ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ
Χειμερινό Εξάμηνο 2016-17

2η Εργαστηριακή Άσκηση:
Εξαγωγή χαρακτηριστικών από φωνή για χρήση σε εφαρμογή αναγνώρισης

ΠΕΡΙΓΡΑΦΗ

Σκοπός είναι η υλοποίηση ενός συστήματος επεξεργασίας και αναγνώρισης φωνής, με εφαρμογή σε αναγνώριση μεμονωμένων λέξεων. Το πρώτο μέρος αποσκοπεί στην εξαγωγή κατάλληλων ακουστικών χαρακτηριστικών από φωνητικά δεδομένα, χρησιμοποιώντας συναρτήσεις οι οποίες σας δίνονται για τους σκοπούς της εργαστηριακής άσκησης. Τα εν λόγω χαρακτηριστικά είναι στην ουσία ένας αριθμός συντελεστών cepstrum που εξάγονται μετά από ανάλυση των σημάτων με μια ειδικά σχεδιασμένη συστοιχία φίλτρων (filterbank). Η συστοιχία αυτή είναι εμπνευσμένη από ψυχοακουστικές μελέτες.

Πιο συγκεκριμένα, το σύστημα που θα αναπτύξετε αφορά σε αναγνώριση μεμονωμένων ψηφίων (isolated digits) στα Αγγλικά. Τα δεδομένα που θα χρησιμοποιήσετε περιέχουν εκφωνήσεις 9 ψηφίων από 15 διαφορετικούς ομιλητές σε ξεχωριστά .wav αρχεία. Συνολικά θα βρείτε 133 αρχεία, αφού 2 εκφωνήσεις θεωρήθηκαν προβληματικές και δεν έχουν συμπεριληφθεί. Τα ονόματα των αρχείων (π.χ. eight8.wav) υποδηλώνουν τόσο το ψηφίο που εκφωνείται (π.χ. eight), όσο και τον ομιλητή (οι ομιλητές είναι αριθμημένοι από 1-15). Οι εκφωνήσεις έχουν ηχογραφηθεί με συχνότητα δειγματοληψίας ίση με Fs = 16kHz και η διάρκειά τους διαφέρει.

ΕΚΤΕΛΕΣΗ

Χρησιμοποιώντας τα δεδομένα που κατεβάσατε ήδη (κατά την προπαρασκευή, από το link http://cvsp.cs.ntua.gr/courses/patrec/labs_material2014/digits2016.zip) καθώς και τον κώδικα που αναπτύξατε από την προπαρασκευή εκτελέστε τα παρακάτω βήματα, χρησιμοποιώντας τα εργαλεία: HMM, KPMtools, KPMstats, netlab, PRTools, Matlab.

Σημείωση: Τα βήματα 1-9 αποτελούν μέρος της προπαρασκευής η οποία έχει προηγηθεί και επαναλαμβάνονται μόνο για λόγους πληρότητας της άσκησης.

Βήμα 1
Εισάγετε τα δεδομένα στο περιβάλλον Matlab και από κάθε ένα .wav αρχείο εξάγετε ένα σήμα φωνής $s_0(n)$.

Βήμα 2
Κάθε σήμα $s_0(n)$ πρέπει να περάσει από ένα σύστημα προέμφασης, δίνοντας στην έξοδο σήμα $s_p(n)$. Το σύστημα προέμφασης έχει την εξής συνάρτηση μεταφοράς:

$$H_{preemph}(z) = 1 - \tilde{a}\, z^{-1}$$

Η παράμετρος είναι $\tilde{a} = 0.97$.

Βήμα 3
Το κάθε σήμα $s_p(n)$ χωρίζεται σε πλαίσια $s_i(n) = s_p(n)\, w(n(i-1)N), i = 1 \ldots N_f$, όπου N_f το πλήθος των πλαισίων. Τα πλαίσια είναι διάρκειας ίσης με $T(ms)$ με $T_{overlap}$ επικάλυψη. Προκειμένου να

ελαχιστοποιηθούν οι ασυνέχειες στα άκρα των πλαισίων, σε κάθε πλαίσιο εφαρμόζεται παραθύρωση Hamming. Το παράθυρο Hamming στο χρόνο έχει μορφή:

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right), 0\le n\le N-1$$

Βήμα 4

Ανάλυση σε συστοιχία φίλτρων σε κλίμακα Mel. Η συστοιχία φίλτρων που θα χρησιμοποιηθεί αποτελείται από έναν αριθμό Q τριγωνικών φίλτρων $H^j, j=1,\ldots,Q$, των οποίων οι κεντρικές συχνότητες f_c^j είναι ισοκατανεμημένες με βάση την κλίμακα mel:

$$f_c^j=2595\cdot\log\left(1+\frac{f^j}{700}\right), j=1\ldots Q$$

όπου f^j είναι οι κεντρικές συχνότητες στη γραμμική κλίμακα. Θεωρείστε επιπλέον οτι $H^j(f_c^j)=1$.

Το εύρος ζώνης b^j του κάθε φίλτρου προσδιορίζεται έτσι ώστε οι συχνότητες αποκοπής του να ταυτίζονται με τις κεντρικές συχνότητες των δυο γειτονικών του φίλτρων στην κλίμακα mel.

Βήμα 5

Υπολογίστε την ενέργεια $E_i(j)$ απόκρισης κάθε (καναλιού) φίλτρου j της συστοιχίας με είσοδο το σήμα $s_i(n)$. Απεικονίστε γραφικά τους συντελεστές $E_i(j)$ για δυο διαφορετικά πλαίσια. Στην ίδια γραφική παράσταση (ξεχωριστά για κάθε πλαίσιο).

Βήμα 6

Στη συνέχεια εξάγετε τους συντελεστές $G_i(j)$, όπου:

$$G_i(j)=\log(E_i(j)), j=1,\ldots,Q$$

Βήμα 7

Εφαρμόστε τον μετασχηματισμό συνημιτόνου (DCT) στούς συντελεστές $G_i(j)$, οπότε και θα προκύψουν οι συντελεστές $C_i(n)$:

$$C_i(n)=\sum_{j=1}^{Q} G_i(j)\cos\left(n\left(j-\frac{1}{2}\right)\frac{\pi}{Q}\right), n=0\ldots N_c-1, N_c<Q$$

Επιλέξτε ένα ψηφίο k_1 . Απεικονίστε γραφικά τα ιστογράμματα των συντελεστών $C_i(n_1)$, $C_i(n_2)$ για το σύνολο των πλαισίων που αντιστοιχούν στις εκφωνήσεις του ψηφίου αυτού. Επαναλάβατε για ένα ψηφίο k_2 .

Βήμα 8

Ανακατασκευάστε την έξοδο της συστοιχίας φίλτρων από τους συντελεστές $C_i(n)$ για τα πλάισια που επιλέξατε στο Βήμα 5. Απεικονίστε γραφικά την ανακατασκευασμένη έξοδο, $\hat{E}_i(j)$. Απεικονίστε το φάσμα ισχύος του σήματος $s_i(n)$ στην ίδια γραφική παράσταση.

Βήμα 9

Για κάθε εκφώνηση του ψηφίου k_1 υπολογίστε τις μέσες τιμές κατά μήκος όλων των πλαισίων των συντελεστών $C_i(n_1)$, $C_i(n_2)$. Επαναλάβατε για κάθε εκφώνηση και για κάθε ψηφίο. Απεικονίστε τα 133 ζεύγη τιμών που βρήκατε στο επίπεδο. Χρησιμοποιήστε διαφορετικό σύμβολο για κάθε ψηφίο (τάξη). Στο ίδιο γράφημα απεικονίστε και τα 9 ζεύγη μέσων τιμών, ένα ζεύγος για κάθε ψηφίο. Σχολιάστε ως προς τη δυνατότητα διαχωρισμού μεταξύ των ψηφίων.

Τα παρακάτω βήματα δεν αποτελοπύν μέρος της προπαρασκευής.

Βήμα 10

Αρχικοποίηση μοντέλων: Τα κρυφά Μαρκοβιανά μοντέλα που θα χρησιμοποιηθούν είναι της μορφής left-right με N_s καταστάσεις. Συγκεκριμένα, αν $A=\{a_{ij}\}$ είναι ο πίνακας μεταβάσεων του μοντέλου, τότε $a_{ij}=0\ για\ j<i$, ενώ οι αρχικές πιθανότητες των καταστάσεων είναι:

$$\pi_i=\begin{cases}0 & i\neq 1\\ 1 & i=1\end{cases}$$

Επιπλέον επιτρέπονται μεταβάσεις μόνο μεταξύ διαδοχικών καταστάσεων, δηλαδή υπάρχει ο περιορισμός $a_{ij}=0\ για\ j>i+1$.

Ένα διάνυσμα ακουστικών χαρακτηριστικών, όπως αυτό εξάγεται από την επεξεργασία ενός πλαισίου φωνής, αποτελεί μια πιθανή παρατήρηση σε κάποια κατάσταση. Λόγω του οτι είναι επιτρεπτές συνεχείς μεταβολές τέτοιων παρατηρήσεων, η πιθανότητα τους μοντελοποιείται με γκαουσιανές κατανομές, μια $(N_m=1)$ για κάθε χαρακτηριστικό.

Βήμα 11

Στη φάση αυτή εκπαιδεύονται τα 9 μοντέλα με χρήση του αλγορίθμου Expectation Maximization. Ο αλγόριθμος εφαρμόζεται για καθορισμένο πλήθος επαναλήψεων N_{iter} ή εως να υπάρξει σύγκλιση. Η σύγκλιση ελέγχεται μέσω της μεταβολής του αλγορίθμου της πιθανοφάνειας (Log Likelihood, πιθανότητα των δεδομενων με γνωστό μοντέλο). Για την εκπαίδευση κάθε μοντέλου (που αντιστοιχεί σε κάποιο ψηφίο) χρησιμοποιείστε όλα τα διαθέσιμα δεδομένα για το ψηφίο αυτό (όπως εξήχθησαν στο πρώτο μέρος).

Βήμα 12

Αναγνώριση μεμονωμένων ψηφίων – Testing. Προκειμένου να εφαρμοστεί η διαδικασία της αναγνώρισης εφαρμόζεται η διαδικασία εξαγωγής χαρακτηριστικών και στα δεδομένα τα οποία προορίζονται για testing. Ολοκληρώνοντας τη διαδικασία της εκπαίδευσης, έχετε καταλήξει στις εκτιμήσεις των παραμέτρων των 9 μοντέλων (δηλαδή ένα μοντέλο για κάθε ψηφίο). Στη συνέχεια υπολογίζεται ο λογάριθμος της πιθανοφάνειας log likelihood για κάθε εκφώνηση η οποία ανήκει στο σύνολο των δεδομένων για αναγνώριση. Το μοντέλο το οποίο δίνει τη μέγιστη πιθανοφάνεια είναι και το αποτέλεσμα της αναγνώρισης για τη συγκεκριμένη εκφώνηση. Τέλος, για κάθε μοντέλο (ψηφίο) υπολογίζεται το πλήθος των αποτελεσμάτων όπως αυτά κατανέμονται στις διαφορετικές κατηγορίες ψηφίων.

Βήμα 13

Για ένα επιλεγμένο ψηφίο k_1 να παραστήσετε γραφικά τη λογαριθμική πιθανοφάνεια ως συνάρτηση του πλήθους των επαναλήψεων.

Βήμα 14

Confusion Matrix: Σχηματίστε έναν πίνακα ο οποίος θα περιέχει τα αποτελέσματα του testing. Πιο συγκεκριμένα ο πίνακας θα περιλαμβάνει για κάθε ψηφίο k_1 9 αριθμούς οι οποίοι θα αντιστοιχούν στα πλήθη των εκφωνήσεων του k_1 , όπως αυτές κατηγοριοποιήθηκαν σε κάθε μια από τος 9 κατηγορίες. Επίσης, υπολογίστε ένα ολικό ποσοστό αναγνώρισης ως το ποσοστό των σωστά κατηγοριοποιημένων εκφωνήσεων.

Βήμα 15

Για κάθε ψηφίο k_1 και για όλες τις εκφωνήσεις που του αντιστοιχούν παραστήστε γραφικά ως συνάρτηση του χρόνου την πιο πιθανή ακολουθία καταστάσεων. Η ακολουθία αυτή υπολογίζεται με χρήση του αλγορίθμου Viterbi.

Προχωρήστε σε κατ'οίκον ολοκλήρωση των βημάτων εκείνων που δεν προλάβατε κατά τη διεξαγωγή του εργαστηρίου.

ΠΑΡΑΔΟΤΕΑ

Αφορά τα εκτός προπαρασκευής βήματα μόνο.

(1) Σύντομη αναφορά (σε pdf) που θα περιγράφει τη διαδικασία που ακολουθήθηκε σε κάθε βήμα, καθώς και τα σχετικά αποτελέσματα.
(2) Κώδικας (συνοδευόμενος από σύντομα σχόλια).

Συγκεντρώστε τα (1) και (2) σε ένα .zip αρχείο το οποίο πρέπει να αποσταλεί μέσω του mycourses.ntua.gr εντός της καθορισμένης προθεσμίας.

ΣΗΜΕΙΩΣΕΙΣ

1. **Training:** Ενδεικτικά αναφέρονται τα ακόλουθα. Το πλήθος των καταστάσεων N_s είναι στο εύρος [5, 9]. Το πλήθος N_m των γκαουσιανών είναι στο εύρος [1, 3]. Το πλήθος των επαναλήψεων N_{iter} είναι στο εύρος [5, 15]. Εναλλακτικές τοπολογίες των μοντέλων θα μπορούσαν να επιτρέπουν μεγαλύτερη ελευθερία μεταβάσεων μεταξύ καταστάσεων. Χρήσιμες συναρτήσεις: mhmm_em, mixgauss_init.

2. **Testing:** Το σύνολο των διαθέσιμων δεδομένων χωρίζεται σε training και testing δεδομένα με αναλογίες ενδεικτικά 70% και 30% αντίστοιχα. Χρήσιμες συναρτήσεις: mhmm_logprob, viterbi.

3. **Toolboxes:** HMM, KPMtools, KPMstats, netlab, PRTools .
4. Στα ερωτήματα στα οποία χρειάζεται επιλογή μεμονωμένου ψηφίου k_1 , επιλέξτε το σε αντιστοιχία με το πρώτο μέρος της εργαστηριακής άσκησης (σύμφωνα με τον αριθμό μητρώου σας).