ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Επεξεργασία Φωνής και Φυσικής Γλώσσας
Χειμερινό εξάμηνο 2020 - 2021
2ο Εργαστήριο: Αναγνώριση φωνής με το KALDi TOOLKiT

Θεοδωρόπουλος Βασίλειος AM: 03117092
Τσακανίκα Χριστίνα AM: 03317012

Ανάπτυξη και σχολιασμός εννοιών MFCCs, Γλωσσικά & Ακουστικά Μοντέλα πρόταση & εκτίμηση βελτίωσης συστήματος

Η κατασκευή ενός συστήματος επεξεργασίας και αναγνώρισης φωνής, θεμελιώνεται σε 4 βασικά βήματα.

Πρώτο Βήμα: Πρόβλημα Αναγνώρισης

Το αρχικό βήμα, αποτελεί η επιλογή του προβλήματος αναγνώρισης, συμπεριλαμβανομένου του λεξικού αναγνώρισης, τους βασικούς ήχους που αναπαριστούν το λεξικό (τις μονάδες δηλαδή της ομιλίας), το γλωσσικό μοντέλο του προβλήματος και τέλος τη σημασιολογία του προβλήματος, εάν αυτή υπάρχει.
Στην περίπτωσή μας, κατά την δεύτερη εργαστηριακή άσκηση, το πρόβλημα αναγνώρισης, που περιγράφεται ως σύστημα αναγνώρισης φωνημάτων από ηχογραφήσεις, χρησιμοποιώντας δεδομένα τεσσάρων ομιλητών, ως βασικές μονάδες αναγνώρισης χρησιμοποιήθηκαν μοντέλα αναγνώρισης φωνημάτων, όπου κάθε λέξη συντέθηκε από ακολουθίες φωνημάτων.

Δεύτερο Βήμα: Σύνολο Χαρακτηριστικών Αναγνώρισης Mel-Frequency Cepstral Coefficients (MFCCs)

Επόμενο βήμα, αποτελεί η επιλογή ενός συνόλου χαρακτηριστικών για την αναγνώριση. Τα πιο δημοφιλή ακουστικά χαρακτηριστικά έχουν υπάρξει οι cepstral συντελεστές στον mel συχνοτικό χώρο (MFCCs) και οι παράγωγοι αυτών. Εκκινώντας από ένα αρχείο ήχου (πχ. usctimit_ema_f1_006.wav), το αναλογικό αυτό σήμα δειγματοληπτείται και κβαντιζεται με ρυθμούς από 8000 δείγματα/δευτερόλεπτο εώς 20000 δείγματα/δευτερόλεπτο, ενώ για την αντιστάθμιση της πτώσης φάσματος στις υψηλές συχνότητες χρησιμοποιείται ένα ανωπερατό φίλτρο πρώτης τάξης. Εν συνεχεία, το φιλτραρισμένο σήμα τμηματοποιείται σε πλαίσια L δειγμάτων, όπου τα διαδοχικά πλασια απέχουν R δείγματα. Οι τυπικές τιμές για τις παραμέτρους L, R αντιστοιχούν σε πλαίσια διαρκείας 15-40ms, όπου η ολίσθηση του παραθύρου είναι συνήθως 10ms. Σε κάθε πλαίσιο εφαρμόζεται ένα παράθυρο Hamming, πριν τη φασματική ανάλυση που χρησιμοποιεί μεθόδους γραμμικής πρόβλεψης. Προαιρετικά, χρησιμοποιούνται απλές μέθοδοι εξάλειψης θορύβου και οι συντελεστές πρόβλεψης που αναπαριστούν το φάσμα βραχέος χρόνου κανονικοποιούνται και μετατρέπονται σε cepstral συντελεστές του mel συχνοτικού χώρου. Απαλείφοντας την μεροληψία των cepstral συντελεστών, προκύπτει ένα σύνολο εξισωμένων mel-cepstrum

συντελεστών, καθώς και των πρώτων και δεύτερων παραγώγων τους. Κατά κύριο λόγο χρησιμοποιούνται 13 mfcc συντελεστές, 13 συντελεστές παραγώγου πρώτης τάξης, 13 συντελεστές παραγώγου δεύτερης τάξης, ενώ καταλήγουμε σε ένα διάνυσμα διαστασης D = 39 για πλαίσια που λαμβάνονται κάθε 10ms από το σήμα ομιλίας.

Τρίτο Βήμα: Εκπαίδευση Ακουστικών & Γλωσσικών Μοντέλων

Το κεντρικό πρόβλημα της Αυτόματης Αναγνώρισης Λόγου (Automatic Speech Recognition), αντιμετωπίζεται ως ένα στατιστικό πρόβλημα απόφασης. Συγκεκριμένα, διατυπώνεται ως μία διεργασία απόφασης με βάση την μέγιστη εκ των υστέρων πιθανότητα, όπου αναζητούμε την ακολουθία λέξεων W που μεγιστοποιεί τη εκ των υστέρων πιθανότητα P(W|X), της ακολουθίας με δεδομένη την ακολουθία των διανυσμάτων χαρακτηριστικών, X, δηλαδή:

$$\hat{W} = \arg\max_{W} P(W|X)$$

Εφαρμόζοντας τον κανόνα του Bayes στην παραπάνω σχέση και και αγνοώντας τον όρο του παρονομαστή, P(X), διότι είναι ανεξάρτητος από την ακολουθία λέξεων W, ως προς την οποία γίνεται η βελτιστοποίηση, λαμβάνουμε διαδοχικά τις σχέσεις:

$$\hat{W} = \arg\max_{W} \frac{P(X|W)P(W)}{P(X)}.$$

$$\hat{W} = \underbrace{\arg\max_{W}}_{\text{Step 3}} \underbrace{P_A(X|W)}_{\text{Step 1}} \underbrace{P_L(W)}_{\text{Step 2}}$$

όπου το πρώτο βήμα, $P_A(X|W)$, είναι ο υπολογισμός της πιθανότητας που σχετίζεται με το ακουστικό μοντέλο των ήχων ομιλίας της πρότασης W, το βήμα 2, $P_L(W)$, είναι ο υπολογισμός της πιθανότητας που σχετίζεται με το γλωσσικό μοντέλο των λέξεων της πρότασης και το βήμα 3 είναι ο υπολογισμός που σχετίζεται με την αναζήτηση μεταξύ όλων των ορθών προτάσεων της γλώσσας του προβλήματος, της ακολουθίας μέγιστης πιθανότητας.

Τέλος, μπορούμε να εκφράσουμε τη βέλτιστα αποκωδικοποιημένη ακολουθία λέξεων W, ως

$$W = \{w_1, w_2, \ldots, w_M\}$$

και το διάνυσμα χαρακτηριστικών, X, ως μία ακολουθία ακουστικών παρατηρήσεων που αντιστοιχούν σε T πλαίσια του σήματος φωνής η οποία έχει τη μορφή

$$X = \{x_1, x_2, \ldots, x_T\},$$

όπου η διάρκεια του σήματος φωνής είναι T πλαίσια (δηλαδή T φορές το βήμα ολίσθησης πλαισίων σε msec). Κάθε πλαίσιο Xt, είναι ένα διάνυσμα ακουστικών χαρακτηριστικών της μορφής:

$$x_t = (x_{t1}, x_{t2}, \ldots, x_{tD}),$$ D: πλήθος των ακουστικών χαρακτηριστικών σε κάθε πλαίσιο (39)

Το πρώτο βήμα της αναζήτησης της ακολουθίας λέξεων W που μεγιστοποιεί τη εκ των υστέρων πιθανότητα P(W|X), (step 1 PA(X|W)), περιγράφεται ως *Ακουστική Μοντελοποιηση*. Στο βήμα αυτό ανατίθενται πιθανότητες στις ακουστικές υλοποιήσεις μιας ακολουθίας λέξεων, δοθέντων των παρατηρούμενων ακουστικών διανυσμάτων. Δηλαδή

χρειάζεται να υπολογίσουμε την πιθανότητα να προήλθε η ακολουθία ακουστικών διανυσμάτων $X=\{X_1,X_2,...,X_T\}$ από την ακολουθία λέξεων $W = \{ W_1,W_2,...,W_M \}$ και αυτός ο υπολογισμός πρέπει να γίνει για όλες τις δυνατές ακολουθίες λέξεων. Ο υπολογισμός εκφράζεται ως εξής:

$$P_A(X|W) = P_A(\{X_1,X_2,...,X_T\}|\{W_1,W_2,...,W_M\})$$

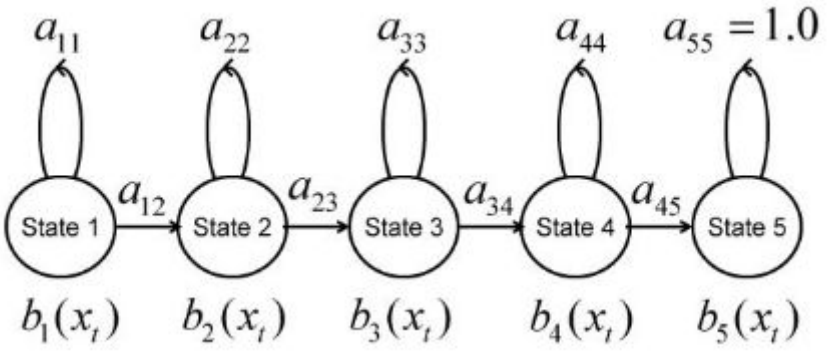
Αν κάνουμε την υπόθεση ότι το πλαίσιο t αντιστοιχίζεται με τον μοντέλο λέξης ι(t) και την κατάσταση j(t) του μοντέλου αυτού μέσω της συνάρτησης w και εάν επίσης υποθέσουμε ότι κάθε πλαίσιο είναι ανεξάρτητο από κάθε άλλο πλαίσιο η παραπάνω εξίσωση εκφράζεται εν τέλει ως το γινόμενο:

$$P_A(X|W) = \prod_{t=1}^{T} P_A\left(\mathbf{x}_t | w_{j(t)}^{i(t)}\right)$$

όπου συσχετίζουμε το κάθε πλαίσιο της X με μία μοναδική λέξη και κατάσταση της ακολουθίας λέξεων. Περισσότερες πληροφορίες για τη βέλτιστη αυτή συσχέτιση (αντιστοίχιση) πλαισίου X και κατάστασης, δίνονται παρακάτω, στην εκτενέστερη περιγραφή του ακουστικού μοντέλου.

Περιγραφή ακουστικού μοντέλου

Η πιο συχνή μέθοδος κατασκευής ακουστικών μοντέλων, τόσο για φωνήματα όσο και λέξεις, είναι το στατιστικό κατασκεύασμα γνωστό ως *κρυφό μοντέλο Markov* (*HMM*). Κάθε κατάσταση του HMM χαρακτηρίζεται από μία μίξη *Gaussian* πυκνοτήτων πιθανότητας (GMM), η οποία αναπαριστά τη στατιστική συμπεριφορά των διανυσμάτων χαρακτηριστικών Xt, στις καταστάσεις του μοντέλου. Επιπλέον, το HMM χαρακτηρίζεται επίσης από ένα σύνολο μεταβάσεων μεταξύ καταστάσεων, $A=\{a_{ij},1\leq i, j\leq Q\}$, για ένα μοντέλο Q καταστάσεων, οι οποίες περιγράφουν την πιθανότητα να γίνει μία μετάβαση από την κατάσταση i στην κατάσταση j σε κάθε πλαίσιο, καθορίζοντας έτσι τη χρονική ακολουθία των διανυσμάτων χαρακτηριστικών κατά την διάρκεια της λέξης. Συνήθως, οι μεταβάσεις από μία κατάσταση στον εαυτό της, aij, είναι μεγάλες (πιθανότητα κοντά στο 1), ενώ οι καταστάσεις άλματος $a_{12},a_{23},a_{34},a_{45}$ είναι μικρές (πιθανότητα κοντά στο μηδέν). Το πλήρες *HMM* μίας λέξης Q καταστάσεων γράφεται στη γενική περίπτωση ως λ(A,B,π), με μητρώο μετάβασης καταστάσεων $A=\{a_{ij}, 1\leq i, j\leq Q\}$, πυκνότητα πιθανότητα παρατήρησης εκπομπών ανά κατάσταση B={bj(xt), 1≤j≤Q} και κατανομή αρχικής κατάστασης, $\pi=\{\pi_i, 1\leq i\leq Q\}$, όπου το π_1 τίθεται ίσο με ένα και όλα τα υπόλοιπα τίθενται ίσα με μηδέν, για τα μοντέλα είδους αριστερά προς τα δεξιά, όπως αυτό του παρακάτω σχήματος.



Το παραπάνω μοντέλο αποτελεί *HMM* με πέντε καταστάσεις και μηδενικές πιθανότητες υπερπήδησης καταστάσεων, δηλαδή $a_{ij} = 0,\ j \geq i+2$. Η συνάρτηση πυκνότητας για την κατάσταση i δηλώνεται ως $b_i(X_t)$, $1 \leq i \leq 5$.

Προκειμένου να γίνει η εκπαίδευση του *HMM* κάθε λέξης χρησιμοποιείται ένα επισημειωμένο σύνολο προτάσεων εκπαίδευσης, με σκοπό να ακολουθηθεί μία αποτελεσματική διαδικασία εκπαίδευσης βάσει του αλγορίθμου *Viterbi.* Ο αλγόριθμος αυτός έχει ως στόχο την εύρεση διαδρομής μέγιστης πιθανότητας που συνδέει κάθε διάνυσμα χαρακτηριστικών της ομιλίας $X=\{x_1,x_2,...,x_T\}$, με μία μοναδική κατάσταση μοντέλου.

Ορίζοντας την ποσότητα $\delta_t(j)$, ως την πιθανότητα να βρεθεί η βέλτιστη διαδρομή στην κατάσταση *i*, στο χρονικό πλαίσιο *t*, αφού έχουν παρέλθει t διανύσματα της πρότασης και δοθέντος του μοντέλου *λ*

$$\delta_t(j) = \max P[q_1, q_2, q_3, \ldots q_{t-1}, q_t = i, X1, X2, X3,... Xt| \lambda]$$

Ο αλγόριθμος *Viterbi* υπολογίζει τη διαδρομή υψηλότερης πιθανότητας που λαμβάνει υπόψη όλα τα Τ διανύσματα χαρακτηριστικών της πρότασης, μέσω αναδρομής.

Στη συνέχεια, για το δεύτερο βήμα της αναζήτησης της ακολουθίας λέξεων W που μεγιστοποιεί τη εκ των υστέρων πιθανότητα P(W|X), (step 2 $P_L(W)$) αναγκαία κρίνεται η περιγραφή του Γλωσσικού Μοντέλου.

Περιγραφή Γλωσσικού Μοντέλου

Το γλωσσικό μοντέλο αναθέτει πιθανότητες σε ακολουθίες λέξεων, με βάση την πιθανοφάνεια της εμφάνισης της ακολουθίας των λέξεων στο περιβάλλον (συμφραζόμενα) του προβλήματος για το οποίο λειτουργεί το σύστημα αναγνώρισης ομιλίας. Όπως ακριβώς προαναφέρθηκε, σκοπός του γλωσσικού μοντέλου είναι να καταστήσει εφικτό τον υπολογισμό της εκ των προτέρων πιθανότητας P_L, του αλφαριθμητικού της λέξης W. Ο πιο δημοφιλής τρόπος κατασκευής του γλωσσικού μοντέλου είναι μέσω της χρήσης μιας στατιστικής γραμματικής λέξεων τύπου N-gram, η οποία εκτιμάται με βάση ένα μεγάλο σύνολο κειμένων εκπαίδευσης, το οποίο έχει προκύψει χειροκίνητα ή έχει παραχθεί ξεκινώντας από μια γενική βάση που έχει εφαρμογή σε ένα μεγάλο εύρος προβλημάτων. Εάν τώρα υποθέσουμε ότι η πιθανότητα μιας συγκεκριμένης λέξης μέσα σε μία πρόταση, εξαρτάται αποκλειστικά από τις προηγούμενες N-1 λέξεις,

έχουμε τη βάση ενός μοντέλου *N-gram* για τη γραμματική. Επομένως, υποθέτουμε ότι μπορούμε να γράψουμε την πιθανότητα της πρότασης W, σύμφωνα με το μοντέλο N-gram της γλώσσας ως:

$$\begin{aligned} P_L(W) &= P_L(w_1, w_2, \ldots, w_M) \\ &= \prod_{n=1}^{M} P_L(w_n | w_{n-1}, w_{n-2}, \ldots, w_{n-N+1}) \end{aligned}$$

και η εκτίμηση αυτής τη πιθανότητας πραγματοποιείται μετρώντας τις σχετικές συχνότητες εμφάνισης N-άδων λέξεων στο σύνολο εκπαίδευσης. Παρατίθεται ένα παράδειγμα υπολογισμού 3-gramm πιθανοτήτων:

$$P(w_n | w_{n-1}, w_{n-2}) = \frac{C(w_{n-2}, w_{n-1}, w_n)}{C(w_{n-2}, w_{n-1})}$$

Μολονότι η μέθοδος των N-gram γραμματικών λειτουργεί ικανοποιητικά καλά σε γενικές γραμμές, δεν μπορεί να υπερνικήσει το πρόβλημα ότι οι μετρήσεις είναι συχνά εσφαλμένες, εξαιτίας της αραιότητας του συνόλου εκπαίδευσης. Ακόμα και για ένα λεξικό αρκετών χιλιάδων λέξεων, ακόμα και αν το σύνολο κειμένων εκπαίδευσης αποτελείται από εκατομμύρια προτάσεις έχει παρατηρηθεί ότι περισσότερο από το 50% των δυνατών τριάδων (N = 3) είναι δυνατόν να εμφανιστεί μία ή και καμία φορά στο σύνολο εκπαίδευσης καθιστώντας την εκτίμηση τριάδων ανακριβή και παραμορφώνοντας τον υπολογισμό της πιθανότητας για πολλές προτάσεις. Για την αποφυγή των τελευταίων, είναι ορθότερο από στατιστικής πλευράς να να εφαρμόσουμε έναν αλγόριθμο λείανσης σε όλες τις εκτιμήσεις τριάδων, σύμφωνα με τον παρακάτω τρόπο:

$$\begin{aligned} \hat{P}(w_n | w_{n-2} w_{n-1}) \;=\; & \lambda_1 P(w_n | w_{n-2} w_{n-1}) \\ & + \lambda_2 P(w_n | w_{n-1}) \\ & + \lambda_3 P(w_n) \end{aligned}$$

όπου όλοι οι συντελεστές λ έχουν άθροισμα 1:

$$\sum_i \lambda_i = 1$$

Τέλος, για το βήμα 3 της αναζήτησης της ακολουθίας λέξεων W που μεγιστοποιεί τη εκ των υστέρων πιθανότητα P(W|X), που δεν είναι άλλο από τον υπολογισμό που σχετίζεται με την αναζήτηση μεταξύ όλων των ορθών προτάσεων της γλώσσας του προβλήματος, της ακολουθίας μέγιστης πιθανότητας, χρησιμοποιούνται δίκτυα πεπερασμένων καταστάσεων *FSN* που ελαττώνουν το υπολογιστικό φορτίο κατά τάξεις μεγέθους, η ανάλυση των οποίων ξεφεύγει των ορίων της δεύτερης εργαστηριακής άσκησης.

Τέταρτο Βήμα: Έλεγχος & Αποτίμηση Επίδοσης Συστήματος Αναγνώρισης πρόταση /εκτίμηση βελτίωσης συστήματος

Στο σημείο αυτό, προκειμένου να βελτιώσουμε την απόδοση οποιουδήποτε συστήματος αναγνώρισης ομιλίας, πρέπει να υπάρχει ένας αξιόπιστος και στατιστικά αξιόλογος τρόπος αποτίμησης της επίδοσης ενός συστήματος αναγνώρισης ομιλίας, με βάση ένα ανεξάρτητο

σύνολο επισημειωμένων περιπτώσεων ομιλίας. Τυπικά, χρησιμοποιούνται ο ρυθμός εσφαλμένων λέξεων, *word error rate* (*phoneme error rate* στην περίπτωσή μας) και ο ρυθμός εσφαλμένων προτάσεων (*sentence/task error rate*), ως μέτρα επιδόσεων αναγνώρισης.
Πιο συγκεκριμένα, για τη βελτίωση της επίδοσης του συστήματος που οι ίδιοι αναπτύξαμε, θα μπορούσαμε να χρησιμοποιήσουμε έναν πιο αυξημένο πληθυσμό ομιλητών (πχ 25 ομιλητές αντί των τεσσάρων) με κάθε ομιλητή να προφέρει κάθε μία από τις δεδομένες προτάσεις του αρχείου *transcription.txt* και στην συνέχεια να αποτιμήσουμε τον ρυθμό εσφαλμένων φωνημάτων κατά επέκταση τον ρυθμό εσφαλμένων προτάσεων.
Τέλος, μία ακόμη μέθοδος που λαμβάνεται σοβαρά υπόψιν για τη βελτίωση της απόδοσης, είναι η επιλογή εύρωστων στο θόρυβο ακουστικών χαρακτηριστικών. Εφοσον, τα προς επεξεργασία αρχεία ήχου, εμπεριέχουν θορύβους και περιόδους σιωπής, το εύρος των μεθόδων που έχουν προταθεί για τη βελτίωση της ευρωστίας συνοψίζεται σε τρεις μεθόδους ταιριάσματος, μία στο επίπεδο σήματος μέσω βελτίωσης της ποιότητας του σήματος φωνής, μία στο επιπεδο χαρακτηριστικών χρησιμοποιώντας κάποιο είδος μεθόδου κανονικοποίησης χαρακτηριστικών, και τέλος, μία στο επί[εδο μοντέλου χρησιμοποιώντας κάποιο είδος μεθόδου προσαρμογής.