ΑΥΤΟΜΑΤΗ ΑΠΟΜΑΓΝΗΤΟΦΩΝΗΣΗ ΣΗΜΑΤΩΝ VIDEO

ΤΣΙΝΑΣ ΜΙΧΑΗΛ

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Ηλεκτρονικών Μηχανικών
και Μηχανικών Υπολογιστών

Επιβλέπων : Διγαλάκης Βασίλης, Καθηγητής

Χανιά, Ιανουάριος 2011.

ΑΥΤΟΜΑΤΗ ΑΠΟΜΑΓΝΗΤΟΦΩΝΗΣΗ ΣΗΜΑΤΩΝ VIDEO

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΤΣΙΝΑ ΜΙΧΑΗΛ

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Ηλεκτρονικών Μηχανικών

και Μηχανικών Υπολογιστών

Εξεταστική επιτροπή:

Διγαλάκης Βασίλης	Ποταμιάνος Αλέξανδρος	Λαγουδάκης Μιχαήλ
Καθηγητής	Αναπληρωτής Καθηγητής	Επίκουρος Καθηγητής

Στους γονείς μου

Περίληψη

Η συνεχής εξέλιξη της τεχνολογίας είχε άμεση επίπτωση στον τομέα της επεξεργασίας φωνής και ειδικότερα της αναγνώρισης φωνής. Τα νέα συστήματα αυξάνουν συνεχώς την υπολογιστική δυνατότητα των αλγορίθμων επεξεργασίας φωνής και παρέχουν νέες δυνατότητες και πιο γρήγορες υπηρεσίες. Η παρούσα διπλωματική εργασία ασχολείται με τη βελτιστοποίηση ενός συστήματος αναγνώρισης ηχητικών σημάτων από τηλεοπτικές εκπομπές. Αρχικά μας δοθήκαν δυο βασικά μοντέλα ανεξάρτητα από τον ομιλητή (Speaker Independed, SI) και εκπαιδευμένα με τα κριτήρια *Maximum likelihood Estimation (MLE)* και *Maximum Mutual Information (MMI)* με δεδομένα από ελληνικές ηλεκτρονικές εφημερίδες. Κατά την αξιολόγηση των βασικών μοντέλων η καλύτερη απόδοση επιτεύχθηκε από το μοντέλο εκπαιδευμένο με το MMI κριτήριο. Στη συνέχεια χρησιμοποιήθηκαν δεδομένα από τηλεοπτικά δελτία ειδήσεων για να προσαρμόσουν τα δυο βασικά μοντέλα. Συγκεκριμένα, εφαρμόστηκαν διάφορες τεχνικές προσαρμογής (MAP,MLLR) και σε διαφορετικό επίπεδο (Ομιλητές, Συνθήκη ομιλίας). Η προσαρμογή στα μοντέλα αυτά έγινε αρχικά στο επίπεδο συνθήκης ομιλίας εφαρμόζοντας MAP Adaptation και βελτιώνοντας την απόδοση της ακρίβειας για το MLE μοντέλο στην καλύτερη περίπτωση έως και 30% ενώ για το MMI έως και 16%. Έπειτα, υλοποιηθήκαν μοντέλα προσαρμοσμένα και στις συνθήκες ομιλίας(MAP Adaptation) αλλά και στους ομιλητές(MLLR Adaptation) που σκοπό είχαν την βελτίωση της ακρίβειας του βασικού μοντέλου εκπαιδευμένο με το MMI κριτήριο. Ύστερα από την υλοποίηση του τελικού πειράματος επήλθε μια βελτίωση για το βασικό MMI μοντέλο η οποία έφτασε στην καλύτερη περίπτωση έως και 26%.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κύριο Βασίλειο Διγαλάκη, ο οποίος μου εμπιστεύτηκε το θέμα της παρούσας διπλωματικής εργασίας. Η βοήθεια και η κατανόηση του στα διάφορα μικρά ή μεγάλα προβλήματα που παρουσιάστηκαν όλον αυτόν τον καιρό, στάθηκε πολύτιμη για την ολοκλήρωση αυτής της εργασίας. Επίσης τον ευχαριστώ πολύ που μου έδωσε την δυνατότητα μέσω της συνεργασίας μας να αποκτήσω γνώσεις και εμπειρίες πάνω σε ένα από τα πιο ενδιαφέροντα θέματα όπως είναι η επεξεργασία φωνής.

Επιπλέον, θα ήθελα να ευχαριστήσω τον κύριο Βασίλη Διακολουκά, ο οποίος ήταν διπλά μου από την πρώτη έως την τελευταία στιγμή της συνεργασίας μας. Η καθοδήγηση και υπομονή του σε όλες τις απορίες μου όσο σημαντικές ή ασήμαντες και αν ήταν στάθηκαν καταλυτικές για την ολοκλήρωση των πειραμάτων.

Τέλος, ευχαριστώ την οικογένεια μου για την υλική και ηθική συμπαράσταση τους σε όλα τα χρόνια της φοιτητικής μου σταδιοδρομίας, όπως επίσης και τους φίλους μου για την υποστήριξη και την βοήθεια τους μέσα και έξω από την σχολή.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ **5**

Λίστα Πινάκων **7**

Λίστα Σχημάτων **9**

ΕΙΣΑΓΩΓΗ **10**

Το θέμα της διπλωματικής 12

Δομή της διπλωματικής 13

ΚΕΦΑΛΑΙΟ 1 .... **14**

ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΑΝΑΓΝΩΡΙΣΗ ΟΜΙΛΙΑΣ **14**

1.1 *Εισαγωγή* 14

1.2 *Η Front –End επεξεργασία* 15

1.3 *Κριτήριο Αναγνώρισης* 18

1.4 *Περιγραφή των Hidden Markov Models* 21

1.4.1 *Τα είδη των HMMs* 22

1.5 *Η Μέθοδος της Προσαρμογής (Adaptation)* 24

1.5.1 *Γενικά* 24

1.5.2 *MAP Adaptation (Maximum a Posteriori)* 25

1.5.3 *MLLR Adaptation (Maximum Likelihood Linear Regression)* 27

ΚΕΦΑΛΑΙΟ 2 .... **28**

ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ **28**

2.1 *Εισαγωγή* 28

2.2 *Τα ακουστικά σήματα* 28

2.3 *Η Διαδικασία της Απομαγνητοφώνησης* 29

2.3.1 *Επίπεδο συνθήκης ομιλίας* 30

2.3.2 *Επίπεδο ομιλητή* 30

2.3.3 *Επίπεδο απομαγνητοφώνησης* 31

2.4 *Δημιουργία της τελικής βάσης δεδομένων* 33

2.4.1 *Επεξεργασία δεδομένων* 33

2.4.2 *Τα δεδομένα εκπαίδευσης(Train Data)* 35

2.4.3 *Τα δεδομένα έλεγχου(Test Data)* 37

ΚΕΦΑΛΑΙΟ 3 .... ***38***

ΤΑ ΒΑΣΙΚΑ ΜΟΝΤΕΛΑ ΑΝΑΓΝΩΡΙΣΗΣ ***38***

3.1 Εισαγωγή *38*

3.2 Σχεδιαστικές επιλογές ακουστικών μοντέλων *38*

3.2.1 Επεξεργασία Front-End *38*

3.2.2 Επιλογή γλωσσικής μονάδας *40*

3.2.3 Τοπολογία HMMs μοντέλων *42*

3.3 Εκπαίδευση ακουστικών μοντέλων *43*

3.4 Τα Βασικά μοντέλα και η MAP προσαρμογή *46*

3.5 Η Αξιολόγηση των μοντέλων *47*

3.5.1 Αξιολόγηση των βασικών μοντέλων (Seed Models) *48*

3.5.2 Αξιολόγηση των μοντέλων μετά από MAP Adaptation *49*

ΚΕΦΑΛΑΙΟ 4 .... ***51***

Η ΜΕΘΟΔΟΣ ΤΗΣ ΠΡΟΣΑΡΜΟΓΗΣ ΣΤΟΝ ΟΜΙΛΗΤΗ ***51***

4.1 Εισαγωγή *51*

4.2 Τα προσαρμοσμένα μοντέλα στον ομιλητή *52*

4.3 Η Αξιολόγηση των μοντέλων *55*

4.3.1 Αξιολόγηση μετά από Supervised Speaker Adaptation *55*

4.3.2 Αξιολόγηση μετά από Unsupervised Speaker Adaptation *61*

ΚΕΦΑΛΑΙΟ 5 .... ***74***

Ανακεφαλαίωση και Μελλοντικές Επεκτάσεις ***74***

Παράρτημα Α ***76***

HTK (Hidden Markov Model Toolkit) ***76***

Παράρτημα Β ***80***

ΠΑΡΑΔΕΙΓΜΑΤΑ ***80***

Βιβλιογραφία ***84***

Λίστα Πινάκων

Πίνακας 2.1 : Συνθήκες ομιλίας 30

Πίνακας 2.2 : Χαρακτηριστικά ομιλητών 31

Πίνακας 2.3 : Συμβάντα κατά την διάρκεια της ομιλίας 32

Πίνακας 2.4 : Μορφή εξόδου επεξεργασμένων Transcriptions 34

Πίνακας 2.5: Κατηγορίες δεδομένων 34

Πίνακας 2.6: Ποσοτική περιγραφή των δεδομένων κάθε κατηγορίας 34

Πίνακας 2.7 : Ποσοτική περιγραφή των δεδομένων εκπαίδευσης σε επίπεδο συνθήκης ομιλίας 35

Πίνακας 2.8 : Ποσοτική περιγραφή των δεδομένων εκπαίδευσης σε επίπεδο ομιλητή 36

Πίνακας 2.9 : Ποσοτική περιγραφή των δεδομένων ελέγχου σε επίπεδο συνθήκης ομιλίας 37

Πίνακας 2.10 : Ποσοτική περιγραφή των δεδομένων ελέγχου σε επίπεδο ομιλητή 37

Πίνακας 3.1 : Τα φωνήματα του συστήματος 41

Πινάκας 3.2 : Συνοπτική παρουσίαση των δεδομένων εκπαίδευσης και ελέγχου 46

Πινάκας 3.3 : Αποτελέσματα των αρχικών μοντέλων 48

Πινάκας 3.4 : Αποτελέσματα των MAP προσαρμοσμένων μοντέλων 49

Πινάκας 4.1 : Συνοπτική παρουσίαση των δεδομένων εκπαίδευση και έλεγχου 52

Πινάκας 4.2 : Αποτελέσματα για τα σύνολα δοκιμών των ομιλητών μετά από MAP και Speaker MLLR Adaptation 56

Πινάκας 4.3 : Αποτελέσματα για Test Spk_F0 μετά από MAP και Supervised Speaker MLLR Adaptation 56

Πινάκας 4.4 : Αποτελέσματα για Test Spk_F1 μετά από MAP και Supervised Speaker MLLR Adaptation 57

Πινάκας 4.5 : Αποτελέσματα για Test Spk_F2 μετά από MAP και Supervised Speaker MLLR Adaptation 57

Πινάκας 4.6 : Αποτελέσματα μετά από Supervised Speaker Adaptation για διαφορετικούς αριθμούς MLLR μετασχηματισμών 58

Πίνακας 4.7 : Αποτελέσματα για Test Spk_F0 μετά από Supervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών 59

Πίνακας 4.8 : Αποτελέσματα για Test Spk_F1 μετά από Supervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών 60

Πίνακας 4.9 : Αποτελέσματα για Test Spk_F2 μετά από Supervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών 60

Πίνακας 4.10 : Αποτελέσματα μετά από Unsupervised Speaker Adaptation για διαφορετικούς αριθμούς MLLR μετασχηματισμών 61

Πίνακας 4.11 : Αποτελέσματα για Test Spk_F0 μετά από Unsupervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών 63

Πίνακας 4.12 : Αποτελέσματα για Test Spk_F1 μετά από Unsupervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών 63

Πίνακας 4.13 : Αποτελέσματα για Test Spk_F2 μετά από Unsupervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών 64

Πίνακας 4.14 : Αποτελέσματα μετά από Unsupervised Speaker Adaptation για μεταβλητό αριθμό MLLR μετασχηματισμών 65

Πίνακας 4.15 : Αποτελέσματα πριν και μετά από Unsupervised Speaker Adaptation για μεταβλητό αριθμό MLLR μετασχηματισμών 65

Πίνακας 4.16 : Αποτελέσματα για Test Spk_F0 μετά από Unsupervised Speaker MLLR Adaptation μεταβλητών MLLR μετασχηματισμών για κάθε ομιλητή. 68

Πίνακας 4.17 : Αποτελέσματα για Test Spk_F1 μετά από Unsupervised Speaker MLLR Adaptation μεταβλητών MLLR μετασχηματισμών για κάθε ομιλητή. 69

Πίνακας 4.18 : Αποτελέσματα για Test Spk_F2 μετά από Unsupervised Speaker MLLR Adaptation μεταβλητών MLLR μετασχηματισμών για κάθε ομιλητή. 69

Πίνακας 4.19 : Αποτελέσματα για το σύνολο των Test sets μετά από Unsupervised Speaker MLLR Adaptation σε τμήματα ομιλητών. 70

Πίνακας 4.20 : Παρουσίαση της προόδου των αποτελεσμάτων για το σύνολο των Test sets ... 72

Πίνακας ΠΒ-1 : Αρχική μορφή προτάσεων ελέγχου 80

Πίνακας ΠΒ-2 : Προτάσεις αναγνώρισης από το MLS και MLS+MAP μοντέλο 81

Πίνακας ΠΒ-3 : Προτάσεις αναγνώρισης από το MMIS και MMIS+MAP μοντέλο 82

Πίνακας ΠΒ-4 : Προτάσεις αναγνώρισης από το MMIS και το MIS+MAP+(SSA+USA_Known_Spk)+(USA_Unknown_Spk) μοντέλο 83

Λίστα Σχημάτων

Σχήμα 1.1: Σύστημα αναγνώρισης φωνής 15

Σχήμα 1.2 Μοντέλο αποκωδικοποίησης 16

Σχήμα 1.3 Front-end processing for MFCC 17

Σχήμα 1.4 : Εποπτική εικόνα ενός συστήματος αναγνώρισης φωνής με στατιστικές μεθόδους 20

Σχήμα 1.5 Παράδειγμα HMM με 5 καταστάσεις 21

Σχήμα 3.1 Βασική δομή ενός φωνητικού HMM 42

Σχήμα 3.2 : Γραφική απεικόνιση της απόδοσης των Seed Models 48

Σχήμα 3.3 : Γραφική απεικόνιση της απόδοσης των Seed Models πριν και μετά από MAP Adaptation 50

Σχήμα 4.1 : Γραφική απεικόνιση αποτελεσμάτων του Supervised Speaker Adaptation για διαφορετικούς αριθμούς MLLR μετασχηματισμών. 58

Σχήμα 4.2 : Γραφική απεικόνιση αποτελεσμάτων του Unsupervised Speaker Adaptation για διαφορετικούς αριθμούς MLLR μετασχηματισμών. 62

Σχήμα 4.3 : Γραφική απεικόνιση αποτελεσμάτων για Spk_F0 σε όλα τα Unsupervised μοντέλα 66

Σχήμα 4.4 : Γραφική απεικόνιση αποτελεσμάτων για Spk_F1 σε όλα τα Unsupervised μοντέλα 67

Σχήμα 4.5 : Γραφική απεικόνιση αποτελεσμάτων για Spk_F2 σε όλα τα Unsupervised μοντέλα 67

Σχήμα 4.6 : Γραφική απεικόνιση της προόδου αποτελεσμάτων για Test F0 70

Σχήμα 4.7 : Γραφική απεικόνιση της προόδου αποτελεσμάτων για Test F1 71

Σχήμα 4.8 : Γραφική απεικόνιση της προόδου αποτελεσμάτων για Test F2 71

Σχήμα ΠΑ-1: Στάδια λειτουργίας του HTK 76

Σχήμα ΠΑ-2 Λειτουργία του HTK 77

ΕΙΣΑΓΩΓΗ

Γενικά

Η ομιλία αποτελεί τον πιο φυσικό και πλέον διαδεδομένο τρόπο επικοινωνίας μεταξύ των ανθρώπων. Η εξέλιξη διαφόρων τομέων της τεχνολογίας και της επιστήμης επέτρεψε την ανάπτυξη συστημάτων και εργαλείων με τα οποία μπορούν να αυτοματοποιηθούν διαδικασίες που σχετίζονται με την ομιλία.

Υπό το φως των νέων τεχνολογικών εξελίξεων το 1950 εμφανίστηκε για πρώτη φορά η αναγνώριση ομιλίας, η οποία αποτέλεσε και συνεχίζει να αποτελεί μια από τις μεγαλύτερες προκλήσεις στον κλάδο της επιστήμης των υπολογιστών. Στις μέρες μας παρατηρούμε στην καθημερινότητα ολοένα και περισσότερο την χρήση συστημάτων αυτόματης αναγνώρισης ομιλίας και αν σκεφτούμε ότι ακόμα βρίσκεται σε πρώιμο στάδιο μπορούμε εύκολα να καταλάβουμε ότι θα είναι μια από τις επιστήμες που θα κυριαρχήσει στο άμεσο μέλλον.

Αναγνώριση Ομιλίας *(ASR: Automatic Speech Recognition)* είναι η διαδικασία μετατροπής ενός ακουστικού ή φωνητικού σήματος της ανθρώπινης ομιλίας σε μια ακολουθία λέξεων με σκοπό την αναγνώριση της από τους υπολογιστές. 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).* Στην τεχνολογία αυτή το σύστημα κατασκευάζει ένα δίκτυο, που υλοποιεί μια γραμματική, όπου για κάθε επιτρεπόμενη πρόταση αντιστοιχίζεται ένα σύνολο από κρυφά μοντέλα Markov (HMMs). Κατά την διαδικασία της αναγνώρισης νέων δεδομένων το σύστημα υπολογίζει τις πιθανότητες τα δεδομένα αυτά να είχαν παραχθεί με βάση καθένα από τα αποθηκευμένα HMMs και εξάγει ως αποτέλεσμα την πρόταση με την μεγαλύτερη πιθανότητα.

Τα συστήματα αυτόματης αναγνώρισης ομιλίας χρησιμοποιούνται στις μέρες μας από διαφορές εφαρμογές σε πολλούς τομείς της καθημερινότητας μας όπως τηλεφωνικές συνομιλίες, συστήματα πλοήγησης, ρομποτική, κλπ. Ανάλογα με τις δυνατότητες τους, μπορούν να κατηγοριοποιηθούν σε:

- *Συστήματα απομονωμένων λέξεων*
- *Συστήματα συνδεδεμένων λέξεων*
- *Συστήματα συνεχούς ομιλίας*
- *Συστήματα αυθόρμητης ομιλίας*

Τα συστήματα απομονωμένων λέξεων είναι περιοριστικοί αναγνωριστές συνήθως όχι τόσο ευέλικτοι και ευχάριστοι προς τους χρήστες μπορούν όμως να λειτουργήσουν ικανοποιητικά σε μια μεγάλη ποικιλία εφαρμογών. Τα συστήματα συνδεδεμένων λέξεων είναι λιγότερο περιοριστικά και έχουν καλές επιδόσεις για μια σειρά από ενδιαφέρουσες εφαρμογές. 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. Τέλος 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 κλπ

Τα συστήματα αναγνώρισης ομιλίας μπορούν να αναπτυχθούν με δεδομένα εκπαίδευσης που είναι είτε ενός ομιλητή (speaker-dependent) είτε έχουν συλλεχθεί 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.

Το θέμα της διπλωματικής

Όπως αναφέραμε και παραπάνω η συνεχής εξέλιξη της τεχνολογίας είχε άμεση επίπτωση στον τομέα της επεξεργασίας φωνής και ειδικότερα της αναγνώρισης φωνής. Τα νέα συστήματα αυξάνουν συνεχώς την υπολογιστική δυνατότητα των αλγορίθμων επεξεργασίας φωνής και παρέχουν νέες δυνατότητες και πιο γρήγορες υπηρεσίες. Η παρούσα διπλωματική εργασία ασχολείται με τη βελτιστοποίηση 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,SI).

Αρχικά μας δοθήκαν ακουστικά σήματα ηχογραφημένα από τηλεοπτικά δελτία ειδήσεων μαζί με τις απομαγνητοφωνήσεις τους . Στη συνέχεια αφού επεξεργαστήκαμε τα δεδομένα αυτά έτσι ώστε να είναι κατάλληλα για την υλοποίηση των πειραμάτων, μας εξαγάγαμε έναν αριθμό ολιγόλεπτων προτάσεων που αποτέλεσαν τη βάση δεδομένων μας. Έχοντας πλέον στη διάθεση μας τα δεδομένα στη μορφή που επιθυμούμε, το επόμενο μας βήμα ήταν να υλοποιήσουμε μια σειρά πειραμάτων που είχαν σαν κύριο θέμα τους την μέθοδο της προσαρμογής*(Adaptation).* Συγκεκριμένα, εφαρμόσαμε διάφορες τεχνικές προσαρμογής (MAP,MLLR) και σε διαφορετικό επίπεδο (Ομιλητές, Συνθήκη ομιλίας). Ως βασικά μοντέλα(Seed Models) χρησιμοποιήσαμε δυο SI μοντέλα εκπαιδευμένα με τα κριτήρια *Maximum likelihood Estimation (MLE)* και *Maximum Mutual Information (MMI).* Η προσαρμογή στα μοντέλα αυτά έγινε αρχικά στο επίπεδο συνθήκης ομιλίας εφαρμόζοντας MAP Adaptation και δημιουργώντας μοντέλα αρκετά ικανά να αναγνωρίσουν δεδομένα από τηλεοπτικές ειδήσεις. Στη συνέχεια εφαρμόσαμε τη μέθοδο της προσαρμογής στο επίπεδο του ομιλητή υλοποιώντας μοντέλα εξαρτώμενα από ομιλητές (Speaker Depended,SD) τα όποια βελτίωσαν την ακρίβεια της αναγνώρισης ακόμα περισσότερο.

Η συνεισφορά αυτής της διπλωματικής εργασίας είναι ιδιαίτερα σημαντική αφού κατάφερε να βελτιώσει σε πολύ μεγάλο βαθμό την ακρίβεια της αναγνώρισης ακουστικών σημάτων ηχογραφημένα από τηλεοπτικές ειδήσεις.

Δομή της διπλωματικής

Στη συνέχεια θα παρουσιάσουμε συνοπτικά την υλη της παρούσας διπλωματικής εργασίας.

- **Κεφάλαιο 1 «Εισαγωγή στην Αναγνώριση Ομιλίας»:** Στο κεφάλαιο αυτό περιγράφονται η αναγνώριση με στατιστικές μεθόδους, η δομή των *Hidden Markov Models(HMMs)* και η μέθοδος της προσαρμογής*(Adaptation).*
- **Κεφάλαιο 2 «Περιγραφή της Βάσης Δεδομένων»:** Στο κεφάλαιο αυτό δίνεται μια περιγραφή ακουστικών σημάτων και του τρόπου απομαγνητοφώνησης τους καθώς και των κατηγοριών που σχηματιστήκαν.
- **Κεφάλαιο 3 «Τα βασικά μοντέλα αναγνώρισης»** : Στο κεφάλαιο αυτό δίνεται μια περιγραφή του τρόπου εκπαίδευσης των βασικών μας μοντέλων καθώς και τα αποτελέσματα της αναγνώρισης τους πριν και μετά την μέθοδο της προσαρμογής.
- **Κεφάλαιο 4 «Η μέθοδος της προσαρμογής στον ομιλητή»** : Στο κεφάλαιο αυτό δίνεται μια περιγραφή των μοντέλων που έγινε η προσαρμογή στον ομιλητή καθώς και τα αποτελέσματα της αναγνώρισης τους.
- **Κεφάλαιο 5 «Ανακεφαλαίωση και Μελλοντικές Επεκτάσεις»** : Στο κεφάλαιο αυτό γίνεται ανακεφαλαίωση των προδιαγραφών του συστήματος αναγνώρισης που παρουσιάστηκαν στην εργασία, καθώς και αναφορά σε θέματα σχετικά με την εξέλιξη και τις μελλοντικές επεκτάσεις του.
- **Παράρτημα «HTK (Hidden Markov Model Toolkit)» :** Στο παράρτημα αυτό γίνεται παρουσίαση της εφαρμογής του συγκεκριμένου εργαλείου στην μέθοδο της προσαρμογής.
- **Βιβλιογραφία :** Δίνονται οι παραπομπές στα επιστημονικά άρθρα που χρησιμοποιήσαμε κατά την ανάπτυξη της εφαρμογής.

ΚΕΦΑΛΑΙΟ 1

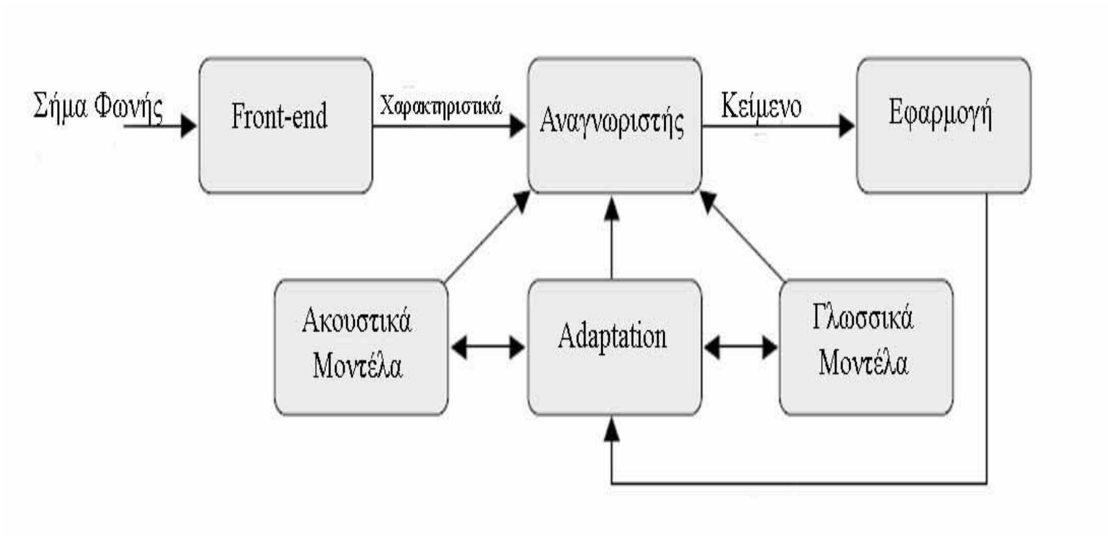
ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΑΝΑΓΝΩΡΙΣΗ ΟΜΙΛΙΑΣ

1.1 Εισαγωγή

Η αυτόματη αναγνώριση ομιλίας αποτελεί μια από τις κλασικές μεθόδους αναγνώρισης προτύπων και σκοπός της είναι η αυτόματη παραγωγή κειμένου από την ανθρώπινη ομιλία. Για να γίνει εφικτό αυτό, εξάγονται από το ακουστικό σήμα μια σειρά από παραμέτρους, οι οποίες σε συνδυασμό με διάφορους αλγόριθμους αναγνώρισης προτύπων, παράγουν το τελικό αποτέλεσμα. Η πλειονότητα των συστημάτων αναγνώρισης φωνής χρησιμοποίει κρυφά μοντέλα Markov (HMMs) ως ακουστικά μοντέλα, παρ' όλα αυτά ο τρόπος κατά τον οποίο εκπαιδεύονται ποικίλλει.

Τα βασικά μέρη ενός συστήματος αναγνώρισης ομιλίας καθώς και ο τρόπος που αυτά αλληλεπιδρούν μεταξύ τους φαίνονται στο Σχήμα 1.1. Το σήμα φωνής συλλέγεται σε ψηφιακή μορφή και στη συνέχεια υπόκειται σε προεπεξεργασία (Front-end) με σκοπό τη εξαγωγή χαρακτηριστικών κατάλληλων για αναγνώριση. Τα χαρακτηριστικά (Features) εισέρχονται στον αναγνωριστή. Έπειτα, το σύστημα κατασκευάζει ένα δίκτυο που υλοποιεί την γραμματική και για κάθε επιτρεπόμενη πρόταση αντιστοιχίζεται ένα σύνολο από 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 έξοδος του αναγνωριστή μπορεί πλέον να χρησιμοποιηθεί από διαφορές ψηφιακές εφαρμογές σε επίπεδο λέξεων, προτάσεων ή ολόκληρου κειμένου.

Ένας πολύ σημαντικός τομέας των συστημάτων αναγνώρισης φωνής είναι η διαδικασία της προσαρμογής (Adaptation), η οποία αυξάνει σε πολύ μεγάλο βαθμό το ποσοστό επιτυχίας των αναγνωριστών. Κατά τη διαδικασία της προσαρμογής το σύστημα «μαθαίνει» να προσαρμόζεται στις εκάστοτε συνθήκες αναγνώρισης (ομιλητής, περιβάλλον, περιεχόμενο) . Υπάρχουν δυο βασικοί τρόποι εφαρμογής της προσαρμογής. Ο πρώτος είναι να γνωρίζουμε εξ’ αρχής το αντίστοιχο κείμενο της ομιλίας (Supervised) και ο δεύτερος να μην έχουμε καμιά πληροφορία για αυτό (Unsupervised).



Σχήμα 1.1: Σύστημα αναγνώρισης φωνής

1.2 Η Front –End επεξεργασία

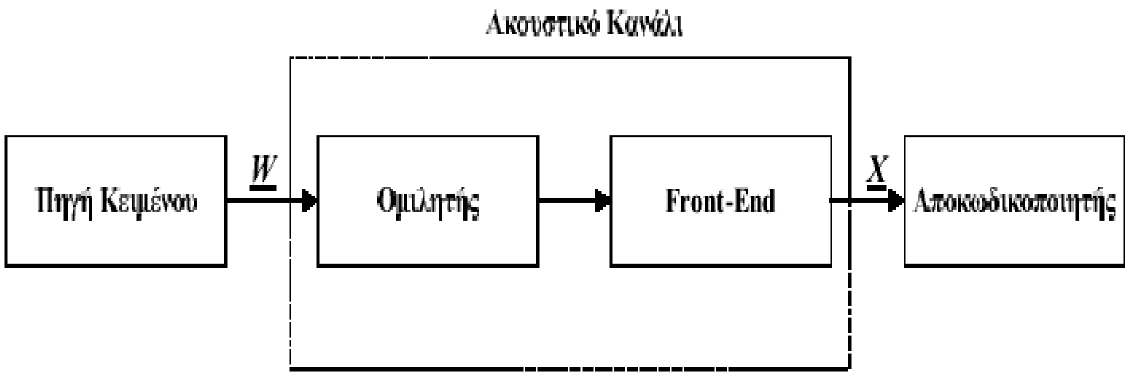
Η Front-end επεξεργασία αποτελεί το πρώτο στάδιο της αναγνώρισης φωνής. Κατά το στάδιο αυτό το σήμα ομιλίας επεξεργάζεται με σκοπό τη εξαγωγή χαρακτηριστικών (Features) κατάλληλων για τα συστήματα αναγνώρισης, τα οποία δεν είναι ευαίσθητα σε εξωτερικούς παράγοντες, όπως τα μικρόφωνα ή ο περιβαλλοντικός θόρυβος. Η διαδικασία του Front-end μπορεί να χωριστεί σε δυο σταδία, το στάδιο της κατάτμησης των ψηφιακών σημάτων (Segmentation) και το στάδιο της εξαγωγής των χαρακτηριστικών (Features extraction). Κατά το πρώτο στάδιο, σκοπός είναι να καταφέρουμε να απομονώσουμε τα ακουστικά σήματα, που περιέχουν την ομιλία και να «κόψουμε» τα σήματα εκείνα που δεν μας είναι χρήσιμα.

Για παράδειγμα, σε ένα σύστημα δελτίων ειδήσεων η διαδικασία της κατάτμησης περιλαμβάνει την αποκοπή των σημάτων έναρξης και τέλους των δελτίων, τις διαφημίσεις, τα σημεία που περιλαμβάνουν μονό μουσική και τις ομιλίες σε γλώσσες που δεν μας ενδιαφέρουν (στην προκειμένη περίπτωση οτιδήποτε μη-Ελληνικό).

Στο στάδιο της εξαγωγής χαρακτηριστικών το σήμα φωνής συλλέγεται και ψηφιοποιείται σε διακριτά δείγματα κατά την διάρκεια του χρόνου 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.



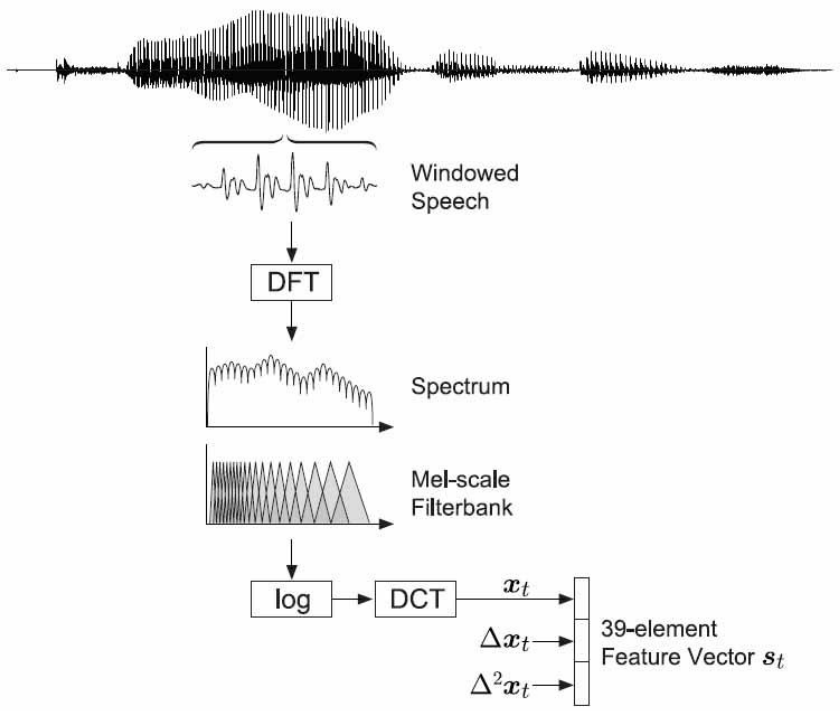
Σχήμα 1.2 Μοντέλο αποκωδικοποίησης

Έστω λοιπόν ότι η πηγή κειμένου παράγει την ακολουθία λέξεων $\underline{W}$ = [w_1 w_2 $w_3 \ldots w_n$]. Το ακουστικό κανάλι, το οποίο αποτελείται από το 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,…,x_T] από παραμετρικά διανύσματα που υπολογίζονται από τον front-end επεξεργαστή του συστήματος αναγνώρισης χρησιμοποιώντας κάποιες από τις μεθόδους που αναφέραμε πιο πριν.

Η πιο 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, η οποία μιμείται και αναπαριστά το ακουστικό εύρος της ανθρώπινης ακοής. Η Mel κλίμακα μπορεί να προσεγγιστεί από τον ακόλουθο τύπο:

$$Mel(f) = 1127 \log_{10}(1 + \frac{f}{700}) \qquad (1.2\text{-}1)$$

Η διαδικασία εξαγωγής διανυσματικών ακολουθιών βασιζόμενα στην Mel-frequency παρουσιάζεται στο σχήμα 1.3.



Σχήμα 1.3 Front-end processing for MFCC.

Αρχικά το σήμα φωνής διαιρείται σε τμήματα (Frames). Η απόσταση ανάμεσα στα Frames είναι τυπικά 10ms και τα τμήματα αυτά επικαλύπτονται για να δώσουν μεγαλύτερης διάρκειας παράθυρο με τυπική τιμή περί τα 25ms. Κάνοντας χρήση των επικαλυπτόμενων παραθύρων (συνήθως Hamming) εξομαλύνεται η μετάβαση μεταξύ των τμημάτων. Ένας διακριτός μετασχηματισμός Fourier (DFT) εφαρμόζεται για να μας δώσει το φάσμα του σήματος (Spectrum). Έπειτα, το παραγόμενο φάσμα του ομαλοποιείται, περνώντας τους φασματικούς συντελεστές από τριγωνοειδείς συχνότητες που καθορίζονται από την Mel –frequency. Στη 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 συνημίτονου (Discrete Cosine Transform - DCT). Συνηθίζεται στις τελικές διανυσματικές ακολουθίες που παράγονται να προστίθενται συντελεστές παραγώγου πρώτης και δεύτερης τάξης και κάποιες φορές να συμπεριλαμβάνεται και μια μέτρηση της ενέργειας του σήματος φωνής.

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, στην προκειμένη περίπτωση πρόκειται για το μέγεθος $P(\underline{W}|\underline{X})$. Επίσης, όπως και σε ένα τυπικό ψηφιακό τηλεπικοινωνιακό σύστημα, θωρείτε ως κριτήριο αποκωδικοποίησης η ελαχιστοποίηση της πιθανότητας σφάλματος. Με βάση το μοντέλο $P(\underline{X}|\underline{W})$, 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 $\underline{X}=[x_1,x_2,\ldots,x_T]$.

Κάνοντας χρήση του κανόνα του Bayes έχουμε:

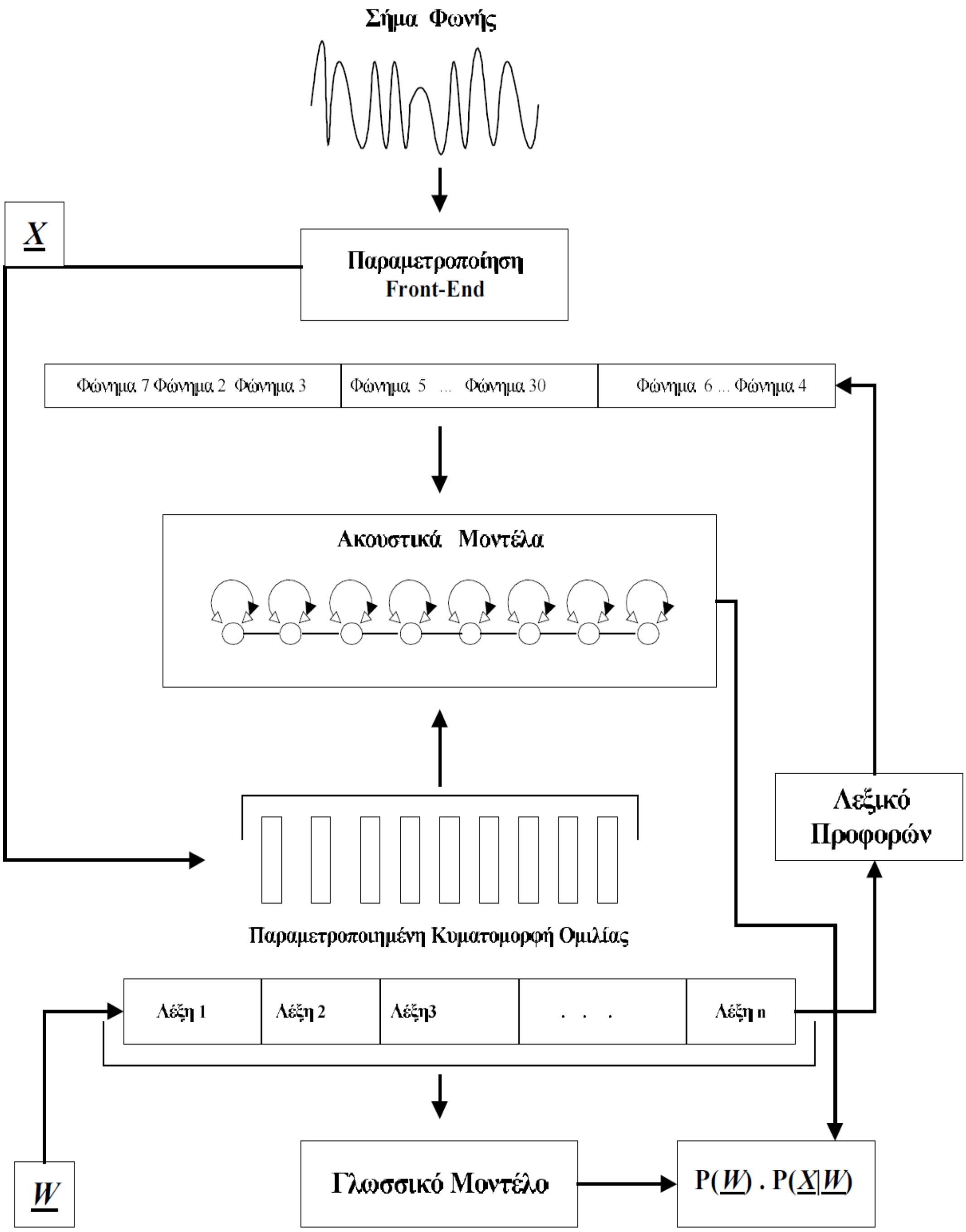
$$\begin{aligned} \hat{W} &= \arg\max_{W} P\left(W \mid X\right) = \arg\max_{W} \frac{P(W)P(X \mid W)}{P(X)} \approx \\ &\approx \arg\max_{W} P(W)P(X \mid W) \end{aligned} \tag{1.3-1}$$

ως argmax συμβολίζεται το όρισμα που μεγιστοποιεί την αντίστοιχη ποσότητα. Δίνοντας μια περιγραφή της εξίσωσης 1.3-1 μπορούμε να πούμε 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 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(X|W) αναπαριστά την πιθανότητα εμφάνισης μιας ακολουθίας διανυσμάτων X δεδομένων μερικών ακολουθιών λέξεων W , και αυτή η πιθανότητα είναι γνωστή ως ακουστικό μοντέλο (*acoustic model*).

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 πιθανοτικά μοντέλα βασικών φωνητικών μονάδων χρησιμοποιείται για να αναπαραστήσουν λέξεις. Οι ακολουθίες των HMMs που χρειάζεται για να αναπαραστήσουν την αρχική έκφραση συνδέονται ώστε να σχηματίσουν ένα απλό μοντέλο και υπολογίζεται η πιθανότητα να παράγει αυτό το μοντέλο την παρατηρούμενη ακολουθία *X.* Αυτή είναι η απαιτούμενη πιθανότητα P(*X*|*W*). Μια ακολουθία από ακουστικές παραμέτρους, προερχόμενες από το σήμα φωνής, αντιμετωπίζεται ως συνδυασμός στοιχειωδών διαδικασιών που περιγράφονται από HMMs.

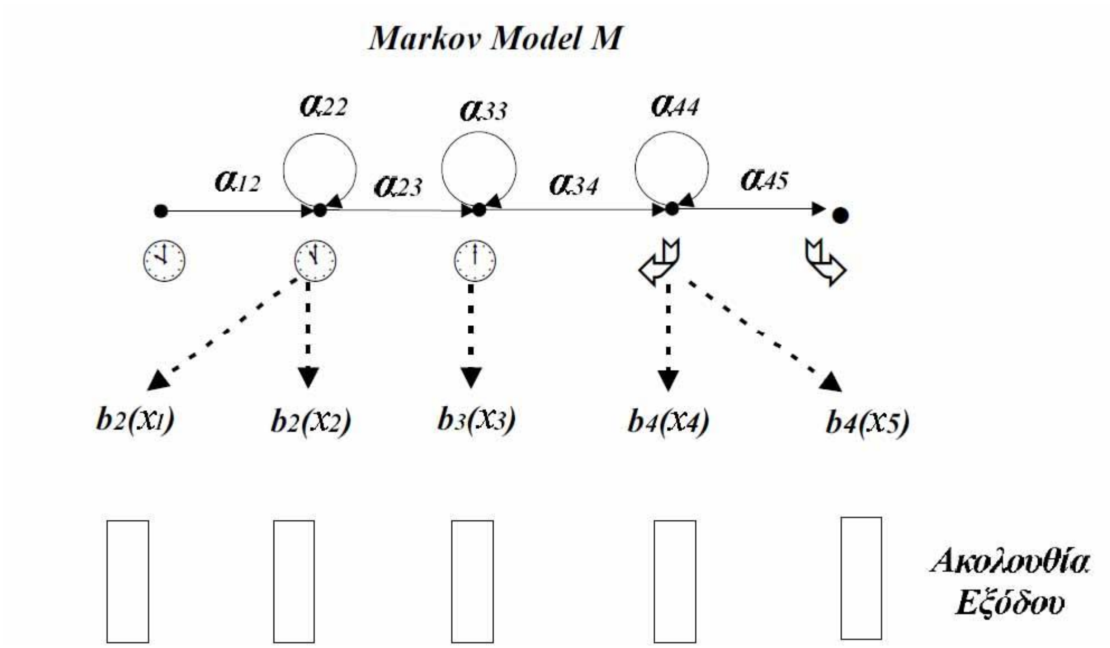
Η πιθανότητα P(X|W) υπολογίζεται χρησιμοποιώντας ένα σύνθετο HMM που αναπαριστά την ακολουθία W και αποτελείται από απλά HMM φωνητικά μοντέλα, συνδεδεμένα σειριακά μεταξύ τους σύμφωνα με τις προφορές στο λεξικό προφορών και υπολογίζεται η πιθανότητα να παράγει αυτό το μοντέλο την παρατηρούμενη ακολουθία X. Η αρχική πιθανότητα P(W) καθορίζεται από το γλωσσικό μοντέλο. Τα παραπάνω φαίνονται αναλυτικά στο Σχήμα 1-4 όπου περιγράφεται η διαδικασία υπολογισμού της πιθανότητας P(X|W) μιας ακολουθίας λέξεων *W* δεδομένου ενός παραμετροποιημένου ακουστικού σήματος *X*.



Σχήμα 1.4 : Εποπτική εικόνα ενός συστήματος αναγνώρισης φωνής με στατιστικές μεθόδους

1.4 Περιγραφή των Hidden Markov Models

Ένα HMM είναι ένα σύνολο από καταστάσεις (states) συνδεδεμένες από μεταβάσεις (βλ. σχήμα 1-5). 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 γνώσιν μας τα προηγούμενα.



$$\mathbf{X} = [\ \mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3 \quad \mathbf{x}_4 \quad \mathbf{x}_5\ ]$$

Σχήμα 1.5 Παράδειγμα HMM με 5 καταστάσεις

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.

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:

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

$$\boxed{\boldsymbol{\lambda = (A, B, \pi)}}$$

1.4.1 Τα είδη των HMMs

Τα HMMs διαφοροποιούνται ανάλογα με το αν διαδικασία που μοντελοποιούμε αποτελείται από συνεχή τυχαία διανύσματα όπως π.χ. οι συντελεστές Cepstral ή εάν η διαδικασία έχει περάσει από κβαντιστή και αποτελείται από διακριτές τυχαίες μεταβλητές. Ανάλογα με τον τύπο της κατανομής εξόδου, τα HMMs ταξινομούνται σε διακριτά και συνεχή.

<u>Διακριτά HMMs (Discrete HMMs)</u>

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 :

$$\sum_{k=1}^{Q} b_j(k) = 1 \qquad (1.4\text{-}1)$$

<u>Συνεχή HMMs (Continuous HMMs)</u>

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

$$b_j(\mathbf{x}_t), \text{ όπου } \mathbf{x}_t = \begin{bmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{dt} \end{bmatrix} \qquad (1.4\text{-}2)$$

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 στη μορφή της συνάρτησης πυκνότητας πιθανότητας έτσι ώστε να εξασφαλίζεται ότι οι παράμετροι της μπορούν να επαναϋπολογιστούν με ακριβή τρόπο. Η συνάρτηση πυκνότητας πιθανότητας, που χρησιμοποιείται για την μέθοδο των συνεχών μειγμάτων γκαουσιανών, έχει την μορφή γραμμικών συνδυασμών (μειγμάτων) από γκαουσιανές :

$$b_j(\mathbf{x}_t) = \sum_{k=1}^{M} c_{jk}\, \boldsymbol{N}(\mathbf{x}_t, \mu_{jk}, \Sigma_{jk}), \qquad 1 \le j \le N \qquad (1.4\text{-}3)$$

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

$$\sum_{k=1}^{M} c_{jk} = 1, \quad 1 \le j \le N \qquad (1.4\text{-}4)$$

$$c_{jk} \geq 0 \,,\; 1 \leq j \leq N,\; 1 \leq k \leq M$$

έτσι ώστε:

$$\int_{-\infty}^{\infty} b_j\,(\mathbf{x}_t) d\mathbf{x}_t = 1 \qquad (1.4\text{-}5)$$

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.

1.5 Η Μέθοδος της Προσαρμογής (Adaptation)

1.5.1 Γενικά

Υπάρχουν διάφορες στρατηγικές που χρησιμοποιούνται για την βελτίωση της ακρίβειας στα συστήματα αναγνώρισης φωνής (*ASR*) και των εφαρμογών που σχετίζονται με αυτά. Η βασικότερη μέθοδος, που ακολουθείται από τα συστήματα αναγνώρισης είναι «εκμάθηση» του συστήματος στις εκάστοτε συνθήκες μέσω μιας διαδικασίας που ονομάζεται Προσαρμογή (Adaptation).

Η μέθοδος της προσαρμογής είναι μια σημαντική διαδικασία για τα ASR και εφαρμόζεται ευρέως στα περισσότερα συστήματα, έτσι ώστε να καταφέρουν να αναγνωρίσουν με όσο το δυνατόν μεγαλύτερη ακρίβεια τα σήματα ανθρώπινης ομιλίας. Η προσαρμογή δίνει την δυνατότητα στους αναγνωριστές να προσαρμόζονται στις διάφορες συνθήκες του σήματος εισόδου όπως:

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

Η τεχνική της προσαρμογής μπορεί να εφαρμοστεί είτε με την μέθοδο της επίβλεψης (Supervised) είτε χωρίς επίβλεψη (Unsupervised). Κατά την μέθοδο της προσαρμογής με επίβλεψη, οι λέξεις ή οι προτάσεις εκπαίδευσης είναι γνωστές στο σύστημα σε αντίθεση με την μέθοδο της προσαρμογής χωρίς επίβλεψη, όπου οι προτάσεις μπορεί να είναι αυθαίρετες. Η μέθοδο της προσαρμογής μπορεί να διαχωριστεί περαιτέρω σε *on-line* και *off-line* τεχνικές. Στην *on-line* τεχνική η προσαρμογή εφαρμόζεται σε πραγματικό χρόνο καθώς το σύστημα είναι σε λειτουργία και προσαρμόζεται αυξητικά, όσο οι προτάσεις εισέρχονται προς αναγνώριση. Η *off-line* προσαρμογή απαιτεί ένας νέος ομιλητής να θέσει ως είσοδο στο σύστημα έναν προκαθορισμένο αριθμό από προτάσεις εκπαίδευσης και στη συνέχεια να τρέξει την αναγνώριση. Κάθε μια από τις παραπάνω μεθόδους μπορεί να εφαρμοστεί με επιτυχία σε κατάλληλα συστήματα, παρόλα αυτά η *on-line* μέθοδος θεωρείται η πιο χρήσιμη διότι δίνει τη δυνατότητα στο σύστημα να προσαρμόζεται σε πραγματικό χρόνο (συνήθως *Unsupervised*) καθώς «ακούει» τις νέες προτάσεις προς αναγνώριση .

Δυο από τις βασικές τεχνικές εφαρμογής του Adaptation είναι η *Maximum a-priori(MAP)* και η *Maximum Likelihood Linear Regression (MLLR).* Στη συνέχεια θα δώσουμε μια θεωρητική περιγραφή των δυο παραπάνω τεχνικών.

1.5.2 MAP Adaptation (Maximum a Posteriori)

Στο MAP Adaptation γίνεται χρήση της παρούσας πληροφορίας από τα υπάρχοντα μοντέλα έτσι ώστε τα περιορισμένα δεδομένα προσαρμογής 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 posteriori πυκνότητας πιθανότητας του λ:

$$\lambda_{MAP} = \arg\max_{\lambda} P(\lambda \mid X) \tag{1.5-1}$$

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

Η σχέση 1.5-1 λόγω του θεωρήματος Bayes γίνεται ως εξής:

$$\lambda_{MAP} = \arg\max_{\lambda} \frac{L(X \mid \lambda) P_0(\lambda)}{P(x)} \tag{1.5-2}$$

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 λόγους ευκολίας, 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:

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\bar{\mu}_{jm} + \frac{\tau}{N_{jm}+\tau}\mu_{jm} \tag{1.5-3}$$

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:

$$N_{jm} = \sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t) \tag{1.5-4}$$

όπου R ο αριθμός των προτάσεων προσαρμογής, T ο αριθμός των frames της r-οστής πρότασης και $L_{jm}^{r}(t)$ είναι η a-posteriori πιθανότητα να βρισκόμαστε στο m-οστό μείγμα της j κατάστασης τη χρονική στιγμή t της r-οστής πρότασης.

Συνεχίζοντας την επεξήγηση των όρων της σχέσης 1.5-3, 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:

$$\bar{\mu}_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t) O_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)} \tag{1.5-5}$$

Τέλος, θα πρέπει να σημειώσουμε πως συνήθως το MAP Adaptation εφαρμόζεται όταν έχουμε πολλά δεδομένα προσαρμογής. Οι μόνοι περιορισμοί και τα μόνα προβλήματα που έχει είναι ότι πρέπει να υπολογιστεί ο όρος $P_0(\lambda)$ της σχέσης 1.5-2 από τα αρχικά εκπαιδευμένα μοντέλα, και ότι αλλάζει μόνο τις παραμέτρους του μοντέλου για τις οποίες έχουν πληροφορία τα δεδομένα προσαρμογής.

1.5.3 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:

$$\hat{\mu}_S = W_S \xi_S \qquad (1.5\text{-}6)$$

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:

$$\xi_S = [\mathrm{w}, \mu_{s1}, \ldots, \mu_{sn}]^t \qquad (1.5\text{-}7)$$

Ο όρος w_S υπολογίζεται από τον τύπο:

$$w_S = \sum_{t=1}^{T} \sum_{r=1}^{R} L_{sr}(t) \sum_{sr}^{-1} O(t) \xi_{sr}^{T} = w_S = \sum_{t=1}^{T} \sum_{r=1}^{R} L_{sr}(t) \sum_{sr}^{-1} W_s \xi_{sr} \xi_{sr}^{T} \qquad (1.5\text{-}8)$$

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

Αντίθετα με τη 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.

ΚΕΦΑΛΑΙΟ 2

ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ

2.1 Εισαγωγή

Ένα αυτόματο σύστημα αναγνώρισης φωνής αποτελείται από δυο βασικά μοντέλα, το γλωσσικό και το ακουστικό μοντέλο. Για την εκπαίδευση και συνεπώς για την λειτουργία των δυο αυτών μοντέλων πρέπει να υπάρχει ένας επαρκής αριθμός δεδομένων εκπαίδευσης. Συγκεκριμένα, για την εκπαίδευση των ακουστικών μοντέλων απαιτείται ένας επαρκής αριθμός ακουστικών σημάτων, που θα χρησιμοποιηθούν για την υλοποίηση του. Σε αυτό το κεφάλαιο θα κάνουμε μια πλήρη περιγραφή της βάσης δεδομένων που οδήγησε στη δημιουργία των ακουστικών μοντέλων. Η βάση δεδομένων μας απαρτίζεται αποκλειστικά και μόνο από ακουστικά σήματα ηχογραφημένα από τηλεοπτικές εκπομπές δελτίων ειδήσεων.

2.2 Τα ακουστικά σήματα

Όπως έχουμε ήδη αναφέρει, η εκπαίδευση των ακουστικών μοντέλων απαιτεί έναν ικανοποιητικό αριθμό από ακουστικά σήματα για να υλοποιηθεί. Συνεπώς 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 ελέγχου.

Τα ακουστικά σήματα που χρησιμοποιηθήκαν για να υλοποιηθεί η βάση δεδομένων αποτελούνται από περίπου 50 ώρες ηχογραφημένων εκπομπών ελληνικών δελτίων ειδήσεων από το το Μάιο του 2006 έως το Δεκέμβριο του 2007 και προέρχονται από τους τηλεοπτικούς σταθμούς ET1, NET, SKAI και MEGA. Τα τεχνικά χαρακτηριστικά των αρχείων ήχου (wav files) είναι τα εξής:

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

Για τη διευκόλυνση των περαιτέρω επεξεργασιών που θα υποστούν τα wav αρχεία έχει οριστεί ένας συγκεκριμένος τρόπος ονομασίας των δελτίων, ο οποίος περιέχει στην ονομασία του κάθε αρχείου την ημερομηνία που παρουσιάστηκε το δελτίο (*χρόνος, μήνας, ημέρα*), το κανάλι από το οποίο έγινε η εγγραφή της εκπομπής, το είδος της εκπομπής (*ειδήσεις, talk show κλπ.)* και τέλος την ώρα παρουσίασης της εκπομπής. Παρακάτω παρουσιάζεται ο τρόπος ονομασίας των δελτίων:

{Έτος}	{Μήνας}	{Ημέρα}	{Σταθμός}	{Είδος εκπομπής}	{Ώρα}.wav
07	***06***	***17***	***NET***	***ED***	***2100.wav***

2.3 Η Διαδικασία της Απομαγνητοφώνησης

Κατά τη διαδικασία της απομαγνητοφώνησης τα ακουστικά σήματα που έχουν συλλεχθεί επεξεργάζονται με τη βοήθεια του εργαλείου απομαγνητοφώνησης για το χαρακτηρισμό τους ως προς το περιεχόμενο της ομιλίας και τον τεμαχισμό τους σε μικρότερες προτάσεις. Το Transcriber βγάζει σαν έξοδο αρχεία μορφής xml, τα οποία μπορούν εύκολα πλέον να υποστούν επεξεργασία με διάφορα προγράμματα ή γλώσσες επεξεργασίας κειμένου (πχ. Perl).

Η απομαγνητοφώνηση των ακουστικών σημάτων φωνής έχει γίνει σε τρία επίπεδα:

- Στο επίπεδο συνθήκης ομιλίας.
- Στο επίπεδο του ομιλητή.
- Στο επίπεδο απομαγνητοφώνησης.

Παρακάτω θα δώσουμε μια συνοπτική περιγραφή των τριών παραπάνω επιπέδων.

2.3.1 Επίπεδο συνθήκης ομιλίας

Το Transcriber έχει τη δυνατότητα του τεμαχισμού 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 τη διάρκεια της ομιλίας. Τα μικρότερα αρχεία ήχου που παράγονται θα χαρακτηρίζονται με μια από τις παρακάτω συνθήκες:

Συνθήκη	Περιγραφή
report	Καθαρή ομιλία από 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 από τα ελληνικά
non trans	Δεν υπάρχει ομιλία

Πίνακας 2.1 : Συνθήκες ομιλίας

2.3.2 Επίπεδο ομιλητή

Τα ακουστικά σήματα των δελτίων ειδήσεων ποικίλλουν λόγω των διαφορετικών συνθηκών ομιλίας που παρουσιάζονται, τους διαφορετικούς ομιλητές και την ποιότητα της ομιλίας. Όπως είναι αναμενόμενο στα δελτία ειδήσεων υπάρχει μεγάλη ποικιλία στους ομιλητές, όπως κεντρικοί παρουσιαστές, σχολιαστές, δημοσιογράφοι, πολιτικοί, απλοί πολίτες, κλπ. Κατά συνέπεια για τον σχεδιασμό μιας αξιόπιστης βάσης δεδομένων δεν θα μπορούσε να μην ληφθεί υπ’ όψιν η αναλυτική περιγραφή του κάθε ομιλητή. Στον *Πίνακα 3.2* παρουσιάζονται αναλυτικά τα χαρακτηριστικά του κάθε ομιλητή.

Συνθήκη	Πιθανές επιλογές
Όνομα	*Name*
Φύλο	*Male/Female*
Εθνικότητα	*Native/Non-Native*
Τύπος Ομιλίας	*Spontaneous/Planned*
Ποιότητα Ομιλίας	*Studio/Telephone*

Πίνακας 2.2 : Χαρακτηριστικά ομιλητών

Τέλος, θα πρέπει να αναφέρουμε την περίπτωση που το όνομα του ομιλητή δεν είναι γνωστό. Σε αυτή την περίπτωση υπάρχουν δυο διαφορετικοί τρόποι που χαρακτηρίζεται ο ομιλητής. Ο πρώτος τρόπος αφορά μόνο τις 20 ώρες εκπομπών, όπου η απομαγνητοφώνηση έγινε από φοιτητές του Πολυτεχνείου Κρήτης και ορίζονται τα ονόματα των άγνωστων ομιλητών ως Speaker#N με N τον αριθμό που εμφανίστηκε ο κάθε ομιλητής στην εκάστοτε εκπομπή. Για παράδειγμα, ο πρώτος άγνωστος ομιλητής που συναντάμε τυχαία σε ένα δελτίο χαρακτηρίζεται ως Speaker#1 ο δεύτερος Speaker#2 κ.ο.κ.

Ο δεύτερος τρόπος αφορά τις υπόλοιπες 30 ώρες όπου η απομαγνητοφώνηση έγινε από το Εθνικό Μετσόβιο Πολυτεχνείο και ως όνομα των άγνωστων ομιλητών ορίζεται η ιδιότητα τους. Για παράδειγμα ο πρώτος ρεπόρτερ που συναντάμε ορίζεται ως Reporter1 ο δεύτερος ως Reporter2 κ.ο.κ. ενώ οι απλοί πολίτες ως ΠεραστικόςN με N τον αριθμό εμφάνισης τους.

2.3.3 Επίπεδο απομαγνητοφώνησης

Το τρίτο και κυριότερο στάδιο κατά τη διαδικασία της απομαγνητοφώνησης είναι η γραπτή αποτύπωση των σημάτων ομιλίας στα ελληνικά. Για την υλοποίηση της παραπάνω διαδικασίας έχουν τεθεί κάποιοι βασικοί κανόνες τους οποίους θα περιγράψουμε περιληπτικά παρακάτω.

1. Όλες οι προτάσεις περιγράφονται μόνο με ελληνικές λέξεις και χρησιμοποιούνται μόνο πεζά γράμματα(ισχύει το ίδιο και για τα ακρωνύμια και τις συντομεύσεις πχ. ΔΕΗ→δεη).
2. Οι αριθμοί γράφονται με λέξεις (πχ 990→εννιακόσια ενενήντα)
3. Οι δισταγμοί(Hesitations) κατά τη διάρκεια της ομιλίας περιγράφονται με το χαρακτήρα @ε@.
4. Οι λανθασμένες προφορές λέξεων περικλείονται από τους χαρακτήρες * και τη σωστή λέξη στο ενδιάμεσο (**σωστή_προφορά**).
5. Οι ατελείς λέξεις χαρακτηρίζονται ως [FRAGMENT].
6. Οι στιγμιαίοι θόρυβοι ως [NOISE].
7. Οι αναπνοές που παίρνουν οι ομιλητές κατά τη διάρκεια της ομιλίας τους χαρακτηρίζονται με [BREATH].
8. Σε οποιαδήποτε άλλη περίπτωση, όπου υπάρχει κάποιο πρόβλημα στην ομιλία, έκτος των παραπάνω, την χαρακτηρίζουμε ως [TAG_BAD_READING].

Στον *Πίνακα 3.3* παρουσιάζονται συνοπτικά οι παραπάνω περιπτώσεις.

Συμβάν	Σύμβολο περιγραφής
Δισταγμοί	*@ε@*
Λανθασμένες προφορές	**σωστή_προφορά**
Ατελείς λέξεις	*[FRAGMENT]*
Στιγμιαίοι θόρυβοι	*[NOISE]*
Αναπνοές	*[BREATH]*
Αδυναμία χαρακτηρισμού	*[TAG_BAD_READING]*

Πίνακας 2.3 : Συμβάντα κατά την διάρκεια της ομιλίας

Τελειώνοντας την περιγραφή της βάσης δεδομένων αξίζει να σημειωθεί πως η συλλογή και η απομαγνητοφώνηση των δελτίων ειδήσεων είχε γίνει στα πλαίσια προηγούμενων διπλωματικών εργασιών σε συνεργασία του Πολυτεχνείου Κρήτης με το Εθνικό Μετσόβιο Πολυτεχνείο(ΕΜΠ). Λόγω της ανάγκης διαφορετικής χρήσης των δεδομένων από τα δυο αυτά ιδρύματα, όπως είναι αναμενόμενο υπήρχαν κάποιες διαφορές στον τρόπο απομαγνητοφώνησης τους. Η βασικότερη διάφορα είναι πως οι 30 ώρες που επεξεργαστήκαν από το ΕΜΠ δεν περιείχαν την πληροφορία για τον τρόπο ομιλίας(Planned/Spontaneous), την ποιότητα(Studio/Telephone) και τον περιβαλλοντικό θόρυβο ή τη μουσική(Background noise/music). Μια ακόμα σημαντική διαφορά είναι πως στα δεδομένα του ΕΜΠ ο διαχωρισμός των προτάσεων είχε γίνει με βάση την αναπνοή που παίρνει ο ομιλητής με αποτέλεσμα να υπάρχουν πολλές μικρές προτάσεις οι οποίες έχαναν το νόημα τους.

Στα πλαίσια της διπλωματικής μας εργασίας επεξεργαστήκαμε και διορθώσαμε τις 10 περίπου από τις 30 ώρες του ΕΜΠ της οποίες μετέπειτα χρησιμοποιήσαμε ως δεδομένα ελέγχου (Test Data). Για τις υπόλοιπες 20 ώρες οι οποίες ενσωματώθηκαν με τις 20 ώρες του Πολυτεχνείου Κρήτης για τη δημιουργία των δεδομένων εκπαίδευσης, χρησιμοποιήσαμε την κατηγοριοποίηση(Classification) που είχε κάνει ο μεταπτυχιακός φοιτητής Τσεργουλας Ορφέας στα πλαίσια του Master του (*A Greek Broadcast News Transcription System[7]*).

2.4 Δημιουργία της τελικής βάσης δεδομένων

Έπειτα από το τέλος τις διαδικασίας απομαγνητοφώνησης έχουμε φτάσει στο στάδιο ολοκλήρωσης της τελικής βάσης δεδομένων. Στο στάδιο αυτό θα επεξεργαστούμε τα Transcriptions έτσι ώστε να είναι εφικτός ο διαχωρισμός τους σε όποια κατηγορία επιθυμούμε και στη συνέχεια θα δημιουργήσουμε για κάθε κατηγορία τα δεδομένα εκπαίδευσης και τα δεδομένα έλεγχου.

2.4.1 Επεξεργασία δεδομένων

Για τη δημιουργία της τελικής βάσης δεδομένων μας έπρεπε να επεξεργαστούμε κατάλληλα τα αρχεία που παράγει το Transcriber Tool ώστε να εξάγουμε από αυτά τις πληροφορίες που θα μας είναι χρήσιμες για την υλοποίηση των πειραμάτων. Αυτές οι πληροφορίες ήταν :

- Η διεύθυνση που είναι αποθηκευμένα τα τεμαχισμένα αρχεία ήχου
- Ο εκάστοτε ομιλητής
- Η συνθήκη ομιλίας (Καθαρή ομιλία, Ομιλία με περιβαλλοντικό θόρυβο ή μουσική και Τηλεφωνική ομιλία)
- Και φυσικά η γραπτή αποτύπωση των σημάτων ομιλίας.

Για την ανάκτηση των παραπάνω πληροφοριών χρησιμοποιήθηκε ένα αρχείο προγραμματισμένο σε γλώσσα Perl όπου διαβάζει τα xml-μορφής αρχεία και «αποσπά» τις συγκεκριμένες πληροφορίες παράγοντας ένα νέο αρχείο της παρακάτω μορφής :

Waveform Path	Speaker	Topic	Mode	Chanel	Transcription
***/full/path.wav**	Name	NoTopic/Noise/music	Planned/Spontaneous	Studio/Telephone	Transcription
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Πίνακας 2.4 : Μορφή εξόδου επεξεργασμένων Transcriptions

Στο σημείο αυτό έχουμε τα δεδομένα μας στην μορφή που επιθυμούμε ώστε να μπορέσουμε να κάνουμε τον τελικό μας διαχωρισμό. Αρχικά, κάναμε το διαχωρισμό με βάση τη συνθήκη ομιλίας, δηλαδή εάν πρόκειται για καθαρή ομιλία, για ομιλία με περιβαλλοντικό θόρυβο ή μουσική και εάν η ομιλία προέρχεται από Τηλεφωνική συσκευή. Στο *Πίνακα 3.5* ορίζονται οι παραπάνω τρεις κατηγορίες.

Κατηγορία	Περιγραφή
F0	Καθαρή ομιλία από στούντιο, χωρίς θόρυβο
F1	Ομιλία με θόρυβο ή μουσική στο background
F2	Τηλεφωνική ομιλία

Πίνακας 2.5: Κατηγορίες δεδομένων

Παρακάτω παρουσιάζεται ο συνολικός αριθμός των προτάσεων για κάθε μια από τις τρεις κατηγορίες:

Κατηγορία	Αριθμός προτάσεων
F0	**16731**
F1	**38077**
F2	**2506**

Πίνακας 2.6: Ποσοτική περιγραφή των δεδομένων κάθε κατηγορίας

Στη συνέχεια προχωρήσαμε στο δεύτερο στάδιο διαχωρισμού των προτάσεων ο όποιος είναι σε επίπεδο ομιλητή. Συγκεκριμένα για κάθε μια από τις κατηγορίες F0, F1 και F2 ομαδοποιήσαμε τις προτάσεις μας με βάση τον εκάστοτε ομιλητή.

Έχοντας πλέων διαχωρίσει τις προτάσεις μας στο επίπεδο συνθήκης της ομιλίας(F0,F1 και F2) και στο επίπεδο του ομιλητή, το επόμενο μας βήμα είναι να χωρίσουμε τα δεδομένα μας σε προτάσεις εκπαίδευσης(Train Data) και σε προτάσεις έλεγχου(Test Data). Για να ανταποκρίνεται το σύστημα μας σε ένα ρεαλιστικό σενάριο θεωρήσαμε σωστό να χωρίσουμε τα δεδομένα μας με βάση το χρόνο. Συγκεκριμένα, θα θεωρήσουμε ως Train Data τις πρώτες 55 χρονικά εκπομπές και τις τελευταίες 10 ως Test Data. Παρακάτω θα μιλήσουμε αναλυτικά για τις συγκεκριμένες κατηγορίες.

2.4.2 Τα δεδομένα εκπαίδευσης(Train Data)

Όπως έχουμε ήδη αναφέρει για την υλοποίηση των ακουστικών μας μοντέλων απαιτείται ένας αρκετά μεγάλος αριθμός δεδομένων εκπαίδευσης. Για τη δημιουργία της βάσης δεδομένων εκπαίδευσης χρησιμοποιηθήκαν οι 55 πρώτες χρονικά εκπομπές. Σε αυτές τις εκπομπές περιέχονται περίπου 20 ώρες όπου η απομαγνητοφώνηση έγινε από φοιτητές του Πολυτεχνείου Κρήτης και άλλες 20 ώρες όπου η απομαγνητοφώνηση έχει γίνει από το ΕΜΠ.

Για κάθε μια από τις τρεις κατηγορίες (F0,F1 και F2) εξαγάγαμε τις προτάσεις εκείνες που περιλαμβάνονταν στις 55 πρώτες εκπομπές και το σύνολο αυτών το θεωρήσαμε ως προτάσεις εκπαίδευσης. Στον *Πίνακα 3.7* φαίνεται αναλυτικά ο αριθμός των προτάσεων εκπαίδευσης για κάθε κατηγορία.

Κατηγορία	**F0**	**F1**	**F2**	**Σύνολο**
Αριθμός προτάσεων εκπαίδευσης	14903	35057	2366	52326

Πίνακας 2.7 : Ποσοτική περιγραφή των δεδομένων εκπαίδευσης σε επίπεδο συνθήκης ομιλίας

Ύστερα, έχοντας πλέον τις προτάσεις εκπαίδευσης στη διάθεση μας προχωρήσαμε στο δεύτερο επίπεδο διαχωρισμού, στο επίπεδο του ομιλητή για να εξάγουμε και εδώ τις προτάσεις εκπαίδευσης. Στα δελτία ειδήσεων υπάρχει μεγάλος αριθμός γνωστών και άγνωστων ομιλητών οι οποίοι μπορεί να διαφέρουν από εκπομπή σε εκπομπή. Κατά συνέπεια, η δημιουργία δεδομένων εκπαίδευσης για όλους τους ομιλητές θα ήταν αδύνατη και ίσως όχι και τόσο χρήσιμη. Για το λόγο αυτό επιλέξαμε να εξάγουμε προτάσεις μόνο για τους ομιλητές εκείνους που είχαν έναν ικανοποιητικό αριθμό προτάσεων καθ' όλη τη διάρκεια των εκπομπών εκπαίδευση και επίσης απαραίτητη προϋπόθεση ήταν οι ομιλητές αυτοί να περιέχονταν και στα δεδομένα ελέγχου. Ύστερα από τους παραπάνω περιορισμούς καταλήξαμε στους ομιλητές που παρουσιάζονται στον *Πίνακα 3.8.*

TRAIN SETS				
Ομιλητές	**Αριθμός προτάσεων**			
	Spk_F0	**Spk_F1**	**Spk_F2**	***Σύνολο ανά ομιλητή***
Γ. Παπανδρέου	41	221	0	***282***
Γαβρά	1466	659	0	***2125***
Κακαβάς	25	75	20	***100***
Κανδύλη	70	150	0	***220***
Καψής	36	356	0	***392***
Κ. Καραμανλής	40	303	0	***343***
Λιναρδάτου	36	197	0	***233***
Λοβέρδος	58	68	0	***126***
Νιότης	65	137	0	***202***
Παπαηλιόπουλος	7	201	0	***208***
Πρετεντέρης	94	761	0	***855***
Σπυράκη	237	244	0	***481***
Τρέμη	1680	886	0	***2566***
Χαρίτος	0	188	29	***217***
Χούκλη	1668	267	0	***1935***
Σύνολο ανά κατηγορία	**5523**	**4713**	**49**	

Πίνακας 2.8 : Ποσοτική περιγραφή των δεδομένων εκπαίδευσης σε επίπεδο ομιλητή

2.4.3 Τα δεδομένα έλεγχου(Test Data)

Για την υλοποίηση της βάσης δεδομένων που θα ελέγχει την αποδοτικότητα του συστήματος χρησιμοποιήθηκαν οι τελευταίες 10 χρονικά εκπομπές. Έχοντας ήδη κατηγοριοποιήσει όλες τις προτάσεις μας με βάση τις συνθήκες ομιλίας (F0,F1,F2) το μόνο που έχουμε να κάνουμε πλέον για τη δημιουργία των δεδομένων έλεγχου είναι να απομονώσουμε τις προτάσεις εκείνες που περιλαμβάνονται στις 10 τελευταίες εκπομπές. Παρακάτω παρουσιάζεται ο αριθμός των προτάσεων για κάθε μια από τις τρεις κατηγορίες έλεγχου.

Κατηγορία	F0	F1	F2	Σύνολο
Αριθμός προτάσεων έλεγχου	1828	3020	140	4988

Πίνακας 2.9 : Ποσοτική περιγραφή των δεδομένων ελέγχου σε επίπεδο συνθήκης ομιλίας

Στη συνέχεια για τις ίδιες αυτές εκπομπές και για κάθε μια από τις παραπάνω κατηγορίες εξαγάγαμε τις προτάσεις ελέγχου για κάθε ομιλητή και τις παρουσιάζουμε στον *Πίνακα 2.10.*

TEST SETS				
Ομιλητές	**Αριθμός προτάσεων**			
	Spk_F0	**Spk_F1**	**Spk_F2**	***Σύνολο ανά ομιλητή***
Γ. Παπανδρέου	1	30	0	**31**
Γαβρά	227	79	0	**306**
Κακαβάς	9	8	17	**34**
Κανδύλη	15	12	0	**27**
Καψής	22	16	0	**38**
Κ. Καραμανλής	21	52	0	**73**
Λιναρδάτου	21	0	0	**21**
Λοβέρδος	17	0	0	**17**
Νιότης	18	9	0	**27**
Παπαηλιόπουλος	15	0	0	**15**
Πρετεντέρης	31	10	0	**41**
Σπυράκη	13	29	0	**42**
Τρέμη	64	12	0	**76**
Χαρίτος	2	23	29	**54**
Χούκλη	306	36	0	**342**
Σύνολο ανά κατηγορία	**782**	**316**	**46**	

Πίνακας 2.10 : Ποσοτική περιγραφή των δεδομένων ελέγχου σε επίπεδο ομιλητή

ΚΕΦΑΛΑΙΟ 3

ΤΑ ΒΑΣΙΚΑ ΜΟΝΤΕΛΑ ΑΝΑΓΝΩΡΙΣΗΣ

3.1 Εισαγωγή

Στο κεφάλαιο αυτό θα περιγράψουμε τη διαδικασία εκπαίδευσης των ακουστικών μας μοντέλων για την οποία έγινε χρήση του εργαλείου HTK. Όπως έχουμε ήδη αναφέρει στο κεφάλαιο 1 για την μοντελοποίηση της ανθρώπινης ομιλίας και συνεπώς των ακουστικών μας μοντέλων χρησιμοποιηθήκαν Hidden Markov Models (HMMs). Κατά τη διάρκεια του κεφαλαίου αυτού θα γίνει μια συνοπτική περιγραφή για το είδος και τον τρόπο εκπαίδευσης των αρχικών μας μοντέλων(Seed Models) όπως επίσης και μια αναφορά στο γλωσσικό μοντέλο που χρησιμοποιήθηκε.

3.2 Σχεδιαστικές επιλογές ακουστικών μοντέλων

3.2.1 Επεξεργασία Front-End

Όπως αναφέραμε αναλυτικά και στο *Κεφάλαιο 1* η αναγνώριση ξεκινά με το ψηφιοποιημένο 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 μας, για να προχωρήσουμε στην εκπαίδευση των ακουστικών μας μοντέλων θα πρέπει πρώτα να εφαρμόσουμε την διαδικασία της Front-end προεπεξεργασίας στα αρχεία ήχου που έχουμε στην κατοχή μας. Για την υλοποίηση της παραπάνω διαδικασίας έγινε χρήση του εργαλείου HTK και συγκεκριμένα της εντολής *HCopy* η οποία μετατρέπει τα σήματα φωνής σε διανυσματικές ακολουθίες. Η εντολή *HCopy* δέχεται σαν είσοδο ένα αρχείο(Configuration File) απαραίτητο για τον προσδιορισμό των παραμέτρων των εξαγόμενων MFCCs. Οι παράμετροι του αρχείου αυτού φαίνονται παρακάτω.

```
SOURCEFORMAT = WAV            # format of wav files
TARGETKIND = MFCC_D_A_C_K_Z_0 # create Mel-Frequency Cepstral Coefficients
                              # 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
```

Έτσι, όπως φαίνεται και στο παραπάνω αρχείο, δημιουργήθηκαν MFCCs για κάθε τεμαχισμένο ακουστικό σήμα της μορφής wav . Το κάθε MFCC περιέχει 39 στοιχεία. Ανάμεσα σε αυτά συντελεστές παραγώγου πρώτης και δεύτερης τάξης και μετρήσεις της ενέργειας του σήματος φωνής. Επίσης εκτελείται και μια κανονικοποίηση των μέσων τιμών των πινάκων εισόδου*(Cepstral Mean Normalization).* Τα MFCC υπολογίσθηκαν ανά 10 ms του σήματος φωνής χρησιμοποιώντας παράθυρο πλάτους 25ms.

3.2.2 Επιλογή γλωσσικής μονάδας

Όταν αποφασίζουμε να χρησιμοποιήσουμε HMMs για να μοντελοποιήσουμε την ανθρώπινη ομιλία, δημιουργείται ένα σημαντικό ερώτημα: ποια γλωσσική μονάδα θα χρησιμοποιηθεί προς μοντελοποιήση; 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' όψιν 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 είναι κακή. Δεν υπάρχει τρόπος για να δημιουργηθούν καινούργιες λέξεις, εφόσον έχουμε ένα καθορισμένο set λέξεων, κάνοντας έτσι τα μοντέλα 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

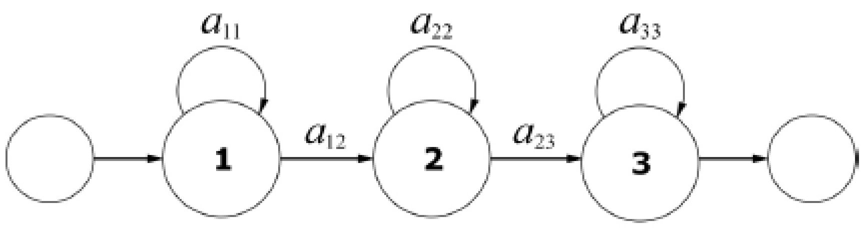
Στα πλαίσια αυτής τις διπλωματικής δεν ασχοληθήκαμε με την παράγωγη γλωσσικού μοντέλου και κάναμε χρήση του γλωσσικού μοντέλου που είχε εκπαιδευτεί από τον Τσεργουλα Ορφέα στα πλαίσια του Master του (*A Greek Broadcast News Transcription System[7]*). Το μοντέλο αυτό είχε εκπαιδευτεί με δεδομένα κειμένου από τις εφημερίδες *«Ελευθεροτυπία», «ΤΑ ΝΕΑ» και «ΤΟ ΒΗΜΑ»* και κάνοντας χρήση ενός συνόλου από 28 φωνήματα σχεδιασμένα για την Ελληνική γλωσσά από ειδικό γλωσσολόγο. Στον *Πίνακα 3.1* παρουσιάζεται το σύνολο των φωνημάτων μαζί με τα επιπλέον ειδικά φωνήματα που αντιπροσωπεύουν συγκεκριμένα γεγονότα.

Βασικά Φωνήματα	A, v, i, s, o, g, E, l, G, m, J, r, n, t, u, z, x, D, k, T, ly, f, p, d, C, b, N, c
Ειδικά Φωνήματα	sil (silence), hes (hesitation), bre (breath), sp (pause) fra (incomplete word), noi (noise), tbr (bad reading)

Πίνακας 3.1 : Τα φωνήματα του συστήματος

3.2.3 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 3.1 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 3.1 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 τρεις 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 3.1. 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.

3.3 Εκπαίδευση ακουστικών μοντέλων

Στα πλαίσια αυτού του κεφαλαίου θα δώσουμε μια σύντομη περιγραφή για τα κριτήρια εκπαίδευσης των δυο αρχικών μας μοντέλων(Seed Models). Τα αρχικά μας μοντέλα εκπαιδευτήκαν με δυο από τα βασικότερα κριτήρια εκπαίδευσης, το *Maximum Likelihood Estimation(MLE)* και το *Discriminative Training*.

Maximum Likelihood Estimation

Το MLE είναι το κριτήριο που εκπαιδεύονται τα περισσότερα συστήματα αναγνώρισης φωνής και χρησιμοποιείται συνήθως ως βασική αρχή για την αξιολόγηση νέων μεθόδων εκπαίδευσης. Η μέθοδος εκτίμησης μέγιστης πιθανοφάνειας έχει διάφορες ελκυστικές ιδιότητες. Κατ' αρχάς, σχεδόν πάντα έχει καλές ιδιότητες σύγκλισης καθώς ο αριθμός δειγμάτων εκπαίδευσης αυξάνει. Επιπλέον, η εκτίμηση μέγιστης πιθανοφάνειας μπορεί συχνά να είναι απλούστερη από τις εναλλακτικές μεθόδους, όπως οι Bayesians τεχνικές ή άλλες μέθοδοι. Συγκεκριμένα για την αναγνώριση φωνής το κριτήριο αυτό εφαρμόζεται ως εξής. Έστω ότι έχουμε ένα σύνολο δεδομένων εκπαίδευσης $\Psi(x_i,w_i)|1 \leq i \leq N$ όπου x_i είναι η ακολουθία των ακουστικών χαρακτηριστικών που έχει παραχθεί από την i-οστη πρόταση εκπαίδευσης και w_i είναι η καταγραφή του σε επίπεδο λέξεων.

Η μέθοδος εκτίμησης μέγιστης πιθανοφάνειας (MLE) στοχεύει στο να υπολογίσει ένα μοντέλο λ^* όπου:

$$\lambda^* = \arg\max_{\lambda} P(\Psi \mid \lambda) = \arg\max_{\lambda} \prod_{i=1}^{N} P(x_i, w_i \mid \lambda) = \arg\max_{\lambda} \sum_{i=1}^{N} \log P(x_i, w_i \mid \lambda) \qquad (3.3\text{-}1)$$

Κάνοντας χρήση του κανόνα αλυσίδας η πιθανότητα $P(x_i, w_i \mid \lambda)$ μπορεί να γραφτεί ως:

$$P(x_i, w_i \mid \lambda) = P(w_i \mid \lambda) P(x_i \mid w_{i,} \lambda) \qquad (3.3\text{-}2)$$

Η ποσότητα $P(w_i \mid \lambda)$ δηλώνει την εκ των προτέρων πιθανότητα να έχει ειπωθεί η ακολουθία λέξεων w_i και η τιμή της παράγεται από το γλωσσικό μοντέλο. Η ποσότητα $P(x_i \mid w_{i,} \lambda)$ δηλώνει την υπό συνθήκη πιθανότητα των παρατηρούμενων ακουστικών χαρακτηριστικών x_i δεδομένου w_i και η τιμή του παράγεται από το ακουστικό μοντέλο.

Για τα περισσότερα συστήματα αναγνώρισης φωνής η εκπαίδευση των γλωσσικών μοντέλων είναι ανεξάρτητη των ακουστικών. Για το λόγο αυτό θα ασχοληθούμε παρακάτω μόνο με το ακουστικό μοντέλο. Έτσι αφαιρώντας το γλωσσικό το MLE κριτήριο για την περιγραφή των ακουστικών μοντέλων είναι:

$$\lambda^* = \arg\max_{\lambda} \sum_{i=1}^{N} \log P(x_i \mid w_i, \lambda) \qquad (3.3\text{-}3)$$

Έχει αποδειχτεί πως εάν η αρχική παραδοχή των στατιστικών μοντέλων είναι σωστή και υπάρχει ένα μεγάλος αριθμός δεδομένων εκπαίδευσης το MLE κριτήριο μπορεί να παράγει μια αρκετά καλή εκτίμηση της υπό συνθήκης πιθανότητας $P(x|w,\lambda)$. Στη πράξη όμως είναι αρκετά δύσκολο να πληρούνται οι παραπάνω υποθέσεις αφού η ποσότητα των δεδομένων εκπαίδευσης συνήθως είναι περιορισμένη και οι αρχικές υποθέσεις είναι εμπειρικές. Συγκεκριμένα στην αναγνώριση φωνής η απομαγνητοφώνηση *w* μιας πρότασης *x* και οι ανταγωνιστικές υποθέσεις *h* με $h \neq w$ μοιράζονται συχνά τις ίδιες λέξεις ή φωνήματα. Η μέθοδος εκτίμησης μέγιστης πιθανοφάνειας στοχεύει στην μεγιστοποίηση της ποσότητας $P(x|w,\lambda)$, μπορεί όμως να αυξήσει σημαντικά και την ποσότητα $P(x|h,\lambda)$.

Discriminative Training

Σε αντίθεση με το MLE κριτήριο όπου η εκτίμηση των παραμέτρων των μοντέλων μιας κατηγορίας εξαρτάτε μόνο από τα δεδομένα εκπαίδευσης της συγκριμένης κατηγορίας, κατά το *Discriminative Training* γίνεται μια προσπάθεια ενσωμάτωσης της επιπλέον γνώσης από τα δεδομένα που σχετίζονται με τις ανταγωνιστικές κλάσεις. Συγκεκριμένα, το *Discriminative* κριτήριο όχι μόνο μεγιστοποιεί την υπό συνθήκη πιθανότητα($P(x|w,\lambda)$), αλλά προσπαθεί συγχρόνως να ελαχιστοποιήσει και την ποσότητα $P(x|h,\lambda)$ για τις ανταγωνιστικές υποθέσεις *h*.

Μια από τις πιο διαδεδομένες μεθόδους εφαρμογής του *Discriminative Training* είναι η Μεγιστοποίηση της Αμοιβαίας Πληροφορίας *(Maximum Mutual Information, MMI)*. Η βασική ιδέα του *MMI* κριτηρίου είναι η ελαχιστοποίηση της εντροπίας $E_\lambda(h|x)$ πάνω σε τυχαίες υποθέσεις της μεταβλητής h δεδομένου του εισερχομένου ακουστικού χαρακτηριστικού x. Αυτό αντιστοιχεί στην εκτίμηση των παραμέτρων του μοντέλου λ παρέχοντας όσο το δυνατόν περισσότερη πληροφορία για την επιθυμητή υπόθεση h δεδομένου του εισερχομένου πρότυπου x.

Η εντροπία για την υπόθεση h ορίζεται ως εξής :

$$\mathrm{E}_{\lambda}(h) = -\sum_{h} Ptrue(h) \log P(h \mid \lambda) \qquad (3.3\text{-}4)$$

Όπου $P_{true}(.)$ ορίζεται η αληθής κατανομή για τα παραγόμενα δεδομένα, ενώ η πιθανότητα $P(.|\lambda)$ ορίζεται ως το εμπειρικό μοντέλο βασισμένο στην εκτίμηση των κατανομών. Η παραπάνω εντροπία αφορά τις υποθέσεις χωρίς να γνωρίζουμε καμιά πληροφορία για τα ακουστικά χαρακτηριστικά. Η υπό συνθήκη εντροπία έχοντας πληροφορία για τα ακουστικά χαρακτηριστικά δίνεται παρακάτω:

$$\mathrm{E}_{\lambda}(h \mid x) = -\sum_{h,x} P_{true}(h,x) \log P(h \mid x, \lambda) \qquad (3.3\text{-}5)$$

Ο υπολογισμός της αμοιβαίας πληροφορίας ορίζεται ως η διάφορα των δυο παραπάνω εντροπιών:

$$\mathrm{I}_{\lambda}(h;x) = E_{\lambda}(h) - E_{\lambda}(h \mid x) = \sum_{h,x} P_{true}(h,x) \log \frac{P(h,x \mid \lambda)}{P(h \mid \lambda) P(x \mid \lambda)} \qquad (3.3\text{-}5)$$

Στόχος του MMI κριτηρίου είναι να μεγιστοποιήσει την αμοιβαία πληροφορία $I_{\lambda}(h;x)$ βελτιστοποιώντας το μοντέλο λ. Η τελική εξίσωση που επιλέγει τις τιμές για την βελτιστοποίηση του μοντέλου λ είναι η (3.3-6).

$$f_{MMI}(\lambda) = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \log \frac{P(h = w_i, x = x_i \mid \lambda)}{P(h = w_i \mid \lambda) P(x = x_i \mid \lambda)} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \log \frac{P(x_i \mid w_i, \lambda)}{\sum_h P(h, x_i \mid \lambda)} \qquad (3.3\text{-}6)$$

Τέλος συγκρίνοντας το MMI και MLE κριτήριο μπορούμε να πούμε πως το MLE ασχολείται μόνο με τη μεγιστοποίηση της ποσότητας $P(x|w,\lambda)$ ενώ από την άλλη το MMI επικεντρώνεται στην μεγιστοποίηση τις διαφοράς μεταξύ της ποσότητας $P(x|w,\lambda)$ και της πιθανότητας $P(x|\lambda)$. Το βασικό πλεονέκτημα του κριτήριου MMI είναι πως η μεγιστοποίηση της ποσότητας $f_{MMI}(\lambda)$ φαντάζει πιο λογική από τη μεγιστοποίηση της πιθανοφάνειας των δεδομένων εκπαίδευσης όταν η αρχική υπόθεση των μοντέλων δεν είναι ακριβής. Όμως το υπολογιστικό κόστος του MMI είναι μεγαλύτερο μιας και χρειάζεται να λαμβάνει υπ’ όψιν του όλες τις πιθανές κλάσεις σε αντίθεση με το MLE που χρειάζεται μονό τη σωστή.

3.4 Τα Βασικά μοντέλα και η MAP προσαρμογή

Στο κεφάλαιο αυτό θα δώσουμε μια περιγραφή των βασικών και των MAP προσαρμοσμένων ακουστικών μοντέλων που εκπαιδευτήκαν με τη βοήθεια του εργαλείου HTKv3.4. Στόχος μας ήταν η βελτίωση τις ακρίβειας των δυο αρχικών μοντέλων, του μοντέλου εκπαιδευμένο με MLE κριτήριο και του μοντέλου εκπαιδευμένο με MMI κριτήριο. Στο σημείο αυτό αξίζει να αναφερθεί πως τα δυο αυτά μοντέλα έχουν εκπαιδευτεί με 72 ώρες καθαρής ομιλίας από τον Δημήτρη Οικονομίδη στα πλαίσια τις μεταπτυχιακής του εργασίας.

Ύστερα από μια σειρά πειραματικών δοκιμών καταλήξαμε στην εφαρμογή της μεθόδου της MAP προσαρμογής(*MAP Adaptation βλέπε 1.5.2*) για τη βελτιστοποίηση της απόδοσης των δυο βασικών μας μοντέλων. Να αναφέρουμε στο σημείο αυτό πως τα δεδομένα εκπαίδευσης που χρησιμοποιηθήκαν για την προσαρμογή των ακουστικών μοντέλων είναι **14903** προτάσεις καθαρής ομιλίας, **35057** προτάσεις με περιβαλλοντικό θόρυβο ή μουσική και **2366** προτάσεις τηλεφωνικής ομιλίας. Στο *Πίνακα 3.2* δίνουμε μια συνοπτική παρουσίαση των δεδομένων αυτών όπως και των δεδομένων ελέγχου.

Κατηγορία	**F0**	**F1**	**F2**	**Σύνολο**
Προτάσεις εκπαίδευσης	14903	35057	2366	**52326**
Προτάσεις έλεγχου	1828	3020	140	**4988**

Πινάκας 3.2 : Συνοπτική παρουσίαση των δεδομένων εκπαίδευσης και ελέγχου

Στη συνέχεια θα δώσουμε μια περιγραφή των βασικών και των προσαρμοσμένων ακουστικών μοντέλων που παρήχθησαν με βάση τα δεδομένα που παρουσιάσαμε παραπάνω.

MLS : Το αρχικό μας μοντέλο εκπαιδευμένο με το MLE κριτήριο από την βάση δεδομένων της Λογοτυπογραφίας με 72 ώρες καθαρής ομιλίας.

MMIS : Ως MMIS έχουμε ονομάσει το μοντέλο που έχει παραχθεί ύστερα από την εφαρμογή του MMI κριτήριου πάνω στο MLS μοντέλο.

<u>MLS+MAP</u> : Για την παράγωγη αυτού του μοντέλου χρησιμοποιήσαμε το σύνολο των προτάσεων εκπαίδευσης (**52326)** από τα δελτία ειδήσεων και εφαρμόσαμε πάνω στο MLS μοντέλο την μέθοδο της MAP προσαρμογής.

<u>MMIS+MAP</u> : Κάνοντας χρήση των ίδιων προτάσεων όπως και παραπάνω εφαρμόσαμε MAP Adaptation στο MMIS και υλοποιήσαμε το MMIS+MAP μοντέλο.

3.5 Η Αξιολόγηση των μοντέλων

Έχοντας πλέον προσαρμόσει τα δυο βασικά μας μοντέλα με τη μέθοδο του MAP Adaptation, το επόμενο μας βήμα είναι η αξιολόγηση τις ακριβείας των μοντέλων αυτών. Φυσικά το αποτέλεσμα της αναγνώρισης φωνής είναι συνέπεια των ακουστικών μοντέλων που προαναφέραμε σε συνάρτηση πάντα με το γλωσσικό μοντέλο. Να αναφέρουμε πως για την αξιολόγηση όλων των ακουστικών μας μοντέλων χρησιμοποιήθηκε ένα **3-gram** γλωσσικό μοντέλο με λεξικό προφορών 100.000 λέξεων. Εφεξής αυτό το γλωσσικό μοντέλο θα θεωρείται δεδομένο και θα αναφέρουμε μόνο τα διάφορα ακουστικά μοντέλα.

Για την αξιολόγηση των μοντέλων μας χρησιμοποιήσαμε τα παρακάτω σύνολα δοκίμων(για την λεπτομερή παρουσίαση τους βλέπε **υποκεφάλαιο 2.4.3**):

- **Test F0 :** Σύνολο που περιέχει προτάσεις που έχουν σαν συνθήκη ομιλίας studio.
- **Test F1 :** Σύνολο που περιέχει προτάσεις που έχουν σαν συνθήκη ομιλίας περιβαλλοντικό θόρυβο ή μουσική
- **Test F2 :** Σύνολο που περιέχει προτάσεις που έχουν σαν συνθήκη ομιλίας ομιλία μέσω τηλεφωνικής συσκευής.

Η αξιολόγηση έγινε με την βοήθεια του εργαλείου HTKv3.4 και συγκεκριμένα με την εντολή HDecode, μια εντολή που προστέθηκε στην τελευταία έκδοση του εργαλείου και αφορά την αναγνώριση συστημάτων μεγάλου λεξιλογίου. Η μέτρηση της ακριβείας της αναγνώρισης σε κάθε περίπτωση έγινε με το εργαλείο HResults.

Στην παρουσίαση αποτελεσμάτων υπολογίζουμε την ακρίβεια των μοντέλων μας με βάση την ποσότητα Accuracy που ορίζεται ως εξής :

$$Accuracy = \frac{H - I}{N} X100\% \qquad (3.5\text{-}1)$$

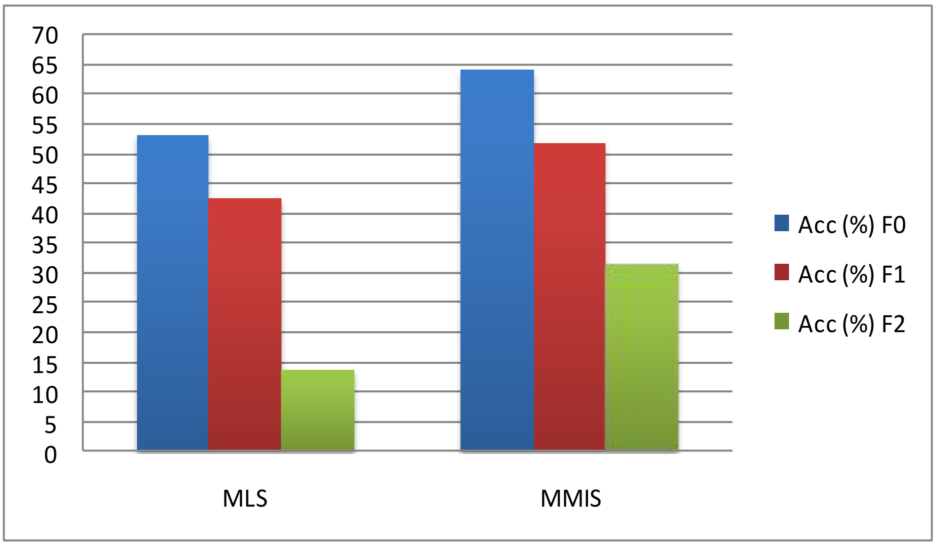
όπου H ο αριθμός των σωστών labels, I ο αριθμός των insertions και N ο συνολικός αριθμός των labels.

3.5.1 Αξιολόγηση των βασικών μοντέλων (Seed Models)

Σε αυτή την ενότητα θα αξιολογήσουμε την απόδοση των δυο αρχικών μας μοντέλων του MLS και του MMIS. Η αξιολόγηση των μοντέλων αυτών θα γίνει και με τα 3 σύνολα δεδομένων έλεγχου που δείξαμε παραπάνω. Στον *Πίνακα 3.3* παρουσιάζονται τα αποτελέσματα της αναγνώρισης των Seed Models για τα **Test F0, Test F1** και **Test F2.**

Models	**Acc (%)**		
	F0	**F1**	**F2**
MLS	53.15	42.58	13.72
MMIS	64.25	51.60	31.42

Πινάκας 3.3 : Αποτελέσματα των αρχικών μοντέλων



Σχήμα 3.2 : Γραφική απεικόνιση της απόδοσης των Seed Models

Όπως παρατηρούμε η ακρίβεια των δυο Seed Models δεν είναι και τόσο ικανοποιητική. Αυτό ήταν αναμενόμενο διότι τα μοντέλα αυτά δεν έχουν εκπαιδευτεί με δεδομένα από τηλεοπτικές εκπομπές δελτίων ειδήσεων αλλά με δεδομένα καθαρής ομιλίας από ηλεκτρονικές εφημερίδες, με αποτέλεσμα να μην ανταποκρίνονται με μεγάλη επιτυχία στην αναγνώριση των δικών μας προτάσεων. Παρόλα αυτά αξίζει να σχολιάσουμε την απόδοση των δυο κριτήριων εκπαίδευσης, του MLE και του MMI. Παρατηρούμε πως το μοντέλο που έχει εκπαιδευτεί με το MMI κριτήριο είναι αποδοτικότερο εκείνου που έχει εκπαιδευτεί με το MLE κριτήριο κάτι που έρχεται σε απόλυτη συμφωνία με την θεωρητική περιγραφή των δυο αυτών κριτήριων που δώσαμε στο προηγούμενο κεφάλαιο. Ειδικά για τα δεδομένα της κατηγορίας **F2** βλέπουμε μια αρκετά μεγάλη αύξηση της τάξης του 18% , όμως και στις υπόλοιπες δυο κατηγορίες η αύξηση περίπου κατά 9% δεν μπορεί σε καμιά περίπτωση να θεωρηθεί αμελητέα.

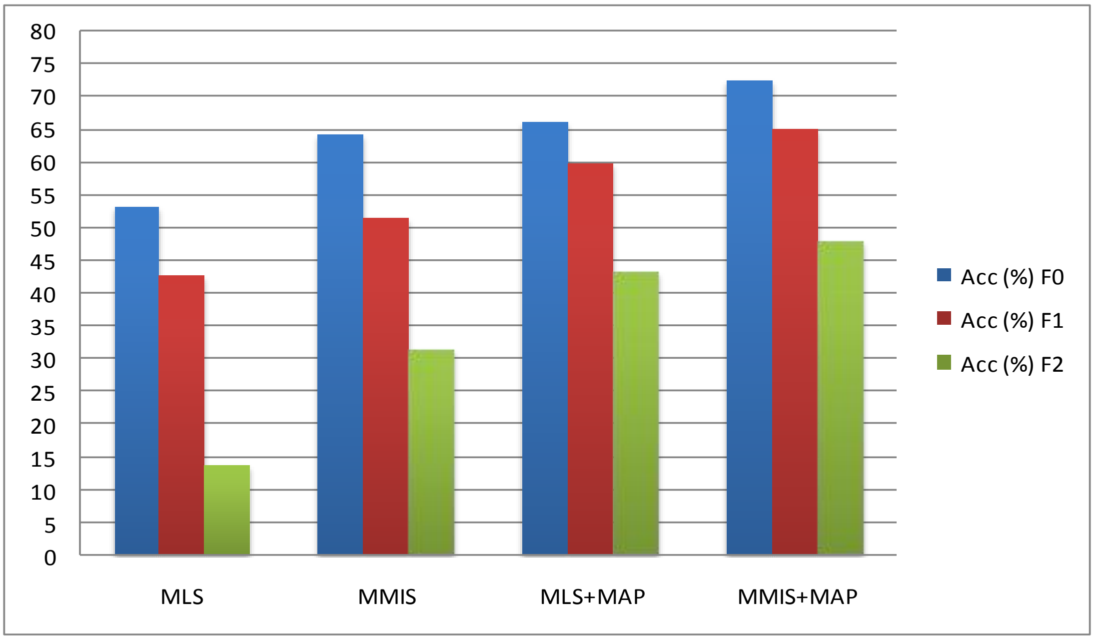
3.5.2 Αξιολόγηση των μοντέλων μετά από MAP Adaptation

Στη συνεχεία αφού έχουμε εφαρμόσει στα δυο βασικά μας μοντέλα τη μέθοδο του MAP Adaptation, όπως περιγράψαμε στην αρχή του κεφαλαίου, προχωρήσαμε στην αξιολόγηση τους. Παρακάτω παρουσιάζονται τα αποτελέσματα των μοντέλων μετά από MAP Adaptation πάνω στα Seed Models για τα σύνολα **Test F0, Test F1** και **Test F2.**

Models	**Acc (%)**		
	F0	**F1**	**F2**
MLS+MAP	66.20	59.68	43.24
MMIS+MAP	72.42	64.92	47.89

Πινάκας 3.4 : Αποτελέσματα των MAP προσαρμοσμένων μοντέλων

Στο *Σχήμα 3.3* παρουσιάζεται μια γραφική απεικόνιση της απόδοσης των αρχικών μοντέλων πριν και μετά από MAP Adaptation.



Σχήμα 3.3 : Γραφική απεικόνιση της απόδοσης των Seed Models πριν και μετά από MAP Adaptation

Όπως ήταν αναμενόμενο προσαρμόζοντας τα δυο αρχικά μοντέλα στα δεδομένα από τηλεοπτικές εκπομπές με την μέθοδο του MAP Adaptation, είχαμε μια αρκετά ικανοποιητική αύξηση της απόδοσης και στις τρεις κατηγορίες. Για το MLS μοντέλο παρατηρούμε αύξηση περίπου από 13% έως 30% ενώ για το MMIS υπάρχει μια αύξηση περίπου από 8% έως 16%. Παρά την αρκετά μεγαλύτερη αύξηση που υπάρχει στο προσαρμοζόμενο MLS μοντέλο, το MMIS+MAP ακόμα έχει την μεγαλύτερη ακρίβεια από τα δυο και στις τρεις κατηγορίες, φτάνοντας μάλιστα στο μέγιστο αποτέλεσμα για την κατηγορία **F0** περίπου στο 72%. Με βάση τα παραπάνω πειράματα καταλήγουμε πως σε κάθε περίπτωση το εκπαιδευμένο μοντέλο με MMI κριτήριο είναι σαφώς καλύτερο από το MLS και για το λόγο αυτό στη συνέχεια ασχοληθήκαμε μόνο με την βελτίωση της ακριβείας του αρχικού μοντέλου MMIS.

ΚΕΦΑΛΑΙΟ 4

Η ΜΕΘΟΔΟΣ ΤΗΣ ΠΡΟΣΑΡΜΟΓΗΣ ΣΤΟΝ ΟΜΙΛΗΤΗ

4.1 Εισαγωγή

Στα αυτόματα συστήματα αναγνώρισης φωνής μεγάλου λεξιλογίου, τα ακουστικά μοντέλα είναι εκπαιδευμένα με δεδομένα από έναν αριθμό διαφορετικών ομιλητών. Αυτή η προσέγγιση τα βοηθά να είναι αρκετά αποδοτικά στο σύνολο του πληθυσμού των ομιλητών. Τα συστήματα αυτά ονομάζονται συστήματα ανεξάρτητων ομιλητών(*Speaker-Independent, SI*) μιας και έχουν τη δυνατότητα να αναγνωρίζουν προτάσεις άγνωστων ομιλητών χωρίς να χρειάζεται να υποβάλλονται σε περαιτέρω εκπαίδευση. Εκτός από τα SI συστήματα υπάρχουν και συστήματα τα οποία έχουν εκπαιδευτεί με δεδομένα συγκεκριμένων ομιλητών, κάνοντας έτσι χρήση ακουστικών μοντέλων αποδοτικότερων για τους συγκεκριμένους ομιλητές. Τα συστήματα αυτά ονομάζονται συστήματα εξαρτώμενα από ομιλητές(*Speaker-Dependent, SD*) και έχουν αρκετά μεγαλύτερη απόδοση από τα SI συστήματα για συγκρίσιμα μεγέθη. Ένας από τους πλέον διαδεδομένους τρόπους για την δημιουργία SD συστημάτων είναι η μέθοδος της προσαρμογής στον ομιλητή(*Speaker Adaptation*). Στόχος του *Speaker Adaptation* είναι να καταφέρει να προσαρμόσει το SI σύστημα όσο το δυνατόν καλύτερα στους συγκεκριμένους ομιλητές δημιουργώντας εστί συστήματα αρκετά κοντά στα πραγματικά SD.

Στα πλαίσια αυτής διπλωματικής εργασίας προσπαθήσαμε να υλοποιήσουμε εξαρτώμενα από τον ομιλητή μοντέλα με τη μέθοδο του Speaker MLLR Adaptation. Η μέθοδος αυτή υλοποιεί την προσαρμογή στον ομιλητή με βάση την MLLR τεχνική(βλέπε υποκεφάλαιο **1.5.3**).

4.2 Τα προσαρμοσμένα μοντέλα στον ομιλητή

Στο κεφάλαιο αυτό θα δώσουμε μια περιγραφή των ακουστικών μοντέλων που υλοποιήσαμε στα πλαίσια της προσαρμογής στον ομιλητή με την μέθοδο του MLLR Adaptation. Να υπενθυμίσουμε στο σημεία αυτό πως κατά την διάρκεια της προσαρμογής χρησιμοποιήσαμε ως βασικά μοντέλα, το MMIS μοντέλο και το προσαρμοσμένο MMIS+MAP αποκλείοντας τα MLS και MLS+MAP που είχαν χειρότερη απόδοση όπως περιγράψαμε στο *Κεφάλαιο 3*.

Όπως έχουμε ήδη αναφέρει και σε προηγούμενο κεφάλαιο η προσαρμογή μπορεί να εφαρμοστεί είτε με τη μέθοδο της επίβλεψης*(Supervised)*, είτε χωρίς επίβλεψη*(Unsupervised)*. Για την υλοποίηση των παραπάνω μεθόδων έγινε χρήση των συνόλων δεδομένων εκπαίδευσης των ομιλητών για την *Supervised* μέθοδο και των συνόλων των δεδομένων έλεγχου για την Unsupervised . Στο *Πίνακα 4.1* δίνουμε μια συνοπτική παρουσίαση των δεδομένων αυτών.

Κατηγορία	**F0**	**F1**	**F2**	**Σύνολο**
Προτάσεις εκπαίδευσης ομιλητών	5523	4713	49	**10285**
Προτάσεις ελέγχου	1828	3020	140	**4988**
Προτάσεις ελέγχου ομιλητών	782	316	46	**1144**

Πινάκας 4.1 : Συνοπτική παρουσίαση των δεδομένων εκπαίδευση και έλεγχου

Παρακάτω θα δώσουμε μια περιγραφή των προσαρμοσμένων στον ομιλητή μοντέλων με την μέθοδο του *Supervised* και *Unsupervised Adaptation.*

MMIS+SSA : Στο πείραμα αυτό εφαρμόσαμε Supervised Speaker Adaptation πάνω στο MMIS μοντέλο κάνοντας χρήση 32 MLLR μετασχηματισμών και ως δεδομένα προσαρμογής χρησιμοποιηθήκαν οι συνολικές προτάσεις εκπαίδευσης για κάθε κατηγορία και για κάθε γνωστό ομιλητή.

MMIS+MAP+SSA : Για την δημιουργία του μοντέλου αυτού χρησιμοποιήσαμε τις προτάσεις εκπαίδευσης των 15 γνωστών ομιλητών και εφαρμόσαμε *Supervised Speaker Adaptation* πάνω στο αρχικό μοντέλο έπειτα από την MAP προσαρμογή(MMIS+MAP). Στα πλαίσια αυτού του πειράματος για τη υλοποίηση του

MLLR Adaptation πειραματιστήκαμε με διαφόρους αριθμούς εξαγόμενων μετασχηματισμών(1,2,4,8,16,24 και 32).

MMIS+MAP+USA : Στα πλαίσια του πειράματος αυτού εφαρμόσαμε την τεχνική της προσαρμογής στον ομιλητή χωρίς επίβλεψη*(Unsupervised Speaker Adaptation)* πάνω στο MMIS+MAP μοντέλο. Για την υλοποίηση του συγκεκριμένου μοντέλου χρησιμοποιήσαμε ως προτάσεις προσαρμογής τις προτάσεις έλεγχου των 15 γνωστών ομιλητών όπως αυτές αναγνωρίστηκαν από το μοντέλο MMIS+MAP. Και σε αυτό το πείραμα υλοποιήσαμε μοντέλα με διαφορετικούς αριθμούς MLLR μετασχηματισμών όπως και παραπάνω.

MMIS+MAP+SSA+USA : Για την δημιουργία του μοντέλου αυτού όπως εύκολα προδίδει και η ονομασία του εφαρμόσαμε την τεχνική του *Unsupervised Speaker Adaptation* πάνω στο ήδη προσαρμοσμένο στον ομιλητή μοντέλο με επίβλεψη **(**MMIS+MAP+SSA).

MMIS+MAP+(SSA+USA)+USA : Στη συνέχεια προχωρήσαμε το παραπάνω μοντέλο ένα βήμα παρακάτω και εφαρμόσαμε *Unsupervised Speaker Adaptation* πάνω στο MMIS+MAP+SSA+USA μοντέλο.

MMIS+MAP+(SSA+USA_Known_Spk)+(USA_Unknown_Spk) : Τέλος με χρήση της εμπειρίας που αποκτήσαμε από την δημιουργία των παραπάνω μοντέλων προχωρήσαμε στην υλοποίηση ενός πιο ρεαλιστικού πειράματος. Στα πλαίσια του πειράματος αυτού υλοποιήσαμε ένα σύστημα το οποίο θα προσαρμόζεται στον εκάστοτε ομιλητή με την μέθοδο του *Unsupervised Adaptation* καθώς νέα τμήματα προτάσεων θα εισέρχονται προς αναγνώριση. Συγκεκριμένα, το σύστημα μας θα δέχεται ως είσοδο ολόκληρα δελτία ειδήσεων προς αναγνώριση και στη συνέχεια καθ' όλη την διάρκεια του δελτίου θα «ακούει» μια μια τις προτάσεις των ομιλητών. Όσο ο ομιλητής παραμένει ο ίδιος το σύστημα ομαδοποιεί τις προτάσεις και εάν ο ομιλητής ανήκει σε έναν από τους 15 γνωστούς ομιλητές θα εφαρμόζει *Unsupervised MLLR Adaptation* πάνω στο MMIS+MAP+SSA. Εάν όχι θα εφαρμόζει *Unsupervised MLLR Adaptation* πάνω στο MMIS+MAP. Όταν αλλάξει ο ομιλητής εφαρμόζεται η ιδία τεχνική για το τμήμα του συγκεκριμένου ομιλητή και συνεχίζεται η ίδια διαδικασία για όλα τα δεδομένα προς αναγνώριση. Τέλος για να είναι το σύστημα μας πιο ευέλικτο, κατά την εφαρμογή του Unsupervised MLLR Adaptation επιλέγεται διαφορετικός αριθμός MLLR μετασχηματισμών ανάλογα με τον αριθμό των συνεχόμενων προτάσεων του εκάστοτε ομιλητή. Παρακάτω παρουσιάζεται ο ψευδοκώδικας που υλοποιεί το παραπάνω πείραμα.

Πάρε ολόκληρα δελτία από τα δεδομένα ελέγχου

Διάβασε μια μια τις προτάσεις

Όσο ο ομιλητής παραμένει ο ίδιος

αποθήκευε τις προτάσεις του

Όταν ο ομιλητής αλλάζει

Αν ο προηγούμενος ομιλητής ανήκει στους 15 γνωστούς ομιλητές

Εφάρμοσε USA πάνω στο MMIS+MAP+SSA με τις αποθηκευμένες προτάσεις

Επιλογή MLLR μετασχηματισμού

- *Όταν* ***«αριθμός προτάσεων ≤ 5»*** *εφαρμόζουμε MLLR-2*
- *Όταν* ***«5< αριθμός προτάσεων ≤10»*** *εφαρμόζουμε MLLR-4*
- *Όταν* ***«10< αριθμός προτάσεων ≤20»*** *εφαρμόζουμε MLLR-8*
- *Όταν* ***«20< αριθμός προτάσεων ≤30»*** *εφαρμόζουμε MLLR-16*
- *Σε διαφορετική περίπτωση εφαρμόζουμε MLLR-32*

Αν ο προηγούμενος ομιλητής ***δεν*** *ανήκει στους 15 γνωστούς ομιλητές*

Εφάρμοσε USA πάνω στο MMIS+MAP με τις αποθηκευμένες προτάσεις

Επιλογή MLLR μετασχηματισμού

- *Όταν* ***«αριθμός προτάσεων ≤ 5»*** *εφαρμόζουμε MLLR-2*
- *Όταν* ***«5< αριθμός προτάσεων ≤10»*** *εφαρμόζουμε MLLR-4*
- *Όταν* ***«10< αριθμός προτάσεων ≤20»*** *εφαρμόζουμε MLLR-8*
- *Όταν* ***«20< αριθμός προτάσεων ≤30»*** *εφαρμόζουμε MLLR-16*
- *Σε διαφορετική περίπτωση εφαρμόζουμε MLLR-32*

Συνέχισε την ίδια διαδικασία για όλα τα δεδομένα ελέγχου

4.3 Η Αξιολόγηση των μοντέλων

Ύστερα από την υλοποίηση των παραπάνω προσαρμοσμένων μοντέλων, το επόμενο μας βήμα ήταν η αξιολόγηση της απόδοσης τους. Για την αξιολόγηση των μοντέλων μας χρησιμοποιήσαμε τα παρακάτω σύνολα δοκιμών(για την λεπτομερή παρουσίαση τους βλέπε **υποκεφάλαιο 2.4.3**):

- **Test F0 :** Σύνολο που περιέχει προτάσεις που έχουν σαν συνθήκη ομιλίας studio.
- **Test F1 :** Σύνολο που περιέχει προτάσεις που έχουν σαν συνθήκη ομιλίας περιβαλλοντικό θόρυβο ή μουσική
- **Test F2 :** Σύνολο που περιέχει προτάσεις που έχουν σαν συνθήκη ομιλίας ομιλία μέσω τηλεφωνικής συσκευής.
- **Test Spk_F0 :** Υποσύνολο του **Test F0** για τους 15 γνωστούς ομιλητές
- **Test Spk_F1 :** Υποσύνολο του **Test F1** για τους 15 γνωστούς ομιλητές
- **Test Spk_F2 :** Υποσύνολο του **Test F2** για τους 15 γνωστούς ομιλητές

4.3.1 Αξιολόγηση μετά από Supervised Speaker Adaptation

Στα πλαίσια αυτού του κεφαλαίου θα παρουσιάσουμε τα αποτελέσματα από μια σειρά πειραμάτων που έγιναν στα πλαίσια τη μεθόδου της προσαρμογής στον ομιλητή με την μέθοδο της *Supervised Adaptation*. Όπως έχουμε αναφέρει και σε προηγούμενο κεφάλαιο θα αξιολογήσουμε τα προσαρμοσμένα μοντέλα μας με δεδομένα έλεγχου από τους 15 γνωστούς ομιλητές για τους οποίους έγινε η προσαρμογή.

Παρακάτω παρουσιάζονται τα αποτελέσματα του αρχικού MMIS μοντέλου, του προσαρμοσμένου MMIS+MAP και των μοντέλων μετά από Speaker MLLR Adaptation 32 μετασχηματισμών πάνω στο MMIS μοντέλο και στο MMIS+MAP για τα **Test Spk_F0, Test Spk_F1** και **Test Spk_F2.**

Models	Acc (%)		
	Spk_F0	**Spk_F1**	**Spk_F2**
MMIS	67.31	57.01	29.91
MMIS+MAP	76.92	72.90	51.14
MMIS+SSA	73.53	69.30	48.97
MMIS+MAP+SSA	78.43	75.02	56.73

Πινάκας 4.2 : Αποτελέσματα για τα σύνολα δοκιμών των ομιλητών μετά από MAP και Speaker MLLR Adaptation

Στη συνέχεια παρουσιάζονται στους παρακάτω πινάκες τα αποτελέσματα των παραπάνω τεσσάρων μοντέλων για κάθε ομιλητή ξεχωριστά και για κάθε κατηγορία.

Ομιλητές	Spk_F0(Acc%)			
	MMIS	**MMIS+SSA**	**MMIS+MAP**	**MMIS+MAP+SSA**
Γ. Παπανδρέου	51.61	**64.52**	61.29	**70.97**
Γαβρά	76.90	**79.93**	83.63	**84.30**
Κακαβάς	58.17	**63.40**	69.93	**69.93**
Κανδύλη	67.43	**75.66**	74.01	**80.26**
Καψής	47.52	**52.48**	60.64	**64.43**
Κ. Καραμανλής	65.54	**82.43**	82.77	**87.16**
Λιναρδάτου	65.00	**75.95**	76.90	**81.19**
Λοβέρδος	37.16	**53.21**	60.09	**65.60**
Νιότης	59.82	**66.57**	74.78	**73.90**
Παπαηλιόπουλος	68.39	**72.58**	70.65	**73.87**
Πρετεντέρης	37.34	**53.08**	60.23	**60.66**
Σπυράκη	69.96	**78.75**	78.02	**82.05**
Τρέμη	65.88	**74.41**	79.31	**80.00**
Χαρίτος	73.68	**80.70**	80.70	**84.21**
Χούκλη	67.27	**72.90**	75.70	**77.13**

Πινάκας 4.3 : Αποτελέσματα για ***Test Spk_F0*** *μετά από MAP και Supervised Speaker MLLR Adaptation*

Ομιλητές	Spk_F1(Acc%)			
	MMIS	**MMIS+SSA**	**MMIS+MAP**	**MMIS+MAP+SSA**
Γ. Παπανδρέου	65.07	**70.09**	76.71	**78.31**
Γαβρά	56.82	**69.10**	71.07	**73.05**
Κακαβάς	79.43	**85.11**	85.11	**90.07**
Κανδύλη	75.56	**79.11**	81.78	**80.44**
Καψής	48.61	**66.93**	68.53	**70.92**
Κ. Καραμανλής	55.03	**72.64**	72.48	**80.35**
Νιότης	69.19	**68.65**	83.78	**82.70**
Πρετεντέρης	35.24	**49.52**	53.33	**57.14**
Σπυράκη	64.14	**75.77**	80.78	**80.61**
Τρέμη	66.84	**71.50**	75.13	**77.20**
Χαρίτος	44.96	**61.83**	64.64	**65.57**
Χούκλη	35.38	**56.15**	63.59	**64.10**

*Πινάκας 4.4 : Αποτελέσματα για **Test Spk_F1** μετά από MAP και Supervised Speaker MLLR Adaptation*

Ομιλητές	Spk_F2(Acc%)			
	MMIS	**MMIS+SSA**	**MMIS+MAP**	**MMIS+MAP+SSA**
Κακαβάς	43.56	**46.32**	58.90	**56.44**
Χαρίτος	21.82	**50.55**	46.55	**56.91**

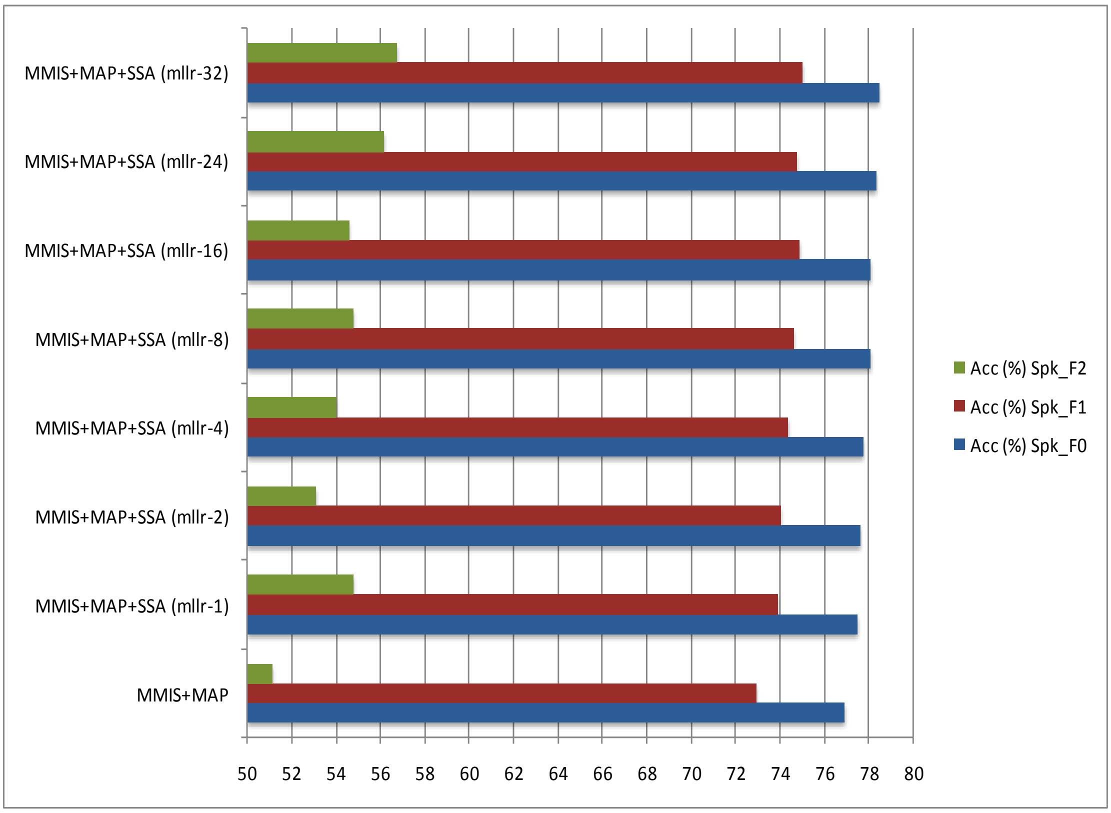
*Πινάκας 4.5 : Αποτελέσματα για **Test Spk_F2** μετά από MAP και Supervised Speaker MLLR Adaptation*

Κατά την αξιολόγηση των παραπάνω μοντέλων παρατηρούμε πως η εφαρμογή του *Supervised Speaker MLLR Adaptation* στο ήδη προσαρμοσμένο MMIS+MAP μοντέλο μας δίνει καλύτερα αποτελέσματα από την προσαρμογή στο MMIS .

Έχοντας υπ' όψιν μας τα παραπάνω καταλήξαμε ότι στη συνέχεια των πειραμάτων θα ορίσουμε ως αρχικό μοντέλο για την προσαρμογή στον ομιλητή το μοντέλο **MMIS+MAP** και για την εφαρμογή του *Supervised Speaker MLLR Adaptation* θα πειραματιστούμε με διαφόρους αριθμούς MLLR μετασχηματισμών. Παρακάτω παρουσιάζονται τα αποτελέσματα στο σύνολο των ομιλητών για τα μοντέλα MMIS+MAP και MMIS+MAP+SSA για διαφορετικούς αριθμούς MLLR μετασχηματισμών.

Models	Acc (%)		
	Spk_F0	**Spk_F1**	**Spk_F2**
MMIS+MAP	76.92	72.90	51.14
MMIS+MAP+SSA (mllr-1)	77.49	73.88	54,80
MMIS+MAP+SSA (mllr-2)	77.62	74.05	53.08
MMIS+MAP+SSA (mllr-4)	77.74	74.35	54.00
MMIS+MAP+SSA (mllr-8)	78.10	74.60	54.79
MMIS+MAP+SSA (mllr-16)	78.07	74.89	54.57
MMIS+MAP+SSA (mllr-24)	78.30	74.72	56.17
MMIS+MAP+SSA (mllr-32)	78.43	75.02	56.73

Πινάκας 4.6 : Αποτελέσματα μετά από Supervised Speaker Adaptation για διαφορετικούς αριθμούς MLLR μετασχηματισμών



Σχήμα 4.1 : Γραφική απεικόνιση αποτελεσμάτων του Supervised Speaker Adaptation για διαφορετικούς αριθμούς MLLR μετασχηματισμών.

Όπως παρατηρούμε και από την παραπάνω γραφική παράσταση, η εξαγωγή 32 μετασχηματισμών κατά την εφαρμογή του Adaptation έχει καλυτέρα αποτελέσματα στο σύνολο των ομιλητών. Αυτό ήταν εν μέρει αναμενόμενο διότι ο αριθμός των προτάσεων εκπαίδευσης για τους περισσότερους ομιλητές ήταν αρκετά μεγάλος ώστε να μπορεί να υποστηρίζει την εφαρμογή του MLLR 32 μετασχηματισμών. Θα είχε όμως ενδιαφέρον να δούμε πως συμπεριφερθήκαν όλοι οι ομιλητές ξεχωριστά για κάθε έναν από τους παραπάνω μετασχηματισμούς. Στους επομένους πινάκες που ακολούθουν παρουσιάζονται τα αποτελέσματα του Supervised Speaker MLLR Adaptation για κάθε ομιλητή στις τρεις κατηγορίες (**Spk_F0,Spk_F1,Spk_F2**) και για κάθε αριθμό MLLR μετασχηματισμών .

Ομιλητές	**Προτάσεις**		**Spk_F0(Acc%)**						
	Train	**Test**	**mllr-1**	**mllr-2**	**mllr-4**	**mllr-8**	**mllr-16**	**mllr-24**	**mllr-32**
Γ. Παπανδρέου	41	1	70.97	74.19	70.97	74.19	64.52	70.97	70.97
Γαβρά	1466	227	83.63	83.75	83.66	84.04	83.68	84.42	84.30
Κακαβάς	25	9	69.93	70.59	69.93	69.93	69.93	69.93	69.93
Κανδύλη	70	15	75.99	76.32	77.30	78.29	76.64	77.96	80.26
Καψής	36	22	63.27	63.85	62.39	64.72	64.43	64.43	64.43
Κ. Καραμανλής	40	21	84.46	84.12	85.81	86.15	87.16	87.16	87.16
Λιναρδάτου	36	21	78.33	78.33	78.10	79.52	80.00	81.19	81.19
Λοβέρδος	58	17	68.81	66.06	68.35	66.51	68.81	67.43	65.60
Νιότης	65	18	76.54	77.42	74.78	76.25	73.31	74.49	73.90
Παπαηλιόπουλος	7	15	72.58	72.90	73.55	73.87	74.19	74.84	73.87
Πρετεντέρης	94	31	60.23	60.09	61.37	60.94	61.80	61.23	60.66
Σπυράκη	237	13	79.85	79.85	80.59	80.59	80.95	82.05	82.05
Τρέμη	1680	64	77.65	78.92	79.22	79.51	80.00	79.51	80.00
Χαρίτος	0	2	84.21	82.46	82.46	82.46	80.70	82.46	84.21
Χούκλη	1668	306	76.28	76.30	76.43	76.71	76.83	77.48	77.13

Πίνακας 4.7 : Αποτελέσματα για ***Test Spk_F0*** *μετά από Supervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών*

Ομιλητές	Προτάσεις		Spk_F1(Acc%)						
	Train	Test	mllr -1	mllr -2	mllr -4	mllr -8	mllr -16	mllr -24	mllr -32
Γ. Παπανδρέου	221	30	76.71	76.26	77.40	76.71	77.40	77.40	78.31
Γαβρά	659	79	73.30	73.22	72.45	72.96	73.22	73.13	73.05
Κακαβάς	75	8	86.52	86.52	88.65	85.82	90.07	90.07	90.07
Κανδύλη	150	12	81.33	80.44	80.44	80.89	80.00	80.44	80.44
Καψής	356	16	70.92	71.71	72.91	74.10	72.51	70.12	70.92
Κ.Καραμανλής	303	52	77.52	77.04	77.67	78.62	79.40	79.56	80.35
Νιότης	137	9	81.62	80.00	81.08	82.16	84.32	83.78	82.70
Πρετεντέρης	761	10	51.43	55.24	54.29	56.19	56.19	56.19	57.14
Σπυράκη	244	29	80.13	81.26	80.78	80.13	80.45	80.29	80.61
Τρέμη	886	12	75.13	75.65	76.68	76.17	76.17	77.20	77.20
Χαρίτος	188	23	63.93	64.64	65.34	66.28	66.28	65.81	65.57
Χούκλη	267	36	62.31	62.82	64.10	64.10	63.33	63.08	64.10

*Πίνακας 4.8 : Αποτελέσματα για **Test Spk_F1** μετά από Supervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών*

Ομιλητές	Προτάσεις		Spk_F2(Acc%)						
	Train	Test	mllr-1	mllr-2	mllr-4	mllr-8	mllr-16	mllr-24	mllr-32
Κακαβάς	20	17	57.06	55.52	56.42	57.36	56.75	56.44	56.44
Χαρίτος	29	29	53.45	51.64	52.55	53.27	53.27	56.00	56.91

*Πίνακας 4.9 : Αποτελέσματα για **Test Spk_F2** μετά από Supervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών*

Εκτός από ελάχιστες περιπτώσεις ομιλητών παρατηρούμε πως το mllr-32 έχει τις καλύτερες επιδόσεις από τους υπόλοιπους αριθμούς μετασχηματισμών. Όμως αξίζει να σημειώσουμε πως σε μερικούς ομιλητές, συνήθως με μικρό αριθμό προτάσεων ή με μικρό αριθμό λέξεων ανά πρόταση, η εξαγωγή 32 μετασχηματισμών δεν απέδωσε στον επιθυμητό βαθμό. Αυτό είναι μια πολύ χρήσιμη πληροφορία που στη συνέχεια θα πρέπει να λάβουμε σοβαρά υπ’ όψιν μας.

4.3.2 Αξιολόγηση μετά από Unsupervised Speaker Adaptation

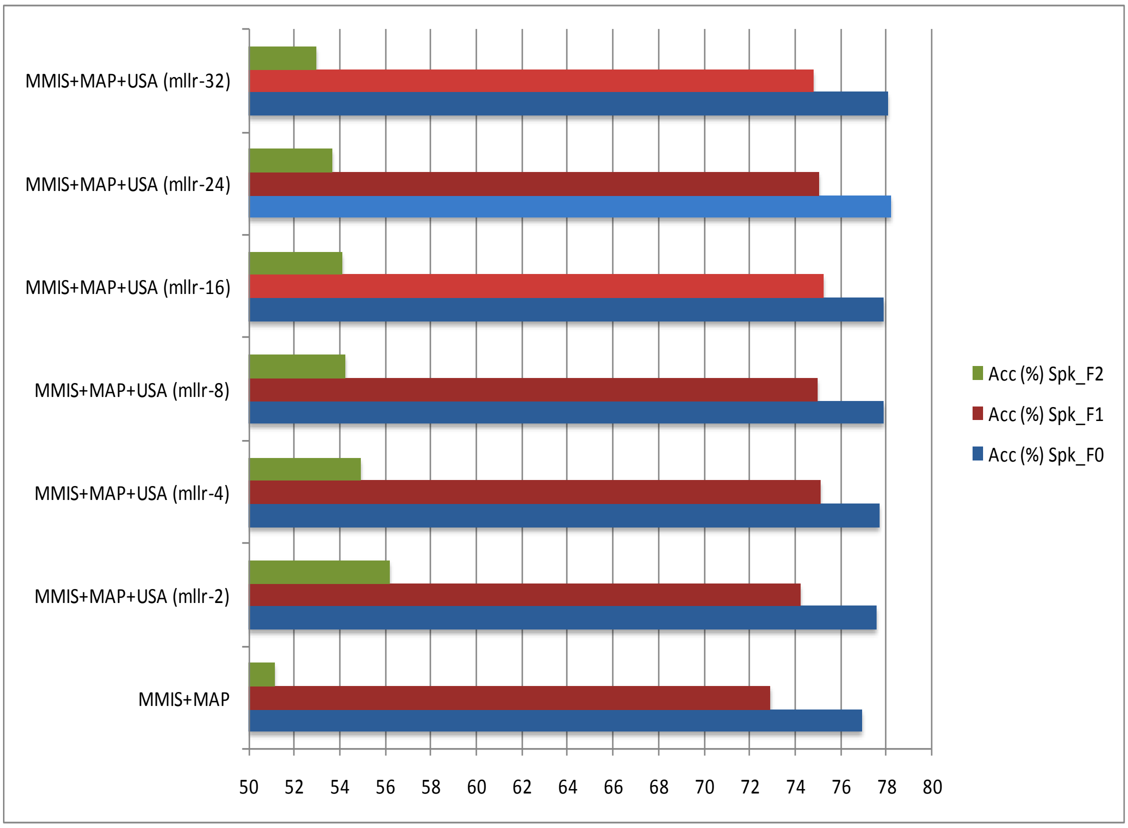
Στο κεφάλαιο αυτό θα παρουσιάσουμε τα αποτελέσματα από μια σειρά πειραμάτων που έγιναν στα πλαίσια της μεθόδου της προσαρμογής στον ομιλητή με την μέθοδο του *Unsupervised Adaptation*. Όπως έχουμε αναφέρει και σε προηγούμενο κεφάλαιο η προσαρμογή και η αξιολόγηση των προσαρμοσμένων μοντέλων μας, έγινε με τα δεδομένα έλεγχου των 15 γνωστών ομιλητών αλλά και με το σύνολο των δεδομένων έλεγχου του συστήματος μας.

Παρακάτω παρουσιάζονται τα αποτελέσματα στο σύνολο των ομιλητών για τα μοντέλα MMIS+MAP και MMIS+MAP+USA για διαφορετικούς αριθμούς MLLR μετασχηματισμών για τα **Test Spk_F0, Test Spk_F1** και **Test Spk_F2.**

Models	Acc (%)		
	Spk_F0	**Spk_F1**	**Spk_F2**
MMIS+MAP	76.92	72.90	51.14
MMIS+MAP+USA (mllr-2)	77.55	74.24	**56.16**
MMIS+MAP+USA (mllr-4)	77.69	75.08	54.91
MMIS+MAP+USA (mllr-8)	77.89	74.96	54.22
MMIS+MAP+USA (mllr-16)	77.91	**75.25**	54.11
MMIS+MAP+USA (mllr-24)	**78.20**	75.06	53.65
MMIS+MAP+USA (mllr-32)	78.09	74.81	52.97

Πίνακας 4.10 : Αποτελέσματα μετά από Unsupervised Speaker Adaptation για διαφορετικούς αριθμούς MLLR μετασχηματισμών

Από τον *Πίνακα 4.10* παρατηρούμε πως η εφαρμογή του mllr-32 δεν είναι η καταλληλότερη επιλογή για την Unsupervised μέθοδο. Συγκεκριμένα και στις τρεις κατηγορίες την καλύτερη απόδοση την είχαν διαφορετικοί αριθμοί μετασχηματισμών, για την κατηγορία Spk_F0 ο mllr-24, για την Spk_F1 ο mllr-16 και για την κατηγορία Spk_F2 ο mllr-2. Στο *Σχήμα 4.2* μπορούμε να δούμε γραφικά την διακύμανση των αποτελεσμάτων ανάλογα με τον αριθμό των μετασχηματισμών.



Σχήμα 4.2 : Γραφική απεικόνιση αποτελεσμάτων του Unsupervised Speaker Adaptation για διαφορετικούς αριθμούς MLLR μετασχηματισμών.

Όπως παρατηρήσαμε και παραπάνω για την εφαρμογή του *Unsupervised Adaptation* για να έχουμε την καλύτερη δυνατή απόδοση θα πρέπει να εφαρμόσουμε διαφορετικό αριθμό MLLR μετασχηματισμών για κάθε κατηγορία. Όμως πριν βγάλουμε συμπεράσματα για τον αριθμό των MLLR μετασχηματισμών που θα πρέπει να χρησιμοποιήσουμε, είναι ενδιαφέρον να ερευνήσουμε σε βάθος την απόδοση του κάθε ομιλητή για κάθε MLLR μετασχηματισμό και για κάθε κατηγορία. Οι αποδόσεις των ομιλητών για τις κατηγορίες Spk_F0 , Spk_F1 και Spk_F2 παρουσιάζονται στους πινάκες *4.11*, *4.12* και *4.13* αντίστοιχα.

Ομιλητές	Προτάσεις	Spk_F0(Acc%)					
	Test	mllr-2	mllr-4	mllr-8	mllr-16	mllr-24	mllr-32
Γ. Παπανδρέου	1	61.29	61.29	61.29	58.06	58.06	58.06
Γαβρά	227	83.85	83.78	84.04	84.06	84.23	84.08
Κακαβάς	9	70.59	69.28	67.32	69.28	68.63	68.63
Κανδύλη	15	76.32	76.32	76.64	75.66	75.66	75.66
Καψής	22	61.22	62.97	61.52	62.97	63.85	62.68
Κ. Καραμανλής	21	85.14	84.46	85.14	84.80	83.45	83.78
Λιναρδάτου	21	78.10	79.05	79.52	79.52	79.76	79.76
Λοβέρδος	17	61.93	61.47	61.47	61.47	61.47	61.47
Νιότης	18	74.49	75.95	74.49	76.54	75.66	75.66
Παπαηλιόπουλος	15	72.58	73.23	73.87	72.90	72.90	72.90
Πρετεντέρης	31	61.95	60.52	62.09	60.66	62.52	62.37
Σπυράκη	13	80.59	80.22	79.85	77.29	78.39	80.22
Τρέμη	64	79.71	80.10	79.80	79.61	80.20	80.10
Χαρίτος	2	82.46	84.21	82.46	82.46	84.21	84.21
Χούκλη	306	76.17	76.47	76.78	77.03	77.30	77.13

*Πίνακας 4.11 : Αποτελέσματα για **Test Spk_F0** μετά από Unsupervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών*

Ομιλητές	Προτάσεις	Spk_F1(Acc%)					
	Test	mllr-2	mllr-4	mllr-8	mllr-16	mllr-24	mllr-32
Γ. Παπανδρέου	30	75.34	78.54	79.45	78.77	78.54	78.54
Γαβρά	79	71.13	73.39	73.91	73.22	74.16	73.99
Κακαβάς	8	82.27	83.69	84.40	85.11	85.11	85.11
Κανδύλη	12	81.33	81.33	81.33	80.44	80.00	80.00
Καψής	16	71.71	72.51	71.71	73.71	73.31	73.71
Κ. Καραμανλής	52	78.46	78.14	77.99	78.93	76.73	76.42
Νιότης	9	82.16	84.32	85.41	84.32	84.32	84.32
Πρετεντέρης	10	55.24	56.19	60.95	55.24	55.24	55.24
Σπυράκη	29	81.26	81.74	81.74	82.39	82.23	81.91
Τρέμη	12	76.68	76.68	76.17	75.65	75.13	75.65
Χαρίτος	23	66.04	67.21	65.81	67.21	67.21	65.81
Χούκλη	36	64.10	64.10	63.33	64.10	63.85	63.33

*Πίνακας 4.12 : Αποτελέσματα για **Test Spk_F1** μετά από Unsupervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών*

Ομιλητές	Προτάσεις	Spk_F2(Acc%)					
	Test	mllr-2	mllr-4	mllr-8	mllr-16	mllr-24	mllr-32
Κακαβάς	17	58.28	60.43	61.35	60.12	59.51	59.51
Χαρίτος	29	54.91	51.64	50.00	50.55	50.18	49.09

Πίνακας 4.13 : Αποτελέσματα για ***Test Spk_F2*** *μετά από Unsupervised Speaker MLLR Adaptation διαφορετικών MLLR μετασχηματισμών*

Από τους παραπάνω πίνακες παρατηρούμε πως τα καλυτέρα αποτελέσματα για κάθε ομιλητή δεν παράγονται από έναν συγκεκριμένο αριθμό μετασχηματισμών. Βλέπουμε πως υπάρχουν ομιλητές που για μεγάλο αριθμό μετασχηματισμών τα αποτελέσματα μειώνονται σημαντικά. Αυτό είναι κάτι το οποίο είναι αναμενόμενο διότι υπάρχουν ομιλητές με μικρό αριθμό προτάσεων όπου η εφαρμογή πολλών μετασχηματισμών είναι πλεονασμός, δηλαδή για να γίνουμε πιο συγκεκριμένοι δεν υπάρχουν αρκετές προτάσεις ώστε να μπορέσουν να εκπαιδεύσουν 32 μετασχηματισμούς. Σαν συνέπεια των παραπάνω αποφασίσαμε να εφαρμόσουμε την μέθοδο του *Unsupervised Adaptation* με μεταβαλλόμενο αριθμό MLLR μετασχηματισμών ανάλογα με τον αριθμό των προτάσεων του κάθε ομιλητή. Σκοπός μας είναι να βρούμε έναν κανόνα για τους MLLR μετασχηματισμούς, ο οποίος θα μας δώσει τα καλύτερα αποτελέσματα για κάθε ομιλητή και συνεπώς για κάθε κατηγορία. Συγκεκριμένα σκοπός μας είναι να φτάσουμε όσο το δυνατόν πιο κοντά στα καλύτερα αποτελέσματα για κάθε κατηγορία που παρουσιάστηκαν στον *Πίνακα 4.10.* Μετά από αρκετές δόκιμες και παρατηρήσεις των αποτελεσμάτων καταλήξαμε στον παρακάτω κανόνα απόφασης για τον αριθμό των MLLR μετασχηματισμών:

- Όταν ***«αριθμός προτάσεων ≤ 10»*** εφαρμόζουμε mllr-2
- Όταν ***«10< αριθμός προτάσεων ≤20»*** εφαρμόζουμε mllr-4
- Όταν ***«20< αριθμός προτάσεων ≤30»*** εφαρμόζουμε mllr-8
- Όταν ***«30< αριθμός προτάσεων ≤40»*** εφαρμόζουμε mllr-16
- Σε διαφορετική περίπτωση εφαρμόζουμε mllr-24

Κάνοντας λοιπόν χρήση του παραπάνω κανόνα τρέξαμε ξανά το πείραμα για το *Unsupervised Adaptation*. *Στον Πίνακα 4.14* παρουσιάζονται ο συνδυασμός των καλύτερων αποτελεσμάτων του *Πίνακα 4.10* και τα αποτελέσματα με μεταβλητό αριθμό MLLR μετασχηματισμών.

Models	**Acc (%)**		
	Spk_F0	**Spk_F1**	**Spk_F2**
MMIS+MAP+USA(Best Combination)	78.20*(mllr-24)*	75.25*(mllr-16)*	56.16*(mllr-2)*
MMIS+MAP+USA(Var-mllr)	78.16	74.99	55.59

Πίνακας 4.14 : Αποτελέσματα μετά από Unsupervised Speaker Adaptation για μεταβλητό αριθμό MLLR μετασχηματισμών

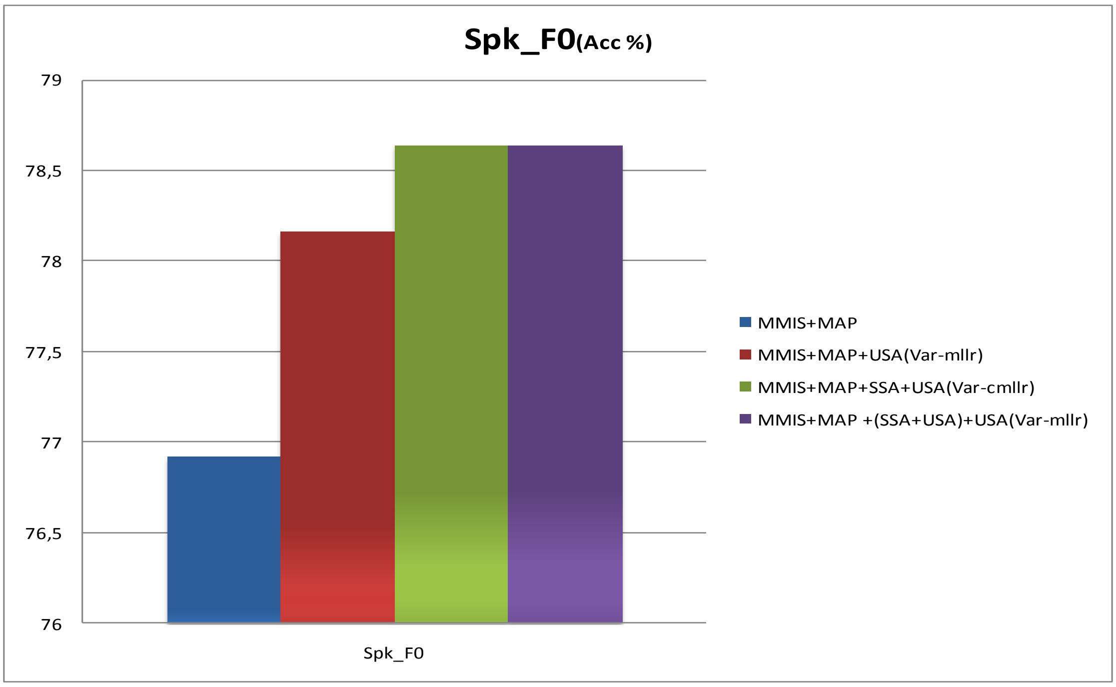
Όπως βλέπουμε από τον *Πινάκα 4.14* ο αρχικός μας στόχος, δηλαδή να φτάσουμε στον καλύτερο δυνατό συνδυασμό αποτελεσμάτων, έχει σχεδόν επιτευχθεί. Οι πολύ μικρές αποκλίσεις οφείλονται στο ότι κάνοντας χρήση του κανόνα απόφασης που προαναφέραμε, δεν λαμβάνουμε υπ’ όψιν μας τον αριθμό των λέξεων ανά πρόταση. Για παράδειγμα μπορεί κάποιος ομιλητής να έχει έναν μεγάλο αριθμό προτάσεων αλλά οι προτάσεις να περιέχουν ελάχιστες λέξεις. Παρ’ όλα αυτά οι αποκλίσεις είναι πολύ μικρές για να τις λάβουμε υπ’ όψιν μας.

Στη συνέχεια για την υλοποίηση των πειραμάτων μας, όπως είναι φυσικό, θα υιοθετήσουμε τον μεταβλητό τρόπο των μετασχηματισμών. Στον *Πίνακα 4.17* παρουσιάζονται τα αποτελέσματα όλων των Unsupervised μοντέλων που εκπαιδεύσαμε στο σύνολο των **Test Spk_F0, Test Spk_F1** και **Test Spk_F2** .

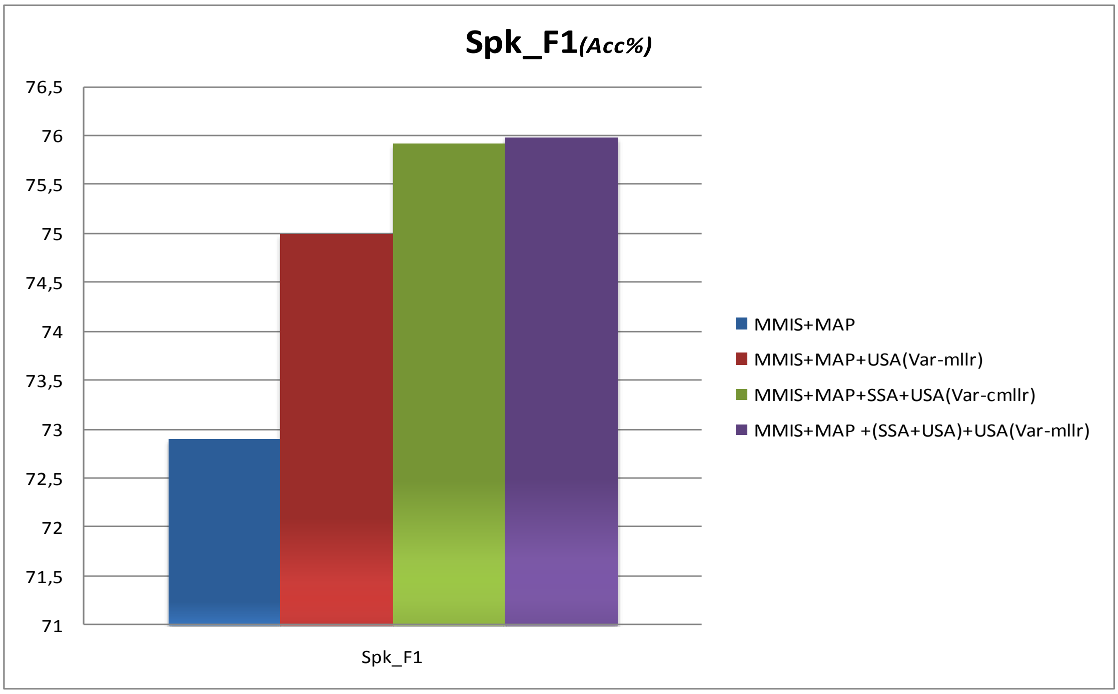
Models	**Acc (%)**		
	Spk_F0	**Spk_F1**	**Spk_F2**
MMIS+MAP	76.92	72.90	51.14
MMIS+MAP+USA(Var-mllr)	78.16	74.99	55.59
MMIS+MAP+SSA+USA(Var-mllr)	78.64	75.92	60.62
MMIS+MAP+(SSA+USA)+USA(Var-mllr)	78.64	75.98	60.96

Πίνακας 4.15 : Αποτελέσματα πριν και μετά από Unsupervised Speaker Adaptation για μεταβλητό αριθμό MLLR μετασχηματισμών

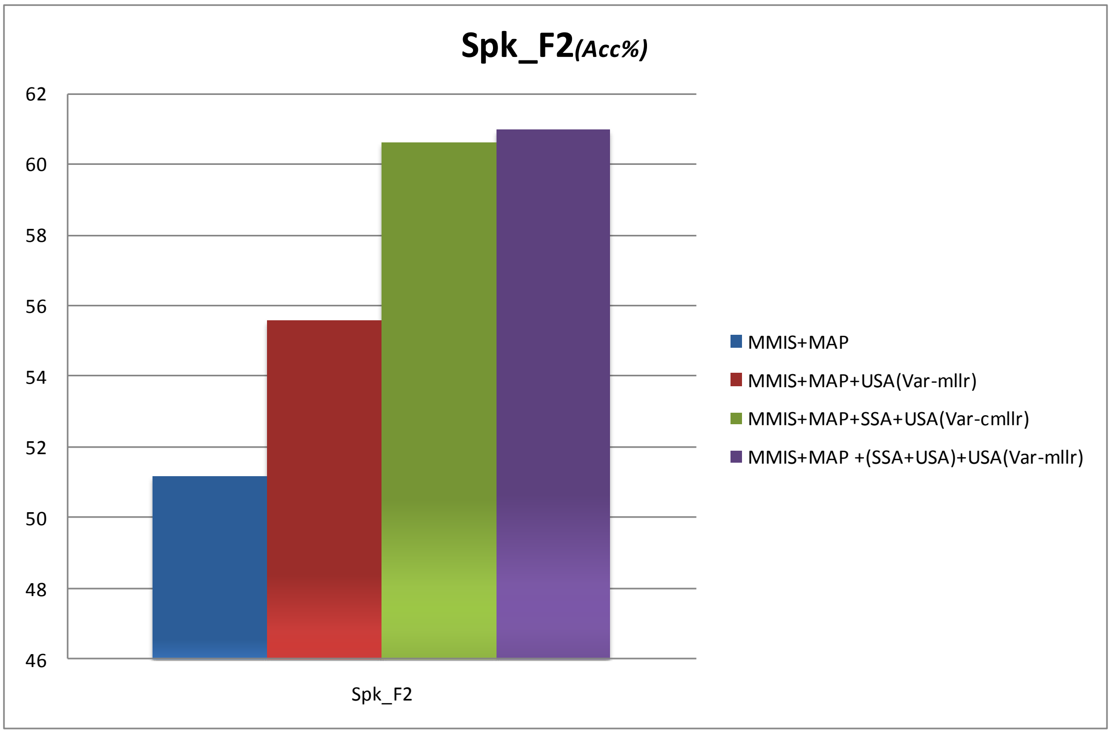
Από τα παραπάνω αποτελέσματα παρατηρούμε ότι με την εφαρμογή του *Unsupervised Speaker Adaptation(USA)* πάνω στο βασικό μας μοντέλο είχαμε μια αύξηση από 2% έως 4% περίπου. Έπειτα κατά την εφαρμογή του USA στο *Supervised* προσαρμοσμένο μοντέλο*(MMIS+MAP+SSA)* παρατηρούμε μια πολύ μικρή αύξηση στις κατηγορίες Spk_F0 και Spk_F1, όμως στην κατηγορία Spk_F2 έχουμε μια ικανοποιητική αύξηση κατά περίπου 5%. Εφαρμόζοντας USA στο προηγούμενο μοντέλο *(MMIS+MAP+(SSA+USA))* παρατηρούμε ότι ουσιαστικά δεν είχαμε καμιά σημαντική αύξηση. Τέλος, πριν βγάλουμε τα τελικά μας συμπεράσματα για τα αποτελέσματα των *Unsupervised* πειραμάτων θα παρουσιάσουμε την γραφική απεικόνιση των παραπάνω αποτελεσμάτων για κάθε κατηγορία ελέγχου.



*Σχήμα 4.3 : Γραφική απεικόνιση αποτελεσμάτων για **Spk_F0** σε όλα τα Unsupervised μοντέλα*



Σχήμα 4.4 : Γραφική απεικόνιση αποτελεσμάτων για ***Spk_F1*** *σε όλα τα Unsupervised μοντέλα*



Σχήμα 4.5 : Γραφική απεικόνιση αποτελεσμάτων για ***Spk_F2*** *σε όλα τα Unsupervised μοντέλα*

Συνοψίζοντας, παρατηρούμε πως η εφαρμογή του USA στα διάφορα μοντέλα μας δεν είχε την ίδια συμπεριφορά και στις τρεις κατηγορίες ελέγχου. Συγκεκριμένα όπως φαίνεται και από το *Σχήμα 4.3* για την κατηγορία Spk_F0 αν αφαιρέσουμε την εφαρμογή του USA στο βασικό μοντέλο (MMIS+MAP), όπου είχαμε μια ικανοποιητική αύξηση, για τα υπόλοιπα μοντέλα η αύξηση ήταν ελάχιστη έως μηδαμινή. Για την κατηγορία Spk_F1 είχαμε μια κάποια σημαντική αύξηση των αποτελεσμάτων μέχρι τη χρήση του μοντέλου MMIS+MAP+SSA+USA στη συνέχεια όμως τα αποτελέσματα και σε αυτή την περίπτωση έμειναν σταθερά. Για την κατηγορία Spk_F2 σχεδόν για κάθε μοντέλο που υλοποιήσαμε είχαμε μια αρκετά καλή απόδοση τον αποτελεσμάτων, όμως και σε αυτή την περίπτωση κατά την χρήση του μοντέλου MMIS+MAP+(SSA+USA)+USA δεν είχαμε σημαντική αύξηση των αποτελεσμάτων. Οι διαφορετικές αυτές διακυμάνσεις στα πειράματα μας είναι αναμενόμενες διότι όπως ήδη γνωρίζουμε η Unsupervised μέθοδος χρησιμοποίει ως προτάσεις προσαρμογής της προτάσεις ελέγχου, από τη στιγμή λοιπόν που οι προτάσεις προσαρμογής έχουν αναγνωριστεί με αρκετά καλή ακρίβεια(πχ.Spk_F0) οι αποδόσεις των μοντέλων μας έχουν πολύ μικρή άνοδο. Από την άλλη όταν τα αποτελέσματα έχουν περιθώρια βελτίωσης(πχ. Spk_F2) η αύξηση είναι μεγαλύτερη.

Στη συνέχεια θα παρουσιάσουμε τα αποτελέσματα για κάθε ομιλητή για όλα τα Unsupervised μοντέλα με μεταβλητό αριθμό MLLR μετασχηματισμών.

Ομιλητές	**Προτάσεις**				
	Test	**MMIS+MAP**	**MMIS+MAP +USA** (Var-mllr)	**MMIS+MAP +SSA+USA** (Var-mllr)	**MMIS+MAP +(SSA+USA) +USA** (Var-mllr)
Γ. Παπανδρέου	1	61.29	64.52	70.97	70.97
Γαβρά	227	83.63	84.13	84.32	84.30
Κακαβάς	9	69.93	71.24	69.93	69.93
Κανδύλη	15	74.01	76.32	79.28	79.28
Καψής	22	60.64	62.39	64.43	63.56
Κ. Καραμανλής	21	82.77	85.47	86.49	86.49
Λιναρδάτου	21	76.90	79.52	81.43	81.19
Λοβέρδος	17	60.09	61.47	64.22	63.76
Νιότης	18	74.78	75.66	74.49	74.19
Παπαηλιόπουλος	15	70.65	73.23	74.52	74.84
Πρετεντέρης	31	60.23	62.52	62.09	62.52
Σπυράκη	13	78.02	80.59	82.42	82.42
Τρέμη	64	79.31	79.61	80.00	80.10
Χαρίτος	2	80.70	80.70	82.46	82.46
Χούκλη	306	75.70	77.15	77.51	77.55

*Πίνακας 4.16 : Αποτελέσματα για **Test Spk_F0** μετά από Unsupervised Speaker MLLR Adaptation μεταβλητών MLLR μετασχηματισμών για κάθε ομιλητή.*

Ομιλητές	Προτάσεις				
	Test	MMIS+MAP	MMIS+MAP +USA (Var-mllr)	MMIS+MAP +SSA+USA (Var-mllr)	MMIS+MAP +(SSA+USA) +USA (Var-mllr)
Γ. Παπανδρέου	30	76.71	79.22	79.45	79.45
Γαβρά	79	71.07	73.82	73.65	74.16
Κακαβάς	8	85.11	87.94	90.78	90.78
Κανδύλη	12	81.78	80.89	80.00	80.00
Καψής	16	68.53	74.50	71.71	71.31
Κ. Καραμανλής	52	72.48	76.42	80.50	80.19
Νιότης	9	83.78	83.78	83.24	84.32
Πρετεντέρης	10	53.33	54.29	58.10	58.10
Σπυράκη	29	80.78	81.42	81.91	81.26
Τρέμη	12	75.13	75.13	76.17	76.17
Χαρίτος	23	64.64	66.74	69.79	70.02
Χούκλη	36	63.59	63.85	64.62	64.87

*Πίνακας 4.17 : Αποτελέσματα για **Test Spk_F1** μετά από Unsupervised Speaker MLLR Adaptation μεταβλητών MLLR μετασχηματισμών για κάθε ομιλητή.*

Ομιλητές	Προτάσεις				
	Test	MMIS+MAP	MMIS+MAP +USA (Var-mllr)	MMIS+MAP +SSA+USA (Var-mllr)	MMIS+MAP +(SSA+USA) +USA (Var-mllr)
Κακαβάς	17	58.90	58.90	62.27	61.96
Χαρίτος	29	46.55	53.64	59.64	60.36

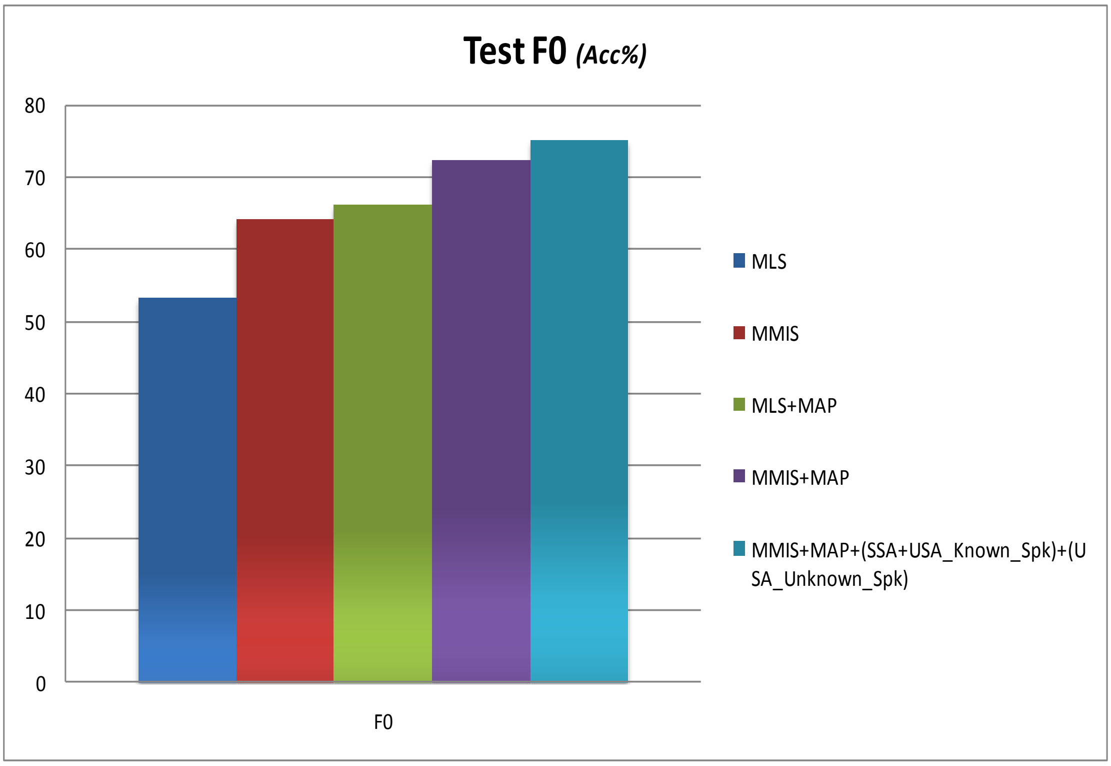
*Πίνακας 4.18 : Αποτελέσματα για **Test Spk_F2** μετά από Unsupervised Speaker MLLR Adaptation μεταβλητών MLLR μετασχηματισμών για κάθε ομιλητή.*

Στη συνέχεια θα παρουσιάσουμε τα αποτελέσματα του τελικού μας πειράματος MMIS+MAP+(SSA+USA_Known_Spk)+(USA_Unknown_Spk) το οποίο περιγράφεται αναλυτικά στη αρχή του κεφαλαίου. Η αναγνώριση για το τελικό μας μοντέλο έγινε στο σύνολο των δεδομένων έλεγχου Test F0, Test F1 και Test F2.

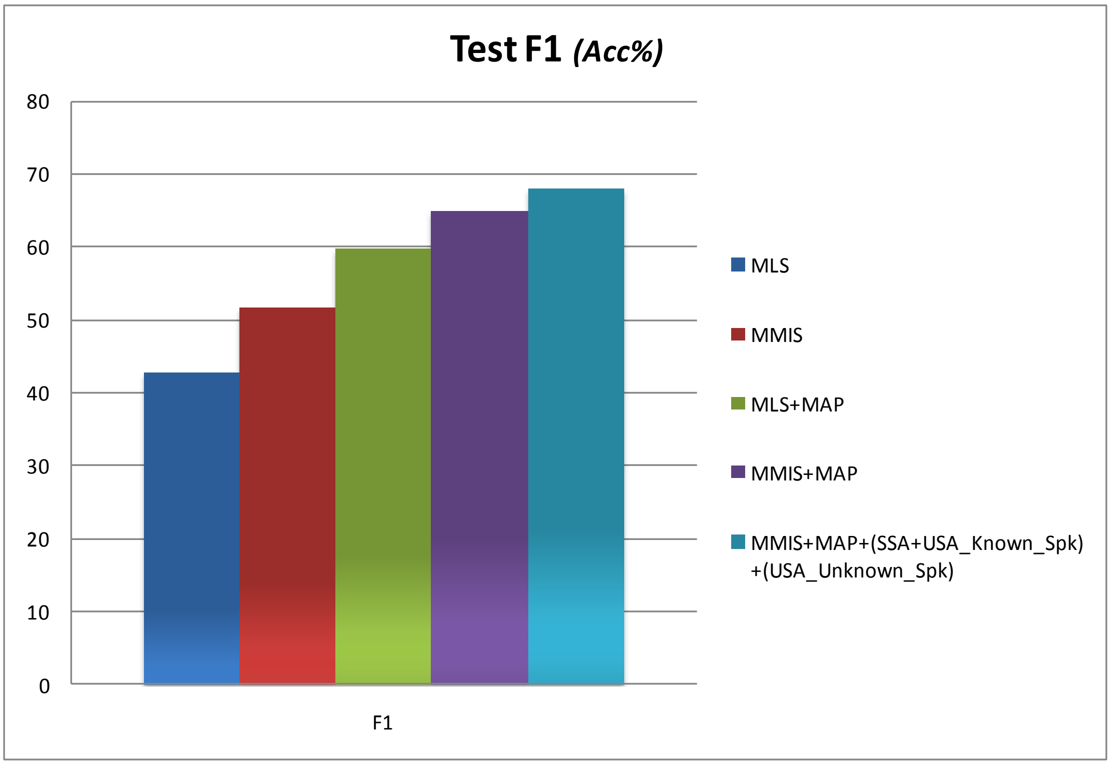
Models	Acc (%)		
	F0	**F1**	**F2**
MMIS+MAP	72.42	64.92	47.89
MMIS+MAP+(SSA+USA_Known_Spk)+(USA_Unknown_Spk)	75.07	67.94	57.59

*Πίνακας 4.19 : Αποτελέσματα για το σύνολο των **Test sets** μετά από Unsupervised Speaker MLLR Adaptation σε τμήματα ομιλητών.*

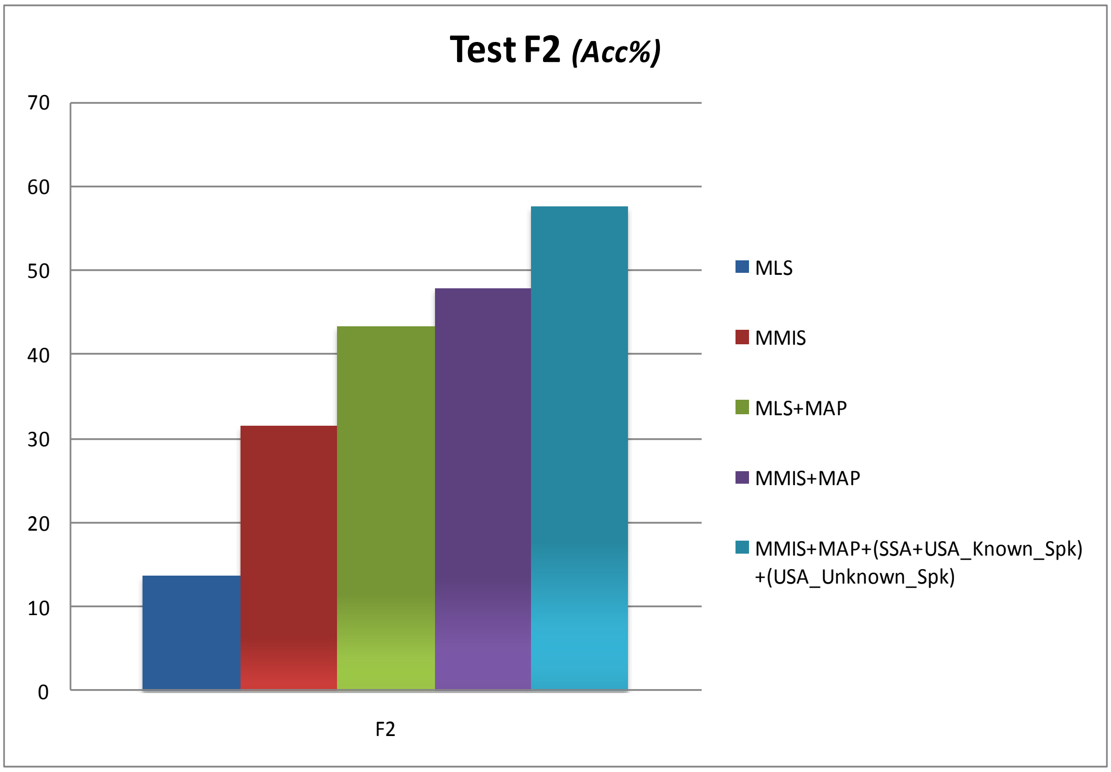
Όπως παρατηρούμε από τον *Πίνακα 4.19* η χρήση της Unsupervised τεχνικής σε τμήματα ομιλίας του ίδιου ομιλητή έχει αρκετά καλά αποτελέσματα. Συγκεκριμένα, παρατηρούμε αύξηση και στις τρεις κατηγορίες από 3% έως 10% περίπου. Αυτό μας δείχνει πως η χρήση του *Speaker Adaptation* μπορεί να επιφέρει μια αρκετά καλή βελτίωση της απόδοσης των αρχικών SI μοντέλων. Πριν όμως βγάλουμε τα τελικά μας συμπεράσματα θα ήταν χρήσιμο να δούμε μια συνολική εικόνα της ακρίβειας των ακουστικών μας μοντέλων σε κάθε στάδιο της εφαρμογής του Adaptation. Παρακάτω παρουσιάζεται η πρόοδος των αποτελεσμάτων μας για τις τρεις κατηγορίες έλεγχου Test F0, Test F1 και Test F2.



*Σχήμα 4.6 : Γραφική απεικόνιση της προόδου αποτελεσμάτων για **Test F0***



Σχήμα 4.7 : Γραφική απεικόνιση της προόδου αποτελεσμάτων για ***Test F1***



Σχήμα 4.8 : Γραφική απεικόνιση της προόδου αποτελεσμάτων για ***Test F2***

Models	Acc (%)		
	F0	**F1**	**F2**
MLS	53.15	42.58	13.72
MMIS	64.25	51.60	31.42
MLS+MAP	66.20	59.68	43.24
MMIS+MAP	72.42	64.92	47.89
MMIS+MAP+(SSA+USA_Known_Spk)+(USA_Unknown_Spk)	75.07	67.94	57.59

Πίνακας 4.20 : Παρουσίαση της προόδου των αποτελεσμάτων για το σύνολο των ***Test sets***

Ο παραπάνω πίνακας μας δίνει μια συνολική εικόνα για την πρόοδο των αποτελεσμάτων μας. Όπως αναφέραμε και στο προηγούμενο κεφάλαιο τα αποτελέσματα των βασικών μας μοντέλων(MLS,MMIS) είναι αρκετά χαμηλά και ο λόγος είναι ότι τα μοντέλα αυτά είναι εκπαιδευμένα για την αναγνώριση μιας μεγάλης γκάμας ομιλητών. Ο αρχικός μας σκοπός ήταν να καταφέρουμε να προσαρμόσουμε τα MLS και MMIS μοντέλα με σκοπό να αναγνωρίζουν με όσο το δυνατόν καλύτερη ακρίβεια δεδομένα από τηλεοπτικές ειδήσεις. Κατά το πρώτο στάδιο της εφαρμογής του Adaptation στο επίπεδο της συνθήκης ομιλίας με βάση την τεχνική της MAP προσαρμογής, έχουμε μια αρκετά μεγάλη αύξηση των αποτελεσμάτων και για τα δυο βασικά μας μοντέλα. Συγκεκριμένα, για το MMIS+MAP μοντέλο στην κατηγορία F0 είχαμε μια αύξηση περίπου κατά 8%, ενώ στις κατηγορίες F1 και F2 είχαμε μεγαλύτερη αύξηση της τάξης του 13% και 16% αντίστοιχα. Τα MAP προσαρμοσμένα μοντέλα μας φτάσανε στην καλύτερη απόδοση τους περίπου στο 72%. Στη συνέχεια κατά το δεύτερο στάδιο με χρήση του Speaker Adaptation(Unsupervised για τους άγνωστους ομιλητές Supervised+Unsupervised για τους γνωστούς) επιτύχαμε μια περαιτέρω αύξηση από 3% έως 10% περίπου. Καταφέραμε έτσι την καλύτερη απόδοση για το σύστημα μας, η οποία έφτασε κοντά στο 75%. Συγκεκριμένα, έπειτα από τα δυο στάδια της προσαρμογής κατορθώσαμε να αυξήσουμε την ακρίβεια του αρχικού μοντέλου MMIS, για της κατηγορία F0 κατά περίπου 11%, για την κατηγορία F1 κατά 16% και για την κατηγορία F2 κατά 26% περίπου. Η αύξηση στις παραπάνω κατηγορίες είναι αρκετά ικανοποιητική παρ' όλα αυτά υπάρχει ακόμα δυνατότητα βελτιστοποίησης τους μιας και υπάρχουν και άλλοι παράγοντες, όπως το γλωσσικό μοντέλο, που επηρεάζουν σημαντικά την απόδοση ενός συστήματος αναγνώρισης φωνής. Στο επόμενο κεφάλαιο θα δώσουμε μια συνοπτική περιγραφή για τις μελλοντικές επεκτάσεις με σκοπό την καλύτερη απόδοση του συστήματος μας.

ΚΕΦΑΛΑΙΟ 5

Ανακεφαλαίωση και Μελλοντικές Επεκτάσεις

Στα πλαίσια της παρούσας διπλωματικής εργασίας μελετήθηκε αναλυτικά η μέθοδος υλοποίησης ενός συστήματος αυτόματης απομαγνητοφώνησης και αναγνώρισης ακουστικών σημάτων ηχογραφημένα από τηλεοπτικές εκπομπές δελτίων ειδήσεων. Στόχος ήταν η υλοποίηση ακουστικών μοντέλων προσαρμοσμένων σε δεδομένα από τηλεοπτικές ειδήσεις με σκοπό την βελτίωση της επίδοσης των βασικών μοντέλων όπου η εκπαίδευση τους είχε γίνει με δεδομένα από τη Λογοτυπογραφία.

Η προσαρμογή των ακουστικών μοντέλων έγινε σε δυο βασικά στάδια. Αρχικά εφαρμόστηκε η μέθοδος του MAP Adaptation η οποία επέφερε μια ικανοποιητική βελτίωση της ακρίβειας από 8% έως 16% για την καλύτερη περίπτωση. Στη συνέχεια εκτελέστηκαν μια σειρά πειραμάτων για την υλοποίηση της προσαρμογής στον ομιλητή πάνω στα ήδη MAP προσαρμοσμένα μοντέλα. Αυτό επέφερε μια επιπλέον βελτίωση στην απόδοση του συστήματος κατά 3% έως 10%. Κατά την εφαρμογή της προσαρμογής στον ομιλητή τα καλύτερα αποτελέσματα επετεύχθησαν με χρήση του συνδυασμού του Supervised Adaptation και στη συνέχεια Unsupervised Adaptation. Ύστερα από τον συνδυασμό των δυο παραπάνω μεθόδων τελικά καταφέραμε μια συνολική βελτίωση του αρχικού συστήματος μας από 11% έως 26%. Να αναφέρουμε στο σημείο αυτό ότι εφαρμόσαμε MAP Adaptation για τη προσαρμογή στο επίπεδο συνθήκης ομιλίας και MLLR στο επίπεδο του ομιλητή. Ύστερα από πειραματικές δοκιμές και βιβλιογραφικές πληροφορίες αποδείχτηκε ότι για την προσαρμογή στις συνθήκες ομιλίας η MAP προσαρμογή είχε καλύτερα αποτελέσματα από την MLLR και ο λόγος είναι ότι σε μεγάλο αριθμό δεδομένων η MAP τεχνική αποδίδει καλύτερα. Από την άλλη, επειδή κατά την προσαρμογή στον ομιλητή, όπως είναι φυσικό, ο αριθμός των προτάσεων προσαρμογής είναι αρκετά μικρός, η MLLR τεχνική αποδίδει αρκετά καλύτερα από την MAP.

Για την επέκταση αυτού του συστήματος με σκοπό πάντα την αύξηση της ακρίβειας των ακουστικών μοντέλων θα μπορούσαν να γίνουν μια σειρά από βελτιώσεις σε κάποιους βασικούς παράγοντες. Αρχικά, θα μπορούσε να γίνει μια αναβάθμιση στο γλωσσικό μοντέλο αυξάνοντας το μέγεθος του κειμένου που έχει εκπαιδευτεί και μια επιπλέον αναβάθμιση στο λεξικό προφορών.

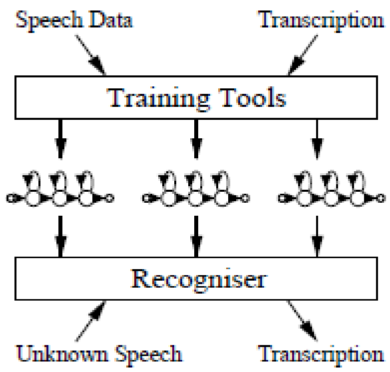
Ένας πολύ σημαντικός παράγοντας, ο οποίος θα μπορούσε να βελτιωθεί σημαντικά, είναι η απομαγνητοφώνηση των ακουστικών σημάτων. Από τις 50 ώρες δεδομένων, όπως έχουμε αναφέρει και σε προηγούμενα κεφάλαια, στις 30 ώρες η απομαγνητοφώνηση είχε γίνει με τρόπο που δεν συμφωνούσε με τις υπόλοιπες 20. Για να γίνουμε πιο συγκεκριμένοι στις 30 ώρες η κατάτμηση του λόγου είχε γίνει με βάση την αναπνοή που παίρνει ο ομιλητής με αποτέλεσμα στις περισσότερες περιπτώσεις να χάνεται το νόημα των προτάσεων. Επίσης, ο χαρακτηρισμός της ποιότητας της ομιλίας δεν ήταν διαθέσιμος. Στα πλαίσια αυτής της διπλωματικής διορθώσαμε τις 10 από τις 30 ώρες κατά συνέπεια η διόρθωση και των υπολοίπων 20 ωρών θα επέφερε μια αρκετά καλή βελτίωση. Ένα ακόμα σημαντικό θέμα κατά την απομαγνητοφώνηση είναι ο ανθρώπινος παράγοντας. Επειδή η αποτύπωση σε λέξεις των ακουστικών σημάτων έγινε χειροκίνητα από προπτυχιακούς φοιτητές, όπως ήταν φυσικό, υπήρχαν αρκετά λάθη στην αποτύπωση των λέξεων. Εν κατακλείδι θα μπορούσε να γίνει ένας έλεγχος σε όλα τα δεδομένα και να διορθωθούν τα τυχόν λάθη.

Τέλος, θα μπορούσαν να εφαρμοστούν διάφορες Adaptive τεχνικές εκπαίδευσης όπως η HLDA και η Semi-Tied οι οποίες υποστηρίζονται πλήρως από το HTK, καθώς και ο συνδυασμός τους με τις τεχνικές που υλοποιήθηκαν στα πλαίσια της παρούσας διπλωματικής εργασίας.

Παράρτημα Α

HTK (Hidden Markov Model Toolkit)

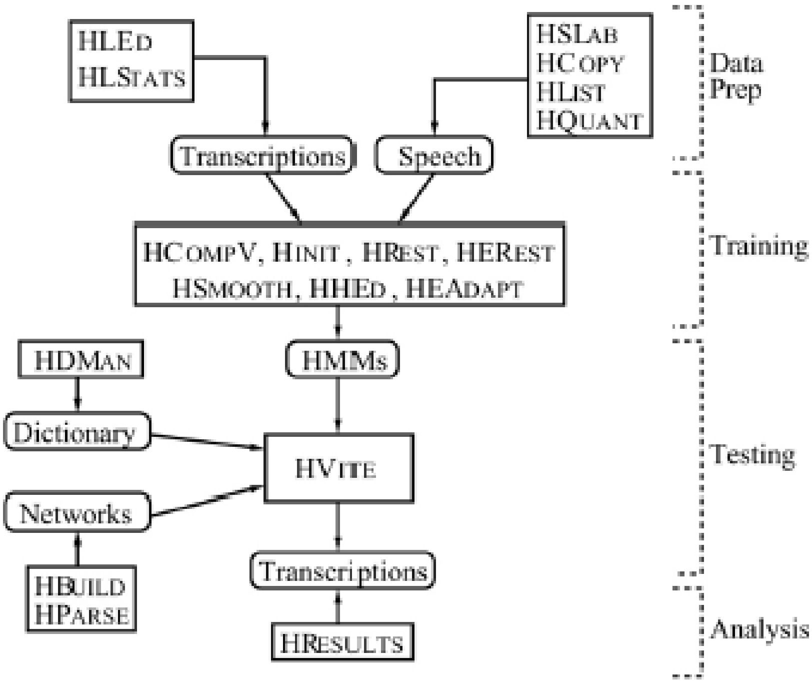
Το HTK είναι μια συλλογή προγραμματιστικών εργαλείων για τη δημιουργία και το χειρισμό των Hidden Markov Models(HMMs). Το HTK ενδείκνυται για ερεύνα πάνω στην αναγνώριση φωνής μιας και μπορεί να μοντελοποιήσει HMMs κάτω από οποιεσδήποτε συνθήκες. Όπως φαίνεται και στο *Σχήμα Π-1* υπάρχουν δυο βασικά στάδια κατά τη λειτουργία του HTK. Στο πρώτο στάδιο γίνεται χρήση των εργαλείων εκπαίδευσης για την εκτίμηση των παραμέτρων των HMMs με βάση της προτάσεις εκπαίδευσης. Στη συνέχεια άγνωστες προτάσεις αναγνωρίζονται κάνοντας χρήση των εργαλείων αναγνώρισης του HTK.



Σχήμα ΠΑ-1: Στάδια λειτουργίας του HTK

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.

Το HTK μας δίνει τη δυνατότητα να υλοποιήσουμε ένα σύστημα αναγνώρισης ομιλίας από το μηδέν χρησιμοποιώντας μια σειρά εργαλείων για την επεξεργασία των δεδομένων*(Data Preparation),* για την εκπαίδευση(*Training)* και για την αξιολόγηση του συστήματος *(Evaluation).* Στο *Σχήμα Π-2* δίνεται μια συνοπτική παρουσίαση των εργαλείων που υλοποιούν τα τρία παραπάνω στάδια.



Σχήμα ΠΑ-2 Λειτουργία του HTK

Ο τρόπος εφαρμογής των εργαλείων αυτών για τη δημιουργία ενός συστήματος αναγνώρισης φωνής έχει παρουσιαστεί κατά κόρον σε παλαιότερες διπλωματικές εργασίες που ασχολήθηκαν με την αναγνώριση ομιλίας. Για το λόγο αυτό θεωρήθηκε σωστό να μην ασχοληθούμε σε βάθος με τον τρόπο λειτουργίας τους. Παρ' όλα αυτά θεωρήθηκε χρήσιμο να παρουσιαστεί ο τρόπος εφαρμογής των εργαλείων του HTK για την υλοποίηση του βασικού κορμού της παρούσας διπλωματικής εργασίας που δεν ήταν άλλος από την μέθοδο της προσαρμογής. Παρακάτω θα δώσουμε μια συνοπτική παρουσίαση για τον τρόπο που υλοποιείται το MAP και το MLLR Adaptation με τα εργαλεία του HTK.

Η μέθοδος της προσαρμογής με τα εργαλεία του HTK

Όπως έχουμε αναφέρει αρκετές φόρες στα πλαίσια αυτής της διπλωματικής εργασίας, υπάρχουν συστήματα αναγνώρισης ομιλίας τα οποία είναι ανεξάρτητα των ομιλητών *(Speaker Independent, SI)* και έχουν μια μέτρια απόδοση για μια μεγάλη γκάμα ομιλητών. Όμως υπάρχουν περιπτώσεις όπου θέλουμε το σύστημα μας να είναι προσαρμοσμένο σε συγκεκριμένη ομάδα ομιλητών *(Speaker Depended,SD)* έτσι ώστε να έχουμε τα καλύτερα δυνατά αποτελέσματα αναγνώρισης.

Το HTK μας δίνει τη δυνατότητα μέσω διαφόρων εργαλείων να υλοποιούμε SD συστήματα με την μέθοδο της προσαρμογής στον ομιλητή αλλά και συστήματα προσαρμοσμένα σε διαφορές άλλες συνθήκες όπως το περιεχόμενο της ομιλίας και το περιβάλλον στο οποίο παράγεται η ομιλία. Η μέθοδος της προσαρμογής εφαρμόζεται με την βοήθεια της εντολής *HERest* είτε κάνοντας χρήση γραμμικών μετασχηματισμών μεγιστοποίησης της πιθανοφάνειας (MLLR,CMLLR), είτε μεγιστοποίησης της a-posteriori πιθανότητας (MAP).

Στη συνέχεια θα δώσουμε μια περιγραφή για τον τρόπο υλοποίησης της τεχνικών αυτών.

MLLR Adaptation

Αφού έχουμε στην διάθεση μας το SI μοντέλο και τα δεδομένα προσαρμογής σε μορφή αναγνωρίσιμη από το HTK στη συνεχεία ακλουθούμε τα παρακάτω βήματα:

- **Βήμα 1 :** Με τη βοήθεια της εντολής *HVite* παράγουμε τις λίστες των προτάσεων προσαρμογής σε επίπεδο φωνημάτων. Στο στάδιο αυτό η *HVite* εφαρμόζει και την τεχνική της εξαναγκασμένης ευθυγράμμισης *(force Alignment)* για την αποφυγή των πολλαπλών προφορών των φωνημάτων.
- **Βήμα 2 :** Στο βήμα αυτό γίνεται χρήση της εντολής HHed για την παραγωγή ενός *regression class tree* μαζί με ένα σύνολο από *baseclasses* από το βασικό SI μοντέλο. Στο σημείο αυτό ορίζεται ο αριθμός των MLLR μετασχηματισμών που θα παραχθούν.
- **Βήμα 3 :** Στη συνέχεια με χρήση της *HERest* και διαφόρων απαραίτητων παραμέτρων υπολογίζεται και αποθηκεύεται ένας global μετασχηματισμός .

- **Βήμα 4** **:** Στο τελευταίο στάδιο της προσαρμογής και πάλι με χρήση της *HERest* ο global μετασχηματισμός που παρήχθει στο βήμα 3 τίθεται ως είσοδος με σκοπό τον μετασχηματισμό των παραμέτρων του αρχικού μοντέλου.
- **Βήμα 5** **:** Τέλος, στο βήμα αυτό γίνεται η αξιολόγηση του προσαρμοσμένου μοντέλου με χρήση της *HDecode.* Απαραίτητη προϋπόθεση για την σωστή λειτουργιά της *HDecode* για τα προσαρμοσμένα μοντέλα είναι η χρήση των παραμέτρων –J και -m για να μπορέσει η εντολή να λάβει υπ' όψιν της τους μετασχηματισμούς που εξαγάγαμε παραπάνω.

MAP Adaptation

- **Βήμα 1 :** Όπως και στο βήμα 1 της MLLR τεχνικής κάνουμε χρήση της *HVite* για παράγουμε τις λίστες των προτάσεων προσαρμογής σε επίπεδο φωνημάτων.
- **Βήμα 2 :** Στη συνέχεια με τη βοήθεια της *HERest* και διαφόρων απαραίτητων συνιστωσών κάνουμε την εκτίμηση των παραμέτρων του βασικού μοντέλου και αποθηκεύουμε το νέο προσαρμοσμένο μοντέλο.
- **Βήμα 3 :** Τέλος, με χρήση της *HDecode* γίνεται η αξιολόγηση του προσαρμοσμένου μοντέλου.

Ύστερα από την εξαγωγή των αναγνωρισμένων λέξεων το HTK δίνει τη δυνατότητα με χρήση της εντολής *HResults* του υπολογισμού της ακρίβειας (*Accuracy)* της αναγνώρισης.

Παράρτημα Β

ΠΑΡΑΔΕΙΓΜΑΤΑ

Καθ’ όλη την διάρκεια της παρούσας διπλωματικής εργασίας παρουσιαστήκαν τα αποτέλεσμα καθώς και η βελτίωση της ακρίβειας των μοντέλων με βάση την ποσότητα Accuracy (βλέπε κεφάλαιο 3.5). Θα ήταν όμως ενδιαφέρον να παρουσιαστεί η βελτίωση σε επίπεδο προτάσεων. Στο παράρτημα αυτό θα παρουσιάσουμε μια σειρά από παραδείγματα για την απόδοση των πειραμάτων μας σε επίπεδο προτάσεων. Στα πλαίσια τις παρουσίασης των παραδειγμάτων αυτών θα παρουσιαστούν τα αποτελέσματα των αναγνωριστών όπως ακριβώς εξάγονται από την εντολή HDEcode του εργαλείου HTK.

Για να μπορεί να γίνει καθολική σύγκριση των αποτελεσμάτων επιλέχτηκαν τυχαία τέσσερεις προτάσεις από τα δεδομένα ελέγχου. Παρακάτω παρουσιάζονται οι τέσσερεις αυτές προτάσεις στην αρχική τους μορφή:

```
070428MEGAED2000_265.lab"
η
απόλυτα
νόμιμη
αγοροπωλησία
οικοπέδων
στη
μύκονο
το
χίλια
εννιακόσια
ενενήντα
εννιά
```

```
070522NETED2100_119.lab"
αυτή
τη
στιγμή
βρισκόμαστε
στην
όλυμπο
```

```
070428MEGAED2000_004.lab"
ο
κύριος
παπαδόπουλος
κατηγορείται
για
εμπλοκή
στο
σκάνδαλο
του
χρηματιστηρίου
και
στα
παιχνίδια
με
τις
μετοχές
φούσκες
```

```
070428MEGAED2000_011.lab"
και
ποια
είναι
η
αφετηρία
αυτής
της
ιστορίας
είναι
οι
αποκαλύψεις
για
τα
έργα
και
τις
ημέρες
του
κυρίου
γιαννακόπουλου
```

Πίνακας ΠΒ-1 : Αρχική μορφή προτάσεων ελέγχου

Παρακάτω παρουσιάζονται τα αποτελέσματα της αναγνώρισης για το MLS μοντέλο πριν και μετά την MAP προσαρμογή:

Αναγνώριση MLS μοντέλου	**Αναγνώριση MLS+MAP μοντέλου**
"*/070428MEGAED2000_265.rec" 1100000 10300000 ήταν -7702.135254 10300000 15200000 όμηροι -3745.005859 15200000 17600000 από -1952.811523 17600000 22800000 απόλυση -4443.944336 22800000 23500000 ή -489.019531 23500000 28700000 αν -3702.037109 28700000 29600000 οι -791.642578 29600000 34700000 κοπέλες -4261.072266 34700000 36300000 με -1229.890625 36300000 40600000 εικονικά -3532.025391 40600000 43300000 δεν -1983.908203 43300000 46400000 άκουσε -2731.332031 46400000 47700000 με -926.265625 47700000 52800000 δέδε -3794.105469	"*/070428MEGAED2000_265.rec" 1200000 2900000 την -1491.431152 2900000 8900000 απόλυτα -4907.105957 8900000 16100000 νόμιμη -5291.013184 16100000 28700000 αγοραπωλησία -9260.219727 28700000 33700000 οικοπέδου -4000.738281 33700000 38800000 γυναικών -3949.248047 38800000 39700000 το -769.611328 39700000 41800000 χίλια -1617.289062 41800000 46400000 εννιακόσια -3509.908203 46400000 49600000 ενενήντα -2214.878906 49600000 52800000 εννιά -2253.343750
"*/070522NETED2100_119.rec" 800000 3200000 και -2161.328613 3200000 7700000 δείχνει -3667.033936 7700000 8500000 να -620.285645 8500000 11100000 έχει -2183.305664 11100000 16200000 μακρινά -4367.431641 16200000 21000000 λοιπόν -4027.825195	"*/070522NETED2100_119.rec" 300000 3300000 αυτή -2619.474121 3300000 4300000 τη -808.796143 4300000 7700000 στιγμή -2694.072021 7700000 13800000 βρισκόμαστε -4765.203613 13800000 15700000 στον -1557.316406 15700000 21000000 ολυμπο -4256.362305
"*070428MEGAED2000_004.rec" 1700000 6100000 οποίος -3611.835693 6100000 12400000 παπαδόπουλος -5006.945801 12400000 19400000 κατηγορητήρια -5668.511719 19400000 22800000 νεκρή -2889.403320 22800000 24100000 στο -1083.542969 24100000 28600000 σκάνδαλο -3499.015625 28600000 29600000 του -883.300781 29600000 35800000 χρηματιστηρίου -4904.027344 35800000 37300000 είτε -1255.142578 37300000 38600000 στα -1229.675781 38600000 44400000 παιχνίδια -4292.792969 44400000 45700000 με -1049.917969 45700000 47200000 την -1316.109375 47200000 51000000 επίθεση -3135.128906 51000000 52600000 που -1347.746094 52600000 54100000 στις -1296.492188	"*070428MEGAED2000_004.rec" 1900000 3800000 ο -1550.003540 3800000 6000000 κύριος -1765.050293 6000000 12200000 παπαδόπουλος -4699.741699 12200000 17500000 κατηγορείται -4133.590820 17500000 19600000 για -1581.698242 19600000 23500000 πράξεις -3118.181641 23500000 24100000 ο -523.423828 24100000 28600000 σκάνδαλο -3438.121094 28600000 29600000 του -905.166016 29600000 36200000 χρηματιστηρίου - 5058.304688 36200000 37300000 και -894.562500 37300000 38600000 στα -1189.046875 38600000 44600000 παιχνίδια -4394.449219 44600000 45300000 με -463.390625 45300000 46700000 τις -1131.347656 46700000 51000000 μετοχές -3481.507812 51000000 54600000 φούσκες -2898.718750
"*070428MEGAED2000_011.rec" 1700000 5000000 βοιωτια -2769.113770 5000000 6100000 η -771.467041 6100000 8000000 μία -1431.252930 8000000 10500000 θέτει -1938.687012 10500000 13000000 δι -1855.006836 13000000 15500000 αυτής -1889.680664 15500000 16700000 της -939.208984 16700000 24600000 θεωρίας -5499.467773 24600000 27300000 είναι -2015.886719 27300000 33100000 διαδοχικά -4508.673828 33100000 34000000 οι -668.328125 34000000 36900000 αιτήσεις -2209.662109 36900000 38400000 για -1232.443359 38400000 40000000 τα -1158.226562 40000000 42600000 έργα -1932.718750 42600000 43700000 και -868.601562 43700000 45000000 τις -1009.652344 45000000 48600000 ημέρες -2813.601562 48600000 49900000 του -1082.156250 49900000 51900000 κυρίου -1520.742188 51900000 54400000 νανά -2072.613281 54400000 56200000 κω -1454.585938	"*070428MEGAED2000_011.rec" 1700000 5200000 δεξιά -2727.877930 5200000 6900000 είναι -1206.439941 6900000 7500000 η -447.593262 7500000 13300000 αφετηρία -4155.712891 13300000 15500000 αυτής -1617.110352 15500000 17000000 της -1073.648438 17000000 24600000 ιστορίας -4980.465820 24600000 28200000 είναι -2539.207031 28200000 29500000 η -908.458984 29500000 36900000 αποκαλύψεις -5296.044922 36900000 38500000 για -1165.654297 38500000 39800000 τα -874.882812 39800000 42600000 έργα -1933.425781 42600000 43700000 και -830.382812 43700000 45000000 τις -938.970703 45000000 48600000 ημέρες -2633.640625 48600000 49900000 του -1035.789062 49900000 52100000 κυρίου -1616.343750 52100000 56100000 γιαννάκου -3142.816406

Πίνακας ΠΒ-2 : Προτάσεις αναγνώρισης από το MLS και MLS+MAP μοντέλο

Παρακάτω παρουσιάζονται τα αποτελέσματα της αναγνώρισης για το MMIS μοντέλο πριν και μετά την MAP προσαρμογή:

Αναγνώριση MMIS μοντέλου	Αναγνώριση MMIS+MAP μοντέλου
070428MEGAED2000_265.rec" 1200000 4600000 η -2941.807861 4600000 8800000 ταινία -3550.817383 8800000 16100000 νόμιμη -5667.319336 16100000 28700000 αγοραπωλησία -9628.739258 28700000 35100000 οικοπέδου -5316.400391 35100000 38800000 ειδικών -2814.738281 38800000 40300000 των -1283.265625 40300000 45200000 φυλακών -3875.513672 45200000 48700000 στενεύει -2695.460938 48700000 52800000 βέβαια -3051.636719	070428MEGAED2000_265.rec" 1200000 2900000 την -1498.485840 2900000 8900000 απόλυτα -4912.119141 8900000 16100000 νόμιμη -5293.798828 16100000 28700000 αγοραπωλησία -9262.062500 28700000 34100000 οικόπεδα -4335.650391 34100000 36100000 στην -1549.652344 36100000 38800000 αιξωνή -2038.500000 38800000 39700000 το -798.058594 39700000 41800000 χίλια -1631.220703 41800000 46400000 εννιακόσια -3523.234375 46400000 49600000 ενενήντα -2213.390625 49600000 52700000 εννιά -2182.289062
070522NETED2100_119.rec" 200000 2200000 αν -1851.612915 2200000 3200000 και -846.065430 3200000 4300000 δεν -869.929443 4300000 8600000 είναι -3528.847168 8600000 13800000 γράμμα -4341.831055 13800000 16900000 κενό -2622.939453 16900000 21800000 λοιπόν -4112.798828	070522NETED2100_119.rec" 300000 3300000 αυτή -2618.468750 3300000 4300000 τη -808.766602 4300000 7700000 στιγμή -2697.986816 7700000 13800000 βρισκόμαστε -4762.321777 13800000 15700000 στον -1560.429688 15700000 21000000 ολυμπο -4252.014648
070428MEGAED2000_004.rec" 1700000 3800000 ο -1655.936401 3800000 6100000 κύριος -1932.078125 6100000 12400000 παπαδόπουλος -4955.271484 12400000 17500000 κατηγορείται -4151.908203 17500000 17800000 η -265.545898 17800000 21900000 γέλινεκ -3330.874023 21900000 24900000 ίσως -2543.544922 24900000 28600000 κάβανο -2933.359375 28600000 29800000 του -1080.533203 29800000 35800000 χρηματιστηρίου -4777.021484 35800000 37300000 είτε -1268.308594 37300000 38600000 στα -1228.152344 38600000 44400000 παιχνίδια -4300.769531 44400000 45600000 με -965.195312 45600000 46700000 τις -939.777344 46700000 50900000 μετοχές -3463.351562 50900000 52300000 που -1175.207031 52300000 55000000 στις -2287.886719	070428MEGAED2000_004.rec" 1900000 3800000 ο -1546.267456 3800000 6000000 κύριος -1746.380859 6000000 12100000 παπαδόπουλος -4649.003418 12100000 17500000 κατηγορείται -4184.915039 17500000 23500000 γιαννάκης -4716.916992 23500000 24100000 ο -502.628906 24100000 28600000 σκάνδαλο -3444.263672 28600000 29900000 του -1172.574219 29900000 36200000 χρηματιστηρίου -4791.378906 36200000 37300000 και -908.585938 37300000 38600000 στα -1189.654297 38600000 44600000 παιχνίδια -4390.511719 44600000 45300000 με -475.691406 45300000 46700000 τις -1133.906250 46700000 51000000 μετοχές -3472.882812 51000000 54500000 φούσκες -2848.136719
070428MEGAED2000_011.rec" 1700000 2700000 και -858.863037 2700000 5100000 ποια -2036.743530 5100000 6800000 είναι -1186.191650 6800000 7400000 η -437.255371 7400000 13300000 αφετηρία -4424.820801 13300000 15500000 αυτής -1725.451172 15500000 17400000 της -1443.903320 17400000 24600000 ιστορίας -4989.178711 24600000 28200000 είναι -2653.980469 28200000 32000000 μια -2929.751953 32000000 34500000 καλή -1896.265625 34500000 36900000 της -1829.597656 36900000 38400000 για -1248.363281 38400000 39800000 τα -1029.568359 39800000 42600000 έργα -2048.769531 42600000 43700000 και -868.234375 43700000 44900000 τις -925.765625 44900000 48600000 ημέρες -2904.210938 48600000 49900000 του -1067.226562 49900000 52100000 κυρίου -1696.621094 52100000 56200000 γιαννάκου -3314.394531	070428MEGAED2000_011.rec" 1700000 5100000 δεξιά -2670.676270 5100000 6800000 είναι -1185.978027 6800000 7500000 η -516.722168 7500000 13300000 αφετηρία -4164.697754 13300000 15500000 αυτής -1617.415039 15500000 17000000 της -1074.101562 17000000 24600000 ιστορίας -4984.381836 24600000 28200000 είναι -2548.232422 28200000 29500000 η -899.099609 29500000 36800000 αποκαλύψεις -5212.501953 36800000 38500000 για -1244.939453 38500000 39800000 τα -875.177734 39800000 42500000 έργα -1850.398438 42500000 43700000 και -908.271484 43700000 45000000 τις -939.808594 45000000 48600000 ημέρες -2636.923828 48600000 49900000 του -1033.863281 49900000 52100000 κυρίου -1616.503906 52100000 56100000 γιαννάκου -3143.976562

Πίνακας ΠΒ-3 : Προτάσεις αναγνώρισης από το MMIS και MMIS+MAP μοντέλο

Παρακάτω παρουσιάζονται τα αποτελέσματα της για το MMIS και για το τελικό μας πείραμα **<u>MIS+MAP+(SSA+USA_Known_Spk)+(USA_Unknown_Spk)</u>***

Αναγνώριση MMIS μοντέλου	**Αναγνώριση <u>τελικού</u>* μοντέλου**
070428MEGAED2000_265.rec" 1200000 4600000 η -2941.807861 4600000 8800000 ταινία -3550.817383 8800000 16100000 νόμιμη -5667.319336 16100000 28700000 αγοραπωλησία -9628.739258 28700000 35100000 οικοπέδου -5316.400391 35100000 38800000 ειδικών -2814.738281 38800000 40300000 των -1283.265625 40300000 45200000 φυλακών -3875.513672 45200000 48700000 στενεύει -2695.460938 48700000 52800000 βέβαια -3051.636719	070428MEGAED2000_265.rec" είναι απόλυτα νόμιμη αγοραπωλησία οικοπέδου στην επον το χίλια εννιακόσια ενενήντα εννιά
070522NETED2100_119.rec" 200000 2200000 αν -1851.612915 2200000 3200000 και -846.065430 3200000 4300000 δεν -869.929443 4300000 8600000 είναι -3528.847168 8600000 13800000 γράμμα -4341.831055 13800000 16900000 κενό -2622.939453 16900000 21800000 λοιπόν -4112.798828	070522NETED2100_119.rec" αυτή τη στιγμή βρισκόμαστε συνολικό
070428MEGAED2000_004.rec" 1700000 3800000 ο -1655.936401 3800000 6100000 κύριος -1932.078125 6100000 12400000 παπαδόπουλος -4955.271484 12400000 17500000 κατηγορείται -4151.908203 17500000 17800000 η -265.545898 17800000 21900000 γέλινεκ -3330.874023 21900000 24900000 ίσως -2543.544922 24900000 28600000 κάβανο -2933.359375 28600000 29800000 του -1080.533203 29800000 35800000 χρηματιστηρίου -4777.021484 35800000 37300000 είτε -1268.308594 37300000 38600000 στα -1228.152344 38600000 44400000 παιχνίδια -4300.769531 44400000 45600000 με -965.195312 45600000 46700000 τις -939.777344 46700000 50900000 μετοχές -3463.351562 50900000 52300000 που -1175.207031 52300000 55000000 στις -2287.886719	070428MEGAED2000_004.rec" ο κύριος παπαδόπουλος κατηγορείται για εμπλοκή στο σκάνδαλο του χρηματιστηρίου είτε στα παιχνίδια με τις μετοχές φούσκες
070428MEGAED2000_011.rec" 1700000 2700000 και -858.863037 2700000 5100000 ποια -2036.743530 5100000 6800000 είναι -1186.191650 6800000 7400000 η -437.255371 7400000 13300000 αφετηρία -4424.820801 13300000 15500000 αυτής -1725.451172 15500000 17400000 της -1443.903320 17400000 24600000 ιστορίας -4989.178711 24600000 28200000 είναι -2653.980469 28200000 32000000 μια -2929.751953 32000000 34500000 καλή -1896.265625 34500000 36900000 της -1829.597656 36900000 38400000 για -1248.363281 38400000 39800000 τα -1029.568359 39800000 42600000 έργα -2048.769531 42600000 43700000 και -868.234375 43700000 44900000 τις -925.765625 44900000 48600000 ημέρες -2904.210938 48600000 49900000 του -1067.226562 49900000 52100000 κυρίου -1696.621094 52100000 56200000 γιαννάκου -3314.394531	070428MEGAED2000_011.rec" και πια είναι η αφετηρία αυτής της ιστορίας είναι οι αποκαλύψεις για τα έργα και τις ημέρες του κυρίου

<u>Πίνακας ΠΒ-4 : Προτάσεις αναγνώρισης από το MMIS και το MIS+MAP+(SSA+USA_Known_Spk)+(USA_Unknown_Spk) μοντέλο</u>

Βιβλιογραφία

[1] S.J. Young, *The HTK Book*, Cambridge University Engineering Department, March 2009

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

[3] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, *Spoken Language Processing*, *Comegie Mellon University*, 2001

[4] Βασίλης Διγαλάκης, *Σημειώσεις του μαθήματος «Εισαγωγή στην Επεξεργασία Φωνής»*

[5] Keith Vertanen. *An Overview of Discriminative Training for Speech Recognition*, Computer Speech, Text and Internet Technology, University of Cambridge

[6] V. Valtchev, J.J. Odell, P.C. Woodland, S.J. Young, *MMIE training of large vocabulary recognition systems*, University Engineering Department, Cambridge

[7] Orfeas Tsergoulas, *A Greek Broadcast News Transcription System*, Technical University of Crete, Chania, November 2009

[8] http://www.voxforge.org

[9] http://www.ling.ohio-state.edu/~bromberg/htk_problems.html

[10] http://htk.eng.cam.ac.uk

[11] Veeru Ramaswamy, *The Importance of Adaptation in Automatic Speech Recognition*, *http://www.speechtechmag.com*, Jul 16, 2010

[12] P.C. Woodland, M.J.F. Gales, D. Pye & S.J. Young, *The Development of the 1996 HTK Broadcast News Transcription Ssystem*, Cambridge University Engineering Department, 2002

[13] S. S. Chen, E. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanvesky, P. Olsen, *Automatic Transcription of Broadcast News*, IBM T.J. Watson Research Center, March 7 2001

[14] Hank Liao, *Uncertainty Decoding for Noise Robust Speech Recognition*, Sidney Sussex College University of Cambridge, September 2007

[15] Joao P. Neto, Ciro Martins, Luıs B. Almeida, *An Incremental Speaker-Adaptation Technique for Hybrid HMM-MLP Recognizer*, Instituto de Engenharia de Sistemas e Computadores (INESC), Portugal , 1996

[16] D. Povey, P.C. Woodland, M.J.F: Gales, *Discriminative MAP for Acoustic Model Adaptation,* Cambridge University Engineering Department, 2003

[17] Veera Venkataramani, William Byrne, *MLLR Adaptation Techniques for Pronunciation Modeling,* Center for Language and Speech processing, The Johns Hopkins University, 2001

[18] Geoffrey Zweig, Michael Picheny, *Advances in Large Vocabulary Continuous Speech Recognition*, IBM T. J. Watson Research Center, 2004

[19] Florian Metze, Alex Waibel, U*sing articulatory information for speaker adaptation,* Carnegie Mellon University, 2003

[20] Arindam Mandal, *Transformation Sharing Strategies for MLLR Speaker Adaptation,* University of Washington, 2007

[21] Man-Wai Mak, Roger Hsiao, Brian Mak, *A Comparison of Various Adaptation Methods for Speaker Verification with Limited Enrollment Data*, The Hong Kong Polytechnic University, 2006