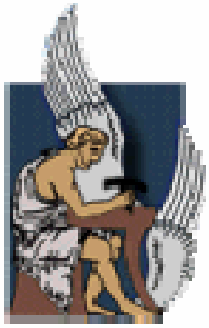
ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ



Τμήμα Ηλεκτρονικών Μηχανικών
και Μηχανικών Υπολογιστών

<u>Διπλωματική εργασία</u>

Αυτόματη απομαγνητοφώνηση ακουστικών σημάτων ηχογραφημένα από τηλεοπτικές εκπομπές

ΤΣΕΡΓΟΥΛΑΣ ΟΡΦΕΑΣ

Εξεταστική επιτροπή:

Διγαλάκης Βασίλης (επιβλέπων), Καθηγητής
Ποταμιάνος Αλέξανδρος, Καθηγητής
Αθανάσιος Λιάβας, Καθηγητής

ΠΕΡΙΕΧΟΜΕΝΑ

Ευχαριστίες

Πρώτα απ’ όλα θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κύριο Βασίλειο Διγαλάκη, που μου εμπιστεύθηκε το θέμα. Η βοήθεια και συμβολή του στάθηκαν πολύτιμες για την ολοκλήρωση αυτής της εργασίας. Επιπλέον μου έδωσε την ευκαιρία να αποκτήσω σημαντικές εμπείριες όλο αυτό το καιρό που συνεργάστηκα μαζί του.

Επίσης θα ήθελα να ευχαριστήσω τους καθηγητές κύριο Ποταμιάνο Αλέξανδρο και κύριο Λιάβα Αθανάσιο για τις υποδείξεις τους από την ανάγνωση της διπλωματικής.

Θερμές ευχαριστίες στον κύριο Δημήτρη Οικονομίδη και Βασίλη Διακολουκά γιατί απλά ήταν πάντα δίπλα μου σε οποιαδήποτε απορία, όσο ανόητη και αν ήταν. Επιπλέον θέλω να ευχαριστήσω και όλα τα παιδιά του εργαστηρίου τηλεπικοινωνιών για την κατανόηση και αγαστή συνεργασία όλη αυτή την περίοδο.

Τέλος, ευχαριστώ την οικογένεια μου για την ηθική και υλική συμπαράσταση από την στιγμή της εισαγωγής μου στο τμήμα μέχρι και την ολοκλήρωση των σπουδών μου.

ΕΙΣΑΓΩΓΗ

Γενικά

Η αναγνώριση φωνής έχει συγκεντρώσει μεγάλο ενδιαφέρον από την προηγούμενη δεκαετία και έπειτα, λόγω των ποικίλων εφαρμογών τους. Οι απαιτήσεις σε ακρίβεια απαιτούν γρήγορες και ακριβείς μεθόδους που να εξασφαλίζουν σταθερότητα σε μεταβαλλόμενες συνθήκες , π.χ. ομιλητή.

Αναγνώριση φωνής είναι η διαδικασία μετατροπής ενός ακουστικού ή φωνητικού σήματος που μπορεί να ληφθεί διάμεσου ποικίλων τρόπων (μικρόφωνο, τηλεφωνικής γραμμή, ραδιοτηλεοπτικές εκπομπές), σε μια ακολουθία λέξεων μέσω ενός αλγορίθμου. Οι αναγνωρισμένες λέξεις μπορούν να είναι τα τελικά αποτελέσματα μιας εφαρμογής, όπως εντολές για έλεγχο ή εισαγωγή δεδομένων. Μπορούν επίσης να χρησιμοποιηθούν και ως είσοδος για μετέπειτα επεξεργασία, προκειμένου να επιτευχθεί κατανόηση. Η πιο επιτυχημένη προσέγγιση στην αναγνώριση ομιλίας βασίζεται στην τεχνολογία της στατιστικής αναγνώρισης προτύπων (statistical pattern recognition), την οποία θα περιγράψουμε αναλυτικά στο επόμενο κεφάλαιο.

Περιγράφοντας συνοπτικά την παραπάνω αναφερθείσα τεχνολογία, το σύστημα κατασκευάζει ένα δίκτυο που υλοποιεί την γραμματική και για κάθε επιτρεπόμενη πρόταση αντιστοιχίζεται ένα σύνολο από μοντέλα Hidden Markov Models (HMMs). Όταν νέα δεδομένα φωνής πρόκειται να αναγνωριστούν, το σύστημα υπολογίζει τις πιθανότητες τα δεδομένα αυτά να είχαν παραχθεί με βάση καθένα από τα αποθηκευμένα HMMs. Το αποτέλεσμα της αναγνώρισης είναι η πρόταση με την μεγαλύτερη πιθανότητα. Η δομή των HMMs, καθώς και οι αλγόριθμοι εκπαίδευσης που έχουν αναπτυχθεί για τον καθορισμό των παραμέτρων αναγνώρισης, παρέχουν αρκετά υψηλές επιδόσεις σε εφαρμογές που λειτουργούν ανεξάρτητα από τον εκάστοτε ομιλητή, εφαρμογές συνεχούς ομιλίας και μεγάλων λεξιλογίων.

Η περιοχή της επεξεργασίας φωνής περιλαμβάνει περιοχές όπως: αναγνώριση, κωδικοποίηση, σύνθεση και τέλος εξακρίβωση ομιλητή. Οι παραπάνω εφαρμογές είναι πολύ δημοφιλείς και χρησιμοποιούνται καθημερινά από εκατομμύρια χρήστες. Τα συστήματα αναγνώρισης φωνής, ανάλογα με τις δυνατότητες τους, μπορούν να κατηγοριοποιηθούν σε **συστήματα απομονωμένων λέξεων** (isolated-word speech recognition systems) που απαιτούν από τον ομιλητή να σταματάει την ομιλία αρκετά συχνά ανάμεσα σε κάθε λέξη, σε **συστήματα συνδεδεμένων λέξεων** και σε **συστήματα συνεχούς ομιλίας** (continuous speech recognition systems), όπου η ομιλία μπορεί να είναι συνεχής και αδιάκοπη. Επίσης υπάρχουν και τα **συστήματα αυθόρμητης ομιλίας**. Υπάρχουν διάφοροι ορισμοί για την έννοια της αυθόρμητης ομιλίας. Γενικά μπορούμε να πούμε πως η αυθόρμητη ομιλία είναι η ομιλία που ακούγεται φυσική και αβίαστη, χωρίς να έχει ετοιμαστεί από πριν. Ένα τέτοιο σύστημα θα πρέπει να είναι αρκετά ευέλικτο και θα πρέπει να αναγνωρίζει ειδικές λέξεις/φράσεις όπως λέξεις κολλημένες μεταξύ τους , δισταγμούς (εμμμμ…), κλπ.

Τα συστήματα απομονωμένων λέξεων είναι περιοριστικοί αναγνωριστές αλλά όμως μπορούν να λειτουργήσουν ικανοποιητικά σε μια μεγάλη ποικιλία εφαρμογών. Τα συστήματα συνδεμένων λέξεων είναι λιγότερο περιοριστικά και έχουν καλές επιδόσεις για μια σειρά από ενδιαφέρουσες εφαρμογές. Τέλος οι αναγνωριστές συνεχούς ομιλίας είναι ελάχιστα περιοριστικοί και απαιτητικοί από τον χρήστη. Με τον χρόνο η επίδοση τους βελτιώνεται και θα μπορούν να χρησιμοποιηθούν σε ιδιαίτερα απαιτητικές εφαρμογές.

Τα συστήματα αναγνώρισης ομιλίας μπορούν να αναπτυχθούν με δεδομένα εκπαίδευσης που είναι είτε ενός ομιλητή (speaker-dependent) ή έχουν συλλεγεί από πλήθος ομιλητών (speaker-independent). Η διαφορά αυτών των δύο ειδών συστημάτων έγκειται στο αν τα λεκτικά πρότυπα κατασκευάζονται με ανάλυση των δεδομένων φωνής ή με επεξεργασία δεδομένων που προέρχονται από ένα ανεξάρτητο και αντιπροσωπευτικό δείγμα ομιλητών. Ειδικά τα συστήματα εξαρτημένα από ομιλητή ,που απαιτούν σαφώς μικρότερο όγκο δεδομένων από τα συστήματα ανεξάρτητα από τον ομιλητή, μπορούν να αναγνωρίσουν ομιλία , έχοντας ένα μεγάλο λεξιλόγιο, με μεγάλη ακρίβεια. Δηλαδή μπορούν να

επιτύχουν αναγνώριση ομιλίας με ακρίβεια 98% , 99% (που σημαίνει από τις 100 λέξεις που είχαν να αναγνωρίσουν μόνο 1, 2 αναγνωρίσθηκαν λανθασμένα).

Σήμερα για να μπορεί να χαρακτηριστεί μια εφαρμογή αναγνώρισης ως εμπορική πρέπει το ποσοστό σφάλματος αναγνώρισης φυσικής γλώσσας (natural language recognition error rate) να είναι κάτω από 5%. Δηλαδή στις 100 προτάσεις που απευθύνεται ο χρήστης προς το σύστημα, αυτό πρέπει να αναγνωρίζει και να εξάγεται το σωστό νόημα από μια πρόταση στο 95% τουλάχιστον των περιπτώσεων.

Το θέμα της διπλωματικής

Όπως ήδη αναφέραμε, ο τομέας της επεξεργασίας φωνής και ιδιαίτερα της αναγνώρισης φωνής έχει γνωρίσει ραγδαία εξέλιξη τα τελευταία χρόνια. Μειώνουν συνεχώς το υπολογιστικό κόστος υλοποίησης των αλγορίθμων επεξεργασίας φωνής και παρέχουν νέες δυνατότητες και πιο γρήγορες υπηρεσίες .Η παρούσα διπλωματική εργασία ασχολείται με την υλοποίηση ενός συστήματος αναγνώρισης ηχητικών σημάτων από τηλεοπτικές εκπομπές. Αυτό το σύστημα μπορεί να χαρακτηριστεί ως σύστημα συνεχούς ομιλίας (continuous speech recognition systems) και ανήκει στα συστήματα που είναι ανεξάρτητα από τον ομιλητή (speaker-independent).

Αρχικά συγκεντρώσαμε αρκετές ώρες οπτικοακουστικών δεδομένων από τηλεοπτικές ειδήσεις 2 διαφορετικών καναλιών, και από αυτά τα δεδομένα απομονώσαμε το ακουστικό κομμάτι (audio stream) έτσι ώστε να το επεξεργαστούμε κατάλληλα. Μετέπειτα, με την βοήθεια ενός προγράμματος/εργαλείου ονόματι Transcriber Tool, πετύχαμε την απομαγνητοφώνηση των δεδομένων και το κόψιμο των μονόωρων δελτίων ειδήσεων σε μικρά αρχεία ήχου, που αντιστοιχούν σε προτάσεις λίγων δευτερολέπτων. Αυτές οι προτάσεις αποτελούν τη βάση δεδομένων που χρησιμοποιήσαμε για την υλοποίηση και κατασκευή του συστήματος-

αναγνωριστή μας. Θα αναφερθούμε διεξοδικά σε επόμενο κεφάλαιο τι δεδομένα περιέχει η βάση μας, από ποιες κατηγορίες αποτελείται, και πως ακριβώς γίνεται η χρήση της.

Στη συνέχεια υλοποιήσαμε, τα δύο κύρια μέρη του συστήματος μας : το γλωσσικό μας μοντέλο (language model) και το ακουστικό μοντέλο (acoustic model). Σε αυτό το σημείο πρέπει να σημειωθεί πως δημιουργήσαμε αρκετά διαφορετικά ακουστικά μοντέλα χάριν πειραματισμού, επιδιώκοντας έτσι όσο το δυνατόν καλύτερα αποτελέσματα αναγνώρισης. Επιπρόσθετα ανάμεσα σε αυτά τα ακουστικά μοντέλα, είναι και κάποια που χρησιμοποιήθηκαν έχοντας σαν βάση τους παλιότερα υλοποιημένα ακουστικά μοντέλα (seed model). Αυτά τα μοντέλα, προσαρμοσθήκαν και εκπαιδεύτηκαν χρησιμοποιώντας δεδομένα από την βάση μας, δημιουργώντας μια σειρά από καινούργια ακουστικά μοντέλα που είναι ικανά να αναγνωρίσουν με ικανοποιητική επιτυχία και ακρίβεια ακουστικά σήματα από τηλεοπτικές ειδήσεις. Αυτά τα καινούργια μοντέλα χαρακτηρίζονται ως προσαρμοσμένα. Όλη αυτή η διαδικασία ονομάζεται προσαρμογή (adaptation).

Η συνεισφορά αυτής της διπλωματικής εργασίας είναι ιδιαίτερα μεγάλη, αφού μπορεί να αναγνωρίσει με αρκετή επιτυχία και ακρίβεια ακουστικά σήματα από ηχογραφημένες τηλεοπτικές ειδήσεις. Πιο συγκεκριμένα, όλο το σύστημα που υλοποιήθηκε παράγει αυτόματα ως έξοδο του απομαγνητοφωνήσεις τηλεοπτικών δελτίων. Η χρησιμότητα αυτών των απομαγνητοφωνήσεων ποικίλει, καθώς μπορούν να χρησιμοποιηθούν για επιστημονική και ερευνητική εφαρμογή.

Οργάνωση της διπλωματικής

Η ύλη που παρουσιάζεται σε αυτήν την διπλωματική εργασία έχει ως εξής:

- **Κεφάλαιο 1** με τίτλο : ‘’Εισαγωγή στην Αναγνώριση Ομιλίας’’ , όπου περιγράφονται η αναγνώριση με στατιστικές μεθόδους καθώς και η δομή των Hidden Markov Models (HMMs) .

- **Κεφάλαιο 2** με τίτλο : ‘’Περιγραφή της Βάσης Δεδομένων των Ακουστικών Σημάτων ’’ , όπου δίνονται αναλυτικά η χρονική διάρκεια των σημάτων που συλλέχθηκαν, οι κατηγορίες δεδομένων που σχηματίστηκαν , καθώς και ο τρόπος επεξεργασίας τους.

- **Κεφάλαιο 3** με τίτλο : ‘’Γλωσσικό Μοντέλο ’’, όπου δίνεται η περιγραφή του γλωσσικού μοντέλου (language model) του συστήματος μας. Παρουσιάζονται βασικές αρχές σχεδιασμού και υλοποίησης ενός γλωσσικού μοντέλου, και αναλυτική περιγραφή του σχεδιασμού του γλωσσικού μοντέλου του συστήματος μας.

- **Κεφάλαιο 4** με τίτλο : ‘’Ακουστικό Μοντέλο’’ , στο οποίο αρχικά παραθέτουμε τα βήματα σχεδίασης και τον τρόπο εκπαίδευσης των δεδομένων με στόχο την υλοποίηση ενός ακουστικού μοντέλου. Αργότερα, παρουσιάζουμε τα διάφορα ακουστικά μοντέλα που δημιουργήθηκαν.

- **Κεφάλαιο 5** με τίτλο : ‘’Αποτελέσματα Αναγνώρισης’’ , όπου παραθέτουμε τα διάφορα αποτελέσματα αναγνώρισης του συστήματος, ανάλογα με το ακουστικό μοντέλο που χρησιμοποιούμε για το σύστημα μας κάθε φορά. Επίσης γίνεται σχολιασμός των εκάστοτε αποτελεσμάτων.

- **Κεφάλαιο 6** με τίτλο : ‘‘Ανακεφαλαίωση και Μελλοντικές Επεκτάσεις’’ , όπου γίνεται ανακεφαλαίωση των προδιαγραφών του συστήματος αναγνώρισης που παρουσιάστηκε στην εργασία, καθώς και αναφορά σε θέματα σχετικά με την εξέλιξη και τις μελλοντικές επεκτάσεις του.

- **Παράρτημα Α** με τίτλο : ‘’HTK (Hidden Markov Model Toolkit) ‘’, στο οποίο γίνεται λεπτομερής παρουσίαση του συγκεκριμένου εργαλείου, το οποίο χρησιμοποιείται για έρευνα πάνω στην αναγνώριση ομιλίας.

- **Παράρτημα Β** με τίτλο : ‘’ Transcriber Tool ’’, όπου γίνεται περιγραφή του συγκεκριμένου εργαλείου το οποίο χρησιμοποιείται για την απομαγνητοφώνηση των ακουστικών σημάτων.

- **Βιβλιογραφία** όπου δίνουμε τις παραπομπές στα επιστημονικά άρθρα που χρησιμοποιήσαμε κατά την ανάπτυξη της εφαρμογής.

Είναι σχεδόν αδύνατο να προβλέψει κανείς την πρόοδο σε ένα επιστημονικό πεδίο. Ωστόσο, αν κρίνουμε από την εξέλιξη στην τελευταία δεκαετία , μοιάζει λογικό να μπορούμε να κάνουμε ορισμένες προβλέψεις για το κοντινό μέλλον στο χώρο της επεξεργασίας και αναγνώρισης φωνής.

Από την προηγούμενη κιόλας δεκαετία, χαρακτηριστικά όπως αυτά της αναγνώρισης συνεχούς ομιλίας και της ραγδαίας αύξησης των υποστηριζόμενων λέξεων σε τάξεις δεκάδων χιλιάδων αποτελούν ήδη γεγονός. Είναι παραπάνω από προφανές ότι μελλοντικά θα επιτευχθεί η ολοκλήρωση της επεξεργασίας φωνής με την επεξεργασία εικόνας, δεδομένων και ασύρματης μετάδοσης με αποτέλεσμα τα συστήματα να παρέχουν παγκόσμια πρόσβαση σε όλους τους χρήστες, οπουδήποτε ,

οποτεδήποτε και με λογικό κόστος, με απώτερο στόχο τη βελτίωση της ποιότητας ζωής. Σήμερα ο σημαντικότερος αποτρεπτικός παράγοντας για την διάδοση εφαρμογών αναγνώρισης φωνής παραμένει το υψηλό κόστος ανάπτυξης και το γεγονός ότι η επιδοση δεν είναι ακόμα ικανοποιητική για πολλές εφαρμογές. Είναι πάντως βέβαιο, ότι η πρόοδος στο χώρο είναι ραγδαία, το κόστος ανάπτυξης έχει μειωθεί δραστικά και η βελτίωση των εργαλείων καθιστά την διάδοση τέτοιων εφαρμογών εξασφαλισμένη, οπότε η εμπορική αποδοχή μπορεί να θεωρείται δεδομένη. Οι εφαρμογές , λοιπόν βελτιώνονται με ραγδαίους ρυθμούς, ωστόσο είναι βέβαιο ότι μετά από λίγα χρόνια αναμένεται τα συστήματα να αποκτήσουν μια ασυμπτωτική μορφή στην πρόοδο τους και τα βήματα, ειδικά στην αύξηση της επίδοσης αναγνώρισης στο μέλλον , να είναι μικρότερα από ότι βλέπουμε σήμερα.

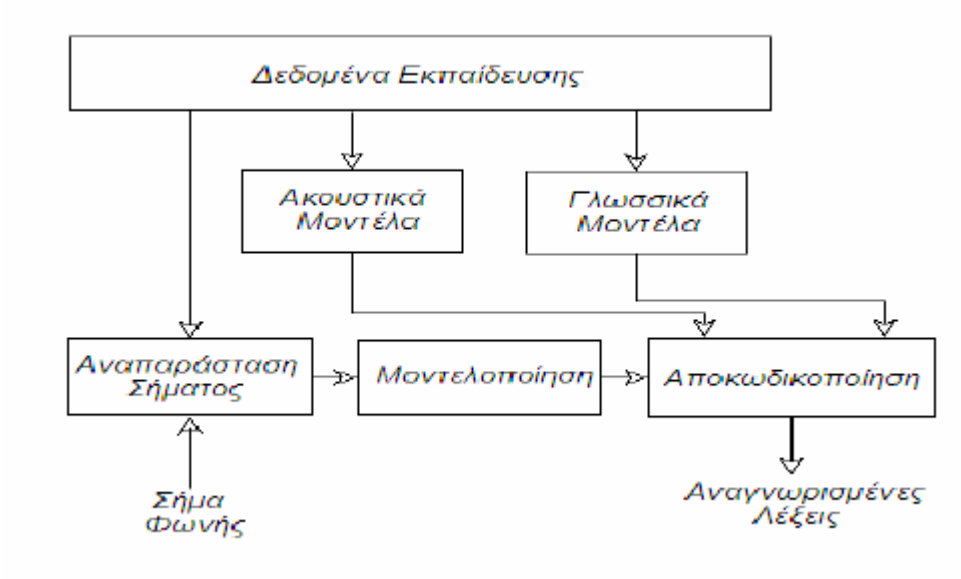
ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΑΝΑΓΝΩΡΙΣΗ ΟΜΙΛΙΑΣ

1.1 Εισαγωγή

Μέχρι πρόσφατα, τα συστήματα με μεγάλο αριθμό λέξεων δεν απευθυνόταν σε ανεξάρτητους ομιλητές και οι λέξεις έπρεπε να δίνονται με κάποια μικρή παύση ανάμεσα τους. Τα σύγχρονα συστήματα βασίζονται σε στατιστικές μεθόδους αναγνώρισης. Το σύστημα χρησιμοποιεί ένα σύνολο προτύπων λέξεων ή φράσεων που δημιουργούνται από ένα πρόγραμμα εκπαίδευσης προτύπων, βασισμένο στην γραμματική αναγνώρισης και για κάθε επιτρεπόμενη πρόταση αντιστοιχίζεται ένα σύνολο από μοντέλα HMMs. Αυτά τα πρότυπα μπορούν να είναι τυπικά φάσματα προτύπων λέξεων, μέσες τιμές προτύπων φασμάτων προτύπων λέξεων διαμέσου διαφορετικών ομιλητών ή εξελιγμένα στατιστικά μοντέλα. Τα μοντέλα αυτά περιλαμβάνουν στατιστικούς μέσους όρους και φασματική μεταβλητότητα που εξαρτάται από την χρονική διάρκεια της λέξης. Όταν νέα δεδομένα φωνής πρόκειται να αναγνωριστούν, το σύστημα υπολογίζει τις πιθανότητες τα δεδομένα αυτά να είχαν παραχθεί με βάση καθένα από τα αποθηκευμένα HMMs. Το αποτέλεσμα της αναγνώρισης είναι η πρόταση με την μεγαλύτερη πιθανότητα.

Η δομή των HMMs, καθώς και οι αλγόριθμοι εκπαίδευσης που έχουν αναπτυχθεί για τον καθορισμό των παραμέτρων αναγνώρισης, παρέχουν υψηλές επιδόσεις σε εφαρμογές που λειτουργούν για οποιοδήποτε ομιλητή χωρίς εκπαίδευση, συνεχούς ομιλίας και μεγάλων λεξιλογίων.



Σχήμα 1.1. Διάγραμμα αναγνώρισης

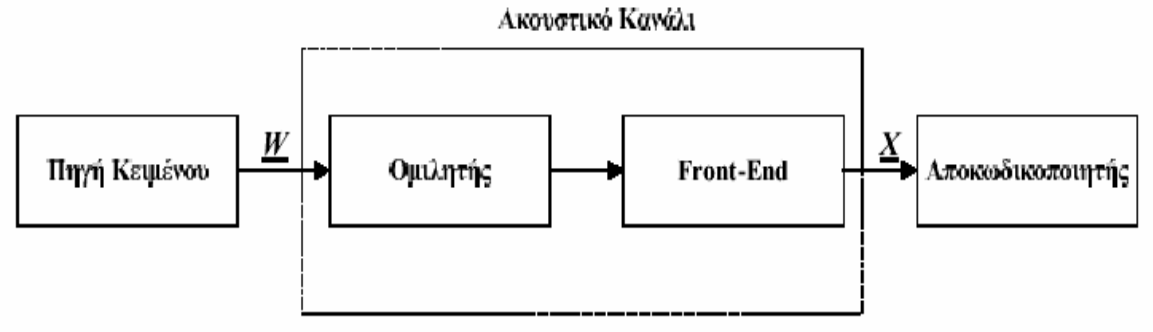
Το σύστημα του Σχήματος 1.1 μπορεί να εφαρμοστεί σε μια ευρύτατη ομάδα προβλημάτων που περιλαμβάνει αναγνώριση απομονωμένων λέξεων ή φράσεων, αναγνώριση συνδεδεμένων λέξεων, ακόμη και αναγνώριση συνεχούς ομιλίας. Παρά την αυξημένη πολυπλοκότητα τέτοιων μεθόδων , το βασικό μοντέλο αναγνώρισης προτύπων είναι η βάση σχεδόν όλων των μεθόδων που χρησιμοποιούνται σήμερα.

1.2 Η Front –End επεξεργασία

Τυπικά η αναγνώριση ξεκινά με το ψηφιακοποιημένο σήμα ομιλίας, το οποίο υπόκειται κατόπιν σε προεπεξεργασία (front-end), μέσα από ποικίλα βήματα, φασματικής συνήθως, επεξεργασίας σήματος. Κάποιες από τις πιο διαδεδομένες μεθόδους είναι οι :

- **Ανάλυση Γραμμικής Πρόβλεψης** (Linear Prediction Analysis - LPC)
- **Εξαγωγή Mel-Frequency Cepstral Coefficients** (MFFC)
- **Μοντελοποίηση κοχλία**

Για κάθε τμήμα ομιλίας (frame) έχουμε εξαγωγή διανυσματικών ακολουθιών. Θεωρούμε ότι μια άγνωστη κυματομορφή σήματος φωνής μετατρέπεται από έναν front-end επεξεργαστή σε μια ακολουθία από ακουστικά διανύσματα. Έτσι για κάθε τμήμα ομιλίας έχουμε εξαγωγή διανυσματικών ακολουθιών. Αξίζει να σημειωθεί πως μερικά συστήματα παράγουν πολλαπλές παράλληλες διανυσματικές ακολουθίες. Καθένα από τα διανύσματα αυτά είναι μια συμπαγής αναπαράσταση του φάσματος στον χρόνο καλύπτοντας τυπικά μια περίοδο 10msec. Έτσι μια έκφραση δέκα λέξεων με διάρκεια γύρω στα 3secs μπορεί να αναπαρασθεί με μια ακολουθία από T= 300 ακουστικά διανύσματα.



Σχήμα 1.2 Μοντέλο αποκωδικοποίησης

Έστω λοιπόν ότι η πηγή κειμένου παράγει την ακολουθία λέξεων $\underline{W} = [w_1 w_2 w_{3....}w_n]$. Το ακουστικό κανάλι (μοντέλο παραγωγής φωνής του ομιλητή μαζί με τον front-end επεξεργαστή) μπορεί να προσομοιωθεί ως την διαμόρφωση και μετάδοση του μηνύματος $\underline{W}$ μέσα από ένα θορυβώδες κανάλι. Στην έξοδο παίρνουμε την ακολουθία $\underline{X} =[x_1,x_2,\ldots,x_T]$ από παραμετρικά διανύσματα που υπολογίζονται από τον front-end επεξεργαστή του συστήματος αναγνώρισης χρησιμοποιώντας κάποιες από τις μεθόδους που αναφέραμε πιο πριν.

1.3 Κριτήριο Αναγνώρισης

Κατά την αποκωδικοποίηση ζητείται να καθοριστεί με βάση κάποιο κριτήριο ότι εστάλη η ακολουθία λέξεων $\underline{W}$, δεδομένου ότι ο αποκωδικοποιητής έλαβε στην είσοδο του την ακολουθία διανυσμάτων $\underline{X}$. Οι στατιστικές μέθοδοι αναγνώρισης προϋποθέτουν την ύπαρξη κάποιου στατιστικού μοντέλου για τον υπολογισμό της πιθανότητας ή συνάρτησης πιθανοφάνειας. Πρόκειται για το μέγεθος $P(\underline{W}|\underline{X})$. Επιπλέον, ως κριτήριο αποκωδικοποίησης, όπως και σε ένα τυπικό ψηφιακό τηλεπικοινωνιακό σύστημα, είναι η ελαχιστοποίηση της πιθανότητας σφάλματος. Με βάση το μοντέλο $P(\underline{W}|\underline{X})$, η πιθανότητα σφάλματος ελαχιστοποιείται αν αποκωδικοποιήσουμε στην ακολουθία εκείνη $\underline{W}$ για την οποία μεγιστοποιείται η a-posteriori πιθανότητα δεδομένου ότι ο αποκωδικοποιητής έλαβε την ακολουθία $\underline{X}=[x_1,x_2,\ldots,x_T]$.

Χρησιμοποιώντας τον κανόνα του Bayes έχουμε:

$$\hat{W} = \arg\max_{W} P\left(W \mid X\right) = \arg\max_{W} \frac{P(W)P(X \mid W)}{P(X)} \approx$$

$$\approx \arg\max_{W} P(W)P(X \mid W) \qquad (1\text{-}3.1)$$

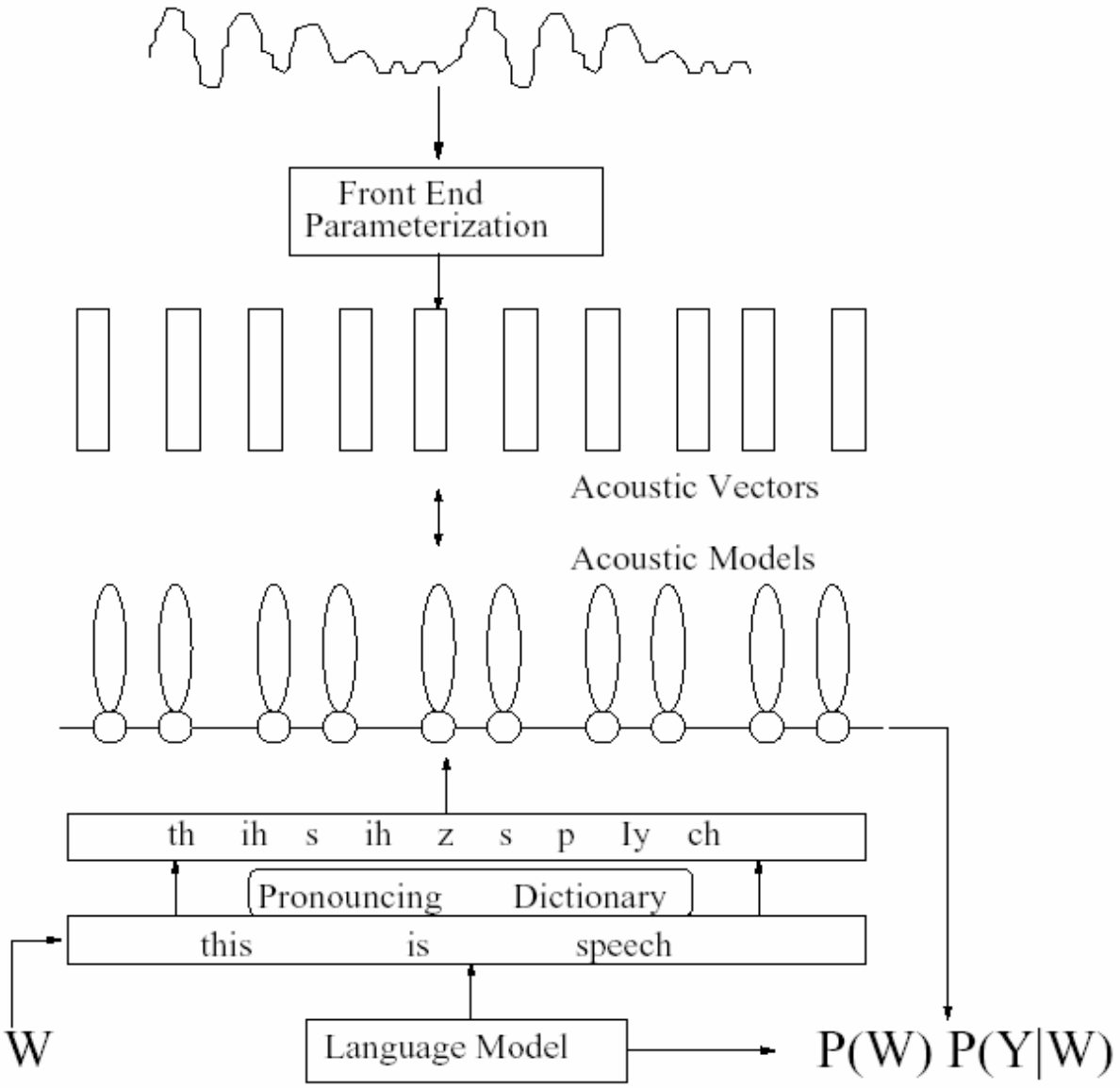
όπου το argmax συμβολίζει το όρισμα που μεγιστοποιεί την αντίστοιχη ποσότητα. Αυτή η εξίσωση δείχνει ότι για να βρεθεί η πιο πιθανή ακολουθία λέξεων $\underline{W}$, πρέπει να βρεθεί η ακολουθία εκείνη που μεγιστοποιεί το γινόμενο $P(\underline{W})*P(\underline{X}|\underline{W})$. Ο όρος $P(\underline{W})$ υπολογίζει την a-priori πιθανότητα της παρατήρησης $\underline{W}$ ανεξάρτητα από το σήμα που παρατηρήθηκε με βάση κάποιο στατιστικό μοντέλο και αυτή η πιθανότητα είναι γνωστή ως γλωσσικό μοντέλο (language model). Ο δεύτερος όρος $P(\underline{X}|\underline{W})$ αναπαριστά την πιθανότητα εμφάνισης μιας ακολουθίας διανυσμάτων $\underline{X}$ δεδομένων μερικών ακολουθιών λέξεων $\underline{W}$, και αυτή η πιθανότητα είναι γνωστή ως ακουστικό μοντέλο (acoustic model). Για το

γλωσσικό και το ακουστικό μοντέλο θα μιλήσουμε εκτενέστερα στα **κεφάλαια 3 και 4** αντίστοιχα.

Η γλωσσική μονάδα που αναπαρίσταται είναι συνήθως η λέξη. Για να υπάρχει δυνατότητα γενίκευσης και να μοντελοποιούνται λέξεις που δεν παρατηρήθηκαν στα δεδομένα εκπαίδευσης, χρησιμοποιούνται μικρότερες γλωσσικές μονάδες όπως το φώνημα (phoneme) ή η συλλαβή. Κάθε λέξη μετατρέπεται σε μια ακολουθία βασικών ήχων, τα φωνήματα που μόλις αναφέραμε, χρησιμοποιώντας ένα λεξικό προφορών (dictionary). Το λεξικό προφορών είναι ένα αρχείο που περιέχει τις ηχητικές αποδόσεις όλων των λέξεων που περιέχονται στη γραμματική και πρέπει να συνταχθεί ώστε να περιγράφει ακριβώς τις προφορές των λέξεων ακόμη και με περισσότερους του ενός τρόπους.

Για καθένα υπάρχει ένα αντίστοιχο στατιστικό μοντέλο HMM. Από στατιστικής πλευράς, ένας κατάλογος από στοχαστικά μοντέλα βασικών φωνητικών μοντέλα βασικών φωνητικών μονάδων χρησιμοποιείται για να αναπαραστήσει λέξεις. Μια ακολουθία από ακουστικές παραμέτρους, προερχόμενες από το σήμα φωνής, αντιμετωπίζεται ως συνδυασμός στοιχειωδών διαδικασιών που περιγράφονται από HMMs.

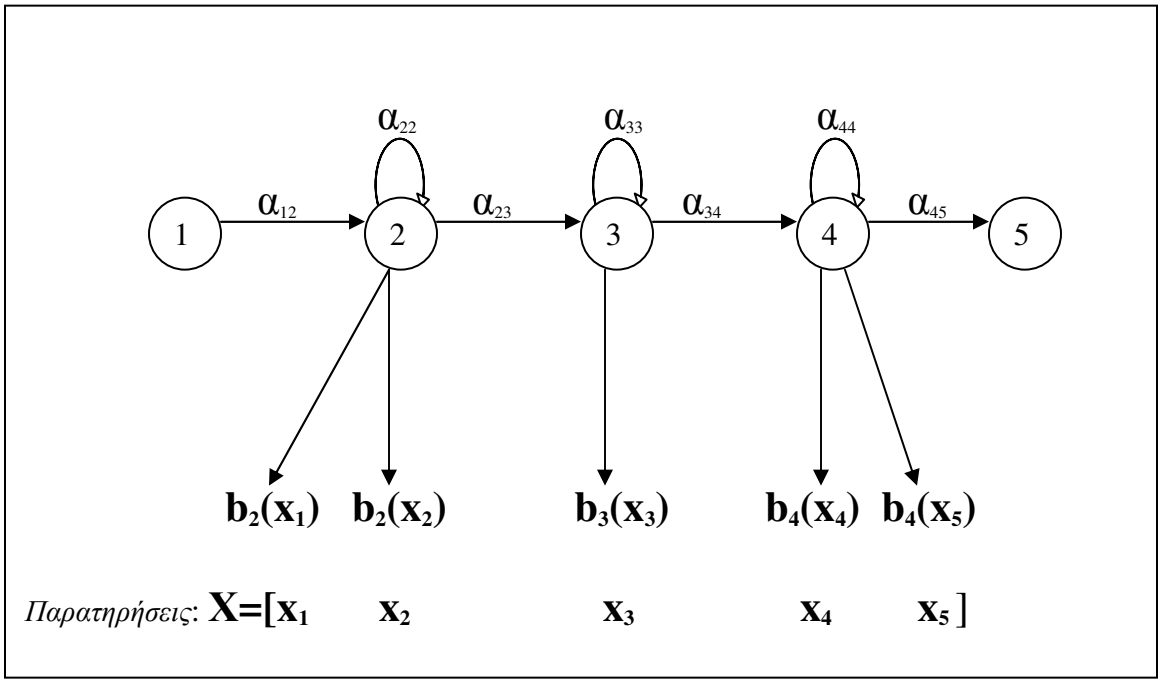
Η πιθανότητα $P(\underline{X}|\underline{W})$ υπολογίζεται χρησιμοποιώντας ένα σύνθετο HMM που αναπαριστά την ακολουθία $\underline{W}$ και αποτελείται από απλά HMM φωνητικά μοντέλα, συνδεδεμένα σειριακά μεταξύ τους σύμφωνα με τις προφορές στο λεξικό προφορών και υπολογίζεται η πιθανότητα να παράγει αυτό το μοντέλο την παρατηρούμενη ακολουθία $\underline{X}$. Η αρχική πιθανότητα $P(\underline{W})$ καθορίζεται από το γλωσσικό μοντέλο.. Τα παραπάνω φαίνονται αναλυτικά στο Σχήμα 1-3 όπου περιγράφεται η διαδικασία υπολογισμού της πιθανότητας $P(\underline{X}|\underline{W})$ και της πιθανότητας $P(\underline{W})$.



Σχήμα 1-3 Αναγνώριση φωνής με στατιστικές μεθόδους

1.4 Περιγραφή των Hidden Markov Models

Ένα HMM είναι ένα σύνολο από καταστάσεις (states) συνδεδεμένα από μεταβάσεις (βλ. σχήμα 1-4). Οι μεταβάσεις μοντελοποιούν την τιμή της εξόδου για ένα τμήμα (frame) ομιλίας. Κάθε μετάβαση του HMM συνδέεται με μια κατανομή εξόδου που ορίζει την πιθανότητα εμφανίσεως του διανύσματος εισόδου που παρατηρήθηκε για δεδομένο frame. Στην πράξη, τα περισσότερα συστήματα συνδέουν την κατανομή εξόδου με τις καταστάσεις, παρά με τις μεταβάσεις. Εφεξής, θα υποθέτουμε ότι η πιθανότητα εξόδου συνδέεται με τον προορισμό της μετάβασης. Η πιθανότητα εξόδου (state) i την χρονική στιγμή t συμβολίζεται με $b_i(t)$. Ουσιαστικά η b_i δεν είναι συνάρτηση του t αλλά συνάρτηση του σήματος ομιλίας, το οποίο είναι συνάρτηση του t. Ωστόσο, θα χρησιμοποιούμε τον συμβολισμό $b_i(t)$ έχοντας εις γνώση μας τα προηγούμενα.



Σχήμα 1.4 Γραφική αναπαράσταση ενός HMM

Ανάλογα με το αν η διαδικασία που μοντελοποιούμε αποτελείται από συνεχή τυχαία διανύσματα (π.χ. συντελεστές cepstral) ή έχει περάσει από κβαντιστή και είναι διαδικασία από διακριτές τυχαίες μεταβλητές, έχουμε διαφορετικά είδη

HMMs, που ταξινομούνται ανάλογα τον τύπο της κατανομής εξόδου. Παρακάτω θα αναλύσουμε σε βάθος τα είδη των κατανομών εξόδου.

Κάθε μετάβαση στα HMM από την κατάσταση *i* στην κατάσταση *j* έχει επίσης μια στατική πιθανότητα μετάβασης (transition probability), που συμβολίζεται με α_{ij} ,και είναι ανεξάρτητη από την είσοδο ομιλίας.

Συνοπτικά ένα Hidden Markov Model ορίζεται από τα εξής στοιχεία:

- *N* - Ο αριθμός των καταστάσεων στο HMM, συμπεριλαμβανομένης της αρχικής και της τελικής κατάστασης.
- *M* - Το πλήθος των διακριτών συμβόλων που μπορούν να παρατηρηθούν ανά κατάσταση.
- *A* - Ένας N x N πίνακας μεταβάσεων, όπου α_{ij} αναπαριστά την πιθανότητα μετάβασης από την κατάσταση *i* στην κατάσταση *j*.
- b_i - Οι κατανομές εξόδου σε μια κατάσταση *i* ($1 < i < N$)
- π - Ένα διάνυσμα μεγέθους *N* με την κατανομή των αρχικών πιθανοτήτων.

Για τον πλήρη καθορισμό ενός HMM απαιτούνται οι παράμετροι *N* και *M* καθώς και ο καθορισμός του συνόλου των συμβόλων παρατήρησης και των τριών πιθανοτικών μεγεθών: *A, B,* π. Για συντομία χρησιμοποιούμε τον πιο συμπαγή συμβολισμό:

$$\lambda = (A, B, \pi)$$

Τα *A, B* και π πρέπει να πληρούν τις παρακάτω ιδιότητες:

$\alpha_{ij} \geq 0$, $b_i \geq 0$, για κάθε i , j

$\sum_j \alpha_{ij} = 1$ για κάθε i

$\sum_k b_i(k) = 1$ για κάθε i , $1 \leq k \leq M$ *(για διακριτά HMMs)*

Τα a και b γράφονται ως εξής:

$$\alpha_{ij} = P(q_{t+1} = j \mid q_t = i)$$

$$b_i(k) = P(\mathbf{x}_t = k \mid q_t = i)$$

όπου $q_t = i$ σημαίνει ότι η μαρκοβιανή αλυσίδα ήταν στην κατάσταση i την χρονική στιγμή t, και $\mathbf{x}_t = k$ σημαίνει ότι το σύμβολο εξόδου την χρονική στιγμή ήταν k. Θα χρησιμοποιήσουμε την τυχαία μεταβλητή **X** για να συμβολίσουμε την πιθανοτική συνάρτηση (probabilistic function) μιας **Q΄** στατικής (stationary) αλυσίδας Markov. Και το **Q** και το **X** παράγονται από ένα κρυφό μαρκοβιανό μοντέλο, ωστόσο το **X,** η ακολουθία εξόδου, παρατηρείται άμεσα ενώ το **Q** η ακολουθία καταστάσεων είναι κρυφή.

Σε ένα πρώτης τάξης κρυφό μαρκοβιανό μοντέλο, υπάρχουν δυο υποθέσεις. Η πρώτη είναι η υπόθεση του Markov.

$$P(q_{t+1} = j \mid q_1^t = j_1^t) = P(q_{t+1} = j \mid q_t = i) \qquad (1.1)$$

Όπου q_i^t αντιπροσωπεύει την ακολουθία καταστάσεων $q_i, q_{i+1}, \ldots, q_t$ και j_i^t αντιπροσωπεύει την ακολουθία τιμών $j_i, j_{i+1}, \ldots, j_t$ των τυχαίων μεταβλητών q. Η ισότητα (1.1) δηλώνει ότι η πιθανότητα η μαρκοβιανή αλυσίδα να είναι σε μια συγκεκριμένη κατάσταση (state) για την χρονική στιγμή $t+1$ εξαρτάται μόνο από την κατάσταση στην οποία βρίσκεται η αλυσίδα την χρονική στιγμή t, και είναι υπό όρους ανεξάρτητη του παρελθόντος.

Η δεύτερη υπόθεση είναι η ανεξαρτησία της εξόδου:

$$P(\mathbf{x}_t = k \mid \mathbf{X}_1^{t-1} = k_1^{t-1}, \; q_1^{t+1} = j_1^{t+1}) = P(\mathbf{x}_t = k \mid q_t = i_t, \; q_{t+1} = j) \qquad (1.2)$$

Η ισότητα (1.2) δηλώνει ότι η πιθανότητα ένα συγκεκριμένο σύμβολο θα παραχθεί την χρονική στιγμή t εξαρτάται μόνο από την μετάβαση που

πραγματοποιείται σ' αυτήν την χρονική στιγμή (από την κατάσταση q_t στην q_{t+1}) και είναι υπό όρους ανεξάρτητη του παρελθόντος.

1.5 HMMs:Βασικά προβλήματα και αλγόριθμοι επίλυσης

Τα HMMs χαρακτηρίζονται από τρία βασικά προβλήματα που είναι:

- Ο υπολογισμός της πιθανότητας μιας ακολουθίας παρατηρήσεων.
- Ο υπολογισμός της πιο πιθανής ακολουθίας καταστάσεων.
- Η εκμάθηση των παραμέτρων τους από δεδομένα εκπαίδευσης.

Με την λύση του πρώτου προβλήματος, του προβλήματος της πιθανότητας μιας ακολουθίας παρατηρήσεων, μπορούμε να δούμε την ταύτιση μεταξύ του μοντέλου και της ακολουθίας παρατηρήσεων. Μια σειρά επαναληπτικών αλγορίθμων (ο αλγόριθμος Forward-Backward στην περίπτωση μας) πρέπει να τρέξει στα δεδομένα εκπαίδευσης για τον υπολογισμό της πιθανότητας να βρισκόμαστε σε μια συγκεκριμένη στιγμή. Τα δεύτερο πρόβλημα, το πρόβλημα της αποκωδικοποίησης που υλοποιείται με τον αλγόριθμο Viterbi, μας δίνει την βέλτιστη ταύτιση της ακολουθίας καταστάσεων δεδομένου της ακολουθίας παρατηρήσεων. Τέλος σχετικά με το τελευταίο πρόβλημα, το πρόβλημα της εκμάθησης των παραμέτρων, χρησιμοποιώντας τον αλγόριθμο του Baum-Welch εκπαιδεύουμε το HMM, δηλαδή τις παραμέτρους που το χαρακτηρίζουν.

1.5.1 Το πρόβλημα υπολογισμού της πιθανότητας μιας ακολουθίας παρατηρήσεων

Το πρόβλημα του υπολογισμού της πιθανότητας μιας ακολουθίας παρατηρήσεων δηλώνεται ως εξής: Δεδομένου του μοντέλου λ, με παραμέτρους A, B και π, υπολόγισε την πιθανότητα $P(\mathbf{X}|\ \lambda)$ παραγωγής της ακολουθίας

παρατηρήσεων $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_{2,\ldots,}\mathbf{x}_T)$. Αυτό επιτυγχάνεται με την άθροιση των πιθανοτήτων για όλες τις πιθανές ακολουθίες καταστάσεων $\mathbf{Q} = (q_1, q_2 ,\ldots,q_t)$:

$$P(\mathbf{X}|\lambda)= \sum_{Q} P(\mathbf{X}|\ \mathbf{Q},\lambda)\ P(\mathbf{Q}\ |\lambda) \qquad (1.3)$$

Το πρώτο μέλος του γινομένου της εξίσωσης (1.3), δηλαδή η πιθανότητα της ακολουθίας παρατηρήσεων **X** δεδομένης της ακολουθίας καταστάσεων **Q** δίνεται από τον τύπο:

$$P(\mathbf{X}|\mathbf{Q},\lambda)=\prod_{t=1}^{T} P(\mathbf{x}_T\ |q_t,\lambda) \qquad (1.4)$$

όπου υποθέτουμε στατιστική ανεξαρτησία των παρατηρήσεων. Οπότε έχουμε:

$$P(\mathbf{X}|\mathbf{Q},\lambda) = b_{q1}(\mathbf{x}_1)\ b_{q2}(\mathbf{x}_2)\ldots\ b_{qT}(\mathbf{x}_T) \qquad (1.5)$$

Η πιθανότητα μιας τέτοιας ακολουθίας καταστάσεων Q γράφεται ως :

$$P(\mathbf{Q}|\lambda) = \pi_{q1}\alpha_{\ q1q2}\alpha_{q2q3}\ldots\alpha_{qT\text{-}1qT} \qquad (1.6)$$

Σύμφωνα με τους παραπάνω τύπους η σχέση (1.3) γράφεται ως εξής:

$$P(\mathbf{X}|\lambda)= \sum_{Q} \pi_{q1}b_{q1}(\mathbf{x}_1)\alpha_{\ q1q2}b_{q2}(\mathbf{x}_2)\ldots\alpha_{qT\text{-}1qT}b_{qT}(\mathbf{x}_T) \qquad (1.7)$$

<u>Ο αλγόριθμος Forward-Backward</u>

Ο αλγόριθμος Forward υπολογίζει την πιθανότητα $P(\mathbf{X}|\lambda)$ με πολυπλοκότητα N^2T. Τα βήματα υπολογισμού φαίνονται συνοπτικά στο σχήμα 1.5. Ορίζουμε την ποσότητα $\alpha_t(j)$ ως την κοινού πιθανότητα να βρισκόμαστε τη χρονική στιγμή t στην κατάσταση $q_t = j$

Στάδιο I. Αρχικοποίηση $\alpha_1(i) = \pi_i b_i(\mathbf{x}_1)$, με $1 \le i \le N$

Στάδιο II.

Για $t = 2,\ldots,T$
  Για $j = 1,\ldots, N$

$$\text{Υπολόγισε: } \alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i)\alpha_{ij}\right] b_i(\mathbf{x}_t)$$

Στάδιο III.

$$\text{Υπολόγισε: } P(\mathbf{X}|\lambda) = \sum_{I}^{N} \alpha_T(i)$$

Σχήμα 1.5 Ο αλγόριθμος Forward

Ο αλγόριθμος Backward υπολογίζει επίσης την πιθανότητα $P(\mathbf{X}|\lambda)$ με πολυπλοκότητα N^2T. Τα βήματα υπολογισμού φαίνονται συνοπτικά στο σχήμα 1.6. Ορίζουμε την πιθανότητα $\beta_T(i)$ ως την πιθανότητα των $\{ \mathbf{x}_{t+1,\ldots,} \mathbf{x}_T \}$, δεδομένου ότι τη χρονική στιγμή t βρισκόμαστε στην κατάσταση $q_t = i$

Στάδιο I. Αρχικοποίηση $\beta_T(i) = 1$, με $1 \le i \le N$

Στάδιο II.

Για $t = T-1, T-2,\ldots,1$
  Για $i = 1,\ldots, N$

$$\text{Υπολόγισε: } \beta_T(i) = \sum_{j} \alpha_{ij} b_j(\mathbf{x}_{t+1})\beta_{t+1}(j)$$

Στάδιο III.

$$\text{Υπολόγισε } P(\mathbf{X}|\lambda) = \sum_{i} \pi_i b_i(\mathbf{x}_i)\beta_1(i)$$

Σχήμα 1.6 Ο αλγόριθμος Backward

1.5.2 Το πρόβλημα της αποκωδικοποίησης

Ενώ ο αλγόριθμος Forward-Backward υπολογίζει την πιθανότητα ένα HMM να έχει παράγει μια ακολουθία παρατηρήσεων, δε δίνει πληροφορία για την ακολουθία καταστάσεων. Αυτή η πληροφορία, η εύρεση της βέλτιστης ακολουθίας καταστάσεων, είναι απαραίτητη για την αναγνώριση ομιλίας.

Δυστυχώς, εξ ορισμού, η ακολουθία καταστάσεων είναι κρυφή σε ένα HMM. Η προσεγγιστική λύση σε αυτό το πρόβλημα είναι η εύρεση της ακολουθίας καταστάσεων με την μεγαλύτερη πιθανότητα να έχει παράγει την ακολουθία παρατηρήσεων. Ο αλγόριθμος μπορεί να παραχθεί με μια μικρή τροποποίηση του αλγορίθμου Forward. Στον αλγόριθμο Forward αθροίζαμε τις πιθανότητες για όλες τις πιθανές ακολουθίες καταστάσεων. Τώρα χρειάζεται να διαλέξουμε και να θυμόμαστε την μέγιστη.

$$v_i(t) = \begin{cases} 0 & t = 0 \quad \gamma\iota\alpha \quad i \neq 1 \\ 1 & t = 0 \quad \gamma\iota\alpha \quad i = 1 \\ MAX_j v_i(t-1) a_{ij} b_i(x_t) & t > 0 \end{cases} \qquad (1.8)$$

Ο αλγόριθμος αυτός είναι γνωστός ως αλγόριθμος του Viterbi.

1.5.3 Το πρόβλημα της εκμάθησης παραμέτρων

Το τρίτο και το πιο δύσκολο πρόβλημα των HMMs είναι η κατάλληλη επιλογή των παραμέτρων λ έτσι ώστε να μεγιστοποιηθεί η πιθανότητα:

$$\max_{\lambda} P(\mathbf{X}|\lambda) \qquad (1.9)$$

Επειδή δεν υπάρχει αναλυτική μέθοδος επίλυσης του προβλήματος καταφεύγουμε συνήθως σε επαναληπτικές μεθόδους. Η πιο διαδεδομένη είναι η μέθοδος του Baum-Welch (γνωστή και ως EM (expectation-maximization)

μέθοδος). Πρόκειται για επαναληπτικό αλγόριθμο που σε κάθε βήμα μεγιστοποιεί την ποσότητα:

$$E\{\log P(\mathbf{X},\mathbf{Q}|\ \lambda_{new})|\ \mathbf{X},\ \lambda_{old}\} \tag{1.10}$$

όπου η αναμενόμενη τιμή υπολογίζεται πάνω σε όλες τις πιθανές ακολουθίες καταστάσεων **Q.** Αυτό γίνεται διότι αυτή δεν είναι γνωστή και δεν μπορεί να μεγιστοποιηθεί απευθείας η ποσότητα:

$$\log P(\mathbf{X}|\mathbf{Q},\lambda) \tag{1.11}$$

Μπορεί να αποδειχτεί ότι ο αλγόριθμος συγκλίνει σε κάποιο τοπικό ακρότατο της συνάρτησης:

$$\log P(\mathbf{X}|\lambda) \tag{1.12}$$

δηλαδή υπολογίζει εκτιμήτριες μέγιστης πιθανοφάνειας (Maximum Likelihood ML) των παραμέτρων λ.

Για να περιγράψουμε την διαδικασία επανυπολογισμού (reestimation) των παραμέτρων του HMM, πρώτα ορίζουμε το $\xi_t(i,j)$, την πιθανότητα να βρισκόμαστε στην κατάσταση *i* την χρονική στιγμή t, και την κατάσταση *j* την χρονική στιγμή t +1, δεδομένου του μοντέλου και την ακολουθία παρατηρήσεων, δηλαδή:

$$\xi_t(i,j) = P(q_t=i,\ q_{t+1}=j|\mathbf{X},\lambda) \tag{1.13}$$

Από τους ορισμούς των μεταβλητών του Forward-Backward αλγόριθμου μπορούμε να εκφράσουμε το $\xi_t(i,j)$ στην μορφή:

$$\xi_t(i,j) = \frac{P(q_t=i,\ q_{t+1}=j|\mathbf{X},\lambda)}{P(\mathbf{X},\lambda)}$$

$$= \frac{\alpha_t(i)\alpha_{ij}b_j(\mathbf{x}_{t+1})\beta_{t+1}(j)}{P(\mathbf{X},\lambda)}$$

$$= \frac{\alpha_t(i)\alpha_{ij}b_j(\mathbf{x}_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)\alpha_{ij}b_j(\mathbf{x}_{t+1})\beta_{t+1}(j)} \quad (1.14)$$

Ορίζουμε επίσης το $\gamma_t(i)$ ως την πιθανότητα να βρισκόμαστε στην κατάσταση *i* την χρονική στιγμή t, δεδομένου του μοντέλου και την ακολουθία παρατηρήσεων, δηλαδή:

$$\gamma_t(i) = P(q_t=i|\mathbf{X},\lambda) \quad (1.15)$$

Μπορούμε να συσχετίσουμε το $\gamma_t(i)$ με το $\xi_t(i,j)$ αθροίζοντας για όλα τα *j,* οπότε:

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \quad (1.16)$$

Αν αθροίσουμε το $\gamma_t(i)$ για όλα τα t, η ποσότητα που παίρνουμε μεταφράζεται ως ο αναμενόμενος (στον χρόνο) αριθμός των μεταβάσεων που έγιναν με αφετηρία τη κατάσταση *i* και προορισμό την κατάσταση *j*.

Με βάση τα παραπάνω, μπορούμε να δώσουμε μια μέθοδο επανυπολογισμού των παραμέτρων ενός HMM.

$\bar{\pi}_j$ = αναμενόμενη συχνότητα στην κατάσταση *i* την χρονική στιγμή (t = 1) = $\gamma_1(i)$ (1.17α)

$$\bar{\alpha}_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma(i)} \quad (1.17b)$$

$$\bar{b}_j(k) = \frac{\sum_{t=1}^{T-1}\gamma_t(j)}{\sum_{t=1}^{T-1}\gamma_t(j)} \qquad \text{όπου } \mathbf{x}_t=\mathbf{v}_k \quad (1.17c)$$

Αν χρησιμοποιήσουμε το μοντέλο $\lambda = (\mathrm{A}, \mathrm{B}, \pi)$ για να υπολογίσουμε τις εξισώσεις (1.17a)-(1.17b), το νέο μοντέλο $\bar{\lambda} = (\bar{\mathrm{A}}, \bar{\mathrm{B}}, \bar{\pi})$ με τις εκτιμημένες παραμέτρους έχει αποδειχθεί ότι το μοντέλο $\bar{\lambda}$ είναι πιο πιθανό από το λ με την έννοια ότι logP(**X**|$\bar{\lambda}$) > logP(**X**|λ) , δηλαδή βρήκαμε ένα νέο μοντέλο $\bar{\lambda}$ από το οποίο η ακολουθία παρατηρήσεων είναι πιο πιθανό να παραχθεί.

Βασισμένοι στην προηγούμενη διαδικασία, αν επαναληπτικά χρησιμοποιήσουμε το $\bar{\lambda}$ στη θέση του λ και επαναλάβουμε τον επαναυπολογισμό, μπορούμε να βελτιώσουμε την πιθανότητα παραγωγής του **X** από το μοντέλο , μέχρις να φτάσουμε σε μερικό ακρότατο.

Οι τύποι επαναυπολογισμού (1.17a)-(1.17b) μπορούν να προκύψουν κατευθείαν με μεγιστοποίηση της βοηθητικής συνάρτησης του Baum:

$$Q(\lambda', \lambda) = \sum_{q} P(\mathbf{X}, q|\, \lambda') \log P(\mathbf{X}, q|\, \lambda) \qquad (1.18)$$

Πάνω στο λ. Επειδή

$$Q(\lambda', \lambda) \geq Q(\lambda', \lambda') \;\Rightarrow\; P(\mathbf{X}, \lambda) \geq P(\mathbf{X}, \lambda') \qquad (1.19)$$

Μπορούμε να μεγιστοποιήσουμε την συνάρτηση πάνω στο λ.

Τα βήματα του γενικού επαναληπτικού αλγορίθμου για διακριτά HMMs φαίνονται συνοπτικά στο σχήμα 1.7

Στάδιο I. (Αρχικοποίηση)

Θέσε: **Π** = π_i, **A** = α_{ij} , **B** = $b_i(k)$

Στάδιο II. (Forward-Backward)

Χρησιμοποιώντας τις παραμέτρους $\lambda = (A, B, \pi)$ και τον αλγόριθμο Forward-Backward:

Για t =1,…,T

Για *i* =1,…,*N*

Υπολόγισε : $P(\mathbf{x}_1, K, x_t, q_t = i \mid \lambda) = \alpha_t(i)$

και: $P(\mathbf{x}_{t+1}, K, x_T, q_t = i \mid \lambda) = \beta_t(i)$

Στάδιο III. (Maximization)

Υπολόγισε νέες τιμές για όλες τις παραμέτρους:

$$\bar{\alpha}_{ij} = \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} = \frac{\sum_{t=1}^{T} \alpha_{t-1}(i) a_{ij} b_j(x_t) \beta_t(j)}{\sum_{t=1}^{T} a_{t-1}(i) \beta_{t-1}(i)}$$

$$\bar{b}_i = \frac{\sum_{t=1}^{T} a_t(i) \beta_t(i) \delta(x_t, v_k)}{\sum_{t=1}^{T} a_t(i) \beta_t(i)}$$

όπου $\delta(x_t, v_k) = 1$ αν $x_t = v_k$

$= 0$ αλλιώς

Στάδιο IV. (Τερματισμός)

Αν το κριτήριο σύγκλισης δεν ικανοποιείται, θέσε :

$\pi_i = \bar{\pi}_i, \alpha_{ij} = \bar{a}_{ij}, b_i(k) = \bar{b}_i(k)$ και πήγαινε στο στάδιο II.

Σχήμα 1.7 Ο γενικός αλγόριθμος Baum-Welch

1.6 Είδη των HMM

Όπως ήδη αναφέραμε, το είδος του HMM εξαρτάται από το αν η διαδικασία που μοντελοποιούμε αποτελείται από συνεχή τυχαία διανύσματα (π.χ. συντελεστές cepstral) ή έχει περάσει από κβαντιστή και είναι διαδικασία από διακριτές τυχαίες μεταβλητές. Τα είδη των HMMs ταξινομούνται ανάλογα με τον τύπο της κατανομής εξόδου και χωρίζονται σε διακριτά και συνεχή.

1.6.1 Διακριτά HMMs (Discrete HMMs)

Αν η διαδικασία $\{\mathbf{x}_t\}$ είναι διακριτή, με $\mathbf{x}_t \in \{1,2,\ldots,Q\}$ τότε η κατανομή εξόδου $b_j(\mathbf{x}_t)$ είναι διακριτή με κατανομή :

$$\sum_{k=1}^{Q} b_j(k) = 1 \tag{1.20}$$

1.6.2 Συνεχή HMMs (Continuous HMMs)

Οι κατανομές εξόδου είναι από κοινού συναρτήσεις πυκνότητας πιθανότητας ενός τυχαίου διανύσματος $\mathbf{x}_t$ με τιμή:

$$b_j(\mathbf{x}_t) \text{ , όπου } \mathbf{x}_t = \begin{bmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{dt} \end{bmatrix} \tag{1.21}$$

και d είναι η διάσταση του $\mathbf{x}_t$ (π.χ. τάξη της ανάλυσης LPC, αριθμός συντελεστών cepstral κ.λ.π.).

Για να χρησιμοποιήσουμε μια συνεχή πυκνότητα παρατηρήσεων, πρέπει να τεθούν κάποιοι περιορισμοί στην μορφή της συνάρτησης πυκνότητας πιθανότητας έτσι ώστε να εξασφαλίζεται ότι οι παράμετροι της μπορούν να επαναυπολογιστούν με ακριβή τρόπο. Η συνάρτηση πυκνότητας πιθανότητας

που χρησιμοποιείται για την μέθοδο των συνεχών μειγμάτων γκαουσιανών έχει την μορφή γραμμικών συνδυασμών (μειγμάτων) από γκαουσιανές :

$$b_j(\mathbf{x}_t) = \sum_{k=1}^{M} c_{jk}\, \boldsymbol{N}(\mathbf{x}_t, \mu_{jk}, \Sigma_{jk}), \qquad 1 \le j \le N \tag{1.22}$$

όπου c_{jk} είναι ο συντελεστής μείγματος (βάρος) για το *k* μείγμα στην κατάσταση *j* και ***N*** μια γκαουσιανή συνάρτηση πιθανότητας με παραμέτρους κατανομής το διάνυσμα μέσων ***μ*** και τον πίνακα διασπορών **Σ**. Οι συντελεστές μείγματος c_{jk} ικανοποιούν τον στοχαστικό περιορισμό

$$\sum_{k=1}^{M} c_{jk} = 1 \ , \ 1 \le j \le N \tag{1.23α}$$

$$c_{jk} \ge 0 \ , 1 \le j \le N, 1 \le k \le M \tag{1.24b}$$

έτσι ώστε:

$$\int_{-\infty}^{\infty} b_j(\mathbf{x}_t) d\mathbf{x}_t = 1 \tag{1.25}$$

Η μέθοδος των συνεχών μειγμάτων γκαουσιανών χρησιμοποιεί μείγματα (γραμμικούς συνδυασμούς) από γκαουσιανές αντί για μια γκαουσιανή έτσι ώστε να μοντελοποιεί επαρκώς την κατανομή του $\mathbf{x}_t$ για μια κατάσταση, ειδικά σε συστήματα αναγνώρισης ανεξάρτητα του ομιλητή.

1.7 Σύγκριση συνεχών και διακριτών HMM

Στην περίπτωση των διακριτών HMMs έχουμε *M* σύμβολα εξόδου, και η συνάρτηση πυκνότητας πιθανότητας εξόδου, $b_i = P(\mathbf{x}_t = k \mid q_t = i)$, μοντελοποιείται επακριβώς. Κάθε τμήμα ομιλίας (frame) αναπαρίσταται με ένα σύμβολο από ένα πεπερασμένο αλφάβητο. Η μέθοδος της διανυσματικής κβαντοποίησης (vector quantization, VQ) που χρησιμοποιείται προσδιορίζει ένα σύνολο από πρότυπα διανύσματα από τα δεδομένα εκπαίδευσης. Έπειτα το σήμα ομιλίας μετατρέπεται

από ένα πολυδιάστατο πραγματικό διάνυσμα, όπως τους συντελεστές cepstral, σε ένα σύμβολο, το οποίο αναπαριστά το πρότυπο διάνυσμα που ταιριάζει περισσότερο περισσότερο με το αρχικό διάνυσμα.

Με τον κβαντισμό και δεδομένου ότι χρησιμοποιούμε πεπερασμένο αριθμό προτύπων διανυσμάτων έχουμε αναπόφευκτα και κάποιο λάθος, το λεγόμενο σφάλμα κβαντισμού. Όσο ο αριθμός των προτύπων διανυσμάτων μεγαλώνει τόσο το σφάλμα κβαντισμού μειώνεται αλλά και τόσο το μέγεθος αποθήκευσης τους μεγαλώνει, Έτσι πρακτικά βρίσκουμε την βέλτιστη λύση ώστε να έχουμε μικρό σφάλμα κβαντισμού και παράλληλα μικρό σχετικά μέγεθος πρότυπων διανυσμάτων. Αυτός ο περιορισμός είναι και ένα από τα βασικά προβλήματα των διακριτών HMMs, δηλαδή η μικρή ανάλυση του ακουστικού χώρου, λόγω πρακτικών προβλημάτων (μεγέθους μοντέλων και χώρου αποθήκευσης τους).

Το πρωταρχικό πλεονέκτημα των συνεχών HMMs είναι η δυνατότητα της απευθείας μοντελοποίησης των παραμέτρων της φωνής, τα οποία συνήθως είναι με την μορφή πολυδιάστατων πραγματικών διανυσμάτων. Έτσι αποφεύγονται τα σφάλματα και επιτυγχάνεται μια μείωση του πλήθους των παραμέτρων.

Από την άλλη όμως, τα συνεχή HMMs απαιτούν σημαντικά μεγαλύτερο χρόνο για την εκπαίδευση των παραμέτρων τους αλλά και για την διαδικασία της αναγνώρισης. Για παράδειγμα, χρησιμοποιώντας διακριτά HMMs απαιτούνται πολλοί μαθηματικοί υπολογισμοί ακόμα και για την απλή περίπτωση της μιας γκαουσιανής. Όπως προαναφέραμε, είναι σχεδόν επιτακτική η ανάγκη χρησιμοποίησης μείγματος γκαουσιανών για την καλύτερη αναπαράσταση των παραμέτρων της φωνής, πράγμα όμως που επιφέρει πρόσθετη πολυπλοκότητα σε εκπαίδευση και αναγνώριση.

ΚΕΦΑΛΑΙΟ 2

Περιγραφή της Βάσης Δεδομένων των Ακουστικών Σημάτων

2.1 Επισκόπιση

Ένα σύστημα αναγνώρισης βασίζει την λειτουργία του στο γλωσσικό και στο ακουστικό μοντέλο. Και για τα δύο αυτά μοντέλα, θα μιλήσουμε αναλυτικά στα επόμενα κεφάλαια (Κεφ. 3 και Κεφ. 4 αντίστοιχα). Για την εκπαίδευση και συνεπώς για την λειτουργία του ακουστικού μοντέλου πρέπει να υπάρχει ένας επαρκής αριθμός δεδομένων, δηλαδή ακουστικών σημάτων, που θα χρησιμοποιηθούν για την υλοποίηση του. Σε αυτό το κεφάλαιο θα κάνουμε μια πλήρη περιγραφή της βάσης δεδομένων που οδήγησε στην δημιουργία των ακουστικών μοντέλων. Η βάση δεδομένων μας απαρτίζεται αποκλειστικά και μόνο από ακουστικά σήματα (acoustic signals) ηχογραφημένα από τηλεοπτικές εκπομπές. Συνεπώς, σε αυτό το κεφάλαιο θα αναφέρουμε και θα αναλύσουμε διεξοδικά τα παρακάτω θέματα σχετικά με τα ακουστικά σήματα:

- τον τρόπο και την πηγή ηχογράφησης τους
- τα χαρακτηριστικά τους
- την απομαγνητοφώνηση τους
- το ‘’κόψιμο‘’ των ακουστικών σημάτων σε προτάσεις μικρής διάρκειας (μερικών μόλις δευτερολέπτων)
- την κατηγοριοποίηση τους

2.2 Συγκέντρωση των ακουστικών σημάτων

Όπως ήδη έχουμε αναφέρει, σε αυτήν την διπλωματική εργασία υλοποιούμε ένα σύστημα αναγνώρισης που θα αναγνωρίζει ακουστικά σήματα ηχογραφημένα από τηλεοπτικές εκπομπές. Συνεπώς, είναι αυτονόητο πως για να δημιουργήσουμε ένα τέτοιο αναγνωριστή θα πρέπει να εκπαιδευτεί ένα ακουστικό μοντέλο, με τέτοια δεδομένα. Όποτε το πρώτο και κύριο μέλημα μας είναι να δημιουργηθεί η βάση τέτοιων ακουστικών σημάτων έτσι ώστε να έχουμε στην κατοχή μας δύο σύνολα δεδομένων: το σύνολο εκπαίδευσης και το σύνολο δοκιμής. Θα εξηγήσουμε παρακάτω σε αυτό το κεφάλαιο την διαδικασία δημιουργίας των παραπάνω συνόλων.

Αρχικά εγγράψαμε μέσω της κάρτας τηλεόρασης για υπολογιστή MPEG PC TV RADIO της CRYPTO 40 ώρες τηλεοπτικών ειδήσεων. Η συγκεκριμένη κάρτα έχει δυνατότητα εγγραφής εικόνας και ήχου σε MPEG 1,2,3 και 4. Επιπλέον έχει δυνατότητα εγγραφής βίντεο που επιτρέπει μεγαλύτερη συμπίεση και καλύτερη ποιότητα εικόνας από την κοινή μέθοδο (AVI).

Σε αυτό το σημείο θα πρέπει να σημειωθεί ότι το αρχεία βίντεο που εγγράψαμε είναι από τα τηλεοπτικά κανάλια ΝΕΤ (Νέα Ελληνική Τηλεόραση), ΕΤ1(Ελληνική Τηλεόραση 1) και ΣΚΑΪ. Τα τεχνικά χαρακτηριστικά αυτών των αρχείων βίντεο είναι τα εξής:

- ανάλυση 640 x 480
- 1411kbps (πληροφορία ήχου)
- 16bit audio sample size (πληροφορία ήχου)
- 44 kHz audio sample rate (πληροφορία ήχου)
- PCM audio format (πληροφορία ήχου)
- 25 frames/sec frame rate (πληροφορία βίντεο)
- 267kbps data rate (πληροφορία βίντεο)

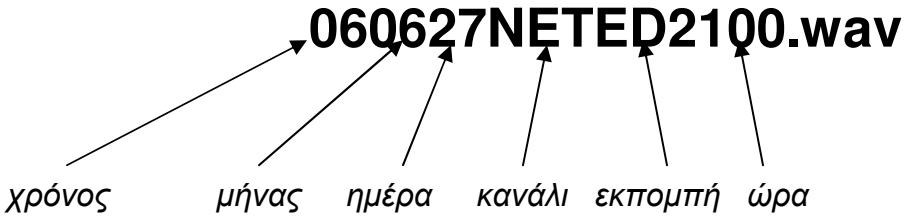
- 16bit video sample size (πληροφορία βίντεο)
- MPEG-4 video format (πληροφορία βίντεο)

Αφού εγγράψαμε τις 40ώρες βίντεο, εξάγαμε από αυτά την ακουστική πληροφορία (audio stream) με την βοήθεια του προγράμματος Virtual Dub (http://www.virtualdub.org/), έτσι ώστε να την επεξεργαστούμε στην συνέχεια κατάλληλα και να οδηγηθούμε τελικά στην δημιουργία του ακουστικού μας μοντέλου. Οπότε τελικά έχουμε 40 αρχεία ήχου (wav files) διάρκειας περίπου μίας ώρας το καθένα, με τα εξής χαρακτηριστικά:

- 256kbps bit rate
- 16bit audio sample size
- 1 channel (mono)
- 16kHz audio sample rate
- PCM audio format

2.3 Απομαγνητοφώνηση και τεμαχισμός των ακουστικών σημάτων

Πριν προχωρήσουμε και εξηγήσουμε πως έγινε η επεξεργασία των δεδομένων, θα δείξουμε πως ονομάσαμε καθένα από τα αρχεία ήχου μας για την πιο εύκολη επεξεργασία τους στη συνέχεια. Τα στοιχεία που θέλαμε να τονίσουμε σε κάθε σήμα φωνής ήταν η ημερομηνία που παρουσιάστηκε το δελτίο (χρόνος, μήνας, ημέρα), το κανάλι από το οποίο γράψαμε την εκπομπή, το είδος της εκπομπής (ειδήσεις, talk show κλπ.) και τέλος την ώρα παρουσίασης της εκπομπής. Ενδεικτικά θα παρουσιάσουμε ένα από τα ονόματα σήματος φωνής που είναι μέρος της βάσης μας:



Αφού πλέον έχουμε τα ακουστικά σήματα στην μορφή που επιθυμούμε, το επόμενο βήμα μας είναι να απομαγνητοφωνήσουμε (transcribe) αυτά τα σήματα και παράλληλα να τα τεμαχίσουμε έτσι ώστε να έχουμε πολλά και μικρότερα αρχεία ήχου. Αυτό γίνεται για να εκπαιδεύσουμε το ακουστικό μοντέλο κατάλληλα και για να είμαστε σε θέση να κατηγοριοποιήσουμε τις προτάσεις σε διάφορες κατηγορίες, ανάλογα με τις συνθήκες που επικρατούν κάθε φορά στην ομιλία. Παρακάτω θα αναφέρουμε ποιες κατηγορίες σχηματίστηκαν και πως τελικά χρησιμοποιήθηκαν.

Το εργαλείο που χρησιμοποιήθηκε για αυτές τις εργασίες είναι το Transcriber Tool (http://trans.sourceforge.net/). Το Transcriber είναι ένα εργαλείο για τον χειροκίνητο χαρακτηρισμό των σημάτων φωνής. Διαθέτει γραφικό περιβάλλον, φιλικό προς τον χρήστη, για τον τεμαχισμό, την απομαγνητοφώνηση και τον χαρακτηρισμό μεγάλων σε διάρκεια ακουστικών σημάτων.

Η απομαγνητοφώνηση των σημάτων ομιλίας έγινε σε 3 επίπεδα:

- σε επίπεδο συνθήκης ομιλίας
- σε επίπεδο ομιλητή
- σε επίπεδο απομαγνητοφώνησης

2.3.1 Επίπεδο συνθήκης ομιλίας

Έχουμε ήδη σημειώσει πως μια από τις πολλές δυνατότητες που έχει το Transcriber Tool είναι ο τεμαχισμός ενός μεγάλου ακουστικού σήματος σε μικρότερα. Αυτά τα μικρότερα σήματα φωνής που δημιουργούνται θα πρέπει να περιγράφονται ειδικότερα, θα πρέπει να καταγράφεται κατά κάποιο τρόπο η συνθήκη που επικρατεί κατά την διάρκεια της ομιλίας. Δηλαδή θα πρέπει να γνωρίζουμε ποια συνθήκη επικρατεί σε κάθε μια από τις προτάσεις, τα μικρά αρχεία ήχου, που θα δημιουργηθούν τελικά.

Οι συνθήκες αυτές είναι οι εξής:

- **report** - καθαρή ομιλία στο στούντιο, χωρίς θόρυβο
- **music** - ομιλία με μουσική στο background
- **noise** - ομιλία με θόρυβο στο background
- **multi_speakers** - ομιλία από πολλούς ομιλητές ταυτόχρονα
- **non_greek** - ομιλία σε άλλη γλώσσα και όχι στα ελληνικά
- **non_trans** - δεν υπάρχει ομιλία

2.3.2. Επίπεδο ομιλητή

Εφόσον τα σήματα φωνής που έχουμε στην κατοχή μας είναι ακουστικά σήματα από δελτία ειδήσεων, οι ομιλητές που περιέχονται σε αυτά είναι παρουσιαστές ειδήσεων, δημοσιογράφοι, απλοί πολίτες κλπ. Οπότε δόκιμο είναι για τον σχεδιασμό μιας αξιόπιστης βάσης δεδομένων να γίνεται περιγραφή του εκάστοτε ομιλητή.

Τα χαρακτηριστικά που έχει ο κάθε ομιλητής είναι τα εξής:

- όνομα (name)
- φύλο (sex)
- εθνικότητα (dialect)

Σε αυτό το επίπεδο ορίζονται επίσης και τα χαρακτηριστικά της ομιλίας:

- αυθόρμητη ή σχεδιασμένη ομιλία
- καθαρή ομιλία (π.χ. σε στούντιο) ή ομιλία χαμηλότερης ποιότητας (π.χ. ομιλία μέσω τηλεφώνου)

Τέλος, θα πρέπει να αναφέρουμε την περίπτωση που το όνομα του ομιλητή δεν είναι γνωστό. Αν ο ομιλητής είναι δημοσιογράφος τότε του δίνουμε το όνομα **rep** μαζί με έναν αριθμό που συμβολίζει τον αριθμό εμφανίσεων διαφορετικών δημοσιογράφων. Για παράδειγμα, για τον πρώτο δημοσιογράφο που θα συναντήσουμε στο σήμα φωνής, θα δώσουμε το όνομα **rep_01**, για τον δεύτερο το όνομα **rep_02** κ.ο.κ. Αν ο ομιλητής δεν είναι δημοσιογράφος (π.χ. αν είναι ένας πολιτικός του οποίου δεν μπορούμε να αναγνωρίσουμε την ταυτότητα του ή ένας κοινός άνθρωπος), τότε του δίνουμε το όνομα **unk** μαζί με έναν αριθμό που συμβολίζει τον αριθμό εμφανίσεων τέτοιων προσώπων. Για παράδειγμα, για τον πρώτο που θα συναντήσουμε στο σήμα φωνής, θα δώσουμε το όνομα **unk_01**, για τον δεύτερο το όνομα **unk_02** κ.ο.κ.

Συνοψίζοντας, στο επίπεδο του ομιλητή καθορίζουμε τα εξής χαρακτηριστικά:

NAME			
SEX	male	or	female
DIALECT	native	or	non-native
MODE	spontaneous	or	planned
CHANNEL	studio	or	telephone

Πίνακας 2-1 Χαρακτηριστικά ομιλητή

2.3.3 Επίπεδο απομαγνητοφώνησης

Τώρα, αφού ήδη έχουμε περιγράψει τα δύο ανώτερα επίπεδα της απομαγνητοφώνησης, δηλαδή το επίπεδο της συνθήκης ομιλίας αλλά και το επίπεδο του ομιλητή, θα περιγράψουμε πως έγινε η ίδια η απομαγνητοφώνηση των ακουστικών σημάτων. Έτσι, θα περιγράψουμε τα βασικά βήματα-οδηγίες των απομαγνητοφωνήσεων που διενεργήθηκαν.

1. Ορθογραφία

Η κύρια και βασικότερη ιδέα στην εργασία των απομαγνητοφωνήσεων είναι να αποτυπώνουμε γραπτώς, στα ελληνικά, ότι ακριβώς ακούμε στα σήματα φωνής. Κατά τη διάρκεια λοιπόν αυτής της διαδικασίας δε γράφουμε με κεφαλαία, σημεία στίξης και επιπλέον γράφουμε τα πάντα με ελληνικούς χαρακτήρες.

Π.χ.

Είναι πολύ trendy	(Λάθος)	είναι πολύ τρέντι	(Σωστό)
Είναι must	(Λάθος)	είναι μαστ	(Σωστό)

2. Ακρωνύμια , Συντομεύσεις κτλ.

Στην περίπτωση ακρονυμίων, συντομεύσεων κτλ. ενεργούμε παρομοίως. Δηλαδή απομαγνητοφωνούμε ότι ακριβώς ακούμε.

Π.χ.

Αν στο σήμα φωνής ακούμε ‘Η ΔΕΗ είναι μια κερδοφόρα επιχείρηση’

Σωστό : ‘Η δεή είναι μια κερδοφόρα επιχείρηση’

Λάθος : ‘Η δημόσια επιχείρηση ηλεκτρισμού είναι μια κερδοφόρα επιχείρηση’

3. Αριθμοί

Παρομοίως, όπως στις παραπάνω περιπτώσεις.

Π.χ.

Αν στο σήμα φωνής ακούμε '101 νύχτες'

Σωστό : 'εκατόν μία νύχτες'

4. Δισταγμοί ομιλητή

Στις περιπτώσεις που παρατηρείται στην ομιλία του ομιλητή δισταγμός γράφουμε την εξής ακολουθία χαρακτήρων @ε@

Π.χ.

Ομιλητής : από την συμπεριφορά των εε καταναλωτών

Απομαγνητοφώνηση : από την συμπεριφορά των @ε@ καταναλωτών

5. Λανθασμένες προφορές λέξεων

Στην περίπτωση που ο ομιλητής προφέρει λανθασμένα μια λέξη (πιο λαϊκά κάνει σαρδάμ), τότε η λέξη αυτή γράφεται με την σωστή της προφορά αλλά περικλείεται από τον χαρακτήρα '*' .

Π.χ.

Ομιλητής : από την συμπαριφορά των καταναλωτών

Απομαγνητοφώνηση : από την *συμπεριφορά* των καταναλωτών

6. Ατελείς λέξεις

Λέξεις που έχουν ειπωθεί ατελώς γράφονται με την εξής ακολουθία

χαρακτήρων : [FRAGMENT]

Π.χ.

Ομιλητής : συμπεριφορά των κατ καταναλωτών

Απομαγνητοφώνηση : συμπεριφορά των [FRAGMENT] καταναλωτών

7. Στιγμιαίοι θόρυβοι

Σε περιπτώσεις όπου στην διάρκεια της ομιλίας έχουμε στιγμιαίους θόρυβους θα αποτυπώνονται σαν [NOISE]. Θα πρέπει να σημειώσουμε πως αυτή η περίπτωση έχει να κάνει μόνο για στιγμιαίους θορύβους και όχι για συνθήκη ομιλίας όπου αναφέραμε παραπάνω. Μερικές περιπτώσεις στιγμιαίων θορύβων είναι οι εξής;

- *side_speech*

Σημασία:
Ομιλία μικρής διάρκειας από άλλον ομιλητή

- *phone_ring*

Σημασία:
Κουδούνισμα τηλεφώνου

- *clear_throat*

Σημασία:
Όταν ο ομιλητής καθαρίζει τον λαιμό του

- *paper_rustle*

Σημασία:
Θόρυβος από χαρτιά

- *paff_noise*

Σημασία:
Όταν ο ομιλητής μιλάει πολύ κοντά στο μικρόφωνο

8. Αναπνοή

Στην περίπτωση που ο ομιλητής εισπνεύσει ή εκπνεύσει και γίνει ηχητικά αντιληπτό, τότε στην απομαγνητοφώνηση γράφουμε [BREATH].

9. Άλλες περιπτώσεις

Σε περίπτωση που μια λέξη ή ένα μέρος της πρότασης δεν διακρίνεται καθαρά λόγω κακής ποιότητας του σήματος φωνής ή αδυναμία του ομιλητή τότε γράφουμε με όμοιο τρόπο με τις παραπάνω περιπτώσεις [TAG_BAD_READING].

Επίσης στην περίπτωση που έχουμε συνθήκη ομιλίας non-Greek είναι αυτονόητο πως δεν γίνεται καμία απομαγνητοφώνηση.

Τελικά αφού έγιναν οι απομαγνητοφωνήσεις 20 μονόωρων δελτίων ειδήσεων, το εργαλείο Transcriber Tool βγάζει σαν έξοδο ένα .trs αρχείο που έχει την μορφή ενός xml αρχείου. Η μορφή του έχει ως εξής:

```
<?xml version="1.0" encoding="ISO-8859-7"?>
<!DOCTYPE Trans SYSTEM "trans-14.dtd">
<Trans scribe="orfeas" audio_filename="" version="21"
version_date="061215">
<Speakers>  ---------------->     //list of speakers
<Speaker id="spk1" name="Xoukli" check="no" type="female"
dialect="native" accent="" scope="local"/>
<Speaker id="spk2" name="Papandreou" check="no" type="male"
dialect="native" accent="" scope="local"/>
<Speaker id="spk3" name="Karamanlis" check="no" type="male"
dialect="native" accent="" scope="local"/>
<Speaker id="spk4" name="Alogoskoufis" check="no" type="male"
dialect="native" accent="" scope="local"/>
<Speaker id="spk5" name="Bakogianni" check="no" type="female"
dialect="native" accent="" scope="local"/>
<Speaker id="spk6" name="Alabanos" check="no" type="male"
dialect="native" accent="" scope="local"/>

                              •

                              •

                              •
</Speakers>

<Topics> ----------------------------> //list of condition level choices
<Topic id="to1" desc="bg_noise"/>
<Topic id="to2" desc="bg_music"/>
<Topic id="to3" desc="NonGreek"/>
</Topics>

<Episode>
<Section type="nontrans" startTime="0" endTime="269.144">
<Turn startTime="0" endTime="269.144">
<Sync time="0"/>
καλησπέρα σας κυρίες και κύριοι [BREATH] ----------->  //transcription
</Turn>
</Section>
<Section type="report" topic="to2" startTime="269.144"
endTime="270.375">
<Turn speaker="spk21" mode="planned" channel="studio"
startTime="269.144" endTime="270.375">
<Sync time="269.144"/>
ας δούμε τα θέματα που μας απασχολήσουν *σήμερα* -->  //transcription
</Turn>

                              •

                              •

                              •
```

Σχήμα 2.1 Μορφή εξόδου του Transcriber Tool

Επιπλέον με μια επιλογή που έχει το συγκεκριμένο πρόγραμμα ‘κόβουμε’ το μεγάλο σήμα φωνής σε μικρότερα, διάρκειας μέχρι μερικών δευτερολέπτων έτσι ώστε να τα επεξεργαστούμε μετέπειτα.

Εφόσον πλέον έχουμε 20 ώρες για απομαγνητοφώνηση, έχουμε και 20 .trs αρχεία που είναι οι έξοδοι του Transcriber Tool. Έτσι όπως δείχνουμε και στο σχήμα 2.1, έχουμε όλη την πληροφορία που χρειαζόμαστε για να επεξεργαστούμε τα δεδομένα μας.

Σχεδιάζοντας ένα perl αρχείο, το *parse.pl*, επεξεργαζόμαστε την κάθε έξοδο του Transcriber Tool και τις φέρνουμε στην μορφή που εμείς επιθυμούμε. Έτσι με την επεξεργασία αυτών των xml αρχείων, έχουμε την εξής μορφοποίηση:

Waveform	**Condition**	**Speaker**	**Sex**	**Dialect**
060625NETED2100_001.wav	report	Xoukli	Female	native
060625NETED2100_002.wav	bg_noise	rep_01	Male	non-native
060625NETED2100_003.wav	bg_music	Alabanos	Male	native
⋮	⋮	⋮	⋮	⋮

Waveform	**Channel**	**Mode**	**Transcription**
060625NETED2100_001.wav	studio	planned	καλησπέρα κυρίες και κύριοι
060625NETED2100_002.wav	telephone	spontaneous	γεια σας από το σεφ
060625NETED2100_003.wav	studio	planned	η πολιτική μας είναι σταθερή
⋮	⋮	⋮	⋮

Σχήμα 2.2 Μορφή εξόδου του αρχείου parse.pl

2.4 Ολοκλήρωση της βάσης δεδομένων

Σε αυτό το σημείο, εφόσον έχουμε τα δεδομένα στην μορφή που επιθυμούμε, μπορούμε να παρουσιάσουμε πόσα και ποια δεδομένα έχουμε εν τέλει στην κατοχή μας. Για την πιο εύκολη και πιο οργανωμένη παρουσίαση της βάσης μας θα ορίσουμε 9 κατηγορίες δεδομένων και στη συνέχεια θα δείξουμε την

ποσότητα των δεδομένων σε κάθε κατηγορία. Οι 9 αυτές κατηγορίες είναι οι παρακάτω:

F0 **-** δεδομένα που έχουν σαν συνθήκη ομιλίας **music** (condition = music)

F1 **-** δεδομένα που έχουν σαν συνθήκη ομιλίας **noise** και η ομιλία είναι καθαρή (condition = noise && channel = studio)

F2 **-** δεδομένα που έχουν σαν συνθήκη ομιλίας **noise** και η ομιλία είναι κακής ποιότητας (condition = noise && channel = telephone)

F3 **-** σήματα που ομιλούν πολλοί ομιλητές ταυτόχρονα

F4 **-** ακουστικά σήματα που δεν περιέχουν ελληνική ομιλία

F5 **-** περιέχουν ελληνική ομιλία από μη αυτόχθονες (dialect = non-native)

F6 **-** δεδομένα που έχουν σαν συνθήκη ομιλίας **studio** και είναι προκαθορισμένη ομιλία (condition = studio && mode =planned)

F7 **-** σήματα που έχουν σαν συνθήκη ομιλίας **studio** και είναι αυθόρμητη ομιλία (condition = studio && mode =spontaneous)

F8 **-** σήματα που περιέχουν ομιλία κακής ποιότητας (channel = telephone)

Ορίζοντας τις κατηγορίες μας, τώρα είμαστε σε θέση να παρουσιάσουμε πόσα δεδομένα έχουμε σε κάθε κατηγορία σε επίπεδο αριθμών προτάσεων και σε επίπεδο χρονικής διάρκειας.

	Αριθμός προτάσεων	Χρονική διάρκεια σε ώρες
F0	417	1.23
F1	3366	9.94
F2	345	1.01
F3	82	0.24
F4	22	0.06
F5	52	0.15
F6	1306	3.85
F7	470	1.38
F8	712	2.10

Πίνακας 2-2 Ποσοτική περιγραφή των δεδομένων κάθε κατηγορίας

ΚΕΦΑΛΑΙΟ 3

ΓΛΩΣΣΙΚΟ ΜΟΝΤΕΛΟ

3.1 Γενικά

Τα γλωσσικά μοντέλα υπολογίζουν την πιθανότητα εμφάνισης μιας λέξης, έχοντας παράλληλα πληροφορία για το περιεχόμενο στο οποίο αναμένεται να βρεθεί η λέξη. Η σημασία των γλωσσικών μοντέλων είναι καθοριστική στην αναγνώριση φωνής, στην μετάφραση μηχανής και στα συστήματα επεξεργασίας φυσικού λόγου (natural language systems).

Όπως αναφέραμε στον τύπο 1- 3.1

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

εργασία του γλωσσικού μοντέλου είναι να υπολογίζει την κατανομή πιθανότητας για τον όρο $P(W)$.

Τα στατιστικά γλωσσικά μοντέλα παρέχουν την κατανομή πιθανότητας, βασιζόμενα σε στατιστικά στοιχεία που έχουν συγκεντρωθεί από ένα αρκετά μεγάλο κείμενο εκπαίδευσης (training text). Για παράδειγμα, ένα μηδενικής τάξης μοντέλο μπορεί να αναθέσει πιθανότητα στην λέξη ‘και’ πιο πολλές φορές από κάθε άλλη λέξη σε ένα ελληνικό κείμενο. Ένα στατιστικό γλωσσικό μοντέλο μπορεί να χρησιμοποιήσει αυτή την πληροφορία όταν κάνει προβλέψεις.

3.2 N-gram μοντέλα

Τα γλωσσικά μοντέλα υπολογίζουν την πιθανότητα μιας ακολουθίας λέξεων, $\hat{P}(w_1, w_2, \ldots, w_m)$, υπολογίζοντας έτσι την πιθανότητα $P(w)$. Η πιθανότητα

$\hat{P}(w_1, w_2, \ldots, w_m)$ μπορεί να αναλυθεί ως γινόμενο από συνδυαστικές πιθανότητες:

$$\hat{P}(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} \hat{P}(w_{i-n+1} \mid w_1, \ldots, w_{i-1}) \qquad \text{για n} \geq 1 \qquad (3.2)$$

Για λόγους έλλειψης δεδομένων, οι συνήθεις τιμές για το n είναι 1 με 4. Μοντέλα που έχουν εύστοχο αλλά περιορισμένο περιεχόμενο σαν και αυτά συνήθως λέγονται **n-gram γλωσσικά μοντέλα**. Τέτοια μοντέλα υποθέτουν ότι η ακολουθία των λέξεων έχει την Μαρκοβιανή ιδιότητα και χρησιμοποιούν τις n -1 λέξεις σαν ιστορία (history) $\phi(h_i) = \{w_i \mid w_{i-n+1}, \ldots, w_{i-1}\}$. Για n=1 το μοντέλο μας χαρακτηρίζεται ως unigram, για n=2 bigram όπως στην περίπτωση μας, και για n=3 trigram. Οι πιθανότητες υπολογίζονται μέσω συχνοτήτων λέξεων στο κείμενο προς εκπαίδευση.

$$\hat{P}(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_i)}{C(w_{i-n+1}, \ldots, w_{i-1})} \qquad (3.3)$$

όπου $C(.)$ είναι ο αριθμός εμφανίσεων μιας συγκεκριμένης ακολουθίας λέξεων στο κείμενο.

Η επιλογή του n είναι καθοριστικής σημασίας, καθώς από αυτό εξαρτάται ο αριθμός των συνδυασμών λέξεων που μπορεί να έχει το μοντέλο μας και το όριο του είναι $|W|^n$, όπου W είναι το σετ από τις λέξεις του γλωσσικού μοντέλου, γνωστό ως λεξικό (vocabulary). Για παράδειγμα, ένα 4-gram μοντέλο με ένα λεξικό των 65000 λέξεων έχει 65000^4 πιθανούς συνδυασμούς λέξεων. Παρόλο αυτά, οι δυνατοί συνδυασμοί λέξεων που εμφανίζονται στο κείμενο εκπαίδευσης είναι αρκετά πιο λίγοι από τον αριθμό που μόλις αναφέραμε, οπότε οι απαιτήσεις σε χώρο είναι πολύ μικρότερες από την θεωρητική εκτίμηση. Ακόμα και μετά από αυτή την μείωση χώρου αποθήκευσης και παρόλο το μεγάλο μέγεθος του κειμένου εκπαίδευσης, υπάρχουν ακόμα πολλοί συνδυασμοί ή ακολουθίες λέξεων που δεν υπάρχουν στο κείμενο εκπαίδευσης και αν υπάρχουν, θα βρεθούν στατιστικά ελάχιστες φορές.

Μεγάλο μέγεθος κειμένου εκπαίδευσης είναι επιθυμητό έτσι ώστε να έχουμε στατιστικά σημαντικές πιθανότητες από ακολουθίες λέξεων. Αυξάνοντας το μέγεθος του κειμένου δίνει μεγαλύτερο βαθμό εμπιστοσύνης στις εκτιμήσεις του γλωσσικού μας μοντέλου, αλλά όμως απαιτεί μεγαλύτερο χώρο αποθήκευσης για το ίδιο το κείμενο και μεγαλύτερο χρόνο ανάλυσης όταν υπολογίζουμε τις παραμέτρους του συστήματος. Αυτό παίζει ιδιαίτερο ρόλο στην σχεδίαση του γλωσσικού μοντέλου, και καθορίζει την ποσότητα των δεδομένων που θα χρησιμοποιηθούν για την κατασκευή του τελικού μας μοντέλου. Επιπρόσθετα, αν το γλωσσικό μοντέλο χρησιμοποιείται για αναγνώριση φωνής, όπως στην περίπτωση μας, τότε δόκιμο είναι να εκπαιδευτεί σε ακριβείς ακουστικές απομαγνητοφωνήσεις για την δημιουργία ενός αξιόπιστου συστήματος αναγνώρισης.

3.3 Bigram γλωσσικό μοντέλο

Ένα από τα πιο πετυχημένα στατιστικά γλωσσικά μοντέλα είναι το bigram γλωσσικό μοντέλο. Σε αυτό το μοντέλο, η πιθανότητα μιας λέξης υπολογίζεται δημιουργώντας ένα Markov μοντέλο, του οποίου η κάθε κατάσταση του βασίζεται στην παρουσία της προηγούμενης λέξης. Για παράδειγμα, σε ένα απλό bigram μοντέλο, η πιθανότητα μιας λέξης w δεδομένης της προηγούμενης λέξης της δίνεται από τον τύπο:

$$P(w \mid w_{-1}) = \frac{c(w_{-1}w)}{c(w_{-1})} \tag{3.4}$$

Η στρατηγική όταν υλοποιούμε ένα μοντέλο bigram, είναι να ενσωματώνουμε στο μοντέλο μας και unigram πιθανότητες σε περιπτώσεις που δεν τις καλύπτει το bigram. Η ερώτηση όμως που τίθεται είναι πως ακριβώς ενσωματώνουμε δύο πηγές πληροφορίας στο μοντέλο μας; Μια επιλογή είναι να εφαρμόσουμε ένα γραμμικό interpolation των δύο κατανομών πιθανότητας. Έτσι η πιθανότητα μιας λέξης w δεδομένης της προηγούμενης λέξης της δίνεται από τον τύπο:

$$P(w \mid w_{-1}) = \lambda_0 + \lambda_1 \frac{c(w)}{N} + \lambda_2 \frac{c(w_{-1}w)}{c(w_{-1})} \quad (3.5)$$

Το N συμβολίζει τον αριθμό των λέξεων στο κείμενο εκπαίδευσης, το $c(.)$ τον αριθμό εμφανίσεων στο κείμενο εκπαίδευσης. Τα λ είναι παράμετροι που υπολογίζονται με την βοήθεια ενός άλλου κειμένου, που ουσιαστικά χρησιμοποιείται για να τεστάρουμε το γλωσσικό μας μοντέλο έτσι ώστε να αποδίδει ικανοποιητικά. Για να είναι η σχέση (3.5) μια έγκυρη κατανομή πιθανότητας, θα πρέπει τα βάρη $\lambda_0, \lambda_1, \lambda_2$ να είναι μη αρνητικά και να αθροίζουν στην μονάδα ($\lambda_0 + \lambda_1 + \lambda_2 = 1$). Τέλος μπορούμε να πούμε πως το bigram μοντέλο είναι ένα αξιόπιστο μοντέλο, με καλές αποδόσεις και η σχεδίαση και η δημιουργία του είναι αρκετά εύκολη.

3.4 Perplexity

Ιδανικά, για να βλέπαμε την απόδοση του γλωσσικού μας μοντέλου θα μπορούσαμε να παρακολουθήσουμε π.χ. την μείωση του λάθους αναγνώρισης σε ένα αντίστοιχο σύστημα. Βέβαια η παραπάνω τακτική δεν είναι εφικτή σε κάθε περίπτωση. Για αυτό το λόγο, ένα μέγεθος που χρησιμοποιείται αρκετά συχνά έτσι ώστε να βλέπουμε πόσο αποδοτικό είναι το γλωσσικό μας μοντέλο είναι το *perplexity.* Το perplexity ορίζεται ως:

$$PP(T) = 2^{H(\hat{P}(T), P(T))} \quad (3.6)$$

Ο όρος $H(\hat{P}(T), P(T))$, δηλαδή η εντροπία $\hat{P}(T)$ που είναι η παρατηρούμενη κατανομή του κειμένου T , σε συνδυασμό με το $P(T)$ που είναι η υπολογιζόμενη κατανομή του κειμένου T από το γλωσσικό μοντέλο μας, ορίζεται ως:

$$H(\hat{P}(T), P(T)) = -\sum_{x \in T} \hat{P}(x) \log P(x) \quad (3.7)$$

Παρόλο που μικρές τιμές perplexity δεν σημαίνουν απαραίτητα μεγαλύτερο σφάλμα στην αναγνώριση μας, το perplexity παραμένει ένα αξιόπιστο μέτρο

σύγκρισης μεταξύ διαφορετικών μοντέλων όταν δοκιμάζονται σε παρόμοια ακουστικά σήματα.

3.5 Το γλωσσικό μοντέλο του συστήματος μας

Όπως έχουμε ήδη αναφέρει, το σύστημα αναγνώρισης φωνής που θα υλοποιήσουμε θα πρέπει να αναγνωρίζει με επιτυχία ακουστικά σήματα ηχογραφημένα από τηλεοπτικές εκπομπές. Συνεπώς το κείμενο για να εκπαιδεύσουμε το γλωσσικό μας μοντέλο θα πρέπει να περιέχει αντίστοιχο υλικό όπως κείμενα από εφημερίδες κλπ.

Οπότε για να υλοποιήσουμε ένα τέτοιο μοντέλο χρησιμοποιήσαμε κείμενο από 3 διαφορετικές εφημερίδες. Με την βοήθεια της γλώσσας προγραμματισμού Perl κατεβάσαμε από το Internet κείμενο από την εφημερίδα «ΤΑ ΝΕΑ» και «ΤΟ ΒΗΜΑ» έτσι ώστε να το επεξεργαστούμε κατάλληλα στην συνέχεια. Συγκεκριμένα έχουμε στην κατοχή μας:

- 215Mb κείμενο από την εφημερίδα «Ελευθεροτυπία»
- 170Mb κείμενο από την εφημερίδα «ΤΑ ΝΕΑ»
- 65Mb κείμενο από την εφημερίδα «ΤΟ ΒΗΜΑ»

Για να έχουμε την δυνατότητα να δημιουργήσουμε το γλωσσικό μας μοντέλο πρέπει αρχικά να 'καθαρίσουμε' το κείμενο 450Mb συνολικά με scripts που υλοποιήθηκαν πάλι με Perl. Συγκεκριμένα θα πρέπει να ακολουθήσουμε κάποιους κανόνες-οδηγίες που χρησιμοποιήθηκαν και στην απομαγνητοφώνηση των ακουστικών σημάτων. Δηλαδή:

1. **Ορθογραφία**

   Πρέπει στο κείμενο να μην υπάρχουν κεφαλαίοι ελληνικοί χαρακτήρες και για αυτό τα μετατρέπουμε σε μικρά. Επιπλέον δεν θα πρέπει να υπάρχουν σημεία στίξης και έτσι όποτε βρίσκουμε σημείο στίξης απλά αλλάζουμε γραμμή στο κείμενο

Π.χ.

Η ΑΕΚ είναι καλή ομάδα (Λάθος)
η άεκ είναι καλή ομάδα (Σωστό)

Ο κύριος Αλοφοσκούφης αναχώρησε σήμερα για Ολλανδία. Εκεί… (Λάθος)
ο κύριος αλογοσκούφης αναχώρησε σήμερα για ολλανδία(\n)
εκεί… (Σωστό)

2. Ακρωνύμια , Συντομεύσεις κτλ.

Στην περίπτωση ακρονυμίων , συντομεύσεων κτλ. ενεργούμε παρομοίως. Δηλαδή όταν βρίσκουμε συντομεύσεις όπως π.χ. , μ. , κ. ,θα πρέπει να γράφουμε ολόκληρη την λέξη.

Π.χ.

Ο κ. Καραμανλής δήλωσε ότι θα γίνουν αυξήσεις. (Λάθος)
ο κύριος καραμανλής δήλωσε ότι θα γίνουν αυξήσεις (Σωστό)

Ο Κώστας Κεντέρης ήταν νικητής στην κούρσα των διακοσίων μ. (Λάθος)
Ο κώστας κεντέρης ήταν νικητής στην κούρσα των διακοσίων μέτρων (Σωστό)

3. Αριθμοί

Παρομοίως, όπως στις παραπάνω περιπτώσεις.

Π.χ.

30 μ. δίπλα από το πτώμα… (Λάθος)
τριάντα μέτρα δίπλα από το πτώμα… (Σωστό)

Τώρα πλέον έχουμε το κείμενο προς εκπαίδευση στην μορφή που επιθυμούμε. Με ειδικά κατασκευασμένα εργαλεία της SRILM (http://www.speech.sri.com/projects/srilm/) κατασκευάζουμε αρχικά το bigram γλωσσικό μας μοντέλο. Το λεξικό του μοντέλου μας περιέχει 60.000 λέξεις. Οι προφορές στο λεξικό μας παράχθηκαν από ένα ειδικό script που υλοποιήθηκε με την βοήθεια γλωσσολόγου, το words2phones.

Θα πρέπει να σημειώσουμε πως εφαρμόσαμε Kneser-Ney smoothing στο γλωσσικό μας μοντέλο. Το Kneser-Ney smoothing είναι μια από τις πολλές back-off τεχνικές. Τέτοιου είδους τεχνικές χρησιμοποιούνται στην περίπτωση που μια ακολουθία λέξεων, 2 λέξεων στην περίπτωση του bigram μοντέλου, δεν βρεθεί στο γλωσσικό μοντέλο και για αυτό το λόγο το μοντέλο μας ανατρέχει πίσω (back-off) στις unigram πιθανότητες. Συνεπώς η Kneser-Ney τεχνική προτείνει το ακόλουθο back-off:

$$p_{KN}(w \mid v) = \frac{\max(0, C(v,w) - D) + Da(v,w)}{\sum_{w} C(v,w)} \tag{3.8}$$

ο back-off παράγοντας είναι

$$a(v,w) = \frac{N(v,\cdot)N(\cdot,w)}{N(\cdot,\cdot)} \tag{3.9}$$

όπου v, w δύο λέξεις από το λεξικό του μοντέλου μας, $D = 1$, $C(v,w)$ είναι ο αριθμός εμφανίσεων των δύο λέξεων, $N(v,\cdot)$ ο αριθμός των bigrams που αρχίζουν με την λέξη v, $N(\cdot,w)$ ο αριθμός των bigrams που τελειώνουν με την λέξη w.

Στη συνέχεια, με την βοήθεια του εργαλείου HTK (Hidden Markov Model Toolkit) μετατρέπουμε το κατασκευασμένο γλωσσικό μοντέλο μας στο format του HTK. Αυτό γίνεται διότι, όπως θα περιγράψουμε και παρακάτω, το σύστημα αναγνώρισης μας θα υλοποιηθεί εξ' ολοκλήρου με το παραπάνω εργαλείο. Για περισσότερες πληροφορίες μπορείτε να ανατρέξετε στο παράρτημα Α.

Τέλος, για να μετρήσουμε με την βοήθεια του HTK την ποιότητα του γλωσσικού μας μοντέλου, θα μετρήσουμε το perplexity του μοντέλου μας καθώς και το ρυθμό λέξεων εκτός λεξιλογίου (OOV - out of vocabulary words). Το OOV μας δείχνει δεδομένου ενός αρχείου δοκιμής, πόσες άγνωστες λέξεις έχει το γλωσσικό μοντέλο. Για να μπορέσουμε να έχουμε συγκρίσιμα μεγέθη

θα υπολογίσουμε τις τιμές αυτών των δύο μεγεθών, για 3 διαφορετικά γλωσσικά μοντέλα:

1) για μοντέλο που δημιουργήθηκε βάση των κειμένων που περιέχουν υλικό της εφημερίδας «Ελευθεροτυπία» **(ΓΜ1)**

2) για μοντέλο που δημιουργήθηκε βάση των κειμένων που περιέχουν υλικό των εφημερίδων «Ελευθεροτυπία» και «ΤΟ ΒΗΜΑ» **(ΓΜ2)**

3) για μοντέλο που δημιουργήθηκε βάση των κειμένων που περιέχουν υλικό των εφημερίδων «Ελευθεροτυπία» και «ΤΟ ΒΗΜΑ» και «ΤΑ ΝΕΑ», που είναι και το γλωσσικό μοντέλο του συστήματος μας **(ΓΜ3)**

Perplexity

ΓΜ#	Perplexity
ΓΜ1	243.5364
ΓΜ2	234.0310
ΓΜ3	211.4075

Πίνακας 3-1 Perplexities τριών γλωσσικών μοντέλων

OOV(%)

ΓΜ#	OOV(%)
ΓΜ1	6.05
ΓΜ2	5.32
ΓΜ3	5.21

Πίνακας 3-2 OOV(%) τριών γλωσσικών μοντέλων

ΚΕΦΑΛΑΙΟ 4

ΑΚΟΥΣΤΙΚΑ ΜΟΝΤΕΛΑ

4.1 Γενικά

Σε αυτό το κεφάλαιο θα περιγράψουμε την διαδικασία δημιουργίας διαφόρων ακουστικών μοντέλων με την βοήθεια του εργαλείου HTK. Όπως ήδη έχουμε αναφέρει θα χρησιμοποιήσουμε Hidden Markov Models για να μοντελοποιήσουμε την ανθρώπινη ομιλία και συνεπώς για να υλοποιήσουμε τα ακουστικά μας μοντέλα. Μπορείτε να ανατρέξετε στο **Κεφάλαιο 1** για την δομή και τα είδη των HMMs. Αρχικά θα αναφερθούμε σε ποια γλωσσική μονάδα βασίζονται τα HMMs και στη συνέχεια θα κάνουμε μια μικρή ανασκόπηση στην τοπολογία των HMM.

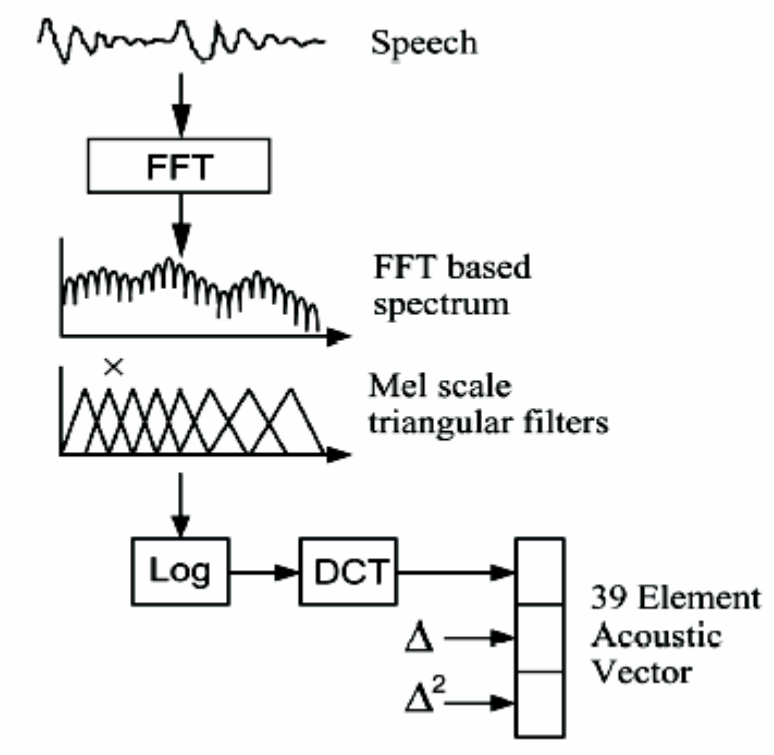
4.2 Σχεδιαστικές επιλογές ακουστικών μοντέλων

4.2.1 Επεξεργασία Front-End

Στο **υποκεφάλαιο 1.2** αναφέραμε πως η αναγνώριση ξεκινά με το ψηφιακοποιημένο σήμα ομιλίας, το οποίο υπόκειται κατόπιν σε προεπεξεργασία (front-end), μέσα από ποικίλα βήματα, φασματικής συνήθως, επεξεργασίας σήματος. Επίσης αναφέραμε πως μία αρκετά διαδεδομένη μέθοδος επεξεργασίας είναι η εξαγωγή Mel-Frequency Cepstral Coefficients (MFFC). Στην περίπτωση της Mel-Scale cepstral ανάλυσης, χρησιμοποιείται μια μη γραμμική κλίμακα, που ονομάζεται Mel κλίμακα, που μιμείται και αναπαριστά το ακουστικό εύρος της ανθρώπινης ακοής. Η Mel κλίμακα μπορεί να προσεγγιστεί από τον ακόλουθο τύπο:

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \qquad (4.1)$$

Η διαδικασία εξαγωγής διανυσματικών ακολουθιών βασιζόμενα στην Mel-frequency παρουσιάζεται στο σχήμα 4.1.



Σχήμα 4-1 Mel-Frequency Cepstral Coefficients

Αρχικά το σήμα μετασχηματίζεται σε φάσμα μέσω ενός μετασχηματισμού Fourier. Έπειτα, το παραγόμενο φάσμα του σήματος φωνής εξομαλύνεται, περνώντας τους φασματικούς συντελεστές από τριγωνοειδείς συχνότητες που καθορίζονται από την Mel –frequency. Στην συνέχεια, η έξοδος αυτού του φίλτρου περνάει από λογαριθμική συμπίεση έτσι ώστε το ενεργειακό φάσμα να γίνεται Γκαουσιανό. Τελικά, στο τελευταίο στάδιο της επεξεργασίας εφαρμόζεται διακριτός μετασχηματισμός συνιμητόνου (discrete consine transform - DCT). Συνηθίζεται στις τελικές διανυσματικές ακολουθίες που παράγονται να προσθέτονται συντελεστές παραγώγου πρώτης και δεύτερης τάξης και κάποιες φορές να συμπεριλαμβάνεται και μια μέτρηση της ενέργειας του σήματος φωνής.

Στην περίπτωση μας, για να προχωρήσουμε στην εκπαίδευση των ακουστικών μας μοντέλων θα πρέπει πρώτα να εφαρμόσουμε την παραπάνω διαδικασία στα αρχεία ήχου που έχουμε στην κατοχή μας. Έτσι, με την βοήθεια

του εργαλείου του HTK, και συγκεκριμένα με την εντολή *HCopy* μετατρέψαμε τα σήματα φωνής σε διανυσματικές ακολουθίες. Το configuration file της εντολής, που καθορίζει το είδος των MFCCs που θα παραχθούν, είναι το εξής:

```
SOURCEFORMAT = WAV        # format of wav files
TARGETKIND = MFCC_E_D_A_Z # C0 + Deltas + Deltas Deltas
                             (acceleration) + Cepstral Mean
                              Normalization
TARGETRATE = 100000   # frame period 10ms (HTK uses 100ns unit)
WINDOWSIZE = 250000   # windows size 25ms
ZMEANSOURCE = TRUE    # zero mean source waveform (removes DC)
PREEMFCOEF = 0.97     # pre-emfasis coefficient
USEHAMMING = TRUE     # use Hamming window
NUMCHANS = 26         # number of filterbank channels
CEPLIFTER = 22        # cepstral liftering coefficient
NUMCEPS = 12          # number of cepstral coefficients
SAVECOMPRESSED = TRUE

ENORMALIZE = TRUE     # perform energy normalization
```

Σχήμα 4-2 Configuration File της HCopy

Έτσι, όπως φαίνεται και στο σχήμα 4-3, δημιουργήθηκαν MFCCs για κάθε σήμα φωνής. Το κάθε MFCC περιέχει 39 στοιχεία, ανάμεσα σε αυτά συντελεστές παραγώγου πρώτης και δεύτερης τάξης , μετρήσεις της ενέργειας του σήματος φωνής. Τα MFCC υπολογίσθηκαν ανά 10 ms του σήματος φωνής χρησιμοποιώντας παράθυρο πλάτους 25ms.

4.2.2 Επιλογή γλωσσικής μονάδας

Όταν αποφασίζουμε να χρησιμοποιήσουμε HMMs για να μοντελοποιήσουμε την ανθρώπινη ομιλία, δημιουργείται ένα σημαντικό ερώτημα: ποια γλωσσική μονάδα θα χρησιμοποιηθεί προς μοντελοποίηση; Υπάρχουν αρκετές επιλογές, μερικές από τις οποίες είναι λέξεις, συλλαβές, φωνήματα κλπ. Καθεμία από αυτές τις επιλογές έχει πλεονεκτήματα αλλά και μειονεκτήματα. Για να επιλέξουμε κατάλληλη γλωσσική μονάδα θα πρέπει να λάβουμε υπ' όψει τα παρακάτω κριτήρια:

- Η μονάδα θα πρέπει να είναι *ακριβής* (accurate) στην ακουστική αναπαράσταση σε διαφορετικά περιεχόμενα.
- Η μονάδα θα πρέπει να είναι *εκπαιδεύσιμη* (trainable). Θα πρέπει να υπάρχουν αρκετά δεδομένα για εκπαίδευση (training data) έτσι ώστε να υπολογιστούν σωστά οι παράμετροι της κάθε μονάδας.
- Η μονάδα θα πρέπει να είναι *γενικεύσιμη* (generalizable) έτσι ώστε να μπορούν να παραχθούν καινούργιες λέξεις.

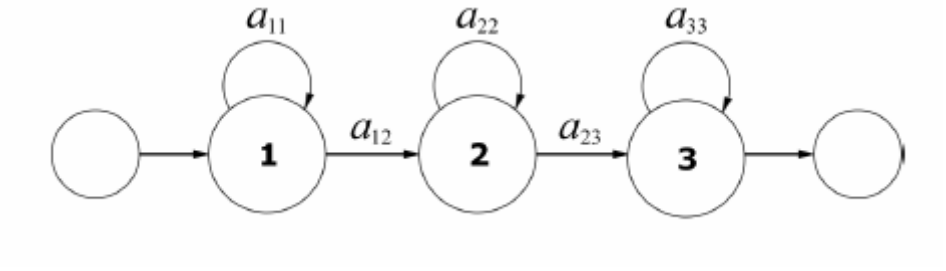
Η πιο φυσική επιλογή είναι να επιλέξουμε να εκπαιδεύσουμε μοντέλα για ολόκληρες λέξεις. Αυτά τα μοντέλα εφόσον εκπαιδευτούν σωστά και εφόσον χρησιμοποιηθούν για συστήματα αναγνώρισης μικρού λεξιλογίου (small-vocabulary recognition systems), έχουν πολύ καλή επίδοση σε σχέση με άλλες γλωσσικές μονάδες. Τα μοντέλα λέξεων πληρούν τα 2 πρώτα κριτήρια, δηλαδή την *ακρίβεια* και την *εκπαιδευσιμότητα* και επιπλέον δεν υπάρχει λόγος να είναι *γενικεύσιμα*. Παρόλα αυτά, για συστήματα αναγνώρισης μεγάλου λεξιλογίου (large-vocabulary recognition systems) η επιλογή των λεξικών μοντέλων λέξεων είναι κακή. Δεν υπάρχει τρόπος για να δημιουργηθούν καινούργιες λέξεις, εφόσον έχουμε ένα καθορισμένο set λέξεων, κάνοντας έτσι τα μοντέλα λέξεων μη-γενικεύσιμα. Επιπλέον κάθε λέξη χρειάζεται να εκπαιδευτεί ξεχωριστά και έτσι απαιτούνται πολλά δεδομένα εκπαίδευσης για να εκπαιδεύσουμε κάθε μονάδα ξεχωριστά.

Ένας εναλλακτικός τρόπος είναι να εκπαιδεύσουμε μοντέλα για κάθε φώνημα. Οι τυπικές ευρωπαϊκές γλώσσες έχουν 40 με 50 διαφορετικά φωνήματα. Ακουστικά μοντέλα που έχουν σαν βάση τους φωνήματα μπορούν να εκπαιδευτούν ικανοποιητικά μόλις με μερικές εκατοντάδες προτάσεις εξασφαλίζοντας έτσι το κριτήριο της *εκπαιδευσιμότητας*. Αυτά τα μοντέλα είναι εξ΄ ορισμού *γενικεύσιμα* και αυτό γιατί τα φωνήματα είναι η βασική μονάδα από τα οποία προέρχονται οι λέξεις. Η *ακρίβεια* είναι ένα σημαντικό θέμα στην περίπτωση των φωνημάτων, διότι κάθε φώνημα εξαρτάται και συγχρόνως επηρεάζεται από το γειτονικό δεξί και αριστερό φώνημα.

Τα φωνητικά μοντέλα που βασίζονται σε φωνήματα μπορούν να γίνουν σημαντικά πιο ακριβή εάν εκπαιδευτούν, παίρνοντας υπ' όψει τα γειτονικά του φωνήματα. Έτσι ενώ πριν είχαμε ξεχωριστό μοντέλο για κάθε φώνημα, δηλαδή μονόφωνα (monophones), τώρα έχουμε πολλαπλά διαφορετικά μοντέλα ανάλογα με την ταυτότητα των γειτονικών φωνημάτων. Έτσι έχουμε την δημιουργία των διφώνων (biphones) και τριφώνων (triphones). Δυστυχώς όμως στην περίπτωση των triphones, η εκπαιδευσιμότητα είναι ένα πρόβλημα γιατί μπορούμε να έχουμε 50 x 50 x 50 = 125,000 triphones για εκπαίδευση. Στην πράξη οι περισσότεροι συνδυασμοί φωνημάτων δεν εμφανίζονται και έτσι έχουμε 10.000 – 20.000 triphones σε συστήματα μεγάλου λεξιλογίου

4.2.3 Τοπολογία HMMs μοντέλων

Η ομιλία είναι ένα σήμα που αλλάζει συνεχώς κατά την διάρκεια του χρόνου. Κάθε κατάσταση ενός HMM έχει την δυνατότητα να ''αιχμαλωτίζει'' κάποια σταθερά τμήματα σε ένα μη-σταθερό σήμα φωνής. Η πιο λογική επιλογή για να μοντελοποιήσουμε ένα σήμα φωνής, είναι μια αριστερά προς τα δεξιά (left-to-right) τοπολογία. Το σχήμα 4.3 δείχνει ένα τυπικό HMM 5 καταστάσεων που είναι κοινό σε πολλά συστήματα αναγνώρισης. Η πρώτη και η τελευταία κατάσταση δεν έχουν συνάρτηση εξόδου και δεν παράγουν παρατηρήσεις για αυτό και λέγονται null-states. Ο σκοπός τους είναι απλά να συνδέουν διαφορετικά μοντέλα.



Σχήμα 4.3 Βασική δομή ενός φωνητικού HMM

Ο αριθμός των εσωτερικών καταστάσεων ενός HMM μπορεί να διαφέρει ανάλογα με την γλωσσική μονάδα που αναπαριστούμε. Για HMMs που αναπαριστούν φώνημα, χρησιμοποιούνται συνήθως 3 με πέντε καταστάσεις. Εάν το HMM αναπαριστά λέξη, απαιτείται ένας σημαντικά μεγαλύτερος αριθμός καταστάσεων. Μπορεί να έχουμε 15 με 25 καταστάσεις, ανάλογα με την προφορά και τη διάρκεια της λέξης. Επίσης είναι πιθανό να έχουμε πιο πολύπλοκες μεταβάσεις από κατάσταση σε κατάσταση από αυτές του σχήματος 4.3. Για παράδειγμα, υπάρχουν περιπτώσεις που οι μεταβάσεις δεν γίνονται σειριακά, δηλαδή μπορεί να γίνει μετάβαση από την κατάσταση 1 στην κατάσταση 3. Έτσι το μοντέλο γίνεται πιο ευέλικτο, αλλά ταυτόχρονα και πιο δύσκολο για να εκπαιδευτεί σωστά.

Η επιλογή της πιθανότητας εξόδου $b_j(x)$ είναι ζωτικής σημασίας για ένα καλό σύστημα αναγνώρισης. Τα πρώτα HMM συστήματα χρησιμοποιούσαν διακριτή πιθανότητα εξόδου σε συνδυασμό με κβαντοποίηση διανυσμάτων. Η κβαντοποίηση διανυσμάτων είναι υπολογιστικά ικανοποιητική, αλλά όμως εισάγει θόρυβο περιορίζοντας έτσι την ακρίβεια που είναι δυνατόν να επιτευχθεί. Τα πιο σύγχρονα συστήματα χρησιμοποιούν παραμετρικές κατανομές πυκνότητας πιθανότητας. Τα μείγματα Γκαουσιανών, που μπορούν να προσεγγίσουν οποιαδήποτε συνεχή συνάρτηση πυκνότητας, είναι οι πιο δημοφιλείς κατανομές στα σύγχρονα συστήματα αναγνώρισης. Σύμφωνα με αυτά το $b_j(x)$ δίνεται από την εξής τύπο:

$$b_j(x) = \sum_{k=1}^{M} c_{jk} N(x, \mu_{jk}, \Sigma_{jk}) = \sum_{k=1}^{M} c_{jk} b_{jk}(x) \qquad (4.2)$$

όπου $N(x, \mu_{jk}, \Sigma_{jk})$, $b_{jk}(x)$ μια πολυδιάστατη Γκαουσιανή συνάρτηση πιθανότητας με μέσο διάνυσμα μ_{jk} και διαγώνιο πίνακα Σ_{jk} για κάθε

κατάσταση j, M ο αριθμός των μειγμάτων των Γκαουσιανών και c_{jk} το βάρος του k-στού μείγματος που ικανοποιεί την εξής συνθήκη:

$$\sum_{k=1}^{M} c_{jk} = 1 \quad \text{με} \quad c_{jk} \geq 0 \tag{4.3}$$

4.2.4 Παραγωγή ακουστικών μοντέλων

Εφόσον πλέον έχουμε παράγει τα MFCCs, μπορούμε να προχωρήσουμε στην δημιουργία των ακουστικών μας μοντέλων (**trainA , trainB , train_mix**). Η εκπαίδευση και ο σχηματισμός αυτών των μοντέλων έγινε αποκλειστικά με την βοήθεια του εργαλείου HTK. Μπορείτε να ανατρέξετε στο **παράρτημα Α** όπου περιγράφουμε αναλυτικά την διαδικασία δημιουργίας των ακουστικών μοντέλων με τις εντολές του συγκεκριμένου εργαλείου. Τα 3 αυτά μοντέλα εκπαιδεύτηκαν με τον ίδιο ακριβώς τρόπο, συνεπώς θα περιγράψουμε γενικά την διαδικασία που ακολουθήσαμε και για τα 3 αυτά μοντέλα.

MONOPHONES

Αρχικά, όπως είπαμε στο **υποκεφάλαιο 4.1.1**, η γλωσσικά μονάδα η οποία θα εκπαιδευτεί είναι τα φωνήματα λόγω του μεγάλου λεξιλογίου. Αρχικά θα εκπαιδεύσουμε μονόφωνα ακουστικά μοντέλα, για να μπορέσουμε στην συνέχεια να εκπαιδεύσουμε triphones μοντέλα. Η λίστα των monophones που θα εκπαιδευτούν είναι η εξής:

monophones
o
t
A
n
G
i
m
J
g
r
s
D
k
E
T
u
l
x
p
d
f
v
z
b
ly
C
N
c

Πίνακας 4-1 Λίστα των monophones

Θα πρέπει ωστόσο να σημειωθεί πως τα παραπάνω μονόφωνα δεν έχουν συγκεντρωθεί αυθαίρετα, αλλά έχουν παραχθεί μετά από έρευνα ειδικού γλωσσολόγου πάνω στην ελληνική γλώσσα και προφορά για τα σύστημα της Λογοτυπογραφίας. Επιπλέον, μαζί με φωνήματα της λίστας του πίνακα 4-1 θα εκπαιδευτούν και κάποια επιπλέον μονόφωνα. Αυτά είναι τα παρακάτω:

Additional monophones	Meaning
sil	σιωπή (silence) στην αρχή και το τέλος κάθε πρότασης
hes	δισταγμός του ομιλητή (εε.. , χμ..)
bre	εισπνοή/εκπνοή του ομιλητή
fra	κομμένη λέξη
noi	στιγμιαίος θόρυβος
tbr	ακαθόριστη ομιλία
sp	παύση ομιλητή

Πίνακας 4-2 Επιπλέον μονόφωνα προς εκπαίδευση

Τώρα έχουμε την πλήρη λίστα με τα μονόφωνα μοντέλα που θα εκπαιδευτούν. Συνεπώς για να αρχίσουμε την εκπαίδευση των μοντέλων, μετατρέπουμε κάθε λέξη που περιέχεται στις απομαγνητοφωνήσεις μας στα φωνήματα που περιγράψαμε στους πίνακες 4-1 και 4-2. Αρχικοποιούμε τα μοντέλα μας και μετά από 10 επαναλήψεις του αλγόριθμου Baum-Welch (βλ. Σχήμα 1-5) έχουμε τα εκπαιδευμένα μονόφωνα μοντέλα μας.

Το κάθε context independent HMM που έχει εκπαιδευτεί περιέχει 5 καταστάσεις. Η πρώτη και η τελευταία κατάσταση δεν έχουν συναρτήσεις εξόδου και χρησιμοποιούνται για την σύνδεση διαφορετικών μοντέλων μεταξύ τους.

TRIPHONES

Ολοκληρώνοντας έτσι την εκπαίδευση των μονόφωνων ακουστικών μας μοντέλων, είμαστε σε θέση να δημιουργήσουμε και να εκπαιδεύσουμε triphones μοντέλα. Με τον ίδιο τρόπο πάλι και χρησιμοποιώντας ακόμα μια φορά το εργαλείο του HTK και συγκεκριμένα την εντολή HLed, μετατρέψαμε τις

απομαγνητοφωνήσεις των ακουστικών μας σημάτων σε triphones και σε biphones. Έτσι όσα triphones και biphones σχηματίζονται θα εκπαιδευτούν προς σχηματισμό καινούργιων ακουστικών μοντέλων.

Σε αυτό το σημείο θα πρέπει να σημειώσουμε πως είναι αδύνατον να εκπαιδευτούν όλα τα biphones και όλα τα triphones που μπορούν να υπάρχουν. Αυτό συμβαίνει γιατί ενδεχομένως έχουμε έλλειψη δεδομένων για να εκπαιδευτούν όλα τα δυνατά ακουστικά μοντέλα, και επίσης γιατί η ελληνική γλώσσα έχει κάποιους περιορισμούς στο πως μπορούν να συνδυαστούν τα σύμφωνα με τα φωνήεντα, τα σύμφωνα μεταξύ τους και τα φωνήεντα μεταξύ τους. Για να δούμε πως είναι σχεδόν αδύνατο να εκπαιδευτούν όλες οι περιπτώσεις των triphones, τα θεωρητικά πιθανά triphones που υπάρχουν είναι : 28 (ο αριθμός των monophones) x 28 x 28 = 21952 triphones!!!

TIED-STATE TRIPHONES

Συνεχίζοντας την διαδικασία της εκπαίδευσης, αρχικοποιούμε τα triphone μοντέλα μας και μετά από 3 επαναλήψεις του αλγόριθμου Baum-Welch έχουμε τα εκπαιδευμένα μοντέλα μας. Για να γίνουν τα ακουστικά μας μοντέλα πιο αξιόπιστα και να μπορούν να χειρίζονται, στην διαδικασία της αναγνώρισης, triphones που δεν έχουν εκπαιδευτεί να αναγνωρίζουν, δημιουργούμε τα tied-states triphones.

Με αυτού του είδους τα μοντέλα, οι καταστάσεις των triphones ‘’δένονται’’ κατά κάποιο τρόπο μεταξύ τους με σκοπό να μοιράσουν πληροφορίες μεταξύ τους και έτσι να κάνουν πιο αξιόπιστους υπολογισμούς ειδικά για φωνήματα που δεν έχουν ξανασυναντήσει. Τα tied-state triphones δημιουργήθηκαν με την βοήθεια της εντολής HHed του HTK και μετά από 3 ακόμα επαναλήψεις του αλγόριθμου Baum-Welch έχουμε τα τελικά tied-state μοντέλα μας.

GAUSSIAN MIXTURES

Θα πρέπει να σημειώσουμε πως μέχρι στιγμής έχει ολοκληρωθεί η εκπαίδευση των tied-state triphones. Αυτά τα HMM περιέχουν και αυτά επίσης 5 καταστάσεις και μια πολυδιάστατη Γκαουσιανή 39 διαστάσεων σε κάθε κατάσταση. Για να βελτιωθεί αισθητά η απόδοση των ακουστικών μας μοντέλων θα πρέπει να αυξήσουμε το αριθμό των μειγμάτων των Γκαουσιανών του κάθε HMM.

Συνεπώς, πάντα με την βοήθεια του HTK , αυξήσαμε σταδιακά τον αριθμό των μειγμάτων. Αφού περάσαμε το πιο πρόσφατο tied-state μοντέλο μας 4 φορές από τον αλγόριθμο Baum-Welch, αυξήσαμε τον αριθμό των Γκαουσιανών σε 2. Από 2 σε 3 και μετά σε 4 μείγματα εφαρμόζοντας πάντα τον παραπάνω αλγόριθμο για τέσσερις φορές. Τελικά, το τελικό μας ακουστικό μοντέλο περιέχει 12 μείγματα Γκαουσιανής από 39 διαστάσεις σε κάθε κατάσταση.

4.3 Εκπαίδευση ακουστικών μοντέλων

Σε αυτό το κεφάλαιο θα παρουσιάσουμε την ακριβή διαδικασία που θα μας οδηγήσει στην δημιουργία διάφορων ακουστικών μοντέλων. Θα δείξουμε από ποια και από πόσα δεδομένα εκπαιδεύτηκαν τα μοντέλα μας καθώς από τι είδους HMM αποτελούνται αυτά τα μοντέλα.

4.3.1 Δεδομένα προς εκπαίδευση (training data)

Στο **κεφάλαιο 2.4** περιγράψαμε τις κατηγορίες από τις οποίες αποτελείται η βάση των ακουστικών μας σημάτων και ειδικότερα, στον πίνακα 2.2, παρουσιάσαμε την ποσότητα των δεδομένων αυτών και σε ποσότητα προτάσεων μα και σε ποσότητα χρονικής διάρκειας.

Το σύνολο αυτών των σημάτων θα πρέπει να το διαχωρίσουμε σε δύο σύνολα δεδομένων. Αρχικά, θα πρέπει να δημιουργήσουμε το σύνολο εκείνο το οποίο θα περιέχει τα δεδομένα εκείνα με τα οποία θα εκπαιδεύσουμε τα ακουστικά μας μοντέλα (**training sets**). Σε δεύτερη φάση, θα δημιουργήσουμε το set εκείνο που θα περιέχει τα δεδομένα εκείνα με τα οποία θα δοκιμάσουμε (evaluation) το

σύστημα αναγνώρισης μας (**test sets**). Αποφασίσαμε πως από την κάθε κατηγορία που έχουμε δημιουργήσει (βλ. Πίνακα 2.2), το 80% των δεδομένων της κάθε κατηγορίας θα το χρησιμοποιήσουμε για την εκπαίδευση των μοντέλων και το υπόλοιπο 20% θα χρησιμοποιηθεί για την διαδικασία της εκτίμησης της επίδοσης. Οπότε παρουσιάζουμε τα δεδομένα εκπαίδευσης της κάθε κατηγορίας.

	Αριθμός προτάσεων	Χρονική διάρκεια σε ώρες
F0	**334**	**0.98**
F1	**2693**	**7.95**
F2	**276**	**0.81**
F3	**66**	**0.19**
F4	**18**	**0.05**
F5	**42**	**0.12**
F6	**1045**	**3.08**
F7	**376**	**1.11**
F8	**570**	**1.68**

Πίνακας 4-3 Παρουσίαση των training data

Αφού πλέον κάναμε τον διαχωρισμό των δεδομένων μας, τώρα μπορούμε να παρουσιάσουμε από ποια συγκεκριμένα δεδομένα θα εκπαιδευτούν τα μοντέλα μας. Θα πρέπει να σημειώσουμε πως ήταν περιττό στον παραπάνω πίνακα να παρουσιάσουμε όλες τις κατηγορίες και αυτό διότι π.χ. η κατηγορία ***F4*** αποτελείται από non-greek δεδομένα και είναι αυτονόητο πως δεν θα πάρουμε δεδομένα από την συγκεκριμένη κατηγορία ούτε για το training set μα ούτε και για το test set. Απλώς θέλαμε να δείξουμε τον διαχωρισμό των όλων δεδομένων σε 80% training data και σε 20% test data.

Θα υλοποιήσουμε τα εξής 3 ακουστικά μοντέλα:

- το **trainA** μοντέλο που θα εκπαιδευτεί μόνο με προτάσεις που έχουν σαν συνθήκη ομιλίας studio (κοίτα σελ. 32), δηλαδή από τις προτάσεις της κατηγορίας ***F6*** και ***F7***
- το **trainB** μοντέλο που θα εκπαιδευτεί με προτάσεις που έχουν σαν συνθήκη ομιλίας studio αλλά και με προτάσεις που έχουν σαν συνθήκη ομιλίας noise (κοίτα σελ. 32), δηλαδή από τις κατηγορίες ***F6*** , ***F7*** , ***F1*** και ***F2***
- το **train_mix** μοντέλο που έχει εκπαιδευτεί με τις προτάσεις όλων των κατηγοριών εκτός της κατηγορίας ***F3*** (multispeaker), και της κατηγορίας ***F4*** (non-greek)

Οπότε πλέον μπορούμε να παρουσιάσουμε την ποσότητα των δεδομένων με τα οποία θα εκπαιδευτούν τα 3 μας ακουστικά μοντέλα

	Αριθμός προτάσεων	χρονική διάρκεια σε ώρες
trainA	**1421**	**4.25**
trainB	**4114**	**12.2**
train_mix	**5594**	**15.8**

Πίνακας 4-4 Παρουσίαση training data του κάθε ακουστικού μοντέλου

Τελικά, σύμφωνα με την διαδικασία που εξηγήσαμε στο προηγούμενο κεφάλαιο εκπαιδεύουμε τα τελικά μας μοντέλα. Συνοπτικά, παρουσιάζουμε ένα πίνακα που δείχνει τον αριθμό των monophones, των biphones και των triphones που έχει εκπαιδευτεί σε κάθε ένα από τα ακουστικά μας μοντέλα.

	Monophones (#)	biphones (#)	triphones(#)
trainA	35	311	2955
trainB	35	358	3432
train_mix	35	364	3606

Πίνακας 4-5 Αριθμός monophones, biphones και triphones κάθε ακουστικού μοντέλου

4.4 Εκπαίδευση ακουστικών μοντέλων με προσαρμογή (Adaptation)

Είναι προφανές το γεγονός πως η ομιλία είναι εξ' ορισμού ποικιλόμορφη. Δηλαδή εξαρτάται από πολλούς παράγοντες και από πληθώρα παραμέτρων. Μπορούμε να πούμε συνοπτικά ότι εξαρτάται από τρία πράγματα:

- Τον ομιλητή.
- Το περιβάλλον στο οποίο παράγεται η ομιλία.
- Τον τρόπο και το περιεχόμενο της ομιλίας.

Για να γίνουμε πιο συγκεκριμένοι, αρχικά είναι αυτονόητο ότι η ομιλία εξαρτάται από τον εκάστοτε ομιλητή γιατί παίζει ρόλο η ηλικία του και το φύλο του. Έπειτα παίζει ρόλο η προφορά του, ο τρόπος ομιλίας του και ενδεχομένως η ταχύτητα με την οποία παράγει τον λόγο. Επιπρόσθετα, όσον αναφορά το περιβάλλον της ομιλίας, σίγουρα όταν βρισκόμαστε σε χώρους με πολύ θόρυβο, όπως δημόσια μέρη, επηρεάζει σημαντικά την ταυτότητα της ομιλίας. Ακόμα και όταν μιλάμε μέσω τηλεφώνου, διαπιστώνουμε και από μόνοι μας ότι αλλοιώνεται δραστικά η φωνή μας. Στην περίπτωση του περιεχομένου της ομιλίας είναι σίγουρο πως είναι διαφορετικός ο τρόπος της ομιλίας μας όταν συνομιλούμε π.χ. με ένα ομιλητή σε χαλαρό ρυθμό από όταν απαγγέλουμε ένα ποίημα.

Παρατηρούμε έτσι πως η ομιλία έχει τεράστια ποικιλομορφία. Στον τομέα της αναγνώρισης φωνής όμως, έχουν βρεθεί ποικίλοι τρόποι για να ειδικεύουμε κατά κάποιον τρόπο τα συστήματα αναγνώρισης με σκοπό να έχουμε καλύτερα αποτελέσματα αναγνώρισης. Συνεπώς όταν έχουμε στην κατοχή μας ένα ‘’γενικό’’ σύστημα αναγνώρισης που αναγνωρίζει με μερική επιτυχία σήματα ομιλίας που παράχθηκαν κάτω από οποιαδήποτε συνθήκες, μπορούμε να προσαρμόσουμε αυτό το γενικό μας σύστημα με διάφορους τρόπους έτσι ώστε να αναγνωρίζει ‘’ειδικά’’ ακουστικά σήματα. Δηλαδή μπορούμε να υλοποιήσουμε ένα ειδικό σύστημα αναγνώρισης με αυτόν τον τρόπο που θα έχει μεγάλη επιτυχία αναγνώρισης πάνω στα δεδομένα που θα προσαρμοστεί. Αυτή όλη η διαδικασία λέγεται **προσαρμογή** (adaptation). Παρακάτω θα παρουσιάσουμε τα δύο κυριότερα είδη προσαρμογής.

4.4.1 MAP Adaptation (Maximum a Posteriori)

Στο MAP adaptation γίνεται χρήση της παρούσας πληροφορίας από τα υπάρχοντα μοντέλα έτσι ώστε τα περιορισμένα adaptation data (δεδομένα που θα χρησιμοποιήσουμε για το adaptation) να αλλάξουν τις παραμέτρους των μοντέλων οδηγημένα πάντα από την ήδη υπάρχουσα πληροφορία των μοντέλων.

Ας υποθέσουμε ότι το λ είναι το σύνολο των παραμέτρων ενός HMM και $P(\lambda)$ είναι η ήδη γνωστή προηγούμενη πληροφορία. Έχοντας σαν παρατήρηση το X, το λ ορίζεται σαν το μέγιστο της posterior πυκνότητας πιθανότητας του λ:

$$\lambda_{MAP} = \arg\max_{\lambda} P(\lambda \mid X) \tag{4.4}$$

Θα πρέπει να σημειωθεί πως αν δεν έχουμε προηγούμενη πληροφορία ο MAP υπολογισμός γίνεται ταυτόσημος με τον maximum likelihood υπολογισμό.

Η σχέση 4-4 λόγω του θεωρήματος Bayes γίνεται ως εξής:

$$\lambda_{MAP} = \arg\max_{\lambda} \frac{L(X \mid \lambda) P_0(\lambda)}{P(x)} \tag{4.5}$$

Το X είναι τα δεδομένα προσαρμογής, το $L(X \mid \lambda)$ η πιθανοφάνεια των δεδομένων προσαρμογής για μια τιμή του λ, $P(x)$ η οριακή πιθανότητα των δεδομένων προσαρμογής, που όμως εξαλείφεται γιατί δεν εξαρτάται από το μοντέλο και τέλος ο όρος $P_0(\lambda)$ είναι η ήδη υπάρχουσα a-priori πυκνότητα πιθανότητας του μοντέλου.

Για ευκολία, το MAP adaptation μπορεί να εφαρμοστεί μόνο στις μέσες τιμές (means) των Γκαουσιανών στα HMM. Έτσι ισχύει ο εξής τύπος για την μέση τιμή του m-οστού μείγματος της j κατάστασης:

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm} + \tau} \bar{\mu}_{jm} + \frac{\tau}{N_{jm} + \tau} \mu_{jm} \tag{4.6}$$

όπου τ το βάρος της παρούσας πληροφορίας, N ο αριθμός δεδομένων προσαρμογής που χρησιμοποιούνται για το m-οστό μείγμα της j κατάστασης. Το N δίνεται από τον εξής τύπο:

$$N_{jm} = \sum_{r=1}^{R} \sum_{t=1}^{T_r} L^r_{jm}(t) \tag{4.7}$$

όπου R ο αριθμός των προτάσεων προσαρμογής, T ο αριθμός των frames της r-οστής πρότασης.

Συνεχίζοντας την επεξήγηση των όρων της σχέσης 4-6 , το μ_{jm} είναι η μέση τιμή του μοντέλου που θα προσαρμοστεί, το $\bar{\mu}_{jm}$ είναι η μέση τιμή των παρατηρήσεων των δεδομένων και ισούται με:

$$\bar{\mu}_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t) O_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)} \tag{4.8}$$

Όπου $L_{jm}^{r}(t)$ είναι η a-posteriori πιθανότητα να βρισκόμαστε στο m-οστό μείγμα της j κατάστασης τη χρονική στιγμή t της r-οστής πρότασης. Αυτή υπολογίζεται με το forward-backward και είναι αντίστοιχη της πιθανότητας $\gamma_t(i)$ που ορίστηκε στην σχέση (1.15) του κεφαλαίου 1.

Τέλος, θα πρέπει να σημειώσουμε πως συνήθως το MAP adaptation εφαρμόζεται όταν έχουμε πολλά δεδομένα προσαρμογής. Οι μόνοι περιορισμοί και προβλήματα που έχει είναι ότι πρέπει να υπολογιστεί ο όρος $P_0(\lambda)$ της σχέσης 4-5 από τα αρχικά εκπαιδευμένα μοντέλα, και ότι αλλάζει μόνο τις παραμέτρους του μοντέλου για τις οποίες έχουν πληροφορία τα δεδομένα προσαρμογής.

4.4.2 MLLR Adaptation (Maximum Likelihood Linear Regression)

Η εφαρμογή του MLLR για προσαρμογή μοντέλων έγκειται στην παραγωγή κάποιων regression-based μετασχηματισμών (transforms) από ορισμένα δεδομένα προσαρμογής. Αυτοί οι μετασχηματισμοί χρησιμοποιούνται αργότερα για να ρυθμίσουν και τελικά να αλλάξουν κάποιες παραμέτρους των HMM που υπόκεινται σε προσαρμογή. Οι MLLR μετασχηματισμοί γενικά εφαρμόζονται πάνω στις μέσες τιμές των μειγμάτων των Γκαουσιανών. Αυτό γίνεται γιατί οι Γκαουσιανές είναι τα πιο βασικά στοιχεία ενός HMM που πρέπει να αναβαθμιστεί όταν προσαρμόζεται σε καινούργια δεδομένα και σε νέες συνθήκες.

Η χρήση της MLLR προσαρμογής για μια Γκαουσιανή προϋποθέτει τον υπολογισμό ενός πίνακα μετασχηματισμού από παρατηρήσεις. Αυτός ο πίνακας

θα χρησιμοποιηθεί για να υπολογιστούν οι προσαρμοσμένες μέσες τιμές. Για μια παρατήρηση της διάστασης n έχω:

$$\hat{\mu}_s = W_s \xi_s \tag{4.9}$$

όπου W_s είναι ένας $n \times (n+1)$ πίνακας μετασχηματισμού, ο ξ_s όρος είναι εκτεταμένο διάνυσμα από means έτσι ώστε για $w=1$ να σημαίνει ότι υπάρχει offset και όταν $w=0$ ότι δεν υπάρχει. Ο όρος ξ_s υπολογίζεται ως εξής:

$$\xi_s = [w, \mu_{s1}, \ldots, \mu_{sn}]^t \tag{4.10}$$

Ο όρος w_s υπολογίζεται από τον τύπο:

$$w_S = \sum_{t=1}^{T} \sum_{r=1}^{R} L_{sr}(t) \sum_{sr}^{-1} O(t) \xi_{sr}^T = w_S = \sum_{t=1}^{T} \sum_{r=1}^{R} L_{sr}(t) \sum_{sr}^{-1} W_s \xi_{sr} \xi_{sr}^T \tag{4.11}$$

όπου $L_{sr}(t)$ η πιθανότητα που υπολογίζεται από τον forward-backward αλγόριθμο που επεξηγήσαμε στην σελίδα 19.

Αντίθετα με το MAP προσαρμογή, η MLLR προσαρμογή αλλάζει και εν τέλει προσαρμόζει τις παραμέτρους όλων των μοντέλων, και συνιστάται η χρήση του όταν έχουμε μικρό αριθμό από δεδομένα προσαρμογής.

4.4.3 Προσαρμογή με τα δικά μας δεδομένα

Όπως αναφέραμε στην αρχή του κεφαλαίου, η ομιλία εξαρτάται από πολλούς παράγοντες. Συνεπώς μπορεί να εφαρμοστούν πολλά είδη προσαρμογής:

- Προσαρμογή ομιλητή (speaker adaptation)
- Προσαρμογή περιβάλλοντος (environment adaptation)

- Προσαρμογή περιεχομένου (task adaptation)

Θα προσπαθήσουμε να προσαρμόσουμε ένα ακουστικό μοντέλο που έχει εκπαιδευτεί μόνο με καθαρά σήματα φωνής (χωρίς θόρυβο) για να αναγνωρίζει ακουστικά σήματα που είναι ηχοχραφημένα από τηλεοπτικές εκπομπές. Δηλαδή θα εφαρμόσουμε task adaptation. Το ακουστικό μοντέλο που θα προσαρμοσθεί στις καινούργιες συνθήκες ονομάζεται seed model. Αυτό το μοντέλο είναι ήδη εκπαιδευμένο από παλαιότερη μεταπτυχιακή εργασία του Δημήτρη Οικονομίδη, υποψήφιου διδάκτωρ του Πολυτεχνείου Κρήτης. Αποτελείται από μια σειρά monophones, biphones και triphones HMMs που έχουν 12 μείγματα Γκαουσιανών ανά κατάσταση. Το συγκεκριμένο ακουστικό μοντέλο έχει εκπαιδευτεί με 72 ώρες καθαρών ακουστικών σημάτων. Αξίζει να σημειωθεί πως έχει ακριβώς τα ίδια monophones με τα δικά μας ακουστικά μοντέλα, γεγονός που μας βοηθάει αρκετά στην όλη διαδικασία. Το seed μοντέλο μας περιέχει:

	monophones(#)	**biphones(#)**	**triphones(#)**
seed model	34	596	6468

Πίνακας 4-6 Αριθμός φωνημάτων του seed μοντέλου

Εφαρμόσαμε δύο ειδών προσαρμογών: MAP προσαρμογή και συνδυασμό MAP και MLLR προσαρμογής. Ειδικότερα για την MLLR προσαρμογή χρησιμοποιήσαμε έναν καθολικό (global) μετασχηματισμό. Στην κάθε περίπτωση υλοποιήσαμε 3 καινούργια ακουστικά μοντέλα. Δηλαδή συνολικά υλοποιήθηκαν 6 καινούργια ακουστικά μοντέλα. Προσαρμόσαμε κάθε φορά στο seed μοντέλο μας σαν δεδομένα προσαρμογής τα δεδομένα εκπαίδευσης από τα δικά μας υλοποιημένα ακουστικά μοντέλα **trainA**, **trainB** και **train_mix**. Συνοπτικά υλοποιήσαμε πάντα με βάση το seed μοντέλο:

- 2 ακουστικά μοντέλα (MAP adaptation και MAP + MLLR adaptation) με adaptation data 1421 προτάσεις που έχουν σαν συνθήκη ομιλίας studio (κοίτα σελ. 32), δηλαδή από τις προτάσεις της κατηγορίας ***F6*** και ***F7***. Το **MAP_A** και το **MAP+MLLR_A**
- 2 ακουστικά μοντέλα (MAP adaptation και MAP + MLLR adaptation) με adaptation data 4114 προτάσεις που έχουν σαν συνθήκη ομιλίας studio αλλά και με προτάσεις που έχουν σαν συνθήκη ομιλίας noise (κοίτα σελ. 32), δηλαδή από τις κατηγορίες ***F6*** , ***F7*** , ***F1*** και ***F2***. Το **MAP_B** και το **MAP+MLLR_B**
- 2 ακουστικά μοντέλα (MAP adaptation και MAP + MLLR adaptation) με adaptation data 5994 προτάσεις όλων των ***F*** κατηγοριών εκτός της κατηγορίας ***F3*** (multispeaker), και της κατηγορίας ***F4*** (non-greek). Το **MAP_mix** και το **MAP+MLLR_mix**

ΚΕΦΑΛΑΙΟ 5

ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑΓΝΩΡΙΣΗΣ

Σε αυτό το κεφάλαιο θα δείξουμε τα διάφορα αποτελέσματα αναγνώρισης του συστήματος. Θα δείξουμε αναλυτικά τις επιδόσεις των εκπαιδευμένων μας ακουστικών μοντέλων, τον μοντέλων που δημιουργήθηκαν μετά από MAP adaptation, και των μοντέλων που έγιναν μετά από MAP+MLLR adaptation. Φυσικά το αποτέλεσμα της αναγνώρισης φωνής είναι συνέπεια των ακουστικών μοντέλων που προαναφέραμε συναρτήσει πάντα με το γλωσσικό μοντέλο που υλοποιήσαμε. Από εδώ και πέρα αυτό το γλωσσικό μοντέλο θα θεωρείται δεδομένο και θα αναφέρουμε μόνο τα διάφορα ακουστικά μοντέλα.

Έχουμε 3 σύνολα δοκιμών με τα οποία θα αξιολογήσουμε το σύστημα μας (evaluation). Αυτά είναι τα εξής:

- το **testA** σύνολο που περιέχει προτάσεις που έχουν σαν συνθήκη ομιλίας studio (κοίτα σελ. 32), δηλαδή από τις προτάσεις της κατηγορίας ***F6*** και ***F7***
- το **test B** σύνολο που περιέχει προτάσεις που έχουν σαν συνθήκη ομιλίας studio αλλά και με προτάσεις που έχουν σαν συνθήκη ομιλίας noise (κοίτα σελ. 32), δηλαδή από τις κατηγορίες ***F6*** , ***F7*** , ***F1*** και ***F2***
- το **test_mix** σύνολο που περιέχει προτάσεις όλων των κατηγοριών εκτός της κατηγορίας ***F3*** (multispeaker), και της κατηγορίας ***F4*** (non-greek)

Παρακάτω δίνεται ένας πίνακας που περιγράφει αριθμητικά τα 3 test set μας:

	Αριθμός προτάσεων	Χρονική διάρκεια σε ώρες
testA	347	1.02
testB	1006	2.97
test_mix	1263	3.73

Πίνακας 5-1 Παρουσίαση των test sets

Στο υποκεφάλαιο 4.2.1 είχαμε τονίσει πως από όλα τα συγκεντρωμένα δεδομένα της βάσης μας, το 80% θα χρησιμοποιούταν για την εκπαίδευση των ακουστικών μας μοντέλων και το 20% για την διαδικασία του evaluation. Συνεπώς τα παραπάνω δεδομένα που παρουσιάζονται στον πίνακα 5-1 είναι το 20% της βάσης των ακουστικών μας δεδομένων.

5.1 Αποτελέσματα των εκπαιδευμένων ακουστικών μοντέλων

Σε αυτή την ενότητα θα υπολογίσουμε την απόδοση και των τριών ακουστικών μας μοντέλων συναρτήσει πάντα του bigram γλωσσικού μοντέλου. Θα τεστάρουμε καθένα από τα ακουστικά μας μοντέλα και με τα τρία test sets που περιγράφτηκαν στην αρχή αυτού του κεφαλαίου. Η μέτρηση του λάθος της αναγνώρισης σε κάθε περίπτωση έγινε με το εργαλείο του HTK , *HResults*.

Στα παρακάτω γραφήματα, έχουμε θεωρήσει σαν κάθετο άξονα την ποσότητα Accuracy που ορίζεται ως εξής:

$$Accuracy = \frac{H - I}{N} X100\% \qquad (5\text{-}1)$$

όπου H ο αριθμός των σωστών labels, I ο αριθμός των insertions και N ο συνολικός αριθμός των labels

Παρακάτω θα παρουσιάσουμε τα αποτελέσματα αναγνώρισης για τα 3 test set, των εκπαιδευμένων μας μοντέλων με μίξεις 10, 12 και 14 Γκαουσιανών σε μορφή πινάκων.

10 Gaussian	testA	testB	test_mix
trainA	47,30	40,27	36,93
trainB	63,28	61,11	58,13
train_mix	65,48	62,97	61,37

Πίνακας 5-2 Αποτελέσματα με μίξεις 10 Γκαουσιανών (Acc%)

12 Gaussian	testA	testB	test_mix
trainA	47,80	40,57	37,23
trainB	63,89	61,67	58,59
train_mix	65,96	63,24	61,58

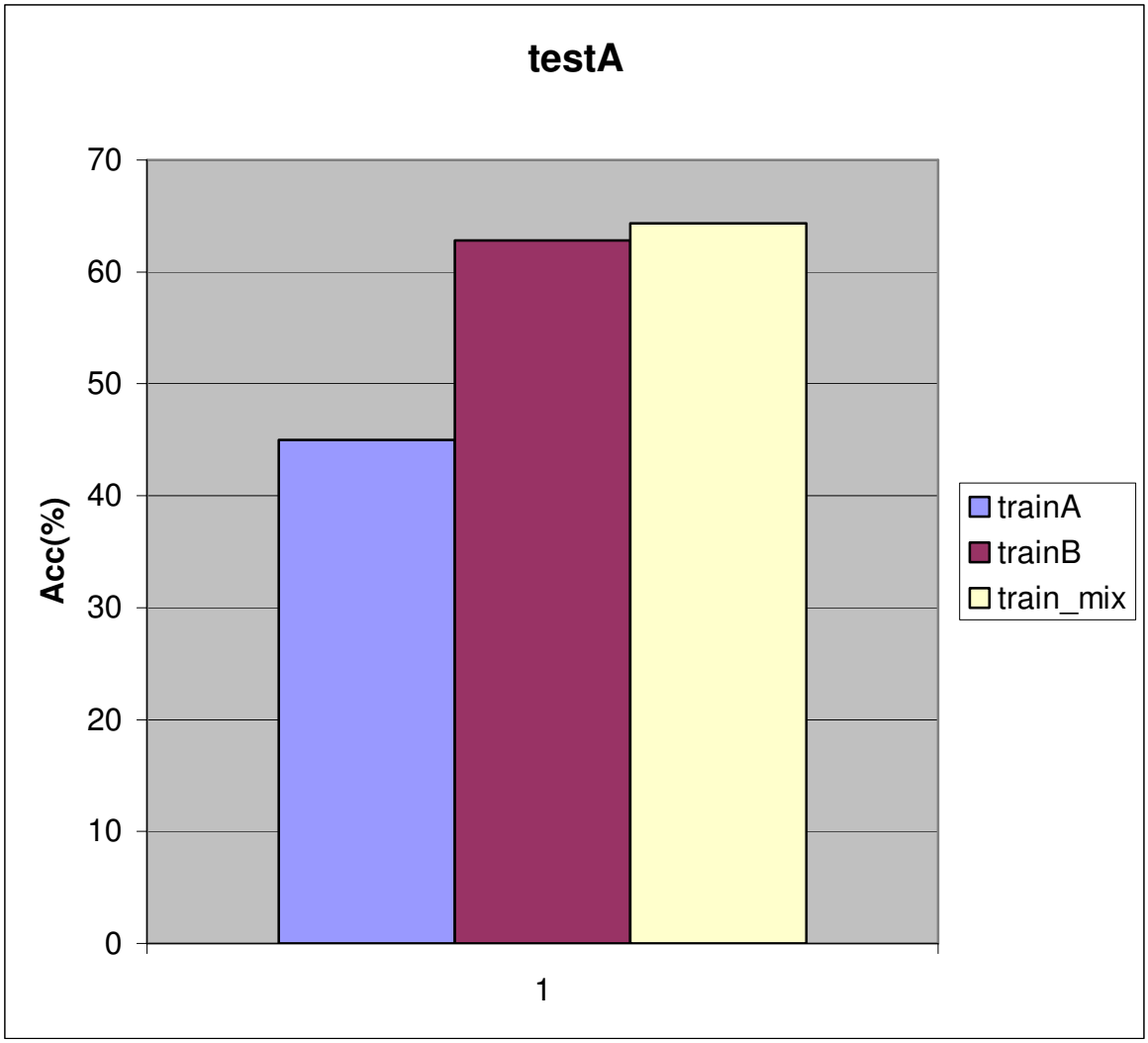
Πίνακας 5-3 Αποτελέσματα με μίξεις 12 Γκαουσιανών (Acc%)

14 Gaussian	testA	testB	test_mix
trainA	43,56	35,14	31,69
trainB	62,75	60,09	57,13
train_mix	64,03	61,71	60,69

Πίνακας 5-4 Αποτελέσματα με μίξεις 14 Γκαουσιανών (Acc%)

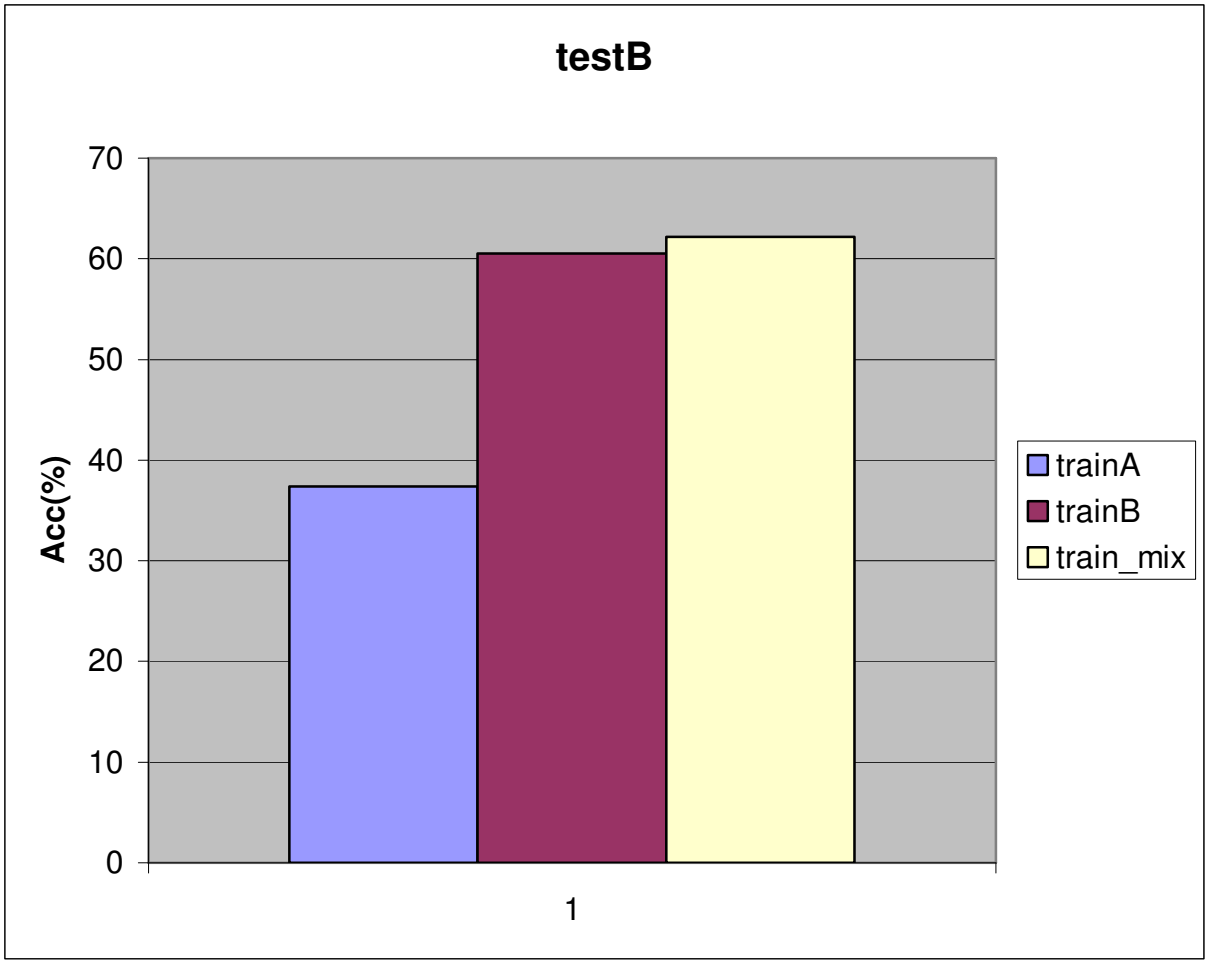
Από ότι βλέπουμε τα καλύτερα αποτελέσματα τα έχουμε για ακουστικά μοντέλα με μίξεις 12 Γκαουσιανών. Παρακάτω παρουσιάζουμε σε μορφή γραφημάτων τα αποτελέσματα με τις μίξεις 12 Γκαουσιανών.

:



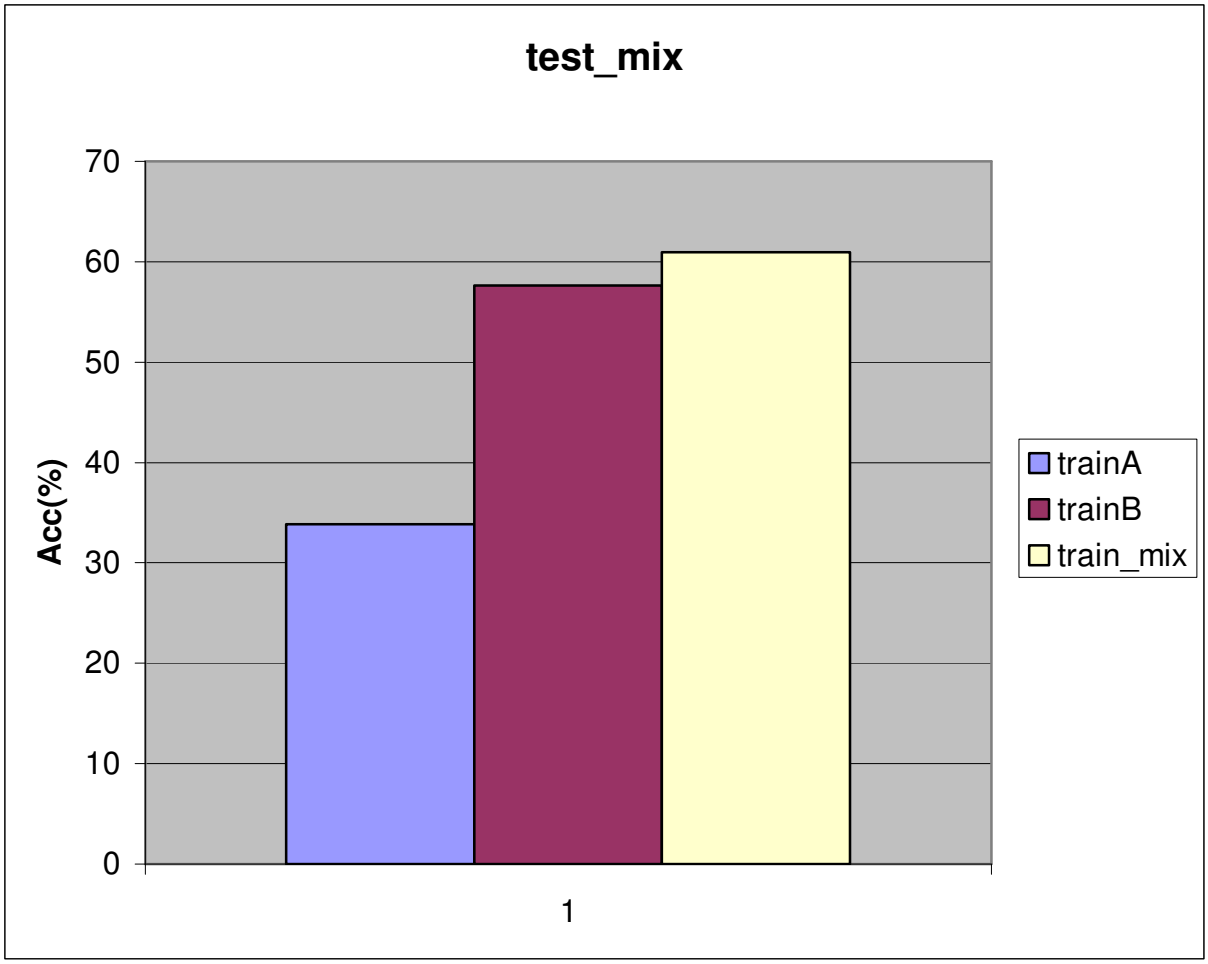
Σχήμα 5-1 Αποτελέσματα του testA set με μίξεις 12 Γκαουσιανών

Αποτελέσματα και των τριών ακουστικών μας μοντέλων αξιολογούνται στο **testB**:



Σχήμα 5-2 Αποτελέσματα του testB set με μίξεις 12 Γκαουσιανών

Αποτελέσματα και των τριών ακουστικών μας μοντέλων αξιολογούνται πάνω στο **test_mix**:



Σχήμα 5-3 Αποτελέσματα του test_mix set με μίξεις 12 Γκαουσιανών

Από τα 3 παραπάνω σχήματα, μπορούμε να βγάλουμε μια σειρά από συμπεράσματα για το σύστημα αναγνώρισης μας. Αρχικά βλέπουμε πως το ακουστικό μοντέλο **trainA** έχει σε κάθε περίπτωση, δηλαδή και στα 3 test sets, απογοητευτικά αποτελέσματα αναγνώρισης. Αυτό συμβαίνει γιατί το συγκεκριμένο μοντέλο, έχει εκπαιδευτεί με πολύ λίγες προτάσεις , μόλις 1421 προτάσεις. Οπότε είναι λογικό να έχουμε τόσο χαμηλά ποσοστά αναγνώρισης.

Επίσης, είναι λογικό τα μοντέλα **trainB** και **train_mix** να έχουν αρκετά πιο μεγάλα ποσοστά αναγνώρισης από το μοντέλο **trainA** εφόσον τα συγκεκριμένα μοντέλα εκπαιδεύτηκαν με πολύ περισσότερες προτάσεις, έχοντας σαν συνέπεια

την καλύτερη εκπαίδευση των φωνημάτων τους. Τέλος το **train_mix** μοντέλο μας πετυχαίνει τα μεγαλύτερα ποσοστά αναγνώρισης και από τα 3 μοντέλα και αυτό γιατί μεν έχει εκπαιδευτεί με τις περισσότερες ώρες από όλα τα άλλα μοντέλα, αλλά και γιατί έχει εκπαιδευτεί σε δύσκολες συνθήκες (telephone ποιότητα κλπ.) γεγονός που το κάνει πιο αξιόπιστο και πιο ακριβή στις προβλέψεις του.

Για να γίνουμε πιο συγκεκριμένοι και για να δείξουμε τις αποδόσεις των ακουστικών μας μοντέλων και σε διαφορετικού είδους σύνολα δοκιμής θα παραθέσουμε παρακάτω 3 πίνακες που παρουσιάζουν την επίδοση των ακουστικών σε 4 διαφορετικά σύνολα δοκιμής:

- το **testA** που αποτελείται από τις προτάσεις της κατηγορίας ***F6*** και ***F7*** (συνθήκη ομιλίας studio)
- το **test_tel_noi** που αποτελείται από τις προτάσεις της κατηγορίας ***F2*** (συνθήκη ομιλίας noise && ποιότητα φωνής telephone)
- το **test_st_noi** το οποίο αποτελείται από τις προτάσεις της κατηγορίας ***F1*** (συνθήκη ομιλίας noise && ποιότητα φωνής studio)
- το **test_other** το οποίο αποτελείται από τις προτάσεις της κατηγορίας ***F0, F5, F8*** (music, non-native, telephone)

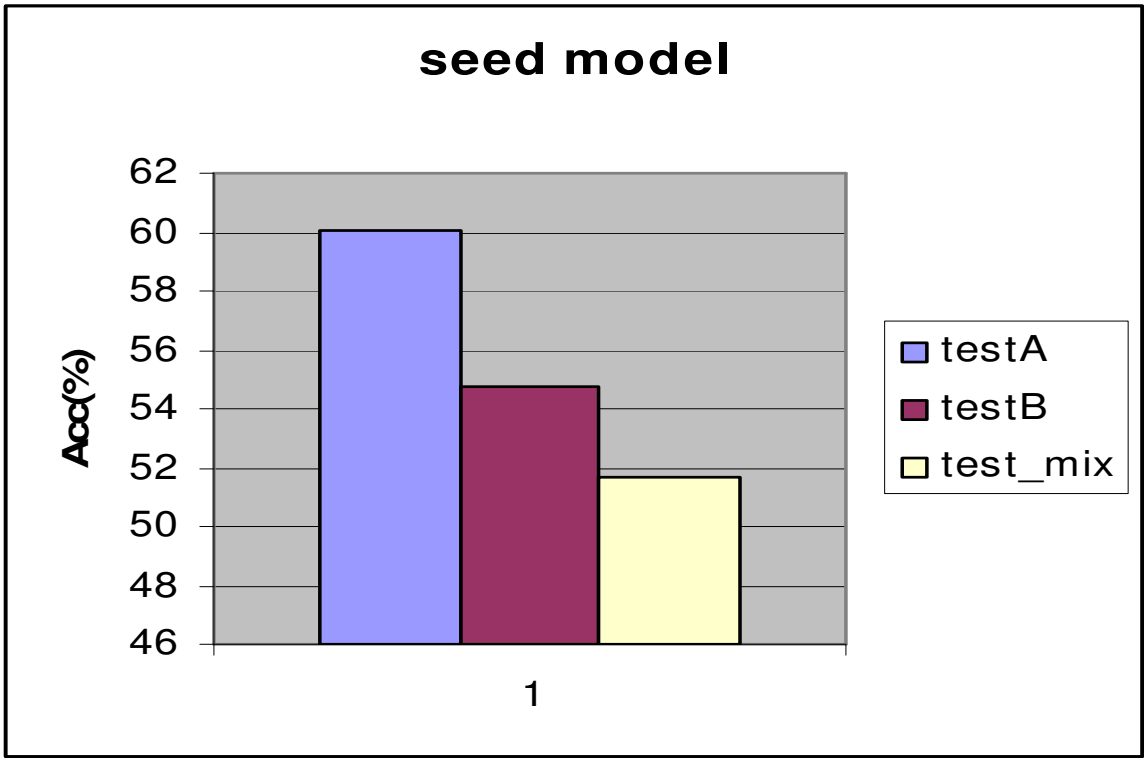
	testA	**test_tel_noi**	**test_st_noi**	**test_other**
trainA	47,80	14,07	31,64	19,06
trainB	63,89	31,85	59,00	44,84
train_mix	65,96	40,00	60,80	55,51

Πίνακας 5-5 Αποτελέσματα των εκπαιδευμένων ακουστικών μοντέλων για 4 σύνολα δοκιμών (Acc%)

5.2 Αποτελέσματα του seed μοντέλου

Πριν περάσουμε στα αποτελέσματα των ακουστικών μοντέλων που παράχθηκαν μετά από προσαρμογή στα δικά μας δεδομένα, θα ήταν χρήσιμο να δούμε ποια είναι τα αποτελέσματα του seed μοντέλου το οποίο κάνουμε προσαρμογή στα σετ δοκιμών μας.

Αυτά τα αποτελέσματα παρουσιάζονται στο σχήμα 5-4.



Σχήμα 5-4 Αποτελέσματα του seed model

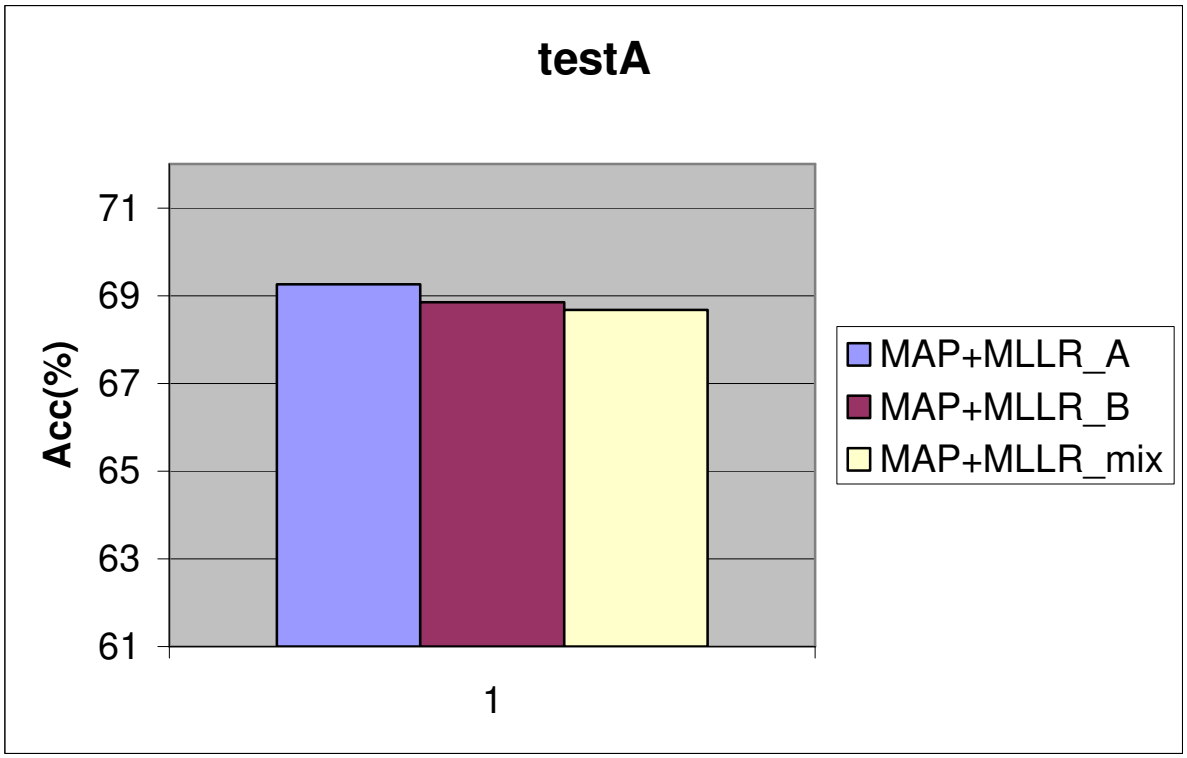
Παρακάτω παραθέτουμε έναν πίνακα που παρουσιάζει τα συγκεντρωτικά αποτελέσματα των ακουστικών μοντέλων που παράχθηκαν μετά από MAP+MLLR προσαρμογή:

	testA	testB	test_mix
seed model	60,02	54,75	51,69

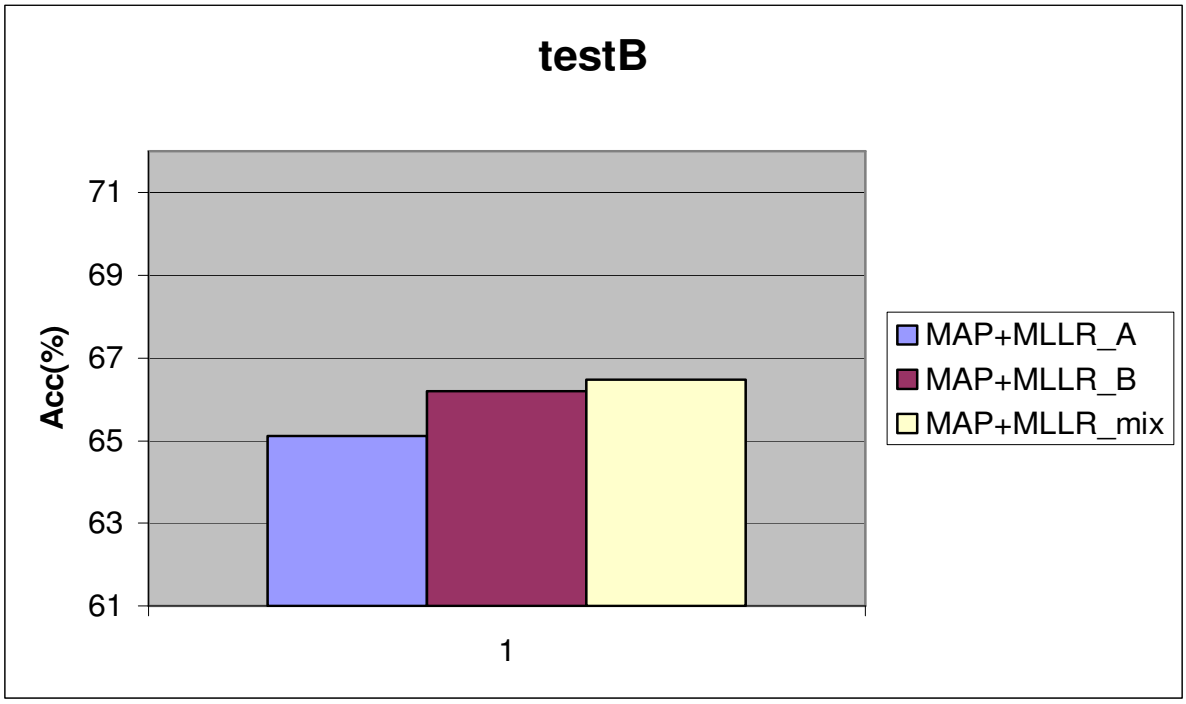
Πίνακας 5-6 Αποτελέσματα seed (Acc%)

5.3 Αποτελέσματα των ακουστικών μοντέλων που παράχθηκαν μετά από MLLR+MAP adaptation

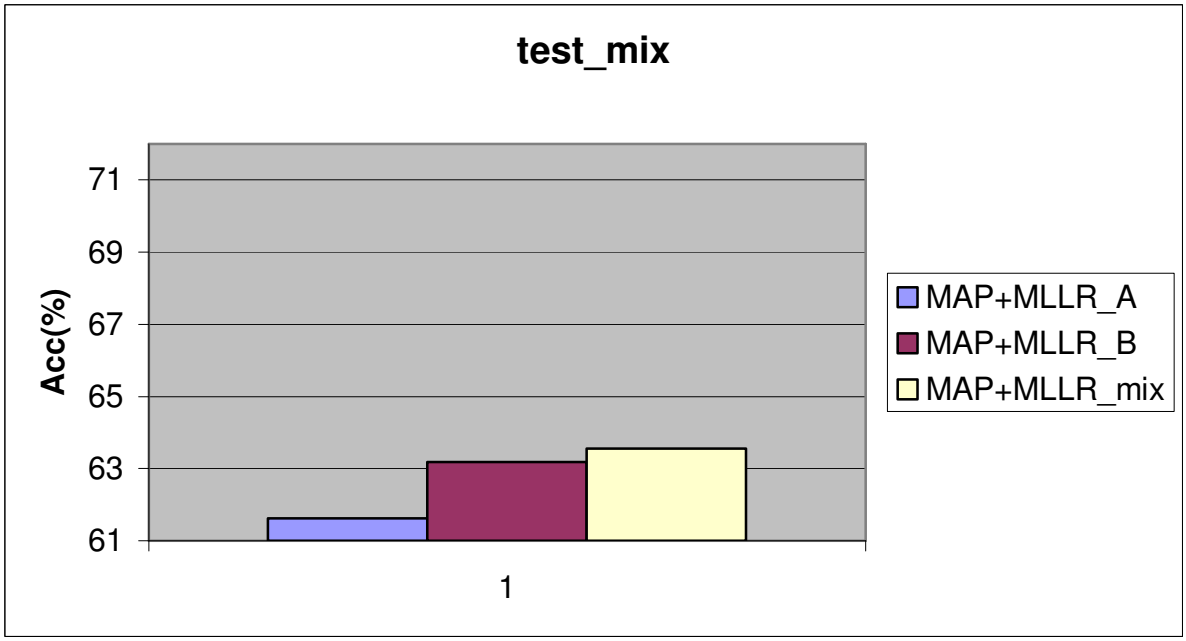
Τα αποτελέσματα της αναγνώρισης των μοντέλων που δημιουργήθηκαν μετά από MLLR και MAP adaptation στα **testA**, **testB** ,**test_mix** παρουσιάζονται στα σχήματα 5-5 ,5-6 και 5-7 αντίστοιχα.



Σχήμα 5-5 Αποτελέσματα του test_A set



Σχήμα 5-6 Αποτελέσματα του test_B set



Σχήμα 5-7 Αποτελέσματα του test_mix set

Παρακάτω παραθέτουμε έναν πίνακα που παρουσιάζει τα συγκεντρωτικά αποτελέσματα των ακουστικών μοντέλων που παράχθηκαν μετά από MAP+MLLR προσαρμογή:

	testA	**testB**	**test_mix**
MAP+MLLR_A	69,26	65,11	61,62
MAP+MLLR_B	68,85	66,19	63,19
MAP+MLLR_mix	68,68	66,14	63,56

Πίνακας 5-7 Αποτελέσματα των ακουστικών μοντέλων μετά από MAP+MLLR προσαρμογή (Acc%)

Παρατηρούμε πως για το **testA**, δηλαδή τα καθαρά ακουστικά σήματα το καλύτερο ακουστικό μοντέλο είναι το **MAP+MLLR_A** και αυτό γιατί με αυτή την μέθοδο adaptation ενισχύεται η ισχύς και η επιρροή του seed μοντέλου μας που ήταν εκπαιδευμένο και αυτό σε καθαρά σήματα φωνής.

Για τα δύο εναπομείναντα test sets τα **testB** και **test_mix** το καλύτερο ακουστικό μοντέλο είναι με διαφορά το **MAP+MLLR_mix** και δικαιολογημένα γιατί είχε τις περισσότερες προτάσεις για adaptation.

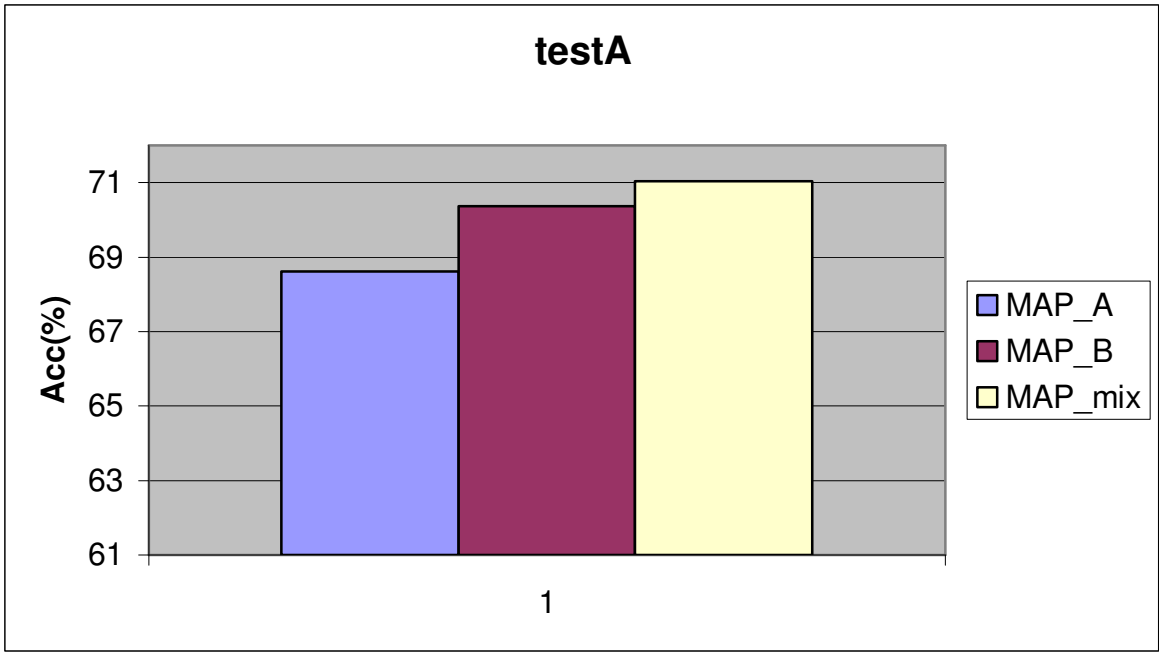
Για να γίνουμε πιο συγκεκριμένοι και για να δείξουμε τις αποδόσεις των ακουστικών μας μοντέλων και σε διαφορετικού είδους σύνολα δοκιμής θα παραθέσουμε παρακάτω 3 πίνακες που παρουσιάζουν την επίδοση των ακουστικών σε 4 διαφορετικά σύνολα δοκιμής που ορίσαμε στην σελίδα 77:

	testA	test_tel_noi	test_st_noi	test_other
MAP+MLLR_A	69,26	34,07	62,41	45,52
MAP+MLLR_B	68,85	45,93	64,67	49,07
MAP+MLLR _mix	68,68	48,15	64,69	51,28

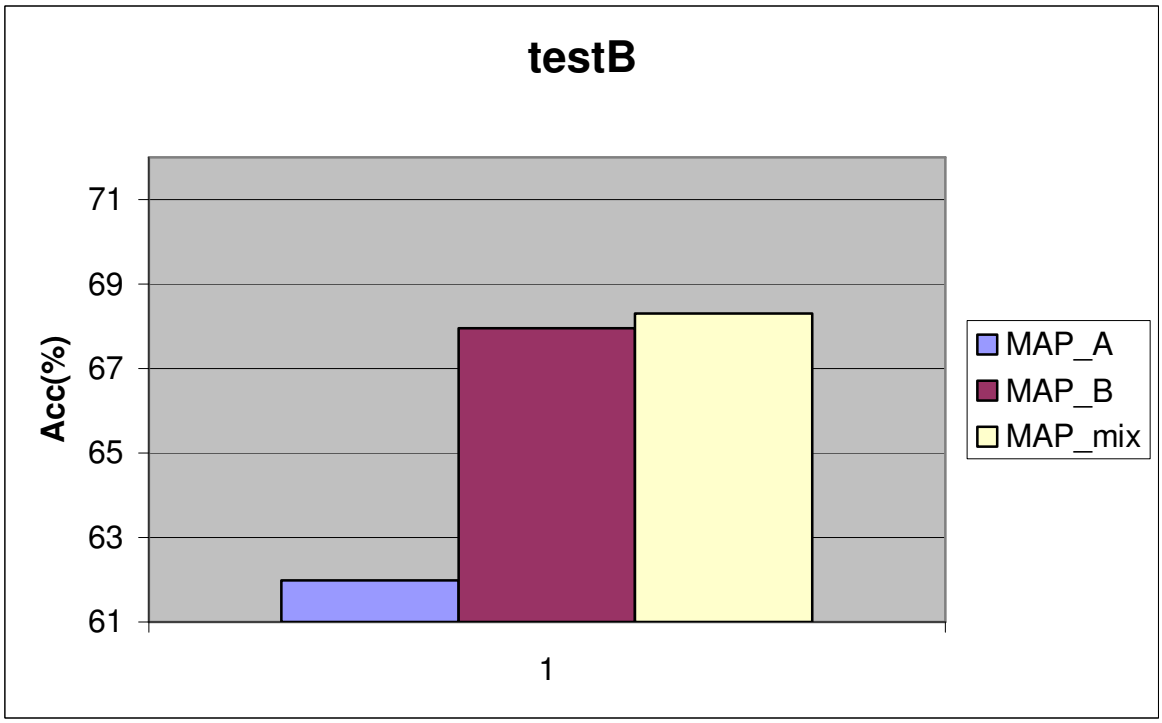
Πίνακας 5-8 Αποτελέσματα ακουστικών μοντέλων μετά από MAP+MLLR προσαρμογή για 4 σύνολα δοκιμών (Acc%)

5.4 Αποτελέσματα των ακουστικών μοντέλων που παράχθηκαν μετά από MAP_adaptation

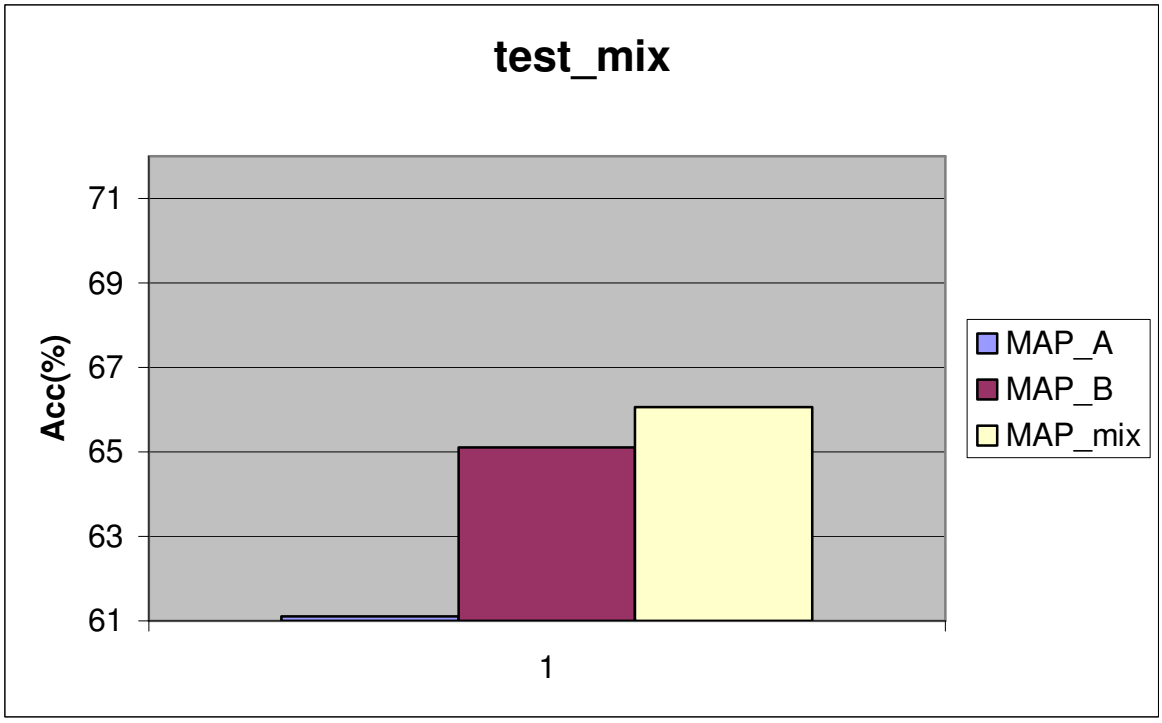
Τα αποτελέσματα της αναγνώρισης των μοντέλων που δημιουργήθηκαν μετά από MAP adaptation στα **testA**, **testB** ,**test_mix** παρουσιάζονται στα σχήματα 5-8 ,5-9 και 5-10 αντίστοιχα.



Σχήμα 5-8 Αποτελέσματα του test_A set



Σχήμα 5-9 Αποτελέσματα του test_B set



Σχήμα 5-10 Αποτελέσματα του test_mix set

Παρακάτω παραθέτουμε έναν πίνακα που παρουσιάζει τα συγκεντρωτικά αποτελέσματα των ακουστικών μοντέλων που παράχθηκαν μετά από MAP προσαρμογή:

	testA	testB	test_mix
MAP_A	68,61	61,98	61,10
MAP_B	70,36	67,96	65,11
MAP_mix	71,04	68,31	66,06

Πίνακας 5-9 Αποτελέσματα των ακουστικών μοντέλων μετά από MAP προσαρμογή (Acc%)

Παρατηρούμε πως το καλύτερο ακουστικό μοντέλο και για τα 3 test sets είναι το **MAP_mix**. Αυτό είναι αναμενόμενο καθώς στο MAP adaptation, ενισχύονται σημαντικά όλα τα τρίφωνα που έχουν εκπαιδευτεί από το seed μοντέλο

αυξάνοντας έτσι σημαντικά τα ποσοστά αναγνώρισης. Έτσι αυτό το μοντέλο έχει τις περισσότερες προτάσεις για adaptation από όλα τα άλλα μοντέλα, γεγονός που το κάνει πιο αξιόπιστο και πιο ακριβή στις προβλέψεις του. Επίσης αξίζει να σημειωθεί πως αυτό το ακουστικό μοντέλο είναι το μοντέλο με τα καλύτερα ποσοστά αναγνώρισης.

Για να γίνουμε πιο συγκεκριμένοι και για να δείξουμε τις αποδόσεις των ακουστικών μας μοντέλων και σε διαφορετικού είδους σύνολα δοκιμής θα παραθέσουμε παρακάτω 3 πίνακες που παρουσιάζουν την επίδοση των ακουστικών σε 4 διαφορετικά σύνολα δοκιμής που ορίσαμε στην σελίδα 77:

	testA	**test_tel_noi**	**test_st_noi**	**test_other**
MAP_A	68,61	26,67	57,57	41,72
MAP_B	70,36	42,22	66,59	51,69
MAP _mix	71,04	45,93	66,71	55,30

Πίνακας 5-10 Αποτελέσματα ακουστικών μοντέλων μετά από MAP προσαρμογή για 4 σύνολα δοκιμών (Acc%)

ΚΕΦΑΛΑΙΟ 6

ΑΝΑΚΕΦΑΛΑΙΩΣΗ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Έχοντας παρουσιάσει όλα τα αποτελέσματα και την απόδοση του συστήματος για όλα τα διαφορετικά ακουστικά μοντέλα, και αφού παρουσιάσαμε και τον τελικό μας αναγνωριστή μπορούμε να πούμε πως έχουμε έναν αξιόπιστο σύστημα αναγνώρισης που απομαγνητοφωνεί αυτόματα ακουστικά σήματα ηχογραφημένα από τηλεοπτικές εκπομπές.

Για την επέκταση αυτού του συστήματος με σκοπό πάντα να αυξήσουμε την απόδοση του θα μπορούσαμε να κάνουμε μια σειρά από ενέργειες. Αρχικά από τις 40 ώρες που έχουμε στην κατοχή μας, μόνο οι 20 τελικά απομαγνητοφωνήθηκαν. Η απόδοση του συστήματος θα αυξηθεί κατά ένα μεγάλο βαθμό αν απομαγνητοφωνήσουμε και τις άλλες 20 ώρες. Έτσι τα ακουστικά μας μοντέλα θα εκπαιδευτούν πολύ καλύτερα πετυχαίνοντας μεγαλύτερα ποσοστά αναγνώρισης. Επίσης υπάρχει πάντα η περίπτωση του να μην έγιναν απολύτως σωστά (με αρκετή ακρίβεια δηλαδή) οι απομαγνητοφωνήσεις των σημάτων μας. Εάν δοθεί μεγαλύτερη προσοχή και ακρίβεια πάνω σε αυτή την εργασία θα έχουμε ακόμα πιο αξιόπιστο σύστημα αναγνώρισης.

Μπορούμε επίσης να αναβαθμίσουμε και το γλωσσικό μας μοντέλο. Μπορούμε να αυξήσουμε το μέγεθος του κειμένου με το οποίο θα εκπαιδευτεί το γλωσσικό μας έτσι ώστε να αυξήσουμε τις πιθανότητες των σωστών λεξικών ακολουθιών. Επιπρόσθετα, θα ήταν εφικτό να κάναμε ένα trigram γλωσσικό μοντέλο (το δικό μας είναι bigram).

Τέλος, μπορούμε να εκπαιδεύσουμε διαφορετικά ακουστικά μοντέλα για κάθε συνθήκη. Δηλαδή θα μπορούμε να εκπαιδεύσουμε ένα ακουστικό μοντέλο μόνο με καθαρά σήματα φωνής, ένα ακουστικό μοντέλο μόνο με σήματα φωνής που περιέχουν θόρυβο κ.ο.κ

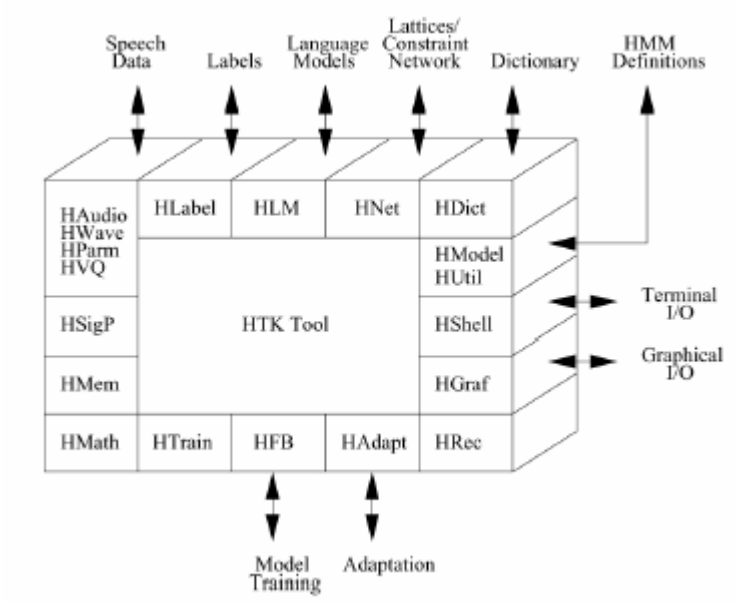
Παράρτημα Α
HTK (Hidden Markov Model Toolkit)

Το HTK (Hidden Markov Model Toolkit) είναι μια συλλογή από προγραμματιστικά εργαλεία για την δημιουργία και τον χειρισμό hidden Markov Models (HMMs). Το HTK ενδείκνυται πρωτίστως για ερεύνα πάνω στην αναγνώριση φωνής, από την στιγμή που μπορεί να μοντελοποιήσει HMMs κάτω από οποιαδήποτε συνθήκες.

Το HTK εξελίχθηκε και δημιουργήθηκε στο *Speech Vision and Robotics Group* του Cambridge University Engineering Department (CUED) για να χτίζει συστήματα αναγνώρισης φωνής μεγάλου λεξιλογίου. Τα δικαιώματα για την πώληση του HTK ανήκουν στην Entropic Research Laboratory Inc. το 1993 και η πλήρης ανάπτυξη του HTK μεταφέρθηκε στο Entropic Research Laboratory Ltd., όταν ιδρύθηκε το 1995. Η Microsoft αγόρασε την Entropic το 1999 και έδωσε το HTK πίσω για ανάπτυξη στο CUED το 2000. Η Microsoft είναι κάτοχος των δικαιωμάτων του κώδικα του HTK, αλλά ο κώδικας είναι ελεύθερα διαθέσιμος για ερευνητικούς σκοπούς.

Η αρχιτεκτονική του HTK

Βασικά, το HTK αποτελείται από μια σειρά εργαλείων, που λειτουργούν μέσω κάποιων βιβλιοθηκών. Αυτές οι βιβλιοθήκες είναι κοινές σε όλα τα εργαλεία, και έτσι εξασφαλίζουν ότι κάθε εργαλείο επικοινωνεί με τον χρήστη με τον ίδιο ακριβώς τρόπο. Επίσης, παρέχουν πρόσβαση, σε κοινές συναρτήσεις. Η αρχιτεκτονική του HTK παρουσιάζεται σχηματικά στο σχήμα Α-1. Η διεπαφή με τον χρήστη αλλά και το λειτουργικό σύστημα ελέγχεται από την *HShell* και η διαχείριση μνήμης γίνεται από την *HMem.* Η μαθηματική υποστήριξη γίνεται από την *HMath* και η λειτουργίες επεξεργασίας σημάτων από την *HSigP.*



Σχήμα Α-1 Αρχιτεκτονική του HTK

Κάθε αρχείο που χρησιμοποιείται από το HTK έχει τις δικές του βιβλιοθήκες και τις δικές του λειτουργικότητες. Η *HLabel* παρέχει το interface για τα label αρχεία, η *HLM* για τα γλωσσικά μοντέλα και τα lattices, η *HDict* για τα λεξικά, η *HVQ* για τα codebooks και η *HModel* για τα HMM definitions.

Η ομιλία χειρίζεται από την *HWave* στο επίπεδο των κυματομορφών και από την *HParm* στο επίπεδο των παραμέτρων. Αυτές οι βιβλιοθήκες υποστηρίζουν ποικίλους τύπους ηχητικών δεδομένων. Η άμεση είσοδος ήχου υποστηρίζεται από την *HAudio* και τα γραφικά από την *HGraf*. Η *Hutil* παρέχει την λειτουργικότητα για τον χειρισμό των HMMs, ενώ η *HTrain* και η HFB παρέχουν τα εργαλεία για την εκπαίδευση των HMMs. Η *HAdapt* παρέχει την δυνατότητα για το adaptation των HMMs. Τέλος, η *HRec* περιέχει τις βασικές συναρτήσεις για την αναγνώριση.

Το σχήμα Α-2 δείχνει ένα παράδειγμα πώς να τρέξουμε ένα τυπικό εργαλείο του HTK. Όλα τα εργαλεία του HTK τρέχουν μέσω command prompt και περιέχουν διάφορες παραμέτρους που αυξάνουν την ευελιξία του εργαλείου. Μερικά ορίσματα που παίρνουν σαν είσοδο μπορούν να είναι και ακέραιοι ή ακόμα και string values. Αν το όνομα μιάς παραμέτρου είναι κέφαλαίο αυτό σημαίνει ότι η συγκεκριμένη παράμετρος είναι η ίδια για όλα τα εργαλεία του

HTK. Στο σχήμα Α-2 η παράμετρος **-T** δείχνει το επιθυμούμενο μέγεθος tracing, και η επιλογή **-S** δείχνει πιο αρχείο θα χρησιμοποιηθεί το οποίο έχει την λίστα με τα αρχεία εισόδου. Το γεγονός ότι το HTK παρέχει text-based command prompt interface έχει αρκετά πλεονεκτήματα: επιτρέπει στα shell scripts να ελέγχουν την λειτουργία του εργαλείου, και αυτό είναι πολύ χρήσιμο όταν χτίζουμε συστήματα μεγάλης κλίμακας που απαιτούν πληθώρα αρχείων.

```
HVite -S mfc_list -i labels.mlf -T 01 -w lattice dict tri
```

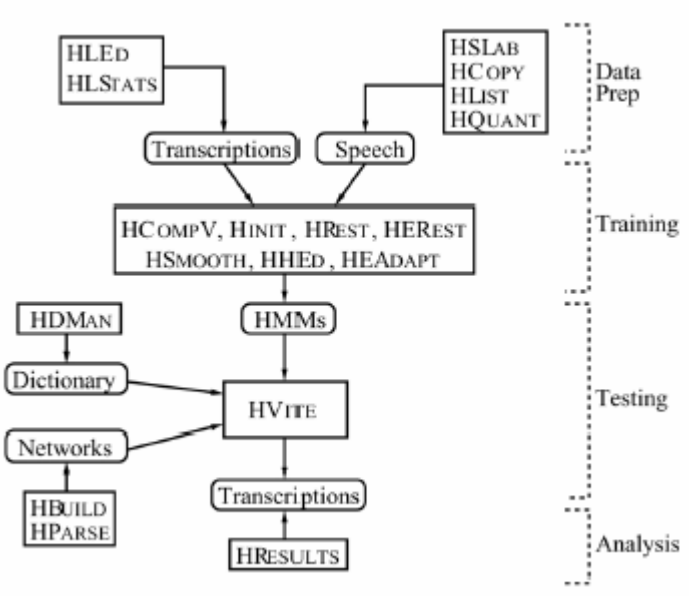
Σχήμα Α-2 Εκτελώντας μια HTK εντολή

Τα εργαλεία του HTK

Εδώ θα περιγραφούν τα εργαλεία του HTK. Τα εργαλεία χωρίζονται σε τέσσερις κατηγορίες που ανταποκρίνονται στις τρεις κύριες φάσεις για την δημιουργία ενός αξιόπιστου συστήματος αναγνώρισης φωνής. Αυτές οι φάσεις είναι οι εξής:

- Επεξεργασία των δεδομένων (Data Preparation)
- Εκπαίδευση (Training)
- Αξιολόγηση του συστήματος (Evaluation)

Τα διάφορα εργαλεία του HTK και ένα συνοπτικό διάγραμμα της χρήσης τους δίνεται στο σχήμα Α-3



Σχήμα Α-3 Λειτουργία του HTK

Εργαλεία προεπεξεργασίας δεδομένων

Προκειμένου να υλοποιήσουμε ένα σύστημα αναγνώρισης χρειάζονται κάποια ακουστικά δεδομένα και κάποιες απομαγνητοφωνήσεις τους. Μια τυπική βάση δεδομένων που περιέχει σήματα φωνής, συνήθως περιέχει αρχεία ήχου από πολλαπλούς ομιλητές, και στις περισσότερες περιπτώσεις είναι αρκετά μεγάλη. Για να μπορέσουμε να εκπαιδεύσουμε σωστά HMMs, πρέπει να μετατρέψουμε τα ακουστικά μας σήματα σε μια συγκεκριμένη μορφή και επίσης θα πρέπει να μετατρέψουμε και τις απομαγνητοφωνήσεις μας σε ένα συγκεκριμένο format. Το HTK μας περιέχει το εργαλείο *HLab* για να ηχογραφούμε ακουστικά σήματα φωνής και παράλληλα να απομαγνητοφωνούμε τα ίδια σήματα φωνής.

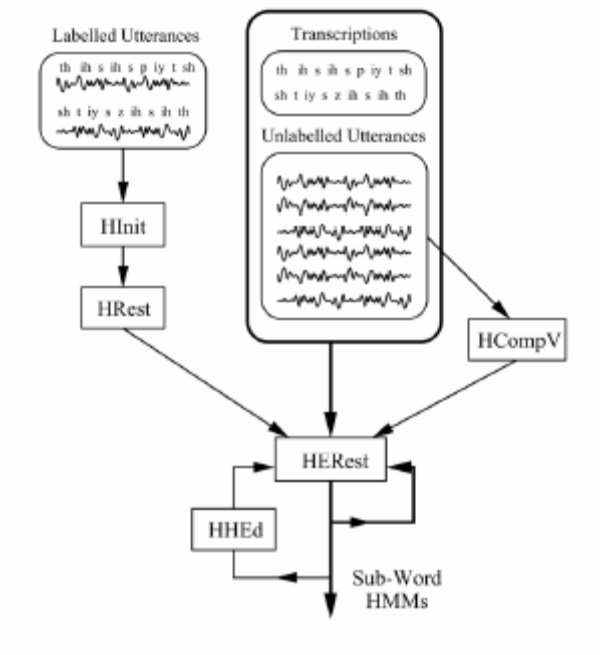
Για να παραμετροποιήσουμε τα αρχεία ήχου χρησιμοποιούμε την *HCopy*. Αυτή η εντολή μετατρέπει το αρχείο ήχου σε κατάλληλες μαθηματικές παραμέτρους έτσι ώστε να μπορούμε να τα επεξεργαστούμε. Έπειτα το εργαλείο *HList* μπορεί να χρησιμοποιηθεί προκειμένου να ελέγξουμε τα περιεχόμενα των αρχείων ήχου και την παραμετρική τους μετατροπή.

Οι απομαγνητοφωνήσεις χρειάζονται και αυτές κάποιες τροποποιήσεις. Για να τις μετατρέψουμε στο κατάλληλο format του HTK χρησιμοποιούμε την *HLed*. Αυτή η εντολή είναι ένας editor που παίρνει σαν είσοδο όλες τις απομαγνητοφωνήσεις που έχουμε στην κατοχή μας και βγάζει ως έξοδο ένα Master Label File (MLF) που περιέχει όλες τις απομαγνητοφωνήσεις μαζί. Ακόμα δύο εργαλείο είναι το *HLStats* που μπορεί να συγκεντρώσει και να παρουσιάσει στατιστικά στοιχεία για τα label files, και η *HQuant* που μπορεί να υλοποιήσει ένα VQ codebook για την δημιουργία διακριτών HMM συστημάτων.

Εργαλεία εκπαίδευσης

Το επόμενο βήμα για την δημιουργία ενός συστήματος φωνής είναι να προσδιορίσουμε την τοπολογία του κάθε HMM σε ένα prototype definition. Το HTK επιτρέπει την δημιουργία των HMMs με τυχαία τοπολογία. Τα HMM prototype definitions αποθηκεύονται σε αρχεία κειμένου και μπορούν να επεξεργαστούν με έναν απλό text editor. Η βασική λειτουργία του prototype definition είναι να προσδιορίζει τα καθολικά χαρακτηριστικά των HMM. Πρέπει να επιλεγούν λογικές τιμές για τις πιθανότητες μετάβασης από κατάσταση σε κατάσταση ή από μια κατάσταση στην ίδια κατάσταση. Μια απλή στρατηγική είναι να δοθούν οι ίδιες πιθανότητες μεταβάσεων για όλες τις περιπτώσεις.

Η εκπαίδευση των HMM γίνεται σταδιακά σε, πολλά στάδια όπως φαίνεται και στο σχήμα Α-4. Το πρώτο στάδιο είναι η δημιουργία των αρχικών μοντέλων. Αν υπάρχουν κάποια δεδομένα εκπαίδευσης (bootstrap data) που να έχουν απομαγνητοφωνηθεί, αυτά μπορούν να χρησιμοποιηθούν για την αρχική εκπαίδευση. Σε αυτήν την περίπτωση χρησιμοποιούνται η *HInit* και η *HRest*. Κάθε HMM δημιουργείται ανεξάρτητα και αυτόνομα. Η *HInit* διαβάζει όλα τα bootstrap data και αρχικοποιεί όλα τα φωνήματα μας. Μετά, τα μοντέλα εκπαιδεύονται επαναληπτικά με τον αλγόριθμο του με την *HRest*. Αν δεν υπάρχουν bootstrap data τότε μπορούμε να χρησιμοποιήσουμε το εργαλείο *HCompV*. Σε αυτή την περίπτωση όλα τα HMM αρχικοποιούνται με τις ίδιες τιμές.



Σχήμα Α-4 Εκπαίδευση των HMM

Αφού πλέον έχουν δημιουργηθεί τα αρχικά μας μοντέλα, το εργαλείο *HERest* χρησιμοποιείται για να εκπαιδεύσει τα HMM με όλα τα δεδομένα μας. Η *HERest* εφαρμόζει τον αλγόριθμο Baum-Welch για όλο το set των HMM ταυτόχρονα. Για κάθε δεδομένο εκπαίδευσης, τα αντίστοιχα φωνητικά μοντέλα συνδέονται και ο forward-backward αλγόριθμος χρησιμοποιείται για να συγκεντρώσει τα απαραίτητα στατιστικά στοιχεία (means, variances κλπ.) για κάθε HMM. Αφού προσπελαστούν όλα τα δεδομένα εκπαίδευσης, τα συγκεντρωμένα στατιστικά στοιχεία χρησιμοποιούνται για να υπολογιστούν και πάλι οι παράμετροι των HMMs. Η *HERest* είναι το κύριο εργαλείο του HTK και έχει μπορούμε να ρυθμίσουμε πάνω σε αυτό πολλές επιλογές, όπως pruning κλπ.

Τυπικά τα πρώτα HMMs που παράγονται είναι μονόφωνα (monophones) μοντέλα που περιέχουν ένα μείγμα Γκαουσιανής ανά κατάσταση. Έτσι με βάση αυτά τα μονόφωνα HMMs μπορούν να δημιουργηθούν δίφωνα και τρίφωνα HMMs με πολλά μείγματα Γκαουσιανών ανά κατάσταση. Για αυτές τις διεργασίες χρησιμοποιείται το εργαλείο *HHed*. Επίσης μπορούμε να κάνουμε adaptation με τα εργαλεία *HEAdapt* και HVite μόλις με λίγα adaptation δεδομένα.

Ένα από τα μεγαλύτερα και πιο συχνά προβλήματα στην εκπαίδευση των HMM είναι η έλλειψη επαρκών δεδομένων. Όσο πιο πολύπλοκα και πολυάριθμα είναι τα HMM που θέλουμε να εκπαιδεύσουμε τόσα πιο πολλά δεδομένα χρειαζόμαστε. Για αυτό μπορούμε να δημιουργήσουμε tied-state μοντέλα έτσι ώστε να μοιράζεται η πληροφορία των HMMs μεταξύ τους και να κάνουν πιο καλούς και σωστούς υπολογισμούς. Το εργαλείο *HSmooth* μπορεί να χρησιμοποιειθεί σε αυτές τις περιπτώσεις για να επιδείξει την έλλειψη δεδομένων.

Εργαλεία αξιολόγησης του συστήματος

Το εργαλείο αναγνώρισης που παρέχεται από το HTK είναι το HVite. Αυτό χρησιμοποιεί ένα αλγόριθμο που λέγεται token passing algorithm για να εφαρμόσει Viterbi-based αναγνώριση ομιλίας. Σαν είσοδο η HVite χρειάζεται ένα γλωσσικό μοντέλο που να προσδιορίζει τις επιτρεπόμενες ακολουθίες λέξεων, ένα λεξικό που να περιέχει την προφορά κάθε λέξης και την λίστα των HMMs. Η HVite μετατρέπει το δίκτυο των λέξεων σε δίκτυο από φωνήματα και επικολλά το κατάλληλο HMM σε κάθε φωνητικό στιγμιότυπο.

Τα γλωσσικά μοντέλα αποθηκεύονται σε ένα ειδικό format του HTK, το lattice και μπορεί να προσπελαστεί από έναν απλό text editor. Το HTK παρέχει δύο εργαλεία για το χτίσιμο τέτοιων γλωσσικών μοντέλων. Την *HBuild* και την *HParse*. Η *HBuild* επιτρέπει την δημιουργία υπο-δικτύων που μπορούν να χρησιμοποιηθούν σε δίκτυα υψηλού επιπέδου και μπορούν να συμβάλλουν στην δημιουργία κυκλικών δικτύων από λέξεις (word loop). Η *HParse* μπορεί να μετατρέψει δίκτυα που είναι γραμμένα σε υψηλό επίπεδο γραμματικής σε lattice format. Αυτό το υψηλό επίπεδο γραμματικής βασίζεται στο Extended Backus Naur Form (EBNF).

Για να δούμε παραδείγματα πιθανών διαδρομών που εμπεριέχονται σε ένα τέτοιο δίκτυο μπορούμε να χρησιμοποιήσουμε το εργαλείο *HSGen*. Αυτό το εργαλείο παίρνει ως είσοδο το δίκτυο και βγάζει σαν έξοδο συμβολοσειρές λέξεων. Αυτές οι συμβολοσειρές μπορούν ύστερα να ελεγχθούν και έτσι να

επιβεβαιώσουμε πως το δίκτυο μας είναι σωστά εκπαιδευμένο. Τέλος μπορούμε να δημιουργήσουμε το λεξικό προφορών μας με την *HDMan*.

Εν τέλει, μπορούμε με την *HResults*, να πάρουμε σαν είσοδο το αποτέλεσμα της αναγνώρισης της HVite και να το συγκρίνουμε με τα σωστά τις σωστές απομαγνητοφωνήσεις έτσι ώστε να δούμε κατά πόσο τις εκατό ήταν σωστά τα αποτελέσματα της αναγνώρισης.

Παράρτημα Β
Transcriber Tool

Όπως ήδη έχουμε αναφέρει, χρησιμοποιήσαμε το εργαλείο **Transcriber Tool** για την απομαγνητοφώνηση των ακουστικών μας σημάτων.

Το Transcriber είναι ένα εργαλείο για τον χειροκίνητο χαρακτηρισμό των σημάτων φωνής. Διαθέτει γραφικό περιβάλλον, φιλικό προς τον χρήστη για τον τεμαχισμό, την απομαγνητοφώνηση και τον χαρακτηρισμό, μεγάλων σε διάρκεια ακουστικών σημάτων. Έχει σχεδιαστεί ειδικά για την δημιουργία βάσης δεδομένων από ακουστικά σήματα ηχογραφημένων από ραδιοτηλεοπτικές εκπομπές (broadcast news) αλλά οι λειτουργίες αυτού του εργαλείου μπορούν να φανούν χρήσιμες και σε άλλες ερευνητικές περιοχές.

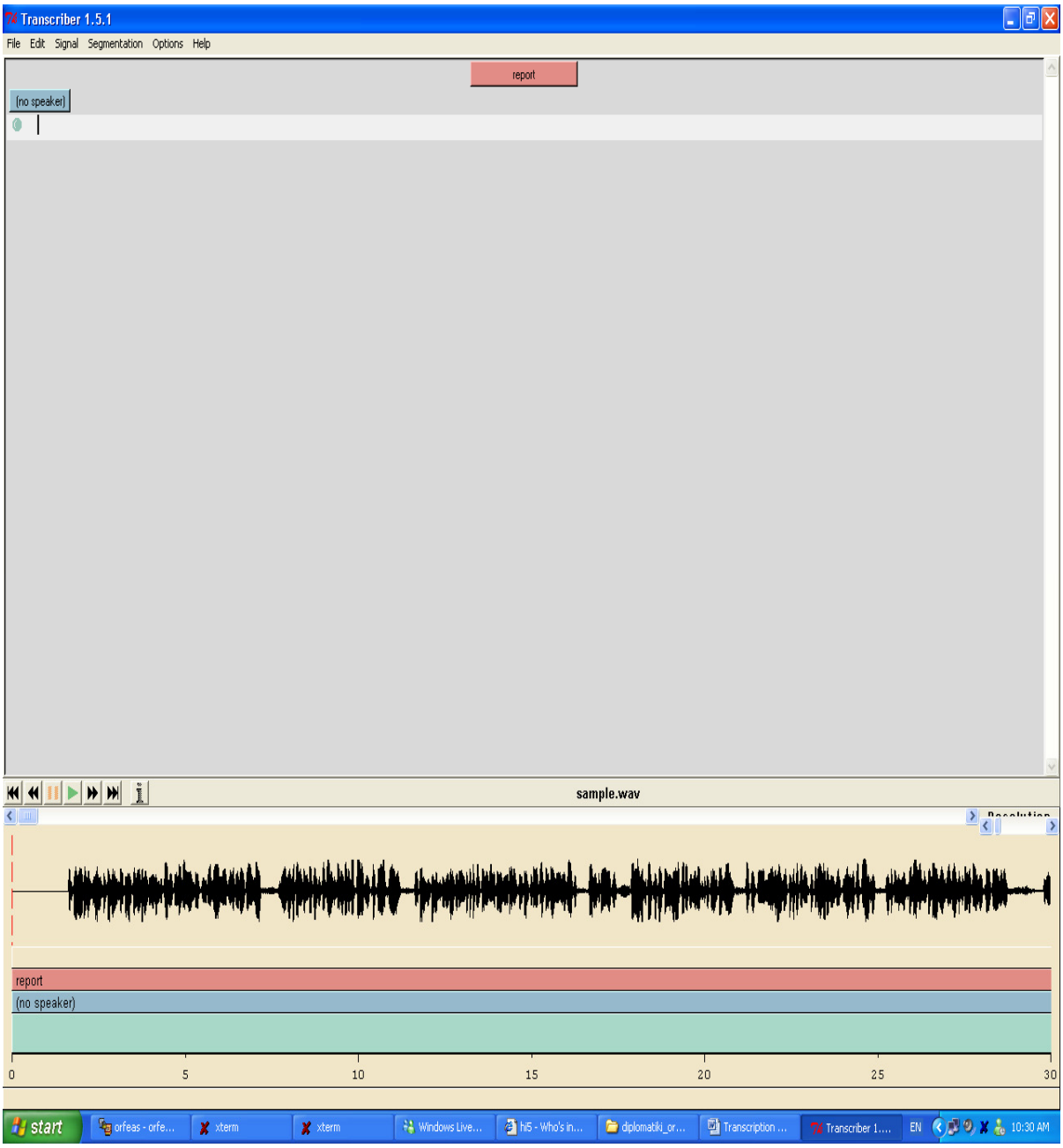
Σε αυτό το παράρτημα θα παρουσιάσουμε περιληπτικά πως χρησιμοποιήσαμε το Transcriber Tool για την απομαγνητοφώνηση των ακουστικών μας σημάτων. Μπορείτε να ανατρέξετε στο Transcriber's Interface και στο Transcriber's User's Manual αν επιθυμείτε να δείτε αναλυτικά όλες τις δυνατότητες του συγκεκριμένου εργαλείου.

Βήμα πρώτο – Άνοιγμα ενός αρχείου ήχου

Για να ξεκινήσουμε την διαδικασία της απομαγνητοφώνησης των αρχείων ήχου που έχουμε στην κατοχή μας, θα πρέπει αρχικά να φορτώσουμε το ακουστικό σήμα προς απομαγνητοφώνηση πατώντας **File -> Open audio file** ή πατώντας **Ctrl + a**.

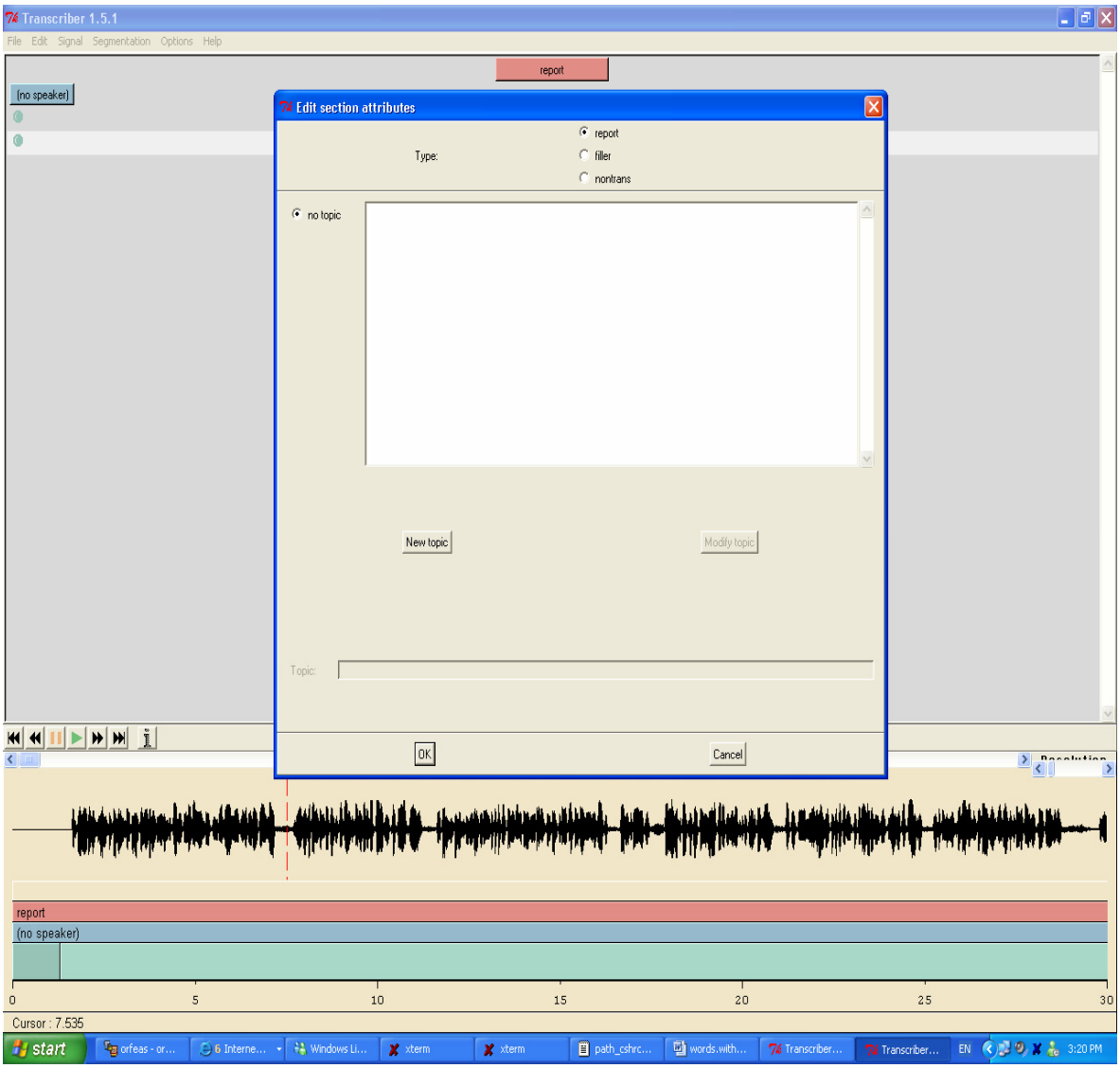
Όταν φορτωθεί το αρχείου ήχου, θα παρουσιαστεί στο κάτω μέρος της οθόνης το σήμα φωνής μας. Μπορείτε να ακούσετε ολόκληρο το σήμα φωνής, ή απλά

ένα μέρος του με τα κουμπιά που βρίσκονται στο πάνω μέρος του σήματος φωνής που μόλις ανοίχτηκε.



Βήμα δεύτερο – Δημιουργία τομέα

Το δεύτερο βήμα μας είναι να δημιουργήσουμε ένα τομέα, στον οποίο θα έχουμε την ίδια συνθήκη ομιλίας. Αυτό μπορούμε να το κάνουμε πατώντας **Segmentation -> Create Section** ή πατώντας **Ctrl + r**.

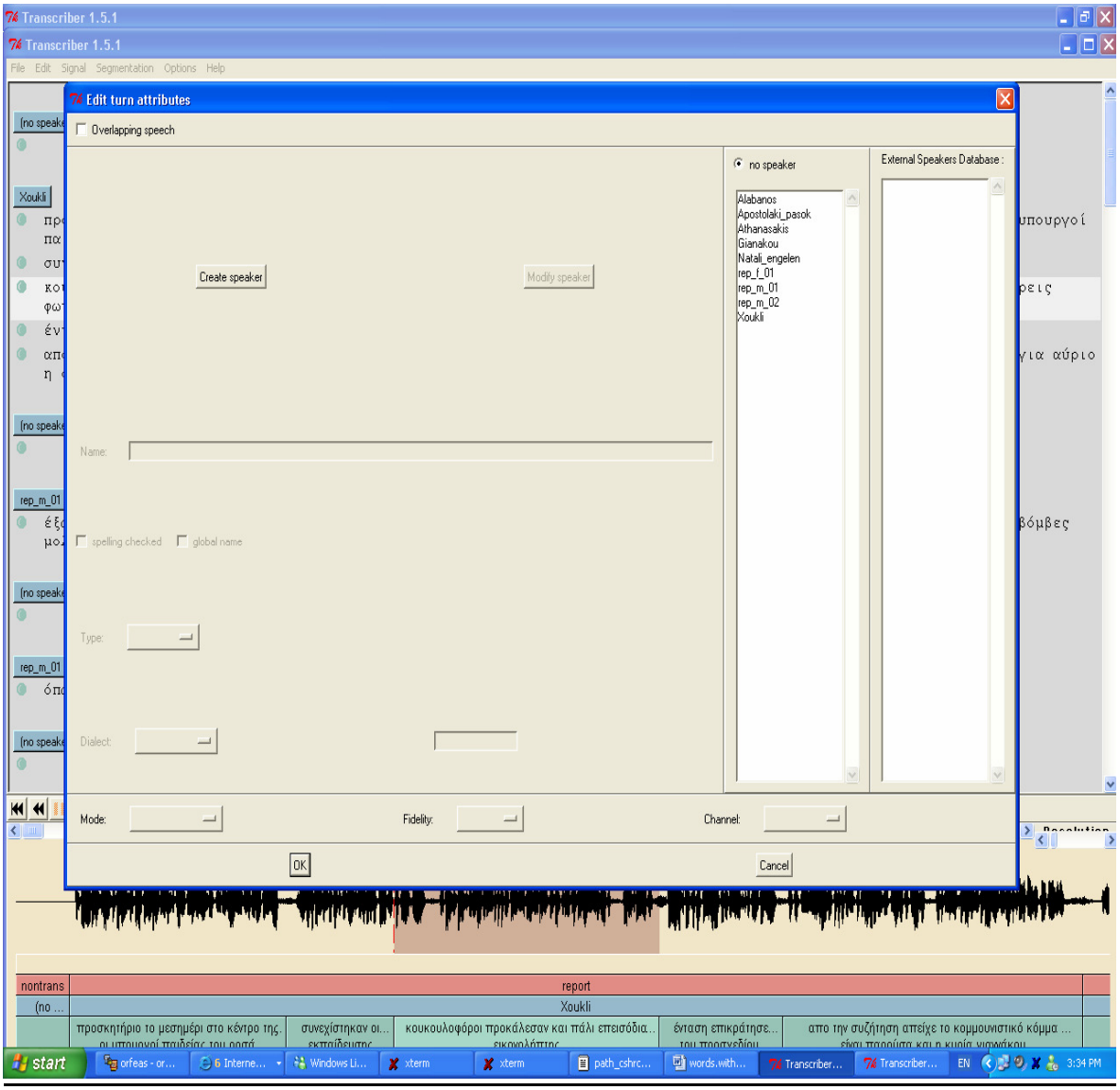


Για περισσότερες πληροφορίες για αυτήν την διαδικασία μπορείτε να ανατρέξετε στο **υποκεφάλαιο 2.3.1**, σελ. 31.

Βήμα τρίτο – Δημιουργία ομιλητή

Το τρίτο βήμα της διαδικασίας της απομαγνητοφώνησης είναι η δημιουργία ενός ομιλητή ή η επιλογή ενός ομιλητή από μια λίστα πιθανών ομιλητών που έχουμε δημιουργήσει και συμπεριλαμβάνει δημοσιογράφους, πολιτικούς, κλπ.

Αυτό μπορεί να γίνει πατώντας **Segmentation - >Create Turn** ή πατώντας **Ctrl + t**.



Για περισσότερες πληροφορίες για αυτήν την διαδικασία μπορείτε να ανατρέξετε στο **υποκεφάλαιο 2.3.2**, σελ. 32.

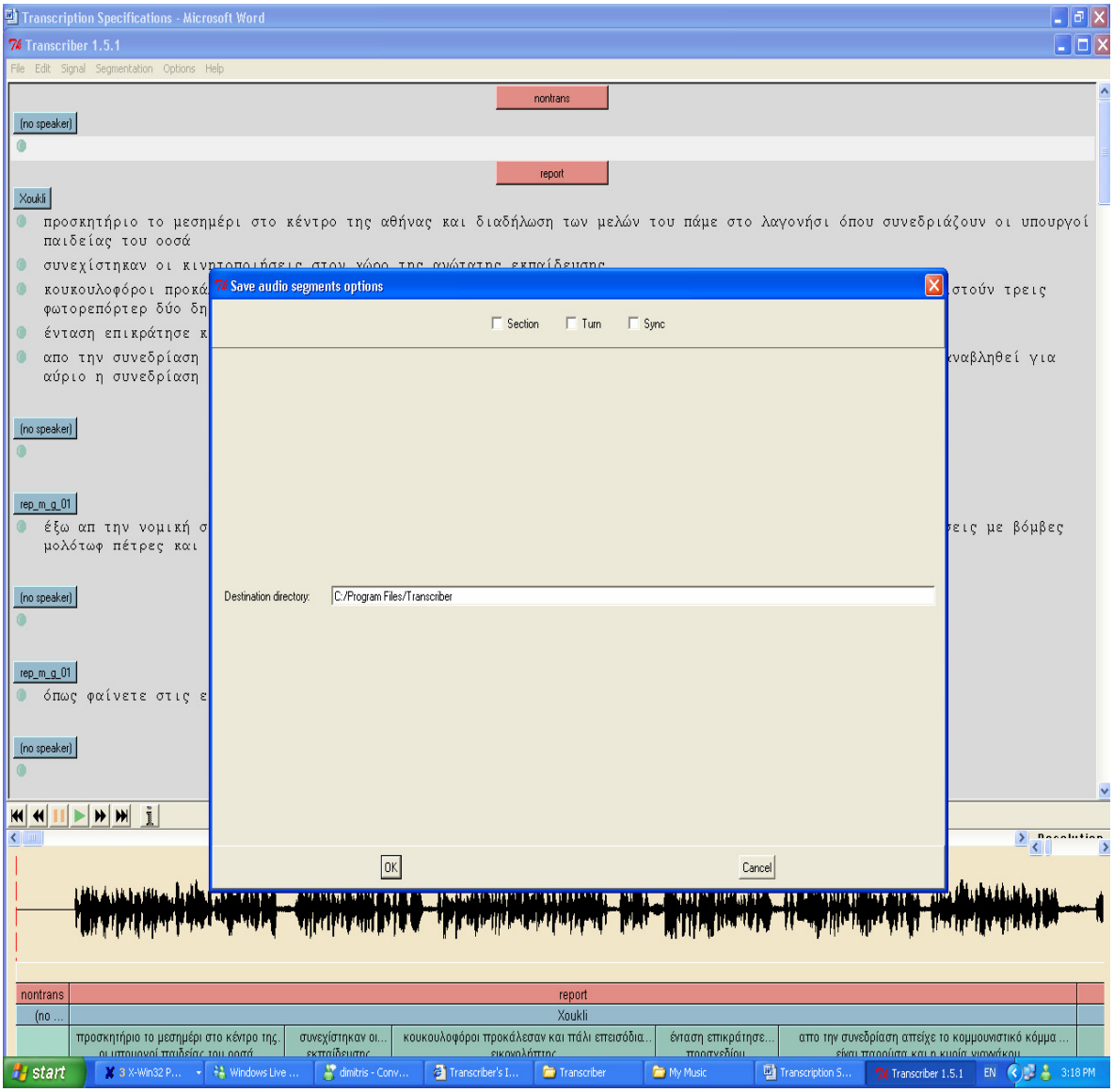
Βήμα τέταρτο – Αποτύπωση της απομαγνητοφώνησης

Εφόσον πλέον έχετε επιλέξει την συνθήκη ομιλίας καθώς και τον ομιλητή, σε αυτό το σημείο μπορείτε να γράψετε την απομαγνητοφώνηση για το εκάστοτε κομμάτι (segment) του σήματος φωνής.

Επιλέγοντας πρώτα το κομμάτι που θα απομαγνητοφωνήσετε, ύστερα θα πρέπει να κάνετε κλικ σε ένα σημείο του σήματος φωνής. Αυτό το σημείο θα είναι το τέλος του κομματιού και μετά από αυτό θα απομαγνητοφωνηθεί άλλο segment του ακουστικού σήματος. Έχοντας επιλέξει ποιο θα είναι το όριο του κομματιού μας, μετά γράφουμε την απομαγνητοφώνηση του κομματιού σύμφωνα με του κανόνες που αναφέραμε στο **υποκεφάλαιο 2.3.3**, σελ. 34, και τέλος πατάμε **Enter**.

Βήμα πέμπτο – Δημιουργία των segments

Εφόσον έχουμε απομαγνητοφωνήσει ένα αρχείου ήχου σύμφωνα με τα παραπάνω βήματα, τώρα θα πρέπει να εξάγουμε τα segment έτσι ώστε να δημιουργήσουμε την βάση των ακουστικών μας σημάτων με την οποία θα δημιουργήσουμε το σύστημα αναγνώρισης. Πατάμε **File -> Save audio segment(s) -> Automatic**.



Τέλος πατάμε Turn και Sync για να τεμαχίσουμε το σήμα φωνής.

Βιβλιογραφία

[1] S.J. Young, *Large Vocabulary Continuous Speech Recognition: a Review,* Cambridge University Engineering Department, April 1996

[2] S.J. Young, *The HTK Book,* Cambridge University Engineering Department, December 2003

[3] Reinhold Haeb-Umbach, *Acoustic Modeling in the Philips Hub-4 Continuous-Speech Recognition System*, Philips Research Laboratories, 1998

[4] L. R. Rabiner, *Fundamentals of Speech Recognition*, Prentice Hall, 1993

[5] M.A Spaans, *On Developing Acoustic Models Using HTK*, Delft University of Technology, December 2004

[6] Raimo Bakis, Scott Chen, Ponami Gopalakrishnan, Ramesh Gopinath, Stephane Maes and Lazaros Polymenakos, *Transcription of Broadcast News – System Robustness Issues and Adaptation Techniques*, Human Language Technologies, Computer Science Department, IBM, November 1995

[7] Man-Wai Mak, Roger Hsiao and Brian Mak, *A Comparison of Various Adaptation Methods for Speaker Verification with Limited Enrollment Data*, The Hong Kong Polytechnic University

[8] http://www.speech.sri.com/projects/srilm/

[9] Enrico Bocchieri, Michael Riley and Murat Saraclar, *Methods for Task Adaptation of Acoustic Models with Limited Transcribed In-Domain Data*, AT&T Labs-Research

[10] Prabhu Raghavan, *Speaker and Environment Adaptation in Continuous Speech Recognition*, CoRE Building, New Jersey, June 1998

[11] http://www.virtualdub.org/

[12] http://trans.sourceforge.net/

[13] Βασίλης Διγαλάκης, *Σημειώσεις του μαθήματος "Εισαγωγή στην Επεξεργασία Φωνής"*, http://www.telecom.tuc.gr/vas/classes/speech/index.html

[14] J. R. Deller, J. G. Proakis and J. L. Hansen, *Discrete-Time Processing of Speech Signals* , Macmillan, 1993

[15] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978

[16] R. A. Cole, *Survey of the State of the Art in Human Language Technology*, http://www.cse.ogi.edu/CSLU/HTLsurvey/

[17] Xuelin Cheng, Han Wang and Zongge Li, Speech Adaptation Using Neural Networks foe Connected Digit Recognition, Department of Computer Science Fudan University