n_mfcc για το συγκεκριμένο ηχητικό.

Τέλος η τιμή του threshold σε κάποιες περιπτώσεις θα πρέπει να είναι διαφορετική από αυτή που έχουμε καθορίσει .Αυτο θα συμβεί λόγω διαφορετικής ποιότητας μικροφώνου ηχογράφησης.