ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Ηλεκτρονικής & Μηχανικών Υπολογιστών

Εργαστήριο Τηλεπικοινωνιών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΟΥΜΠΙΩΤΗΣ ΒΛΑΣΙΟΣ

Αναγνώριση Φωνής με τεχνικές κανονικοποίησης

ΧΑΝΙΑ

Αύγουστος 1998

Διπλωματική Εργασία

Αναγνώριση Φωνής με τεχνικές κανονικοποίησης

Τμήμα Ηλεκτρονικής & Μηχανικών Υπολογιστών

Πολυτεχνείο Κρήτης

Εισηγητής:

Διγαλάκης Βασίλης, Επίκουρος Καθηγητής

Επιτροπή Παρακολούθησης *:*

Διγαλάκης Βασίλης, Επίκουρος Καθηγητής

Μαράς Ανδρέας, Αναπληρωτής Καθηγητής

Πατεράκης Μιχάλης, Αναπληρωτής Καθηγητής

Ευχαριστίες

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

Επίσης θα ηθελα να ευχαριστήσω τους καθηγητές κύριο Πατεράκη Μιχάλη και κύριο Μαρά Ανδρέα για τις συζητήσεις και υποδείξεις τους, από την ανάγνωση της διπλωματικής.

Θερμές ευχαριστίες σε όλα τα παιδιά του εργαστηρίου τηλεπικοινωνιών για τη κατανόηση και αγαστή συνεργασία όλη αυτή την περίοδο. Εκτός εργαστηρίου θα ήθελα να ευχαριστήσω προσωπικά τους Αφεντάκη Θέμη καθώς και πολλούς άλλους για την βοήθεια τους.

Τέλος, ευχαριστώ την οικογένειά μου για την ηθική και υλική συμπαράσταση, η οποία στάθηκε δίπλα μου όλο αυτό τον καιρό, από τη στιγμή της εισαγωγής μου στο τμήμα μέχρι και την ολοκλήρωση των σπουδών μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Ηλεκτρονικής & Μηχανικών Υπολογιστών

Εργαστήριο Τηλεπικοινωνιών

ΠΑΡΟΥΣΙΑΣΗ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

" Αναγνώριση Φωνής με τεχνικές κανονικοποίησης "

ΔΟΥΜΠΙΩΤΗΣ Δ ΒΛΑΣΙΟΣ

Δευτέρα 24 Αυγούστου 1998, Ώρα:11.00
Αίθουσα Ε33.01

Εξεταστική Επιτροπή : Επίκ. Καθηγητής Διγαλάκης Βασίλης (Επιβλέπων)

Αναπλ. Καθηγητής Μαράς Ανδρέας

Αναπλ. Καθηγητής Πατεράκης Μιχάλης

ΠΕΡΙΛΗΨΗ

Η αναγνώριση φωνής έχει συγκεντρώσει μεγάλο ενδιαφέρον τα τελευταία χρόνια εξ' αιτίας των ποικίλων εφαρμογών που μπορούν να βασισθούν σ' αυτή. Προκειμένου να εξασφαλιστεί σταθερότητα σε μεταβαλλόμενες συνθήκες πχ ομιλητή απαιτούνται γρήγορες και ακριβείς μεθόδους.

Η παρούσα διπλωματική εργασία ασχολείται με την ανάπτυξη αλγορίθμων κανονικοποίησης που έχουν σαν στόχο τη μείωση του σφάλματος αναγνώρισης σε συστήματα αναγνώρισης συνεχούς ομιλίας. Στην περίπτωσή μας, μια από τις κύριες αιτίες μείωσης της επίδοσης στα συστήματα αυτά είναι η μεταβλητότητα του μήκους του φωνητικού σωλήνα. Η κανονικοποίηση του μήκους του φωνητικού σωλήνα μέσω της συχνότητας (*vocal tract length normalization via frequency warping*) είναι μια δημοφιλής τεχνική όπου ο άξονας της συχνότητας επεκτείνεται ή συρρικνώνεται πριν την εξαγωγή των *cepstral* συντελεστών κατά τη διάρκεια επεξεργασίας του σήματος της φωνής. 'Ετσι επιλέγεται ένας συντελεστής για κάθε ομιλητή με σκοπό την επίτευξη καλύτερης απεικόνισης και τη μείωση του σφάλματος αναγνώρισης. Η διαδικασία αυτή εφαρμόζεται στο *front end.* Μελετάμε διάφορους μετασχηματισμούς στο πεδίο της συχνότητας όπως γραμμικούς, γραμμικούς κατά περιοχές και μετασχηματισμούς μεταφοράς. Στη συνέχεια αναπτύσσουμε αλγορίθμους για την εύρεση του βέλτιστου συντελεστή για κάθε ομιλητή στο *test set* για την επίτευξη καλύτερης επίδοσης αναγνώρισης μαζί με τα κριτήρια πιθανοφάνειας που πρέπει να ικανοποιούνται κατά την εκπαίδευση του GMM. Τέλος έγινε επανεκπαίδευση των αρχικών μοντέλων.

Μετά από σειρά πειραμάτων με δεδομένα που συγκεντρώθηκαν από ομιλητές που δεν είχαν λάβει μέρος στην εκπαίδευση του συστήματος, έγινε εφικτό το επί τοις εκατό σφάλμα αναγνώρισης να περιοριστεί σε ποσοστό 12**%** στους άντρες και 14**%** στις γυναίκες**.**.

ΕΙΣΑΓΩΓΗ

Ο τομέας της επεξεργασίας φωνής έχει γνωρίσει ραγδαία εξέλιξη τα τελευταία χρόνια, χάρη στις νέες μεθόδους οι οποίες μειώνουν το υπολογιστικό κόστος υλοποίησης των αλγορίθμων επεξεργασίας φωνής και στην ραγδαία εξελισσόμενη τεχνολογία υλικού και λογισμικού, που παρέχουν νέες δυνατότητες και πιο γρήγορες υπηρεσίες. Οι παραπάνω παράγοντες έχουν καταστήσει εφικτή την έξοδο της αναγνώρισης ομιλίας από το πειραματικό στάδιο και την εμφάνισή της στην αγορά ως εμπορικό προιόν.

Οι εφαρμογές της αναγνώρισης ομιλίας έχουν σχέση είτε με μείωση κόστους (αυτές δηλαδή οι οποίες αντικαθιστούν ανθρώπους από αναγνωριστές ομιλίας), είτε με δημιουργία νέων δυνατοτήτων και πιο γρήγορων υπηρεσιών, όπως πρόσβαση σε βάσεις δεδομένων και πληροφοριών που έχουν σχέση με κλείσιμο αεροπορικών θέσων, δελτία καιρού, χρηματιστήριο, υπηρεσίες τραπεζικών συναλλαγών μέσω φωνής και πλήθος άλλες εφαρμογές.

Η αναγνώριση ομιλίας (*speech recognition*) είναι η διαδικασία μετατροπής ενός ακουστικού σήματος, που λαμβάνεται μέσω μικροφώνου ή τηλεφωνικής γραμμής, σε μια ακολουθία λέξεων. 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*). Το σύστημα κατασκευάζει ένα δίκτυο, που υλοποιεί τη γραμματική και για κάθε επιτρεπόμενη πρόταση αντιστοιχίζεται ένα σύνολο από μοντέλα 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 τη μεγαλύτερη πιθανότητα. Η δομή των HMMs, καθώς και οι αλγόριθμοι εκπαίδευσης που έχουν αναπτυχθεί για τον καθορισμό των παραμέτρων αναγνώρισης, παρέχουν υψηλές επιδόσεις σε εφαρμογές που λειτουργούν

ανεξάρτητα από τον ομιλητή, εφαρμογές συνεχούς ομιλίας και μεγάλων λεξιλογίων. Στην προκειμένη περίπτωση το σύστημα, που χρησιμοποιήθηκε, βασίζει τη λειτουργία του σε εξελιγμένα HMMs, που αναπτύχθηκαν στο ερευνητικό ινστιτούτο SRI.

Η περιοχή της επεξεργασίας φωνής περιλαμβάνει τις εξής περιοχές: αναγνώριση, κωδικοποίηση, σύνθεση και τέλος εξακρίβωση ομιλητή. Οι παραπάνω εφαρμογές χρησιμοποιούνται καθημερινά από εκατομμύρια πελάτες. Τα συστήματα αναγνώρισης φωνής, ανάλογα με τις δυνατότητες τους, συχνά κατηγοριοποιούνται σε συστήματα απομονωμένων λέξεων/φράσεων, συστήματα συνδεδεμένων λέξεων και συστήματα συνεχούς ομιλίας.Τα συστήματα απομονωμένων λέξεων/φράσεων είναι οι πιο περιοριστικοί αναγνωριστές, αλλά μπορούν να λειτουργήσουν ικανοποιητικά σε μια μεγάλη ποικιλία εφαρμογών. Τα συστήματα συνδεδεμένων λέξεων είναι λιγότερο περιοριστικά και αρχίζουν να αποκτούν επίδοση κατάλληλη για μια σειρά από ενδιαφέρουσες εφαρμογές. Οι αναγνωριστές συνεχούς ομιλίας είναι ελάχιστα περιοριστικοί και απαιτητικοί από το χρήστη. Με το χρόνο η επίδοση τους βελτιώνεται και θα μπορούν να χρησιμοποιηθούν σε ιδιαίτερα απαιτητικές εφαρμογές. Είναι λοιπόν θέμα χρόνου να επιτευχθεί η αξιόπιστη επικοινωνία ανάμεσα σε ανθρώπους και μηχανές με στόχο την παροχή βέλτιστων υπηρεσιών στο χρήστη.

Τα συστήματα αναγνώρισης ομιλίας μπορούν να αναπτυχθούν με δεδομένα εκπαίδευσης που είναι είτε εξαρτημένα από ομιλήτη (*speaker-dependent*) ή έχουν συλλεγεί ανεξάρτητα από αυτόν (*speaker-independent*). Η διαφορά έγκειται στο αν τα λεκτικά πρότυπα κατασκευάζονται με ανάλυση των δεδομένων φωνής των ιδίων των χρηστών ή με επεξεργασία δεδομένων που προέρχονται από ένα ανεξάρτητο και αντιπροσωπευτικό δείγμα ομιλητών. Η ποσότητα των δεδομένων εκπαίδευσης που απαιτούνται για συστήματα εξαρτημένα από ομιλητή είναι κατά πολύ μικρότερη από αυτήν για κατασκευή συστήματος για ανεξάρτητους ομιλητές.

Το θέμα της διπλωματικής

Η αναγνώριση φωνής έχει συγκεντρώσει μεγάλο ενδιαφέρον τα τελευταία χρόνια εξ' αιτίας των ποικίλων εφαρμογών που μπορούν να βασισθούν σ' αυτή. Οι απαιτήσεις σε ακρίβεια απαιτούν γρήγορες και ακριβείς μεθόδους που να εξασφαλίζουν σταθερότητα σε μεταβαλλόμενες συνθήκες, π.χ ομιλητή. Η παρούσα διπλωματική εργασία ασχολείται με την ανάπτυξη αλγορίθμων κανονικοποίησης που έχουν σαν στόχο τη μείωση του σφάλματος αναγνώρισης σε συστήματα αναγνώρισης συνεχούς ομιλίας. Στην περίπτωσή μας, μια από τις κύριες αιτίες μείωσης της επίδοσης στα συστήματα αυτά είναι η μεταβλητότητα του μήκους του φωνητικού σωλήνα. Η κανονικοποίηση του μήκους του φωνητικού σωλήνα μέσω της συχνότητας (*vocal tract length normalization via frequency warping*) είναι μια δημοφιλής τεχνική όπου ο άξονας της συχνότητας επεκτείνεται ή συρρικνώνεται πριν την εξαγωγή των *cepstral* συντελεστών κατά τη διάρκεια της επεξεργασίας του σήματος της φωνής. 'Ετσι με την κανονικοποίηση επιλέγεται ένας συντελεστής ανά ομιλητή ή ανά πρόταση προκειμένου να επιτευχθεί καλύτερη απεικόνιση των ακουστικών χαρακτηριστικών του ομιλητή και κατά συνέπεια καλύτερη επίδοση αναγνώρισης. Η διαδικασία αυτή εφαρμόζεται στο *front-end*. Μελετάμε διάφορους μετασχηματισμούς στο πεδίο της συχνότητας, όπως γραμμικούς, γραμμικούς κατά περιοχές και μετασχηματισμούς μεταφοράς. Στη συνέχεια επιδιώκεται η ολοκλήρωση του συστήματος αναγνώρισης συνεχούς ομιλίας. Για το σκοπό αυτό αναπτύσσουμε αλγορίθμους για την εύρεση του βέλτιστου συντελεστή για κάθε ομιλητή στο *test set* για την επίτευξη καλύτερης επίδοσης αναγνώρισης. Οι αλγόριθμοι αυτοί περιλαμβάνουν παράλληλη αναγνώριση, αναγνώριση με χρήση GMM (*gaussian mixture model*) και τέλος αναγνώριση με επανεκπαίδευση των μοντέλων.

Οργάνωση της διπλωματικής

Η ύλη που παρουσιάζεται σε αυτή την εργασία έχει ως εξής:

Κεφάλαιο 1 με τίτλο: "Εισαγωγή στην Αναγνώριση Ομιλίας", όπου περιγράφονται η αναγνώριση με στατιστικές μεθόδους, η δομή των *HMMs* και ο φωνητικός σωλήνας.

Κεφάλαιο 2 με τίτλο: "Αρχιτεκτονική του Συστήματος Αναγνώρισης Ομιλίας", όπου περιγράφονται τα εξής: το υποσύστημα εξαγωγής παραμέτρων *front-end*, τα ακουστικά μοντέλα, το γλωσσικό μοντέλο και η διαδικασία της αποκωδικοποίησης.

Κεφάλαιο 3 με τίτλο: "Αναγνώριση Φωνής με τεχνικές κανονικοποίησης", όπου περιγράφονται οι αλγόριθμοι κανονικοποίησης του μήκους του φωνητικού σωλήνα μέσω της συχνότητας (*vocal tract length normalization via frequency warping*). Ο άξονας της συχνότητας επεκτείνεται ή συρρικνώνεται πριν την εξαγωγή των *cepstral* συντελεστών κατά τη διάρκεια επεξεργασίας του σήματος της φωνής. Μελετάμε διάφορους μετασχηματισμούς στο πεδίο της συχνότητας (γραμμικούς, γραμμικούς κατά περιοχές και μετασχηματισμούς μεταφοράς). Επιπλέον περιγράφουμε τους αλγορίθμους για την εύρεση του βέλτιστου συντελεστή για κάθε ομιλητή στο *test set* για την επίτευξη καλύτερης επίδοσης αναγνώρισης και τέλος τα κριτήρια πιθανοφάνειας που πρέπει να ικανοποιούνται κατά την επανεκπαίδευση των αρχικών μοντέλων.

Κεφάλαιο 4 με τίτλο: "Πειράματα Αναγνώρισης", όπου δίνουμε τις παραμέτρους των φίλτρων στις οποίες έγιναν επεμβάσεις για την επίτευξη καλύτερης επίδοσης αναγνώρισης του συστήματος καθώς επίσης και πειράματα με εφαρμογή αλγορίθμων για την εύρεση του βέλτιστου συντελεστή για κάθε ομιλητή στο *test set* .

Βιβλιογραφία, όπου δίνουμε τις παραπομπές στα επιστημονικά άρθρα που χρησιμοποίησαμε κατά την ανάπτυξη της εφαρμογής.

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

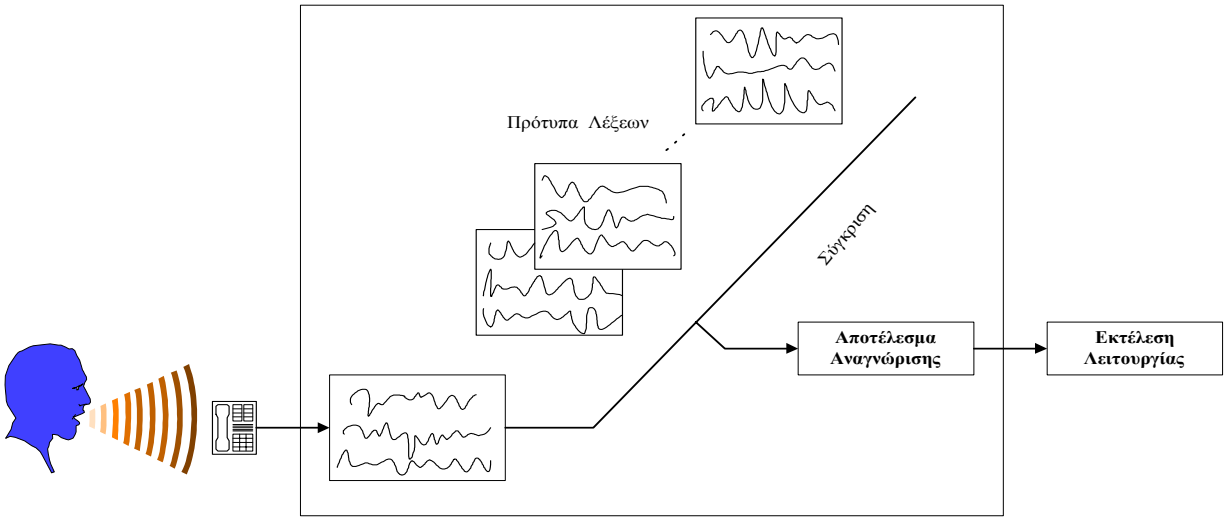
Τη δεκαετία 1990-2000 που διανύουμε, χαρακτηριστικά όπως αυτά της αναγνώρισης συνεχούς ομιλίας και της ραγδαίας αύξησης των υποστηριζόμενων λέξεων σε τάξεις δεκάδων χιλιάδων αποτελούν ήδη γεγονός. Είναι προφανές ότι μελλοντικά θα επιτευχθεί η ολοκλήρωση της επεξεργασίας φωνής με την επεξεργασία εικόνας, δεδομένων και ασύρματης μετάδοσης με αποτέλεσμα τα συστήματα να παρέχουν παγκόσμια πρόσβαση σε όλους τους χρήστες, οπουδήποτε, οποτεδήποτε και με λογικό κόστος, με απώτερο στόχο τη βελτίωση της ποιότητας ζωής. Σήμερα ο σημαντικότερος αποτρεπτικός παράγοντας για την διάδοση εφαρμογών αναγνώρισης φωνής παραμένει το υψηλό κόστος ανάπτυξης. Είναι πάντως βέβαιο, ότι η πρόοδος στο χώρο είναι ραγδαία, το κόστος ανάπτυξης αρχίζει να μειώνεται δραστικά και η βελτίωση των εργαλείων καθιστά την διάδοση τέτοιων εφαρμογών εξασφαλισμένη, οπότε η εμπορική αποδοχή μπορεί να θεωρείται δεδομένη. Οι εφαρμογές, λοιπόν, βελτιώνονται με ραγδαίους ρυθμούς, ωστόσο, είναι βέβαιο ότι μετά από λίγα χρόνια αναμένεται τα συστήματα να αποκτήσουν μια ασυμπτωτική μορφή στην πρόοδο τους και τα βήματα, ειδικά στην αύξηση της επίδοσης αναγνώρισης στο μέλλον, να είναι μικρότερα από ότι βλέπουμε σήμερα.

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΑΝΑΓΝΩΡΙΣΗ ΟΜΙΛΙΑΣ

1.1 Σύστημα Αναγνώρισης

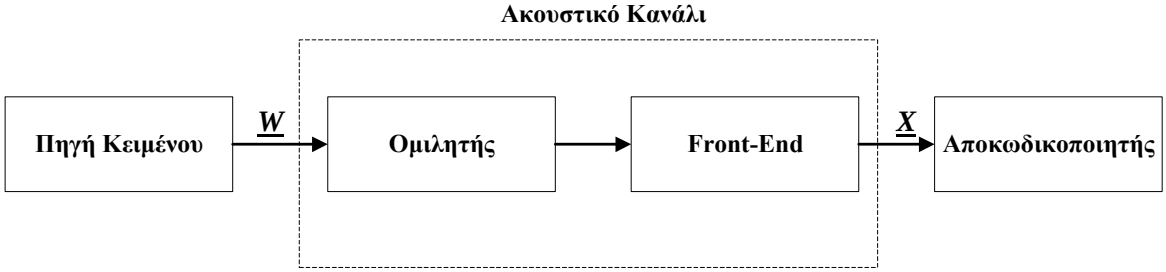
Τα σύγχρονα συστήματα βασίζονται σε στατιστικές μεθόδους αναγνώρισης. Το σύστημα χρησιμοποιεί ένα σύνολο προτύπων λέξεων ή φράσεων που δημιουργούνται από ένα πρόγραμμα εκπαίδευσης προτύπων, βασισμένο στη γραμματική αναγνώρισης και για κάθε επιτρεπόμενη πρόταση αντιστοιχίζεται ένα σύνολο από μοντέλα HMMs. Αυτά τα πρότυπα μπορούν να είναι τυπικά φάσματα προτύπων λέξεων, μέσες τιμές προτύπων φασμάτων προτύπων λέξεων δια μέσου διαφορετικών ομιλητών ή εξελιγμένα στατιστικά μοντέλα. Τα μοντέλα αυτά περιλαμβάνουν στατιστικούς μέσους όρους και φασματική μεταβλητότητα που εξαρτάται από τη χρονική διάρκεια της λέξης.



Σχήμα 1.1. Διάγραμμα αναγνώρισης

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

Η αναγνώριση ξεκινά με το ψηφιακοποιημένο σήμα ομιλίας, το οποίο υπόκειται σε επεξεργασία (*front-end*). Θεωρούμε ότι μια άγνωστη κυματομορφή σήματος φωνής μετατρέπεται από έναν front-end (Παρ.2.2) επεξεργαστή σε μια ακολουθία από ακουστικά διανύσματα $\underline{X} = [x_1, x_2, \ldots, x_T]$. Καθένα από αυτά τα διανύσματα είναι μια αναπαράσταση του φάσματος στο χρόνο καλύπτοντας τυπικά μία περίοδο 10msecs. Έτσι μια έκφραση δέκα λέξεων με διάρκεια γύρω στα 3secs μπορεί να αναπαρασταθεί με μια ακολουθία από T=300 ακουστικά διανύσματα.



Σχήμα 1-2 Μοντέλο αποκωδικοποίησης

Έστω λοιπόν ότι η **πηγή κειμένου** παράγει την ακολουθία λέξεων $\underline{W} = [w_1 w_2 \ldots w_n]$. Το **ακουστικό κανάλι** (μοντέλο παραγωγής φωνής του ομιλητή (παράγραφος 1.7) μαζί με τον *front-end* επεξεργαστή) αναλαμβάνει τη διαμόρφωση και μετάδοση του μηνύματος $\underline{W}$ μέσα από ένα θορυβώδες κανάλι. Στην έξοδο παίρνουμε την ακολουθία $\underline{X} = [x_1, x_2, \ldots, x_T]$ από παραμετρικά διανύσματα που υπολογίζονται από τον *front-end* επεξεργαστή του συστήματος αναγνώρισης χρησιμοποιώντας διάφορες μεθόδους, π.χ **Ανάλυση Γραμμικής Πρόβλεψης** (*Linear Prediction Analysis* - LPC), η **Εξαγωγή Mel-Frequency Cepstral Coefficients** (MFCC) και άλλες. Το επόμενο στάδιο περιλαμβάνει αναγνώριση φωνημάτων, ακολουθιών φωνημάτων και λέξεων. Σκοπός του αναγνωριστή είναι να μετατρέψει το σήμα φωνής που αναπαρίσταται με τη διανυσματική ακολουθία παρατήρησης $\underline{X}$, στην αντίστοιχη πρόταση σε γραπτή μορφή.

1.2 Κριτήριο Αναγνώρισης

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

Χρησιμοποιώντας τον κανόνα του Bayes έχουμε:

$$\hat{W} = \arg\max_{\underline{w}} P(\underline{W} \mid \underline{X}) = \arg\max_{\underline{w}} \frac{P(\underline{W})P(\underline{X} \mid \underline{W})}{P(\underline{X})} \qquad (1\text{-}2.1)$$

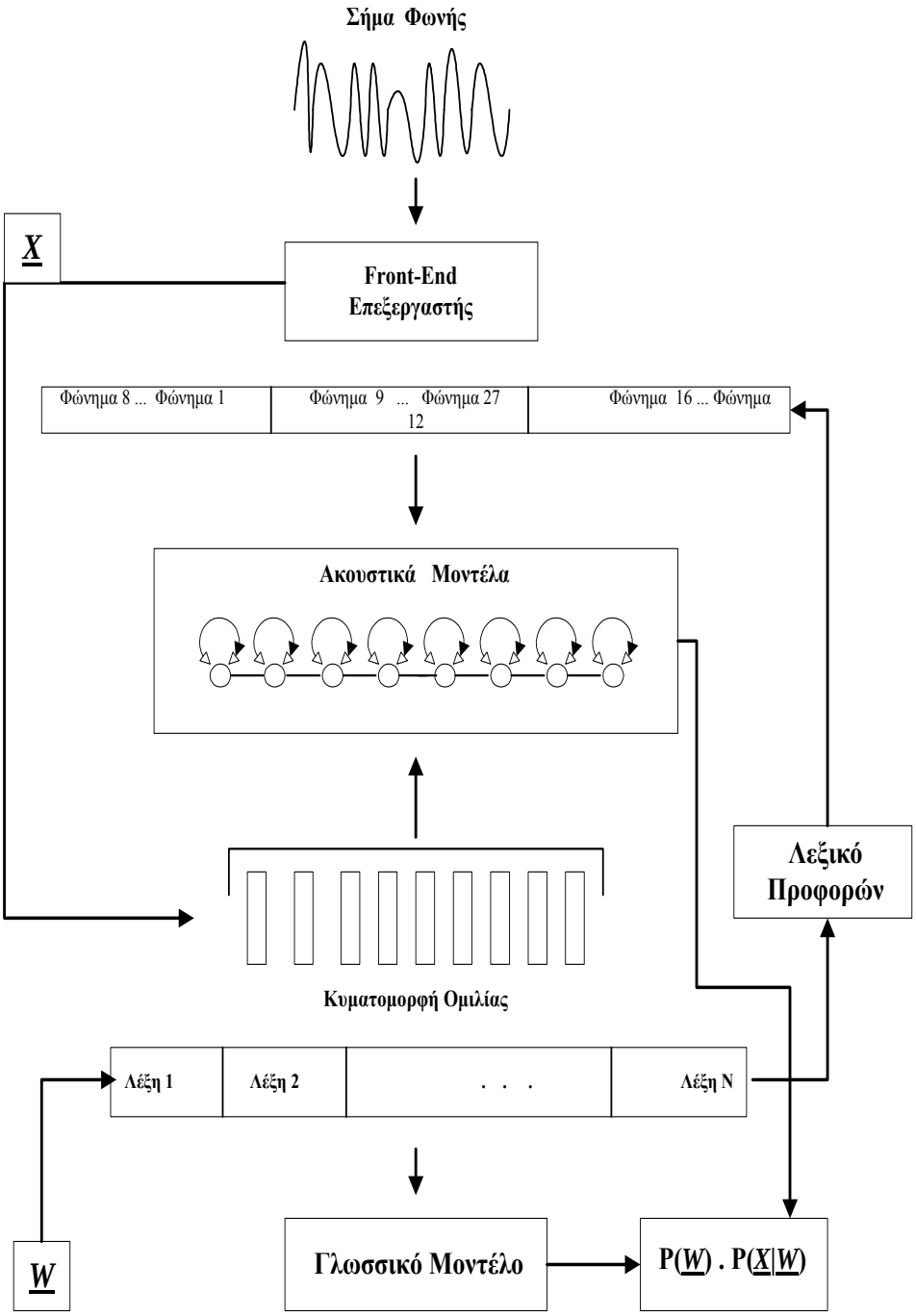
όπου το *argmax* συμβολίζει το όρισμα που μεγιστοποιεί την αντίστοιχη ποσότητα. Αυτή η εξίσωση δείχνει ότι για να βρεθεί η πιο πιθανή ακολουθία λέξεων $\underline{W}$, πρέπει να βρεθεί η ακολουθία εκείνη που μεγιστοποιεί το γινόμενο του $P(\underline{W})$ και του $P(\underline{X}|\underline{W})$. Ο πρώτος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

Η γλωσσική μονάδα που αναπαριστάται είναι συνήθως η λέξη. Για να υπάρχει η δυνατότητα γενίκευσης και να μοντελοποιούνται λέξεις που δεν παρατηρήθηκαν στα δεδομένα εκπαίδευσης, χρησιμοποιούνται μικρότερες γλωσσικές μονάδες όπως το **φώνημα** (*phoneme*) ή η συλλαβή. Η όλη διαδικασία ονομάζεται σχεδίαση του ακουστικού μοντέλου και περιγράφεται

στην παράγραφο 2-3. Κάθε λέξη μετατρέπεται σε μία ακολουθία βασικών ήχων, τα φωνήματα, 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 κάθε φώνημα υπάρχει ένα αντίστοιχο στατιστικό μοντέλο HMM. Από στατιστικής πλευράς, ένας κατάλογος από στοχαστικά μοντέλα βασικών φωνητικών μονάδων χρησιμοποιείται για να αναπαραστήσει λέξεις. Μία ακολουθία από ακουστικές παραμέτρους, προερχόμενες από το σήμα φωνής, αντιμετωπίζεται ως συνδυασμός στοιχειωδών διαδικασιών που περιγράφονται από HMMs.

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



Σχήμα 1-3 Αναγνώριση φωνής με στατιστικές μεθόδους

1.3 Περιγραφή των Hidden Markov Models

Ένα HMM είναι ένας συνδυασμός δύο στοχαστικών διαδικασιών ($\underline{Q}$, $\underline{X}$), μίας $\underline{Q}$ κρυφής αλυσίδας Markov (*hidden Markov chain*), η οποία περιγράφει χρονική μεταβλητότητα (*temporal variability*), και μίας $\underline{X}$ φανερής, η οποία περιγράφει τη φασματική μεταβλητότητα (*spectral variability*).Κάθε HMM χαρακτηρίζεται από τα εξής στοιχεία:

1. Αριθμός καταστάσεων: N

2. Πλήθος διακριτών συμβόλων που μπορούν να παρατηρηθούν ανά κατάσταση: M για διακριτά HMMs ή άπειρο για συνεχή HMMs, δηλαδή όταν x_t είναι πραγματικός αριθμός ή διάνυσμα πραγματικών αριθμών.

3. Πιθανότητες μετάβασης: η διαδικασία $\{q_t\}$ μοντελοποιείται ως αλυσίδα Markov με πιθανότητες μετάβασης $A=\{a_{ij}\}$, όπου $a_{ij}=P(q_{t+1}=j|q_t=i)$, με $1\le i\le N$.

4. Κατανομές εξόδου σε μία κατάσταση j. Σε κάθε χρονική στιγμή δημιουργείται μία παρατήρηση (τυχαίο διάνυσμα ή τυχαία μεταβλητή (διακριτή ή συνεχής)) με βάση μια κατανομή που εξαρτάται από την κατάσταση στην οποία βρισκόμαστε. Για διακριτά HMMs είναι $B=\{b_j(x_t)\}$, όπου το μέγεθος $b_j(x_t)=P(x_t|q_t=j)$ είναι η κατανομή εξόδου με $1\le j\le N$ και $1\le x_t\le M$.

5. Αρχικές πιθανότητες: $\Pi=\{\pi_i\}$, όπου $\pi_i=P(q_0=i)$, με $1\le i\le N$, για την ακολουθία καταστάσεων: $q_0,q_1,q_2,\ldots q_t,\ldots$ όπου $q_t\in\{1,2,\ldots,N\}$.

Έτσι, με κατάλληλες τιμές των μεγεθών *N, M, A, B* και π, το HMM μπορεί να χρησιμοποιηθεί σαν μία γεννήτρια που παράγει ακολουθίες εξόδου της μορφής:

$$\underline{X}=[x_1x_2\ldots x_T], \qquad (1\text{-}3.1)$$

όπου x_t είναι η παρατήρηση ενός συμβόλου και Τ είναι το πλήθος των παρατηρήσεων στην συγκεκριμένη ακολουθία. Για την πλήρη λοιπόν περιγραφή ενός HMM απαιτούνται οι παράμετροι Ν και Μ καθώς και ο καθορισμός του συνόλου των συμβόλων παρατήρησης και των τριών πιθανοτικών μεγεθών: Α, Β, π. Για συντομία χρησιμοποιούμε τον συμβολισμό:

$$\lambda=(A,B,\pi) \qquad (1\text{-}3.2).$$

1.4 Τα Hidden Markov Models στην αναγνώριση

Ο συνδυασμός της κρυφής αλυσίδας Markov με τις τυχαίες παρατηρήσεις έχει αποδειχθεί ότι παρέχει υψηλές επιδόσεις σε εφαρμογές που λειτουργούν ανεξάρτητα από τον ομιλητή και επιτρέπει ανάπτυξη εφαρμογών αναγνώρισης με λεξικά δεκάδων χιλιάδων λέξεων.

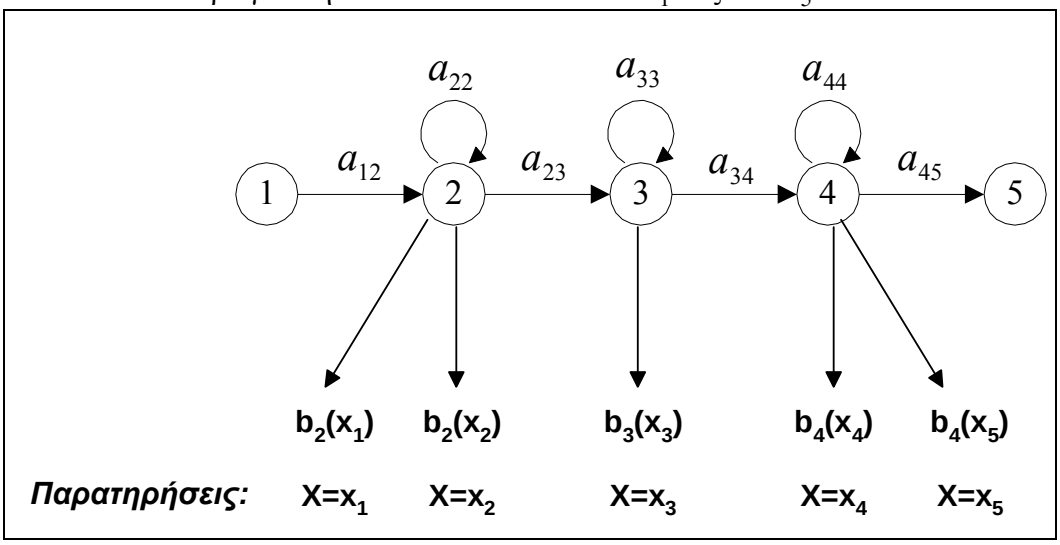
Η χρονική μεταβλητότητα του σήματος ομιλίας αναπαρίσταται από μια ακολουθία καταστάσεων που μοντελοποιείται ως στάσιμη αλυσίδα Markov. Η βασική πιθανότητα μετάβασης από τη μια κατάσταση στην άλλη εξαρτάται μόνο από τις δύο αυτές καταστάσεις και όχι τις προηγούμενές τους.

Οι προτάσεις χωρίζονται σε λέξεις, με βάση τη γραμματική της εφαρμογής. Οι λέξεις χωρίζονται σε φωνήματα, με βάση το λεξικό προφορών. Τα φωνήματα χωρίζονται σε αρχή, μέση και τέλος. Κάθε ανεξάρτητο φώνημα αναπαρίσταται από ένα HMM. Ένα HMM έχει έναν αριθμό από καταστάσεις που συνδέονται με βέλη. Τα μοντέλα φωνημάτων με HMMs έχουν τυπικά τρεις καταστάσεις, που μπορούν να δώσουν τιμή εξόδου (*emitting states*). Τα τρία αυτά επίπεδα ακολουθίας συνδυάζονται, με αποτέλεσμα τη δημιουργία ενός δικτύου, οι καταστάσεις του οποίου μοντελοποιούνται με μια αλυσίδα Markov.

Σημειώνεται ότι κάθε κατάσταση σχετίζεται με μια διαφορετική κατάσταση του **φωνητικού σωλήνα**. Οι καταστάσεις εισόδου και εξόδου προορίζονται για να κάνουν εύκολη την σύνδεση μεταξύ των μοντέλων. Αυτό σημαίνει ότι η κατάσταση εξόδου ενός μοντέλου μπορεί να συγχωνευτεί με την κατάσταση εισόδου ενός άλλου, δημιουργώντας ένα σύνθετο HMM. Αυτό επιτρέπει στα μοντέλα φωνημάτων να μπορούν να συνδεθούν μεταξύ τους παράγοντας λέξεις

και οι λέξεις με τη σειρά τους να μπορούν να συνδεθούν ώστε να αναπαραστήσουν πλήρεις εκφράσεις. Τα HMMs έχουν τη δυνατότητα να μοντελοποιούν και τη διάρκεια του ήχου, στοιχείο απαραίτητο εξαιτίας της μεταβλητότητας του ρυθμού της ομιλίας που εμφανίζεται ανάμεσα στους ομιλητές.

Ένα HMM γίνεται ευκολότερα κατανοητό αν θεωρηθεί ως μια γεννήτρια ακολουθιών διανυσμάτων. Είναι ένα διάγραμμα πεπερασμένων καταστάσεων (*finite-state machine-FSM*), το οποίο αλλάζει κατάσταση κάθε χρονική μονάδα, *t,* εισέρχεται σε μία κατάσταση, *j,* και παράγεται ένα ακουστικό διάνυσμα x_t με κατανομή εξόδου $b_j(x_t)$. Επιπλέον, η μετάβαση από μια κατάσταση, *i,* σε μια κατάσταση *j* είναι και αυτή πιθανοτική και καθορίζεται από μια διακριτή πιθανότητα a_{ij}. Στο Σχήμα 1-4 φαίνεται ένα παράδειγμα αυτής της διαδικασίας όπου το μοντέλο κινείται από την ακολουθία καταστάσεων $\underline{Q}$*=[1, 2, 2, 3, 4, 4, 5]* με σκοπό να παράγει την ακολουθία από το x_1 ως το x_5.



Σχήμα 1-4 Σχήμα HMM με 5 καταστάσεις

Η πιθανότητα μιας ακολουθίας διανυσμάτων $\underline{X}$ και ακολουθίας καταστάσεων *Q*, δεδομένου κάποιου μοντέλου M, υπολογίζεται απλά ως το γινόμενο πιθανοτήτων μετάβασης και των πιθανοτήτων εξόδου. Έτσι, για την ακολουθία καταστάσεων *Q* του σχήματος είναι:

$$P(\underline{X},\underline{Q}|M) = a_{12}b_2(x_1)a_{22}b_2(x_2)a_{23}b_3(x_3)... \qquad (1\text{-}4.1)$$

Πιο γενικά, η συνδυασμένη πιθανότητα μιας ακολουθίας ακουστικών διανυσμάτων $\underline{X}$ της αντίστοιχης ακολουθίας καταστάσεων $\underline{Q} = [q_1,q_2,\ldots,q_T]$ δίνεται από:

$$P(\underline{X},\underline{Q}|M) = a_{q(0)q(t)} \prod_{t=1}^{T} b_{q_t}(x_t) a_{q_t q_{t+1}}, \qquad (1\text{-}4.2)$$

όπου q_0 προορίζεται να είναι η κατάσταση εισόδου του μοντέλου και q_{T+1} προορίζεται να είναι η κατάσταση εξόδου του μοντέλου. Στην πράξη, βέβαια, μόνο η ακολουθία παρατηρήσεων $\underline{X}$ είναι γνωστή ενώ η ακολουθία $\underline{Q}$ είναι κρυφή. Ωστόσο, η απαιτούμενη πιθανότητα $P(\underline{X}|M)$ υπολογίζεται αθροίζοντας την τελευταία εξίσωση πάνω σε **όλες** τις δυνατές ακολουθίες καταστάσεων.

1.5 Κατηγορίες των HMMs

Ανάλογα με το αν η διαδικασία που μοντελοποιούμε αποτελείται από συνεχή τυχαία διανύσματα (π.χ παράμετροι LPC, συντελεστές *cepstral* κ.λ.π.), ή έχει περάσει από κβαντιστή και είναι διαδικασία από διακριτές τυχαίες μεταβλητές έχουμε διαφορετικά είδη HMMs, που ταξινομούνται ανάλογα με τον τύπο της κατανομής εξόδου. Έτσι τα HMMs χωρίζονται σε διακριτά και συνεχή.

Διακριτά HMMs (Discrete HMMs)

Αν η διαδικασία $\{x_t\}$ είναι διακριτή, με $x_t \in \{1,\ldots,Q\}$ τότε και η κατανομή εξόδου $b_j(x_t)$ είναι διακριτή με κατανομή:

$$\sum_{k=1}^{Q} b_j(k) = 1. \qquad (1\text{-}5.1)$$

Συνεχή HMMs (Continuous HMMs)

Οι κατανομές εξόδου είναι από κοινού συναρτήσεις πυκνότητας πιθανότητας ενός τυχαίου διανύσματος x_t με τιμή:

$$b_j(x_t), \qquad \text{όπου } x_t = \begin{bmatrix} x_{1t} \\ x_{2t} \\ \dots \\ x_{dt} \end{bmatrix} \tag{1-5.2}$$

και d είναι η διάσταση του x_t (π.χ τάξη της ανάλυσης LPC, αριθμός συντελεστών cepstral κ.λ.π.).

Τα συνεχή HMMs διακρίνονται με τη σειρά τους σε Gaussian HMMs και Gaussian Mixture HMMs.

- **Gaussian HMMs**

Τα Gaussian HMMs έχουν κατανομές εξόδου με τιμή:

$$b_j(x_t) = N(x_t; \mu_j; \Sigma_j) = \frac{1}{(2\pi)^{d/2} \left|\Sigma_j\right|^{1/2}} e^{-\frac{1}{2}(x_t-\mu_j)^T \Sigma_j^{-1} (x_t-\mu_j)} \tag{1-5.3}$$

με παραμέτρους κατανομής τις ποσότητες μ_j, Σ_j ($|\ |$ ορίζουσα).

Συχνά, για πιο ευέλικτη αναπαράσταση των δεδομένων χρησιμοποιείται διαγώνιος πίνακας συμμεταβλητότητας, οπότε οι κατανομές εξόδου γράφονται στη μορφή:

$$b_j(x_t) = \frac{1}{(2\pi)^{d/2} \prod_{k=1}^{d} \sigma_{jk}} e^{-\frac{1}{2}\sum_{k=1}^{d} \frac{(x_{kt}-\mu_{jk})^2}{\sigma^2{}_{jk}}} \tag{1-5.4}$$

όπου $\mu_j = E\{x_t | s_t = j\} = \begin{bmatrix} \mu_{j1} \\ .... \\ \mu_{jd} \end{bmatrix}$

και οι μεταβλητές σ_{ik} φυλάσσονται σε διαγώνιο πίνακα Σ_j με στοιχεία διαγωνίου:

$$\Sigma_j = \sigma^2_{j1} \sigma^2_{j2} \cdots \sigma^2_{jd}, \tag{1-5.5}$$

όπου $\sigma_{jk}^{2} = E\{(x_{kt}-\mu_{jk})^2\}$.

•Gaussian Mixture HMMs

Επειδή η μια Γκαουσιανή μπορεί να μην επαρκεί για να μοντελοποιήσει την κατανομή του X_t σε μια κατάσταση, ειδικά σε συστήματα αναγνώρισης ανεξάρτητα του ομιλητή, χρησιμοποιούνται μείγματα (γραμμικοί συνδυασμοί) από Γκαουσιανές ως κατανομές εξόδου. Αυτές οι κατανομές εξόδου ενός Gaussian Mixture HMM περιγράφονται από τη σχέση:

$$b_j(x_t) = \sum_{m=1}^{M} c_{jm} N(x_t; \mu_{jm}, \Sigma_{jm}). \qquad (1\text{-}5.6)$$

Παραπάνω ισχύει:

$$\sum_{m=1}^{M} c_{jm} = 1, \qquad (1\text{-}5.7)$$

με $c_{jm} = P(m|s_t = j)$, έτσι ώστε:

$$\int b_j(x_t)\, dx_t = 1. \qquad (1\text{-}5.8)$$

Παρόλο που ο υπολογισμός των πιθανοτήτων με διακριτά HMMs είναι ταχύτερος από τα συνεχή HMMs, τα τελευταία επιλέγονται λόγω της καλύτερης επίδοσης τους. Συνεχή μοντέλα HMMs είναι τα **Genones** και τα **PTM Models** (*Phonetically Tied Mixture*). Μεταξύ τους τα δύο αυτά μοντέλα διαφέρουν ως προς τις απαιτήσεις υπολογισμών και την επίδοση. Συγκεκριμένα, τα PTM Models χαρακτηρίζονται από μεγαλύτερη ταχύτητα, αλλά και μικρότερη ακρίβεια από τα Genones, γιατί τα πρώτα βασίζονται σε λιγότερες Γκαουσιανές ανά φώνημα.

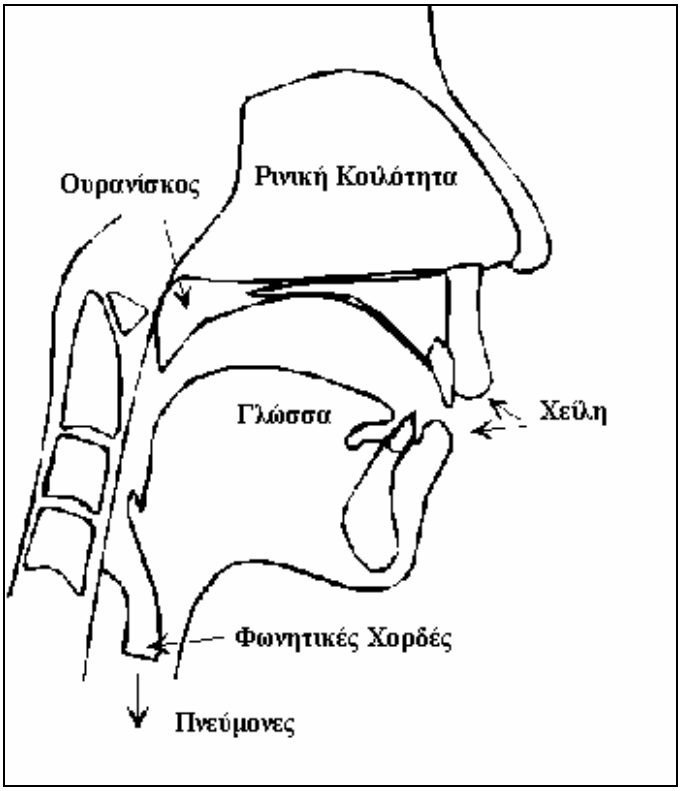
1.6 Αλγόριθμοι στην Αναγνώριση

Η χρήση των HMMs στην αναγνώριση προυποθέτει τα εξής: υπολογισμό της πιθανότητας μιας ακολουθίας παρατηρήσεων, υπολογισμό της πιο πιθανής ακολουθίας καταστάσεων και τέλος την εκμάθηση των παραμέτρων τους από δεδομένα εκπαίδευσης. Για το σκοπό αυτό έχουν αναπτυχθεί κατάλληλοι αλγόριθμοι.

Στη φάση της εκπαίδευσης λοιπόν θέλουμε να εκτιμήσουμε τις τιμές των παραμέτρων λ=(A,B,π) από την ακολουθία παρατηρήσεων $\underline{X} = [x_1,x_2,\ldots,x_T]$. Μια σειρά επαναληπτικών αλγορίθμων πρέπει να τρέξει πάνω στα δεδομένα εκπαίδευσης (*training data*), ώστε να διαθέτουμε τα απαραίτητα στοιχεία για τον υπολογισμό τους. Ο **αλγόριθμος Forward-Backward** είναι μία επαναληπτική μέθοδος για τον υπολογισμό της πιθανότητας να βρισκόμαστε σε μια συγκεκριμένη κατάσταση σε μια συγκεκριμένη χρονική στιγμή. Ο **αλγόριθμος Baum-Welch,** είναι μια μέθοδος για την εύρεση εκτιμητριών μέγιστης πιθανοφάνειας των συνόλων παραμέτρων του HMM. Ο **αλγόριθμος Viterbi** χρησιμοποιείται για να βρούμε τη συνολικά βέλτιστη ακολουθία καταστάσεων, δηλαδή την πιο πιθανή ακολουθία $\underline{Q} = [q_1,q_2,\ldots,q_T]$, δεδομένου ότι έχουμε παρατηρήσει την ακολουθία $\underline{X} = [x_1,x_2,\ldots,x_T]$. Το πρόβλημα μπορεί να λυθεί χρησιμοποιώντας **αποκωδικοποίηση *trellis***. Στη συνέχεια θα μελετήσουμε το φωνητικό σωλήνα.

1.7 Φωνητικός σωλήνας

Όλοι οι φωνητικοί ήχοι παράγονται όταν αέρας που προέρχεται από τους πνεύμονες διεγείρει τις φωνητικές χορδές, περάσει μέσα από το φωνητικό σωλήνα. Για έναν ενήλικα το μέσο μήκος του φωνητικού σωλήνα είναι 17 cm και κυμαίνεται από 13 (γυναίκες) μέχρι 20 (άντρες) cm. Το γεγονός ότι το μήκος του φωνητικού σωλήνα διαφέρει ανάλογα με το φύλο του ομιλητή παίζει σημαντικό παράγοντα στην αναγνώριση μιας και οι τεχνικές κανονικοποίησης που μελετάμε στην παρούσα διπλωματική στηρίζονται στη διαπίστωση αυτή. Ένα απλό διάγραμμα του φωνητικού σωλήνα φαίνεται στο σχήμα 1.5:



Σχήμα 1.5 Ο φωνητικός σωλήνας

Επίσης υπάρχει και ένας δευτερεύοντας σωλήνας, ο ρινικός, ο οποίος είναι συνδεδεμένος με το φωνητικό μέσω του ουρανίσκου. Οι φωνητικές χορδές είναι δύο μεμβράνες στο λάρυγγα που μεταβάλλουν την επιφάνεια της τραχείας. Έτσι ενώ κατά την αναπνοή, οι φωνητικές χορδές παραμένουν ανοιχτές, κατά την παραγωγή της ομιλίας ανοιγοκλείνουν με κάποιο ρυθμό. Τελικά η παραγωγή της ομιλίας είναι η συνισταμένη των μεταβολών στο φωνητικό σωλήνα (κίνηση του στόματος και της γλώσσας) και της ταλάντωσης των φωνητικών χορδών.

Η διέγερση είναι: είτε μια κρουστική παλμοσειρά (εύφωνος ήχος), είτε ένα σήμα θορύβου (άφωνος ήχος). Οι εύφωνοι ήχοι (πχ. φωνήεντα) δημιουργούνται στη γλωττίδα. Αυτά τα σήματα είναι γενικά μεγάλου πλάτους, σχεδόν περιοδικά και δίνουν έμφαση στις χαμηλές συχνότητες. Η ενέργεια του φάσματος είναι συγκεντρωμένη στις αρμονικές συχνότητες ταλάντωσης ή συχνότητες συντονισμού του φωνητικού σωλήνα, οι τέσσερις πρώτες αρμονικές ταλάντωσης βρίσκονται στην περιοχή 0-4kHz με τη πρώτη αρμονική περίπου στα 700Hz. Οι άφωνοι ήχοι (πχ s, f,) παράγονται από

ασταθή ροή αέρα που δημιουργείται από κάποια σύσπαση σε κάποιο μέρος του φωνητικού σωλήνα. Τα σήματα αυτά είναι συνήθως μικρού πλάτους και έχουν ευρύ και επίπεδο φάσμα όπως ο θόρυβος. Υπάρχουν ήχοι που μπορεί να είναι εύφωνοι ή άφωνοι και είναι αποτέλεσμα της συγκέντρωσης πίεσης σε κάποιο κλειστό σημείο, συνοδευόμενης από απότομη απελευθέρωση. Αυτοί οι ήχοι ονομάζονται 'Εκρηκτικοί' ήχοι(εύφωνοι: b, d, άφωνοι: p, t, k) και μοντελοποιούνται σαν μια βηματική συνάρτηση.

Η ομιλία είναι ένας συνδυασμός από τις παραπάνω κατηγορίες ήχων. Όμως ο παραπάνω διαχωρισμός είναι χρήσιμος για την περιγραφή της διέγερσης και του φωνητικού σωλήνα όπως θα δούμε παρακάτω.

•**Διέγερση:** Η άφωνη διέγερση μπορεί να αναπαρασταθεί σαν τυχαίος θόρυβος με Γκαουσιανό πλάτος και επίπεδο φάσμα. Στην περίπτωση αυτή το φάσμα της φωνής στην έξοδο του συστήματος διαμορφώνεται μόνο από την απόκριση συχνότητας του φωνητικού σωλήνα. Αντίθετα η εύφωνη διέγερση παίζει σημαντικό ρόλο και μοντελοποιείται γενικά σαν μία παλμοσειρά, η οποία διεγείρει το γλωττικό φίλτρο. Στο πεδίο της συχνότητας, το φίλτρο έχει μια πολύ χαμηλή συχνότητα αποκοπής και μία πτώση κατά 12 db/octave. Το φάσμα εξόδου στην περίπτωση αυτή διαμορφώνεται από το φίλτρο του φωνητικού σωλήνα και από τα γλωττικά φίλτρα.

•**Φωνητικός σωλήνας:** Για να εξάγουμε τις παραμέτρους του φίλτρου του φωνητικού σωλήνα, υποθέτουμε επίπεδη διάδοση των ηχητικών κυμάτων, φωνητικό σωλήνα μήκους 17 cm ομοιόμορφης διατομής με ανελαστικά τοιχώματα ο οποίος διαρρέεται από αέριο χωρίς τριβές. Αν εφαρμόσουμε τους νόμους της φυσικής, όπως η αρχή διατήρησης μάζας και ενέργειας, οδηγούμαστε σε μια απλή λύση των κυματικών εξισώσεων. Η συνάρτηση μεταφοράς του φίλτρου αυτού είναι:

$$V(j\Omega) = 1/\cos(\Omega l / c) \qquad (1\text{-}7.1)$$

όπου l είναι το μήκος του σωλήνα, c η ταχύτητα του ήχου και Ω η συχνότητα. Στο πεδίο της συχνότητας έχουμε έναν άπειρο αριθμό πόλων στον jΩ άξονα στις συχνότητες $F_i = \frac{(2i-1)c}{4l} = 500,1500,2500,....Hz$. Αυτές αποτελούν

τις συχνότητες συντονισμού του ομοιόμορφου φωνητικού σωλήνα χωρίς απώλειες. Οι συχνότητες συντονισμού εξαρτώνται από το μήκος του φωνητικού σωλήνα (αντιστρόφως ανάλογες). Για μεγαλύτερο μήκος του φωνητικού σωλήνα(άντρες) έχουμε μικρότερες συχνότητες συντονισμού, το αντίστροφο ισχύει για τις γυναίκες. Η παραπάνω μελέτη είναι χρήσιμη όταν θα περιγραφεί η εξαγωγή των *cepstral* συντελεστών από τον *frond-end* επεξεργαστή. Στο επόμενο κεφάλαιο θα μελετήσουμε τον τρόπο εφαρμογής των παραπάνω στοιχείων σε πρακτικό σύστημα (αρχιτεκτονική του συστήματος αναγνώρισης).

ΚΕΦΑΛΑΙΟ 2

ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΑΝΑΓΝΩΡΙΣΗΣ ΟΜΙΛΙΑΣ

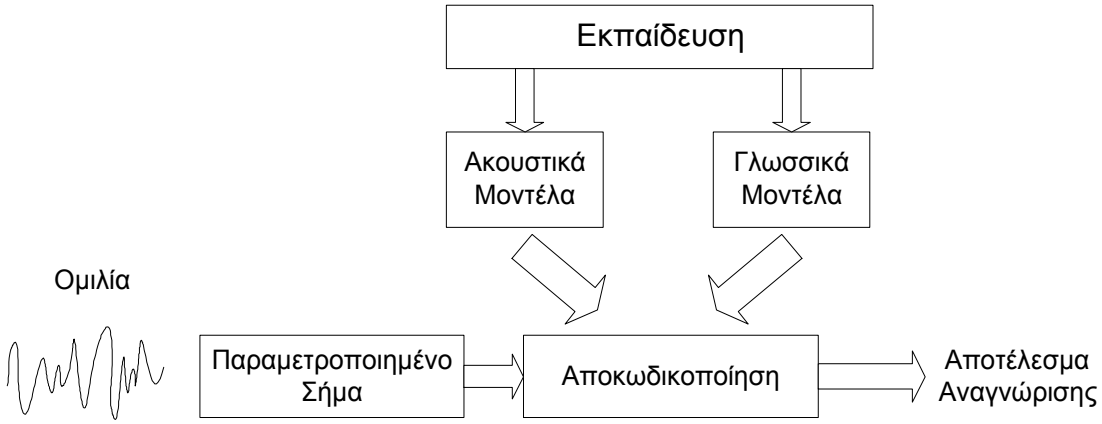
2.1 Γενική αρχιτεκτονική του συστήματος

Όπως προαναφέρθηκε τα σύγχρονα συστήματα αναγνώρισης φωνής βασίζονται στις αρχές της στατιστικής αναγνώρισης προτύπων. Οι αρχές αυτές πρωτοεφαρμόστηκαν το 1970 όταν εισήχθει το μοντέλο πηγής-καναλιού *(source-channel model)*. Από τότε τα περισσότερα συστήματα αναγνώρισης φωνής χρησιμοποιούν το μοντέλο αυτό με ελάχιστες παραλλαγές. Σύμφωνα με το μοντέλο πηγής-καναλιού, το σύστημα αναγνώρισης φωνής πρέπει να περιλαμβάνει τα εξής:

- **Υποσύστημα εξαγωγής παραμέτρων της φωνής**: Το σύστημα αυτό χρησιμοποιεί αλγορίθμους ψηφιακής επεξεργασίας σήματος για την εξαγωγή παραμέτρων χαμηλότερου *bit-rate* από το αρχικό σήμα και συμβατών με τα ακουστικά μοντέλα του αναγνωριστή.
- **Τα ακουστικά μοντέλα**: Τα ακουστικά μοντέλα αποτελούνται από HMM (*Hidden Markov Models*) τα οποία αναπαριστούν τα φωνήματα και έχουν εκπαιδευθεί με κάποια δεδομένα εκπαίδευσης. Τα HMM πρέπει να αναπαριστούν όχι μόνο τα ίδια τα φωνήματα, αλλά και τη συμπεριφορά τους όταν βρίσκονται ανάμεσα σε άλλα φωνήματα (*context dependent*) για καλύτερη μοντελοποίηση.
- **Το γλωσσικό μοντέλο**: Αποτελεί το ‘συντακτικό’ του συστήματος αναγνώρισης. Θα πρέπει να εκπαιδευθεί ώστε να δίνει όσο το δυνατόν καλύτερες εκτιμήσεις βάσει των δεδομένων εκπαίδευσης αλλά και της συντακτικής τους δομής μιας και είναι σίγουρο ότι τα δεδομένα εκπαίδευσης δε θα είναι αρκετά για να καλύψουν όλες τις πιθανές ακολουθίες λέξεων.
- **Ο αποκωδικοποιητής** : Είναι το μέρος του συστήματος που ψάχνει για όλες τις πιθανότερες ακολουθίες λέξεων των οποίων πρέπει να υπολογιστεί η πιθανότητα $P(\underline{W})$. Όπως είναι φυσικό, δεν είναι δυνατό να εξεταστούν όλες οι

δυνατές ακολουθίες λέξεων σε πραγματικό χρόνο. Ο αποκωδικοποιητής ψάχνει παράλληλα και απορρίπτει βάσει κάποιου κριτηρίου τις ακολουθίες $\underline{W}$ που καθίστανται απίθανες.

Η εξάρτηση των παραπάνω στοιχείων ενός συστήματος αναγνώρισης φαίνεται στο Σχήμα 2-1.



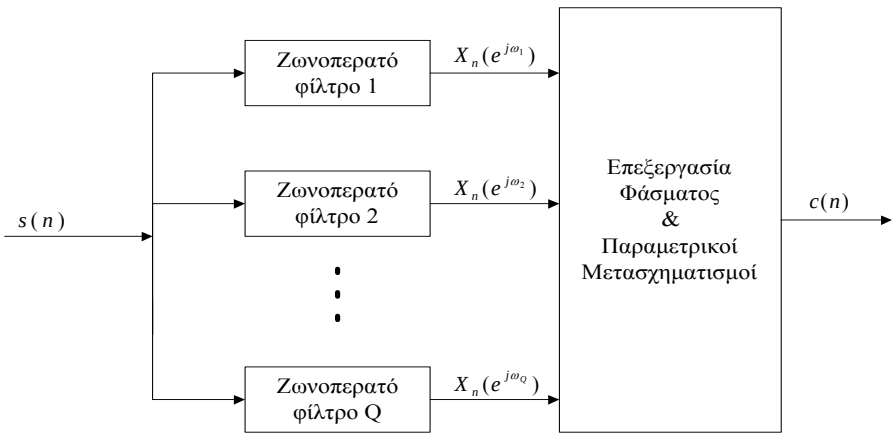
Σχήμα 2.1 Στοιχεία αναγνωριστή

Παρακάτω θα αναλύσουμε καθένα από τα επιμέρους συστήματα.

2.2 Ψηφιακή Επεξεργασία σήματος

2.2.1 Φασματική ανάλυση

Το σχήμα του μοντέλου φασματικής ανάλυσης με σετ φίλτρων φαίνεται παρακάτω:

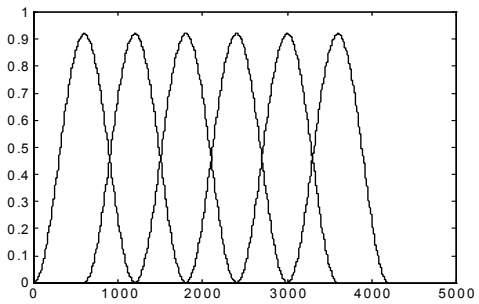


Σχήμα 2.2 Η φασματική ανάλυση

Το σήμα φωνής s(*n)* περνά από έναν αριθμό Q ζωνοδιαβατών φίλτρων (*filterbank*) τα οποία καλύπτουν το εύρος συχνοτήτων του σήματος εισόδου που μας ενδιαφέρει (π.χ. 100-3000 Hz για σήματα τηλεφωνικής ποιότητας και 100-8000 Hz για σήματα υψηλής ποιότητας). Τα φίλτρα μπορεί να αλληλοκαλύπτονται μεταξύ τους στη συχνότητα. Η έξοδος του i-οστού φίλτρου $X_n\left(e^{j\omega_i}\right)$, (όπου $\boldsymbol{\omega_i}$ είναι η κανονικοποιημένη συχνότητα $\mathbf{2\pi f_i/F_s}$ και $\mathbf{F_s}$ η συχνότητα δειγματοληψίας) είναι η φασματική αναπαράσταση βραχέως χρόνου του σήματος *s(n)* τη χρονική στιγμή *n*, όπως εξάγεται από το i-οστό ζωνοδιαβατό φίλτρο με κεντρική συχνότητα $\boldsymbol{\omega_i}$. Στο μοντέλο αυτό η επεξεργασία του σήματος γίνεται παράλληλα και ανεξάρτητα για κάθε μπάντα συχνοτήτων παράγοντας τη φασματική αναπαράσταση X_n. Τις περισσότερες φορές όμως δεν αρκεί η απλή φασματική αναπαράσταση αλλά οι παράμετροι X_n υφίστανται περαιτέρω επεξεργασία για ακόμα καλύτερη απόδοση.

Υπάρχουν διάφορα είδη φίλτρων που χρησιμοποιούνται:

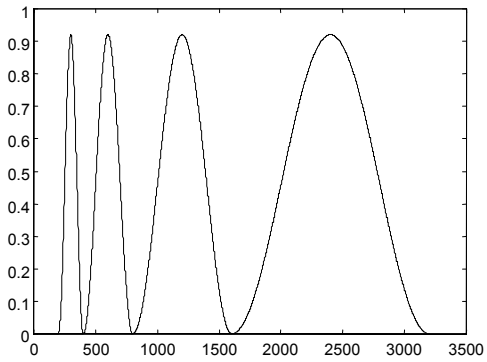
•Ομοιόμορφα κατανεμημένη *filterbank*, όπου η κεντρική συχνότητα του i-οστού ζωνοδιαβατού φίλτρου ορίζεται σαν $f_i = \frac{F_s}{N} i, \quad 1 \le i \le Q$ όπου F_s είναι η συχνότητα δειγματοληψίας του σήματος και N ο αριθμός των ομοιόμορφα κατανεμημένων φίλτρων που απαιτούνται για την κάλυψη όλου του εύρους φάσματος που μας ενδιαφέρει και Q ο πραγματικός αριθμός των φίλτρων που χρησιμοποιούνται. Ο αριθμός αυτός (λόγω της επικάλυψης μεταξύ τους) είναι $Q \le N/2$ με την ισότητα να ισχύει όταν όλο το εύρος του σήματος χρησιμοποιείται. Το εύρος φάσματος b_i του i-οστού φίλτρου ικανοποιεί τη σχέση $b_i \ge F_s/N$, με την ισότητα να ισχύει όταν δεν υπάρχει επικάλυψη μεταξύ των φίλτρων. Στο Σχ. 2.3 φαίνεται ένα σετ 6 φίλτρων.



Σχήμα 2.3 Ομοιόμορφη filterbank

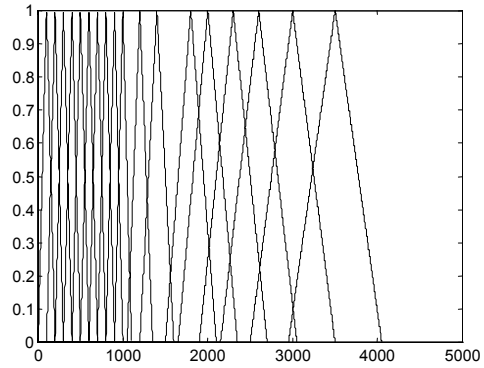
Ισχύει ότι $Q=6$, Fs/N=600 Hz, $b_i=2$(Fs/N)=1200Hz, δηλαδή καλύπτει την περιοχή $0 \le f \le Q(F_s/N)+F_s/N$ ή $0Hz \le f \le 4200Hz$.

•Μη ομοιόμορφα κατανεμημένη *filterbank*. Συνήθως οι κεντρικές συχνότητες των φίλτρων τοποθετούνται σε λογαριθμική κλίμακα. Το Σχ. 2.4 δείχνει ένα σετ 4 φίλτρων οκτάβας (δηλαδή το εύρος του i φίλτρου είναι διπλάσιο από το εύρος του i-1) που καλύπτουν ένα εύρος 200-3200 Hz.



Σχήμα 2.4 Μη ομοιόμορφη filterbank

•Υβριδικά κατανεμημένη *filterbank*. Παράδειγμα είναι η Mel Scale κλίμακα Στην περίπτωση αυτή, η κλίμακα μεταβολής του εύρους είναι γραμμική μέχρι τα 1000Hz και κατόπιν γίνεται λογαριθμική. Η Mel κλίμακα χρησιμοποιεί τριγωνικά φίλτρα των οποίων το εύρος μεταβάλλεται ως προς την κεντρική συχνότητα (Σχ. 2.5). Από διάφορες μελέτες, έχει βρεθεί ότι χρησιμοποιώντας τη Mel κλίμακα έχουμε τα καλύτερα αποτελέσματα.



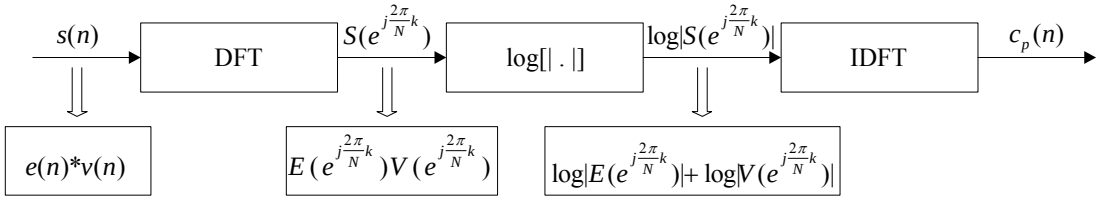
Σχήμα 2.5 Mel scale κλίμακα

Κατόπιν θ'αναφερθούμε στην επεξεργασία που υφίσταται το φάσμα προκειμένου να εξαχθούν οι παράμετροι του φωνητικού σήματος.

2.2.2 Η ανάλυση Cepstrum

Ας θεωρήσουμε την παραγωγή φωνής σαν ένα γραμμικό χρονικά μεταβαλλόμενο σύστημα s(n)=v(n)*e(n) ή στο πεδίο της συχνότητας *S(z)=E(z)V(z)* όπου V είναι η συνάρτηση μεταφοράς του φωνητικού σωλήνα και Ε είναι η διέγερση. Έχουμε ήδη αναφέρει (**Κεφάλαιο 1**) ότι η διέγερση είναι είτε μια κρουστική παλμοσειρά (εύφωνος ήχος), είτε ένα σήμα θορύβου (άφωνος ήχος). Το σήμα αυτό στο πεδίο της συχνότητας δρα πολλαπλασιαστικά στη συνάρτηση μεταφοράς του φωνητικού σωλήνα. Μας ενδιαφέρει να έχουμε παραμετρικές αναπαραστάσεις οι οποίες να έχουν άμεση σχέση με το αργά μεταβαλλόμενο μοντέλο του φωνητικού σωλήνα (άρα χαμηλό *bit rate*), να διαχωρίζουν τους εύφωνους από τους άφωνους ήχους, αλλά και να αγνοούν την επίδραση της τονικής περιόδου που υπεισέρχεται από τη διέγερση (εύφωνη διέγερση). Πρέπει λοιπόν να διαχωρίσουμε επιτυχώς τη διέγερση *e(n)* από το μοντέλο του σωλήνα *v(n)*. Αυτό γίνεται με την ανάλυση *cepstrum* που ανήκει στην ομομορφική επεξεργασία σήματος.

Η ανάλυση *cepstrum* αναπαριστά το λογάριθμο του πλάτους του φάσματος (ισοδύναμα την ισχύ του φάσματος μιας και η ισχύς είναι το τετράγωνο του πλάτους) αντί για το ίδιο το φάσμα. Οποιοδήποτε είδος κέρδους (η διέγερση στη συγκεκριμένη περίπτωση) εφαρμοστεί στο λογαριθμικό φάσμα V(z), δεν αλλάζει τη μορφή του φάσματος αλλά προσθέτει μια **DC** συνιστώσα σ'αυτό. Έτσι ο πολλαπλασιασμός της διέγερσης με το φωνητικό μοντέλο στο γραμμικό φάσμα, ανάγεται σε πρόσθεση στο λογαριθμικό φάσμα και συνεπώς καθίσταται πιο εύκολος ο διαχωρισμός τους.



Σχήμα 2.6 Εξαγωγή των συντελεστών cepstrum

Οι σχέσεις που μας δίνουν το *cepstrum* δεδομένου του Διακριτού Μετασχηματισμού Fourier (Discrete Fourier Transform(DFT)) του σήματος φωνής, $S_p(k)$ (αντιστοιχεί στη συνάρτηση μεταφοράς S(z) δειγματοληπτημένη στο μοναδιαίο κύκλο από N σημεία δηλ., $z = e^{j\frac{2\pi}{N}k}$ $k=0,\ldots,N-1$) είναι οι εξής:

$$S_p(k)= \sum_{n=0}^{N-1} s(n)e^{-j\frac{2\pi}{N}kn} \qquad 0\le k \le N-1 \tag{2-2-2.1}$$

$$\widehat{S}_p(k) = \log\left[S_p(k)\right] \qquad 0\le k \le N-1. \tag{2-2-2.2}$$

Τελικά υπολογίζουμε το *cepstrum* σαν το αντίστροφο DFT του λογαρίθμου της φασματικής ισχύος (μέτρο):

$$c_p(n)= \frac{1}{N}\sum_{n=0}^{N-1} \log|S_p(k)|e^{j\frac{2\pi}{N}kn} \quad 0\le n \le N-1. \tag{2-2-2.3}$$

Παρατηρώντας ότι το λογαριθμικό φασματικό πλάτος είναι μια πραγματική συμμετρική συνάρτηση, η παραπάνω εξίσωση μπορεί να απλοποιηθεί σε έναν διακριτό μετασχηματισμό συνημιτόνου.

Το μέρος του cepstrum που αντιστοιχεί σε μικρότερο πλάτος οφείλεται στη συνεισφορά του αργά μεταβαλλόμενου μοντέλου του φωνητικού σωλήνα. Έτσι, μπορούμε να διαχωρίσουμε εύκολα το *e(n)* από το *v(n)*. Ο πρώτος όρος c(0) έχει και το μεγαλύτερο πλάτος. Αυτό συμβαίνει επειδή ο c(0) είναι η μέση τιμή του λογαριθμικού φασματικού πλάτους δηλαδή εκφράζει την ενέργεια του σήματος φωνής. Οι υπόλοιποι συντελεστές *cepstrum* είναι ανεξάρτητοι από την ενέργεια της φωνής. Ο συντελεστής c(1) είναι ένα μέτρο της ισορροπίας στο πλάτος μεταξύ των χαμηλών και υψηλών συχνοτήτων. Οι υπόλοιποι πρώτοι cepstrum συντελεστές χαρακτηρίζουν την κυματομορφή πλάτους (*envelope*) του φάσματος της φωνής.

Το *Mel-cepstrum* αποτελεί μια παραλλαγή του *cepstrum* και συνδυάζει τον υπολογισμό των παραμέτρων *cepstrum* με τη μέθοδο των σετ φίλτρων που παρουσιάσαμε προηγουμένως. Οι συντελεστές *Mel-cepstrum* υπολογίζονται βάσει της σχέσης:

$$MFCC_i = \sum_{k=1}^{Q} X_k \cos\left[i\left(k - \frac{1}{2} \right) \frac{\pi}{20} \right], \qquad i = 1,2,\ldots,M \qquad (2\text{-}2\text{-}2.4)$$

όπου M είναι ο επιθυμητός αριθμός των cepstrum συντελεστών, Q ο αριθμός των φίλτρων στη Mel scale κλίμακα και X_k, $k=1,\ldots,Q$, είναι ο **λογάριθμος** της ενέργειας (πλάτους) του κ-οστού φίλτρου του σετ. Στην περίπτωση αυτή το φάσμα εξομαλύνεται εξαιτίας της λογαριθμικής φύσης του σετ φίλτρων.

Τα διανύσματα παραμέτρων *cepstrum* που έχουμε δει μέχρι στιγμής, υπολογίζονται για κάθε πακέτο (*frame*) φωνής και γι' αυτό ονομάζονται και διανύσματα βραχυπρόθεσμων παραμέτρων. Μας ενδιαφέρει όμως η πληροφορία για περισσότερα του ενός *frame* φωνής, που ονομάζονται διανύσματα μακροπρόθεσμων παραμέτρων. Ο πιο δημοφιλής τρόπος υπολογισμού των μακροπρόθεσμων παραμέτρων είναι με τη μέθοδο των διαφορών *(delta cepstrum)*.

Ο όρος παράμετρος δέλτα (*delta feature*) γενικά έχει σχέση με χρονική παράγωγο και μας δίνει πληροφορία σχετικά με τις αλλαγές του σήματος φωνής ως προς το χρόνο. Ο πιο απλός τρόπος αναπαράστασης δυναμικής πληροφορίας είναι παίρνοντας τη διαφορά μεταξύ παραμετρικών διανυσμάτων συνεχόμενων πακέτων (*frames*). Όμως με τις απλές διαφορές υπάρχει πρόβλημα όταν έχουμε τυχαίες αλλαγές μεταξύ διαδοχικών πακέτων. Στην περίπτωση αυτή μπορούμε να χρησιμοποιήσουμε προβλεπτικές (*regression*) μεθόδους όπου προσεγγίζουμε το διαφοριστή με κάποιο φίλτρο γραμμικής φάσης παίρνοντας έτσι εκτιμήσεις πάνω σε πέντε ή επτά συνεχόμενα πακέτα. Κάποια συστήματα παίρνουν υπόψη τους και μεταβολές του ρυθμού μεταβολής, δηλαδή παραγώγους δεύτερης τάξης. Αυτό επιτυγχάνεται εφαρμόζοντας τη μέθοδο των διαφορών στις δέλτα παραμέτρους (είτε με διαφορές, είτε με *autoregression* φίλτρα) και χρειάζονται ακόμα περισσότερα *frames* για τον υπολογισμό τους.

2.2.3 Εξαγωγή συντελεστών στο DECIPHER

Η διαδικασία εξαγωγής παραμέτρων από το *front-end* του **DECIPHER** έχει ως εξής: Το αρχικό ψηφιακό σήμα φωνής περνά από ένα υψιπερατό φίλτρο προέμφασης που στόχο έχει να ανεβάσει το πλάτος της περιοχής των υψηλών συχνοτήτων που είναι γενικά εξασθενημένο λόγω της χειλικής επίδρασης και να

καταστήσει το φάσμα της φωνής επίπεδο (*flat*). Το σήμα που προκύπτει χωρίζεται σε πακέτα των 10 ms με τη χρήση αλληλοεπικαλυπτόμενων παραθύρων Hamming εύρους 25 ms, και κατόπιν για το κάθε πακέτο υπολογίζεται ο Διακριτός Μετασχηματισμός Fourier (DFT). Κατόπιν οι συντελεστές Fourier εξομαλύνονται περνώντας από ένα σετ 25 Mel τριγωνικών φίλτρων και το πλάτος της εξόδου του κάθε φίλτρου λογαριθμίζεται για να ακολουθήσει η ανάλυση *cepstra*.

Οι λογαριθμικοί συντελεστές-έξοδοι του σετ φίλτρων περνούν κατόπιν από ένα διακριτό μετασχηματισμό συνημιτόνου (2-2-2.4) και έτσι προκύπτουν τελικά οι συντελεστές Mel-cepstrum $MFCC_i$, i=0,...,K, όπου $MFCC_0$ είναι η ενέργεια του φάσματος. Αυτό έχει σαν αποτέλεσμα να συμπιεστεί η πληροφορία του φάσματος στους K+1 συντελεστές *cepstrum* οι οποίοι αποτελούν και το βασικό ακουστικό παραμετρικό διάνυσμα. Η τιμή του K εξαρτάται από το αν έχουμε τηλεφωνικά δεδομένα ή δεδομένα υψηλής ποιότητας. Στην περίπτωση μας έχουμε high quality δεδομένα K=12 (F_s=16KHz). Έτσι, το βασικό παραμετρικό διάνυσμα έχει διάσταση K+1=13.

Πέρα από τους στατικούς(βραχυπρόθεσμους) *cepstrum* συντελεστές που υπολογίζονται για κάθε πακέτο, υπολογίζονται και δυναμικοί (μακροπρόθεσμοι) πρώτης και δεύτερης τάξης πάνω σε ένα φίλτρο παραθύρου (*regression filter*) που καλύπτει τα 2 προηγούμενα και τα 2 επόμενα πακέτα για το συγκεκριμένο πακέτο. Με τον υπολογισμό των μακροπρόθεσμων παραμέτρων διαμορφώνεται το τελικό ακουστικό διάνυσμα που αποτελείται από 3x13=39 συντελεστές:

$$\mathbf{y}=\left[c_0,c_1,....c_{12},\Delta c_0,\Delta c_1,....\Delta c_{12},\Delta^2 c_0,\Delta^2 c_1,.....\Delta^2 c_{12}\right]$$

Σχήμα 2-7 Ακουστικό διάνυσμα

2.3 Ακουστικά Μοντέλα

Ο σκοπός των ακουστικών μοντέλων είναι να παρέχουν μία μέθοδο υπολογισμού της πιθανότητας κάθε ακολουθίας διανυσμάτων ***X*,** δεδομένης μιας λέξης *W*. Έτσι η απαιτούμενη κατανομή πιθανότητας μπορεί να προσδιοριστεί με την εύρεση πολλών παραδειγμάτων της κάθε λέξης *W* και τη συλλογή στατιστικών των αντίστοιχων ακολουθιών διανυσμάτων. Στην πράξη όμως,

αυτό είναι ανέφικτο να γίνει σε πραγματικό χρόνο για συστήματα με μεγάλα λεξικά (*large vocabulary systems*), αντίθετα προτιμούμε οι ακολουθίες λέξεων να αποσυντίθενται σε βασικούς ήχους, τα φωνήματα. Όπως είδαμε και στην Παράγραφο 1.2 η αντιστοίχηση μεταξύ λέξεων και των προφορών τους που ορίζεται με ακολουθίες φωνημάτων είναι απαραίτητη για το μηχανισμό αναγνώρισης.

Μέχρι τώρα θεωρήθηκε ότι μόνο ένα **HMM** απαιτείται για κάθε φώνημα, και καθώς προσεγγιστικά 45 φωνήματα απαιτούνται για γλώσσες όπως Ελληνικά και Αγγλικά, μπορεί να νομίσει κάποιος ότι αρκεί να εκπαιδευτούν αυτά τα 45 **HMM**s. Στην πράξη, ωστόσο, η επίδραση από τα συμφραζόμενα μπορεί να οδηγήσει σε μεγάλη ποικιλία τον τρόπο με τον οποίο διαφορετικοί ήχοι μπορούν να παραχθούν. Αυτό σημαίνει ότι, για να επιτευχθεί καλή φωνητική διακριτοποίηση, πρέπει να εκπαιδευτούν διαφορετικά **HMM**s για κάθε διαφορετική λέξη. Η πιο διαδεδομένη προσέγγιση στο πρόβλημα είναι η χρήση των ***triphones***, όπου κάθε φώνημα έχει ένα διαφορετικό **HMM** μοντέλο για κάθε μοναδικό ζεύγος από αριστερούς και δεξιούς γείτονες. Για παράδειγμα, υποθέτουμε ότι ο συμβολισμός x[y]z αναπαριστά την εμφάνιση του φωνήματος y μετά από ένα x και πριν από ένα z.

Η χρήση των **μειγμάτων** (γραμμικών συνδυασμών) **Γκαουσιανών κατανομών εξόδου** (*Gaussian mixture output distributions*) επιτρέπει κάθε κατανομή κατάστασης να μοντελοποιηθεί με μεγάλη ακρίβεια. Ωστόσο, όταν χρησιμοποιούνται *triphones*, το αποτέλεσμα είναι ότι το σύστημα έχει πάρα πολλές παραμέτρους που πρέπει να εκπαιδευτούν. Για παράδειγμα, ένα σύστημα με μεγάλο λεξικό θα χρειαστεί τυπικά περίπου 60.000 *triphones* (με 45 φωνήματα υπάρχουν 45^3 =91.125 πιθανά *triphones* τα οποία όμως δεν μπορούν να εμφανιστούν στην πράξη όλα, λόγω φωνητικών περιορισμών στη γλώσσα). Ωστόσο, η συχνότητα εμφάνισης τους σχετίζεται με τις λέξεις που χρησιμοποιούνται στη γλώσσα, οι οποίες δεν κάνουν ισοκατανεμημένη χρήση των ήχων. Η λύση στην αραιότητα των δεδομένων είναι η χρήση της μεθόδου ***back-off***. Με τον όρο ***back-off*** εννοούμε την διαδικασία αντικατάστασης της πιθανότητας ενός *triphone* από την πιθανότητα ενός **biphone** (συνδυασμός ενός

φωνήματος ή με το προηγούμενο ή με το επόμενό του) στην περίπτωση που υπάρχουν πάρα πολύ λίγα triphones.

2.4 Γλωσσικά Μοντέλα

Ο σκοπός του γλωσσικού μοντέλου είναι να δώσει ένα μηχανισμό εκτίμησης της πιθανότητας $P(\mathbf{W})$ μιας ακολουθίας γλωσσικών μονάδων (λέξεις ή φωνήματα ανάλογα με το σύστημα) $\mathbf{W} = \{\mathbf{w}_1,....,\mathbf{w}_n\}$. Η πιθανότητα αυτή μπορεί να γραφεί:

$$P(\mathbf{W}) = P(\mathbf{w}_1,....,\mathbf{w}_n) = \prod_{i=1}^{n} P(\mathbf{w}_i|\mathbf{w}_0,....,\mathbf{w}_{i-1}) \qquad (2\text{-}4.1)$$

όπου η $\mathbf{w}_0$ επιλέγεται κατάλληλα για να ικανοποιεί την αρχική συνθήκη. Η πιθανότητα της επόμενης γλωσσικής μονάδας εξαρτάται από όλες τις προηγούμενες που έχουν ειπωθεί. Παρόλ' αυτά, η πολυπλοκότητα του μοντέλου αυξάνεται εκθετικά με το μέγεθος του παραθύρου των προηγούμενων λέξεων που κοιτάμε. Έτσι χρησιμοποιούνται τα λεγόμενα N-grams τα οποία κωδικοποιούν ταυτόχρονα το συντακτικό και σημαντική ανάλυση (semantics) και επικεντρώνονται στις τοπικές εξαρτήσεις της ομιλίας. Τα N-grams με N=3 έχει αποδειχθεί ότι δίνουν την καλύτερη απόδοση σε σχέση με το υπολογιστικό κόστος και ονομάζονται trigrams. Η πιθανότητα $P(\mathbf{W})$ βρίσκεται από τη σχέση:

$$P(\mathbf{W}) = \prod_{i=1}^{n} P(\mathbf{w}_i|\mathbf{w}_{i-2},\mathbf{w}_{i-1}) \qquad (2\text{-}4.2)$$

Η εκτίμηση των πιθανοτήτων trigram γίνεται με τη χρήση ενός μεγάλου αρχείου κειμένου. Το γεγονός ότι έχουμε αρχεία κειμένου και όχι φωνής αποτελεί πλεονέκτημα, διότι έτσι δεν υπάρχει απαίτηση για αυστηρούς γλωσσικούς κανόνες. Στην ουσία αυτό που υπολογίζεται είναι απλά συχνότητες εμφάνισης: $f(\mathbf{w}_3|\mathbf{w}_1,\mathbf{w}_2) = \frac{c_{123}}{c_{12}}$ όπου c_{123} είναι ο αριθμός εμφανίσεων του trigram $\{\mathbf{w}_1,\mathbf{w}_2,\mathbf{w}_3\}$ και c_{12} ο αριθμός εμφανίσεων του bigram $\{\mathbf{w}_1,\mathbf{w}_2\}$. Για ένα μέγεθος λεξιλογίου V, υπάρχουν V^3 trigrams! Πολλά από αυτά τα trigrams δεν υπάρχουν στο αρχείο κειμένου με αποτέλεσμα να έχουμε μηδενικές πιθανότητες γι' αυτά αν χρησιμοποιήσουμε την απλή μέθοδο των συχνοτήτων εμφάνισης. Το

πρόβλημα αυτό αντιμετωπίζεται με γραμμική παρεμβολή των συχνοτήτων, ή με άλλους αλγορίθμους εκτίμησης των πιθανοτήτων αυτών.

2.5 Ο αποκωδικοποιητής

Όπως είδαμε και στην εισαγωγή, το πρόβλημα της αναγνώρισης φωνής συνοψίζεται στην εξίσωση.

$$\hat{W} = \arg\max_{\underline{w}} P(\underline{W} \mid \underline{X}) = \arg\max_{\underline{w}} \frac{P(\underline{W})P(\underline{X} \mid \underline{W})}{P(\underline{X})} \qquad (2\text{-}5.1)$$

Μπορούμε να φανταστούμε ένα δενδρικό δίκτυο έτσι ώστε σε κάθε αρχικό κόμβο να υπάρχει κάθε πιθανή αρχική λέξη. Όλες οι πρώτες λέξεις ενώνονται με όλες τις λέξεις που πιθανά μπορεί να ακολουθούν και πάει λέγοντας. Θεωρητικά αν αυτό το δέντρο επεκταθεί αρκετά βαθιά μπορεί να αναπαραστήσει όλες τις πιθανές ακολουθίες λέξεων **W**. Κανονικά το σύστημα θα έπρεπε να αναζητήσει τη βέλτιστη ακολουθία $\hat{W}$ πάνω σε όλες τις πιθανές ακολουθίες λέξεων **W**. Όπως είναι εύκολα αντιληπτό, κάτι τέτοιο είναι υπολογιστικά ακριβό (εώς αδύνατο) αν λάβουμε υπόψη μας τη μεγάλη ποσότητα των γλωσσικών παραμέτρων και ότι έχουμε περιορισμούς πραγματικού χρόνου. Ο αποκωδικοποιητής χρησιμοποιεί τα ακουστικά και γλωσσικά μοντέλα για να βρει τις πιθανότερες ακολουθίες λέξεων **W** που θα χρησιμοποιηθούν για να βρεθεί η τελική ακολουθία λέξεων $\hat{W}$. Στην ουσία πρόκειται για ένα πρόβλημα αναζήτησης και συνήθως χρησιμοποιούνται διάφορες τεχνικές δενδρικής αναζήτησης.Βέβαια πιο γνωστός από όλους τους αλγορίθμους αναζήτησης είναι ο αλγόριθμος Viterbi.

Το σύστημα DECIPHER χρησιμοποιεί τεχνικές προοδευτικής αναζήτησης *(progressive search techniques)*, μία μέθοδο αναζήτησης πολλαπλών βημάτων (*multiple-pass*). Στην πράξη τα συστήματα με μεγάλα λεξιλόγια είναι πολύπλοκα και απαιτείται ένα είδος απόρριψης (pruning) ώστε ο χώρος αναζήτησης να περιορίζεται σταδιακά. Κάθε βήμα λοιπόν καταλήγει στη δημιουργία ενός λεκτικού πλέγματος (*word-lattice*) το οποίο χρησιμεύει σαν γραμματική για το επόμενο βήμα περιορίζοντας το χώρο αναζήτησης (*search space*). Σε κάθε βήμα χρησιμοποιούνται προοδευτικά πιο ακριβείς αλγόριθμοι αναζήτησης, οι οποίοι όμως εφαρμόζονται σε πιο μικρούς χώρους αναζήτησης.

ΚΕΦΑΛΑΙΟ 3

ΑΝΑΓΝΩΡΙΣΗ ΦΩΝΗΣ ΜΕ ΤΕΧΝΙΚΕΣ ΚΑΝΟΝΙΚΟΠΟΙΗΣΗΣ

3.1 Εισαγωγή

Είναι προφανής η ανάγκη να διατηρηθεί η ακρίβεια αναγνώρισης στην περίπτωση που το σήμα φωνής παραμορφώνεται ή στην περίπτωση που τα ακουστικά, φωνητικά χαρακτηριστικά του περιβάλλοντος εκπαίδευσης διαφέρουν από αυτά του τεστ περιβάλλοντος. Τα εμπόδια που υπάρχουν περιλαμβάνουν ακουστικές παραμορφώσεις που προέρχονται από προσθετικό θόρυβο, από επίδραση γραμμικού φιλτραρίσματος, από μη γραμμικότητα στη μεταγωγή ή μεταφορά, καθώς και παρουσία πηγών θορύβου μεγάλης έντασης.

Οι διαφορές από ομιλητή σε ομιλητή είναι μια άλλη μορφή μεταβλητότητας, η οποία έχει να κάνει με αλλαγές στο ρυθμό ομιλίας, συνάρθρωση, συμφραζόμενα και διάλεκτο. Συστήματα τα οποία έχουν σχεδιαστεί να είναι ανεξάρτητα από ομιλητή, παρουσιάζουν πολλές φορές δραματικές πτώσεις σε ακρίβεια αναγνώρισης όταν οι συνθήκες εκπαίδευσης και χρήσης διαφέρουν. Στην περίπτωσή μας, μια από τις κύριες αιτίες μείωσης της επίδοσης σε συστήματα αναγνώρισης συνεχούς ομιλίας είναι η μεταβλητότητα του μήκους του φωνητικού σωλήνα. Σύμφωνα με **το Κεφάλαιο 1** η συνάρτηση μεταφοράς ενός στοιχειώδους μοντέλου του φωνητικού σωλήνα είναι

$$V(j\Omega) = 1 / \cos(\Omega l / c) \qquad (3\text{-}1.1)$$

όπου l είναι το μήκος του σωλήνα, c η ταχύτητα του ήχου και Ω η συχνότητα. Στο πεδίο της συχνότητας έχουμε έναν άπειρο αριθμό πόλων στον jΩ άξονα στις συχνότητες $F_i = \frac{(2i-1)c}{4l} = 500,1500,2500,....Hz$. Αυτές αποτελούν τις συχνότητες συντονισμού του σωλήνα χωρίς απώλειες. Οι συχνότητες συντονισμού εξαρτώνται από το μήκος του φωνητικού σωλήνα (αντιστρόφως ανάλογες). Για μεγαλύτερο μήκος του φωνητικού σωλήνα (άντρες) έχουμε μικρότερες συχνότητες συντονισμού, το αντίστροφο ίσχυει για τις

γυναίκες.Για παράδειγμα αν υποθέσουμε ένα φωνητικό σωλήνα με μήκος L τότε κάθε formant frequency θα είναι ανάλογο του 1/L. 'Ετσι ενώ το μήκος του φωνητικού σωλήνα κυμαίνεται από 13cm(γυναίκες) σε 18cm(άντρες) ταυτόχρονα οι formant frequencies διαφέρουν μέχρι και 25% μεταξύ των ομιλητών.

Συγκεκριμένα ενώ στους άντρες ψάχνουμε για 4 formant frequencies στο πεδίο 300-3300 Hz στις γυναίκες ψάχνουμε για 3 formant frequencies στο ίδιο πεδίο. Ως formant frequency ορίζουμε τη συχνότητα που μεταφέρει το μεγαλύτερο μέρος της ακουστικής ενέργειας από την πηγή στην έξοδο. Αν λοιπόν αλλάξουμε το μήκος του φωνητικού σωλήνα από L σε KL τότε έχουμε ένα scaling του άξονα της συχνότητας κατά 1/K.

Υπάρχουν διάφορες μέθοδοι που χρησιμοποιούνται για την επίλυση των προβλημάτων που αναφέρθηκαν. Η πρώτη μέθοδος περιλαμβάνει τεχνικές κανονικοποίησης, όπου τα χαρακτηριστικά ενός ομιλητή μετασχηματίζονται έτσι ώστε να ταιριάζουν καλύτερα με αυτά ενός πρότυπου ομιλητή (speaker normalization). Η δεύτερη μέθοδος επιδρά στο μοντέλο του αναγνωριστή και τον προσαρμόζει στα καινούργια δεδομένα, είτε αυτά προέρχονται από θορυβώδες κανάλι, είτε από κάποιο διαφορετικό ομιλητή.

3.2 Τεχνικές κανονικοποίησης στον ομιλητή (speaker normalization)

'Οπως προαναφέρθηκε, η επίδοση ενός συστήματος αναγνώρισης συνεχούς ομιλίας μπορεί να βελτιωθεί, όταν τα χαρακτηριστικά βραχέως χρόνου (short time features) ενός ομιλητή μετασχηματίζονται έτσι ώστε να ταιριάζουν καλύτερα με αυτά ενός πρότυπου ομιλητή (speaker normalization).

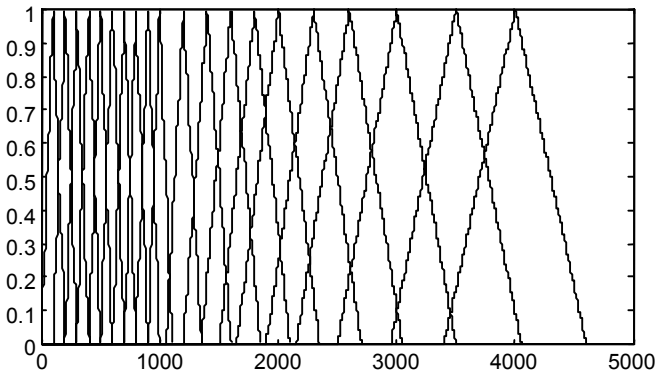
Η κανονικοποίηση του μήκους του φωνητικού σωλήνα μέσω της συχνότητας (vocal tract length normalization via frequency warping) είναι μια δημοφιλής τεχνική, όπου ο άξονας της συχνότητας επεκτείνεται ή συρρικνώνεται πριν την εξαγωγή των cepstral συντελεστών κατά τη διάρκεια επεξεργασίας του σήματος της φωνής. 'Ετσι με την κανονικοποίηση επιλέγεται ένας συντελεστής ανά ομιλητή ή ανά πρόταση προκειμένου να επιτευχθεί καλύτερη απεικόνιση των ακουστικών χαρακτηριστικών του ομιλητή και κατά συνέπεια καλύτερη επίδοση αναγνώρισης. Για τις γυναίκες χρησιμοποιούνται συντελεστές>=1 ενώ

οι άντρες παρουσιάζουν συντελεστές <=1. Η διαδικασία αυτή εφαρμόζεται στο front-end.

3.2.1 Εξαγωγή cepstral συντελεστών

Το ψηφιακό σήμα φωνής s(n) περνά πρώτα από προέμφαση μετά από ένα παράθυρο (preemphasize end window) και τέλος από ένα σετ Q τον αριθμό ζωνοδιαβατών φίλτρων τα οποία καλύπτουν το εύρος συχνοτήτων του σήματος εισόδου που μας ενδιαφέρει (π.χ. 100-3000 Hz για σήματα τηλεφωνικής ποιότητας και 100-8000 Hz για σήματα υψηλής ποιότητας). Τα φίλτρα μπορεί να αλληλοκαλύπτονται μεταξύ τους στη συχνότητα. Η έξοδος του i-οστού ζωνοδιαβατού φίλτρου, είναι η φασματική αναπαράσταση βραχέως χρόνου του σήματος s(n) τη χρονική στιγμή n.

Χρησιμοποιούνται μη ομοιόμορφα κατανεμημένα σετ φίλτρων. Συνήθως οι κεντρικές συχνότητες των φίλτρων τοποθετούνται σε λογαριθμική κλίμακα επειδή το ανθρώπινο αυτί επιδεικνύει ανάλογη απόκριση. Η Mel scale κλίμακα χρησιμοποιεί τριγωνικά φίλτρα των οποίων το εύρος μεταβάλλεται ως προς την κεντρική συχνότητα (Σχ. 3.1) και η κλίμακα μεταβολής του εύρους είναι γραμμική μέχρι τα 1000Hz και κατόπιν γίνεται λογαριθμική.

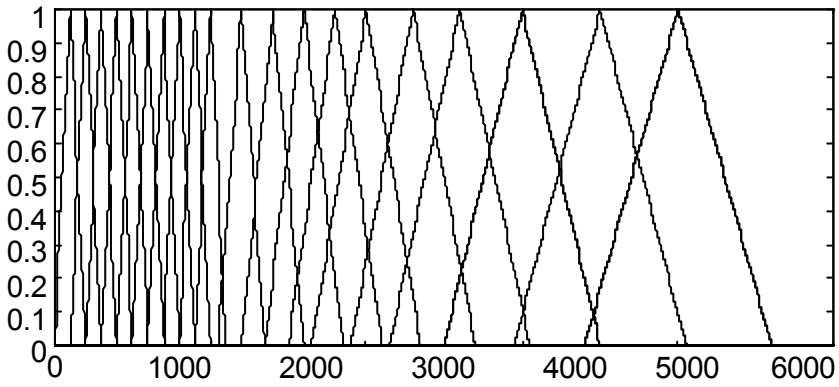


Σχήμα 3.1 Mel φίλτρα

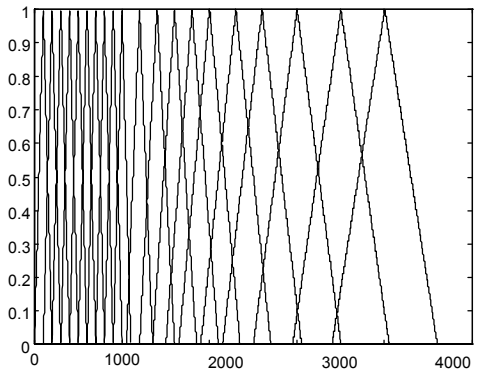
Έπειτα υπολογίζουμε το λογάριθμο του πλάτους του φάσματος (ισοδύναμα την ισχύ του φάσματος μιας και η ισχύς είναι το τετράγωνο του πλάτους) αντί

για το ίδιο το φάσμα. Επειδή το λογαριθμικό φασματικό πλάτος είναι μια πραγματική συμμετρική συνάρτηση, οι συντελεστές cepstrum προκύπτουν με ένα διακριτό μετασχηματισμό συνημιτόνου.

Τροποποιούμε λοιπόν τα φίλτρα, μεταβάλλοντας τις κεντρικές συχνότητες και το εύρος τους οπότε έχουμε ένα scaling του άξονα της συχνότητας. 'Ετσι ο άξονας της συχνότητας επεκτείνεται ή συρρικνώνεται πριν την εξαγωγή των cepstral συντελεστών κατά τη διάρκεια επεξεργασίας του σήματος της φωνής. Επιτυγχάνεται έτσι η κανονικοποίηση του μήκους του φωνητικού σωλήνα και η φωνή των ομιλητών φαίνεται σαν να έχει παραχθεί από ένα φωνητικό σωλήνα σταθερού μήκους. Στα παρακάτω σχήματα φαίνεται η επέκταση (σχήμα 3.2) ή η συρρίκνωση (σχήμα 3.3) του άξονα της συχνότητας.



Σχήμα 3.2 Επέκταση του άξονα της συχνότητας



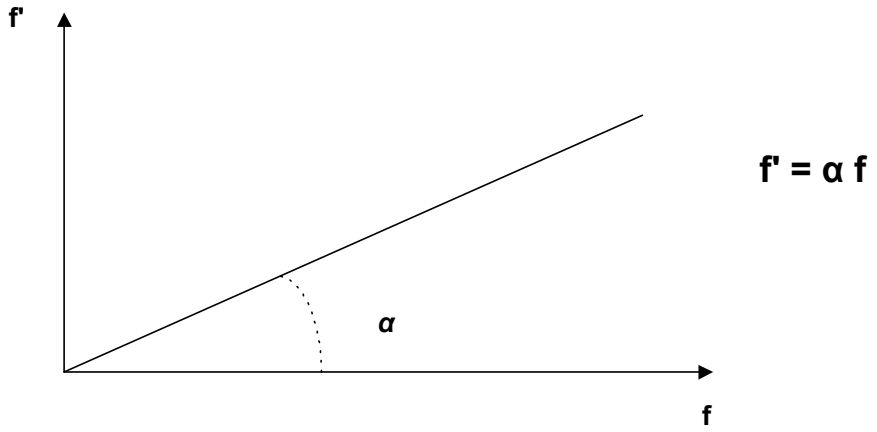
Σχήμα 3.3 Συρρίκνωση του άξονα της συχνότητας

Στη συνέχεια θα αναφερθούμε στους διάφορους μετασχηματισμούς που εφαρμόζουμε στον άξονα της συχνότητας των φίλτρων προκειμένου να επιτύχουμε την επέκταση ή συρρίκνωση του άξονα της συχνότητας και την οποία εφαρμόζουμε στο πειραματικό μέρος.

3.2.2 Διάφοροι Μετασχηματισμοί

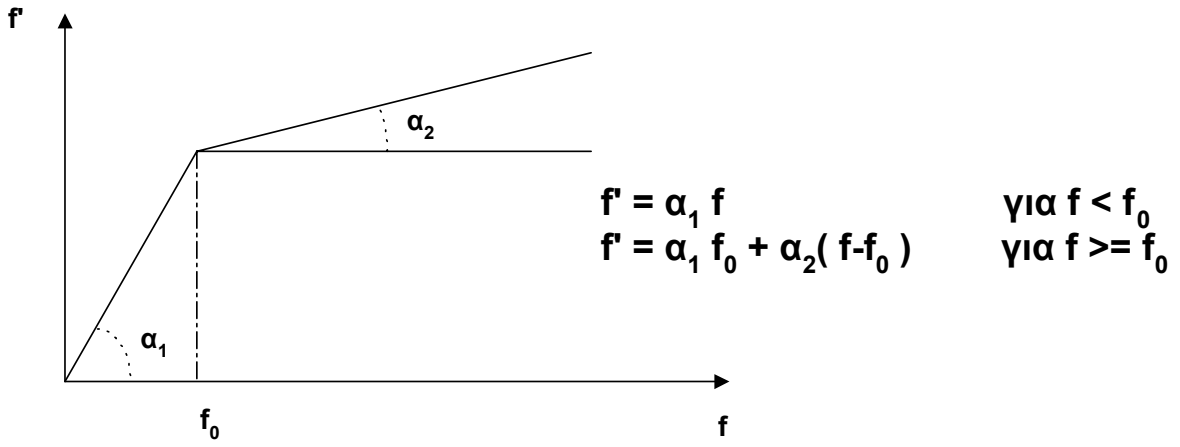
Οι μετασχηματισμοί τους οποίους εφαρμόζουμε στα φίλτρα είναι οι εξής:

•Γραμμικός μετασχηματισμός. Με τον παραπάνω μετασχηματισμό αλλάζει η κεντρική συχνότητα των φίλτρων και το εύρος τους.



Σχήμα 3.4 Γραμμικός μετασχηματισμός

•Γραμμικός μετασχηματισμός κατά περιοχές όπου ορίζουμε μια συχνότητα αποκοπής και έχουμε δύο διαφορετικούς συντελεστές, έναν συντελεστή για συχνότητες μικρότερες από την συχνότητα αποκοπής και έναν άλλο συντελεστή για συχνότητες μεγαλύτερες από την συχνότητα αποκοπής. Με τον παραπάνω μετασχηματισμό αλλάζει η κεντρική συχνότητα των φίλτρων και το εύρος τους.



Σχήμα 3.5 Γραμμικός μετασχηματισμός κατά περιοχές

•Μετασχηματισμός μεταφοράς, όπου μεταφέρουμε την κεντρική συχνότητα των φίλτρων και διατηρούμε σταθερό το εύρος των φίλτρων.

$$f_c' = \alpha f_c$$

3.3 Ολοκλήρωση του Συστήματος Αναγνώρισης

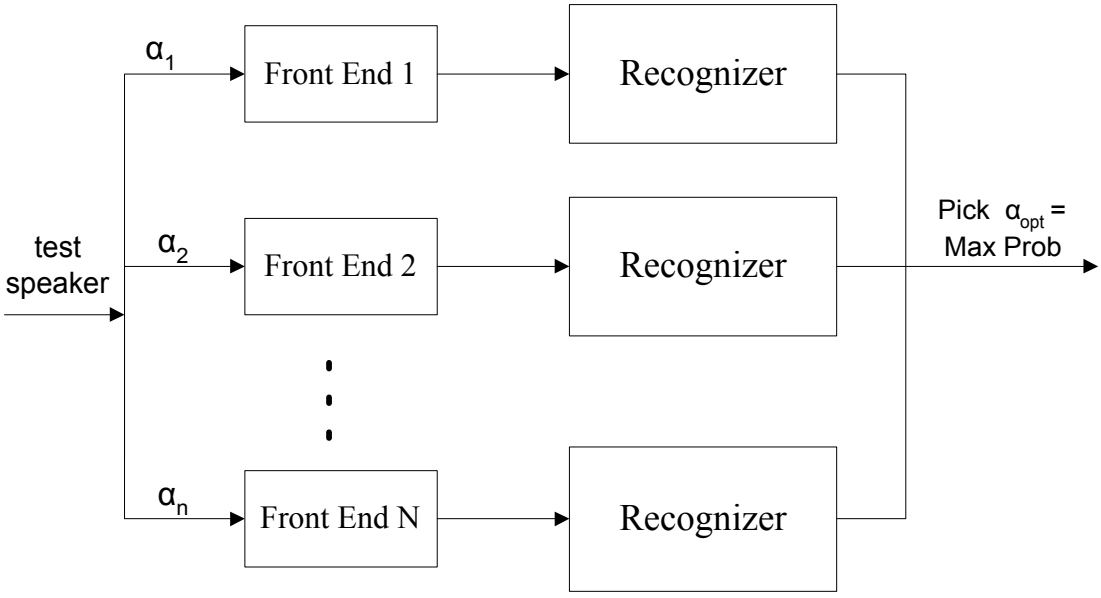
Αφού μελετήσουμε τους διάφορους μετασχηματισμούς, στη συνέχεια επιδιώκεται η ολοκλήρωση των παραπάνω εφαρμογών στο σύστημα αναγνώρισης συνεχούς ομιλίας. Αναπτύσσουμε δηλαδή αλγορίθμους για την εύρεση του βέλτιστου συντελεστή για κάθε ομιλητή στο test set για την επίτευξη καλύτερης επίδοσης αναγνώρισης. Στην περίπτωση αυτή βρίσκουμε το βέλτιστο συντελεστή α_i, βάση ενός κριτηρίου. Σκοπός μας είναι να επιλέξουμε το κατάλληλο κριτήριο και να το χρησιμοποιήσουμε στην αναγνώριση. Οι αλγόριθμοι τους οποίους εφαρμόζουμε είναι οι εξής:

- Παράλληλη Αναγνώριση
- Αναγνώριση με χρήση GMM (gaussian mixture model).

Τέλος για την επίτευξη συμβατότητας κατά την εκπαίδευση και κατά την αναγνώριση επιδιώκεται η επανεκπαίδευση των μοντέλων. Παρακάτω θα μελετήσουμε κάθε μια μέθοδο χωριστά.

3.4 Παράλληλη Αναγνώριση

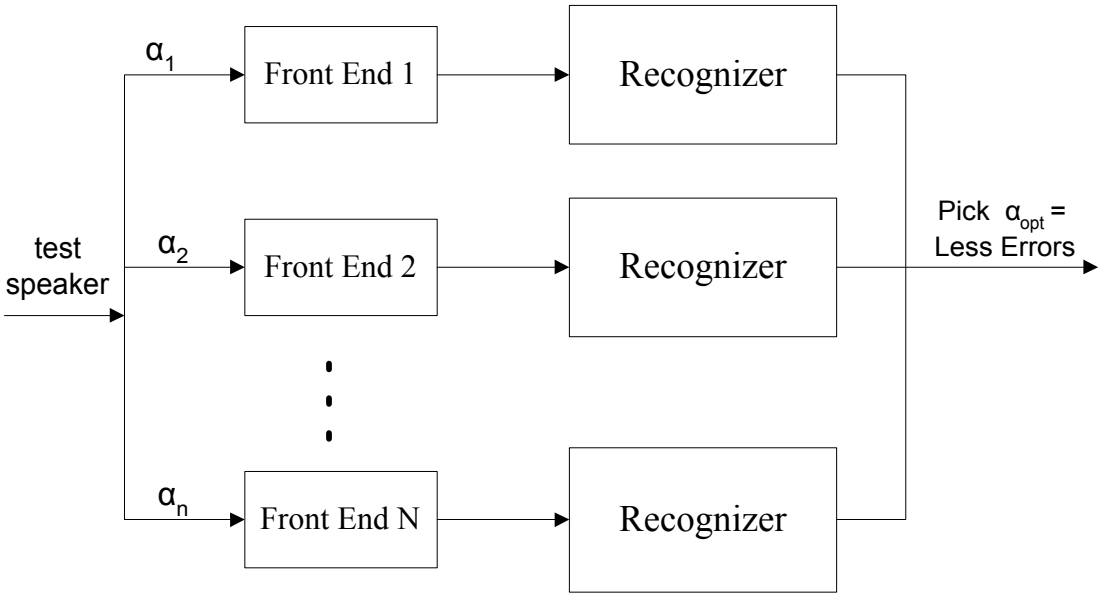
Στην περίπτωση αυτή κατά την αναγνώριση τρέχουν ταυτόχρονα πολλοί recognizers για τους διάφορους συντελεστές αi. Σκοπός μας είναι να επιλέξουμε με ένα κριτήριο κάποιον recognizer και να τον χρησιμοποιήσουμε κατά την αναγνώριση. Συνήθως βρίσκουμε τον καλύτερο συντελεστή αi χρησιμοποιώντας το πεδίο PROB (maximum likelihood). Η γραμμή PROB επιστρέφει την πιθανότητα με την οποία επιλέχθηκε η προηγούμενη απάντηση. Όσο λιγότερο αρνητική τιμή έχει αυτός ο αριθμός τόσο μεγαλύτερη είναι η πιθανότητα. Μπορούμε να επιλέξουμε το βέλτιστο συντελεστή α_i, για κάθε πρόταση ανεξάρτητα ή ένα συντελεστή για όλες τις προτάσεις του κάθε ομιλητή. Παρακάτω φαίνεται το σχήμα αναγνώρισης με το κριτήριο μέγιστης πιθανοφάνειας και παράλληλη αναγνώριση.



Σχ 3.6.1 Παράλληλη αναγνώριση με κριτήριο PROB

Μια άλλη μέθοδος είναι χρησιμοποιώντας τα πεδίο ERROR του αναγνωριστή. Υποθέτουμε ότι γνωρίζουμε (μαγικά GENIE) ποιος συντελεστής α_i, δίνει τα καλύτερα αποτελέσματα ως προς τα λάθη (βέλτιστη υπόθεση) και τον χρησιμοποιούμε στην αναγνώριση. Παρακάτω φαίνεται το σχήμα αναγνώρισης με το κριτήριο GENIE και παράλληλη αναγνώριση. Το σχήμα

αυτό φυσικά, δεν είναι πραγματοποιήσιμο, αλλά χρησιμεύει για την εκτίμηση ενός κάτω φράγματος για το σφάλμα αναγνώρισης.



Σχήμα 3.6.2 Παράλληλη αναγνώριση με κριτήριο GENIE

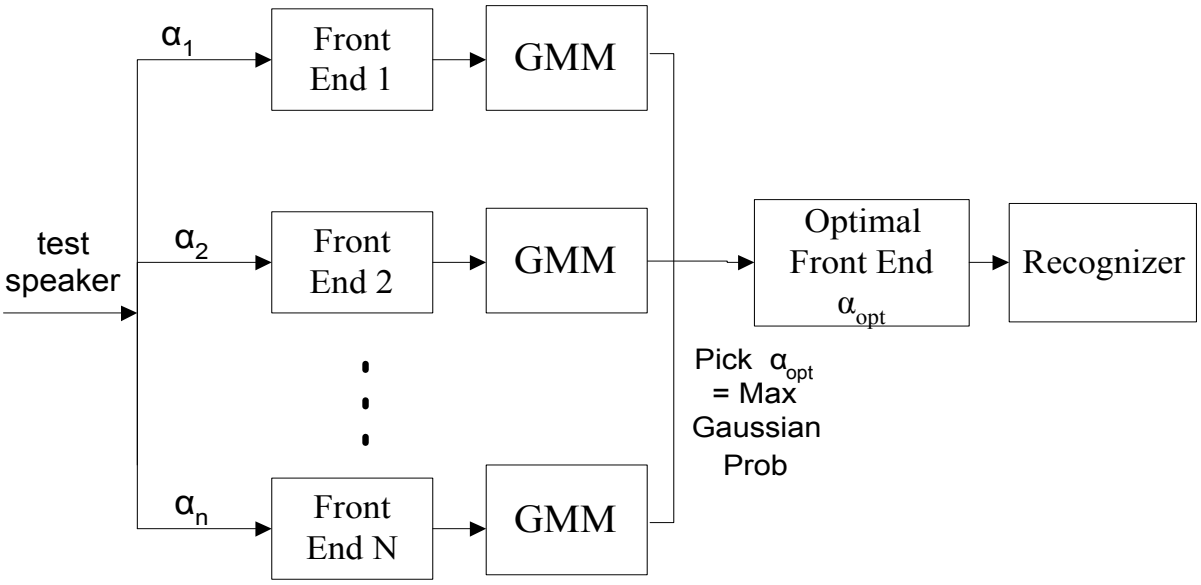
Η παραπάνω προσέγγιση έχει το μειονέκτημα ότι τρέχουν παράλληλα πολλοί recognizers(υπολογιστική σπατάλη) για τους διάφορους συντελεστές α_i και η βέλτιστη επιλογή γίνεται στο τέλος. Πρέπει λοιπόν να βρίσκουμε το βέλτιστο συντελεστή α_i πριν την αναγνώριση, με τον τρόπο αυτό χρησιμοποιούμε έναν μόνο αναγνωριστή και αποφεύγουμε την παραπάνω διαδικασία των παράλληλων recognizers.

3.5 Αναγνώριση με χρήση GMM (gaussian mixture model)

Στην περίπτωση αυτή βρίσκουμε το βέλτιστο συντελεστή α_i πριν την αναγνώριση, βάση ενός κριτηρίου μέγιστης πιθανοφάνειας (maximum likelihood), όπου οι παράμετροι, α_i, επιλέγονται ώστε να μεγιστοποιηθεί η πιθανότητα:

$$\max_{\alpha_i} P(\underline{X^{\alpha_i}} | M)$$ όπου **M** είναι το **GMM** (gaussian mixture model)

Με τον τρόπο αυτό χρησιμοποιούμε έναν μόνο αναγνωριστή. Κατασκευάζεται δηλαδή αρχικά ένα στοχαστικό μοντέλο και στην περίπτωσή μας είναι ένα μείγμα (γραμμικός συνδυασμός) γκαουσιανών **GMM**. Μπορούμε να επιλέξουμε το βέλτιστο συντελεστή α_i, για κάθε πρόταση ανεξάρτητα ή ένα συντελεστή για όλες τις προτάσεις του κάθε ομιλητή. Παρακάτω φαίνεται το σχήμα αναγνώρισης με χρήση GMM(gaussian mixture model).



Σχήμα 3.7 Αναγνώριση με χρήση GMM

Κατά την αναγνώριση χρησιμοποιούνται τα voiced frames του κάθε ομιλητή για τον υπολογισμό της πιθανότητας:

$$P(\underline{X}^{\alpha_i}|M) = \prod_{t=1}^{T} p(x_t^{\alpha_i}) \tag{3-5.1}$$

δηλαδή το γινόμενο των πιθανοτήτων όλων των voiced frames $x_t^{\alpha_i}$, της πρότασης για το συντελεστή α_i όπου

$$p(x_t^{\alpha_i}) = \sum_{m=1}^{M} w_m N(x_t^{\alpha_i}; \mu_m, \sum_m) \tag{3-5.2}$$

όπου $$\sum_{m=1}^{M} w_m = 1 \tag{3-5.3}$$

και M το πλήθος των Gaussians, w_m τα βάρη των Gaussians , $N(x_t;\mu_m,\sum_m)$ η Gaussian, με μέση τιμή μ_m και $\sum_m$ πίνακας συμμεταβλητότητας.

Με βάση το κριτήριο $\max_{\alpha_i} P(\underline{X^{\alpha_i}}|M)$ βρίσκουμε το βέλτιστο συντελεστή α_i και προχωρούμε έπειτα στην αναγνώριση. Για τους παραπάνω υπολογισμούς χρειάζεται το GMM (Gaussian Mixture Model) το οποίο πρέπει να το εκπαιδεύσουμε. Παρακάτω περιγράφεται η διαδικασία εκπαίδευσης.

3.6 Εκπαίδευση του GMM (gaussian mixture model)

Η πιο συνήθης προσέγγιση είναι αυτή της μέγιστης πιθανοφάνειας (maximum likelihood), όπου οι παράμετροι, α_i, επιλέγονται ώστε να μεγιστοποιηθεί η πιθανότητα:

$$\max_{\alpha_i} P(\underline{X^{\alpha_i}}|M) \qquad (3\text{-}6.1)$$

Για τον υπολογισμό της πιθανότητας ισχύουν οι προηγούμενοι τύποι. Καταλήγουμε, λοιπόν, στον αλγόριθμο, του οποίου τα βήματα του φαίνονται συνοπτικά παρακάτω.

Στάδιο I. Αρχικοποίηση: Υπολόγισε αρχικό GMM με ουδέτερες παραμέτρους (χωρίς κανονικοποίηση).

Στάδιο II. Για κάθε ομιλητή i=1,…,N. Υπολόγισε βέλτιστο warp factor α_i με βάση το κριτήριο μέγιστης πιθανοφάνειας $\max_{\alpha_i} P(\underline{X^{\alpha_i}}|M)$ σε όλες τις προτάσεις του ομιλητή.

Στάδιο III. Υπολόγισε νέο GMM με τις καινούργιες παραμέτρους για κάθε ομιλητή.

Στάδιο IV. Επανέλαβε το **Στάδιο II** και **Στάδιο III:**

Ο παραπάνω αλγόριθμος προυποθέτει ομαδοποίηση όλων των προτάσεων για κάθε ομιλητή. Επιπλέον χρησιμοποιούνται τα voiced frames του κάθε ομιλητή. Αξίζει να σημειωθεί ότι ο αλγόριθμος συγκλίνει, έπειτα από 3 επαναλήψεις έχουμε βρει το βέλτιστο warp factor α_i για κάθε ομιλητή που ανήκει στο σύνολο εκπαίδευσης. Επιπλέον η ορθότητα του αλγορίθμου επαληθεύεται από το γεγονός ότι οι γυναίκες παρουσιάζουν warp factor>=1, ενώ οι άντρες παρουσιάζουν warp factor<1.Θα πρέπει να σημειωθεί ότι κατά την πειραματική διαδικασία προέκυψαν τα καλύτερα αποτελέσματα για αριθμό Gaussians ίσο με 64.

3.7 Εκπαίδευση των αρχικών μοντέλων

Στη φάση της εκπαίδευσης λοιπόν θέλουμε να εκτιμήσουμε τις τιμές των παραμέτρων λ=(A,B,π) από την ακολουθία παρατηρήσεων $\underline{X} = [x_1, x_2, \ldots, x_T]$. Μια σειρά επαναληπτικών αλγορίθμων πρέπει να τρέξει πάνω στα δεδομένα εκπαίδευσης (*training data*), ώστε να διαθέτουμε τα απαραίτητα στοιχεία για τον υπολογισμό τους. Στην προκειμένη περίπτωση εκπαιδεύουμε τα αρχικά μοντέλα χρησιμοποιώντας τον αλγόριθμο Baum-Welch λαμβάνοντας ταυτόχρονα υπόψη τα διαφορετικά front-ends για την εξαγωγή των ακουστικών παραμέτρων. Δηλαδή οι ακουστικές παράμετροι $\underline{X^a} = [x_1^a, x_2^a, \ldots, x_T^a]$ όπου $\underline{X^a}$ οι παράμετροι του α front-end α=1..N χρησιμοποιούνται κατά την εκπαίδευση. Για την εύρεση του front-end του κάθε ομιλητή επιδιώκεται να μεγιστοποιηθεί η πιθανότητα:

$$\max_{\alpha_i} P(\underline{X^{\alpha_i}} | M) \qquad (3\text{-}7.1)$$

Ο αλγόριθμος Baum-Welch είναι εφαρμογή του αλγορίθμου Expectation Maximization σε HMMs. Πρόκειται για επαναληπτικό αλγόριθμο που σε κάθε βήμα μεγιστοποιεί την ποσότητα:

$$E\{\log P(\underline{X}, \underline{Q} | \lambda_{new}) | \underline{X}, \lambda_{old}\}, \qquad (3.7\text{-}2)$$

και υπολογίζεται πάνω σε όλες τις πιθανές ακολουθίες καταστάσεων $\underline{Q}$ διότι αυτή δεν είναι γνωστή και δεν μπορεί να μεγιστοποιηθεί απευθείας η ποσότητα:

$$\log P(\underline{X},\underline{Q}|\lambda). \qquad (3.7\text{-}3)$$

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

$$\log P(\underline{X}|\lambda), \qquad (3.7\text{-}4)$$

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

Ο αλγόριθμος μπορεί να εφαρμοστεί τόσο σε Discrete HMMs όσο και Gaussian Continuous Density HMMs. Αμέσως μετά δίνουμε τα βήματα του αλγορίθμου για Gaussian Continuous Density HMMs, που χρησιμοποιήθηκε στην εκπαίδευση των μοντέλων του συστήματος μας.

Όταν η ακολουθία καταστάσεων $\underline{Q}$ δεν είναι γνωστή, ο αλγόριθμος EM μας λέει ότι για τον υπολογισμό των νέων τιμών για τις πιθανότητες μετάβασης $\hat{a}_{kl}$ θα πρέπει να χρησιμοποιήσουμε τις πιθανότητες της τρέχουσας μετάβασης, δεδομένων των παρατηρήσεων $\underline{X}^a = [x_1^a, x_2^a, \ldots, x_T^a]$ του α *front-end* και των προηγούμενων τιμών των παραμέτρων λ.

Στάδιο I. (Αρχικοποίηση) Θέσε: $\Pi = \{\pi_i\}$, $A = \{a_{kl}\}$ και $B = \{b_i(l)\}$

Στάδιο II. (Forward-Backward) Χρησιμοποιώντας τις παραμέτρους λ=(A,B,π) και τον αλγόριθμο Forward-Backward:

Για t=1,…,T

Για i=1,…,N

Υπολόγισε: $P(x_1,,\ldots,x_t,q_t = i|\lambda) = a_t(i)$

και $P(x_{t+1},,\ldots,x_T|q_t = i,\lambda) = \beta_t(i)$

Στάδιο III. (Maximization).Υπολόγισε νέες τιμές για όλες τις παραμέτρους.

$$\hat{\mu}_i = \frac{\sum_{t+1}^{T} P(q_t = i \mid \underline{X}^a, \lambda) x_t}{\sum_{t=1}^{T} P(q_t = i \mid \underline{X}^a, \lambda)} = \frac{\sum_{t=1}^{T} \gamma_t(i) x_t}{\sum_{t=1}^{T} \gamma_t(i)}$$

$$\hat{\sum}_i = \frac{\sum_{t+1}^{T} \gamma_t(i)(x_t - \hat{\mu}_i)(x_t - \hat{\mu}_i)^T}{\sum_{t=1}^{T} \gamma_t(i)}$$

$$\gamma_t(i) = \frac{a_t(i)\beta_t(i)}{\sum_{i=1}^{N} a_t(i)\beta_t(i)}$$

$$\hat{a}_{kl} = \frac{\sum_{t=1}^{T-1} P(q_t = k, q_{t+1} = l \mid \underline{X}^a, \lambda)}{\sum_{t=1}^{T-1}\sum_{l=1}^{N} P(q_t = k, q_{t+1} = l \mid \underline{X}^a, \lambda)}$$

δηλαδή

$$a_{kl} = \frac{\text{πιθανότητα μετάβασης από την κατάσταση k στην κατάσταση l}}{\text{πιθανότητα μετάβασης από την κατάσταση k σε οποιαδήποτε κατάσταση}}$$

ή σαν συνάρτηση των α και β γράφεται:

$$\hat{a}_{kl} = \frac{\sum_{t=1}^{T-1} \xi_t(k,l)}{\sum_{t=1}^{T-1}\sum_{l=1}^{N} \xi_t(k,l)}, \quad \text{όπου } \xi_t(k,l) = \frac{a_t(k) a_{kl} b_l(x_{t+1}) \beta_{t+1}(l)}{\sum_{k=1}^{N}\sum_{l=1}^{N} a_t(k) a_{kl} b_l(x_{t+1}) \beta_{t+1}(l)}$$

Στάδιο IV. Τερματισμός. Αν το κριτήριο σύγκλισης δεν ικανοποιείται,

θέσε: $\pi_i = \hat{\pi}_i$, $a_{ij} = \hat{a}_{ij}$, $b_i(k) = \hat{b}_i(k)$ και πήγαινε στο **στάδιο II**.

ΚΕΦΑΛΑΙΟ 4

ΠΕΙΡΑΜΑΤΑ ΑΝΑΓΝΩΡΙΣΗΣ

4.1 Εισαγωγή

Στο κεφάλαιο αυτό θα δώσουμε τα αποτελέσματα μαζί με τις παραμέτρους των φίλτρων στις οποίες έγιναν επεμβάσεις για την επίτευξη καλύτερης επίδοσης αναγνώρισης. Επιπλέον δίνουμε τα αποτελέσματα από την εκπαίδευση των GMM(*gaussian mixture models*), ενώ δίνουμε και τα αποτελέσματα από την επανεκπαίδευση των αρχικών μοντέλων ύστερα από την κανονικοποίηση των ομιλητών στα δεδομένα εκπαίδευσης. Τα αποτελέσματα, τα παραθέτουμε και με τρόπο ώστε να παρέχεται η δυνατότητα άμεσης σύγκρισης.

Για τον **έλεγχο της επίδοσης** χρησιμοποιήθηκαν:

- 400 εκφράσεις, από 34 ομιλητές
- οι εκφράσεις ήταν ισοκατανεμημένες ανάμεσα στα δύο φύλα
- οι ομιλητές ήταν διαφορετικοί από εκείνους που συμμετείχαν στα δεδομένα για εκπαίδευση.

4.2 Έλεγχος της Επίδοσης Αναγνώρισης

Για τον έλεγχο της επίδοσης προτιμήθηκε το εργαλείο *recognizer* του *Decipher* του *SRI*, κυρίως επειδή πολλά δεδομένα ομιλίας μπορούν να ελεγχθούν ταυτόχρονα και μπορούν να επεξεργαστούν για περισσότερες από μία φορές.

Έτσι ορίζουμε μία βασική επίδοση (*baseline*) και μπορούμε να πειραματιστούμε έπειτα με τις παραμέτρους των φίλτρων. Παράδειγμα εξόδου του *recognizer* δίνουμε παρακάτω :

```
SENTENCE: 1

SENTENCE: 1 pass1

FILENAME:/telecom9/speech/atis/01_tables/waveforms/slsdev-
atis/spcd-atis-dec93-eval/8k3011ss.wav

REF:  FIND ME A FLIGHT THAT FLIES FROM MEMPHIS TO TACOMA

INFORMATION:init_gstate_to_active_hash_n_best:Made
state_to_active_hash, size 8192
```

```
GRAMMAR PROB: -500000000, -2310

INFORMATION: end_sentence_standard: prob = -64986, gprob = -2310

HYP:  FIND ME A FLIGHT THAT FLIES FROM MEMPHIS TO TACOMA

PROB: -64986

PROB_PER_FRAME: -161

PFSG_REALS: 1418 active, 398 ends, 2612 starts, 1230 saved, 494 pruned (34.9%), 32 bt

PFSG_NULLS: 226 active (13.7%), 352 ends, 464 starts, 3 levels

PFSG_REJECTS: 0 active (0.0%), 0 ends, 0 starts (0.0%)

ERRORS:  0 ins 0 del 0 sub 10 wds 0.00% err

TOTAL_ERRORS:  (1) 0 ins 0 del 0 sub 10 wds 0.00% err 0.0% sent

TIMES: 22.8 secs (21.3p 1.4g) 5.7xRT (22.7u 0.0s 5.7xcpuRT)

TOTAL_TIMES: 22.8 secs (21.3p 1.4g) 5.7xRT

GAUSS: -3.09408e+08

GAUSS_PER_FRAME: -767761

ACTIVE_GAUSSIANS: frame_single_feature 1833.9

GAUSSIANS_STARTED: frame_single_feature 19163.2

GENONES_PER_FRAME: frame_single_feature 598.9
```

Όπως φαίνεται, κατά την εκτέλεση του *recognizer* έχουμε στοιχεία για την κάθε πρόταση αλλά και το συνολικό αριθμό προτάσεων μέχρι εκείνη τη στιγμή. Ας εξηγήσουμε τι σημαίνουν τα αποτελέσματα αυτά.

Η πρώτη γραμμή **FILENAME** δίνει το όνομα του αρχείου φωνής που αναγνωρίζεται. Η γραμμή **REF** περιέχει την περιγραφή του αρχείου φωνής, όπως διαβάστηκε από το αρχείο εισόδου που είχαμε δημιουργήσει. Η γραμμή **GRAMMAR** καθορίζει τη γραμματική την οποία θέλουμε να χρησιμοποιήσουμε στην αναγνώριση. Η γραμμή **HYP** επιστρέφει το αποτέλεσμα αναγνώρισης σε επίπεδο λέξης με κεφαλαία γράμματα.

Η γραμμή **PROB** επιστρέφει την πιθανότητα με την οποία επιλέχθηκε η προηγούμενη απάντηση. Η τιμή της παραμέτρου **PROB_PER_FRAME** είναι το αποτέλεσμα της διαίρεσης του **PROB** με τον συνολικό αριθμό των frames των δεδομένων που επεξεργάστηκαν. Δίνεται ο λογάριθμος της πιθανοφάνειας. Όσο λιγότερο αρνητική τιμή έχει αυτός ο αριθμός τόσο μεγαλύτερη είναι η πιθανότητα.

Με την παράμετρο **TIMES** δίνονται οι τιμές των χρόνων, πραγματικές και CPU αντίστοιχα, που χρειάστηκαν για την αναγνώριση. Επιπλέον με την παράμετρο **TOTAL_TIMES** έχουμε τον συνολικό πραγματικό χρόνο που απαιτήθηκε για την αναγνώριση μέχρι αυτήν την πρόταση.

Οι γραμμές **ERRORS** και **TOTAL_ERRORS** περιγράφουν τα στατιστικά αποτελέσματα για το τρέχον αρχείο και όσα έχουν εξεταστεί μέχρι εκείνη τη στιγμή αντίστοιχα. Η αναγνώριση ήταν σωστή και γι' αυτό η γραμμή ERRORS δείχνει 0 **ins**(ertions), 0 **del**(etions) και 0 **sub**(stitutions). Με τους τρεις αυτούς όρους (εισαγωγή, διαγραφή, αντικατάσταση) περιγράφονται τα αποτελέσματα της σύγκρισης της πρότασης που αναγνωρίστηκε και της πρότασης που είχε δοθεί ως περιγραφή. Συγκεκριμένα, μια εισαγωγή σημαίνει ότι το αποτέλεσμα της αναγνώρισης περιέχει μια επιπλέον λέξη, μια διαγραφή σημαίνει ότι το αποτέλεσμα της αναγνώρισης περιέχει μια λέξη λιγότερη, ενώ μια αντικατάσταση σημαίνει ότι το αποτέλεσμα της αναγνώρισης έχει αντικαταστήσει μία λέξη με κάποια άλλη μέσα στην πρόταση.

Το ***Word Error*** περιγράφεται από τη σχέση:

$$W_{Err}(\%) = \frac{I + D + S}{N} 100, \tag{4-1}$$

όπου ***N*** είναι ο συνολικός αριθμός των προς αναγνώριση λέξεων και ***I, D, S*** ο αριθμός των εισαγωγών, διαγραφών και αντικαταστάσεων, αντίστοιχα. Στο παράδειγμα που περιγράψαμε, υπάρχουν λέξεις για αναγνώριση και το σφάλμα για το συγκεκριμένο αρχείο ήταν 0.00%. Το συνολικό σφάλμα σε επίπεδο λέξης μέχρι αυτό το σημείο είναι 0.0%.

4.3 Πειράματα με γραμμικούς μετασχηματισμούς

Στα παρακάτω πειράματα εφαρμόζουμε τους γραμμικούς μετασχηματισμούς όπως αυτοί περιγράφηκαν στο προηγούμενο κεφάλαιο.

4.3.1 Βασική επίδοση

Επίδοση χωρίς κανονικοποίηση, συντελεστής α=1.0. Παραθέτουμε τις εισαγωγές, διαγραφές, αντικαταστάσεις στο σύνολο των λέξεων καθώς και το ποσοστό σφάλματος.

Άντρες

Ομιλητής	INS	DEL	SUB	WORDS	%ERROR
Spk_8k3	0	2	3	123	**4.06**
Spk_8kc	2	1	1	157	**2.54**
Spk_8kn	0	3	0	154	**1.94**
Spk_8ko	1	2	5	126	**6.34**
Spk_8ku	4	0	3	167	**4.19**
Spk_8le	3	3	7	137	**9.48**
Spk_8li	3	0	3	54	**11.11**
Spk_g05	1	0	1	16	**12.5**
Spk_goj	3	1	9	95	**13.68**
Spk_i07	2	0	1	82	**3.65**
Spk_i0e	5	1	3	146	**6.16**
Spk_i0k	2	1	2	84	**5.95**
Spk_tl0	0	0	1	75	**1.33**
Spk_tr0	0	0	4	121	**3.30**
Spk_x06	2	3	15	143	**13.98**
Spk_x0s	7	0	1	59	**13.55**
Spk_x11	5	0	3	93	**8.60**
Spk_x20	0	0	0	91	**0**

Γυναίκες

Ομιλητής	INS	DEL	SUB	WORDS	%ERROR
Spk_8k8	7	0	10	172	**9.88**
Spk_8kp	3	0	6	104	**8.65**
Spk_8lh	4	0	2	116	**5.17**
Spk_8li	1	0	2	136	**2.20**
Spk_g02	1	0	1	81	**2.46**
Spk_g05	3	1	4	172	**4.65**
Spk_g0d	1	0	1	126	**1.58**
Spk_g0j	3	2	11	92	**19.56**
Spk_i0b	4	0	2	98	**6.12**
Spk_i0k	0	0	2	3	**66.66**
Spk_q03	3	0	2	91	**5.49**
Spk_s30	1	0	0	74	**1.351**
Spk_tm0	7	3	7	167	**10.17**
Spk_x0s	2	0	3	144	**3.47**
Spk_x0z	12	1	6	264	**7.19**
Spk_x1e	7	8	6	174	**12.06**

Συνολική βασική επίδοση (χωρίς κανονικοποίηση)

Σύνολο	INS	DEL	SUB	WORDS	%ERROR
άντρες	40	17	62	1923	**6.18**
γυναίκες	59	15	67	2014	**7.0**
άντρες γυναίκες	99	32	129	3937	**6.60**

4.3.2 Βέλτιστη επίδοση ανά ομιλητή

Παραθέτουμε την επίδοση για το καλύτερο αποτέλεσμα μαζί με τις εισαγωγές, διαγραφές, αντικαταστάσεις στο σύνολο των λέξεων καθώς και το ποσοστό σφάλματος.

Άντρες

Ομιλητής	Συντελεστής	INS	DEL	SUB	WORDS	%ERROR
Spk_8k3	**REC_1.06**	0	2	2	123	**3.25**
Spk_8kc	**REC_1.0**	2	1	1	157	**2.54**
Spk_8kn	**REC_0.94**	0	1	0	154	**0.64**
Spk_8ko	**REC_0.92**	1	1	3	126	**3.96**
Spk_8ku	**REC_0.92,1.04**	2	0	2	167	**2.39**
Spk_8le	**REC_0.94**	4	3	5	137	**8 .75**
Spk_8li	**REC_0.96**	3	0	2	54	**9.25**
Spk_g05	**REC_1.0**	1	0	1	16	**12.5**
Spk_goj	**REC_0.98**	3	1	8	95	**12.63**
Spk_i07	**REC_1.0**	2	0	1	82	**3.65**
Spk_i0e	**REC_1.02**	2	2	2	146	**4.10**
Spk_i0k	**REC_0.96**	3	0	1	84	**4.76**
Spk_tl0	**REC_1.12**	0	0	0	75	**0**
Spk_tr0	**REC_0.90**	0	0	2	121	**1.65**
Spk_x06	**REC_0.98**	1	3	12	143	**11.18**
Spk_x0s	**REC_1.0**	7	0	1	59	**13.55**
Spk_x11	**REC_1.02**	5	0	2	93	**7.52**
Spk_x20	**REC_1.0**	0	0	0	91	**0**

Βλέπουμε ότι οι άντρες έχουν συντελεστές στην περιοχή του 1 και διαφορετικούς μεταξύ τους δηλαδή διαφορετικό φωνητικό σωλήνα.

Γυναίκες

Ομιλητής	Συντελεστής	INS	DEL	SUB	WORDS	%ERROR
Spk_8k8	**REC_1.10**	3	0	3	172	**3.48**

Spk_8kp	**REC_1.12**	3	0	1	104	**3.84**
Spk_8lh	**REC_1.0**	4	0	2	116	**5.17**
Spk_8li	**REC_1.02**	0	0	2	136	**1.47**
Spk_g02	**REC_1.0**	1	0	1	81	**2.46**
Spk_g05	**REC_0.98**	3	0	4	172	**4.06**
Spk_g0d	**REC_1.0**	1	0	1	126	**1.58**
Spk_g0j	**REC_1.02**	2	2	11	92	**16.30**
Spk_i0b	**REC_1.02**	4	0	1	98	**5.10**
Spk_i0k	**REC_1.0**	0	0	2	3	**66.66**
Spk_q03	**REC_1.04**	2	0	1	91	**3.29**
Spk_s30	**REC_1.02**	0	0	0	74	**0**
Spk_tm0	**REC_1.02**	7	3	5	167	**8.98**
Spk_x0s	**REC_1.04**	1	0	3	144	**2.77**
Spk_x0z	**REC_0.96**	9	1	5	264	**5.68**
Spk_x1e	**REC_1.0**	7	8	6	174	**12.06**

Οι γυναίκες έχουν καλύτερα αποτελέσματα για συντελεστές μεγαλύτερους από τους άντρες όπως έχουμε αναφέρει στη θεωρία.

Βέλτιστο αποτέλεσμα

Φύλο	**INS**	**DEL**	**SUB**	**WORDS**	**%ERROR**
άντρες	36	14	45	1923	**4.94**
γυναίκες	47	14	48	2014	**5.41**
Σύνολο	83	28	93	3937	**5.18**

4.3.3 Επιλογή ενός συντελεστή ανά φύλο

Χρησιμοποιούμε έναν κοινό συντελεστή για όλους τους ομιλητές.

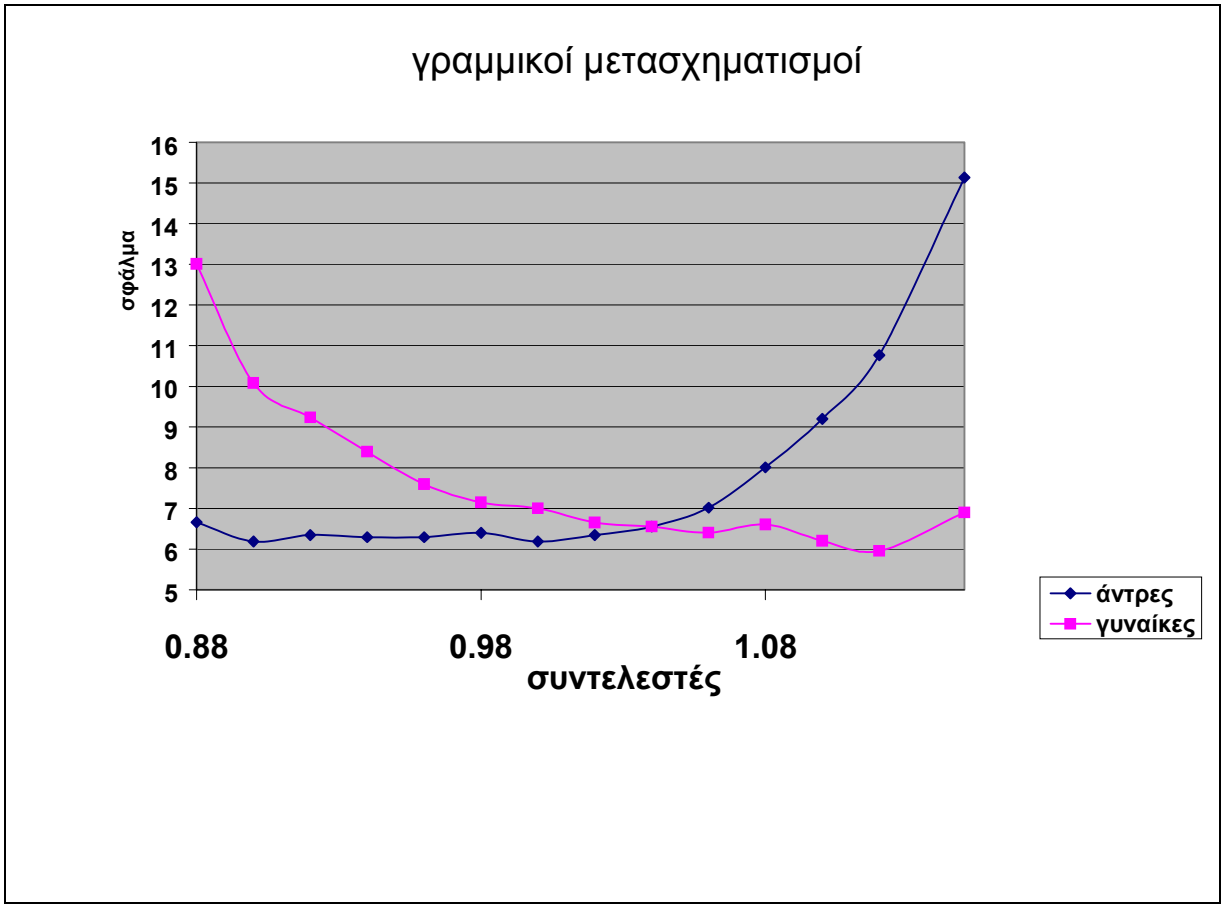
Άντρες

Συντελεστής	INS	DEL	SUB	WORDS	%ERROR
REC_0.88	40	17	71	1923	**6.65**
REC_0.90	41	15	63	1923	**6.18**
REC_0.92	41	16	65	1923	**6.34**
REC_0.94	42	15	64	1923	**6.29**
REC_0.96	40	16	65	1923	**6.29**
REC_0.98	44	16	63	1923	**6.39**
REC_1.0	40	17	62	1923	**6.18**
REC_1.02	41	17	64	1923	**6.34**
REC_1.04	41	17	68	1923	**6.55**
REC_1.06	40	17	78	1923	**7.02**
REC_1.08	43	18	93	1923	**8.00**
REC_1.10	48	18	111	1923	**9.20**
REC_1.12	57	23	127	1923	**10.76**
REC_1.15	72	31	188	1923	**15.13**

Γυναίκες

Συντελεστής	INS	DEL	SUB	WORDS	%ERROR
REC_0.88	94	27	141	2014	**13**
REC_0.90	74	23	106	2014	**10.07**
REC_0.92	74	20	92	2014	**9.23**
REC_0.94	67	20	82	2014	**8.39**
REC_0.96	60	20	72	2014	**7.59**
REC_0.98	60	16	68	2014	**7.14**
REC_1.0	59	15	67	2014	**7**
REC_1.02	58	16	59	2014	**6.65**
REC_1.04	57	16	59	2014	**6.55**
REC_1.06	57	15	57	2014	**6.40**
REC_1.08	59	14	60	2014	**6.60**
REC_1.10	57	15	53	2014	**6.20**
REC_1.12	52	14	54	2014	**5.95**
REC_1.15	59	15	65	2014	**6.90**

Γραφική επίδοση για γραμμικούς μετασχηματισμούς



Πίνακας με καλύτερο συντελεστή ανά φύλο

Συνδυάζοντας τα πειράματα έχουμε

ΦΥΛΟ	Συντελεστής	INS	DEL	SUB	WORDS	%ERROR
ΑΝΤΡΕΣ	**REC_1.0**	40	17	62	1923	**6.18**
ΓΥΝΑΙΚΕΣ	**REC_1.12**	52	14	54	2014	**5.95**

Από τα παραπάνω πειράματα διαπιστώνουμε ότι οι άντρες ομιλητές παρουσιάζουν τα καλύτερα αποτελέσματα για συντελεστή 1, ενώ για συντελεστές>1 έχουμε βαθμιαία επιδείνωση στην επίδοση αναγνώρισης. Στις γυναίκες ομιλητές έχουμε βελτίωση στην επίδοση αναγνώρισης για συντελεστές>1, ενώ τα καλύτερα αποτελέσματα παρουσιάζονται για συντελεστή 1.12. Τα αποτελέσματα αυτά επιβεβαιώνουν τη θεωρητική μελέτη του προηγούμενου κεφαλαίου, ενώ το μήκος του φωνητικού σωλήνα κυμαίνεται από 13cm (γυναίκες) σε 18cm (άντρες) ταυτόχρονα οι *formant frequencies* διαφέρουν μέχρι και 25% μεταξύ των ομιλητών. Με συντελεστές>1 έχουμε καλύτερη κάλυψη για γυναίκες με συνέπεια καλύτερα αποτελέσματα, ενώ στους άντρες χάνουμε σε ακρίβεια με συνέπεια χειρότερα αποτελέσματα.

4.3.4 Πειράματα GENIE στο σύνολο των ομιλητών

Στα πειράματα αυτά χρησιμοποιούμε τα πεδίο **ERROR** του αναγνωριστή. Τα πειράματα αυτά έχουν θεωρητική αξία. Υποθέτουμε ότι γνωρίζουμε εκ των προτέρων (μαγικά **GENIE**) ποιος συντελεστής δίνει τα καλύτερα αποτελέσματα ανά πρόταση ως προς **ERROR** και τον χρησιμοποιούμε στην αναγνώριση. Χρησιμοποιούμε όλους τους συντελεστές και κάθε πρόταση ανεξάρτητα.

Φύλο	**INS**	**DEL**	**SUB**	**WORDS**	**%ERROR**
άντρες	28	13	41	1923	**4.26**
γυναίκες	42	10	41	2014	**4.61**

Τα αποτελέσματα είναι καλύτερα όταν χρησιμοποιούμε το πεδίο **ERROR** κάτι που είναι αναμενόμενο. Τα αποτελέσματα αυτά μας δίνουν ένα άνω όριο της βέλτιστης επίδοσης. Στα παραπάνω πειράματα είναι δυνατόν κάθε πρόταση να έχει διαφορετικό βέλτιστο συντελεστή για αυτό έχουμε καλύτερα αποτελέσματα αντί να επιλέγουμε με βάση όλες τις προτάσεις (επόμενος πίνακας).

Στον πίνακα αυτό έχουμε επίδοση με όλες τις προτάσεις ανά ομιλητή και όχι κάθε πρόταση ανεξάρτητα.

Φύλο	**INS**	**DEL**	**SUB**	**WORDS**	**%ERROR**
άντρες	36	14	45	1923	**4.94**
γυναίκες	47	14	48	2014	**5.41**

4.3.5 Πειράματα με το κριτήριο Maximum Likelihood

Βρίσκουμε τον καλύτερο συντελεστή χρησιμοποιώντας το πεδίο **PROB**.

ΠΕΙΡΑΜΑ Α Όλες οι προτάσεις ανά ομιλητή (πίνακας με επιλογή ανά ομιλητή).Καλύτερος συντελεστής ανά ομιλητή. Παραθέτουμε το συντελεστή **α** καθώς και το ποσοστό σφάλματος.

ΠΕΙΡΑΜΑ Β Κάθε πρόταση ανεξάρτητα (πίνακας με επιλογή ανά πρόταση). Καλύτερος συντελεστής ανά πρόταση. Παραθέτουμε το συντελεστή **α** καθώς και το ποσοστό σφάλματος κάθε ομιλητή.

Παραθέτουμε τις εισαγωγές διαγραφές αντικαταστάσεις στο σύνολο των λέξεων καθώς και το ποσοστό σφάλματος. Τα αποτελέσματα μετά την αναγνώριση φαίνονται στον Πίνακα.

Γυναίκες

FEMALE	INS	DEL	SUB	WORDS	%ERROR
PROB A	55	13	59	2014	**6.30**
PROB B	57	16	57	2014	**6.45**

Ομιλητής	FR_WARP (A)	%ERROR PROB (A)	FR_WARP (B)	%ERROR PROB(B)
Spk_8k8	1.15	**5.23**	1.143	**5.23**
Spk_8kp	1.15	**4.80**	1.141	**4.80**
Spk_8lh	1.15	**6.03**	1.123	**5.17**
Spk_8li	1.06	**2.20**	1.065	**2.20**
Spk_g02	1.15	**2.46**	1.081	**3.70**
Spk_g05	0.98	**4.06**	1.02	**6.97**
Spk_g0d	1.06	**1.58**	1.06	**1.58**
Spk_g0j	1.06	**16.30**	1.043	**16.30**
Spk_i0b	1.08	**6.12**	1.081	**5.10**
Spk_i0k	0.98	**66.67**	0.98	**66.67**
Spk_q03	1.08	**3.29**	1.081	**3.29**
Spk_s30	1.08	**2.70**	1.069	**0**
Spk_tm0	1.08	**10.17**	1.062	**9.58**
Spk_x0s	0.98	**3.47**	1.0	**3.47**
Spk_x0z	1.04	**7.19**	1.042	**7.57**
Spk_x1e	1.06	**13.21**	1.041	**13.79**

Στις τιμές της 4 στήλης του πίνακα παίρνουμε την μέση τιμή των συντελεστών για κάθε πρόταση για αυτό έχουμε δεκαδικές τιμές. Παρατηρούμε ότι οι γυναίκες έχουν συντελεστές >1.

Άντρες

MALE	INS	DEL	SUB	WORDS	%ERROR
PROB A	39	17	60	1923	**6.03**

PROB B	39	17	63	1923	**6.188**

Ομιλητής	FR_WARP (A)	%ERROR PROB(A)	FR_WARP (B)	%ERROR PROB(B)
Spk_8k3	1.02	**4.06**	1.013	**4.06**
Spk_8kc	1.02	**2.54**	1.02	**2.54**
Spk_8kn	0.96	**1.29**	0.965	**1.29**
Spk_8ko	0.98	**6.34**	0.99	**6.34**
Spk_8ku	0.98	**4.19**	0.998	**4.19**
Spk_8le	0.96	**9.48**	0.968	**10.21**
Spk_8li	1.08	**11.11**	1.086	**11.11**
Spk_g05	1.06	**12.5**	1.073	**12.5**
Spk_goj	1.02	**13.68**	1.036	**12.63**
Spk_i07	1.02	**4.87**	0.991	**4.87**
Spk_i0e	1.02	**4.10**	1.023	**4.79**
Spk_i0k	1.04	**7.14**	1.094	**7.14**
Spk_tl0	1.04	**1.33**	1.037	**1.33**
Spk_tr0	0.96	**3.30**	0.96	**3.30**
Spk_x06	0.98	**11.18**	0.964	**11.18**
Spk_x0s	0.98	**20.33**	1.002	**23.72**
Spk_x11	0.92	**7.52**	0.92	**7.26**
Spk_x20	1.02	**0.0**	1.003	**0.0**

Στις τιμές της 4 στήλης του πίνακα παίρνουμε την μέση τιμή των συντελεστών για κάθε πρόταση για αυτό έχουμε δεκαδικές τιμές. Οι άντρες έχουν συντελεστές στην περιοχή του 1. Επίσης μπορούμε να πούμε ότι έχουμε καλύτερα αποτελέσματα, όταν αποφασίζουμε για όλες τις προτάσεις αντί για κάθε πρόταση χωριστά μιας όταν χρησιμοποιούμε όλες τις προτάσεις παίρνουμε απόφαση με μεγαλύτερη ακρίβεια βέβαια οι διαφορές είναι μικρές.

4.4 Πειράματα με γραμμικούς μετασχηματισμούς κατά περιοχές (piecewise linear)

Στα παρακάτω πειράματα εφαρμόζουμε τους γραμμικούς μετασχηματισμούς κατά περιοχές όπως αυτοί περιγράφτηκαν στη θεωρία.

4.4.1 Επιλογή ενός συντελεστή ανά φύλο

Παραθέτουμε τη συχνότητα αποκοπής, τους δυο συντελεστές, τις εισαγωγές, διαγραφές, αντικαταστάσεις στο σύνολο των λέξεων καθώς και το

ποσοστό σφάλματος. Χρησιμοποιούμε ένα κοινό συντελεστή για όλους τους ομιλητές.

Άντρες

Fo-A1-A2	INS	DEL	SUB	WORDS	%ERROR
0.25-0.92-0.92	41	16	65	1923	**6.34**
0.25-0.92-1.0	39	15	63	1923	**6.08**
0.25-0.92-1.08	35	14	65	1923	**5.92**
0.25-1.0-0.92	39	18	65	1923	**6.34**
0.25-1.0-1.0	40	17	62	1923	**6.18**
0.25-1.0-1.08	43	15	63	1923	**6.29**
0.25-1.08-0.92	42	18	81	1923	**7.33**
0.25-1.08-1.0	45	20	89	1923	**8**
0.25-1.08-1.08	43	18	93	1923	**8**
0.43-0.92-0.92	41	16	65	1923	**6.34**
0.43-0.92-1.0	41	15	64	1923	**6.24**
0.43-0.92-1.08	37	15	65	1923	**6.08**
0.43-1.0-0.92	36	18	64	1923	**6.13**
0.43-1.0-1.0	40	15	62	1923	**6.18**
0.43-1.0-1.08	42	15	62	1923	**6.18**
0.43-1.08-0.92	46	21	87	1923	**8**
0.43-1.08-1.0	44	19	88	1923	**7.85**
0.43-1.08-1.08	43	18	93	1923	**8**
0.625-0.92-0.92	41	16	65	1923	**6.34**
0.625-0.92-1.0	42	16	68	1923	**6.55**
0.625-0.92-1.08	42	14	63	1923	**6.18**
0.625-1.0-0.92	39	16	62	1923	**6.08**
0.625-1.0-1.0	40	17	62	1923	**6.18**
0.625-1.0-1.08	45	17	61	1923	**6.39**
0.625-1.08-0.92	45	18	94	1923	**8.16**
0.625-1.08-1.0	43	18	97	1923	**8.21**
0.625-1.08-1.08	43	18	93	1923	**8**

Γυναίκες

Fo-A1-A2	INS	DEL	SUB	WORDS	%ERROR
0.25-0.92-0.92	74	20	92	2014	**9.23**
0.25-0.92-1.0	69	18	78	2014	**8.19**
0.25-0.92-1.08	60	19	76	2014	**7.69**
0.25-1.0-0.92	61	17	74	2014	**7.54**
0.25-1.0-1.0	59	15	67	2014	**7**
0.25-1.0-1.08	58	16	62	2014	**6.75**
0.25-1.08-0.92	62	15	65	2014	**7.05**
0.25-1.08-1.0	59	16	61	2014	**6.75**
0.25-1.08-1.08	59	14	61	2014	**6.60**
0.43-0.92-0.92	74	20	92	2014	**9.23**
0.43-0.92-1.0	71	19	85	2014	**8.68**
0.43-0.92-1.08	67	20	78	2014	**8.19**
0.43-1.0-0.92	61	15	70	2014	**7.24**
0.43-1.0-1.0	59	15	67	2014	**7**
0.43-1.0-1.08	60	15	66	2014	**7**
0.43-1.08-0.92	59	15	56	2014	**6.45**
0.43-1.08-1.0	59	15	55	2014	**6.40**
0.43-1.08-1.08	59	14	60	2014	**6.60**
0.625-0.92-0.92	74	20	92	2014	**9.23**
0.625-0.92-1.0	70	20	86	2014	**8.73**
0.625-0.92-1.08	69	20	86	2014	**8.68**
0.625-1.0-0.92	60	15	67	2014	**7.05**
0.625-1.0-1.0	59	15	67	2014	**7**
0.625-1.0-1.08	58	15	65	2014	**6.85**
0.625-1.08-0.92	58	15	55	2014	**6.35**
0.625-1.08-1.0	57	15	54	2014	**6.25**
0.625-1.08-1.08	59	14	60	2014	**6.60**

Αξίζει να σημειωθεί ότι όταν οι δυο συντελεστές είναι ίδιοι η επίδοση είναι εκείνη των γραμμικών μετασχηματισμών (επαλήθευση αποτελεσμάτων). Επίσης όταν οι συντελεστές>1 έχουμε επιδείνωση των αποτελεσμάτων στους άντρες. Στις γυναίκες για συντελεστές>1 έχουμε βελτίωση των αποτελεσμάτων κάτι που είναι αναμενόμενο. Καθώς το εύρος τιμών για τους δυο συντελεστές είναι μικρότερο από το εύρος τιμών στους γραμμικούς μετασχηματισμούς δεν έχουμε τα ίδια βέλτιστα αποτελέσματα. Ισχύει ότι και στους γραμμικούς μετασχηματισμούς.

Καλύτερος συντελεστής ανά φύλο

Συνδυάζοντας τα πειράματα έχουμε

Συνολικά βέλτιστο αποτέλεσμα

ΦΥΛΟ	FACTOR	INS	DEL	SUB	WORDS	%ERROR
ΑΝΤΡΕΣ	**0.25-0.92-1.08**	35	14	65	1923	**5.92**
ΓΥΝΑΙΚΕΣ	**0.625-1.08-1.0**	57	15	54	2014	**6.25**

Αξίζει να σημειωθεί ότι όταν η συχνότητα αποκοπής αυξάνεται έχουμε επιδείνωση των αποτελεσμάτων στους άντρες. Στις γυναίκες αντίθετα έχουμε βελτίωση των αποτελεσμάτων αφού μας ενδιαφέρουν οι ψηλότερες συχνότητες.

4.4.2 Πειράματα GENIE στο σύνολο των ομιλητών

Στα πειράματα αυτά χρησιμοποιούμε τα πεδίο **ERROR** του αναγνωριστή. Τα πειράματα αυτά έχουν θεωρητική αξία. Υποθέτουμε ότι γνωρίζουμε εκ των προτέρων (μαγικά **GENIE**) ποιος συντελεστής δίνει τα καλύτερα αποτελέσματα ως προς **ERROR** και τον χρησιμοποιούμε στην αναγνώριση. Χρησιμοποιούμε όλους τους συντελεστές και κάθε πρόταση ανεξάρτητα.

Συνολικά βέλτιστο αποτέλεσμα

Φύλο	INS	DEL	SUB	WORDS	%ERROR
άντρες	28	14	45	1923	**4.52**
γυναίκες	47	13	43	2014	**5.11**

Τα αποτελέσματα μας δίνουν ένα άνω όριο της βέλτιστης επίδοσης. Επιλέγουμε από όλους τους αναγνωριστές εκείνη την υπόθεση με το μικρότερο σφάλμα για όλες τις συχνότητες αποκοπής.

4.4.3 Πειράματα με το κριτήριο Maximum Likelihood

Βρίσκουμε τον καλύτερο συντελεστή χρησιμοποιώντας το πεδίο **PROB**. Παραθέτουμε τις εισαγωγές διαγραφές αντικαταστάσεις στο σύνολο των λέξεων καθώς και το ποσοστό σφάλματος.

ΠΕΙΡΑΜΑ Α Όλες οι προτάσεις ανά ομιλητή. Καλύτερος συντελεστής ανά ομιλητή.

ΠΕΙΡΑΜΑ Β Κάθε πρόταση ανεξάρτητα Καλύτερος συντελεστής ανά πρόταση.

MALE	INS	DEL	SUB	WORDS	%ERROR
PROB A	41	15	55	1923	**5.77**
PROB B	38	15	61	1923	**5.92**

Ομιλητής	FR_WARP (A)	%ERROR PROB(A)	FR_WARP (B)	%ERROR PROB(B)
Spk_8k3	0.625-1.0-1.08	**4.06**	0.42-1.02-1.0	**4.06**
Spk_8kc	0.25-1.0-1.08	**2.54**	0.29-1.01-1.03	**3.18**
Spk_8kn	0.43-0.92-1.08	**0.64**	0.34-0.93-1.08	**0.64**
Spk_8ko	0.25-0.92-1.08	**4.761**	0.28-0.92-1.08	**4.76**
Spk_8ku	0.43-1.0-1.08	**2.39**	0.43-0.99-1.04	**3.59**
Spk_8le	0.625-0.92-1.08	**9.48**	0.51-0.92-1.0	**10.21**
Spk_8li	0.25-1.08-1.08	**11.11**	0.25-1.08-1.08	**11.11**
Spk_g05	0.625-1.08-0.92	**12.5**	0.31-1.08-1.02	**12.5**
Spk_goj	0.25-1.08-1.0	**12.63**	0.46-1.06-1.0	**13.68**
Spk_i07	0.25-1.0-1.08	**6.09**	0.26-0.95-1.053	**6.09**
Spk_i0e	0.25-1.0-1.08	**5.47**	0.28-1.0-1.06	**4.10**
Spk_i0k	0.25-1.0-1.08	**5.95**	0.35-1.0-1.0	**7.14**
Spk_tl0	0.625-1.08-1.0	**1.33**	0.41-1.05-0.97	**1.33**
Spk_tr0	0.43-0.92-1.08	**3.30**	0.49-0.93-1.06	**2.47**
Spk_x06	0.43-0.92-1.08	**11.18**	0.40-0.92-1.06	**11.18**
Spk_x0s	0.43-1.08-0.92	**20.33**	0.41-1.0-0.97	**20.33**
Spk_x11	0.25-0.92-0.92	**7.52**	0.29-0.92-1.0	**7.52**
Spk_x20	0.25-0.92-1.08	**0.0**	0.29-0.96-1.08	**0.0**

Στις τιμές της 4 στήλης του πίνακα παίρνουμε την μέση τιμή των συντελεστών για κάθε πρόταση για αυτό έχουμε δεκαδικές τιμές. Οι άντρες έχουν συντελεστές στην περιοχή του 1.

Γυναίκες

FEMALE	INS	DEL	SUB	WORDS	%ERROR
PROB A	60	14	58	2014	**6.55**
PROB B	58	15	55	2014	**6.35**

Ομιλητής	FR_WARP (A)	%ERROR PROB (A)	FR_WARP (B)	%ERROR PROB(B)
Spk_8k8	0.625-1.08-0.92	**4.65**	0.46-1.08-0.98	**4.06**
Spk_8kp	0.625-1.08-0.92	**6.73**	0.51-1.08-0.97	**6.73**
Spk_8lh	0.625-1.08-0.92	**5.17**	0.58-1.08-0.93	**5.17**
Spk_8li	0.625-1.08-1.0	**2.20**	0.43-1.06-1.04	**2.2**
Spk_g02	0.25-1.08-1.08	**2.46**	0.34-1.06-1.01	**2.46**
Spk_g05	0.43-1.08-0.92	**6.39**	0.44-1.04-0.95	**6.39**
Spk_g0d	0.25-1.08-1.08	**1.58**	0.29-1.08-1.01	**1.58**
Spk_g0j	0.25-1.08-1.08	**16.3**	0.28-1.03-1.06	**16.3**
Spk_i0b	0.25-1.08-1.08	**6.12**	0.30-1.08-1.04	**6.12**
Spk_i0k	0.43-1.08-1.0	**66.67**	0.43-1.08-1.0	**66.67**
Spk_q03	0.625-1.08-1.0	**3.29**	0.50-1.08-1.02	**3.29**
Spk_s30	0.25-1.08-1.08	**2.7**	0.43-1.08-0.98	**0**
Spk_tm0	0.25-1.08-1.08	**10.17**	0.46-1.08-0.98	**9.58**
Spk_x0s	0.625-1.0-0.92	**3.47**	0.41-1.0-0.98	**4.16**
Spk_x0z	0.625-1.08-0.92	**6.81**	0.35-1.05-0.97	**6.81**
Spk_x1e	0.43-1.08-1.0	**14.36**	0.42-1.06-1.0	**13.79**

Στις τιμές της 4 στήλης του πίνακα παίρνουμε την μέση τιμή των συντελεστών για κάθε πρόταση. Παρατηρούμε ότι οι γυναίκες έχουν συντελεστές >1. Επίσης μπορούμε να πούμε ότι έχουμε καλύτερα αποτελέσματα, όταν αποφασίζουμε για κάθε πρόταση χωριστά αντί για όλες τις προτάσεις βέβαια οι διαφορές είναι μικρές.

4.5 Πειράματα με μετασχηματισμούς μεταφοράς

Στα παρακάτω πειράματα εφαρμόζουμε τους μετασχηματισμούς μεταφοράς όπως αυτοί περιγράφτηκαν στο προηγούμενο κεφάλαιο.

4.5.1 Επιλογή ενός συντελεστή ανά φύλο

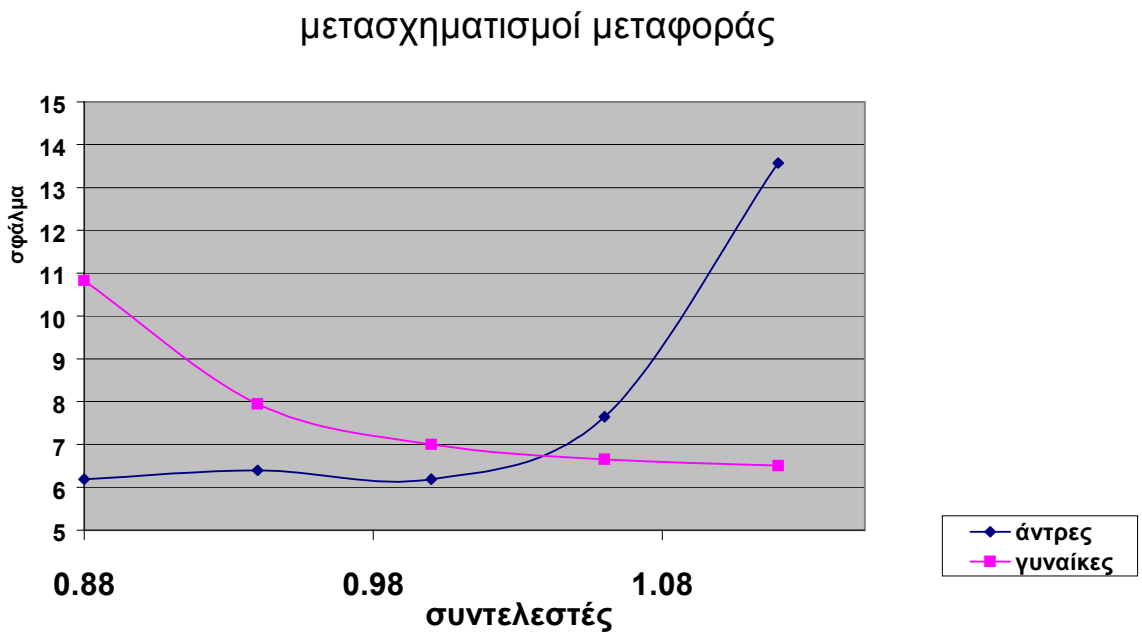
Παραθέτουμε τις εισαγωγές διαγραφές αντικαταστάσεις στο σύνολο των λέξεων καθώς και το ποσοστό σφάλματος.

Άντρες

Fo	INS	DEL	SUB	WORDS	%ERROR
0.88	37	15	67	1923	**6.18**
0.94	41	15	67	1923	**6.39**
1.0	40	17	62	1923	**6.18**
1.06	44	18	85	1923	**7.64**
1.12	65	27	169	1923	**13.57**

Γυναίκες

Fo	INS	DEL	SUB	WORDS	%ERROR
0.88	79	21	118	2014	**10.82**
0.94	61	22	77	2014	**7.94**
1.0	59	15	67	2014	**7**
1.06	58	16	60	2014	**6.65**
1.12	58	16	57	2014	**6.5**



Γραφική επίδοση για μετασχηματισμούς μεταφοράς

Από τα παραπάνω πειράματα διαπιστώνουμε ότι οι άντρες ομιλητές παρουσιάζουν τα καλύτερα αποτελέσματα για συντελεστή==1, ενώ για συντελεστές>1 έχουμε βαθμιαία επιδείνωση στην επίδοση αναγνώρισης. Στις

γυναίκες ομιλητές έχουμε βελτίωση στην επίδοση αναγνώρισης για συντελεστές>1, ενώ τα καλύτερα αποτελέσματα παρουσιάζονται για συντελεστή==1.12. Καθώς μεταβάλλουμε μόνο την κεντρική συχνότητα των φίλτρων ενώ το εύρος τους παραμένει σταθερό δεν έχουμε τα ίδια βέλτιστα αποτελέσματα όπως στους γραμμικούς μετασχηματισμούς.

Καλύτερος συντελεστής ανά φύλο

Συνδυάζοντας τα πειράματα έχουμε

Φύλο	FACTOR	INS	DEL	SUB	WORDS	%ERROR
άντρες	**1.0**	40	17	62	1923	**6.18**
γυναίκες	**1.12**	58	16	57	2014	**6.5**

Παρατηρούμε ότι δεν έχουμε τα ίδια βέλτιστα αποτελέσματα όπως στους γραμμικούς μετασχηματισμούς όπως εξηγήσαμε πριν.

4.5.2 Πειράματα GENIE στο σύνολο των ομιλητών

Στα πειράματα αυτά χρησιμοποιούμε τα πεδίο **ERROR** του αναγνωριστή. Τα πειράματα αυτά έχουν θεωρητική αξία. Υποθέτουμε ότι γνωρίζουμε εκ των προτέρων (μαγικά **GENIE**) ποιος συντελεστής δίνει τα καλύτερα αποτελέσματα ως προς **ERROR** και τον χρησιμοποιούμε στην αναγνώριση. Χρησιμοποιούμε όλους τους συντελεστές και κάθε πρόταση ανεξάρτητα. Βρίσκουμε τον καλύτερο συντελεστή ανά πρόταση χρησιμοποιώντας το πεδίο **ERROR.**

Φύλο	INS	DEL	SUB	WORDS	%ERROR
άντρες	29	13	45	1923	**4.52**
γυναίκες	44	12	43	2014	**4.91**

Τα αποτελέσματα μας δίνουν ένα άνω όριο της βέλτιστης επίδοσης. Επιλέγουμε από όλους τους αναγνωριστές εκείνη την υπόθεση με το μικρότερο σφάλμα για τις διάφορες συχνότητες αποκοπής.

4.5.3 Πειράματα με το κριτήριο Maximum Likelihood

Βρίσκουμε τον καλύτερο συντελεστή χρησιμοποιώντας το πεδίο **PROB**. Παραθέτουμε τις εισαγωγές διαγραφές αντικαταστάσεις στο σύνολο των λέξεων καθώς και το ποσοστό σφάλματος.

ΠΕΙΡΑΜΑ Α Όλες οι προτάσεις ανά ομιλητή. Καλύτερος συντελεστής ανά ομιλητή.

ΠΕΙΡΑΜΑ Β Κάθε πρόταση ανεξάρτητα Καλύτερος συντελεστής ανά πρόταση.

MALE	INS	DEL	SUB	WORDS	%ERROR
PROB A	36	14	65	1923	**5.98**
PROB B	37	14	65	1923	**6.03**

Ομιλητής	FR_WARP (A)	%ERROR PROB(A)	FR_WARP (B)	%ERROR PROB(B)
Spk_8k3	0.88	**4.87**	0.915	**4.87**
Spk_8kc	0.88	**3.82**	0.907	**3.82**
Spk_8kn	0.88	**1.29**	0.89	**1.29**
Spk_8ko	0.88	**3.96**	0.88	**3.96**
Spk_8ku	0.88	**2.39**	0.89	**2.39**
Spk_8le	0.94	**9.25**	0.88	**9.48**
Spk_8li	0.94	**12.5**	0.92	**9.25**
Spk_g05	0.94	**12.5**	0.98	**12.5**
Spk_goj	0.88	**13.68**	0.89	**13.68**
Spk_i07	0.88	**4.87**	0.88	**4.87**
Spk_i0e	0.88	**6.16**	0.88	**6.16**
Spk_i0k	0.88	**5.95**	0.88	**5.95**
Spk_tl0	0.88	**1.33**	0.925	**1.33**
Spk_tr0	0.88	**1.62**	0.88	**1.62**
Spk_x06	0.88	**11.18**	0.884	**11.18**
Spk_x0s	0.94	**23.72**	0.913	**25.42**
Spk_x11	0.88	**7.52**	0.88	**7.52**
Spk_x20	0.88	**1.09**	0.88	**1.09**

Στις τιμές της 4-στήλης του πίνακα παίρνουμε την μέση τιμή των συντελεστών για κάθε πρόταση. Οι άντρες έχουν συντελεστές μικρότερους του 1.

Γυναίκες

FEMALE	INS	DEL	SUB	WORDS	%ERROR
PROB A	59	16	62	2014	**6.8**
PROB B	60	16	58	2014	**6.65**

Ομιλητής	FR_WARP (A)	%ERROR PROB (A)	FR_WARP (B)	%ERROR PROB (B)
Spk_8k8	1.06	**4.65**	1.072	**2.9**
Spk_8kp	1.06	**6.73**	1.048	**4.8**
Spk_8lh	0.94	**8.62**	0.94	**8.62**
Spk_8li	0.94	**2.2**	0.92	**2.2**
Spk_g02	1.00	**2.46**	0.985	**2.46**
Spk_g05	0.94	**5.23**	0.982	**5.23**
Spk_g0d	1.0	**1.58**	1.00	**2.38**
Spk_g0j	0.88	**23.91**	0.89	**21.73**
Spk_i0b	1.0	**6.12**	0.967	**7.14**
Spk_i0k	0.88	**66.67**	0.88	**66.67**
Spk_q03	1.06	**3.29**	1.05	**3.29**
Spk_s30	1.0	**1.31**	1.018	**1.313**
Spk_tm0	1.000	**10.17**	1.021	**8.98**
Spk_x0s	0.94	**4.16**	0.91	**6.25**
Spk_x0z	0.94	**6.81**	0.905	**6.81**
Spk_x1e	1.000	**12.06**	0.996	**12.64**

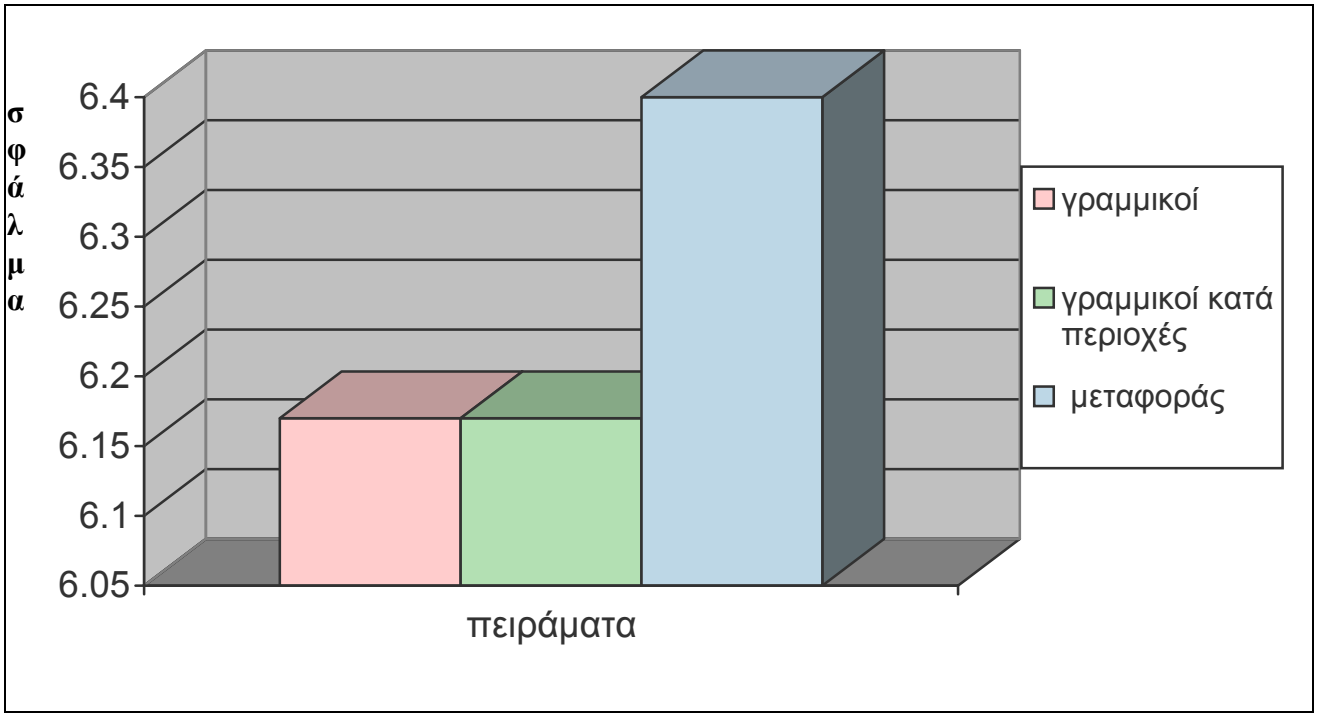
Στις τιμές της 4 στήλης του πίνακα παίρνουμε την μέση τιμή των συντελεστών για κάθε πρόταση. Παρατηρούμε ότι οι γυναίκες έχουν συντελεστές >1. Οι διαφορές είναι μικρές μεταξύ των 2 πειραμάτων.

4.6 Σύγκριση μεθόδων με το κριτήριο Maximum Likelihood

Όλες οι προτάσεις ανά ομιλητή. Καλύτερος συντελεστής ανά ομιλητή.

Μέθοδος	INS	DEL	SUB	WORDS	%ERROR
Γραμμικοί	94	30	119	3937	**6.17**
Γραμμικοί κατά περιοχές	101	29	113	3937	**6.17**
Μεταφοράς	95	30	127	3937	**6.40**

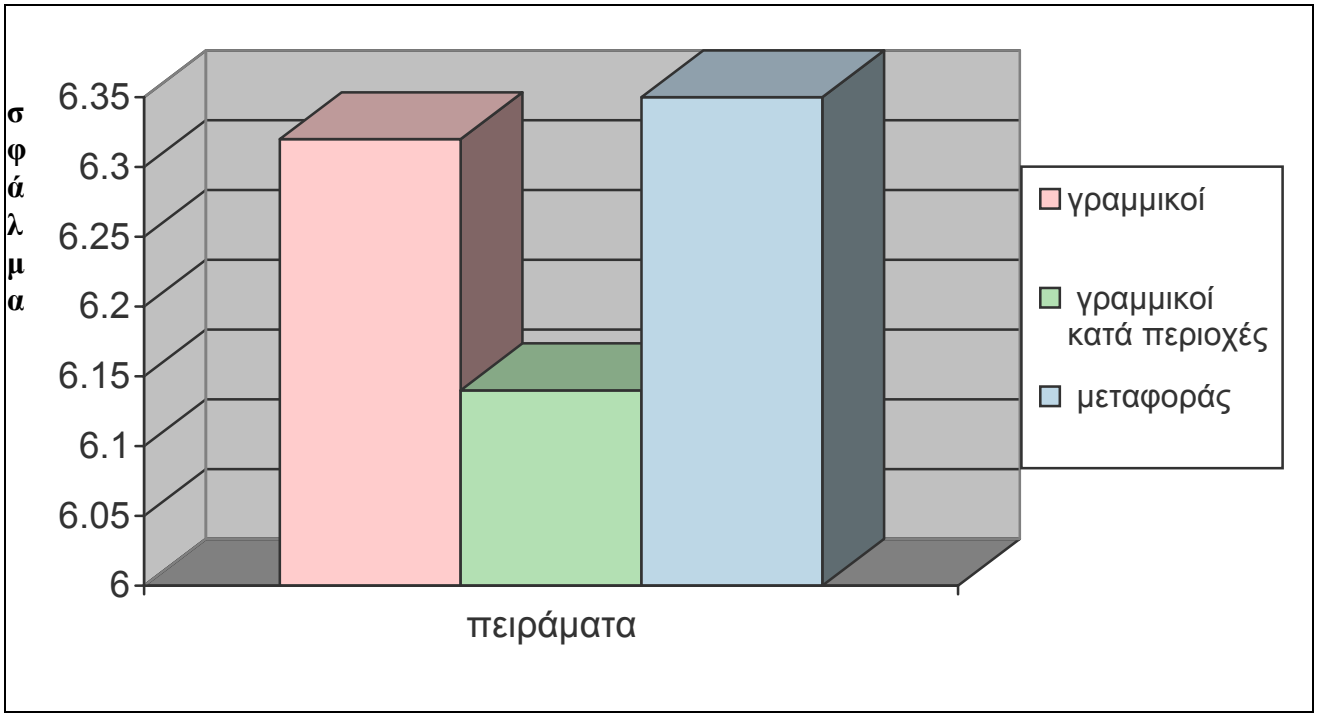
Γραφική Επίδοση όλες οι προτάσεις ανά ομιλητή



Κάθε πρόταση ανεξάρτητα. Καλύτερος συντελεστής ανά πρόταση.

Μέθοδος	INS	DEL	SUB	WORDS	%ERROR
γραμμικοί	96	33	120	3937	**6.32**
Γραμμικοί κατά περιοχές	96	30	116	3937	**6.14**
μεταφοράς	97	30	123	3937	**6.35**

Γραφική Επίδοση κάθε πρόταση ανεξάρτητα



Επειδή η φύση των γραμμικών μετασχηματισμών(απλός, κατά περιοχές) είναι η ίδια, οι διαφορές στην επίδοση μεταξύ όλων των προτάσεων και κάθε πρόταση ξεχωριστά είναι της τάξης του 1.5 και 2%. Καλύτερο αποτέλεσμα έχουμε για τους γραμμικούς κατά περιοχές που εξηγείται ότι επιλέγουμε από περισσότερα πειράματα 27 έναντι 14. Καθώς μεταβάλλουμε μόνο την κεντρική συχνότητα των φίλτρων ενώ το εύρος τους παραμένει σταθερό δεν έχουμε τα ίδια βέλτιστα αποτελέσματα στους μετασχηματισμούς μεταφοράς όπως στους γραμμικούς. Σε όλες τις περιπτώσεις η επίδοση είναι καλύτερη από τη βασική επίδοση χωρίς κανονικοποίηση. Θέλουμε όμως να αποφύγουμε την ταυτόχρονη εκτέλεση των *recognizers* για αυτό πρέπει να επιλέγουμε βάση ενός κριτηρίου το βέλτιστο συντελεστή και στη συνέχεια να κάνουμε αναγνώριση. Αυτό επιτυγχάνεται με χρήση **GMM.**

4.7 Πειράματα με GMM (gaussian mixture model) και γραμμικούς μετασχηματισμούς

Η **εκπαίδευση των μοντέλων** αναλυτικά μπορεί να περιγραφεί από τα εξής στοιχεία:

- 5000 προτάσεις, από 150 ομιλητές
- οι ομιλητές ήταν ισοκατανεμημένοι ανάμεσα στα δύο φύλα

- **εκπαίδευση με παραμέτρους 0.90 1.0 1.12**
- **εκτελέστηκαν πειράματα με τις παρακάτω παραμέτρους,**

3 παραμέτρους 0.90 1.0 1.12

5 παραμέτρους 0.90 0.94 1.0 1.06 1.12

6 παραμέτρους 0.90 0.94 1.0 1.06 1.10 1.12

8 παραμέτρους 0.88 0.92 0.96 1.0 1.04 1.08 1.12.

Πειράματα με όλες τις προτάσεις

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	5.77	5.9	5.84
5	5.82	5.9	5.86
6	5.82	6.10	5.96
8	6.24	6.2	6.22

Πειράματα με μια πρόταση

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	5.82	6	5.91
5	5.98	6	5.99
6	5.98	6.05	6.01
8	6.29	6.50	6.40

Από τα παραπάνω πειράματα διαπιστώνουμε ότι τόσο οι άντρες ομιλητές όσο και οι γυναίκες ομιλητές παρουσιάζουν τα καλύτερα αποτελέσματα για 3-5 παραμέτρους, ενώ για 6-8 παραμέτρους έχουμε επιδείνωση στην επίδοση αναγνώρισης. Αυτό οφείλεται στο γεγονός ότι αυξάνοντας τον αριθμό των παραμέτρων χάνεται η ακρίβεια με την οποία το **GMM** επιλέγει το βέλτιστο συντελεστή μιας και πρέπει να επιλέξει από μεγαλύτερο αριθμό υποθέσεων, ενώ έχει εκπαιδευτεί με 3 παραμέτρους. Αξίζει να σημειωθεί ότι έχουμε καλύτερα αποτελέσματα αντί να χρησιμοποιούμε το πεδίο Prob του αναγνωριστή. Επίσης έχουμε καλύτερα αποτελέσματα, όταν χρησιμοποιούμε 64 gaussians μιας με λιγότερες gaussians επιτυγχάνουμε καλύτερη εκπαίδευση των παραμέτρων τους για το ίδιο πλήθος δεδομένων εκπαίδευσης. Επίσης μπορούμε να πούμε ότι έχουμε καλύτερα αποτελέσματα, όταν

αποφασίζουμε για όλες τις προτάσεις αντί για κάθε πρόταση χωριστά μιας όταν χρησιμοποιούμε όλες τις προτάσεις παίρνουμε απόφαση με μεγαλύτερη ακρίβεια, βέβαια οι διαφορές είναι μικρές.

4.8 Πειράματα με GMM (gaussian mixture model) και μετασχηματισμούς μεταφοράς

Η **εκπαίδευση των μοντέλων** αναλυτικά μπορεί να περιγραφεί από τα εξής στοιχεία:

- 5000 προτάσεις , από 150 ομιλητές
- οι ομιλητές ήταν ισοκατανεμημένοι ανάμεσα στα δύο φύλα
- **εκπαίδευση με παραμέτρους 1.0 1.06 1.12**
- **εκτελέστηκαν πειράματα με τις παρακάτω παραμέτρους,**

3 παραμέτρους 1.0 1.06 1.12

5 παραμέτρους 0.88 0.94 1.0 1.06 1.12.

Πειράματα με όλες τις προτάσεις

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	6.18	6.35	6.27
5	6.18	7.49	6.85

Πειράματα με μια πρόταση

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	6.18	6.10	6.14
5	6.03	7.0	6.52

Από τα παραπάνω πειράματα διαπιστώνουμε ότι οι γυναίκες ομιλητές παρουσιάζουν τα καλύτερα αποτελέσματα για 3 παραμέτρους, ενώ για 5 παραμέτρους έχουμε επιδείνωση στην επίδοση αναγνώρισης. Μάλιστα η επιδείνωση στις γυναίκες είναι μεγάλη όταν μεταβαίνουμε από 3 σε 5 παραμέτρους που οφείλεται στο γεγονός ότι παίρνοντας 5 παραμέτρους έχουμε και συρρίκνωση του άξονα της συχνότητας. Αυτό οφείλεται στο γεγονός ότι αυξάνοντας τον αριθμό των παραμέτρων χάνεται η ακρίβεια με την οποία το **GMM** επιλέγει το βέλτιστο συντελεστή μιας και πρέπει να

επιλέξει από μεγαλύτερο αριθμό υποθέσεων, ενώ έχει εκπαιδευτεί με 3 παραμέτρους. Αντίθετα στους άντρες δεν παρουσιάζεται σημαντική μεταβολή μιας και οι παράμετροι>1. Αξίζει να σημειωθεί ότι έχουμε καλύτερα αποτελέσματα αντί να χρησιμοποιούμε το πεδίο Prob του αναγνωριστή. Καθώς μεταβάλλουμε μόνο την κεντρική συχνότητα των φίλτρων ενώ το εύρος τους παραμένει σταθερό δεν έχουμε τα ίδια βέλτιστα αποτελέσματα όπως στους γραμμικούς μετασχηματισμούς.

4.9 Πειράματα με GMM (gaussian mixture model) και γραμμικούς μετασχηματισμούς κατά περιοχές και συχνότητα αποκοπής 0.25

Η **εκπαίδευση των μοντέλων** αναλυτικά μπορεί να περιγραφεί από τα εξής στοιχεία:

- 5000 προτάσεις, από 150 ομιλητές
- οι ομιλητές ήταν ισοκατανεμημένοι ανάμεσα στα δύο φύλα
- **εκπαίδευση με παραμέτρους 0.25-0.92-1.08, 0.25-1.0-1.0, 0.25-1.08-1.08**
- **εκτελέστηκαν πειράματα με τις παρακάτω παραμέτρους,**

3 παραμέτρους 0.25-0.92-1.08 ,0.25-1.08-1.08 ,0.25-1.08-1.0
5 παραμέτρους 0.25-0.92-1.08 ,0.25-1.0-1.0, 0.25-0.92-1.0, 0.25-1.08-1.0,0.25-1.08-1.08

Πειράματα με όλες τις προτάσεις

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	5.92	6.6	6.27
5	5.82	6.6	6.22

Πειράματα με μια πρόταση

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	5.92	6.75	6.34
5	5.77	6.6	6.19

4.10 Πειράματα με GMM (gaussian mixture model) και γραμμικούς μετασχηματισμούς κατά περιοχές και συχνότητα αποκοπής 0.43.

Η **εκπαίδευση των μοντέλων** αναλυτικά μπορεί να περιγραφεί από τα εξής στοιχεία:

- 5000 προτάσεις, από 150 ομιλητές
- οι ομιλητές ήταν ισοκατανεμημένοι ανάμεσα στα δύο φύλα
- **εκπαίδευση με παραμέτρους 0.43.1.0.0.92, 0.43.0.92.1.08,0.43.1.08.1.0**
- **εκτελέστηκαν πειράματα με τις παρακάτω παραμέτρους,**

3 παραμέτρους 0.43.1.0.0.92, 0.43.0.92.1.08, 0.43.1.08.1.0
5 παραμέτρους 0.43.1.0.0.92, 0.43.0.92.1.08, 0.43.1.08.1.0, 0.43.1.0.1.0, 0.43.1.08.0.92

Πειράματα με όλες τις προτάσεις

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	5.92	6.65	6.29
5	5.92	6.55	6.24

Πειράματα με μια πρόταση

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	6.03	6.55	6.29
5	6.08	6.55	6.32

4.11 Πειράματα με GMM (gaussian mixture model) και γραμμικούς μετασχηματισμούς κατά περιοχές και συχνότητα αποκοπής 0.625.

Η **εκπαίδευση των μοντέλων** αναλυτικά μπορεί να περιγραφεί από τα εξής στοιχεία:

- 5000 προτάσεις, από 150 ομιλητές
- οι ομιλητές ήταν ισοκατανεμημένοι ανάμεσα στα δύο φύλα

- **εκπαίδευση με παραμέτρους 0.625.1.0.0.92, 0.625.1.0.1.0, 0.625.1.08.1.0**

- **εκτελέστηκαν πειράματα με τις παρακάτω παραμέτρους,**

3 παραμέτρους 0.625.1.0.0.92, 0.625.1.0.1.0, 0.625.1.08.1.0
5 παραμέτρους 0.625.1.0.0.92, 0.625.1.0.1.0, 0.625.1.08.1.0, 0.625.92.1.08, 0.625.1.08.0.92.

Πειράματα με όλες τις προτάσεις

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	6.08	6.3	6.19
5	5.72	6.4	6.07

Πειράματα με μια πρόταση

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	6.08	6.25	6.17
5	5.82	6.35	6.09

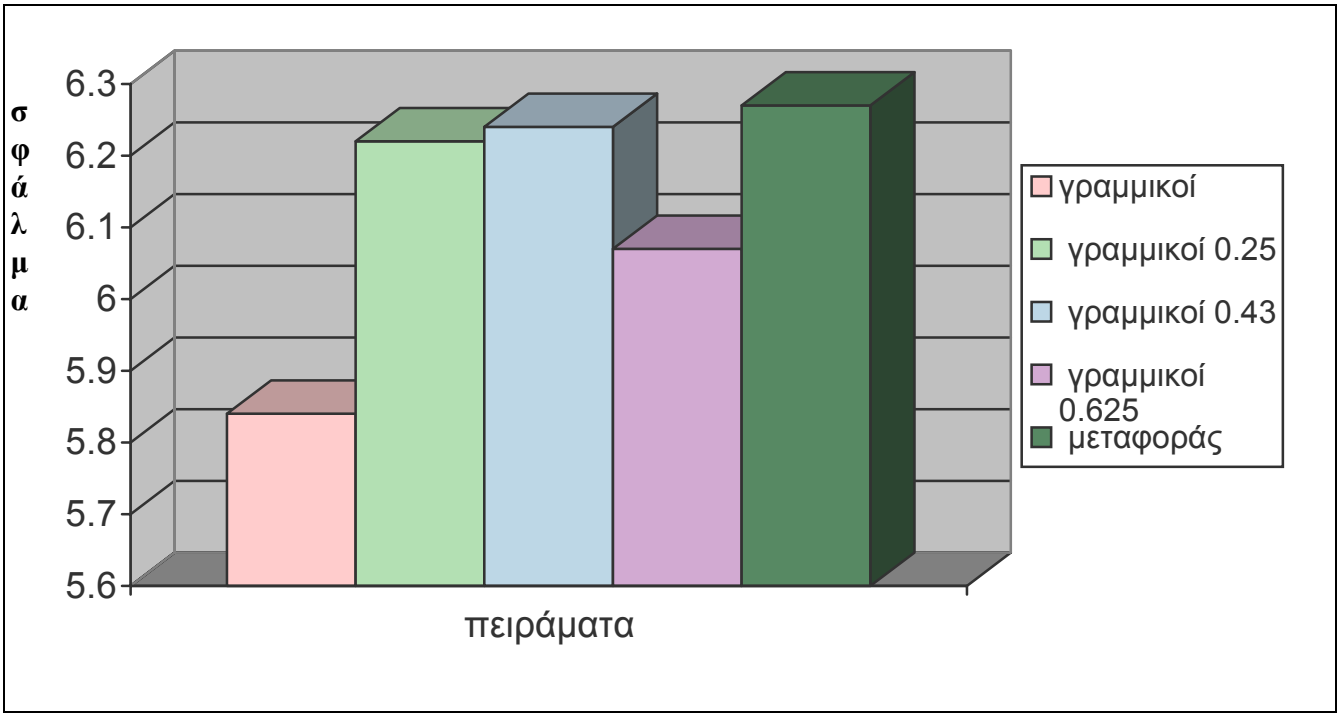
Από τα παραπάνω πειράματα **4.9 4.10 4.11** αξίζει να σημειωθεί ότι έχουμε καλύτερα αποτελέσματα αντί να χρησιμοποιούμε το πεδίο Prob του αναγνωριστή. Επίσης έχουμε καλύτερα αποτελέσματα, όταν χρησιμοποιούμε 64 gaussians μιας με λιγότερες gaussians επιτυγχάνουμε καλύτερη εκπαίδευση των παραμέτρων τους για το ίδιο πλήθος δεδομένων εκπαίδευσης. Από τα παραπάνω πειράματα μπορούμε να πούμε ότι έχουμε τα ίδια περίπου αποτελέσματα όταν αποφασίζουμε για όλες τις προτάσεις αντί για κάθε πρόταση χωριστά Αξίζει να σημειωθεί ότι όταν η συχνότητα αποκοπής αυξάνεται έχουμε επιδείνωση των αποτελεσμάτων στους άντρες. Στις γυναίκες αντίθετα έχουμε βελτίωση των αποτελεσμάτων.

4.12 Συνολικά βέλτιστα αποτελέσματα για όλες τις μεθόδους (άντρες γυναίκες) με χρήση GMM

Όλες οι προτάσεις ανά ομιλητή. Καλύτερος συντελεστής ανά ομιλητή.

Μέθοδος	#παράμετροι	INS	DEL	SUB	WORDS	%ERROR
γραμμικοί	**3**	90	29	111	3937	**5.84**
γραμμικοί 0.25	**5**	95	31	119	3937	**6.22**
γραμμικοί 0.43	**5**	95	30	121	3937	**6.24**
γραμμικοί 0.625	**5**	98	29	112	3937	**6.07**
μεταφοράς	**3**	95	32	120	3937	**6.27**

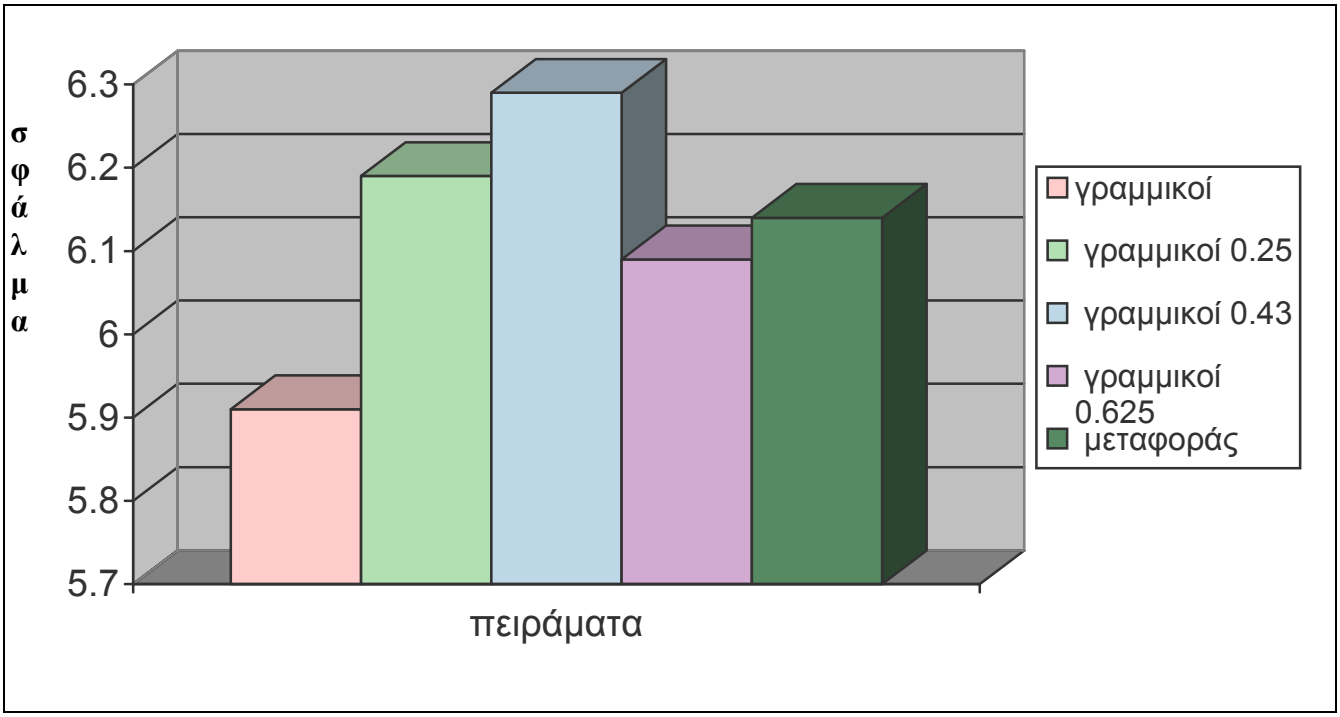
Γραφική Επίδοση όλες οι προτάσεις ανά ομιλητή



Κάθε πρόταση ανεξάρτητα Καλύτερος συντελεστής ανά πρόταση.

Μέθοδος	#παράμετροι	INS	DEL	SUB	WORDS	%ERROR
γραμμικοί	**3**	91	29	113	3937	**5.91**
γραμμικοί 0.25	**5**	94	31	119	3937	**6.19**
γραμμικοί 0.43	**3**	95	30	123	3937	**6.29**
γραμμικοί 0.625	**5**	99	28	113	3937	**6.09**
μεταφοράς	**3**	93	32	117	3937	**6.14**

Γραφική Επίδοση κάθε πρόταση ανεξάρτητα



Καλύτερα αποτελέσματα

Πειράματα με όλες τις προτάσεις

Μέθοδος	#παράμετροι	INS	DEL	SUB	WORDS	%ERROR
γραμμικοί	**3**	90	29	111	3937	**5.84**

Πειράματα με μια πρόταση

Μέθοδος	#παράμετροι	INS	DEL	SUB	WORDS	%ERROR
γραμμικοί	**3**	91	29	113	3937	**5.91**

Από τα παραπάνω πειράματα έχουμε τα καλύτερα αποτελέσματα για γραμμικούς μετασχηματισμούς και αριθμό παραμέτρων ίσο με 3. Επιτύχαμε την μείωση του σφάλματος αναγνώρισης από 6.60(χωρίς κανονικοποίηση) σε 5.84 με χρήση **GMM** βελτίωση 11.5% ενώ πλέον τα αποτελέσματα είναι καλύτερα εκείνων με χρήση του πεδίου PROB.Αυτό που έχει επιπλέον σημασία είναι ότι αποφύγαμε την ταυτόχρονη εκτέλεση των *recognizers.* Επίσης οι διαφορές είναι μικρές όταν επιλέγουμε με όλες τις προτάσεις και όταν επιλέγουμε για κάθε πρόταση ανεξάρτητα.

4.13 Πειράματα με επανεκπαίδευση των αρχικών μοντέλων

Στο σημείο αυτό προκειμένου να επιτευχθεί συμβατότητα μεταξύ των αρχικών μοντέλων και της διαδικασίας αναγνώρισης γίνεται επανεκπαίδευση όπως περιγράφηκε στο προηγούμενο κεφάλαιο.

Πείραμα Α εκπαίδευση με 3 παραμέτρους 0.90 1.0 1.12

Για την επανεκπαίδευση:

•χρησιμοποιήθηκαν 21261 προτάσεις

•για παράμετρο 0.90 9449 προτάσεις, 1.0 5369 προτάσεις, 1.12 6443 προτάσεις αντίστοιχα.

Πείραμα Β εκπαίδευση με 6 παραμέτρους 0.90 0.94 1.0 1.04 1.08 1.12

Για την επανεκπαίδευση:

•με παράμετρο 0.90 6221 προτάσεις,0.94 4606 προτάσεις 1.0 737 προτάσεις, 1.04 4066 προτάσεις 1.08 2891 προτάσεις 1.12 2760 προτάσεις.

Πείραμα Γ εκπαίδευση με 5 παραμέτρους 0.90 1.0 1.04 1.08 1.12

Για την επανεκπαίδευση:

•με παράμετρο 0.90 10807 προτάσεις, 1.0 737 προτάσεις, 1.04 4066 προτάσεις 1.08 2891 προτάσεις 1.12 2760 προτάσεις αντίστοιχα.

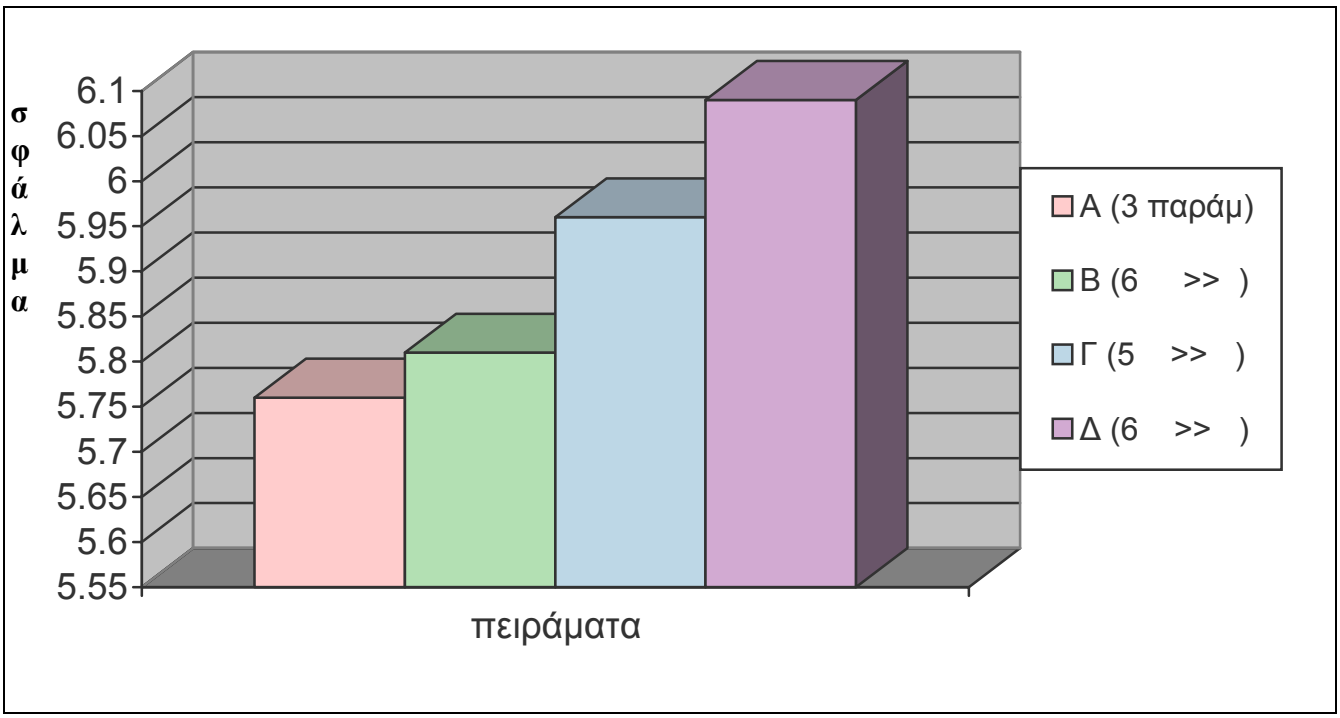
Πείραμα Δ εκπαίδευση με 6 παραμέτρους 0.98 1.0 1.02 1.04 1.06 1.08

Για την επανεκπαίδευση:

•με παράμετρο 0.98 11727 προτάσεις, 1.0 0 προτάσεις, 1.02 1206 προτάσεις 1.04 1691 προτάσεις 1.06 1913 προτάσεις 1.08 4724 προτάσεις αντίστοιχα.

Πειράματα με GMM (gaussian mixture model)

Πειράματα	άντρες	γυναίκες	Σύνολο
Α	**5.46**	**6.05**	**5.76**
Β	**5.61**	**6**	**5.81**
Γ	**5.66**	**6.25**	**5.96**
Δ	**5.82**	**6.35**	**6.09**



Βέλτιστο αποτέλεσμα με χρήση GMM

ΠΕΙΡΑΜΑ	#παράμετροι	INS	DEL	SUB	WORDS	%ERROR
παλιά μοντέλα	**3**	90	29	111	3937	**5.84**
νέα μοντέλα	**3**	96	28	103	3937	**5.76**

Εκπαιδεύοντας λοιπόν τα μοντέλα με 3 παραμέτρους 0.90 (συρρίκνωση) 1.0 (χωρίς αλλαγή) και 1.12 (επέκταση) έχουμε καλύτερα αποτελέσματα στους άντρες από 5.77 σε 5.46 και ελάχιστα χειρότερα αποτελέσματα στις γυναίκες 5.90 σε 6.0 σε σύγκριση με τη χρήση των παλιών μοντέλων. Συνολικά έχουμε βελτίωση της επίδοσης αναγνώρισης κατά 12% στους άντρες και 14% στις γυναίκες σε σχέση με τα παλιά μοντέλα χωρίς κανονικοποίηση. Επίσης η επίδοση με χρήση GMM είναι 5.84 (παλιά μοντέλα) σε 5.76 (νέα μοντέλα) πράγμα που σημαίνει ότι το GMM έχει εκπαιδευτεί επαρκώς.

4.14 Πειράματα με μετασχηματισμούς μεταφοράς και εκπαίδευση με 3 παραμέτρους 1.0 1.06 1.12

Για την επανεκπαίδευση:

•με παράμετρο 1.0 18364 προτάσεις 1.06 1206 προτάσεις 1.12 1691 προτάσεις.

Πειράματα με GMM (gaussian mixture model) και 3 παραμέτρους

ΦΥΛΟ	INS	DEL	SUB	WORDS	%ERROR
ΑΝΤΡΕΣ	34	15	63	1923	**5.82423**
ΓΥΝΑΙΚΕΣ	61	19	56	2014	**6.75273**
ΣΥΝΟΛΟ	95	34	119	3937	**6.29**

Βέλτιστο αποτέλεσμα με χρήση GMM

ΠΕΙΡΑΜΑ	#παράμετροι	INS	DEL	SUB	WORDS	%ERROR
παλιά μοντέλα	**3**	95	32	120	3937	**6.27**
νέα μοντέλα	**3**	95	34	119	3937	**6.29**

4.15 Πειράματα με γραμμικούς μετασχηματισμούς κατά περιοχές και εκπαίδευση με 3 συχνότητες αποκοπής

Πειράματα με GMM (gaussian mixture model) και συχνότητα αποκοπής 0.25

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	**6.03**	**6.7**	**6.37**
5	**6.18**	**6.75**	**6.47**

Πειράματα με GMM (gaussian mixture model) και συχνότητα αποκοπής 0.43

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	**5.94**	**6.7**	**6.32**
5	**5.94**	**6.7**	**6.32**

Πειράματα με GMM (gaussian mixture model) και συχνότητα αποκοπής 0.625

#παράμετροι	άντρες	γυναίκες	Σύνολο
3	**6.24**	**6.45**	**6.35**
5	**5.82**	**6.45**	**6.19**

Καλύτερα αποτελέσματα

Παλιά μοντέλα

Μέθοδος	#παράμετροι	INS	DEL	SUB	WORDS	%ERROR
γραμμικοί 0.25	**5**	95	31	119	3937	**6.22**
γραμμικοί 0.43	**5**	95	30	121	3937	**6.24**
γραμμικοί 0.625	**5**	98	29	112	3937	**6.07**

Νέα μοντέλα

Μέθοδος	#παράμετροι	INS	DEL	SUB	WORDS	%ERROR
γραμμικοί 0.25	**3**	97	29	125	3937	**6.37**
γραμμικοί 0.43	**3-5**	100	30	119	3937	**6.32**
γραμμικοί 0.625	**5**	97	29	118	3937	**6.19**

Οι διαφορές στην επίδοση μεταξύ παλιών και νέων μοντέλων είναι της τάξης του 1.5 και 2%. Μολονότι η φύση των γραμμικών μετασχηματισμών(απλός, κατά περιοχές) είναι η ίδια δεν έχουμε τα ίδια βέλτιστα αποτελέσματα για στους απλούς είχαμε εύρος 0.90-1.12 ενώ στους κατά περιοχές είχαμε εύρος 0.92-1.08 για λιγότερα πειράματα.

4.16 Βέλτιστο αποτέλεσμα για όλες τις μεθόδους επανεκπαίδευσης

Συγκρίνοντας τις επιδόσεις από όλα τα πειράματα επανεκπαίδευσης έχουμε το καλύτερο αποτέλεσμα για γραμμικούς μετασχηματισμούς πείραμα 4.13 Α. Παραθέτουμε επίσης και το βέλτιστο αποτέλεσμα για τα παλιά μοντέλα για σύγκριση.

Βέλτιστο αποτέλεσμα με χρήση GMM

ΠΕΙΡΑΜΑ	#παράμετροι	INS	DEL	SUB	WORDS	%ERROR
παλιά μοντέλα	**3**	90	29	111	3937	**5.84**
νέα μοντέλα	**3**	96	28	103	3937	**5.76**

4.17 Ανακεφαλαίωση Συμπεράσματα

Από τα παραπάνω πειράματα μελετήσαμε, διάφορους μετασχηματισμούς για να επιτύχουμε την κανονικοποίηση του μήκους του φωνητικού σωλήνα όπως:

- γραμμικούς μετασχηματισμούς
- γραμμικούς μετασχηματισμούς κατά περιοχές
- μετασχηματισμούς μεταφοράς.

Οι άντρες ομιλητές παρουσιάζουν τα καλύτερα αποτελέσματα για συντελεστή==1, ενώ για συντελεστές>1 έχουμε βαθμιαία επιδείνωση στην επίδοση αναγνώρισης. Στις γυναίκες ομιλητές έχουμε βελτίωση στην επίδοση αναγνώρισης για συντελεστές>1. Τα αποτελέσματα αυτά επιβεβαιώνουν τη θεωρητική μελέτη του προηγούμενου κεφαλαίου. Με συντελεστές>1 έχουμε καλύτερη κάλυψη για γυναίκες με συνέπεια καλύτερα αποτελέσματα, ενώ στους άντρες χάνουμε σε ακρίβεια με συνέπεια χειρότερα αποτελέσματα. Από τα παραπάνω πειράματα είχαμε καλύτερα αποτελέσματα χρησιμοποιώντας γραμμικούς μετασχηματισμούς και γραμμικούς μετασχηματισμούς κατά περιοχές αντί για μετασχηματισμούς μεταφοράς. Καθώς μεταβάλλουμε μόνο την κεντρική συχνότητα των φίλτρων ενώ το εύρος τους παραμένει σταθερό στους μετασχηματισμούς μεταφοράς δεν έχουμε τα ίδια βέλτιστα αποτελέσματα όπως στους γραμμικούς μετασχηματισμούς.

Στη συνέχεια επιδιώχθηκε η ολοκλήρωση των παραπάνω παρατηρήσεων στο σύστημα αναγνώρισης συνεχούς ομιλίας. Αναπτύσσουμε δηλαδή αλγορίθμους για την εύρεση του βέλτιστου συντελεστή για κάθε ομιλητή στο test set για την επίτευξη καλύτερης επίδοσης αναγνώρισης. Στην περίπτωση αυτή βρίσκουμε το βέλτιστο συντελεστή α_i, βάση ενός κριτηρίου. Σκοπός μας είναι να επιλέξουμε το κατάλληλο κριτήριο και να το χρησιμοποιήσουμε στην αναγνώριση. Οι αλγόριθμοι τους οποίους εφαρμόζουμε είναι οι εξής:

- Παράλληλη Αναγνώριση
- Αναγνώριση με χρήση GMM (gaussian mixture model).

Θα πρέπει να σημειωθεί ότι χρησιμοποιώντας αναγνώριση με χρήση GMM αντί για παράλληλη αναγνώριση (χρήση του πεδίου PROB) επιτύχαμε όχι μόνο βελτίωση της επίδοσης αλλά αποφύγαμε και την παράλληλη εκτέλεση πολλών recognizers (υπολογιστική σπατάλη) για τους διάφορους συντελεστές α_i. Βρίσκουμε το βέλτιστο συντελεστή α_i πριν την αναγνώριση, με χρήση GMM (gaussian mixture model) και με τον τρόπο αυτό χρησιμοποιούμε έναν μόνο αναγνωριστή.

Τέλος για την επίτευξη συμβατότητας μεταξύ φάσης εκπαίδευσης και φάσης αναγνώρισης έγινε επανεκπαίδευση των μοντέλων και επιτεύχθηκαν ανάλογα αποτελέσματα με αυτά της χρήσης GMM (gaussian mixture model) και βελτίωση της επίδοσης αναγνώρισης κατά 12% στους άντρες και 14% στις γυναίκες σε σχέση με τα παλιά μοντέλα χωρίς κανονικοποίηση. Μελλοντικά μπορεί να επιδιωχθεί η αύξηση των δεδομένων εκπαίδευσης του GMM (gaussian mixture model) για την επίτευξη μεγαλύτερης ακρίβειας, ή η εκπαίδευση ενός GMM (gaussian mixture model) για την εύρεση του φύλου του ομιλητή ή η εκπαίδευση ενός GMM (gaussian mixture model) για κάθε συντελεστή.

Βιβλιογραφία

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

[2] R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen and V. Zue **"Survey of the State of the Art in Human Language Technology"**, *November 1995.*

[3] E.Eide & H.Gish ,*A parametric approach to vocal tract length normalization ICASSP 96 Vol 1: pg 346-348, 1996* .

[4] S.Wegmann ,Don M Allaster,Jeremy Orloff & Barbara Peskin ,*Speaker normalization on conventional telephone Speech. Proc ICASSP 96 Vol 1 pg 339-341 Atlanta.*

[5] Kamm, T Andreou & Cohen J, *Vocal tract length Normalization in Speech Recognition, Proc CAIP Workshop: frontiers in speech recognition II 1994.*

[6] L.R Rabiner, *The impact of Voice Processing on modern telecommunications* Nuance Speech Recognition System Version. 5 **Developer's Manual**. *Nuance Communications, 1996.*

[7] L.R. Rabiner, **"A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition"** *Proceedings of the IEEE, 77 Vol(2):pg 257-286, February 1989.*

[8]. D.Pye & P.C.Wooland *"Experiments in speaker normalisation an adaptation or large vocabulary speech recognition ", Proc ICASSP 1997 Vol 2 pg 1047-1050.*

[9] A. P. Dempster, N. M. Laird, and D.B. Rubin **"Maximum Likelihood from Incomplete Data via the EM"** *J. Roy. Stat. Soc., Vol. 39, no. 1, pp. 1-38, 1977.*

[10] V. Digalakis, P. Monaco and H. Murveit, *Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers,* IEEE Transactions on Speech and Audio Processing, June 1996.

[11] L.Lee & Richard C.Rose, *Speaker normalization using efficient frequency warping procedures ICASSP 96 Vol 1: pg 346-348 ,1996.*

[12] A.Potamianos & Richard C.Rose ,*On combining frequency warping and spectral shaping in HMM based speech Recognition ICASSP 97 pg 1275-1278.*

[13] P. Zhan & M Westphal ,*Speaker normalization based on frequency warping, Proc ICASSP 97 Munich Germany.*

[14] A.Potamianos , Richard C.Rose & L.Lee, *A feature-space transformation for telephone based speech recognition Eurospeech 1995.*

[15] T.Kamm ,A.G.Andreou & Dod Fort Meade ,*Vocal tract normalization in Speech Recognition: Compensating for Systematic Speaker Variability CAIP workshop August 1994 "frontiers in Speech Processing II".*

[16] John R Deller ,John G Proakis & John H .L. Hansen ,*Discrete time processing of speech signals New York 1993.*

[17] Steeve Young, **"A Review of Large Vocabulary Continuous Speech Recognition"** *IEEE Signal Processing Magazine pp. 45-57, September 1996.*

[18] H. Murveit, J. Butzberger, V. Digalakis, M. Weintraub, *Large Vocabulary dictation using SRI's DECIPHER speech recognition system : Progressive Search techniques.*

[19] V. Digalakis, D. Rtischev & L. Neumeyer, *Fast Speaker Adaptation using constrained estimation of Gaussian Mixtures.* IEEE Transactions on Speech and Audio Processing, Sept. 1995, vol.3, pp. 357-366

[20] John W.Mac Donough *Speaker normalization with all-pass transforms. Research Notes No 28 March 1998.*