ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Επεξεργασία Φωνής & Φυσικής Γλώσσας
2η Εργαστηριακή Άσκηση: Αναγνώριση Φωνής με το KALDI TOOLKIT

Μία αναφορά των φοιτητών:

- Σερλή Εμμανουήλ – Αναστάσιου (Α.Μ. 03118125)
- Αβράμη Στέφανου (Α.Μ. 03121724)

1.Εισαγωγικά

Αναλυτικές οδηγίες για το πώς θα τρέξουν τα επιμέρους κομμάτια της εργαστηριακής άσκησης μπορείτε να βρείτε στο Read.md αρχείο που βρίσκεται εντός του zip.

2. Θεωρητικό Υπόβαθρο

- **MFCCs:** Τα MFCCs η Mel-Frequency Cepstral Coefficients αποτελούν ορισμένα εκ των πιο δημοφιλών χαρακτηριστικών στο πεδίο της αναγνώρισης φωνής. Συγκεκριμένα, η εξαγωγή τους βασίζεται στο Mel-Frequency Cepstrum, που αποτελεί αναπαράσταση short-term ενεργειακού φάσματος ενός ηχητικού σήματος. Πιο αναλυτικά, τα βήματα που ακολουθούνται για την εξαγωγή των εν λόγω χαρακτηριστικών είναι τα εξής:

Α) Εφαρμογή **pre-emphasis φίλτρου** για ενίσχυση των υψηλών συχνοτήτων. Στόχος του φιλτραρίσματος αυτού είναι η μείωση της διαφοράς πλάτους ανάμεσα σε υψηλές και χαμηλές συχνότητες καθώς και η αποφυγή προβλημάτων αριθμητικής αστάθειας στο βήμα υπολογισμού του μετασχηματισμού Fourier. Η εφαρμογή του pre-emphasis φίλτρου δίνεται από την κάτωθι σχέση

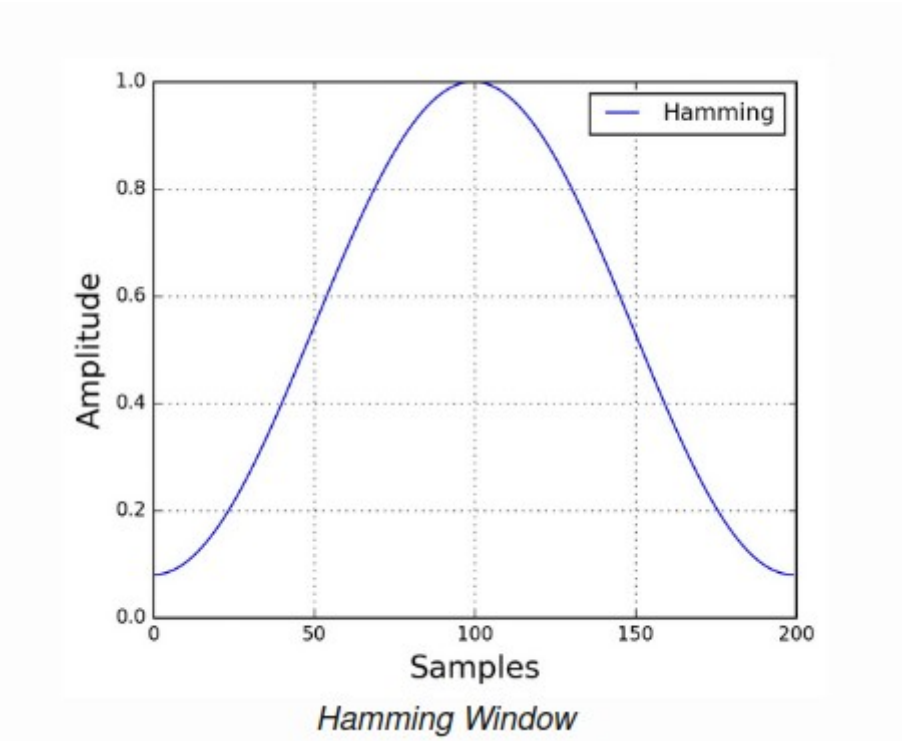
$$y(t) = x(t) – ax(t-1)$$

Μια εναλλακτική των pre-emphasis φίλτρων είναι η εφαρμογή κανονικοποίησης στο φωνητικό σήμα.

Β) Διαχωρισμό του σήματος σε επιμέρους βραχύχρονα **frames.** Ο λόγος για την ύπαρξη του εν λόγω βήματος έγκειται στο γεγονός ότι ένα πραγματικό φωνητικό σήμα έχει διαρκώς μεταβαλλόμενες συχνότητες για μεγάλα χρονικά διαστήματα. Συνεπώς, η εφαρμογή μετασχηματισμού Fourier έχει νόημα σε χρονικά διαστήματα αρκετά μικρά, όπου μπορεί να γίνει η παραδοχή ότι εντός του εν λόγω παραθύρου το συχνοτικό περιεχόμενο παραμένει σταθερό.

Τυπικές τιμές dt για τα εν λόγω frames κυμαίνονται από 20 εως 40 ms με 50% επικάλυψη μεταξύ διαδοχικών frames.

Γ) Εφαρμογή **συνάρτησης παραθύρου** σε καθένα από τα εν λόγω frames, με κύριο στόχο την μείωση της τιμής του φάσματος στα χρονικά άκρα του κάθε frame. Μια εκ των πιο δημοφιλών είναι η συνάρτηση Hamming η οποία αναπαρίσταται παρακάτω:

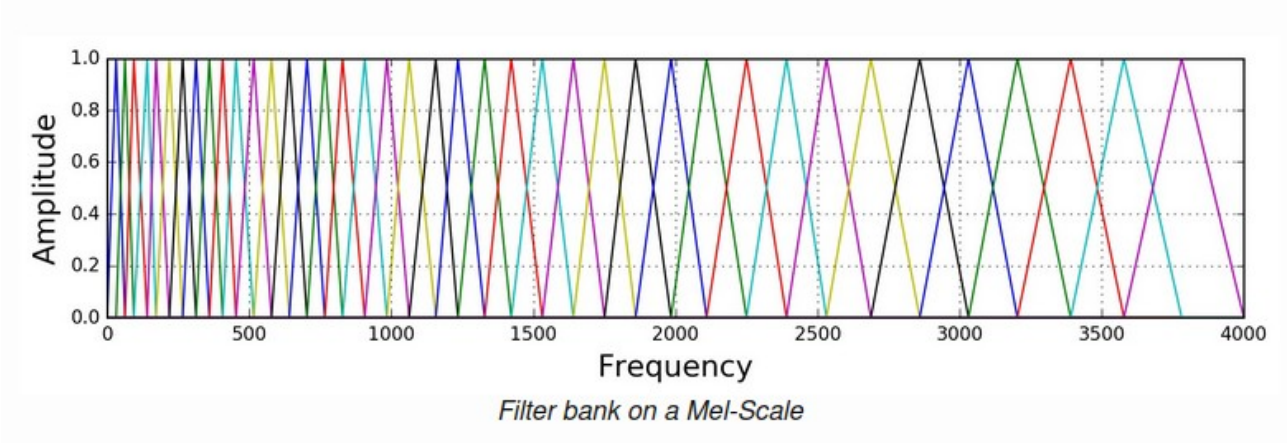


Hamming Window

Δ) Εφαρμογή του **Short-Time Fourier Transform (STFT)** σε καθένα από τα windowed frames και εξαγωγή του **περιοδογράμματος**, μέσω της σχέσης:

$$P = \frac{|FFT(x_i)|^2}{N}$$

Ε) Υπολογισμός των **filter banks** μέσω εφαρμογής επικαλυπτόμενων τριγωνικών φίλτρων, τα οποία έχουν τιμή 1 στην κεντρική συχνότητα f_0 και γραμμικώς μειούμενη τιμή πλάτους μέχρι τις συχνότητες f_0 - a και f_0 + a, όπως φαίνεται παρακάτω:



Filter bank on a Mel-Scale

Αξίζει να σημειωθεί ότι ο υπολογισμός των filter banks γίνεται στην κλίμακα συχνοτήτων Mel αντί για Hertz, μιας και η πρώτερη μοντελοποιεί την καλύτερη διακριτική ικανότητα του ανθρώπινου αυτιού στις χαμηλές συχνότητες η οποία μειώνεται στις υψηλότερες.

ΣΤ) Εφαρμογή του **γραμμικού μετασχηματισμού Discrete Cosine Transform (DCT)** με σκοπό την μείωση της συσχέτισης μεταξύ των filter banks. Αυτή η τελική συμπιεσμένη αναπαράστασή τους μας δίνει τα MFCCs, εκ των οποίων κρατάμε τους 2 με 13 πρώτους συντελεστές (μαζί με τις πρώτες και τις δεύτερες παραγώγους αυτών).

Η χρήση των MFCCs ήταν ευρεώς διαδεδομένη την εποχή που χρησιμοποιούνταν μοντέλα αναγνώρισης φωνής τύπου GMM – HMM, μιας και απαντούνταν πρώτα η αποσυσχέτιση των filter banks πρωτού αυτές μπουν ως είσοδοι στο μοντέλο. Ωστόσο, στην περίπτωση χρήσης Deep Learning νευρωνικών δικτύων η άνωθι απαίτηση ασυσχέτιστων εισόδων παύει να ισχύει, μιας και τα εν λόγω μοντέλα αποδίδουν καλά ακόμα και με τα αρχικά filter banks (χωρίς δηλαδή την εφαρμογή DCT).

- **<u>Γλωσσικά μοντέλα:</u>** Ως γλωσσικά ορίζονται τα μοντέλα τα οποία κωδικοποιούν ακολουθίες λέξεων μέσω πιθανοτικών κατανομών της μορφής $P(w_1,w_2,\ldots,w_n)$, οι οποίες προκύπτουν μέσα από την προπόνηση σε corpora κειμένων. Δύο εκ των πιο συνηθισμένων γλωσσικών μοντέλων είναι τα εξής:

A) **n-grams:** Μοντελοποιούν τις αλληλουχίες των λέξεων σαν αλυσίδες Markov, υποθέτωντας ότι η πιθανότητα εμφάνισης της επόμενης λέξης εξαρτάται μόνο από το σύνολο λέξεων που βρίσκονται σε ένα παράθυρο σταθερού μήκους n-1. Παραδείγματα τέτοιων μοντέλων είναι τα unigrams και τα bigrams με τιμές μήκους παραθύρου 0 και 1 αντίστοιχα. Στην σημερινή εποχή, τα n-grams έχουν πάψει να χρησιμοποιούνται για state-of-the-art language μοντέλα.

B)**Neural Networks:** Χρησιμοποιούν συνεχείς αναπαραστάσεις των λέξεων, οι οποίες ονομάζονται word embeddings. Στόχος τους είναι να προσεγγίσουν την απόδοση ενός γλωσσικού μοντέλου το οποίο έχει εκθετικά αυξημένο αριθμό από πιθανούς συνδιασμούς, οι οποίοι αναπαρίστανται εντός του δικτύου ως μη-γραμμικοί συνδιασμοί από βάρη.

Όσον αφορά την δομή τους χρησιμοποιούνται τόσο feedforward όσο και recurrent δίκτυα, ενώ ως έξοδο έχουν την πιθανότητα $P(w_t \mid \text{context})$, όπου το context μπορεί να είναι είτε η παρελθοντική ακολουθία λέξεων είτε ο συνδιασμός παρελθοντικής και μελλοντικής ακολουθίας (Bag-Of-Words). Μία ακόμα εναλλακτική είναι η προπόνηση νευρωνικών δικτύων ονόματι skip-gram models, τα οποία καλούνται να βρουν το context, έχοντας ως είσοδο την λέξη w_t

- **Φωνητικά μοντέλα:** Τα φωνητικά μοντέλα χρησιμοποιούνται στην αναγνώριση φωνής και καλούνται να αναπαραστήσουν την σχέση ανάμεσα σε ένα ακουστικό σήμα εισόδου και στα φωνήματα τα οποία αποτελούν το εν λόγω σήμα. Η προπόνηση τέτοιων μοντέλων γίνεται μέσω συνόλου δεδομένων που περιλαμβάνει τα σήματα ήχου μαζί με τα αντίστοιχα transcripts κειμένου. Τελική έξοδος του μοντέλου είναι η πιθανότητα υπάρξης φωνητικών χαρακτηριστικών για κάθε frame ακουστικής εισόδου.

  Μία συνηθισμένη μέθοδος δημιουργίας ακουστικών μοντέλων είναι η χρήση Hidden Markov Models, τα οποία κωδικοποιούν την πιθανότητα μετάβασης από το ένα φώνημα στο επόμενο και λαμβάνουν ως παρατηρήσεις acoustic features, όπως τα MFCCs που αναφέρθηκαν παραπάνω. Τελευταία, μεγάλες βελτιώσεις στα εν λόγω μοντέλα έχουν έρθει από την χρήση LSTMs και CNNs.

3. Βήματα Προπαρασκευής

Αρχικά, πραγματοποιήθηκε εγκατάσταση του kaldi και εξοικείωση μέσω των παραδειγμάτων της εκφώνησης. Στην συνέχεια, έγινε download του συνόλου δεδομένων που περιλαμβάνει τα αρχεία ήχου καθώς και τα κείμενα που αντιστοιχούν σε κάθε αρχείο ήχου, ενώ το εν λόγω σύνολο είναι χωρισμένο σε train, dev και test.

Το downloaded folder usc προστέθηκε στο directory kaldi/egs, όπου πραγματοποιήθηκε η υλοποίηση του αρχικού σκελετού (για καθένα από τους υποφακέλους data/train, data/dev, data/test) καθώς και των επόμενων βημάτων. Όσον αφορά τον αρχικό σκελετό, αρχικά έγινε η δημιουργία των ζητούμενων αρχείων uttids, utt2spk, wav.scp και text.

Τέλος, έγινε χρήση του του lexicon.txt, που περιλαμβάνει την ακολουθία φωνημάτων που αντιστοιχεί σε κάθε λέξη, έτσι ώστε να γίνει αντικατάσταση των λέξεων των text files σε ακολουθίες φωνημάτων. Έτσι, αφού πρώτα έγινε μετατροπή κάθε πρότασης σε lower case και αφαίρεση των αντίστοιχων ειδικών χαρακτήρων, δημιουργήθηκαν 2 αρχεία, το text με τις ζητούμενες ακολουθίες φωνημάτων και το text_init με τις αρχικές ακολουθίες λέξεων.

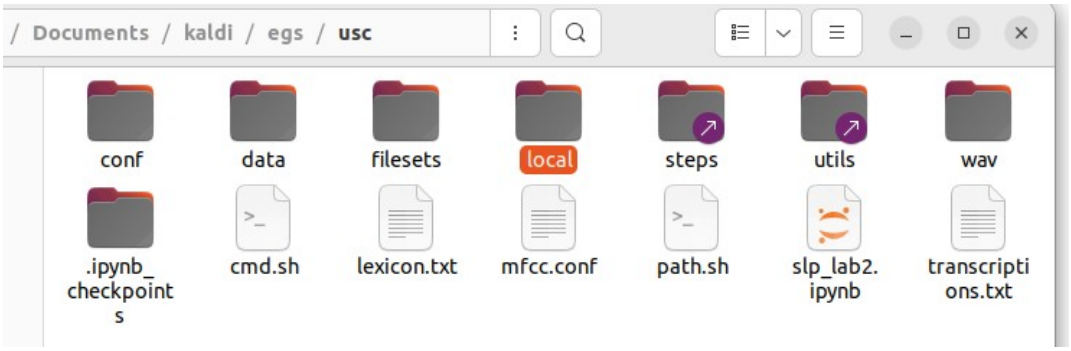
4. Βήματα Κυρίους Μέρους

4.1 Προετοιμασία διαδικασίας αναγνώρισης φωνής για τη USC-TIMIT

Στόχος του εν λόγω βήματος του κυρίως μέρους είναι η δημιουργία του κατάλληλου kaldi environment, με βάση το οποίο θα γίνει η διαδικασία εκπαίδευσης και αξιολόγησης των φωνητικών μοντέλων. Συνοπτικά, έγιναν τα εξής:

- Αντιγραφή των αρχείων wsj/path.sh και wsj/cmd.sh στο usc/ directory με τις κατάλληλες τροποποιήσεις.
- Δημιουργία των φακέλων conf, local, data/local/dict , data/local/lm_tmp και data/local/nist_lm.
- Δημιουργία soft links για τους φακέλους wsj/steps και wsj/utils
- Αντιγραφή του step/score_kaldi.sh εντός του φακέλου local και του δοθέντος mfcc.conf εντός του φακέλου conf.

Η δημιουργία των προαναφερθέωντων αρχείων και των συντομεύσεων φαίνεται παρακάτω:

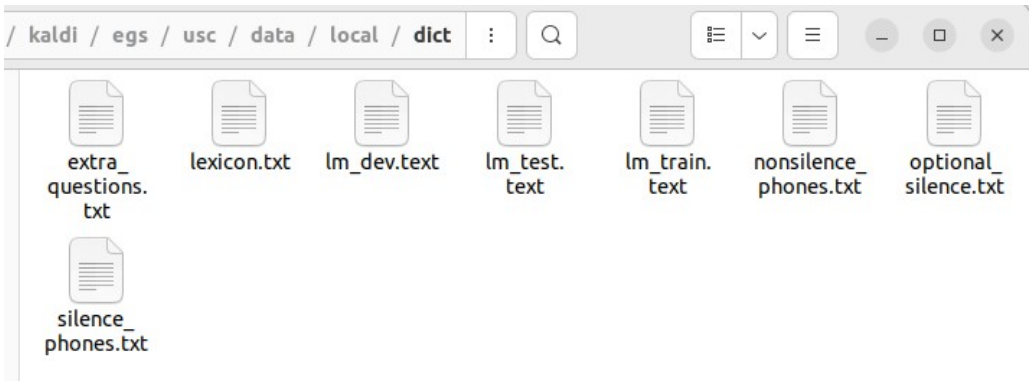


4.2 Προετοιμασία γλωσσικού μοντέλου

Αρχικά προετοιμάζουμε τον φάκελο dict με τα αρχεία τα οποία θα χρειαστούν για τη δημιουργία του γλωσσικού μοντέλου. Συγκεκριμένα, έγινε η δημιουργία των κάτωθι αρχείων:

- **silence_phones.txt, optional_silence.txt:** Περιλαμβάνει μόνο το silence phoneme (sil)
- **nonsilence_phones.txt:** Περιλαμβάνει όλα τα υπόλοιπα φωνήματα, τα οποία πάρθηκαν από το lexicon.txt (προστέθηκαν αφού πρώτα έγιναν sorted).
- **lexicon.txt:** Περιλαμβάνει την 1 -1 αντιστοίχιση κάθε φωνήματος με τον εαυτό του (συμπεριλαμβανομένου και του sil).
- **lm_train.text, lm_test.text, lm_dev.text:** Περιλαμβάνει κάθε πρόταση φωνημάτων (χωρίς τα utterance_ids) από το text αρχείο της προπαρασκευής, μαζί με τους χαρακτήρες <s> και </s> στην αρχή και στο τέλος κάθε ακολουθίας.
- **extra_questions.txt:** Αποτελεί κενό αρχείο

Η τελική δομή του dict folder φαίνεται παρακάτω:



Στην συνέχεια, εντός του φακέλου local/lm_tmp, έγινε η ενδιάμεση μορφή του γλωσσικού μοντέλου, μέσω της εντολής build-lm.sh που παρέχει το πακέτο IRSTLM. Έτσι, σχηματίστηκαν για κάθε set 2 μοντέλα, ένα unigram (n=1) και ένα bigram (n=2). Αντίστοιχα, εντός του φακέλου local/nist_tmp, τα άνωθι γλωσσικά μοντέλα έγιναν compiled μέσω της εντολής compile-lm σε μορφή arpa.

Μετ’έπειτα, από το αρχείο prepare_lang.sh έγινε ο σχηματισμός του αυτόματου του λεξικού (L.fst) εντός του φακέλου data/lang, μαζί με το sorting των αρχείων wav.scp, text και utt2spk από το βήμα της προπαρασκευής. Μέσω του βοηθητικού script utils/utt2spk_to_spk2utt.pl, δημιουργήθηκε το αρχείο spk2utt, ενώ μέσω του timid_format_data.sh έγιναν 2 αυτόματα γραμματικής για κάθε γλωσσικό μοντέλο, ονόματι G_ug.fst και G_bg.fst.

Ερώτημα 1:

Για να υπολογίσουμε το perplexity στο dev και στο test set χρησιμοποιούμε -όπως και παραπάνω - την εντολή compile-lm, προσθέτοντας και το –eval ως όρισμα. Τα αποτελέσματα για τους διαφορετικούς συνδιασμούς μοντέλων και testing δεδομένων αναγράφονται παρακάτω:

```
inpfile: data/local/lm_tmp/lm_phone_ug.ilm.gz
outfile: lm_phone_ug.ilm.blm
evalfile: data/local/dict/lm_test.text
loading up to the LM level 1000 (if any)
dub: 10000000
OOV code is 1362
OOV code is 1362
Start Eval
OOV code: 1362
%% Nw=13162 PP=54.56 PPwp=19.75 Nbo=0 Noov=367 OOV=2.79%
inpfile: data/local/lm_tmp/lm_phone_bg.ilm.gz
outfile: lm_phone_bg.ilm.blm
evalfile: data/local/dict/lm_test.text
loading up to the LM level 1000 (if any)
dub: 10000000
OOV code is 1362
OOV code is 1362
Start Eval
OOV code: 1362
%% Nw=13162 PP=27.65 PPwp=10.01 Nbo=606 Noov=367 OOV=2.79%
inpfile: data/local/lm_tmp/lm_phone_ug.ilm.gz
outfile: lm_phone_ug.ilm.blm
evalfile: data/local/dict/lm_dev.text
loading up to the LM level 1000 (if any)
dub: 10000000
OOV code is 1362
OOV code is 1362
Start Eval
OOV code: 1362
%% Nw=5078 PP=56.46 PPwp=21.16 Nbo=0 Noov=148 OOV=2.91%
inpfile: data/local/lm_tmp/lm_phone_bg.ilm.gz
outfile: lm_phone_bg.ilm.blm
evalfile: data/local/dict/lm_dev.text
loading up to the LM level 1000 (if any)
dub: 10000000
OOV code is 1362
OOV code is 1362
Start Eval
OOV code: 1362
%% Nw=5078 PP=28.54 PPwp=10.70 Nbo=199 Noov=148 OOV=2.91%
```

Όπως ήταν αναμενόμενο το bigram μοντέλο έχει πολύ μικρότερο perplexity(PP) τόσο στο development όσο και στο validation set, αφού χρησιμοποιεί ένα μοντέλο μήκους δύο λέξεων για να κάνει προβλέψεις και άρα μπορεί να χρησιμοποιήσει context για να κάνει καλύτερες προβλέψεις. Το perplexity εκφράζει πόσο καλά προβλέπει μία επόμενη λέξη το μοντέλο, δηλαδή όσο μικρότερο είναι, τόσο πιο confident είναι το μοντέλο για τις προβλέψεις που κάνει.

4.3 Εξαγωγή ακουστικών χαρακτηριστικών

Στο βήμα αυτό, πραγματοποιείται η δημιουργία των χαρακτηριστικών MFCCs για το ακουστικό μοντέλο του βήματος 4.4, για καθένα από τα 3 σετ δεδομένων, μέσω των βοηθητικών εντολών make_mfcc.sh και compute_cmvn_stats.sh. Αξίζει να σημειωθεί ότι αρχικά έγινε εκ νέου η εξαγωγή των αρχείων spk2utt, με παρόμοιο τρόπο με αυτόν του βήματος 4.2.

Ερώτημα 2:

Η διαδικασία του Cepstral Mean and Variance Normalization (CMVN), αποτελεί μια υπολογιστική μέθοδο μέσω της οποίας τα δεδομένα εισόδου (εν προκειμένω τα wav files) αποκτούν μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση. Σκοπός της εν λόγω μεθόδου είναι η ελαχιστοποίηση της παραμόρφωσης των χαρακτηριστικών λόγω θορύβου, μέσω γραμμικού μετασχηματισμού τους, βελτιστοποιώντας έτσι της διαδικασία εξαγωγής τους. Η εν λόγω μέθοδος είναι ιδιαίτερα χρήσιμη σε περιβάλλοντα με μεταβαλλόμενα ακουστικά χαρακτηριστικά.

Μαθηματική εξήγηση:

Έστω ότι έχουμε ένα σήμα εισόδου x(t) και τον συνοδευόμενο θόρυβο n(t), τα οποία δίνουν στη έξοδο ένα σήμα της μορφής y(t) = x(t) + n(t). Έστω ότι τροποποιούμε το σήμα x(t) έτσι ώστε να έχει την μορφή x'(t) = [x(t) – mean(x(t))] / std(x(t)) που αντιστοιχεί σε σήμα εξόδου y'(t) = x'(t) + n(t) = [x(t) – mean(x(t))] / std(x(t)) + n(t). Τότε, η διαφορά μεταξύ των 2 σημάτων εξόδου είναι

y(t) – y'(t) = x(t) + n(t) – [x'(t) + n(t)] = x(t) – x'(t), δηλαδή **ανεξάρτητη του αρχικού θορύβου**.

Ερώτημα 3:

Μέσω των εντολών feat-to-dim και feat-to-length που παρέχει η KALDI, εξήχθη το μήκος των frames για καθεμία από τις 5 πρώτες προτάσεις του training set:

```
feat-to-dim ark:data/train/data/raw_mfcc_train.1.ark -
13
feat-to-len scp:data/train/feats.scp ark,t:data/train/feats.lengths
f1_003 317
f1_004 371
f1_005 399
f1_007 328
f1_008 464
```

Η διάσταση των features ταυτίζεται με το πλήθος των MFCCs που κρατάμε, τα οποία είναι 13.

4.4 Εκπαίδευση ακουστικών μοντέλων και αποκωδικοποίηση προτάσεων

Αρχικά, πραγματοποιήθηκε η προπόνηση ενός GMM-HMM μοντέλου πάνω στα train δεδομένα. Στην συνέχεια, για καθέναν από τους 2 γράφους γραμματικής (G_ug.fst και G_bg.fst), πραγματοποιήθηκε ο HCLG γράφος καθώς και η αποκωδικοποίηση των προτάσεων μέσω του αλγόριθμου Viterbi. Τέλος, έγινε χρήση του script local/score.sh για την εξαγωγή του Phone Error Rate(PER), με τα αποτελέσματα να ακολουθούν στον κάτωθι πίνακα:

	Unigram_mono	Bigram_mono
Dev	52.15%	51.90%
Test	45.46%	**44.96%**

Παρατηρούμε ότι τα bigram μοντέλα έχουν κατά λίγο χαμηλότερο PER σε σχέση με τα unigram μοντέλα, γεγονός που εκ νέου επαληθεύει την ανωτερότητά τους σε επίπεδο απόδοσης. Επιπλέον, αξίζει να τονισθεί, ότι οι υπερπαράμετροι που προκύπτουν από την score.sh είναι το ελάχιστο και το μέγιστο LM-weight για το lattice resourcing, όπου στην περίπτωση του bigram μοντέλου, πήραν τιμή 7 και 19 αντίστοιχα.

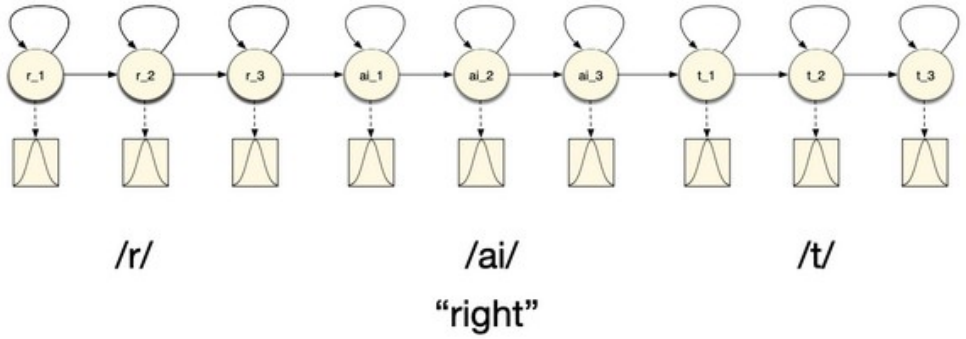
Τέλος, κάνοντας alignment των φωνημάτων μέσω του μονοφωνικου μοντέλου (steps/align_si.sh), έγινε training ενός τριφωνικού μοντέλου, το οποίο στην συνέχεια χρησιμοποιήθηκε – ομοίως με άνωθι – για τον σχηματισμό HCLG γράφων για καθένα από τα unigram και bigram μοντέλα και για την κωδικοποίησή τους. Παρακάτω ακολουθούν τα PER scores για όλους τους συνδιασμούς μοντέλων και συνόλων δεδομένων.

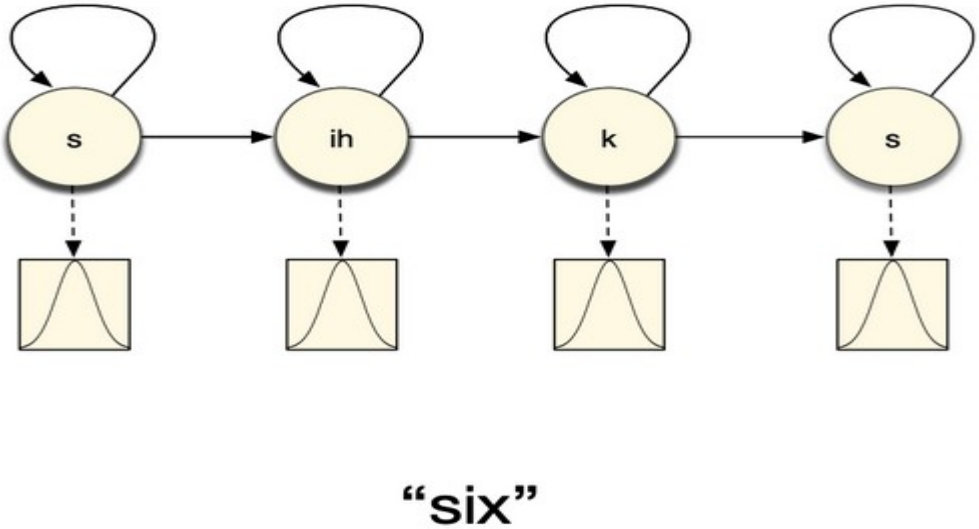
	Unigram_tri	Bigram_tri
Dev	39.27%	35.86%
Test	38.54%	**34.78%**

Παρατηρούμε ότι τα νέα τριφωνικά μοντέλα αποδίδουν αισθητά καλύτερα από τα αντίστοιχα μονοφωνικά, με το καλύτερο αποτέλεσμα να προκύπτει εκ νέου από τον συνδιασμό bigram με test.

Ερώτημα 4:

Ο συνδιασμός GMM-HMM αποτελεί μία ιδιαίτερα συχνή κατηγορία ακουστικών μοντέλων, σκοπός των οποίων είναι να αντιστοιχίσουν την ηχητική πληροφορία σε ακολουθία φωνημάτων. Όσον αφορά το κομμάτι του HMM, αυτό βασίζεται σε ένα αυτόματο πεπερασμένων καταστάσεων, στο οποίο κάθε φώνημα αναπαρίσταται από μία ή περισσότερες κρυμμένες καταστάσεις, όπως φαίνεται παρακάτω:





Στα άνωθι σχήματα, το μείγμα των γκαουσιανών εκπαιδεύεται, ώστε να προσφέρει ως παρατήρησεις στο HMM την πιθανότητα ενός φωνήματος να βρίσκεται σε μία λέξη και χρησιμοποιείται ως ένα σύνολο παρατηρήσεων, που επηρεάζει τον πίνακα μετάβασης από τις προηγούμενες καταστάσεις του HMM στην επόμενη.

Η εκπαίδευση των GMM-HMM models γίνεται μέσω του αλγορίθμου Viterbi,ο οποίος μεγιστοποιεί την posterior πιθανότητα της πιο πιθανής αλληλουχίας κρυφών καταστάσεων. Η εκπαίδευση αυτού του μοντέλου γίνεται επαναληπτικά πάνω σε ένα train set το οποίο βοηθά το GMM με βάσει τα αποτελέσματα του HMM για τις νέες εισόδους που δέχεται να κάνει ομαδοποίηση σε νέες κλάσεις φωνημάτων.

Ερώτημα 5:

Με βάση τον τύπο του Bayes, λαμβάνουμε την κάτωθι posterior πιθανότητα για την εμφάνιση ενός φωνήματος:

$$\mathbf{P(W|X) = P(X|W)*P(W) / P(X)}, \text{ όπου:}$$

- W, X: φώνημα και διάνυσμα χαρακτηριστικών αντίστοιχα
- P(W|X): posterior πιθανότητα εμφάνισης του φωνήματος, με δεδομένο το διάνυσμα χαρακτηριστικών X (ζητείται)
- P(X|W): η πιθανότητα εμφάνισης του X, με δεδομένη την ύπαρξη του φωνήματος W (υπολογίζεται από το εκάστοτε ακουστικό μοντέλο)
- P(W): πιθανότητα εμφάνισης φωνήματος
- P(X): πιθανότητα εμφάνισης διανύσματος χαρακτηριστικών (αγνοείται στον υπολογισμό μιας και είναι ίδιο για τις posterior όλων των διαφορετικών φωνημάτων)

Για τον υπολογισμό του πιο πιθανού φωνήματος, αρκεί να κρατήσουμε την μέγιστη $P(W_i | X)$ ή αλλιώς:

$$\mathbf{W = argmax_{1<i<n} \{P(W_i | X) \}}$$

Ερώτημα 6:

Ο γράφος HCLG είναι γράφος-αποκωδικοποιητής της KALDI και αποτελείται από 4 επιμέρους κομμάτια:

- H → Μετάβαση από HMM μεταβάσεις σε context-depedent labels
- C → Μετάβαση από context-depedent labels σε φωνήματα
- L → Γλωσσικό μοντέλο για αντιστοίχιση φωνημάτων σε λέξεις (ή πάλι σε φωνήματα στην περίπτωσή μας)
- G → Αποδοχέας γραμματικής