ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Επεξεργασία Φωνής και Φυσικής Γλώσσας

Προπαρασκευή 2ου Εργαστηρίου: Αναγνώριση φωνής με το Kaldi toolkit

1 Περιγραφή

Σκοπός της άσκησης αυτής είναι η υλοποίηση ενός συστήματος επεξεργασίας και αναγνώρισης φωνής με το εργαλείο Kaldi, το οποίο χρησιμοποιείται ευρέως στον ερευνητικό τομέα, αλλά και όχι μόνο, για την εκπαίδευση state-of-the-art συστημάτων αναγνώρισης φωνής.

Πιο συγκεκριμένα, το σύστημα που θα αναπτύξετε αφορά σε αναγνώριση φωνημάτων (Phone Recognition) από ηχογραφήσεις της USC-TIMIT. Θα σας δοθούν δεδόμενα audio από 4 διαφορετικούς ομιλητές, με τα αντίστοιχα transcriptions, ώστε να εκπαιδεύσετε και να εκτιμήσετε το σύστημά σας.

Η διαδικασία σχεδιασμού του συστήματος μπορεί να χωριστεί σε 4 μέρη. Το πρώτο μέρος αποσκοπεί στην εξαγωγή κατάλληλων ακουστικών χαρακτηριστικών από τα φωνητικά δεδομένα (Mel-Frequency Cepstral Coefficients). Τα εν λόγω χαρακτηριστικά είναι στην ουσία ένας αριθμός συντελεστών cepstrum που εξάγονται μετά από ανάλυση των σημάτων σημάτων φωνής με μια ειδικά σχεδιασμένη συστοιχία φίλτρων (Mel filterbank). Η συστοιχία αυτή είναι εμπνευσμένη από το μη γραμμικό τρόπο που το ανθρώπινο αυτί αντιλαβάνεται τον ήχο και ειδικά σχεδιασμένη από ψυχοακουστικές μελέτες. Το δεύτερο μέρος αφορά τη δημιουργία γλωσσικών μοντέλων από τα transcriptions του σετ δεδομένων, τα οποία θα δίνουν την a priori πιθανότητα στο τελικό σύστημα. Το τρίτο μέρος αφορά την εκπαίδευση των ακουστικών μοντέλων χρησιμοποιώντας τα ακουστικά χαρακτηριστικά τα οποία εξήχθησαν. Τέλος, συνδυάζοντας τις παραπάνω μονάδες μπορεί να κατασκευαστεί το τελικό σύστημα αναγνώρισης φωνής, το οποίο δεδομένου ενός σήματος φωνής, εξάγει τα ακουστικά χαρακτηριστικά και τα χρησιμοποιεί ώστε να αποκωδικοποιήσει το σήμα σε μία ακολουθία φωνημάτων ή λέξεων.

2 Θεωρητικό υπόβαθρο

Κατά την προπαρασκευή θα πρέπει να εξοικειωθείτε με συγκεκριμένες έννοιες που θα χρησιμοποιηθούν κατά τη διεξαγωγή του εργαστηρίου. Συγκεκριμένα, θα θέλαμε να γνωρίζετε για τις παρακάτω έννοιες:

1. Mel-frequency Cepstral Coefficients (MFCCs)
2. Γλωσσικά Μοντέλα (Language Models)
3. Φωνητικά Μοντέλα (Acoustic Models)

Στην τελική σας αναφορά, θα θέλαμε εν συντομία να αναπτύξετε τόσο τις παραπάνω έννοιες όσο και να σχολιάσετε την απόδοση τους. Μην μείνετε στα βήματα των βασικών αλγορίθμων αλλά προσπαθείστε να προτείνετε/εκτιμήσετε πως θα μπορούσε να βελτιωθεί το σύστημα αναγνώρισης φωνής που έχετε αναπτύξει.

Ως προετοιμασία για το εργαστήριο διαβάστε τα παρακάτω:

- Κεφάλαιο 14 από το βιβλίο του μαθήματος **[R&S] Theory and Applications of Digital Speech Processing** των Lawrence R. Rabiner and Ronald W. Schafer (Pearson, 2011), σχετικά με Automatic Speech Recognition (ASR) συστήματα.
- Mel-frequency Cepstral Coefficients (MFCCs)
- GMM-HMM for acoustic modeling
- Kaldi tutorial 1
- Kaldi tutorial 2
- Kaldi tutorial 3

3 Βήματα προπαρασκευής

1. Εγκαταστήστε το Kaldi σύμφωνα με τις οδηγίες που θα δωθούν στις διευκρινίσεις του helios.

2. Εξοικειωθείτε με το εργαλείο Kaldi . Η γλώσσα με την οποία έχει αναπτυχθεί είναι C++, αλλά οι κύριες λειτουργίες που μας ενδιαφέρουν καλούνται από bash scripts. Υπάρχουν ήδη υλοποιημένες διαδικασίες για την ανάπτυξη μοντέλων αναγνώρισης φωνής για πολλά συνηθισμένα σετ δεδομένων μέσα στο φάκελο *egs* του Kaldi. Παρ' όλα αυτά, για το σετ δεδομένων που θα σας δοθεί θα πρέπει να το υλοποιήσετε μόνοι σας τη διαδικασία από την αρχή.

3. Κατεβάστε τα δεδομένα από το παρακάτω link:

   `https://drive.google.com/file/d/1_mIoioHMeC2HZtIbGs1LcL4kkIF696nB/view?usp=sharing`

   Τα δεδομένα περιλαμβάνουν ηχογραφήσεις από 4 ομιλητές με ονόματα: m1, m3 (άντρες) και f1, f5 (γυναίκες). Σε κάθε ομιλητή αντιστοιχούν 460 προτάσεις (Προσοχή: στον ομιλητή m1 λείπουν οι προτάσεις 231 έως 235 λόγω σφάλματος στην ηχογράφηση).

   Τα αρχεία ήχου βρίσκονται στο φάκελο *wav*, χωρισμένα σε φακέλους ανάλογα με το όνομα του ομιλητή και το όνομα κάθε αρχείου περιγράφει σε ποιον ομιλητή και σε ποια πρόταση αντιστοιχεί. Στο αρχείο *transcription.txt* θα βρείτε το κείμενο που εκφωνούν οι ομιλητές σε κάθε πρόταση (1η γραμμή → 1η πρόταση, 2η γραμμή → 2η πρόταση κ.ο.κ.) και στο φάκελο *filesets* θα βρείτε ποιές προτάσεις αντιστοιχούν στο σετ εκπαίδευσης, στο σετ επαλήθευσης και στο σετ αποτίμησης (training, validation, testing).

4. Κατασκευή αρχικού σκελετού:

   - Μέσα στο φάκελο *egs* δημιουργείστε ένα φάκελο *usc*, μέσα στον οποίο θα εργάζεστε από εδώ και πέρα.
   - Δημιουργείστε το φάκελο *data* και τους υποφακέλους *data/train*, *data/dev*, *data/test*, μέσα στους οποίους θα δημιουργήσετε αρχεία-δείκτες τα οποία θα περιγράφουν τα δεδομένα εκπαίδευσης, επαλήθευσης και αποτίμησης αντίστοιχα.
   - Μέσα σε κάθε έναν από αυτούς τους 3 φακέλους θα πρέπει να δημιουργήσετε τα εξής αρχεία:
     - uttids: περιέχει στην κάθε του γραμμή ένα μοναδικό συμβολικό όνομα για κάθε πρόταση του συγκεκριμένου συνόλου δεδομένων (δηλαδή το περιεχόμενο των αρχείων στο φάκελο *filesets*) τα οποία από εδώ και πέρα θα αναφέρουμε ως utterance_ids
     - utt2spk: περιέχει σε κάθε γραμμή τον ομιλητή που αντιστοιχεί σε κάθε πρόταση και είναι της μορφής:

       utterance_id_1 <κενό> speaker_id
       utterance_id_2 <κενό> speaker_id
       κ.ο.κ.

       όπου ως speaker_id επιλέγουμε αντίστοιχα τα m1, m3, f1, f5
     - wav.scp: περιέχει τη θέση του αρχείου ήχου που αντιστοιχεί σε κάθε πρόταση και είναι της μορφής:

       utterance_id_1 <κενό> /path/to/wav1
       utterance_id_2 <κενό> /path/to/wav2
       κ.ο.κ.

     - text: περιέχει το κείμενο που αντιστοιχεί στην κάθε πρόταση και είναι της μορφής:

       utterance_id_1 <κενό> <utterance 1 text>
       utterance_id_2 <κενό> <utterance 2 text>
       κ.ο.κ.

   - Τέλος, για κάθε αρχείο text που δημιουργήσατε πρέπει να αντικαταστήσετε τις λέξεις που περιέχουν οι προτάσεις με τις αντίστοιχες αλληλουχίες φωνημάτων. Για το λόγο αυτό σας δίνεται μαζί με τα υπόλοιπα δεδομένα το λεξικό (lexicon.txt), το οποίο αντιστοιχίζει κάθε λέξη της αγγλικής γλώσσας στην αλληλουχία φωνημάτων που της αντιστοιχεί. Προσέξτε σε αυτό το βήμα να μετατρέψετε αρχικά όλους το χαρακτήρες σε lower case, καθώς και να αφαιρέσετε τους ειδικούς χαρακτήρες (π.χ. τελείες, παύλες κτλ.) εκτός από τα single quotes (’). Επίσης, στην αρχή και στο τέλος πρέπει να προσθέσετε το φώνημα της σιωπής (sil). Δίνεται το παράδειγμα για την 1η πρόταση του ομιλητή f1:

     This was easy for us.

     Θα πρέπει να μετασχηματιστεί σε:

     sil dh ih s w ao z iy z iy f r er ah s sil