ΕΠΕΞΕΡΓΑΣΙΑ ΣΗΜΑΤΩΝ ΦΩΝΗΣ ΚΑΙ ΗΧΟΥ

Μαρόπουλος Παρασκευάς Π15086

Περίληψη

Υλοποίηση συστήματος αναγνώρισης ομιλίας(ASR-automatic speech recognition) όπου αναγνωρίζει 4-10 ψηφία που έχουν ειπωθεί σε μία πρόταση.

Περιγραφή

Τα σύστημα αναγνώρισης ομιλίας χρησιμοποιούν μεθόδους μηχανικής μάθησης όπου εξάγωγουν χαρακτηριστικά από δεδομένα με σκοπό να δημιουργήσουν μια αναπαράσταση του περιεχομένου.Στην περίπτωση της αναγνώρισης ομιλίας σκοπός είναι να βρούμε καλύτερη σειρά λέξεων που αντιστοιχούν στον ήχο βάση κάποιον ακουστικών και γλωσσικών μοντέλων. Λόγω του ότι αναφερόμαστε μόνο σε ψηφία δεν θα χρειαστούμε γλωσσικό μοντέλο(απαραίτητο για την ορθή σύνταξη στην περίπτωση ολόκληρων προτάσεων).

Βασικά σημεία για την κατασκευή ενός συστήματος ASR είναι:

- Εκπαίδευση ενός ακουστικού μοντέλου
- Εξαγωγή χαρακτηριστικών του σήματος.
- Χρήση των χαρακτηριστικών για την εύρεση της καλύτερης αντιστοιχίας με την χρήση των HMM και GMM.
- Πρόβλεψη της λέξης
- Εκτίμηση του συστήματος βάσει των επιδόσεων του.

MFCC

Αρχικά μετατρέπουμε το ηχητικό σήμα σε αναλογικό μέσω του μικροφώνου.Στη συνέχεια εξάγουμε τα χαρακτηριστικά του.
Για την εξαγωγή χαρακτηριστικών ενός σήματος χρησιμοποιούμε την μέθοδο Mel-frequency Cepstral Coefficients(MFCC) η οποία έχει ώς αποτέλεσμα 39 χαρακτηριστικά για κάθε “πλάνο” του σήματος.

Μετατροπή

Το πρώτο βήμα της μεθόδου είναι η μετατροπή του σήματος από αναλογικό σήμα σε σήμα διακριτού χρόνου.Τις περισσότερες φορές με συχνότητα δειγματοληψίας 8 ή 16 kHz(8000-16000 Hz).

Πρό-Έμφαση

Στη συνεχεια εφαρμόζουμε ένα φίλτρο προέμφασης όπου ενισχύει την ποσότητα ενέργεια στις υψηλές συχνότητες.Αυτό γίνεται για δύο λόγους, ο πρώτος είναι ότι βοηθάει στην καλύτερη αναγνώριση των φωνημάτων και δεύτερων οι ανθρωποι δυσκολευονται να κατανοήσουν τους ήχους όταν δεν υπάρχουν αυτές οι υψηλές συχνότητες.Για την ενίσχυση αυτή ισχύει η σχέση:

$x'[t] = x[t] - ax[t-1]$ με $0.95 < a < 0.99$

Παραθυροποίηση

Σε αυτό το κομμάτι ο αλγόριθμος κόβει την κυματομορφή του ακουστικού σήματος σε συρόμενα διαστήματα.Το πιό σύνηθες είναι τα διαστήματα να διαρκούν 20 με 25 ms και να ξεκινάνε ανα 10ms.Για να μην είναι απότομη η κοπή του διαστήματος χρησιμοποιείτε κάποιο από τα παρακάτω παράθυρα..

- Hamming
- Hanning
- Rectangular

Στην δική μας περίπτωση χρησιμοποιείται το παράθυρο του Hamming όπου στις άκρες του τμήματος του σήματος το πλάτος δεν έχει απότομη αλλαγή.
Κάθε πλαίσιο μπορει ορίζεται από τη σχέση

$x[n] = w[n]s[n]$

x: κομμένο πλαίσιο
s: αρχικο ηχητικό σήμα

Το παράθυρο ορίζεται ως
$w[n] = (1-a) - a \cos(2\pi n/(L-1))$ με L το πλάτος του παραθύρου.

DFT

Μετά εφαρμόζουμε τον διακριτό μετασχηματισμό Φουριέ για να πάρουμε πληροφορίες στον τομέα των συχνοτήτων.

Mel Filterbank

Για να μιμηθούμε τον τρόπο με τον οποίο ακούει ο άνθρωπος μετατρεπουμε τις πληροφορίες από τον τομέα συχνοτήτων εφαρμόζοντας τριγωνικά ζωνοπερατά φίλτρα.
Πρώτα όμως υψώνουμε στο τετράγωνο τα αποτελέσματα του DFT που αντιστοιχούν στην δύναμη της κάθε συχνότητας(ον. σπεκτογραμμα δύναμης DFT).
Μετά εφαρμόζουμε τα τριγωνικά φίλτρα Mel-Scale

$M(f) = 1127\ln(1 + f/700)$

Bark Scale

$b(f) = 13\arctan(0.00076f) + 3.5\arctan((f/500)^2)$

και το αποτέλεσμα είναι το σπεκτογραμμα δύναμης Mel-scale.

Log

Επειδή ο άνθρωπος είναι λιγότερο ευαίσθητος στις μικρές αλλαγές ενέργειας υψηλών ενεργειών από' ότι στις μικρές αλλαγές στις χαμηλές ενέργειες, το επόμενο βήμα είναι να αφαιρέσουμε τον λογάριθμο απο το σπεκτογραμμα δύναμης.

Cepstrum-IDFT

Τέλος για να αφαιρέσουμε τον τόνο του σήματος αλλά και να κάνουμε τους ήχους ανεξάρτητους μεταξύ τους εφαρμόζουμε το αντίστροφο διακριτό μετασχηματισμό φουριέ.Το αποτέλεσμα που έχουμε είναι μια σειρά από τιμές που όμως εμείς χρειαζόμαστε μόνος τις 12 πρώτες.
Για τις υπόλοιπες τιμές/χαρακτηριστικά ισχύει:
13η:Η συνολική ενέργεια κάθε πλαισίου.
14-26:Οι αλλαγές χαρακτηριστικών κάθε πλαισίου από το προηγούμενο,με τύπο

$$d(t) = \frac{c(t+1) - c(t-1)}{2}$$

27-39:Οι δυναμικές αλλαγές του d (t) από το τελευταίο πλαίσιο στο επόμενο πλαίσιο.

GMM-HMM

Για την εύρεση της καλύτερης ακολουθίας λέξεων(αριθμών στη δική μας περίπτωση) θα χρησιμοποιήσουμε μοντέλα HMM(hidden markov model) για να αναπαραστήσουμε το ακουστικό μοντέλο P(X|W).Για την μοντελοποίηση των χαρακτηριστικών (MFCC) μαθαίνουμε ένα GMM μοντέλο.

HMM

Εκπαίδευση

Η πιθανότητα παρατήρησης ενός παρατηρήσιμου δεδομένου μιας εσωτερικής κατάστασης ονομάζεται πιθανότητα εκπομπής. Η πιθανότητα μετάβασης από μια εσωτερική κατάσταση σε άλλη ονομάζεται πιθανότητα μετάβασης.
Στην περίπτωση των συστημάτων ASR τα παρατηρήσιμα δεδομένα είναι το περιεχόμενο του κάθε ηχητικού πλαισίου όπου τα αναπαριστούμε με τις παραμέτρους MFCC που βρήκαμε προηγουμένως.

Το πρώτο βήμα για την εκπαίδευση του μοντέλου είναι η συλλογή δεδομένων εκπαιδευσης.Στην περίπτωσή μας αρχεία ήχου με τα νούμερα 0-9.

Το μοντέο HMM έχει 5 παραμέτρους:
M: αριθμός καταστάσεων
N: αριθμός παρατηρήσεων σε κάθε κατάσταση
A: Πίνακας πιθανοτήτων μεταβάσεων
B: Πίνακας πιθανοτήτων παρατήρησης ανάλογα την κατάσταση
π: Αρχική κατάσταση.

Για την εκπαίδευση του Hidden Markov Model ανα λέξη χρησιμοποιούμε τον αλγόριθμο Baum-Welch(forward-backward algorithm). Ο Baum Welch χρησιμοποιεί τον αλγόριθμο προσδοκίας-μεγιστοποίησης για να βρεί την μέγιστη πιθανοφάνεια των παραμέτρων του HMM.
Ο αλγόριθμος αυτό υπολογίζει:
α(τ): Την εμπρόσθια πιθανότητα των καταστάσεων για κάθε βήμα στον άξονα του χρόνου.Δηλαδή ποια είναι η πιθανότητα να είμαστε στην κατάσταση S βάση όλες τις προηγούμενες καταστάσεις.
β(τ): Την “οπίσθια” πιθανότητα ,δηλαδή την πιθανότητα να είμαστε στην κατάσταση S ξέροντας τι επόμενος k καταστάσεις.
γ(τ): ο συνδυασμός των α,β δηλαδή την πιθανότητα να είμαστε στην κατάσταση S ξέροντας τις προηγούμενες αλλα και τις επόμενες καταστάσεις.
ξ(i,j): Η πιθανότητα την στιγμή t να είμαστε στην κατάσταση Si και την στιγμή t+1 να είμαστε στην κατάσταση Sj.

Εφόσον γτ(ι): η πιθανότητα να είμαστε στην κατάσταση Si την στιγμή t
Το σύνολο για όλες τις τιμες t = $\sum_{1}^{\tau-1} \gamma_\tau(i)$ ορίζεται ο αριθμός πρόβλεψης για το πόσες φορές η κατάσταση Si θα επισκεφθεί.
Ενώ το σύνολο των μετακινήσεων από την κατάσταση Si στην Sj, δηλαδη το σύνολο των ξτ(i,j)

εκφράζει το φορές η κατάσταση Si θα μετακινηθεί στην Sj.

Τέλος έχοντας ένα μοντέλο λ(A,B,π) ο αλγόριθμος Baum-Welch αναδιαμορφώνει του πινακες ανάλογα με τα δεδομένα(παρατηρήσεις) που υπάρχουν στα δεδομένα μάθησης.

Με α_hat = Σύνολο ξτ(i,j)/Σύνολο γτ(i)
β_ hat(u) = Σύνολο γτ(i,j){μόνο εάν έχει παρατηρηθεί η κατάσταση u}/Σύνολο γτ(i)
π_hat = γ1(τ)
Αυτό επαναλαμβάνεται μέχρι να πετύχουμε το επιθυμητό επίπεδο σύγκλισης.

Πρόβλεψη

Αφού έχουμε εκπαιδεύσει τα μοντέλα για κάθε ψηφίο για την πρόβλεψη ενός νέου ήχου χρησιμοποιούμε τον αλγοριθμο Viterbi.
Πρώτα εξάγουμε τα χαρακτηριστικά mfcc του σήματος και συνέχεια υπολογίζουμε τα σκοπ πιθανοφάνειας για κάθε ένα από τα 11 ψηφία και επιλέγουμε το μέγιστο σκόπ πιθανοφάνειας.
Για τον υπολογισμό των σκοπ ο αλγόριθμος Viterbi στοιχίζει με βέλτιστο τρόπο τα διανύσματα χαρακτηριστικών X του σήματος με τις καταστάσεις των μοντέλων HMM.

Απόδοση

Για να μπορέσουμε να μετρήσουμε την επιτυχία ενός συστήματος ASR πρέπει να αποτιμήσετε την απόδοσή του.
Οι τύποι σφαλμάτων σε ένα μοντέλο αναγνώρισης ομιλίας είναι 3:

1. Εισαγωγή λέξεων: Δηλαδή όταν αναγνωρίζονται περισσότερες λέξει από ότι ειπώθηκαν
2. Αντικατάσταση λέξεων: Σφάλμα σε μία λέξη, δηλαδή αναγνώριση λέξης διαφορετική από την σωστή.
3. Διαγραφή λέξεων: Παράλληψη λέξεων.

Στην περίπτωση αυτή ο ρυθμός επιτυχίας υπολογίζεται από τον τύπο

$$WER = \frac{NI + NS + ND}{|W|}$$

NI:πλήθος εισαγωγών
NS:πλήθος αντικαταστάσεων
ND:πλήθος διαγραφών
W:πλήθος λέξεων της πρότασης

Όμως στο δικό μας παράδειγμα(μόνο ψηφία και μεμονωμένα) το μέτρο απόδοσης είναι ο ρυθμό σφαλμάτων.Δηλαδή WER = F/|W|

Πρόγραμμα

Το πρόγραμμα είναι γραμμένο σε γλώσσα προγραμματισμού Python.
Για την υλοποίηση σου χρησιμοποιήθηκαν οι παρακάτω βιβλιοθήκες.:

- Numpy(https://numpy.org/doc/)
- Os(https://docs.python.org/3/library/os.html)
- Hmmlearn(https://hmmlearn.readthedocs.io/en/latest/):Δημιουργία εκπαίδευση και πρόβλεψη με βάση τα MFCC με την χρήση μοντέλων HMM,GMM
- Scipy.io.wavfile(https://kite.com/python/docs/scipy.io.wavfile)
- pickle(https://docs.python.org/3/library/pickle.html):Αποθήκευση και διάβασμα των μοντέλων HMM απο/σε αρχεία .pkl

Για το πρόγραμμα εύρεσης χαρακτηριστικών MFCC:

- Scipy.fft(https://docs.scipy.org/doc/scipy/reference/fft.html#module-scipy.fft):Λειτουργίες για τους μετασχηματισμού φουριέ.

Για την καταγραφή του ομιλητή:

- SoundDevice(https://python-sounddevice.readthedocs.io/en/0.3.15/)

Για το dataset των αρχείων wav:

- https://github.com/Jakobovski/free-spoken-digit-dataset

Εκτίμηση:

- Σε 400 δείγματα (1600 για εκπαίδευση)
  - 310 σωστα
  - 90 λάθος
  - WER = 22.5% (word error rate)