ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών
Εργαστήριο Τηλεπικοινωνιών και Πληροφορίας & Δικτύων

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Μελέτη τεχνικών μείωσης αριθμού παραμέτρων με χρήση ομαδοποίησης σε επίπεδο υποδιανύσματος για κρυφά Μαρκοβιανά μοντέλα μειγμάτων διακριτών κατανομών

(Clustering of distributions at the subvector level for Discrete Mixture HMMs.)

ΓΑΒΑΛΑΚΗΣ ΠΕΤΡΟΣ

Επιβλέπων: Καθηγητής Διγαλάκης Βασίλης

Εξεταστική Επιτροπή: Καθηγητής Πατεράκης Μιχάλης

Αναπλ. καθηγητής Ποταμιάνος Αλέξανδρος

Ιούλιος 2004

ΧΑΝΙΑ

Ευχαριστίες:

Στο σημείο αυτό θα ήθελα να ευχαριστήσω ιδιαίτερα τον επιβλέποντα και εισηγητή της εργασίας καθηγητή Διγαλάκη Βασίλη για τη βοήθεια και τη στήριξη που είχα σε όλη τη διάρκεια εκπόνησης της μεταπτυχιακής μου εργασίας. Εποικοδομητικές επίσης ήταν οι απόψεις και η βοήθεια του συμφοιτητή μου (διδακτορικού φοιτητή πλέον) Χαριζάκη Κώστα καθώς και των υπόλοιπων μελών της ομάδας αναγνώρισης ομιλίας.

Θα ήθελα επίσης να ευχαριστήσω και την υπόλοιπη εξεταστική επιτροπή, αποτελούμενη από τον καθηγητή Πατεράκη Μιχάλη και τον αναπληρωτή καθηγητή Ποταμιάνο Αλέξανδρο για την ανάγνωση της εργασίας και τις υποδείξεις τους.

1 Εισαγωγή

1.1 Επιγραμματικά

Τα τελευταία χρόνια η τεχνολογία της αναγνώρισης ομιλίας μεγάλου λεξιλογίου ανεξαρτήτως ομιλητή, έχει κάνει πολύ σημαντικά βήματα και έχει πλέον πολλές πρακτικές εμπορικές εφαρμογές. Παρόλα αυτά απέχει πολύ από τον αρχικά υποσχόμενο πανταχού παρών τρόπο επικοινωνίας με τον υπολογιστή αλλά και τις άλλες «έξυπνες» συσκευές που σιγά-σιγά κατακλύζουν την καθημερινή μας ζωή.

Πίσω από τα εντυπωσιακά αποτελέσματα στο επίπεδο μείωσης του σφάλματος αναγνώρισης σε επίπεδο λέξης αλλά και σημασίας, βρίσκεται ένα πολύ μεγάλο υπολογιστικό κόστος. Για να επιτευχθεί το χαμηλό σφάλμα αναγνώρισης, τα συστήματα αναγνώρισης ομιλίας τελευταίας γενιάς «αναγκάζονται» να «τρέχουν» σε μια τάξη μεγέθους μίας έως δύο φορές πιο αργά από τις απαιτήσεις πραγματικού χρόνου (real time), ενώ απαιτούν πολύ ισχυρούς επεξεργαστές και μεγάλα ποσά μνήμης.

Από τη στιγμή όμως που η δυνατότητα αναγνώρισης ομιλίας απαιτείται σε ολοένα μικρότερες συσκευές (π.χ. φορητές συσκευές), με μικρότερη υπολογιστική ισχύ αλλά και μειωμένες δυνατότητες μνήμης ανοίγονται δύο δρόμοι για την αντιμετώπιση των απαιτήσεων:

1.1.1 Κατανεμημένη αναγνώριση ομιλίας (Distributed Speech Recognition)

Στην περίπτωση αυτή, υπολογιστικά συστήματα με μεγάλη υπολογιστική ισχύ και δυνατότητες μνήμης αναλαμβάνουν το δύσκολο κομμάτι της αναγνώρισης λαμβάνοντας από τη συσκευή με τις περιορισμένες δυνατότητες (thin client) το σήμα της φωνής και στέλνοντας πίσω το κείμενο που αντιστοιχεί στα λεγόμενα του χρήστη.

Για να γίνει κάτι τέτοιο απαιτείται διασύνδεση (connectivity) αλλά και η κατάλληλη ποιότητα (Quality of Service) στη μετάδοση της ομιλίας, ώστε η όλη διαδικασία να γίνεται σε πραγματικό χρόνο. Προς αυτή την κατεύθυνση και την μείωση του όγκου πληροφορίας που πρέπει να μεταδοθεί από την συσκευή-πελάτη στον εξυπηρετητή έχουν ήδη γίνει κάποιες εργασίες στο επίπεδο της διανυσματικής κβαντοποίησης του διανύσματος χαρακτηριστικών (feature vector) που εξάγεται από το σήμα της ομιλίας [3].

1.1.2 Εύρεση τρόπων μείωσης των απαιτήσεων μνήμης και απαιτούμενης υπολογιστικής ισχύος χωρίς όμως να υπάρξει θυσία της επίδοσης αναγνώρισης.

Για να επιτευχθεί μεγαλύτερη ταχύτητα αναγνώρισης και ταυτόχρονα μικρές απαιτήσεις μνήμης χωρίς όμως να υπάρξει κόστος στην επίδοση του συστήματος (κάτι που φαίνεται αλληλοσυγκρουόμενο), πρέπει να υπάρξουν πολύ προσεκτικές επιλογές κατά την παραμετροποίηση των διαφόρων τμημάτων του συστήματος (ακουστικό μοντέλο, γλωσσικό μοντέλο, αλγόριθμος αναζήτησης στο χώρο των υποθέσεων κ.λπ.).

Πολλές τεχνικές έχουν εφαρμοστεί για να μειώσουν τις απαιτήσεις σε μνήμη, όπως για παράδειγμα χρησιμοποιώντας απλούστερα αλλά λιγότερο ακριβή μοντέλα.

Παράλληλα, υπάρχουν και πολλές τεχνικές για μείωση της υπολογιστικής πολυπλοκότητας, όπως για παράδειγμα εφαρμογή pruning μεθόδων για να μειωθεί το εύρος του ψαξίματος ή ο υπολογισμός των πιθανοφανειών κατάστασης (state likelihoods) μόνο για ένα μικρό υποσύνολο των πιο σχετικών με την συγκεκριμένη κατάσταση κατανομών (τεχνική shortlists).

Ένα μεγάλο ποσοστό του χρόνου αναγνώρισης (το οποίο όμως μειώνεται με τον όγκο του λεξιλογίου) έχει παρατηρηθεί ότι καταναλώνεται στον υπολογισμό των πιθανοφανειών των διαφόρων καταστάσεων (state likelihoods) του ακουστικού μοντέλου, κάτι που δεν είναι περίεργο μιας και πλέον χρησιμοποιούνται αρκετά πολύπλοκα ακουστικά μοντέλα με πολύ μεγάλο αριθμό παραμέτρων ώστε να καλυφθεί η απαιτούμενη ανάλυση του ακουστικού χώρου.

Προς την κατεύθυνση των μειωμένων απαιτήσεων υπολογιστικής ισχύος κινείται και η προσέγγιση των Κρυφών Μαρκοβιανών Μοντέλων (HMM) με διακριτά μείγματα (Discrete Mixture HMMs – DMHMM), όπου ο υπολογισμός των πιθανοφανειών κατάστασης ανάγεται σε πράξεις πάνω σε προϋπολογισμένους πίνακες (table lookups) και μάλιστα στο λογαριθμικό πεδίο ώστε να αντικαθίστανται οι απαιτούμενοι πολλαπλασιασμοί με προσθέσεις. Το κόστος στην περίπτωση αυτή είναι στην απαιτούμενη μνήμη μιας και οι διακριτές κατανομές γενικά χρειάζονται πολύ μεγαλύτερο αριθμό παραμέτρων για να προσδιοριστούν πλήρως σε σχέση με τις κανονικές (γκαουσιανές) κατανομές που χρησιμοποιούνται στα μοντέλα συνεχών κατανομών.

Αυτός ο μεγάλος αριθμός ελεύθερων παραμέτρων οδηγεί στα ακόλουθα προβλήματα:

- Μεγαλύτερες απαιτήσεις σε μνήμη
- Μικρότερη ταχύτητα αναγνώρισης (κάτι που αποφεύγεται στην περίπτωση των DMHMM)
- Απαίτηση για περισσότερα δεδομένα εκπαίδευσης
- Περισσότερα δεδομένα προσαρμογής στο περιβάλλον ή στον ομιλητή

Το πρόβλημα των μη επαρκών δεδομένων εκπαίδευσης ώστε να εκτιμηθεί σωστά το μεγάλο πλήθος των παραμέτρων του μοντέλου, αντιμετωπίστηκε στην εργασία [7] με την εφαρμογή διαφόρων *τεχνικών ομαλοποίησης* των διακριτών κατανομών, με αποτέλεσμα να επιτυγχάνεται τελικά η ίδια επίδοση ως προς τα συνεχή μοντέλα με αντίστοιχο αριθμό μειγμάτων.

1.2 Σκοπός της παρούσας εργασίας

Αν λοιπόν υπήρχε η δυνατότητα να μειωθεί ο αριθμός των ελεύθερων παραμέτρων των κρυφών μαρκοβιανών μοντέλων με μείγματα διακριτών κατανομών, θα μπορούσε να αντιμετωπιστεί τόσο το πρόβλημα της μνήμης όσο και το πρόβλημα της ταχύτητας. Παράλληλα, θα μπορούσε να επιτευχθεί και η απαίτηση για εύρωστη εκπαίδευση (robust training) με μικρό αριθμό δεδομένων εκπαίδευσης.

1.2.1 Ώρα για περισσότερο «μοίρασμα» (parameter tying)

Η πιο γνωστή προσέγγιση στο πρόβλημα της μείωσης του αριθμού των παραμέτρων στα ακουστικά μοντέλα είναι η «συσχέτιση» των παραμέτρων (parameter tying). Γίνεται προσπάθεια να εντοπιστούν «παρόμοιες» δομές, οι οποίες στη συνέχεια συνδέονται κατάλληλα ώστε να «μοιράζονται» την ίδια τιμή. Στο παρελθόν η παραπάνω μέθοδος έχει εφαρμοστεί σε διάφορα επίπεδα, όπως στο επίπεδο φωνημάτων (π.χ. generalized tri-phones, context-independent phones), καταστάσεων (π.χ. tied-state HMMs), κατανομών παρατήρησης (π.χ. tied mixtures, genones [1], phonetically-tied HMMs) ακόμα και σε επίπεδο

διανύσματος χαρακτηριστικών (περίπτωση επιλογής υποδιανυσμάτων κατά τη διαδικασία vector quantization).

Στην παρούσα εργασία, προσπαθούμε να επεκτείνουμε την παραπάνω προσέγγιση στο επίπεδο των κατανομών που αντιστοιχούν στα διάφορα υποδιανύσματα του διανύσματος χαρακτηριστικών (feature vector) σε ακουστικά μοντέλα μειγμάτων διακριτών κατανομών (DMHMM), και να διερευνήσουμε τη δυνατότητα μείωσης των ελεύθερων παραμέτρων διατηρώντας παράλληλα την απαιτούμενη ανάλυση του ακουστικού χώρου ώστε να μη μειωθεί η επίδοση της αναγνώρισης.

Για το σκοπό αυτό, αρχικά προτείνονται κάποιες τεχνικές, οι οποίες στη συνέχεια εφαρμόζονται και εξετάζεται η επίδοσή τους. Στο τέλος παρουσιάζονται και σχολιάζονται τα αποτελέσματα που προκύπτουν.

2 Αναγνώριση ομιλίας με τη βοήθεια Κρυφών Μαρκοβιανών Μοντέλων (HMM)

2.1.1 Το πρόβλημα της αναγνώρισης ομιλίας

Ορίζοντας το πρόβλημα της αναγνώρισης (αποκωδικοποίησης) φωνής (ομιλίας), λέμε ότι δίδεται ένα ακουστικό σήμα ***X*** και ζητείται να καθοριστεί με βάση κάποιο κριτήριο ότι προφέρθηκε η ακολουθία λέξεων ***W***. Κριτήριο της αποκωδικοποίησης, όπως και σε ένα τυπικό ψηφιακό τηλεπικοινωνιακό σύστημα, είναι η *ελαχιστοποίηση της πιθανότητας σφάλματος*.

Οι στατιστικές μέθοδοι αναγνώρισης προϋποθέτουν την ύπαρξη κάποιου αντίστοιχου στατιστικού μοντέλου για τον υπολογισμό της ζητούμενης πιθανότητας (ή συνάρτησης πιθανοφάνειας). Η πιθανότητα σφάλματος ελαχιστοποιείται αν αποκωδικοποιήσουμε στην ακολουθία εκείνη λέξεων $\hat{\underline{W}}$ για την οποία μεγιστοποιείται η *a-posteriori* πιθανότητα δεδομένου ότι ο αναγνωριστής (αποκωδικοποιητής) "έλαβε" την ακολουθία ακουστικών παρατηρήσεων $\underline{X} = [x_1, x_2, \ldots, x_T]$.

Εφαρμόζοντας τον κανόνα του Bayes, έχουμε:

$$\hat{W} = \arg\max_{\underline{w}} P(\underline{W} \mid \underline{X}) = \arg\max_{\underline{w}} \frac{P(\underline{W})P(\underline{X} \mid \underline{W})}{P(\underline{X})}$$
$$= \arg\max_{\underline{w}} P(\underline{W}) \cdot P(\underline{X} \mid \underline{W}) \quad ,$$

όπου ο τελεστής *argmax* συμβολίζει το όρισμα που μεγιστοποιεί την αντίστοιχη ποσότητα.

Αυτή η εξίσωση δείχνει ότι για να βρεθεί η πιο πιθανή ακολουθία λέξεων ***W***, πρέπει να βρεθεί η ακολουθία αυτή 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 ακολουθίας λέξεων $\underline{W}$ 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 $\underline{X}$ για μια δεδομένη ακολουθία λέξεων $\underline{W}$, και αυτή η πιθανότητα είναι γνωστή ως ***ακουστικό μοντέλο*** (*acoustic model*).

Γενικά η εύρεση του ορίσματος στην παραπάνω σχέση, εμπλέκει μια διαδικασία αναζήτησης ανάμεσα σε ένα μεγάλο αριθμό πιθανών ακολουθιών λέξεων.

2.1.2 Γλωσσικό Μοντέλο

Το γλωσσικό μοντέλο είναι η κατανομή πιθανοτήτων συνδυασμών λέξεων, η οποία προσπαθεί να απεικονίσει τη συχνότητα με την οποία κάθε τέτοιος συνδυασμός αναμένεται να απαντηθεί σε κάποιο κείμενο.

Το πιο ευρέως χρησιμοποιούμενο γλωσσικό μοντέλο είναι το trigram μοντέλο. Στην περίπτωση αυτή, γίνεται η θεώρηση ότι η πιθανότητα εμφάνισης μιας λέξης εξαρτάται μόνο από τις δύο προηγούμενες λέξεις. Έτσι:

$$\Pr(w) = \prod_{i=1}^{n} \Pr(w_i \mid w_{i-1} w_{i-2})$$

2.1.3 Ακουστικό Μοντέλο

Η γλωσσική μονάδα που αποτελούσε, αρχικά, τη βάση του ακουστικού μοντέλου ήταν η λέξη. Όμως, για να υπάρχει η δυνατότητα γενίκευσης και να καλύπτονται και λέξεις που δεν παρατηρήθηκαν στα δεδομένα εκπαίδευσης, χρησιμοποιούνται μικρότερες γλωσσικές μονάδες όπως το ***φώνημα*** (*phoneme*) ή η συλλαβή.

Το σύστημά μας χρησιμοποιεί triphones αντί για φωνήματα και έτσι μοντελοποιεί τα διάφορα φωνήματα λαμβάνοντας υπ' όψη τα γειτονικά τους. Στην περίπτωση αυτή κάθε φώνημα αντιστοιχεί σε ένα διαφορετικό κρυφό μαρκοβιανό μοντέλο (HMM) για κάθε διαφορετικό ζεύγος από αριστερούς και δεξιούς γείτονες.

Με τον τρόπο αυτό, μπορεί να μοντελοποιηθεί το γεγονός ότι τα συμφραζόμενα μπορούν να αλλοιώσουν σημαντικά τον τρόπο με τον οποίο ένα φώνημα μπορεί να ειπωθεί.

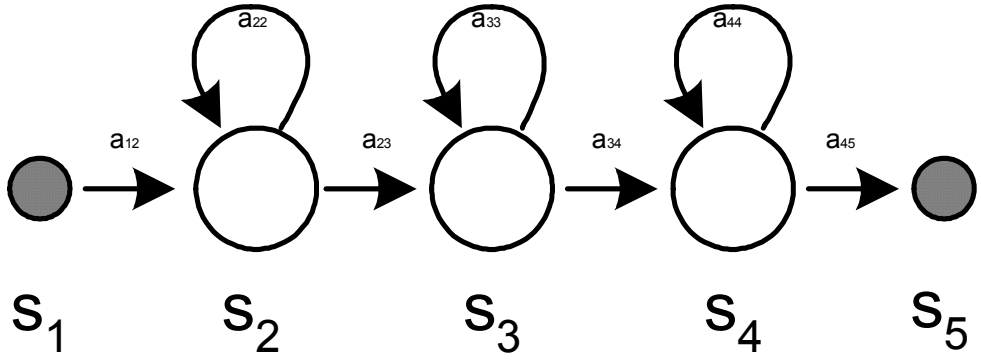
Από στατιστικής πλευράς, ένας κατάλογος από πιθανοτικά μοντέλα βασικών φωνητικών μονάδων (π.χ. triphones) χρησιμοποιείται για να αναπαραστήσει λέξεις (με τη βοήθεια ενός λεξικού προφορών).

Μία ακολουθία από ακουστικές παρατηρήσεις, προερχόμενη μετά από επεξεργασία από το σήμα φωνής, αντιμετωπίζεται ως συνδυασμός στοιχειωδών διαδικασιών που περιγράφονται από Κρυφά Μαρκοβιανά μοντέλα (ΚΜΜ) ή Hidden Markov Models (HMM).

Επιστρέφοντας στην ανάλυση για την αναγνώριση ομιλίας σε ακουστικό κανάλι, οι ακολουθίες των HMMs που χρειάζονται για να αναπαραστήσουν την αρχική έκφραση συνδέονται σειριακά μεταξύ τους σύμφωνα με το λεξικό προφορών ώστε να

σχηματίσουν ένα πιο σύνθετο μοντέλο που αναπαριστά την ακολουθία ***W*** και στο επόμενο στάδιο υπολογίζεται η πιθανότητα να παράγει αυτό το σύνθετο μοντέλο την παρατηρούμενη ακολουθία ακουστικών παρατηρήσεων ***X***. Η πιθανότητα αυτή είναι η ζητούμενη πιθανότητα *P(X|W).*

2.2 Λίγα λόγια για τα Κρυφά Μαρκοβιανά Μοντέλα - Hidden Markov Models (HMM)



Σχήμα 2-1 Αναπαράσταση καταστάσεων και μεταβάσεων για ένα μοντέλο HMM

Η μέθοδος των HMM είναι η περισσότερο διαδεδομένη μέθοδος μοντελοποίησης των φασματικών ιδιοτήτων των τμημάτων της φωνής για συστήματα μεγάλου λεξιλογίου και ανεξάρτητα από ομιλητή.

Ένα HMM είναι ένα σύνολο από καταστάσεις (states) συνδεδεμένες με μεταβάσεις. Κάθε κατάσταση συνδέεται με μια κατανομή εξόδου, η οποία προσπαθεί να μοντελοποιήσει στατιστικά την τιμή κάποιων χαρακτηριστικών παραμέτρων (features), που αντιστοιχούν σε κάθε τμήμα (frame) ομιλίας.

Τα μοντέλα HMM είναι γενίκευση των γνωστών μοντέλων Markov: στα μοντέλα Markov κάθε μετάβαση από κατάσταση σε κατάσταση έχει ως συνέπεια την έξοδο ενός συμβόλου ντετερμινιστικά. Στα μοντέλα HMM κάθε τέτοια μετάβαση σχετίζεται με μια πιθανοτική κατανομή πάνω σε ένα σύνολο πιθανών «συμβόλων εξόδου».

Ένα κρυφό μαρκοβιανό μοντέλο (HMM) ορίζεται συνοπτικά από τα εξής στοιχεία:

- N: Αριθμός καταστάσεων.
- Πλήθος συμβόλων εξόδου: Πλήθος διακριτών συμβόλων που μπορούν να παρατηρηθούν ανά κατάσταση: M για διακριτά HMMs ή άπειρο για συνεχή HMMs.
- A: ένας NxN πίνακας μεταβάσεων,όπου a_{ij} είναι η πιθανότητα μετάβασης από την κατάσταση i στην κατάσταση j.
  $A = \{a_{ij}\}$, όπου $a_{ij} = P(q_{t+1} = j \mid q_t = i)$, με $1 \le i \le N$
- Κατανομές εξόδου (output distributions) για κάθε κατάσταση j:
  Σε κάθε χρονική στιγμή δημιουργείται μία παρατήρηση (τυχαίο διάνυσμα ή τυχαία μεταβλητή (διακριτή ή συνεχής)) με βάση μία κατανομή που εξαρτάται από την κατάσταση στην οποία βρισκόμαστε.
  Για διακριτά HMMs είναι $B = \{b_j(x_t)\}$, όπου το μέγεθος $b_j(x_t) = P(x_t \mid q_t = j)$ είναι η κατανομή εξόδου με $1 \le j \le N$ και $1 \le x_t \le M$.
- 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 $q_t \in \{1,2,\ldots,N\}$.

Για τα A,B και π πρέπει να ισχύουν οι παρακάτω σχέσεις:

$$a_{ij} \ge 0, b_i \ge 0, \forall i, j$$
$$\sum_j a_{ij} = 1, \forall i$$
$$\sum_k b_i(k) = 1, \forall i, 1 \le k \le M$$

Για τον πλήρη ορισμό ενός HMM μοντέλου απαιτείται ο καθορισμός των παραπάνω **παραμέτρων** του μοντέλου. Για συντομία χρησιμοποιείται ο συμβολισμός:

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

Βασικό χαρακτηριστικό των HMM μοντέλων (από όπου προκύπτει και το όνομά τους), είναι ότι παρόλο που η έξοδός τους είναι παρατηρήσιμη, δεν είναι παρατηρήσιμη (είναι **κρυφή**) η ακολουθία των καταστάσεων.

Σκοπός της αναγνώρισης ομιλίας όμως είναι να βρεθεί η πιθανότερη ακολουθία καταστάσεων (και άρα η ακολουθία των στοιχειωδών γλωσσικών μονάδων) που θα μπορούσε να έχει «γεννήσει» τα σύμβολα που παρατηρούμε στην έξοδο.

Έτσι με τα μοντέλα HMM συνδέονται τα παρακάτω **τρία βασικά** προβλήματα:

1. Ο υπολογισμός της πιθανότητας μιας ακολουθίας παρατηρήσεων.
2. Ο υπολογισμός της πιο πιθανής ακολουθίας καταστάσεων κατά την αποκωδικοποίηση (αναγνώριση).
3. Η αυτόματη εκμάθηση (εκπαίδευση) των παραμέτρων τους από δεδομένα εκπαίδευσης.

Στη φάση της εκπαίδευσης, η οποία μας ενδιαφέρει στην εργασία αυτή, σκοπός είναι να εκτιμηθούν οι τιμές των παραμέτρων $\lambda = (\mathrm{A}, \mathrm{B}, \pi)$ από ένα σύνολο δεδομένων εκπαίδευσης. Το σύνολο αυτό αποτελείται από προτάσεις (τμήματα ομιλίας) για τις οποίες είναι γνωστή η αντίστοιχη ακολουθία γλωσσικών μονάδων, δηλαδή τι έχει προφερθεί (labeled data). Έτσι οι αλγόριθμοι εκπαίδευσης ανήκουν στην ευρύτερη κατηγορία της εποπτευόμενης μάθησης (supervised training).

Μια σειρά επαναληπτικών αλγορίθμων πρέπει να τρέξει πάνω στα δεδομένα εκπαίδευσης (*training data*), ώστε να διαθέτουμε τα απαραίτητα στοιχεία για την εκτίμηση των παραμέτρων του μοντέλου.

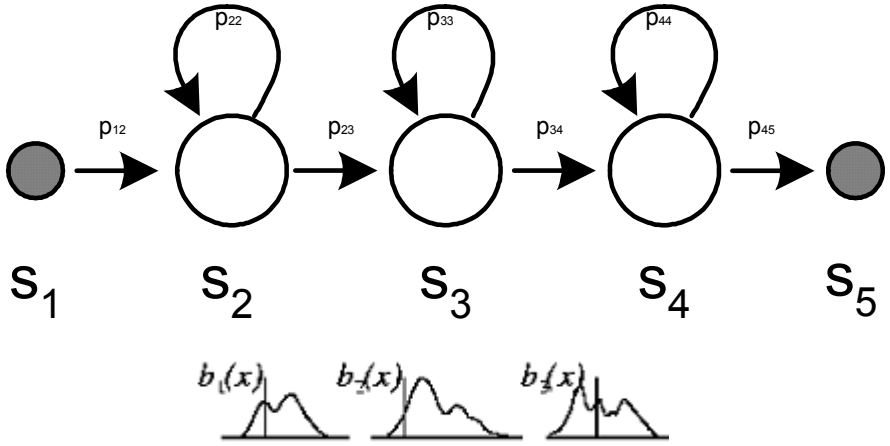
Ο **αλγόριθμος *Forward-Backward*** είναι μία αποδοτική επαναληπτική μέθοδος για τον υπολογισμό της πιθανότητας να βρισκόμαστε σε μια συγκεκριμένη κατάσταση σε μια συγκεκριμένη χρονική στιγμή. **Ο αλγόριθμος *Baum-Welch,*** είναι μία διαδικασία για την εύρεση εκτιμητριών μέγιστης πιθανοφάνειας των παραμέτρων ενός HMM.

Βέβαια γίνεται χρήση και του αλγορίθμου Viterbi που παίζει πολύ σημαντικό ρόλο στη διαδικασία της αναγνώρισης-αποκωδικοποίησης. Ο αλγόριθμος αυτός βασίζεται σε μεθόδους δυναμικού προγραμματισμού και με τη βοήθειά του βρίσκεται η βέλτιστη ακολουθία καταστάσεων.

2.3 Είδη 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 σε:

2.3.1 Συνεχή HMMs (Continuous HMMs)



Σχήμα 2-2 Αναπαράσταση ενός συνεχούς HMM

Στην περίπτωση αυτή, όπως φαίνεται και στο παραπάνω σχήμα, οι κατανομές εξόδου είναι από κοινού συναρτήσεις πυκνότητας πιθανότητας ενός τυχαίου διανύσματος x_t με τιμή:

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

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

Στην κατηγορία αυτή ανήκουν και τα κρυφά μαρκοβιανά μοντέλα μειγμάτων κανονικών κατανομών (Gaussian Mixture HMMs). Η κατανομή εξόδου για κάθε κατάσταση έχει τη μορφή:

$$b_j(x_t) = \sum_{k=1}^{M} c_{jk} N(x_t, \overline{\mu}_{jk}, \Sigma_{jk}),$$

όπου M ο αριθμός των μειγμάτων (mixtures) ανά κατάσταση και c_{jk} το βάρος για το k-οστό μείγμα. Ακόμη, $N(x_t, \overline{\mu}_{jk}, \Sigma_{jk})$ είναι η πολυδιάστατη κανονική (γκαουσιανή) κατανομή με παραμέτρους κατανομής τις ποσότητες $\underline{\mu}_{jk}$ (μέσες τιμές) και Σ_{jk} (διαγώνιος πίνακας συμμεταβλητότητας).

Τα βάρη c_{jk} πρέπει να ικανοποιούν τις σχέσεις:

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

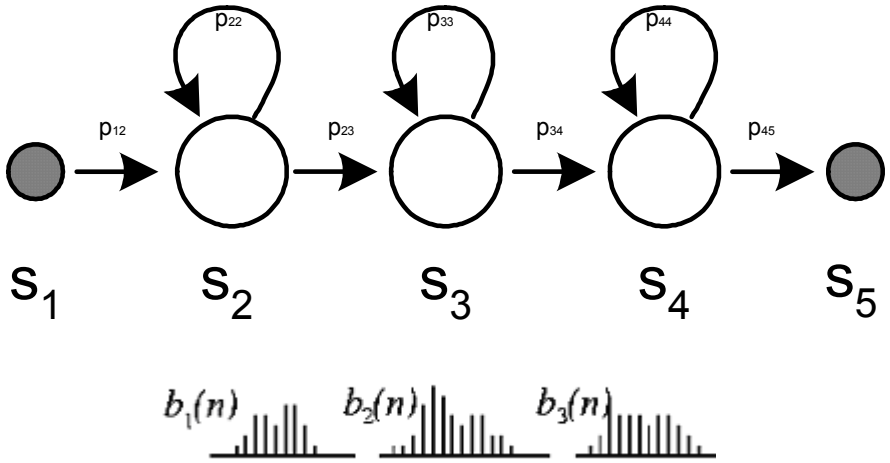
$$c_{jk} \ge 0 \quad , 1 \le j \le N \ , 1 \le k \le M \ ,$$

ώστε να ισχύει:

$$\int_{-\infty}^{+\infty} b_j(x_t)\ dx_t = 1$$

2.3.2 Διακριτά HMMs (Discrete HMMs)

Αν η μοντελοποιούμενη διαδικασία $\{x_t\}$ είναι διακριτή με $x_t \in \{1,...,M\}$, τότε και η κατανομή εξόδου $b_j(x_t)$ είναι διακριτή, όπως φαίνεται και στο ακόλουθο σχήμα:



Σχήμα 2-3 Αναπαράσταση ενός διακριτού HMM

Και στην περίπτωση, βέβαια, αυτή θα ισχύει: $\sum_{k=1}^{M} b_j(k) = 1$

2.4 Σύντομη σύγκριση συνεχών και διακριτών HMMs

Το βασικό πλεονέκτημα των συνεχών HMM μοντέλων είναι η απευθείας μοντελοποίηση των παραμέτρων της φωνής, έτσι ώστε αποφεύγεται ο κβαντισμός. Τα συνεχή HMM μειονεκτούν όμως τόσο στο μεγαλύτερο χρόνο εκπαίδευσης όσο και στην πολυπλοκότητα και την αυξημένη υπολογιστική ισχύ αλλά και τον αυξημένο χρόνο κατά την αναγνώριση.

Ενώ δηλαδή για τα διακριτά μοντέλα ο υπολογισμός μιας πιθανότητας εξόδου από μια κατάσταση έχει να κάνει μόνο με ένα μικρό αριθμό από table-lookups και την πρόσθεσή τους, στα συνεχή απαιτούνται πολλοί υπολογισμοί για τον υπολογισμό των πολυδιάστατων κανονικών κατανομών (γκαουσιανών).

Η πολυπλοκότητα αυξάνεται ακόμη πιο πολύ στα συστήματα gaussian mixture HMM εξαιτίας του μεγαλύτερου αριθμού γκαουσιανών. *Όμως η χρήση των τελευταίων επιβάλλεται από την καλύτερη μοντελοποίηση που παρέχουν κυρίως όσον αφορά στα συστήματα αναγνώρισης ανεξάρτητα από ομιλητή.*

Για τα διακριτά μοντέλα το βασικό πρόβλημα είναι τα σφάλματα κβαντισμού, μιας και για να αποθηκευτεί η αντίστοιχη πληροφορία με τα συνεχή, απαιτείται πολύ μεγαλύτερος αριθμός παραμέτρων και άρα περισσότερος αποθηκευτικός χώρος (μνήμη). Έτσι, για καλύτερη ανάλυση του ακουστικού χώρου και άρα μικρότερο σφάλμα κατά την κβάντιση, απαιτείται απαγορευτικός πολλές φορές αριθμός παραμέτρων.

3 Εκπαίδευση ακουστικών μοντέλων HMM

3.1 Υλοποίηση της διαδικασία εκπαίδευσης

Το περιβάλλον το οποίο χρησιμοποιήθηκε για τα πειράματα και την αξιολόγηση των εξεταζόμενων τεχνικών ομαδοποίησης αποτελεί μετεξέλιξη του εκπαιδευτή (trainer) του συστήματος SRI DECIPHER.

Η διαδικασία εκπαίδευσης είναι μια σταδιακή (iterative) διαδικασία που περιλαμβάνει την εκπαίδευση των παρακάτω τύπων μοντέλων μέχρι την «παραγωγή» HMM με μείγματα διακριτών κατανομών (DM-HMMs).

3.1.1 Μοντέλα PTM (Phonetically Tied Mixtures)

Στα PTM μοντέλα, για όλα τα tri-phones που έχουν το ίδιο κεντρικό φώνημα (δηλαδή για καταστάσεις της μορφής ***[phoneme]***-*) έχουμε κοινές πολυδιάστατες κανονικές κατανομές.

Για παράδειγμα, η μορφή της κατανομής εξόδου της κατάστασης b[a]c-1 είναι:

$$p(x \mid b[\alpha]c-1) = \sum_i \lambda_{b[a]c-1}(i) \cdot \mathrm{N}_{[a]}(),$$

και της κατάστασης [a]-1 του context-independent μοντέλου είναι:

$$p(x \mid [\alpha]-1) = \sum_i \lambda_{[\alpha]-1}(i) \cdot \mathrm{N}_{[a]}().$$

Βλέπουμε δηλαδή ότι χρησιμοποιούνται κοινές κανονικές κατανομές αλλά με διαφορετικά βάρη για κάθε κατάσταση. Ο αριθμός τέτοιων ομάδων κανονικών κατανομών που επαναχρησιμοποιούνται εξαρτάται από τον αριθμό των βασικών φωνημάτων.

3.1.2 Μοντέλα Genones

Στα μοντέλα genones, η κατανομή εξόδου για μια κατάσταση είναι της μορφής:

$$b_j(\overline{x}_t) = \sum_{k=1}^{M} c_{jk} \cdot N(\overline{x}_t, \mu_{jk}, \Sigma_{jk}) .$$

Οι διάφορες καταστάσεις (και άρα κατανομές εξόδου μιας και κάθε κατάσταση συνδέεται με μια κατανομή εξόδου) ***ομαδοποιούνται*** και κάθε ομάδα χρησιμοποιεί κοινά μείγματα κανονικών κατανομών. Ένας αριθμός από (τυπικά M={64,32}) τέτοια μείγματα αποτελούν ένα genone. Κάθε κατάσταση έχει διαφορετικά βάρη για τα μείγματα-μέλη του genone, παρόλο που με το ίδιο genone μπορούν να συνδέονται και πολλές άλλες καταστάσεις.

Όσο για τον αριθμό των καταστάσεων, χρησιμοποιούνται context-dependent triphones, δηλαδή διαφορετικά HMM για κάθε κεντρικό φώνημα όταν αυτό έχει διαφορετικά γειτονικά φωνήματα. Κάθε triphone μοντελοποιείται από ένα HMM τριών καταστάσεων (tri-state HMM), ενώ υπάρχουν και δυο ακόμα βοηθητικές καταστάσεις για τη σύνδεση με άλλα triphones.

Κατά τη διαδικασία της αναγνώρισης πολλά τέτοια στοιχειώδη HMM που αναπαριστούν τα triphones συνδυάζονται για να μοντελοποιήσουν λέξεις και προτάσεις *W* ώστε να εξεταστούν οι διάφορες υποθέσεις.

Για τα triphones χρησιμοποιείται ο συμβολισμός α[β]γ και κάθε κατάσταση του αντίστοιχου HMM συμβολίζεται από το όνομα του triphone και τον αριθμό της κατάστασης ως εξής: α[β]γ-{0,1,2}.

Ένας τυπικός αριθμός βασικών κεντρικών φωνημάτων σε μια γλώσσα είναι 40. Αυτό σημαίνει ότι ο αριθμός των δυνατών συνδυασμών τριάδων φωνημάτων (triphones) είναι πολύ μεγάλος: 40^3.

Επειδή πολλά από τα triphones είναι απίθανο να συναντηθούν κατά τη διαδικασία της αναγνώρισης λόγω των φωνοτακτικών περιορισμών κάθε γλώσσας, ανάλογα με το σύνολο δεδομένων εκπαίδευσης "επιβιώνει" ένα μικρό μέρος που είναι πιο πιθανό να απαντηθούν στο συγκεκριμένο task. Έτσι υπάρχουν περίπου 12.000 (ακριβώς: 11977) δυνατές καταστάσεις στο μοντέλο μας.

Αναλυτική περιγραφή της οργάνωσης και του τρόπου εκπαίδευσης μοντέλων της μορφής αυτής μπορεί να βρεθεί στην εργασία [1].

3.1.3 Το μοντέλο Discrete-Mixture HMM (DM-HMM)

Βασικό χαρακτηριστικό των DMHMM μοντέλων είναι ότι δεν υπάρχει (όπως συμβαίνει στα απλά διακριτά HMM) η περιοριστική υπόθεση της ανεξαρτησίας των διαφόρων features ή υποδιανυσμάτων. Με τη χρήση των μειγμάτων μοντελοποιείται η παρατηρούμενη αυτή συσχέτιση.

Ακόμη, αφού η διακριτή κατανομή μπορεί να αναπαραστήσει καλύτερα μια αυθαίρετη κατανομή (όπως είναι αυτή των δεικτών στα centroids), αναμένεται να πλεονεκτούν στην ακριβή αναπαράσταση ως προς τα mixtures συνεχών γκαουσιανών κατανομών.

Το σημαντικότερο πλεονέκτημα των DMHMM είναι ότι επιτυγχάνουν απόδοση στα επίπεδα των συνεχών αλλά παρέχουν τόσο μείωση στο χρόνο της εκπαίδευσης όσο και κυρίως στο χρόνο αναγνώρισης εξαιτίας του μειωμένου αριθμού των υπολογισμών. Το τελευταίο επιτυγχάνεται επειδή για να υπολογιστεί για το μοντέλο αυτό μια πιθανότητα (στο λογαριθμικό πεδίο) απαιτείται μια απλή αναζήτηση σε έναν πίνακα προϋπολογισμένων ποσοτήτων (table look-up).

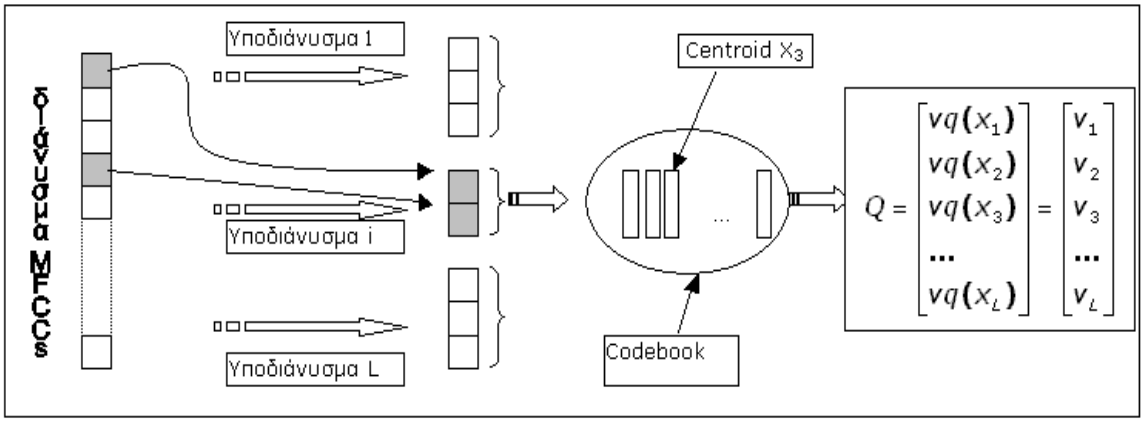
Το front-end σε ένα σύστημα αναγνώρισης ομιλίας έχει τη δυνατότητα να μας παράσχει ένα (τουλάχιστον) διάνυσμα ακουστικών παρατηρήσεων (feature vector) για το εξεταζόμενο frame ομιλίας (ένα frame είναι ένα «παράθυρο» της τάξης των 10 msec στο πεδίο του χρόνου). Στο σύστημά μας προκύπτει ένα feature με την επεξεργασία του σήματος ομιλίας από το front-end όπου λαμβάνεται, μετά από βραχέως χρόνου ανάλυση Fourier (Short Time Fourier Transform) και σταδιακή επεξεργασία σε κάθε τμήμα (frame) της ομιλίας, ένα διάνυσμα cepstral παραμέτρων (Mel-Frequency Cepstral Coefficients – MFCCs). Ο αριθμός των στοιχείων του διανύσματος είναι 27 και είναι άμεσα συνδεδεμένος με τη διάσταση των πολυδιάστατων κανονικών κατανομών και με τον αριθμό των παραμέτρων για την πλήρη αναπαράστασή τους.

3.1.3.1 Διανυσματική κβαντοποίηση - (Sub)Vector Quantization

Το διάνυσμα χαρακτηριστικών (feature vector) , που προέρχεται από το front-end, αφού χωριστεί σε υποδιανύσματα, κωδικοποιείται με ξεχωριστά codebooks και με διαφορετικό αριθμό bits για κάθε υποδιάνυσμα (Vector Quantization). Το αποτέλεσμα της όλης διαδικασίας είναι ένα διάνυσμα μήκους ίσου με τον αριθμό των υποδιανυσμάτων (subvectors), που περιέχει είτε τα centroids (οπότε μιλάμε για

centroid-coded κβαντισμό) είτε τους δείκτες στα αντίστοιχα (για κάθε υποδιάνυσμα) centroids (index-coded κβαντισμός).

Στο σχήμα (3.1) φαίνεται εποπτικά η διαδικασία διανυσματικής κβαντοποίησης.



Σχήμα 3-1 Σχηματική αναπαράσταση της διαδικασίας διανυσματικής κβαντοποίησης

Μέσω της διανυσματικής κβαντοποίησης επιτυγχάνονται:

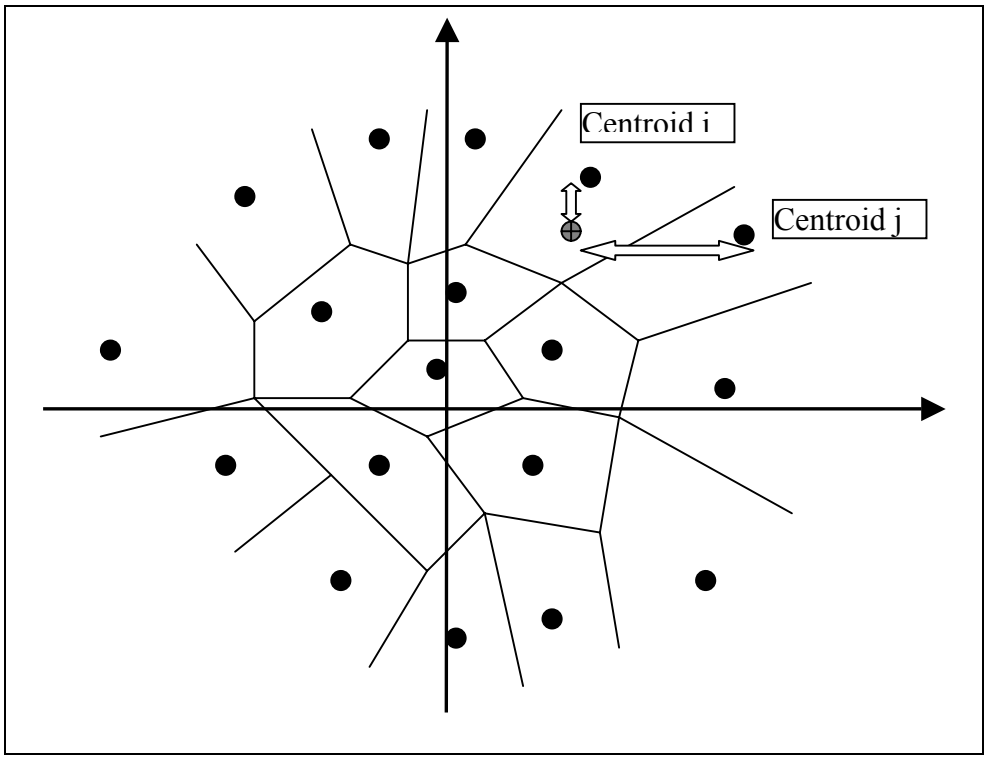
- Μεγάλη ανάλυση του ακουστικού χώρου.
- Μικρό σφάλμα κβαντισμού.
- Σημαντική μείωση του ρυθμού μετάδοσης.

3.1.3.2 Διαδικασία δημιουργίας codebooks – Εκτίμηση centroids

Για την διαδικασία της διανυσματικής κβαντοποίησης σημαντικό ρόλο παίζει ο υπολογισμός των codebooks και των centroids που αυτά «περιέχουν» (ένα codebook περιέχει όλα τα centroids που έχουμε στο χώρο). Με τη διαδικασία αυτή ουσιαστικά διαχωρίζεται ο N-διάστατος (εδώ N=27) χώρος σε M περιοχές. Κάθε μια από αυτές τις περιοχές έχει ένα σημείο αναφοράς, το centroid. Όλα τα σημεία που ανήκουν σε μια περιοχή λέμε ότι μπορούν να κβαντιστούν στο centroid αυτής της περιοχής. Το κριτήριο βάσει του οποίου τα σημεία του χώρου κβαντίζονται σε κάποιο centroid και σχηματίζουν έτσι μια περιοχή, είναι μια μετρική απόστασης.

Για την εκπαίδευσή των codebooks (δηλαδή την εκτίμηση της τιμής του centroid κάθε περιοχής) χρησιμοποιείται ένας επαναληπτικός k-means αλγόριθμος. Αναλυτική

περιγραφή της διαδικασίας αυτής μπορεί να βρεθεί στην εργασία [3]. Στο παρακάτω σχήμα φαίνεται εποπτικά ο διαχωρισμός ενός 2-διάστατου χώρου (N=2) σε περιοχές και η συσχέτιση ενός centroid με κάθε περιοχή. Κάθε centroid είναι ένα σημείο στον N-διάστατο χώρο.



Σχήμα 3-2 Παράδειγμα υπολογισμού codebook για N=2. Το σύνολο των N-διάστατων centroids αποτελεί το codebook

Η διαδικασία της εκτίμησης των codebooks και της εύρεσης των centroids γίνεται για κάθε συστατικό υποδιάνυσμα που περιέχεται στο τελικό διάνυσμα που προκύπτει από τη διαδικασία διανυσματικού κβαντισμού.

Σημαντική παράμετρος της όλης διαδικασίας είναι ο αριθμός των centroids κάθε περιοχής (που αντιστοιχεί σε κάθε υποδιάνυσμα), κάτι που επηρεάζει τον αριθμό των bits που θα χρησιμοποιηθούν για την κωδικοποίηση του δείκτη προς το centroid στην περίπτωση του index-coded κβαντισμού.

Για την ελαχιστοποίηση τόσο του σφάλματος κβαντισμού αλλά και του αριθμού των χρησιμοποιούμενων bits για την κωδικοποίηση του τελικού διανύσματος που

προκύπτει από την διανυσματική κβαντοποίηση εφαρμόζονται τεχνικές bit-allocation. Μια trial-and-error τεχνική του τύπου αυτού παρουσιάζεται στην εργασία [3].

3.1.3.3 Μοντέλα μειγμάτων διακριτών κατανομών

Βάσει της παραπάνω κωδικοποίησης και για την εκμετάλλευση των πλεονεκτημάτων των μοντέλων διακριτών κατανομών και των genonic μοντέλων προτάθηκαν στην εργασία [2] τα μοντέλα μειγμάτων διακριτών κατανομών (DMHMMs). Στα μοντέλα αυτά η έξοδος (το παρατηρούμενο «σύμβολο») κάθε κατάστασης (ή κάθε μετάβασης από κατάσταση σε κατάσταση βάσει μερικών ορισμών) είναι ένα διάνυσμα από centroids ή από δείκτες σε centroids.

Η μορφή της κατανομής εξόδου για κάθε κατάσταση του μοντέλου DMHMM είναι:

$$b_j(x_t) = \sum_{k=1}^{M} c_{jk} \prod_{i=1}^{L} P_{jki}(vq(x_{it}))$$

,

όπου η συνάρτηση $vq(x_{it})$ επιστρέφει το δείκτη του centroid στο οποίο κβαντίζεται το υποδιάνυσμα x_{it}.

Τα βάρη c_{jk} πρέπει να ικανοποιούν τις σχέσεις:

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

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

Σύμφωνα με την παραπάνω μορφή κάθε κατάσταση συνδέεται μέσω κάποιων βαρών c_{jk} (mixture weights) με ένα άθροισμα (μείγμα) από «πολυδιάστατες διακριτές κατανομές». Ο αριθμός των πολυδιάστατων κατανομών ανά μείγμα (mixture) είναι τυπικά 32 ή 8. Κάθε πολυδιάστατη κατανομή είναι ένα γινόμενο διακριτών κατανομών (vector distributions) για κάθε ένα από τα (τυπικά 12) υποδιανύσματα (subvectors).

Βασική υπόθεση που «κρύβεται» πίσω από το γινόμενο των διακριτών κατανομών των υποδιανυσμάτων είναι η υπόθεση ανεξαρτησίας τους. Μέρος της διαδικασίας διανυσματικής κβαντοποίησης είναι και η εξασφάλιση της παραπάνω υπόθεσης μέσω της επιλογής κατάλληλων υποδιανυσμάτων του αρχικού διανύσματος των MFCC συντελεστών.

Στην παραπάνω σχέση, c_{jk} είναι ο συντελεστής (βάρος) της πολυδιάστατης κατανομής k στην κατάσταση j και $P_{jki}(vq(x_{it}))$ η πιθανότητα του διακριτού συμβόλου (centroid) $vq(x_{it})$ για το υποδιάνυσμα i της πολυδιάστατης κατανομής k. Για να διατηρηθεί ο αριθμός των παραμέτρων του μοντέλου σε χαμηλά επίπεδα, ο αριθμός των μειγμάτων είναι πολύ μικρότερος από τον αριθμό των καταστάσεων. Έτσι απαιτείται κάποια ομαδοποίηση (clustering) των διαφόρων καταστάσεων σε ομάδες, οι οποίες συνδέονται με το ίδιο μείγμα από πολυδιάστατες διακριτές κατανομές αλλά με διαφορετικά βάρη c_{jk} για κάθε κατάσταση.

Μιας και κάθε υποδιάνυσμα τυπικά κωδικοποιείται χρησιμοποιώντας 3-10 bits, κάθε διακριτή κατανομή των δεικτών στα centroids είναι ένα διάνυσμα πιθανότητας (probability vector) που αποτελείται από 8 (2^3) μέχρι 1024 (2^{10}) στοιχεία.

3.1.3.4 Διακριτοποίηση ακουστικών μοντέλων

Τα μοντέλα μειγμάτων διακριτών κατανομών προκύπτουν μετά από μια διαδικασία διακριτοποίησης των αντίστοιχων μοντέλων μειγμάτων συνεχών (κανονικών) κατανομών [2]. Στη διαδικασία αυτή χρησιμοποιείται η ίδια ομαδοποίηση των cepstrum παραμέτρων του αρχικού feature vector βάσει της οποίας γίνεται και η διανυσματική κβαντοποίηση ενώ γίνεται χρήση της αντιστοιχίας των μορφών των κατανομών εξόδου στις δύο περιπτώσεις:

$$CDHMMs: \quad b_j(x_t) = \sum_{k=1}^{M} c_{jk} \cdot N(x_t, \mu_{jk}, \Sigma_{jk})$$

$$DMHMMs: \quad b_j(x_t) = \sum_{k=1}^{M} c_{jk} \cdot \prod_{i=1}^{L} P_{jki}(vq(x_{it}))$$

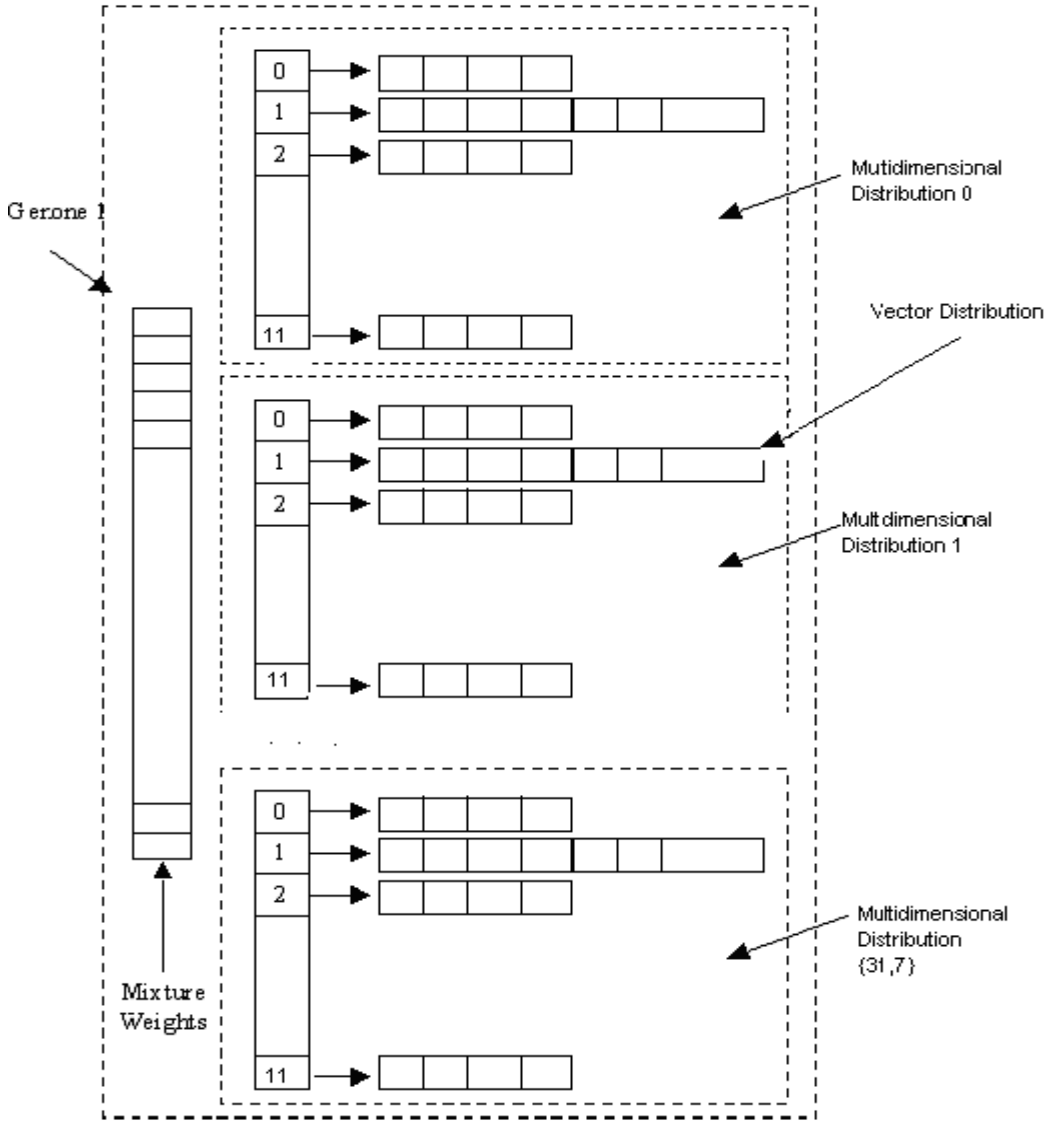
Στο τέλος της διαδικασίας ο αριθμός των μειγμάτων παραμένει ο ίδιος με το αντίστοιχο genonic μοντέλο, όμως αλλάζει πλέον η μορφή των συστατικών κατανομών κάθε μείγματος. Αντί για την πολυδιάστατη Γκαουσιανή κατανομή, εισάγεται μια «πολυδιάστατη διακριτή» κατανομή, αποτελούμενη από L (=αριθμός υποδιανυσμάτων κατά την διανυσματική κβαντοποίηση) υποκατανομές (διανυσματικές κατανομές) με διαφορετικό αριθμό στοιχείων η καθεμιά.

Για να εκτιμηθούν οι *l* τιμές για το i υποδιάνυσμα χρησιμοποιείται η ακόλουθη σχέση:

$$P_{jki}(l) = \frac{N(v_{il1}; \mu_{jk1}, \sigma^2_{jk1}) \cdot \ldots \cdot N(v_{ild}; \mu_{jkd}, \sigma^2_{jkd})}{\sum_{l} N(v_{il1}; \mu_{jk1}, \sigma^2_{jk1}) \cdot \ldots \cdot N(v_{ild}; \mu_{jkd}, \sigma^2_{jkd})}$$

Στον αριθμητή το γινόμενο αποτελείται από εκείνες τις κανονικές κατανομές που αντιστοιχούν στα στοιχεία του αρχικού διανύσματος που συνεισφέρουν στο συγκεκριμένο υποδιάνυσμα, ενώ το άθροισμα του παρανομαστή γίνεται σε όλα τα $l = 2^B$ (B=αριθμός bits για την κωδικοποίηση) δυνατά centroids για το συγκεκριμένο υποδιάνυσμα για λόγους κανονικοποίησης. Έτσι η διάσταση κάθε κατανομής είναι ίση με τον αριθμό των centroids (ή φύλλων – leaves - στην περίπτωση του δενδροειδούς διανυσματικού κβαντιστή (binary tree vector quantizer).

Στο ακόλουθο σχήμα φαίνεται εποπτικά ο τρόπος υλοποίησης της μορφής μιας κατανομής εξόδου ενός DMHMM:



Σχήμα 3-3 Εποπτική αναπαράσταση της υλοποίησης της κατανομής εξόδου στην περίπτωση DMHMM ακουστικού μοντέλου

3.1.3.5 Εκμάθηση παραμέτρων DMHMM μοντέλων

Μετά τη διακριτοποίηση και τη δημιουργία των ακουστικών μοντέλων μειγμάτων διακριτών κατανομών, μπορεί να παρουσιαστεί η ανάγκη για περαιτέρω εκπαίδευσή τους μέσω εφαρμογής επιπλέον iterations του forward-backward αλγορίθμου. Τέτοια ανάγκη προκύπτει αν επιλεγεί η διακριτοποίηση όχι στο τελικό στάδιο της εκπαίδευσης

των συνεχών μοντέλων αλλά πιο νωρίς (όπως στην περίπτωση της εναλλακτικής διαδικασίας εκπαίδευσης που παρουσιάζεται στο κεφάλαιο 4).

Η εκμάθηση των παραμέτρων από δεδομένα εκπαίδευσης ισοδυναμεί με την κατάλληλη επιλογή των παραμέτρων λ του μοντέλου ώστε να μεγιστοποιηθεί η πιθανότητα:

$$\max_{\lambda} P(\underline{X} \mid \lambda) .$$

Με τον τρόπο αυτό ικανοποιείται το κριτήριο της μέγιστης πιθανοφάνειας (Maximum Likelihood). Η μέθοδος που ακολουθείται για τη λύση του παραπάνω προβλήματος είναι ο επαναληπτικός αλγόριθμος Baum-Welch (forward-backward), ειδική περίπτωση της εφαρμογής του Expectation-Maximization (EM) αλγορίθμου στο πρόβλημα της αναγνώρισης ομιλίας. Για την εφαρμογή του EM αλγορίθμου μπορούμε να θεωρήσουμε την κατάσταση ως την κρυφή πληροφορία και τα σύμβολα εξόδου ως τα παρατηρούμενα σύμβολα. Ο Expectation-Maximization αλγόριθμος σε κάθε βήμα μεγιστοποιεί την ποσότητα:

$$E\left\{\log P(\overline{X}, \overline{Q} \mid \lambda_{new}) \mid \overline{X}, \lambda_{old}\right\}$$

και η αναμενόμενη αυτή τιμή υπολογίζεται πάνω σε όλες τις πιθανές ακολουθίες καταστάσεων Q. Αυτό επιβάλλεται από το γεγονός ότι η ακολουθία καταστάσεων δεν είναι γνωστή (είναι κρυφή) και έτσι δεν μπορεί να μεγιστοποιηθεί απευθείας η ποσότητα:

$$\log P(\overline{X}, \overline{Q} \mid \lambda) ,$$

που είναι η λογαριθμική συνάρτηση πιθανοφάνειας (log-likelihood function). Ο EM αλγόριθμος προσπαθεί να βρει τις παραμέτρους του μοντέλου που μεγιστοποιούν τη συνάρτηση πιθανοφάνειας πάνω στα δεδομένα εκπαίδευσης.

Οι τύποι επανυπολογισμού (re-estimation formulas) των παραμέτρων μπορούν να προκύψουν μεγιστοποιώντας τη βοηθητική συνάρτηση του Baum. Στην περίπτωση των DMHMM μοντέλων, πέρα από την κρυφή αλυσίδα των καταστάσεων υπάρχει και η κρυφή ακολουθία των δεικτών στις πολυδιάστατες κατανομές (components) που αποτελούν ένα μείγμα (mixture).

Για τα μοντέλα αυτά, η βοηθητική συνάρτηση του Baum γράφεται ως εξής:

$$Q(\lambda', \lambda) = \mathrm{E}\{\log P(\overline{X}, Q, \Omega \mid \lambda) \mid \overline{\mathrm{X}}, \lambda'\} =$$
$$= \sum_{Q,\Omega} P(Q, \Omega \mid \overline{X}, \lambda') \cdot \log P(\overline{X}, Q, \Omega \mid \lambda) =$$
$$= \frac{1}{P(\overline{X} \mid \lambda')} \sum_{Q,\Omega} P(\overline{X}, Q, \Omega \mid \lambda') \log P(\overline{X}, Q, \Omega \mid \lambda) =$$
$$= \sum_{Q,\Omega} P(Q, \Omega \mid \overline{X}, \lambda') \left\{ \log \pi_{q_0} + \sum_{t=1}^{T} \log \alpha_{q_{t-q} q_t} + \sum_{t=1}^{T} \log \prod_{i=1}^{L} P_{jki}(vq(x_{it})) \right\}$$

Στην παραπάνω σχέση το διάνυσμα $\underline{X}$ εκφράζει το σύνολο των ακουστικών παρατηρήσεων, το Q και το Ω είναι το σύνολο των κρυφών καταστάσεων και των δεικτών στα components αντίστοιχα, λ και λ’ είναι οι νέες και οι παλιές εκτιμήσεις για τις διάφορες παραμέτρους του μοντέλου, π_q είναι η αρχική πιθανότητα της κατάστασης q, $a_{qq'}$ είναι η πιθανότητα μετάβασης από την κατάσταση q στην κατάσταση q’, $j=q_t$ είναι η κρυφή κατάσταση τη χρονική στιγμή t και $k=\omega_t$ είναι ο δείκτης μείγματος τη χρονική στιγμή t.

Οι δύο πρώτοι όροι μέσα στην αγκύλη αντιστοιχούν στις αρχικές πιθανότητες και τις πιθανότητες μετάβασης. Για τη μεγιστοποίηση ως προς τις πιθανότητες εξόδου, αρκεί η μεγιστοποίηση του τρίτου όρου μέσα στην αγκύλη.

Έτσι έχουμε:

$$Q'(\lambda', \lambda) = \sum_{t=1}^{T} \sum_{\substack{q_t = j \\ \omega_\tau = k}} P(q_t = j \mid \overline{X}, \lambda') P(\omega_t = k \mid q_t = j, \overline{X}, \lambda') \cdot \log \prod_{i=1}^{L} P_{jki}(vq(x_{it})) =$$
$$= \sum_{t=1}^{T} \sum_{\substack{q_t = j \\ \omega_\tau = k}} P(q_t = j \mid \overline{X}, \lambda') P(\omega_t = k \mid q_t = j, \overline{X}, \lambda') \cdot \sum_{i=1}^{L} \log P_{jki}(vq(x_{it})).$$

Εφαρμόζοντας τον κανόνα του Bayes και τον ορισμό της μορφής της κατανομής εξόδου, έχουμε ότι:

$$P(\omega_t = k \mid q_t = j, \overline{X}, \lambda') = \frac{P(\omega_t = k, \overline{X} \mid q_t = j, \lambda')}{P(\overline{X} \mid q_t = j, \lambda')} =$$
$$= \frac{c_{jk} \prod_{i=1}^{L} P_{j,k,i}(vq(x_{it}))}{\sum_{k=1}^{M} c_{jk} \prod_{i=1}^{L} P_{j,k,i}(vq(x_{it}))}$$

Αν συμβολίσουμε με $\gamma_t(i,k)$ την πιθανότητα να εξετάζεται η κατάσταση i και το component k τη χρονική στιγμή t, δηλαδή:

$$\gamma_t(i,k) = P(q_t = i, \omega_t = k \mid \overline{X}, \lambda) = \left[\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \right] \left[\frac{c_{jk} \prod_{i=1}^{L} P_{j,k,i}(vq(x_{it}))}{\sum_{k=1}^{M} c_{jk} \prod_{i=1}^{L} P_{j,k,i}(vq(x_{it}))} \right]$$

όπου για τις βοηθητικές ποσότητες *α* και *β,* οι οποίες υπολογίζονται στο Forward και Backward στάδιο του αλγορίθμου, αντίστοιχα, ισχύουν οι σχέσεις:

$$\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) \cdot \alpha_{ij} \right] \cdot b_i(x_t)$$

$$\beta_t(i) = \sum_j \alpha_{ij} \cdot b_j(x_{t+1}) \cdot \beta_{t+1}(j),$$

καταλήγουμε στην παρακάτω έκφραση για τη βοηθητική συνάρτηση:

$$Q(\lambda', \lambda) = \sum_{t:x_{t,l}=v} \sum_{i,k} \gamma_t(i,k) \sum_{i=1}^{L} \log P_{jki}(vq(x_{it}))$$

Γενικά παρατηρούμε ότι οι βοηθητικές συναρτήσεις που προσπαθούμε να μεγιστοποιήσουμε έχουν όλες τη μορφή:

$$\sum_{j=1}^{N} w_j \cdot \log y_j$$

Η συνάρτηση αυτή, σαν συνάρτηση του $\{y_j\}_{j=1}^{N}$ δεδομένου του περιορισμού $\sum_{j=1}^{N} y_j = 1$, $y_j \geq 0$, παρουσιάζει ολικό μέγιστο στο σημείο:

$$y_j = \frac{w_j}{\sum_{i=1}^{N} w_i}$$

Άρα και δεδομένου του περιορισμού: $\sum_{vq} P_{jki}(vq(x_{it})) = 1$, η βοηθητική συνάρτηση Q(λ',λ) μεγιστοποιείται για την παρακάτω τιμή της πιθανότητας εξόδου:

$$P_{jki}(l) = \frac{\sum_{t:vq(x_{it})=l} \gamma_t(j,k)}{\sum_{t} \gamma_t(j,k)},$$

δηλαδή η ζητούμενη πιθανότητα για την κατάσταση j υπολογίζεται από την πιθανότητα να εμφανιστεί το *l*-centroid στο i-οστό υποδιάνυσμα και το k-οστό component.

Η ποσότητα $N_{jki}(l) \equiv \sum_{t:vq(x_{it})=l} \gamma_t(j,k)$ αποτελεί τον αριθμό των παρατηρήσεων (*counts*) του *l*-οστού συμβόλου (δείκτη centroid) για την κατανομή P_{jki}. Αντίστοιχα, η ποσότητα $N_{jki} \equiv \sum_{t} \gamma_t(j,k)$ αποτελεί το σύνολο των παρατηρήσεων (denom) των συμβόλων της κατανομής P_{jki} .

3.1.3.6 Τεχνική Ομαλοποίησης Delta Smoothing

Μετά την εφαρμογή των παραπάνω σχέσεων για την εκτίμηση των παραμέτρων κάθε διακριτής κατανομής μετά από ένα iteration εφαρμογής του αλγορίθμου των Baum-Welch, υπάρχει η δυνατότητα εφαρμογής κάποιων τεχνικών ομαλοποίησης των διακριτών (διανυσματικών) κατανομών για να αποφευχθεί η ανάθεση μηδενικής πιθανότητας σε centroids που δεν παρατηρήθηκαν στα δεδομένα εκπαίδευσης. Τέτοιες τεχνικές παρουσιάζονται αναλυτικά στην εργασία [7], ενώ στην παρούσα εργασία που τα δεδομένα εκπαίδευσης θεωρούνται αρκετά χρησιμοποιήθηκε η στοιχειώδης τεχνική

ομαλοποίησης delta smoothing. Σύμφωνα με την τεχνική αυτή για την εκτίμηση των τιμών μιας διανυσματικής κατανομής χρησιμοποιείται η ακόλουθη σχέση:

$$p_{jki}(l) = \frac{N(l) + \delta}{N + \delta \cdot 2^B},$$

όπου: $N = \sum_l N(l)$, $N(k) = \sum_{t:vq(x_{it})=l} \gamma_t(j,k)$ και B ο αριθμός των bits για την κωδικοποίηση του συγκεκριμένου υποδιανύσματος.

Όπως είναι εμφανές, προστίθεται σε όλα τα στοιχεία της κατανομής μια μικρή ποσότητα μάζας πιθανότητας δ, ενώ γίνεται και η απαραίτητη κανονικοποίηση.

4 Ομαδοποίηση διακριτών κατανομών

4.1 Επισκόπηση

Η διαδικασία της εκπαίδευσης των μοντέλων με την εφαρμογή του αλγορίθμου Baum-Welch σε κάθε στάδιο (iteration) της εκπαίδευσης, όπως είδαμε στο προηγούμενο κεφάλαιο έχει σαν αποτέλεσμα έναν αριθμό «παρατηρήσεων» (counts) για κάθε σύμβολο εξόδου του μοντέλου μας.

Στο μοντέλο DMHMM, τα σύμβολα εξόδου είναι διανύσματα (vectors) από centroids ή δείκτες σε centroids, των οποίων το πλήθος για κάθε υποδιάνυσμα εξαρτάται από τον αριθμό των bits που χρησιμοποιούνται για την κωδικοποίηση του αντίστοιχου υποδιανύσματος (subvector).

Για τον πλήρη ορισμό του μοντέλου, πρέπει να καθοριστούν οι αντίστοιχες κατανομές από τις παρατηρήσεις (counts) κάθε συμβόλου.

Όπως είδαμε σε προηγούμενο κεφάλαιο οι κατανομές των συμβόλων εξόδου στο μοντέλο DM-HMM είναι της μορφής:

$$b_j(x_t) = \sum_{k=1}^{M} c_{jk} \prod_{i=1}^{L} P_{jki}(vq(x_{it}))$$

Ειδικότερα, θα μας απασχολήσουν οι κατανομές της μορφής:

$$P_{jki}(vq(x_{it})) = p(centroid \mid index \mid state, mixture, subvector)$$

τις οποίες θα ονομάζουμε από εδώ και στο εξής διανυσματικές κατανομές (vector distributions).

Ένα μεγάλο πρόβλημα που σχετίζεται άμεσα με την εκπαίδευση των παραμέτρων των μοντέλων DM-HMM μέσω των εξισώσεων επανεκτίμησης Μέγιστης Πιθανοφάνειας (Maximum Likelihood re-estimation) είναι ο μεγάλος αριθμός ελεύθερων παραμέτρων του μοντέλου.

Μιας και τα δεδομένα εκπαίδευσης είναι αναγκαστικά πεπερασμένα, πολλά «γεγονότα» δεν παρατηρούνται σ' αυτά. Ένας από τους βασικούς στόχους των διαδικασιών εκπαίδευσης είναι να καλυφθούν και τα «γεγονότα» που δεν παρατηρήθηκαν στα δεδομένα εκπαίδευσης, αλλά που θα μπορούσαν να παρατηρηθούν με κάποια μη-μηδενική πιθανότητα σε πραγματικές συνθήκες.

Μια από τις προσεγγίσεις είναι και αναδιανομή μάζας πιθανότητας (probability mass) στα γεγονότα που δεν παρατηρήθηκαν κατά την εκπαίδευση. Προς αυτή την κατεύθυνση κινούνται οι διάφορες τεχνικές ομαλοποίησης (smoothing) των διακριτών διανυσματικών κατανομών και μια συγκριτική μελέτη τους έγινε στα πλαίσια της εργασίας [7].

Μια άλλη συνέπεια του προβλήματος του μεγάλου αριθμού των παραμέτρων είναι και οι απαιτήσεις σε μνήμη, τόσο κατά τη διάρκεια της εκπαίδευσης του μοντέλου όσο και κατά τη διαδικασία της αναγνώρισης.

Στην εργασία αυτή στόχος μας είναι η μείωση του αριθμού των ελεύθερων παραμέτρων του μοντέλου και κατά συνέπεια των απαιτήσεων του σε μνήμη αλλά και εμμέσως και των απαιτούμενων δεδομένων εκπαίδευσης, χωρίς να έχουμε σημαντική επίπτωση στην επίδοση του μοντέλου.

Η κατεύθυνση στην οποία κινηθήκαμε είναι η επέκταση της ήδη υπάρχουσας «συσχέτισης» (tying) των παραμέτρων του μοντέλου: Ήδη σε επίπεδο genones, υπάρχει «μοίρασμα» (sharing) κοινών genone για τις διάφορες κατανομές κατάστασης (state distributions) [1]. Αυτά που διαφοροποιούνται για κάθε κατανομή κατάστασης που μοιράζεται το ίδιο genone με πολλές άλλες, είναι τα βάρη για τα συστατικά (components) του genone.

Αρχικά μελετήθηκε η διαδικασία συσχέτισης «παρόμοιων» διανυσματικών κατανομών με στόχο τη μείωση του συνολικού αριθμού των, ενώ στη συνέχεια η μελέτη επεκτάθηκε και στη δυνατότητα συσχέτισης των centroid με στόχο τη μείωση της μέσης διάστασης των διανυσματικών κατανομών.

Ο τρόπος επίτευξης αυτής της «συσχέτισης» των παραμέτρων ανατίθεται σε μια διαδικασία ομαδοποίησης κατανομών (distribution clustering).

4.2 Μετρικές «απόστασης» μεταξύ δύο κατανομών

Βασικό ρόλο στη διαδικασία την ομαδοποίησης κατανομών, παίζει η επιλογή της μετρικής της απόστασης μεταξύ δυο κατανομών. Για την εύρεση της κατάλληλης μετρικής, καταφεύγουμε σε κάποιες βασικές έννοιες της θεωρίας πληροφοριών.

Καταρχήν, σαν εντροπία μιας τυχαίας μεταβλητής X με πυκνότητα πιθανότητας $p(x)$ με

$$p(x_i) = P(X = x_i) = p_i$$

ορίζεται η ποσότητα H(X), όπου:

$$H_p(X) \equiv -\sum_x p(x) \log p(x)$$

Η εντροπία είναι μέτρο του πληροφοριακού περιεχομένου ή της αβεβαιότητας της τυχαίας μεταβλητής X.

Βασικές ιδιότητες της εντροπίας είναι οι ακόλουθες:

- $H_p(X) \geq 0$
- $H_p(X) = 0 \quad iff \quad p(x_i) = 1 \ \gamma\iota\alpha \ \acute{\epsilon}\nu\alpha \ i$
- $H_p(X) \leq \log |N|$

Μέγιστη εντροπία (και άρα ελάχιστη πληροφορία) έχουμε όταν η κατανομή είναι ομοιόμορφη. Τότε η τιμή της εντροπίας δίνεται από τη σχέση:

$H_{\max} = \log |N|$,

όπου N είναι ο αριθμός των στοιχείων της κατανομής.

Σαν **σχετική εντροπία** (relative entropy) ή απόσταση Kullback-Leibler (KL divergence) μεταξύ δύο κατανομών πυκνότητας πιθανότητας (πιθανοτικών κατανομών) p(x) και p(y), ορίζεται η ποσότητα:

$$KL(p,q) \equiv \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Η απόσταση KL θα μπορούσε να θεωρηθεί σαν ένα μέτρο της «ομοιότητας» μεταξύ δύο κατανομών πυκνότητας πιθανότητας (probability density functions) μιας και στην περίπτωση που p(x) = q(x), η απόσταση KL είναι μηδενική:

$$KL(p \,||\, q) = 0 \Leftrightarrow p(x) = q(x), \forall x$$

Όμως η παραπάνω ποσότητα δεν μπορεί να οριστεί σαν πραγματική μετρική μιας και δεν είναι συμμετρική και δεν υπακούει στην τριγωνική ανισότητα (βλ. ορισμό μετρικής απόστασης [4, p.18]).

Για το λόγο αυτό αναζητούμε μια διαφορετική μετρική. Έτσι εισάγεται η **απόσταση Jensen-Shannon (JS)** μεταξύ δύο κατανομών πυκνότητας πιθανότητας p(x) και q(x), η οποία ορίζεται ως εξής:

$$\begin{aligned} JS(p,q) &= \pi_1 KL(p, \pi_1 p + \pi_2 q) + \pi_2 KL(q, \pi_1 p + \pi_2 q) \\ &= H_{\pi_1 p + \pi_2 q}(X) - \pi_1 H_p(X) - \pi_2 H_q(X) \end{aligned} ,$$

$$\text{όπου } \pi_1 + \pi_2 = 1, \pi_i \geq 0 .$$

Αντίθετα από ότι ισχύει με τη σχετική εντροπία, η απόσταση JS είναι προφανώς μια συμμετρική μετρική αλλά και φραγμένη [5] και καλύπτει και τις υπόλοιπες από τις παραπάνω ιδιότητες ώστε να πληροί τα κριτήρια μιας μετρικής απόστασης. Μια επιπλέον θετική ιδιότητα της απόστασης Jensen-Shannon είναι η δυνατότητα γενίκευσής για περισσότερες από δύο κατανομές πυκνότητας πιθανότητας για την εύρεση της «ανομοιότητας» μεταξύ περισσότερων κατανομών.

4.2.1 Επιλογή μετρικής και ομαδοποίησης

Ακολουθώντας τα βήματα του παραπάνω ορισμού της απόσταση Jensen-Shannon, το κριτήριο απόστασης μεταξύ δύο διακριτών κατανομών που επιλέχθηκε τελικά είναι η αύξηση της *counts-weighted* εντροπίας των δύο κατανομών που παρατηρείται με την ομαδοποίησή τους. Ουσιαστικά χρησιμοποιούνται τα counts των δύο κατανομών σαν εκτιμήτριες των ποσοτήτων π_1, π_2 της απόστασης Jensen-Johnson.

Η μείωση της πληροφορίας (distortion) και άρα αύξηση της εντροπίας, που παρατηρείται όταν δύο κατανομές p και q ομαδοποιούνται στην κατανομή s ορίζεται σαν:

$$d(p,q) = (n_1 + n_2)H_s(X) - n_1H_p(X) - n_2H_q(X), \qquad (4.1)$$

όπου n_1, n_2 είναι ο αριθμός των παρατηρήσεων (counts) που χρησιμοποιήθηκαν για την εκτίμηση των κατανομών πυκνότητας πιθανότητας p και q , αντίστοιχα.

Κριτήριο απόστασης, λοιπόν, θα είναι η παραπάνω ποσότητα και στόχος η μείωσή της και άρα η εύρεση εκείνης της ομαδοποίησης που ελαχιστοποιεί την πληροφορία που χάνεται κατά την ομαδοποίηση των κατανομών.

Πέρα όμως από το μέτρο της ομοιότητας δύο κατανομών (similarity measure), πρέπει να καθοριστεί και ο τρόπος εκτίμησης της νέας κατανομής που προκύπτει από την ομαδοποίηση των δύο κατανομών αφού αποφασιστεί ότι οι αρχικές κατανομές είναι αρκετά «όμοιες». Σύμφωνα με τα παραπάνω, η νέα κατανομή s εκτιμάται από τη σχέση:

$$s(x) = \frac{n_p}{n_p + n_q} p(x) + \frac{n_q}{n_p + n_q} q(x)$$

4.3 Τεχνικές ομαδοποίησης κατανομών

4.3.1 Agglomerative clustering

Για να επιτευχθεί η ζητούμενη ομαδοποίηση των κατανομών και να προκύψει ο επιθυμητός τελικός αριθμός ομάδων, καταφεύγουμε σε ένα *ιεραρχικό* σχήμα ομαδοποίησης «από κάτω προς τα πάνω» (bottom-up hierarchical scheme), το οποίο συναντάται στη βιβλιογραφία και με τον όρο agglomerative clustering [6].

Ο λόγος είναι ότι μια εξαντλητική διαδικασία αναζήτησης των ομαδοποιήσεων έχει ιδιαίτερα αυξημένο υπολογιστικό κόστος μιας και οι δυνατοί διαχωρισμοί αυξάνονται εκθετικά με τον αριθμό των αρχικών στοιχείων. Πράγματι, υπάρχουν περίπου $\frac{C^n}{c!}$ τρόποι για να διαχωριστούν *n* στοιχεία (εδώ κατανομές) σε *c* ομάδες.

Έτσι επιλέγεται μια διαδικασία επαναληπτικής βελτιστοποίησης (iterative optimization) όπως η παραπάνω. Στις διαδικασίες αυτές η βασική ιδέα είναι η εύρεση μιας λογικής αρχικής κατάταξης σε ομάδες και η διαδοχική εξέταση μετακινήσεων των διαφόρων στοιχείων από μια ομάδα σε μια άλλη ώστε να βελτιστοποιηθεί το κριτήριο που έχει τεθεί (criterion function). Το κρίσιμο θέμα στις προσεγγίσεις αυτές είναι η επιλογή του σημείου εκκίνησης μιας και οι διαδικασίες είναι υπο-βέλτιστες.

Η διαδικασία είναι επαναληπτική:

- *Αρχικά κάθε μια από τις x υπό εξέταση κατανομές θεωρείται ότι αποτελεί και μια ξεχωριστή ομάδα (cluster).*
- *Οι κατανομές που έχουν τη μικρότερη «απόσταση» μεταξύ τους σε κάθε στάδιο, συνενώνονται (merge) σε μια νέα, οι παράμετροι της οποίας επανεκτιμούνται, ενώ ξανά-υπολογίζονται και οι «αποστάσεις» από τις υπόλοιπες κατανομές.*
- *Η ομαδοποίηση δύο κατανομών γίνεται όταν η απόσταση όπως ορίστηκε στη σχέση (3.1) είναι μικρότερη ενός κατωφλιού, που είτε καθορίζεται είτε υπολογίζεται βάσει άλλων κριτηρίων (όπως για παράδειγμα ο απαιτούμενος αριθμός τελικών ομάδων).*
- *Η όλη διαδικασία συνεχίζεται μέχρι να ικανοποιηθεί κάποιο κριτήριο μέσης αύξησης της εντροπίας ή ενός τελικού αριθμού ομάδων.*

Ο χαρακτηρισμός του αλγορίθμου ως ιεραρχικού (hierarchical), έγκειται στο γεγονός ότι αν δύο κατανομές ομαδοποιήθηκαν σε μία σε κάποιο στάδιο του αλγόριθμου, θα παραμείνουν στην ίδια ομάδα σε όλα τα μετέπειτα στάδια.

Οι απαιτήσεις μιας τέτοιας προσέγγισης σε υπολογιστική ισχύ εξακολουθεί να είναι σχετικά υψηλές μιας και σε κάθε στάδιο, για κάθε κατανομή θα πρέπει να ελεγχθεί η δυνατότητα ομαδοποίησής με τις υπόλοιπες εναπομένουσες κατανομές. Βέβαια, οι δυνατοί συνδυασμοί μειώνονται σε κάθε στάδιο.

4.3.1.1 Υπολογιστική Πολυπλοκότητα – Απαιτήσεις σε μνήμη

Πιο αναλυτικά, αν αρχικά έχουμε n κατανομές και στόχος μας είναι να τις ομαδοποιήσουμε σε c κλάσεις, αρχικά θα πρέπει να υπολογιστούν n(n-1) «αποστάσεις» μεταξύ των κατανομών.

Λόγω της συμμετρίας της χρησιμοποιούμενης μετρικής, ο αριθμός των αρχικών συνδυασμών είναι:

$$\frac{n(n-1)}{2}$$

δηλαδή η πολυπλοκότητα για το στάδιο αυτό είναι $O(n^2)$.

Τα αποτελέσματα πρέπει να αποθηκευτούν για χρήση σε επόμενα στάδια του αλγορίθμου. Για την εύρεση του ζευγαριού κατανομών με τη μικρότερη «απόσταση» και πρέπει να ομαδοποιηθεί σε μία κατανομή θα πρέπει να διατρέξει κανείς ολόκληρο τον πίνακα και να κρατά κάθε φορά το ζευγάρι με τη μικρότερη απόσταση. Εναλλακτικά μπορεί να εφαρμοστεί κάποιος αλγόριθμος τύπου *quicksort* για την ταχύτερη εύρεση της ελάχιστης απόστασης. Για ένα οποιοδήποτε άλλο στάδιο, για παράδειγμα από τη μετάβαση από $\hat{c}$ κλάσεις σε $\hat{c}-1$, θα πρέπει να υπολογιστούν $\frac{\hat{c}(\hat{c}-1)}{2}$ «αποστάσεις».

4.3.2 Εναλλακτικοί αλγόριθμοι ομαδοποίησης

Εναλλακτική προς την παραπάνω περιγραφόμενη διαδικασία είναι η περιγραφόμενη ως divisive ή «από πάνω προς τα κάτω» (top-down) ομαδοποίηση. Η διαδικασία είναι κι αυτή επαναληπτική με τη διαφορά ότι αρχικά θεωρούμε ότι όλες οι κατανομές-δείγματα ανήκουν στην ίδια μεγάλη ομάδα, και οι επιμέρους ομάδες σχηματίζονται σταδιακά διαιρώντας την αρχική. Γενικά οι υπολογιστικές απαιτήσεις για την μετάβαση από το ένα στάδιο στο άλλο, είναι μεγαλύτερες στην περίπτωση της divisive ομαδοποίησης. Όμως αν ο αριθμός των επιθυμητών τελικών ομάδων είναι σχετικά μικρός, η προσέγγιση αυτή εξοικονομεί πολλά στάδια.

4.4 Περιγραφή προσεγγίσεων

Παρακάτω περιγράφονται αναλυτικά οι διάφοροι τρόποι που εξετάσθηκαν για την επιλογή των κατανομών για ομαδοποίηση.

4.4.1 Ομαδοποίηση σε επίπεδο διανυσματικών κατανομών (Vector Clustering)

Βασική ιδέα της κατηγορίας αυτής τεχνικών που περιγράφονται παρακάτω, είναι η «συσχέτιση» (tying) των κατανομών που αντιστοιχούν στο ίδιο υποδιάνυσμα (και άρα έχουν την ίδια διάσταση) με τις αντίστοιχες διαφορετικών μειγμάτων (genones). Με τον τρόπο αυτό μειώνεται ο συνολικός αριθμός διανυσματικών κατανομών που υπάρχει στο σύστημα, ενώ ο αριθμός των μειγμάτων και των βαρών μένει ο ίδιος.

Για κάθε υποδιάνυσμα, ο συνολικός αρχικός αριθμός τέτοιων κατανομών είναι ίσος με τον αριθμό των μειγμάτων επί τον αριθμό των «πολυδιάστατων» κατανομών ανά genone.

Η διάστασή τους διαφέρει για κάθε υποδιάνυσμα (και εξαρτάται από τον αριθμό των bits που χρησιμοποιήθηκαν για κάθε υποδιάνυσμα κατά την διανυσματική κβαντοποίηση).

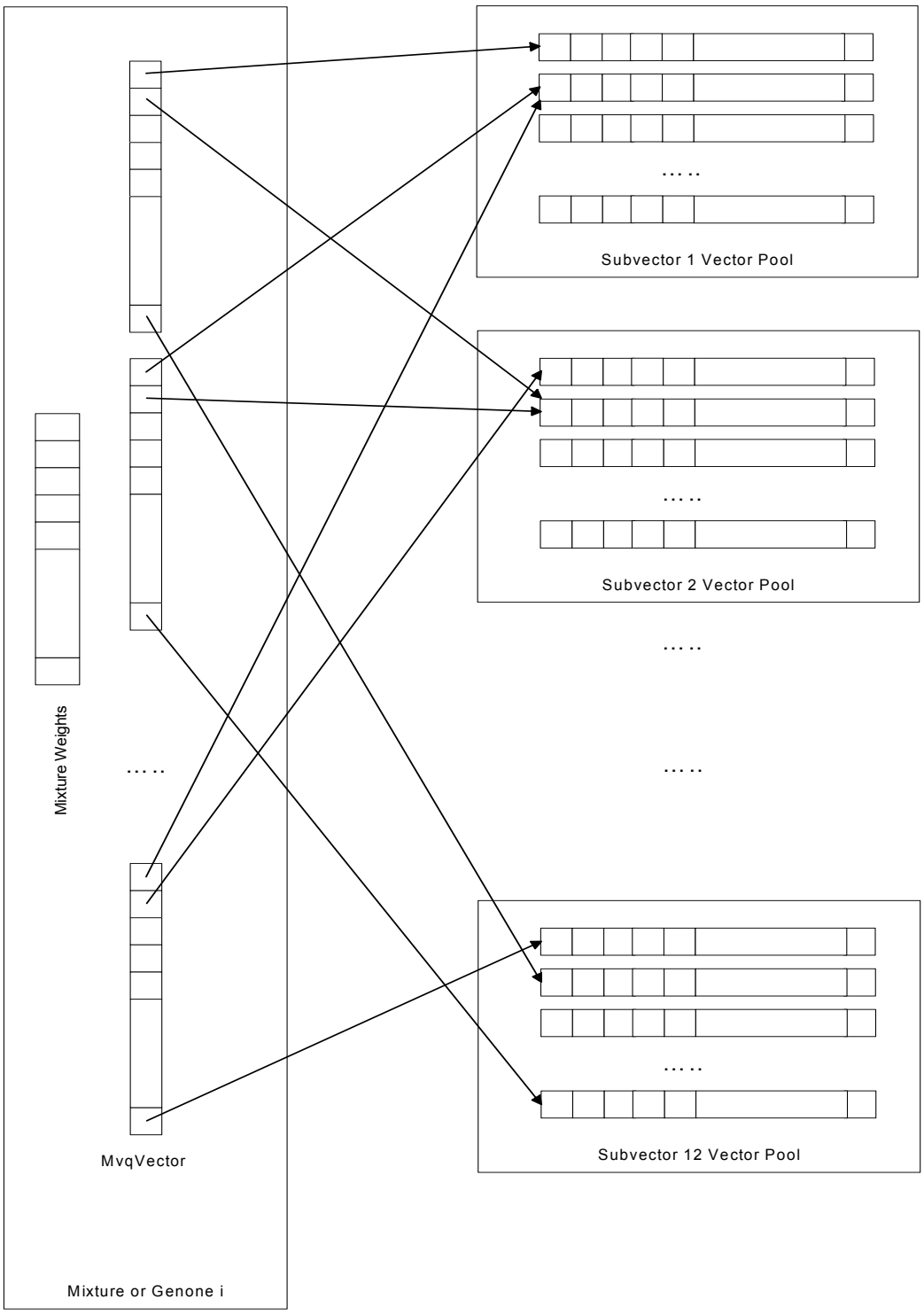
Μετά την ομαδοποίηση, οι τελικές κατανομές θα «μοιράζονται» για τον υπολογισμό των κατανομών εξόδου, αλλά με διαφορετικά βάρη για κάθε κατάσταση (state) του μοντέλου.

Εξετάστηκαν διάφορες εκδοχές του αλγόριθμου ως προς τον χωρισμό του συνόλου των κατανομών σε μικρότερες υποομάδες, ώστε να αποφευχθεί η εξαιρετικά μεγάλη πολυπλοκότητα του αλγόριθμου όταν απαιτείται ομαδοποίηση μεγάλου αριθμού κατανομών.

4.4.1.1 Single pool

Η πιο προφανής εκδοχή είναι να ληφθούν υπ’ όψιν όλες οι κατανομές για την ομαδοποίηση (που αντιστοιχούν πάντα στο ίδιο υποδιάνυσμα). Έτσι δημιουργείται ένας κοινός «κουβάς» (single pool) κατανομών για κάθε υποδιάνυσμα.

Το σχήμα αυτό ομαδοποίησης φαίνεται παραστατικά στο ακόλουθο διάγραμμα:



Σχήμα 4-1 Ομαδοποίηση διανυσματικών κατανομών με ένα pool ανά υποδιάνυσμα

4.4.1.2 Multiple pools

Δημιουργούνται πολλαπλοί «κουβάδες» κατανομών, εκμεταλλευόμενοι τη θέση της κατανομής στο μείγμα και ο διαχωρισμός των κατανομών σε πολλαπλά pools γίνεται ανάλογα με τη θέση αυτή. Ο αριθμός των pools είναι ίδιος με τον αριθμό των «πολυδιάστατων» διακριτών κατανομών σε κάθε genone, ενώ πλέον κάθε pool περιέχει μικρότερο αριθμό κατανομών για ομαδοποίηση (και ίσο με τον συνολικό αριθμό μειγμάτων στο μοντέλο). Τυπικές τιμές για τον αριθμό των pools είναι 8 ενώ για τον αριθμό των κατανομών που ομαδοποιούνται κάθε φορά: 1000.

4.4.1.3 Acoustically related pools

Μια εναλλακτική εκδοχή είναι ο αρχικός διαχωρισμός των κατανομών σε υπο-ομάδες (pools) βάσει της ακουστικής τους «συγγένειας». Κριτήριο είναι η ακουστική συγγένεια των φωνημάτων με τα οποία σχετίζονται τα αντίστοιχα genones στα οποία «ανήκουν» αρχικά οι διανυσματικές κατανομές.

Στην περίπτωση αυτή το αποτέλεσμα είναι μια ιεραρχική σχέση μεταξύ των τελικών ομάδων που προκύπτουν. Η εκδοχή αυτή έχει το πλεονέκτημα της ταχύτητας σε σχέση με τις υπόλοιπες μιας και ο αριθμός των κατανομών που πρέπει να ομαδοποιηθούν είναι ακόμα μικρότερος για κάθε υποομάδα.

4.4.2 Ομαδοποίηση σε επίπεδο φύλλου (Leaves ή centroid clustering)

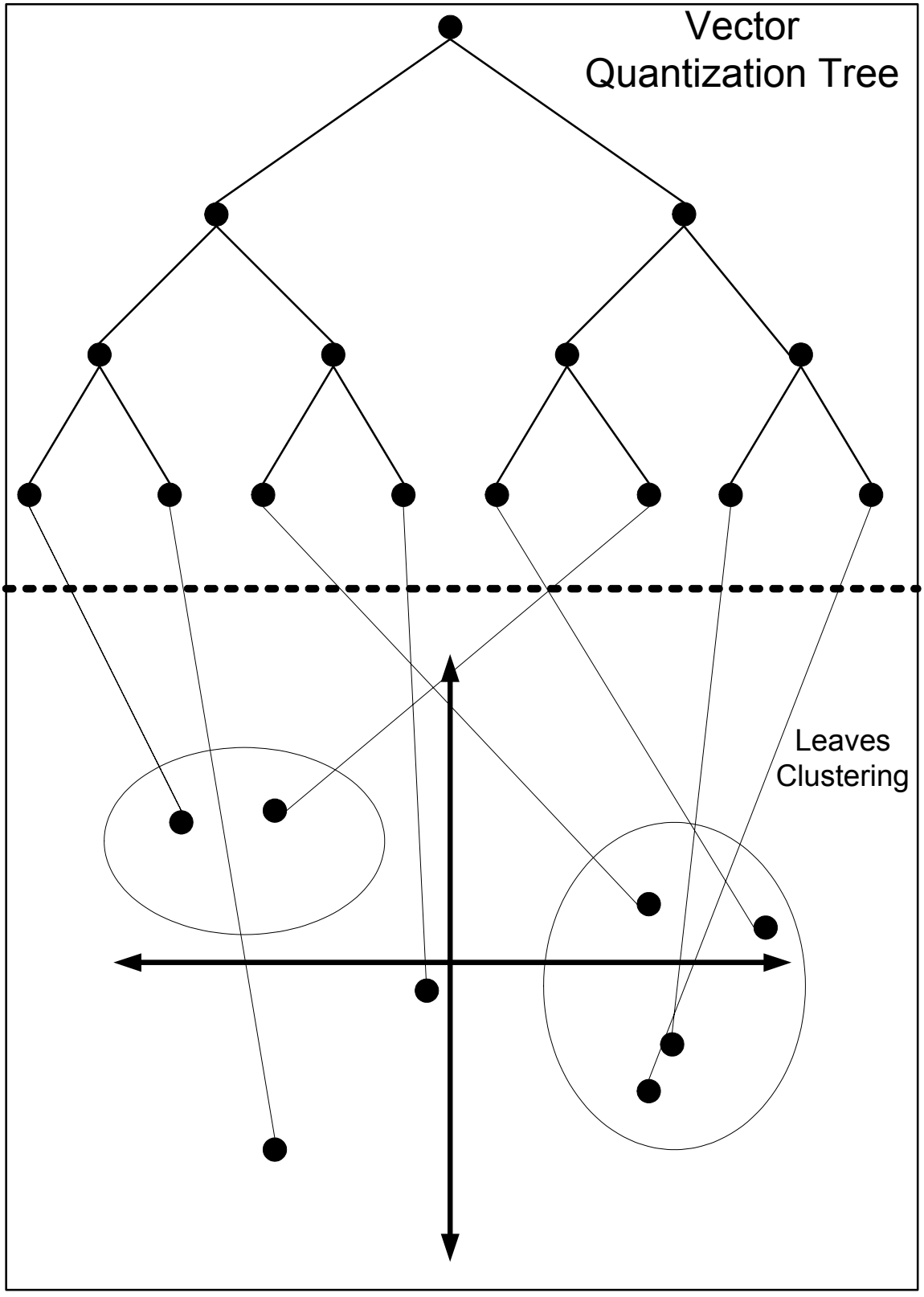
Στην προσέγγιση αυτή, βασική ιδέα είναι η ομαδοποίηση των «κοντινών» centroids (ή leaves στην περίπτωσή μας, όπου έχει εφαρμοστεί δενδροειδής διανυσματική κβαντοποίηση) με στόχο την μείωση της διάστασης των κατανομών για κάθε υποδιάνυσμα και άρα του συνολικού αριθμού ελεύθερων παραμέτρων στο μοντέλο μας.

Κατά την front-end επεξεργασία, για κάθε υποδιάνυσμα χρησιμοποιείται ένα δυαδικό δέντρο για την κβαντοποίηση κάποιων στοιχείων (elements) του αρχικού διανύσματος cepstrum χαρακτηριστικών. Ο αριθμός των φύλλων του κάθε δυαδικού δέντρου εξαρτάται από τον αριθμό των bits που χρησιμοποιούνται για την κωδικοποίηση κάθε υποδιανύσματος.

Ο αριθμός αυτός έχει προκύψει από μια διαδικασία που προσπαθεί να καλύψει μια απαιτούμενη ελάχιστη ανάλυση (resolution) του ακουστικού χώρου αλλά και ελαχιστοποίησης του συνολικού αριθμού των χρησιμοποιούμενων bits [3].

Στο ακόλουθο σχήμα εικονίζεται ένα παράδειγμα τέτοιου δυαδικού δέντρου διανυσματικής κβαντοποίησης (VQ Tree). Για την κωδικοποίηση του αντίστοιχου υποδιανύσματος χρησιμοποιούνται 3 bits οπότε προκύπτουν:

$$3bits \Rightarrow 2^3 = 8 \ centroids \ ή \ leaves$$



Σχήμα 4-2 Ομαδοποίηση φύλλων (leaves clustering)

Όπως φαίνεται στο παραπάνω σχήμα, αναζητείται κάποιος «μετασχηματισμός», με τη βοήθεια του οποίου μπορεί να εκτιμηθεί η «απόσταση» μεταξύ των centroid ώστε να προχωρήσουμε σε ομαδοποίησή τους.

Όπως περιγράφηκε παραπάνω, για το i-οστό υποδιάνυσμα έχουμε $N \cdot 1000$ διανυσματικές κατανομές της μορφής:

$$P_d(j) \equiv P(j \mid d),$$

όπου $1 \le j \le K$,($K = 2^B$ ο συνολικός αριθμός των centroids του εξεταζόμενου υποδιανύσματος) και $1 \le d \le N \cdot 1000$, d:ο δείκτης της διανυσματικής κατανομής και 1000 ο τυπικός αριθμός genones στο σύστημά μας.

Προσπαθώντας να αναπαραστήσουμε τη στατιστική συμπεριφορά ενός συγκεκριμένου centroid j του εξεταζόμενου υποδιανύσματος, προτείνουμε τον υπολογισμό της ακόλουθης κατανομής:

$$\begin{bmatrix} P(1 \mid j) \\ P(2 \mid j) \\ .... \\ P(N \cdot 1000 \mid j) \end{bmatrix} \qquad (4.2)$$

(όπου d_q είναι η q συνιστώσα κατανομή μείγματος και N είναι ο αριθμός των συστατικών «πολυδιάστατων» κατανομών κάθε μείγματος).

Στόχος μας είναι να βρεθούν και να ομαδοποιηθούν τα «κοντινά» centroids με τη χρήση των αλγορίθμων ομαδοποίησης που περιγράφηκαν παραπάνω και τη βοήθεια της παραπάνω κατανομής για κάθε centroid.

Κάθε στοιχείο της παραπάνω κατανομής ορίζεται να είναι η a-posteriori πιθανότητα της αντίστοιχης από τις $N \cdot 1000$ διαθέσιμες διανυσματικές κατανομές του συστήματός μας. Για την εκτίμηση της παραπάνω κατανομής με χρήση των κανόνων του Bayes και της ολικής πιθανότητας, έχουμε τα ακόλουθα βήματα:

$$P(q \mid j) = \frac{P(j \mid q) \cdot P(q)}{P(j)} = \frac{P(j \mid q) \cdot P(q)}{\sum_j P(j \mid q) \cdot P(q)}$$

Η ποσότητα $P(j \mid q)$ εκτιμάται άμεσα από τις «παρατηρήσεις» (counts) της διανυσματικής κατανομής βάσει των εξισώσεων επανεκτίμησης μέγιστης πιθανοφάνειας, ενώ για την εκτίμηση της ποσότητας $P(q)$,

υπάρχουν οι επιλογές:

$$P(q) = \sum_m P(q \mid g_m) \cdot P(g_m)$$

ή με τη βοήθεια των counts:

$$P(q) \equiv \frac{N_q}{\sum_q N_q} .$$

4.4.2.1 Κατανομές μικρότερης διάστασης

Για την αποφυγή των κατανομών μεγάλης διάστασης (παραπάνω η διάσταση ήταν N*1000), εναλλακτικά μπορεί να αναζητηθεί η εκτίμηση μιας κατανομή για κάθε centroid j της μορφής:

$$\begin{bmatrix} P(1 \mid j) \\ \dots \\ P(q \mid j) \\ \dots \\ P(1000 \mid j) \end{bmatrix},$$

όπου q είναι ο δείκτης σε ένα από τα μείγματα (genones) του συστήματος.

Όπως παρατηρούμε η διάσταση της παραπάνω κατανομής είναι ίση με τον αριθμό των genones στο μοντέλο μας (τυπική τιμή 1000).

Για την εκτίμηση της παραπάνω κατανομής, οι εξισώσεις παίρνουν τη μορφή:

$$P(q \mid j) = \frac{P(j \mid q) \cdot P(q)}{P(j)} = \frac{P(j \mid q) \cdot P(q)}{\sum_j P(j \mid q) \cdot P(q)}$$

Για την εκτίμηση της κατανομής $P(q)$, και πάλι χρησιμοποιείται πληροφορία από τα counts,

$$P(q) \equiv \frac{N_q}{\sum_q N_q}$$

Στην περίπτωση αυτή όμως, η ποσότητα $P(j \mid q)$,δεν μπορεί να εκτιμηθεί απευθείας από τα counts των διανυσματικών κατανομών, οπότε εκτιμάται βάσει της παρακάτω σχέσης:

$$P(i \mid q) = \sum_{d_1}^{d_N} P(j, d_x \mid q) = \sum_{d_1}^{d_N} P(j \mid d_x, q) \cdot P(d_x \mid q) = \sum_{d_1}^{d_N} P(j \mid d_x) \cdot P(d_x \mid q)$$ (όπ

ου Ν είναι ο αριθμός των συστατικών «πολυδιάστατων» κατανομών κάθε μείγματος).

Πλέον η ποσότητα $P(j \mid d_x)$ εκτιμάται απευθείας από τα counts των διανυσματικών κατανομών, ενώ για την εκτίμηση της ποσότητας $P(d_x \mid q)$ μπορούμε να καταφύγουμε σε μια μέση τιμή των αντίστοιχων mixture weights ή στη σχέση:

$$P(d_x \mid q) \equiv \frac{N_{d_x|q}}{\sum_x N_{d_x|q}}$$

Με τη βοήθεια των παραπάνω σχέσεων, εφαρμόζεται ο αλγόριθμος ομαδοποίησης σε ένα αριθμό βοηθητικών κατανομών που είναι ίσος με τον αριθμό των centroids κάθε υποδιανύσματος, ενώ η διάστασή τους είναι ίση με τον αριθμό μειγμάτων του μοντέλου.

4.5 Συνοψίζοντας – Αναμενόμενες επιπτώσεις στον αριθμό των παραμέτρων

Ο συνολικός αριθμός παραμέτρων του μοντέλου προκύπτει από το άθροισμα των a-priori πιθανοτήτων κατάστασης, των παραμέτρων που αφορούν στα βάρη για τα μείγματα (mixture-weights) και των παραμέτρων που αφορούν τις κατανομές (distribution parameters).

Στη μετάβαση από genonic HMMs σε DMHMMs, παρατηρείται μια πολύ μεγάλη αύξηση των ελεύθερων παραμέτρων μιας και αντί για κάθε πολυδιάστατη γκαουσιανή κατανομή για την οποία απαιτείται ένας αριθμός παραμέτρων 52 (27=διάσταση διανύσματος συντελεστών cepstrum * 2=παράμετροι μέσης τιμής και διασποράς) για την πλήρη περιγραφή της, χρειάζεται μια «πολυδιάστατη διακριτή κατανομή» με συνολικά 1648[1] παραμέτρους προς εκτίμηση.

Αυτό το μεγάλο αριθμό παραμέτρων έρχονται να «καταπολεμήσουν» οι τεχνικές ομαδοποίησης που περιγράψαμε. Η πρώτη από αυτές στοχεύει στη μείωση του συνολικού αριθμού διανυσματικών κατανομών, ενώ η δεύτερη στη μείωση της μέσης διάστασης των διανυσματικών κατανομών.

Συνοπτικά στον παρακάτω πίνακα παρουσιάζεται η επίπτωση των τεχνικών ομαδοποίησης σε διάφορα επίπεδα στον τελικό αριθμό των ελεύθερων παραμέτρων του μοντέλου. Οι αριθμοί των καταστάσεων, μειγμάτων και συστατικών κατανομών κάθε μείγματος είναι *τυπικοί* της τάξης μεγέθους και της διαδικασίας εκπαίδευσης που ακολουθήθηκε.

[1] Για 12 υποδιανύσματα και συγκεκριμένο αριθμό από bits για κάθε subvector. Βλέπε στο κεφάλαιο με τα πειραματικά αποτελέσματα το σχήμα που επιλέχθηκε κατά την διανυσματική κβαντοποίηση (κωδικοποίηση).

Τύπος Μοντέλου **Τύπος παραμέτρων**	**CI-PTM**	**Tri-phone CD-PTM**	**Genone HMMs**	**DMHMM**	**Vector Clustered DMHMMs**	**Leaf Clustered DMHMMs**
Phones	46	46	46	46	46	46
States (=Dists)	138	15200	15200	15200	15200	15200
Gaussians /Distribution (=weights/dist)	100	100	{32,8}	{32,8}	*Variable*	{32,8}
Mixtures (Genones)	46	46	1000	1000		1000
Multidimensional Gaussian	2*27	2*27	2*27	-	-	-
Subvectors	-	-	-	12	12	12
Sum of Subvector Dimensions	-	-	-	1648	1648	*Variable*

Πίνακας 4-1 Αριθμός παραμέτρων σε κάθε στάδιο εκπαίδευσης

Ως προς τις απαιτήσεις μνήμης, γενικά οι παράμετροι παίρνουν τιμές στο πεδίο των πραγματικών αριθμών, κάτι που κατά την υλοποίηση απαιτεί τη δέσμευση μνήμης αρκετής για την αποθήκευση ενός αριθμού κινητής υποδιαστολής (float) για κάθε μια από αυτές. Για λόγους εξοικονόμησης μνήμης και παρά το σφάλμα κβαντισμού που εισάγεται, γενικά επιλέγεται η κβάντιση και μάλιστα στο πεδίο τιμών 0-255 (character quantization), ώστε να απαιτείται η δέσμευση μόλις ενός byte για κάθε παράμετρο. Για λόγους ταχύτητας κατά τη διάρκεια της αναγνώρισης, η τιμή που αποθηκεύεται τελικά για κάθε πιθανότητα είναι προ-υπολογισμένη στο λογαριθμικό πεδίο.

5 Πειράματα Εκπαίδευσης και Αναγνώρισης

5.1 Επισκόπηση

Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα της εφαρμογής των τεχνικών που περιγράφηκαν στο προηγούμενο κεφάλαιο στην περίπτωση ενός ακουστικού μοντέλου για την Αγγλική γλώσσα.

Η όλη διαδικασία έγινε στο περιβάλλον μιας εξελιγμένης έκδοσης του περιβάλλοντος SRI Decipher™.

Ο αριθμός των προτάσεων που χρησιμοποιήθηκαν στο στάδιο της εκπαίδευσης ήταν της τάξης των 97.600 προτάσεων κάτι που αντιστοιχεί σε περίπου 73 ώρες ομιλίες.

Για την πιο αξιόπιστη μέτρηση της επίδοσης των υπό εξέταση τεχνικών ομαδοποίησης, παρουσιάζονται τα αποτελέσματα σε διάφορα σύνολα προτάσεων ελέγχου (test-sets). Τα σύνολα αυτά προτάσεων αφορούν σε διαφορετικές ανάγκες και διαφορετικής δυσκολίας εφαρμογών αναγνώρισης ομιλίας (tasks). Στη συγκεκριμένη περίπτωση τα σύνολα των προτάσεων ελέγχου ήταν από το πεδίο της αναγνώρισης αριθμών (διακριτών ψηφίων), επιβεβαίωσης (ναι, όχι και συναφών εκφράσεων) καθώς και ονομάτων (εφαρμογή αυτόματου τηλεφωνητή). Τέλος οι προτάσεις των δύο συνόλων (εκπαίδευσης και ελέγχου) ανήκουν σε *διαφορετικούς* ομιλητές.

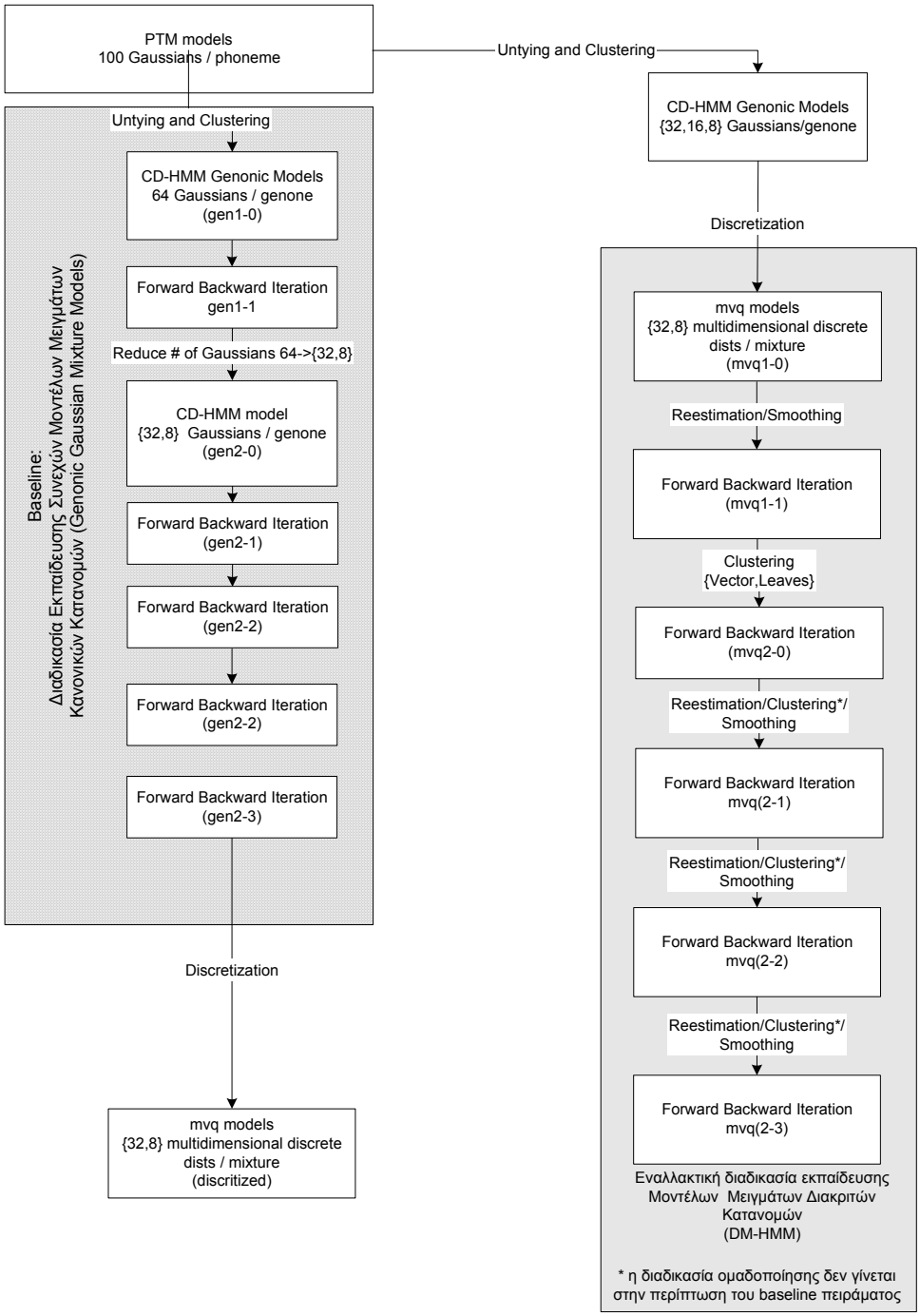
Ο αριθμός των προτάσεων σε κάθε σύνολο προτάσεων ελέγχου είναι τέτοιος ώστε οι όποιες διαφορές να μπορούν να θεωρηθούν στατιστικά σημαντικές. Ο συνολικός αριθμός προτάσεων ελέγχου ανέρχεται σε 7.636.

Έγινε εκπαίδευση των μοντέλων εξαρχής (δηλαδή ξεκινώντας από την εκπαίδευση του PTM συστήματος) και φτάνουμε σε μοντέλα DMHMM μετά από τη διακριτοποίηση genonic μοντέλων (σε διαφορετικά στάδια της όλης διαδικασίας όπως θα περιγραφεί παρακάτω). Η ακριβής διαδικασία εκπαίδευσης που ακολουθήθηκε παρουσιάζεται στην παράγραφο 2.

Η διαδικασία και τα πειράματα επαναλήφθηκαν και για διαφορετικό αριθμό κατανομών (components) σε κάθε genone ή mixture (32,8).

5.2 Σχηματική αναπαράσταση της διαδικασίας εκπαίδευσης

Οι διαδικασίες εκπαίδευσης που έχουν παρουσιαστεί φαίνονται σχηματικά στο ακόλουθο διάγραμμα, ενώ περιγράφονται αναλυτικά στις επόμενες παραγράφους:



Σχήμα 5-1 Σχηματική αναπαράσταση διαδικασίας εκπαίδευσης

5.3 Μετρικές Απόδοσης

Στη βιβλιογραφία μπορεί να βρει κανείς τις ακόλουθες μετρικές απόδοσης:

Correctness:

$$Co = (\frac{TOTAL\_WORDS - DEL - SUBST}{TOTAL\_WORDS}) \times 100\%$$

Ακρίβεια (Accuracy):

$$Ac = (\frac{TOTAL\_WORDS - DEL - SUBST - INS}{TOTAL\_WORDS}) \times 100\%$$

Word Error Rate:

$$WER = (\frac{DEL + SUBST + INS}{TOTAL\_WORDS}) \times 100\%$$

NL (Natural Language) Error Rate:

$$NL\ ER = (\frac{SLOTS_{DEL} + SLOTS_{SUBST} + SLOTS_{INS}}{TOTAL\_SLOTS}) \times 100\%$$

Στα πειράματα μας ως βασική μετρική απόδοσης της αναγνώρισης χρησιμοποιείται η *WER* (word error rate), που είναι το ποσοστό επί τοις % των λέξεων που αναγνωρίστηκαν «λάθος» (δηλαδή που η εφαρμογή του αναγνωριστή (recognizer) έχει προσθέσει, αφαιρέσει ή αντικαταστήσει σε σχέση με τις *λέξεις* που είχαν ειπωθεί στην εξεταζόμενη πρόταση), ενώ σε πολλές περιπτώσεις παρουσιάζεται και το NL Error rate, που είναι η αντίστοιχη μετρική για τις διαφορές σε επίπεδο natural language interpretation (slots).

Τα αποτελέσματα των διαφόρων πειραμάτων είναι της μορφής:

39 ins, 246 del, 580 sub = 30.06% of 2878 words (31.21% of 1775 files).

Τα πεδία στο παραπάνω ενδεικτικό αποτέλεσμα αναπαριστούν:

39 ins	246 del	580 sub	**30.06%**	2878 words	31.21%	1775 files
Αριθμός λέξεων που προστέθηκαν λάθος	Αριθμός λέξεων που αφαιρέθηκαν λάθος	Αριθμός λέξεων που αντικαταστάθηκαν λάθος	**WER**	Σύνολο λέξεων	Ποσοστιαίο λάθος σε επίπεδο πρότασης (Sentence Error Rate)	Αριθμός προτάσεων ελέγχου

Για λόγους συντομίας, στους πίνακες και τα γραφήματα παρουσιάζονται μόνο τα ποσοστά ***WER*** και ***NL Error Rate*** για κάθε ένα από τα πειράματα αναγνώρισης.

5.4 CM-HMM Baseline πειράματα

Αρχικά ακολουθήθηκε η προτεινόμενη από την βιβλιογραφία διαδικασία εκπαίδευσης συνεχών μοντέλων μέσω της σταδιακής μετάβασης από context-independent μοντέλα σε context-dependent (tri-phone) και τελικά σε genonic μοντέλα.

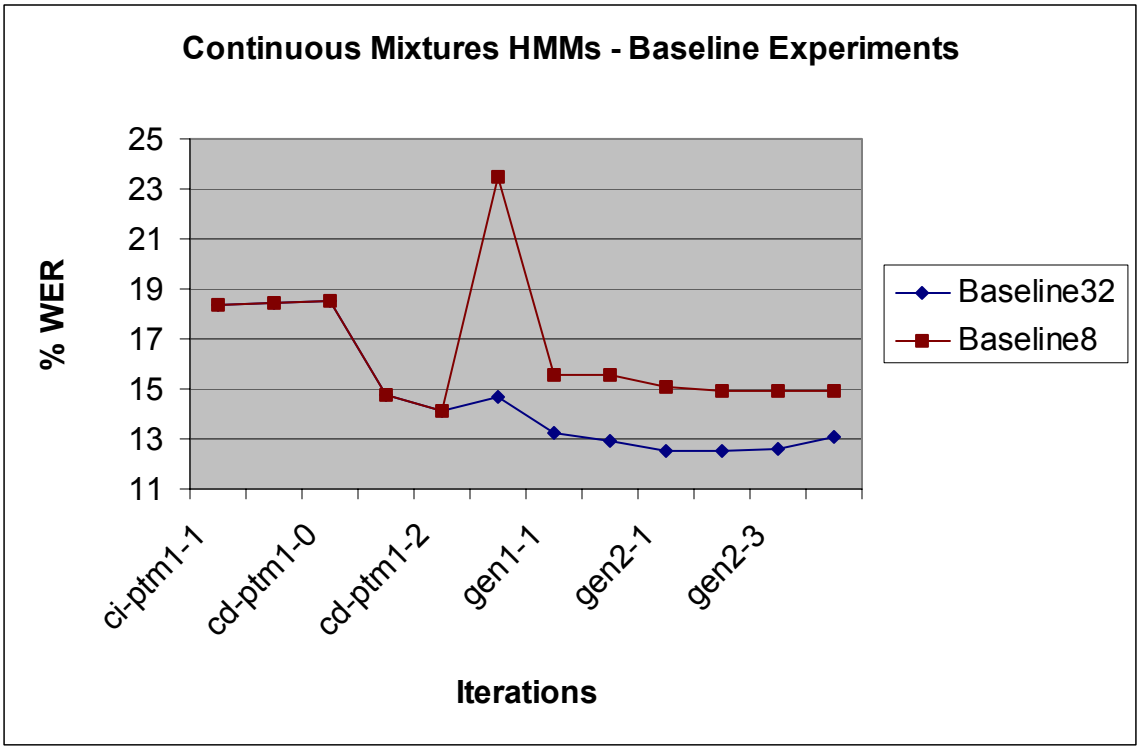
Η διαδικασία παρουσιάζεται εποπτικά στο σχήμα 5.1 ενώ πιο αναλυτικά τα στάδια είναι τα ακόλουθα:

- 2 Forward/Backward iterations CI-PTM (Context-Independent PTM) (100 Gaussians/phone)
- 2 Forward/Backward iterations CD-PTM (Context-Dependent PTM)
- Clustering, Untying, Gaussian Reduction -> Genones (64 Gaussians/genone)
- 1 Forward/Backward iteration GEN (64 Gaussians/genone)
- Gaussian Reduction 64 → {32,8}
- 2 Forward/Backward iterations GEN ({32,8} Gaussians/genone)
- 1 Forward/Backward iteration for shortlist generation

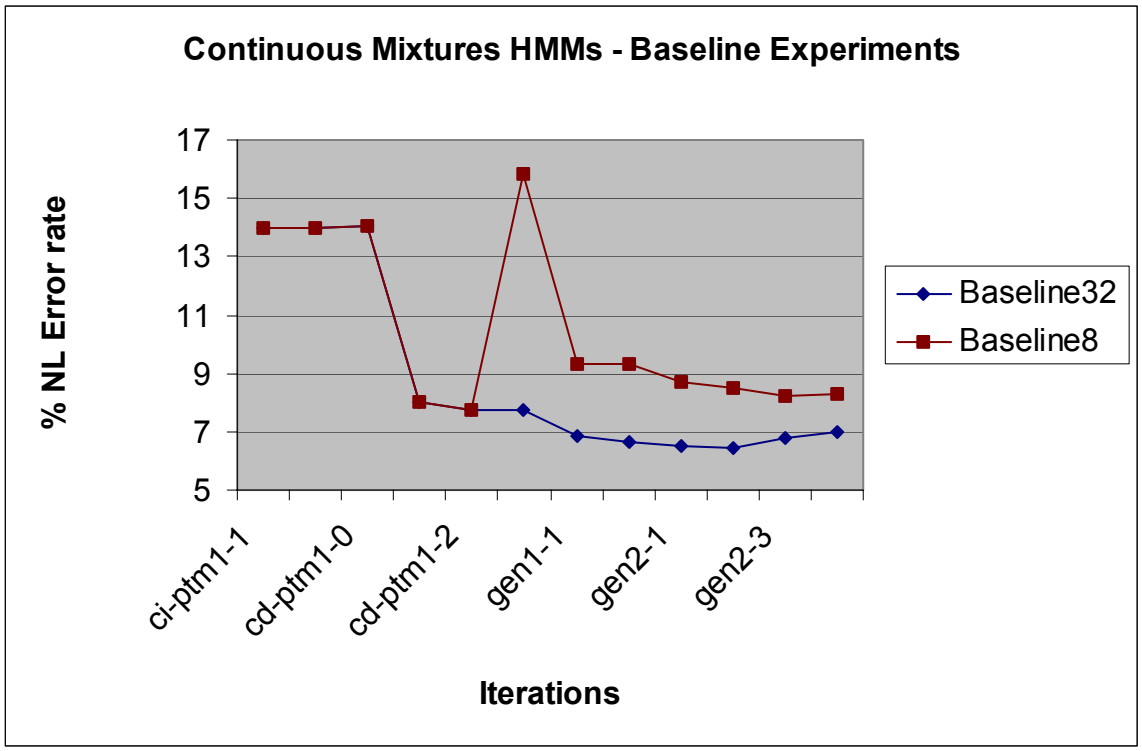
Τα τελικά μοντέλα αποτελούνται από 1000 genones, ενώ δημιουργήθηκαν μοντέλα τόσο με 32 όσο και 8 πολυδιάστατες κανονικές κατανομές (Gaussians) για κάθε genone.

Η τελική (gen2-3-final) εκδοχή του κάθε genone μοντέλου χρησιμοποιεί την τεχνική των Gaussian Shortlists [1] για να επιτευχθεί συντόμευση του χρόνου αναγνώρισης (xcpuRT) με κάποιο μικρό κόστος στην ποιότητα αναγνώρισης.

Στην ονοματολογία των μοντέλων στα διάφορα στάδια της εκπαίδευσης, εκτός από την κατηγορία του μοντέλου σημειώνεται το στάδιο και ο αριθμός του iteration.



Σχήμα 5-2 Επίδοση συνεχών μοντέλων στα διάφορα στάδια της εκπαίδευσης



Σχήμα 5-3

Για την περίπτωση του baseline8 πειράματος (των μοντέλων με 8 πολυδιάστατες κανονικές κατανομές ανά genone), το σφάλμα κατά τη μετάβαση από τα 100 components στο στάδιο των context-dependent PTM μοντέλων είναι σημαντικά αυξημένο λόγω της απευθείας μετάβασης σε 8 components στο στάδιο των genone μοντέλων.

Αντίθετα, στην περίπτωση του baseline32 πειράματος, η μετάβαση σε μειωμένο αριθμό components (Gaussian reduction) είναι σταδιακή (αρχικά υπάρχει η μετάβαση σε 64 και στο επόμενο στάδιο σε 32). Γενικά παρατηρούμε ότι το genonic μοντέλο με 32 components και συνολικά $32 \cdot 1000$ γκαουσιανές κατανομές υπερτερεί του PTM μοντέλου με $46 \cdot 100$ γκαουσιανές, το οποίο με τη σειρά έχει καλύτερη επίδοση από το genone μοντέλο με 8 components και συνολικά $8 \cdot 1000$ γκαουσιανές.

5.4.1 Διακριτοποίηση των συνεχών μοντέλων μετά το τελικό στάδιο (gen2-3-final)

Σύμφωνα με την προσέγγιση αυτή, η εκπαίδευση του ακουστικού μοντέλου γίνεται με τον συνηθισμένο τρόπο, παράγεται το μοντέλο μειγμάτων συνεχών (κανονικών) κατανομών και στο τελικό στάδιο προστίθεται μια διαδικασία διακριτοποίησής του, ώστε να προκύψει το ζητούμενο ακουστικό μοντέλο μειγμάτων («πολυδιάστατων») διακριτών κατανομών (DMHMM).

Για να γίνει κάτι τέτοιο, αρχικά – σύμφωνα με τις διαδικασίες που περιγράφηκαν στο προηγούμενο κεφάλαιο – γίνεται η εκπαίδευση των codebooks για την διανυσματική κβαντοποίηση (vector quantization) του - παραγόμενου από το front-end στάδιο επεξεργασίας - διανύσματος των cepstrum συντελεστών. Στόχος είναι η επιλογή του σχήματος που επιτυγχάνει πολύ καλή επίδοση αναγνώρισης, ενώ ταυτόχρονα δεν έχει πολλές απαιτήσεις σε χώρο αποθήκευσης.

Αποφασίστηκε η χρήση των 12 υποδιανυσμάτων και τα στοιχεία του αρχικού cepstrum διανύσματος που αντιστοιχούν στο κάθε υποδιάνυσμα φαίνεται στον ακόλουθο πίνακα:

Subvector	**Number of feature vector elements**	**Feature vector elements**
1	2	9,14
2	3	2,8,24
3	2	17,25
4	2	19,27
5	2	1,3
6	2	6,22
7	2	10,11
8	2	13,15
9	2	21,26
10	3	7,12,23
11	3	2,5,18
12	3	4,16,20

Πίνακας 5-1 Ανάθεση στοιχείων του αρχικού cepstrum διανύσματος στα διάφορα υποδιανύσματα

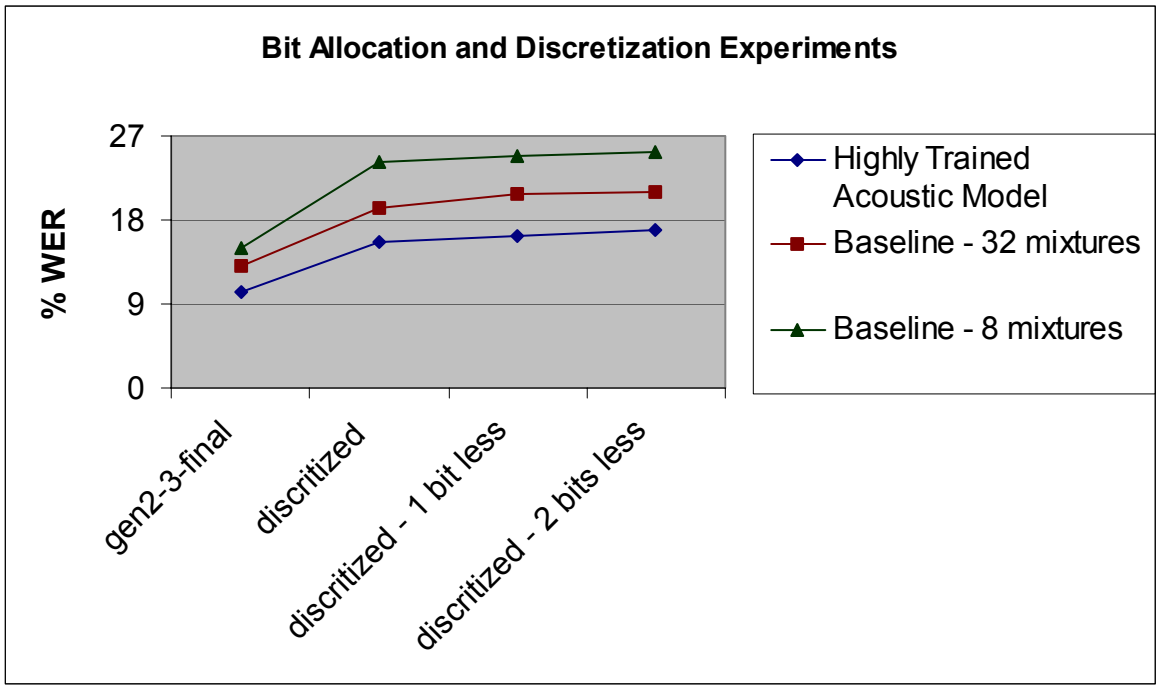
Σε επόμενο στάδιο θα πρέπει να αποφασιστεί ο αριθμός των bits (και άρα ο αριθμός των centroids) που θα χρησιμοποιηθεί για κάθε υποδιάνυσμα. Εφαρμόζοντας τον αλγόριθμο bit allocation [3] στο τελικό στάδιο των συνεχών μοντέλων προέκυψε το ακόλουθο configuration:

Subvector	Bits	Centroids	1bit less	Centroids
1	7	128	6	64
2	6	64	5	32
3	6	64	5	32
4	7	128	6	64
5	6	64	5	32
6	4	16	3	8
7	8	256	7	128
8	5	32	4	16
9	7	128	6	64
10	8	256	7	128
11	8	256	7	128
12	8	256	7	128
Total		**1648**		**824**

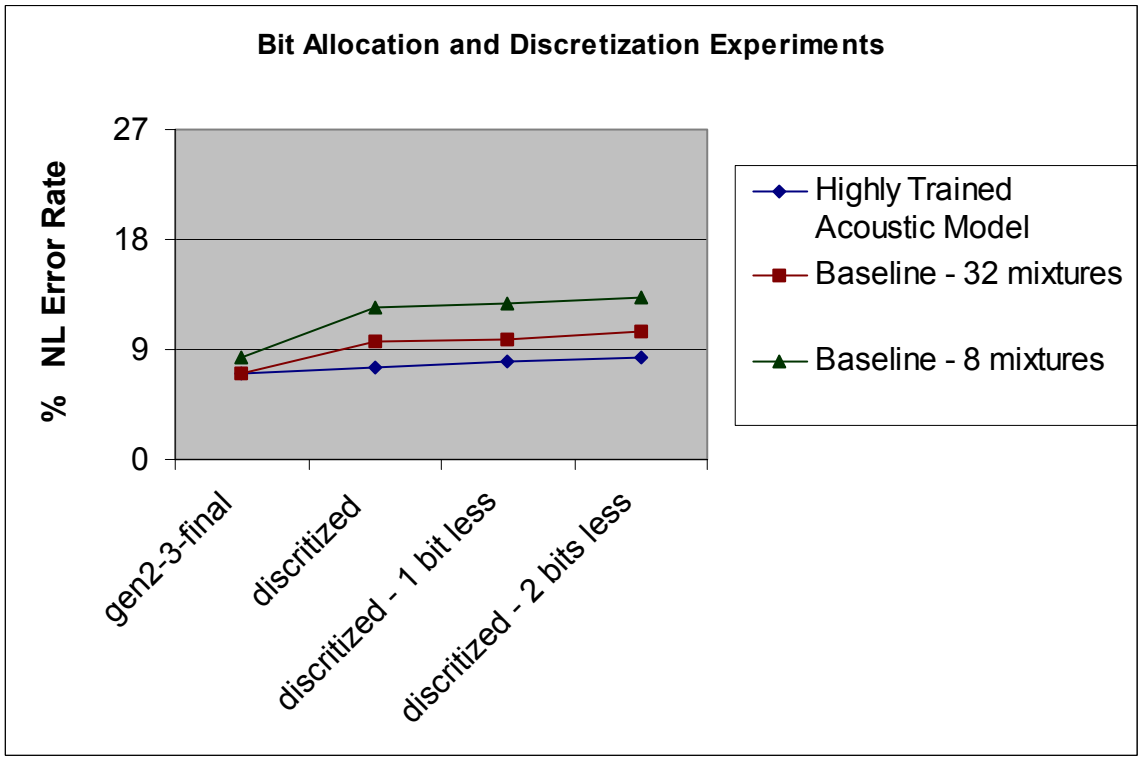
Πίνακας 5-2 Αριθμός bits για την κωδικοποίηση κάθε υποδιανύσματος

Εφαρμόστηκε η διαδικασία της διακριτοποίησης τόσο στα συνεχή μοντέλα 32 όσο και 8 components ανά genone, ενώ η ίδια διαδικασία επαναλήφθηκε και στην περίπτωση ενός αντίστοιχου ακουστικού μοντέλου που έχει προκύψει με αντίστοιχη διαδικασία και έχει τα χαρακτηριστικά του τελικού μοντέλου του baseline32 πειράματος αλλά για την εκπαίδευσή του έχει χρησιμοποιηθεί πολύ μεγαλύτερος αριθμός προτάσεων.

Η διαδικασία έγινε τόσο για τον αριθμό των bits για κάθε υποδιάνυσμα που φαίνεται στον παραπάνω πίνακα αλλά και για τις περιπτώσεις που έχουμε 1 και 2 bits λιγότερα για την κωδικοποίηση κάθε υποδιανύσματος (1 bit less και 2 bits less,αντίστοιχα). Αυτό σημαίνει ότι ο αριθμός των centroids πέφτει στο μισό και στο ένα τέταρτο των αρχικών, αντίστοιχα.



Σχήμα 5-4



Σχήμα 5-5

Παρατηρούμε ότι εισάγεται σφάλμα διακριτοποίησης το οποίο, όπως αναμένεται, αυξάνεται όσο μεταβαίνουμε σε μικρότερο αριθμό bits για την κωδικοποίηση κάθε υποδιανύσματος.

5.5 Εναλλακτικός τρόπος εκπαίδευσης μοντέλων μειγμάτων διακριτών κατανομών

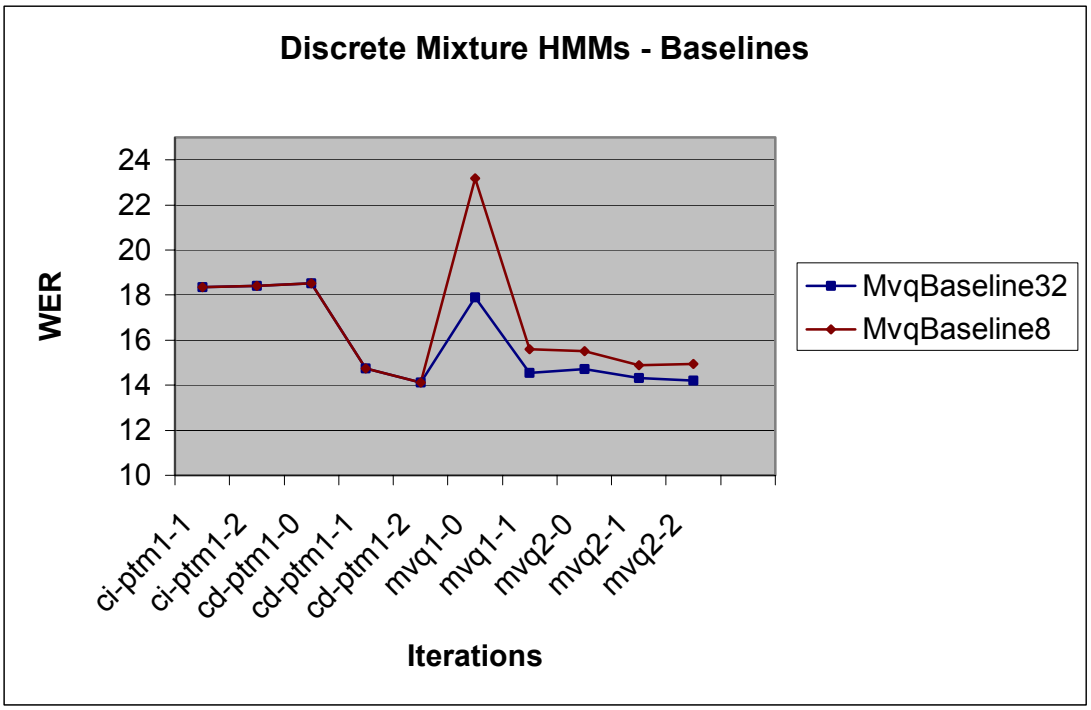
Για την αποφυγή της παραπάνω αδυναμίας «ανάκαμψης» από το σφάλμα που εισάγεται κατά τη διακριτοποίηση στο τελευταίο στάδιο εκπαίδευσης, ακολουθήθηκε μια εναλλακτική διαδικασία, σύμφωνα με την οποία η διακριτοποίηση γίνεται σε σχετικά πρώιμο στάδιο (αμέσως μετά την μετατροπή των context-dependent ptm μοντέλων σε genonic).

Η εναλλακτική αυτή διαδικασία φαίνεται εποπτικά στο σχήμα 5.1 ενώ τα αναλυτικά βήματα της διαδικασίας είναι τα παρακάτω:

- 2 Forward/Backward iterations CI-PTM (Context-Independent PTM) (100 Gaussians/phone)
- 2 Forward/Backward iterations CD-PTM (Context-Dependent PTM)
- Clustering, Untying, Gaussian Reduction -> Genones ({32,8} Gaussians/genone)
- Discretization
- 3 Forward/Backward Iteration MVQ ({32,8} mvqvectors/mixture)

Όπως εξηγήθηκε στο προηγούμενο κεφάλαιο, οι πολυδιάστατες κανονικές κατανομές (Gaussians) έχουν αντικατασταθεί από «πολυδιάστατες» διακριτές κατανομές (mvqvectors). Μετά από κάθε iteration του Forward/Backward αλγορίθμου και πριν την εφαρμογή των εξισώσεων επανεκτίμησης (re-estimation), εφαρμόζεται η στοιχειώδης τεχνική ομαλοποίησης των διακριτών κατανομών delta smoothing, που περιγράφηκε στην εργασία [7]. Εφαρμόστηκε το ίδιο bit-allocation σχήμα με το προηγούμενο πείραμα παρόλο που στη γενική περίπτωση αυτό δεν είναι απαραίτητο να ισχύει μιας και η διακριτοποίηση γίνεται σε διαφορετικό στάδιο.

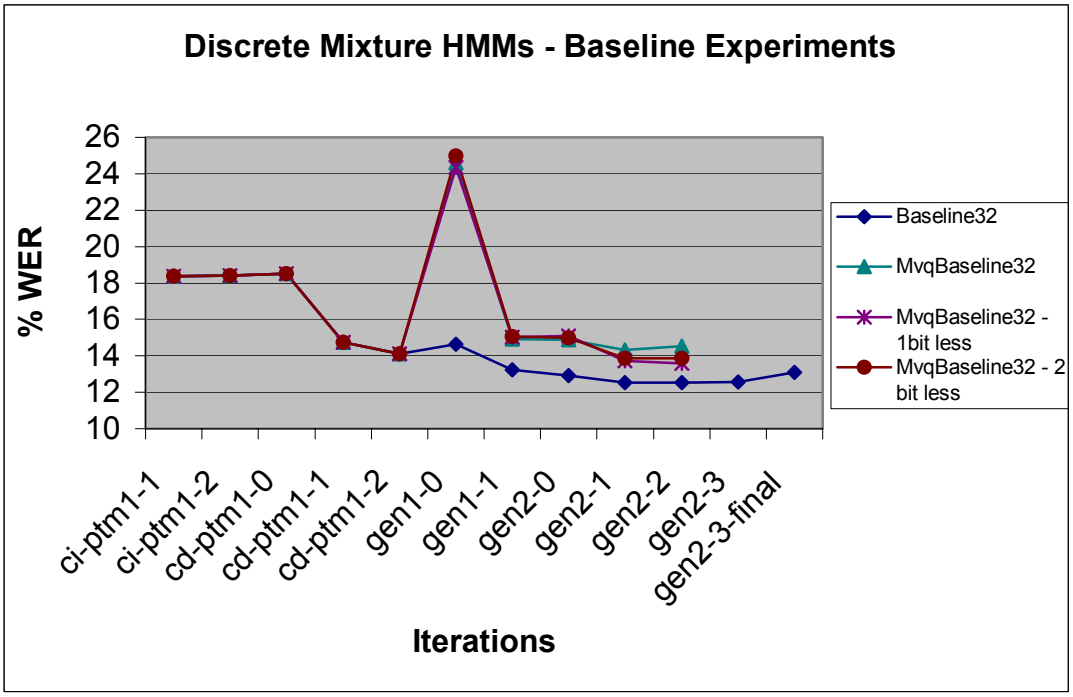
5.5.1 Discrete Mixture HMM Baseline πειράματα



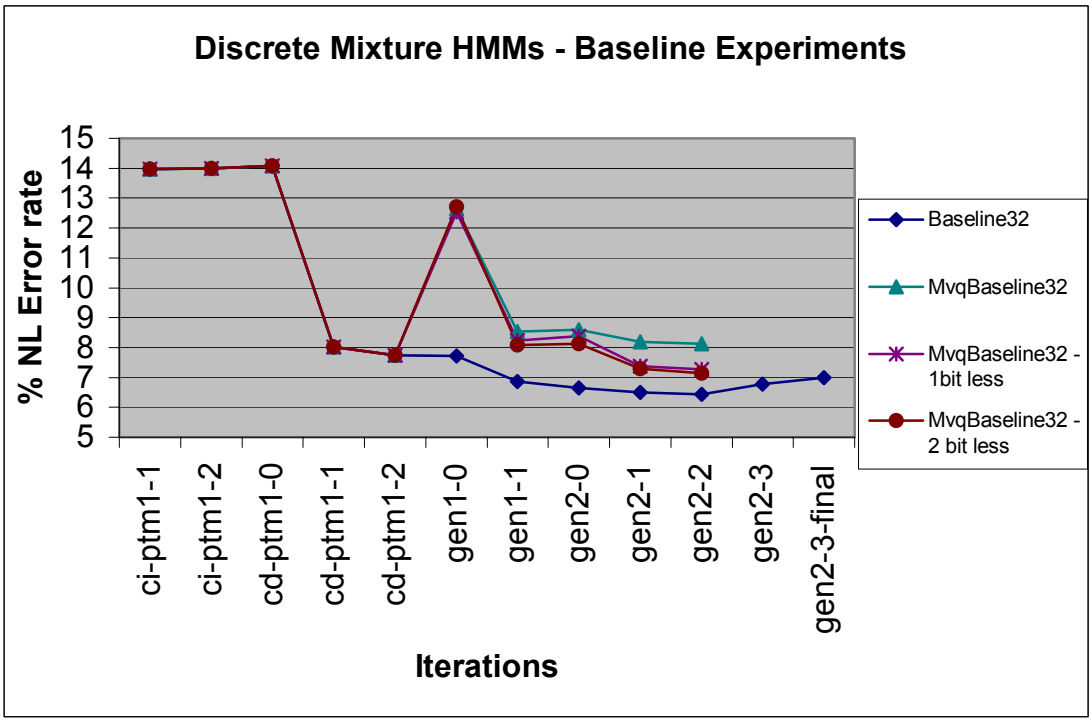
Σχήμα 5-6

Παρατηρούμε ότι κατά τη διακριτοποίηση υπάρχει μια αντίστοιχη αύξηση του σφάλματος με τη διαδικασία μετάβασης από cd-ptm σε genone στην περίπτωση των συνεχών μοντέλων. Η αύξηση αυτή, όμως, «απορροφάται» με τα επιπλέον iterations και το σφάλμα φτάνει στα επίπεδα του αντίστοιχου συνεχούς μοντέλου (δηλαδή με τον ίδιο αριθμό πολυδιάστατων κατανομών ανά μείγμα).

Επικεντρώνοντας στα μοντέλα με 32 components, έγιναν πειράματα με ανάθεση διαφορετικού αριθμού bits ανά υποδιάνυσμα:



Σχήμα 5-7



Σχήμα 5-8

Τα αποτελέσματα επιβεβαιώνουν ότι το bit-allocation σχήμα που υιοθετήθηκε δεν ήταν το βέλτιστο μιας και βασίστηκε σε εφαρμογή της trial-and-error τεχνικής που περιγράφεται στην εργασία [3] σε διαφορετικά μοντέλα. Εφαρμόζοντας διαφορετικό bit-allocation σχήμα με λιγότερα centroids ανά υποδιάνυσμα επιτυγχάνεται καλύτερη επίδοση αλλά και εξοικονόμηση μνήμης.

5.6 Πειράματα Ομαδοποίησης

Οι διάφορες τεχνικές ομαδοποίησης εφαρμόζονται σε κάποιο στάδιο μετά τη διακριτοποίηση και τη δημιουργία των μοντέλων μειγμάτων διακριτών κατανομών. Η διαδικασία της εκπαίδευσης φαίνεται παραστατικά στο σχήμα 5.1 ενώ τα αναλυτικά βήματα της διαδικασίας είναι τα παρακάτω:

- 2 Forward/Backward iterations CI-PTM (Context-Independent PTM) (100 Gaussians/phone)
- 2 Forward/Backward iterations CD-PTM (Context-Dependent PTM)
- Clustering, Untying, Gaussian Reduction -> Genones ({32,8} Gaussians/genone)
- Discretization
- 1 Forward/Backward Iteration MVQ ({32,8} mixtures/genone)
- {Vector,Leaf} Clustering
- 2 Forward/Backward Iterations Clustered MVQ ({32,8} mixtures/genone)

Τα ακόλουθα γραφήματα αναφέρονται στην περίπτωση των *32* components ανά μείγμα.

5.6.1 Ομαδοποίηση διανυσματικών κατανομών (vector clustering)

5.6.1.1 Θέματα υλοποίησης

Εφαρμόστηκε η τεχνική ομαδοποίησης διανυσματικών κατανομών μετά από ένα iteration και εφαρμογή των εξισώσεων επανεκτίμησης των μοντέλων μειγμάτων διακριτών κατανομών, όπως φαίνεται και στο σχήμα 5.1.
Πρέπει να σημειωθεί ότι με τον τρόπο που υλοποιήθηκε η τεχνική αυτή στόχος ήταν η εκτίμηση της επίδοσής της μιας και στην πράξη λόγω της τρέχουσας σειριακής δομής αποθήκευσης των μοντέλων δεν ήταν δυνατό να εισαχθεί το απαιτούμενο redirection κατά την αναγνώριση.

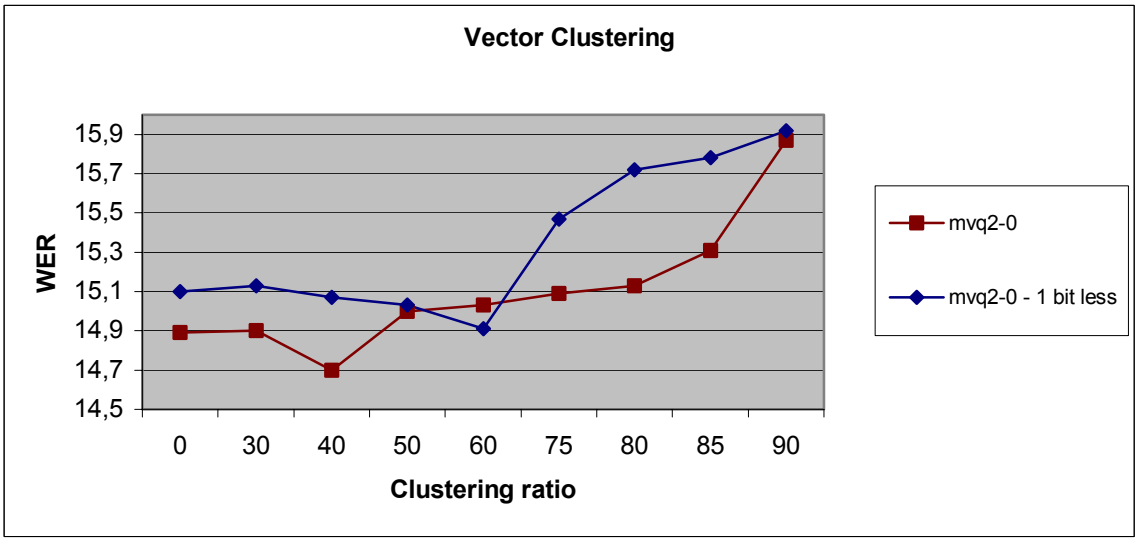
Η τεχνική διανυσματικής ομαδοποίησης εφαρμόστηκε στο επίπεδο της δομής όπου αποθηκεύονται τα counts για κάθε κατανομή, επιλέχθηκε ο αριθμός κατανομών που αντιστοιχεί στον εκάστοτε βαθμό ομαδοποίησης ενώ στο πεδίο των μοντέλων υιοθετήθηκε η επανάληψή τους (replication) στα απαιτούμενα σημεία.
Αυτό έχει σαν συνέπεια πρακτικά να μην υπάρχει μείωση στο μέγεθος του ακουστικού μοντέλου και στην απαιτούμενη μνήμη κατά τη διάρκεια της αναγνώρισης, ενώ επίσης δεν μπορεί να εκτιμηθεί με τον τρόπο αυτό και το επιπλέον υπολογιστικό κόστος που θα είχε η τυχαία προσπέλαση στη μνήμη για να βρεθεί η κατάλληλη κατανομή κατά τη διάρκεια της αναγνώρισης. Βέβαια ο στόχος της μείωσης των ελεύθερων παραμέτρων του μοντέλου επιτεύχθηκε. Η εκδοχή που δοκιμάστηκε περιγράφεται στο προηγούμενο κεφάλαιο (χρήση πολλαπλών pools). Στη συγκεκριμένη περίπτωση ο συνολικός αριθμός των pools διανυσματικών κατανομών ανέρχεται σε 12 (ίσος με τον αριθμό των υποδιανυσμάτων).

5.6.1.2 Αποτελέσματα και παρατηρήσεις

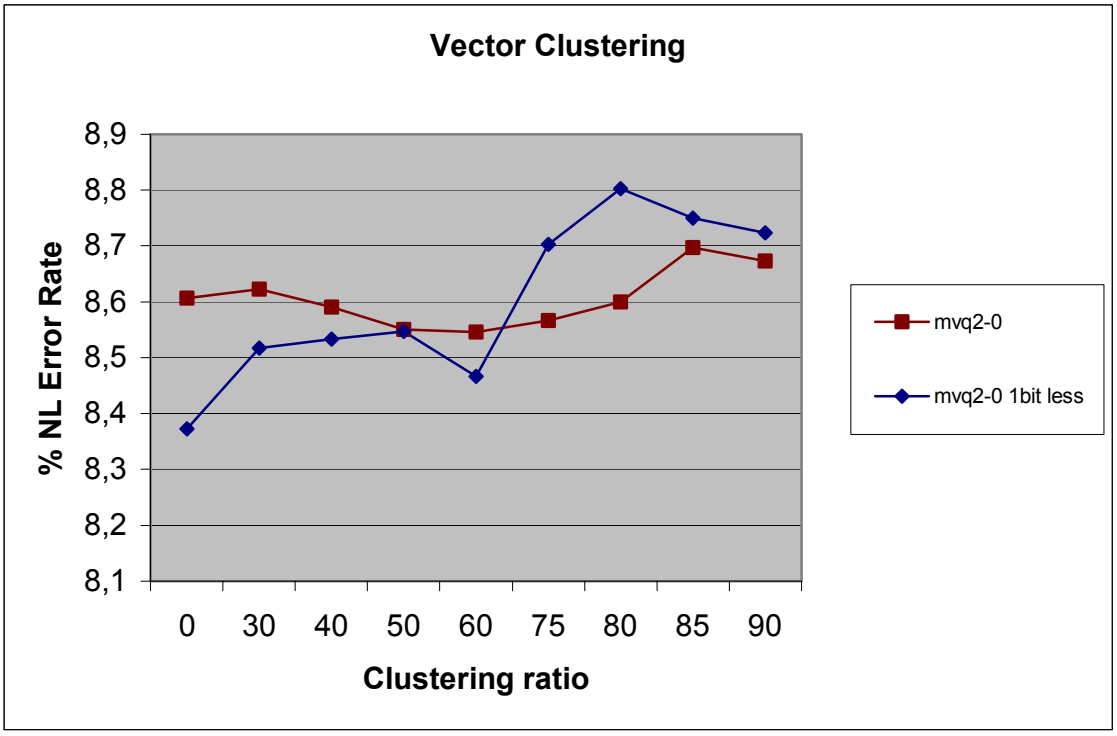
Αρχικά, έγιναν πειράματα εφαρμόζοντας την τεχνική διανυσματικής ομαδοποίησης σε ένα συγκεκριμένο ακουστικό μοντέλο που προκύπτει στο τέλος του σταδίου ***mvq1-1*** για διάφορες τιμές βαθμού ομαδοποίησης ώστε να προκύψει το μοντέλο του σταδίου ***mvq2-0***. Ο βαθμός ομαδοποίησης (clustering ratio) ουσιαστικά καθορίζει τον τελικό αριθμό διανυσματικών κατανομών που ανήκουν σε κάθε pool, βάσει της σχέσης:

$$clustering\ ratio = (1 - \frac{\text{final size of pool}}{\text{original size of pool}}) \cdot 100\%$$

Στη δικιά μας περίπτωση ο αρχικός αριθμός διανυσματικών κατανομών ήταν 1000 (ίσος με τον αριθμό μειγμάτων στο σύστημα), ενώ ο τελικός αριθμός κατανομών σε κάθε pool είναι ίδιος και υπολογίζεται από την παραπάνω σχέση.

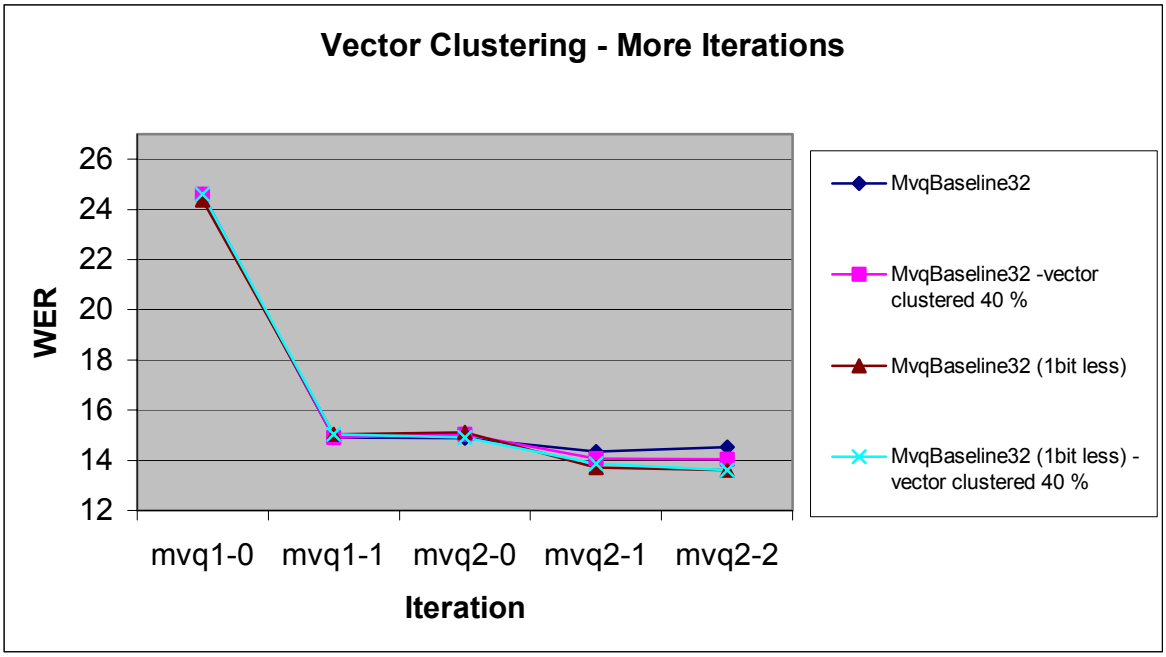


Σχήμα 5-9

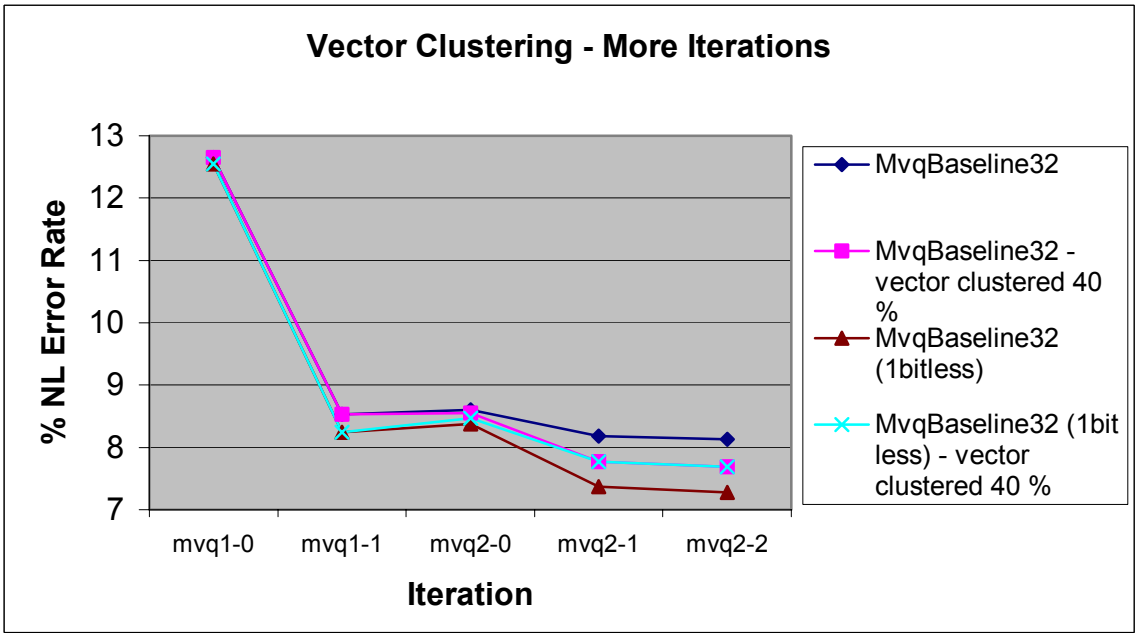


Σχήμα 5-10

Στα παραπάνω γραφήματα, παρατηρούμε ότι επιτρέπεται ένας σημαντικός βαθμός ομαδοποίησης και άρα μείωσης του αριθμού των ελεύθερων παραμέτρων του μοντέλου πριν αυτό να έχει σημαντική επίπτωση στην επίδοση του μοντέλου.
Για να εκτιμηθεί η συμπεριφορά του ομαδοποιημένου ακουστικού μοντέλου για περισσότερα training iterations, επιλέχθηκε ο βαθμός ομαδοποίησης που επιτρέπει 400 κατανομές για κάθε pool (μείωση στο 40% των αρχικών) και εφαρμόστηκε η διαδικασία που περιγράφεται στο σχήμα 5.1. Όπως φαίνεται και στο ακόλουθο γράφημα, η «καλή» αυτή συμπεριφορά των «συμπιεσμένου» μοντέλου διατηρείται και για περισσότερα στάδια:



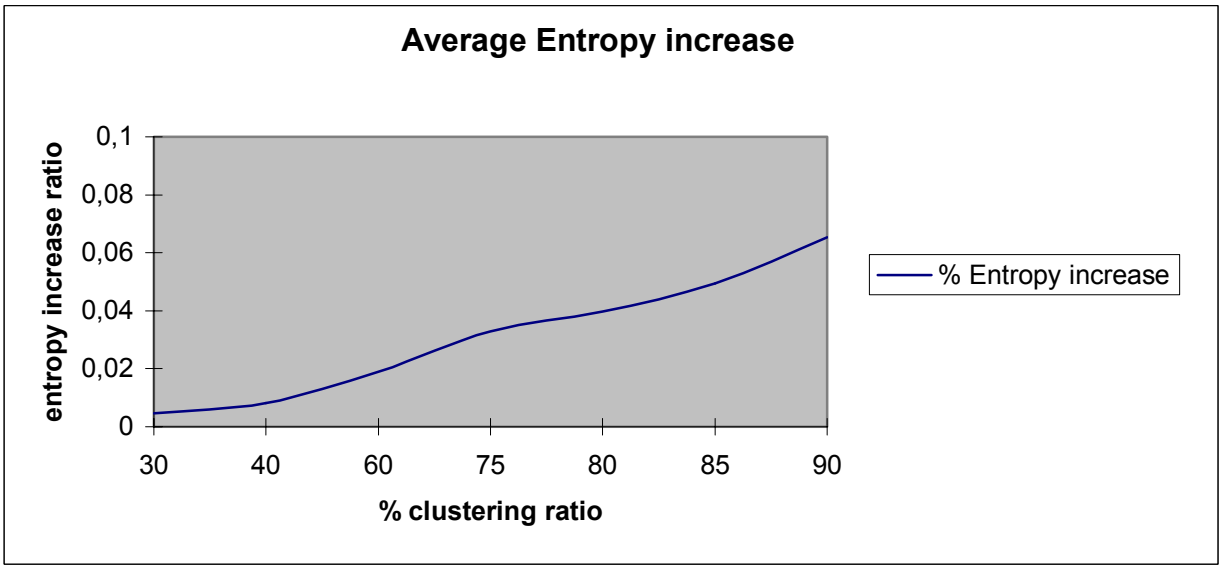
Σχήμα 5-11



Σχήμα 5-12

Κατά τη διαδικασία της εκτίμησης της μορφής των τελικών κατανομών που προκύπτουν από την ομαδοποίηση των αρχικών εισέρχεται μείωση στην πληροφορία των κατανομών (information loss) και μια ομαλοποίησή τους. Αυτό καταγράφεται και μέσω της αύξησης της εντροπίας των τελικών κατανομών σε σχέση με τις αρχικές όπως φαίνεται και στο ακόλουθο γράφημα:

$$entropy\ increase\ ratio = (1 - \frac{\text{final average entropy}}{\text{original average entropy}})$$



Σχήμα 5-13 Μέση αύξηση εντροπίας των τελικών κατανομών για κάθε βαθμό ομαδοποίησης

5.6.2 Ομαδοποίηση σε επίπεδο φύλλου (leaves clustering)

5.6.2.1 Θέματα υλοποίησης

Εφαρμόστηκε η τεχνική ομαδοποίησης σε επίπεδο φύλλου και στις δύο εκδοχές της όπως αυτές παρουσιάζονται στην αντίστοιχη παράγραφο του προηγούμενου κεφαλαίου. Και στην περίπτωση αυτή η ομαδοποίηση εφαρμόζεται μετά από ένα iteration και εφαρμογή των εξισώσεων επανεκτίμησης των μοντέλων μειγμάτων διακριτών κατανομών, όπως φαίνεται και στο σχήμα 5.1.

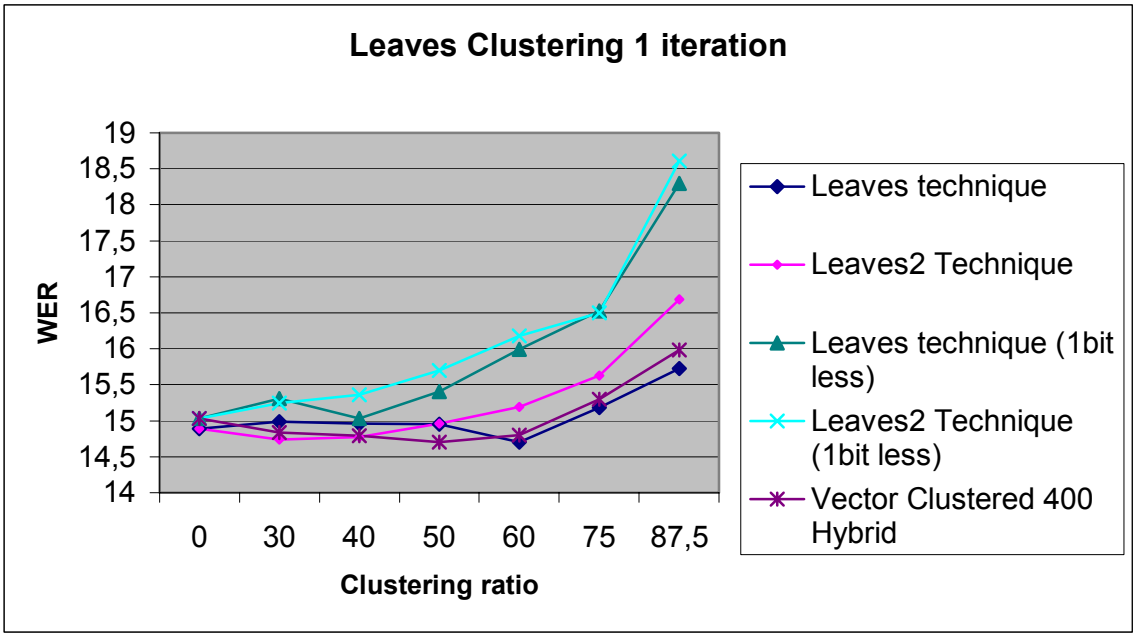
Αφού στο στάδιο ***mvq2-0*** αποφασιστεί το σχήμα της ομαδοποίησης σε επίπεδου φύλλου και για τα επόμενα iterations, το front-end τμήμα της επεξεργασίας φροντίζει να αντιστοιχίζει τα στοιχεία του διανύσματος που προκύπτει από το στάδιο της διανυσματικής κβαντοποίησης στα κατάλληλα leaves (centroids) που έχουν «επιβιώσει» της διαδικασίας ομαδοποίησης.

5.6.2.2 Αποτελέσματα και παρατηρήσεις

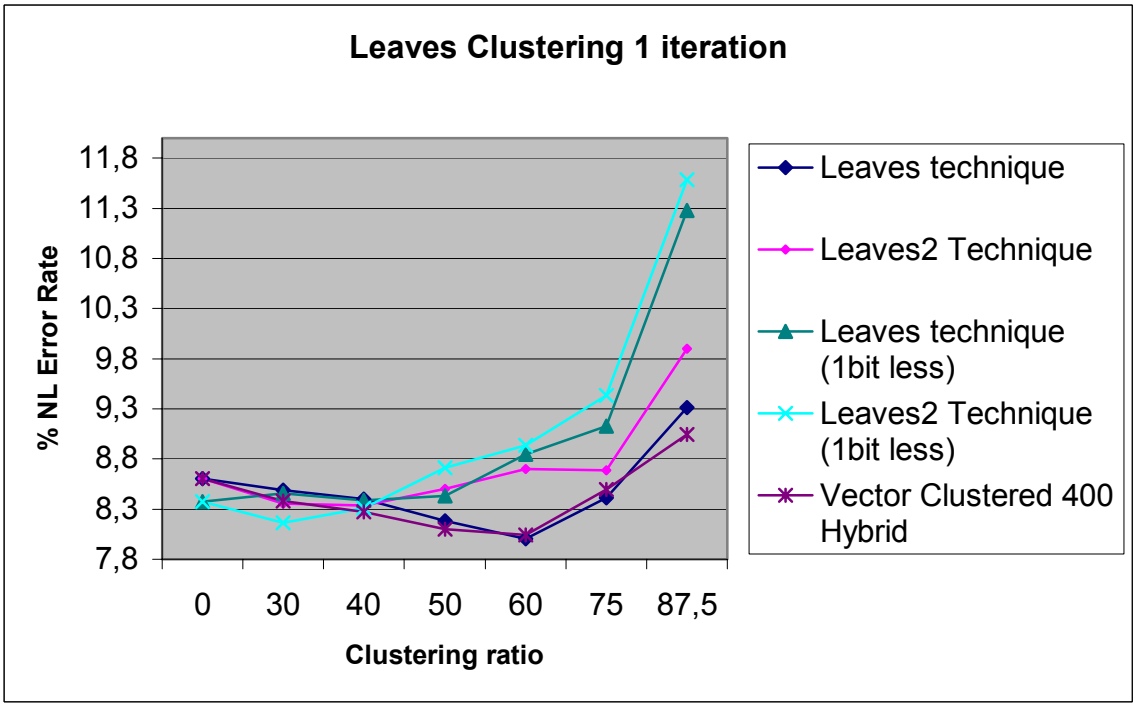
Αρχικά μελετήθηκαν οι τεχνικές ομαδοποίησης σε επίπεδο φύλλου σε ακουστικά μοντέλα που προκύπτουν μετά από το στάδιο mvq1-1 της διαδικασίας εκπαίδευσης για διάφορες τιμές βαθμού ομαδοποίησης, ο οποίος στην περίπτωση αυτή καθορίζει τον τελικό αριθμό των φύλλων που «επιβιώνουν» για κάθε υποδιάνυσμα, βάσει της σχέσης:

$$clustering\ \ ratio = (1 - \frac{\text{final number of leaves}}{\text{original number of leaves}}) \cdot 100\%$$

Υπενθυμίζουμε ότι ο αριθμός των φύλλων είναι διαφορετικός σε κάθε υποδιάνυσμα και εξαρτάται από το σχήμα διανυσματικής κβαντοποίησης που έχει επιλεγεί. Σαν leaves τεχνική αναφέρεται η τεχνική με τις κατανομές που βασίζονται στην a-posteriori πιθανότητα βάσει των διανυσματικών κατανομών, ενώ το εναλλακτικό σχήμα ομαδοποίησης φύλλων με μικρότερες κατανομές αναφέρεται σαν leaves2.

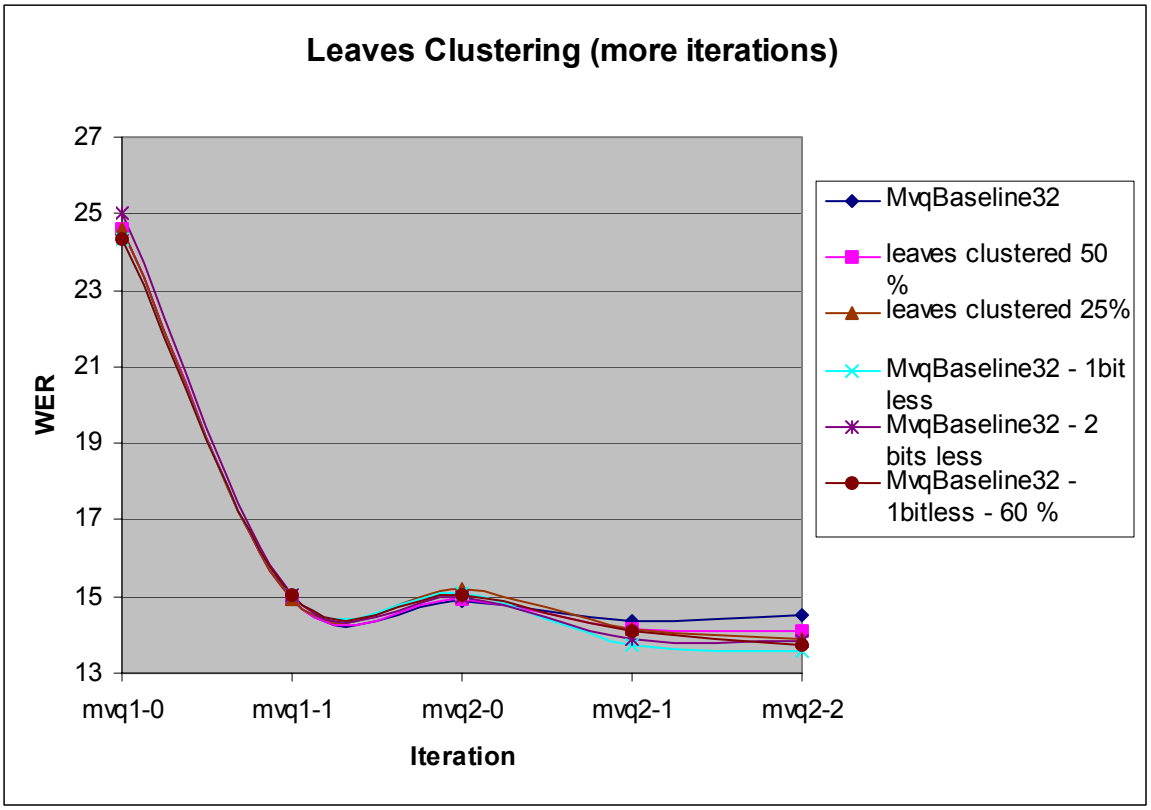


Σχήμα 5-14

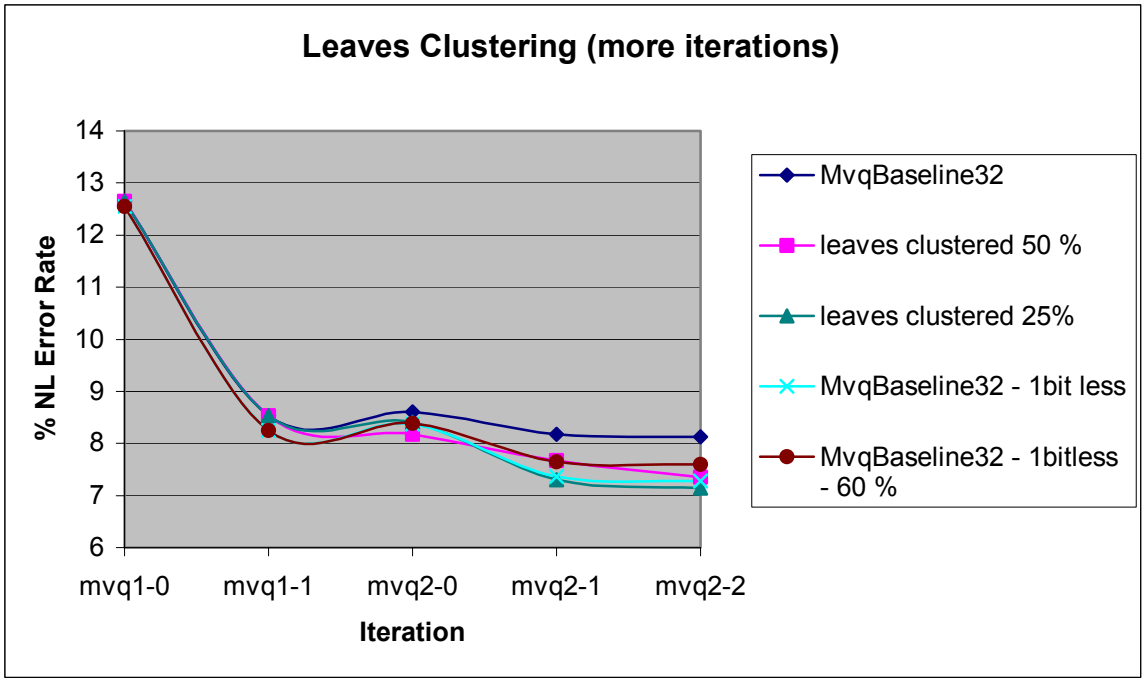


Σχήμα 5-15

Και με την τεχνική αυτή παρατηρούμε ότι υπάρχει μεγάλο περιθώριο μείωσης του αριθμού των ελεύθερων παραμέτρων του μοντέλου πριν αυτό αρχίσει να εμφανίζει μειωμένη επίδοση σε σχέση με το αρχικό. Η συμπεριφορά αυτή των ομαδοποιημένων (clustered) μοντέλων παραμένει και μετά από επιπλέον iterations εφαρμογής των εξισώσεων επανεκτίμησης:

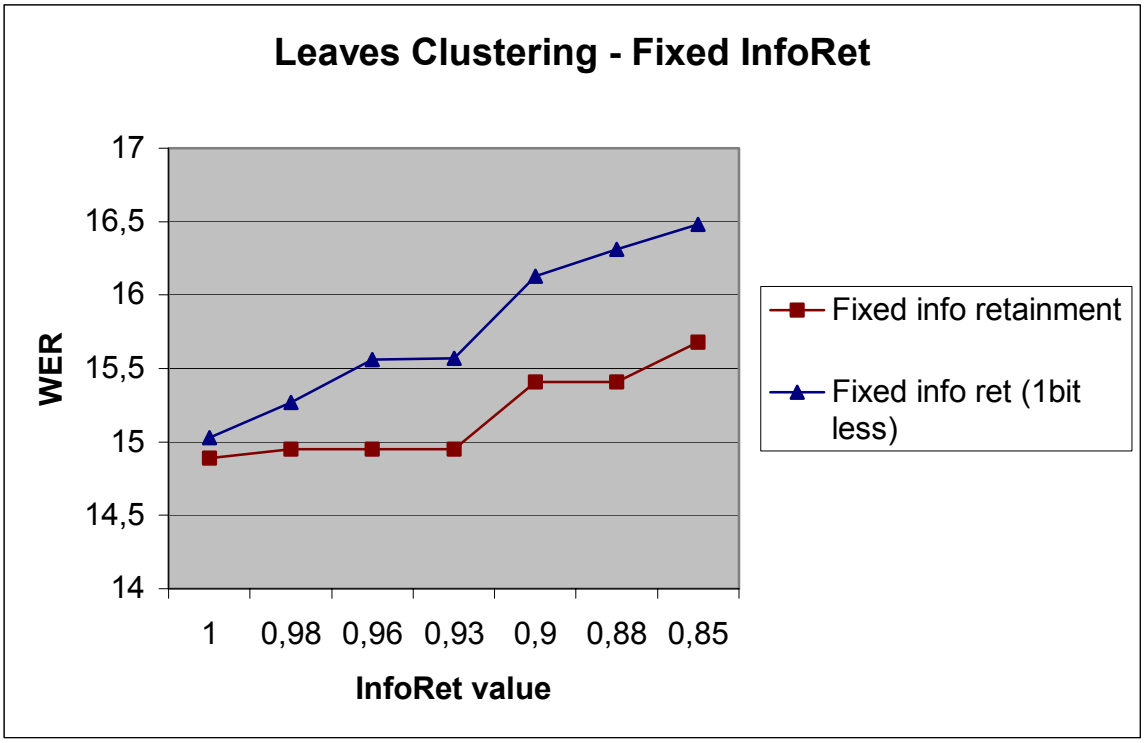


Σχήμα 5-16



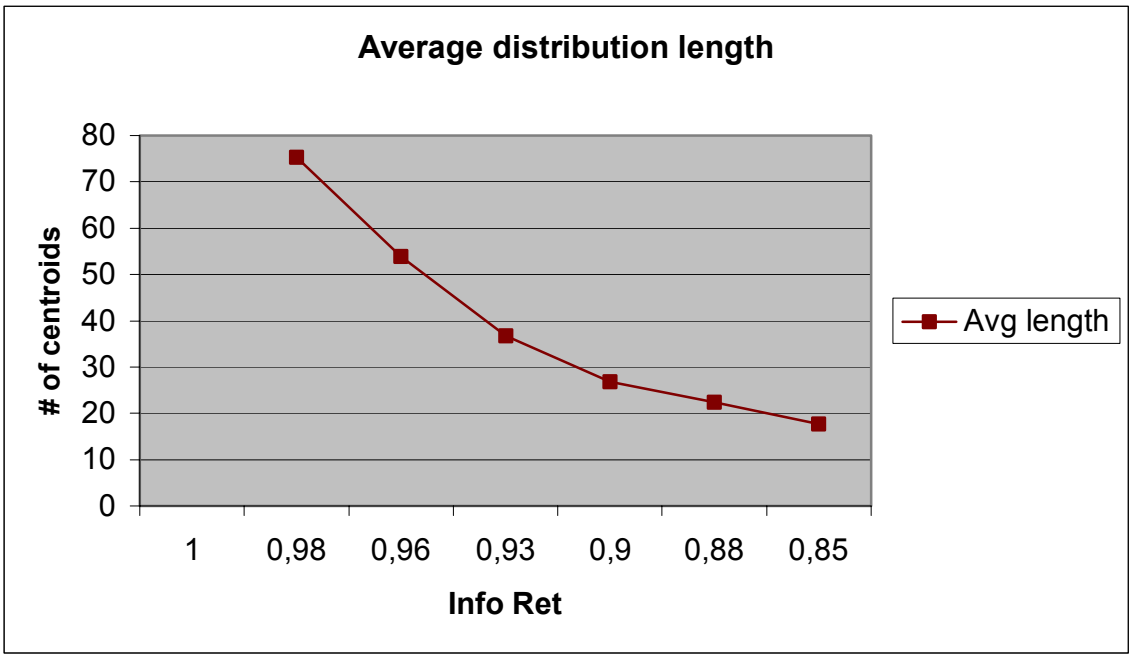
Σχήμα 5-17

Στα παραπάνω πειράματα ο τελικός αριθμός φύλλων (centroids) για κάθε υποδιάνυσμα είναι κλάσμα του αρχικού αριθμού, κάτι που βάσει της παραπάνω σχέσης για τον βαθμό ομαδοποίησης είναι το ίδιο για όλα τα υποδιανύσματα. Στα παρακάτω γραφήματα παρουσιάζονται τα αποτελέσματα για διαφορετικό κλάσμα τελικού και αρχικού αριθμού φύλλων για κάθε υποδιάνυσμα. Εξετάζεται η επίδοσή τους για διαφορετικές τιμές του ποσοστού μείωσης της πληροφορίας (και άρα αύξησης της εντροπίας των κατανομών) (*InfoRet* value) που προκύπτουν από τη διαδικασία ομαδοποίησης σε σχέση με την μέση εντροπία των αρχικών κατανομών. Σημειώνεται ότι όλα τα πειράματα έγιναν στο ακουστικό μοντέλο που προέκυψε από το στάδιο ***mvq1-1*** της διαδικασίας του σχήματος 5.1 για 32 components ανά μείγμα.


Leaves Clustering - Fixed InfoRet
WER
17
16,5
16
15,5
15
14,5
14
1
0,98
0,96
0,93
0,9
0,88
0,85
InfoRet value
Fixed info retainment
Fixed info ret (1bit less)

Σχήμα 5-18


Leaves Clustering - Fixed InfoRet
% NL Error Rate
10
9,5
9
8,5
8
7,5
1
0,98
0,96
0,93
0,9
0,88
0,85
InfoRet value
Fixed info retainment
Fixed info ret (1bit less)



Σχήμα 1-19 Μέσο μήκος των κατανομών που προκύπτουν για σταθερή μείωση της πληροφορίας (InfoRet)

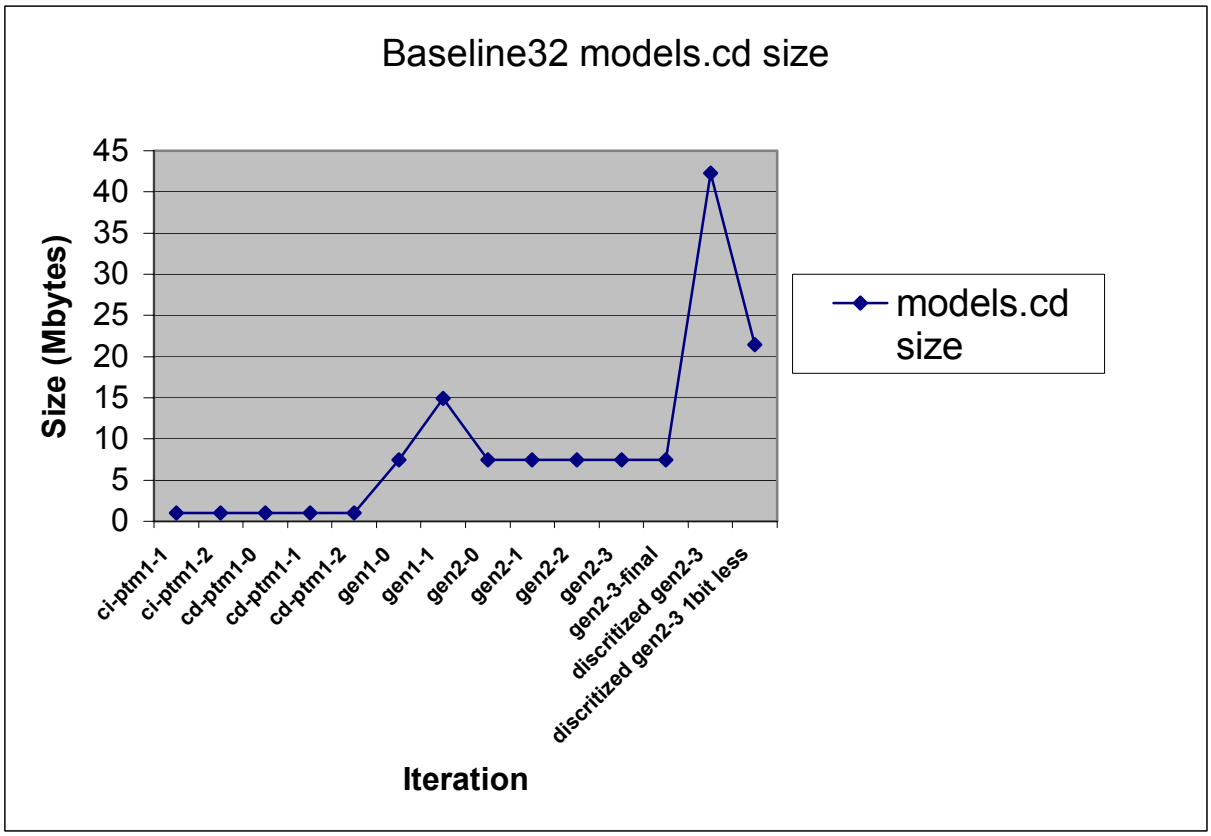
5.7 Μελέτη μεγέθους των τελικών μοντέλων

Οι παραπάνω τεχνικές όπως τονίστηκε και στο προηγούμενο κεφάλαιο στόχο είχαν τη μείωση του αριθμού των παραμέτρων του ακουστικού μοντέλου. Η μείωση αυτή επιτεύχθηκε μέσω της μείωσης του αριθμού αλλά και της διάστασης των διανυσματικών κατανομών του μοντέλου. Όπως εξηγήθηκε και στο προηγούμενο κεφάλαιο οι παράμετροι των διανυσματικών κατανομών αποτελούν τη συντριπτική πλειοψηφία των συνολικών παραμέτρων του μοντέλου. Στον ακόλουθο πίνακα φαίνονται ενδεικτικά τα μεγέθη των αρχείων των διαφόρων παραμέτρων του μοντέλου:

Τύπος παραμέτρων	**Αρχείο**	**Μέγεθος αρχείου (bytes)**	**% Ποσοστό μείωσης**	**Μέγεθος αρχείου (bytes)**	**Ποσοστό**
Κατανομές	models.cd	27444313	92,8	16686089	88,68941
Καταστάσεις και πιθανότητες μετάβασης	models.models	761488	88,7	763856	4,060025
Βάρη μειγμάτων	models.mw	1356757	4,06003	1361002	7,233958
a-priori πιθανότητες	models.priors	3124	7,23396	3124	0,016605

Πίνακας 5-3 Ενδεικτικά μεγέθη αρχείων παραμέτρων ακουστικών μοντέλων

Στο παρακάτω γράφημα φαίνεται ενδεικτικά το μέγεθος του αρχείου των παραμέτρων που αφορούν τις (πολυδιάστατες κανονικές ή διακριτές) κατανομές του μοντέλου στα διάφορα στάδια εκπαίδευσης όπως αυτή εφαρμόστηκε για το baseline πείραμα.



Σχήμα 5-20

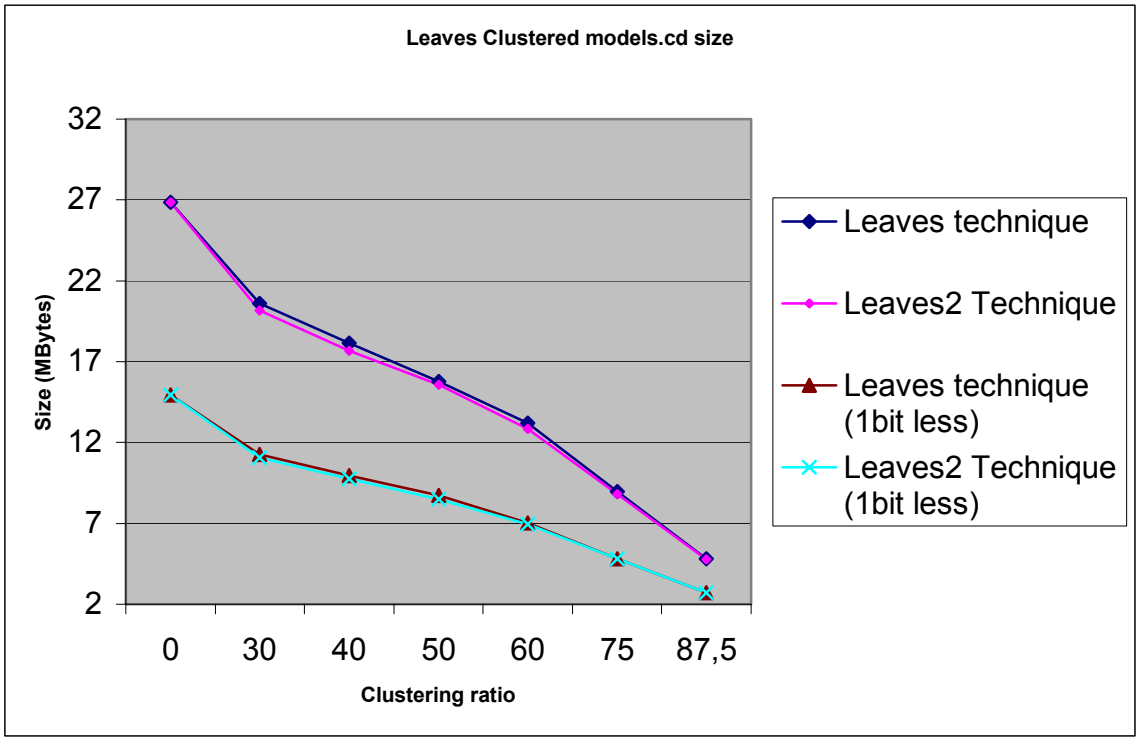
Παρατηρείται η μεγάλη αύξηση στο μέγεθος των αρχείων των μοντέλων στο στάδιο της διακριτοποίησης. Καθοριστικό ρόλο παίζει ο αριθμός των bits ανά υποδιάνυσμα. Η διάσταση των διαφόρων κατανομών κατά τη διακριτοποίηση με 1 bit λιγότερο ανά υποδιάνυσμα πέφτει στο μισό των αρχικών κάτι που φαίνεται και στο μέγεθος του αντίστοιχου αρχείου.

Αντίστοιχα γραφήματα παρουσιάζονται παρακάτω και για τις περιπτώσεις της εναλλακτικής διαδικασίας εκπαίδευσης αλλά και της εφαρμογής των τεχνικών ομαδοποίησης.

***Σημείωση*:** Για λόγους εξοικονόμησης χώρου στο δίσκο τα αρχεία, στην περίπτωση των μοντέλων μειγμάτων διακριτών κατανομών, συμπιέζονται κατά Lempel-Zipf και

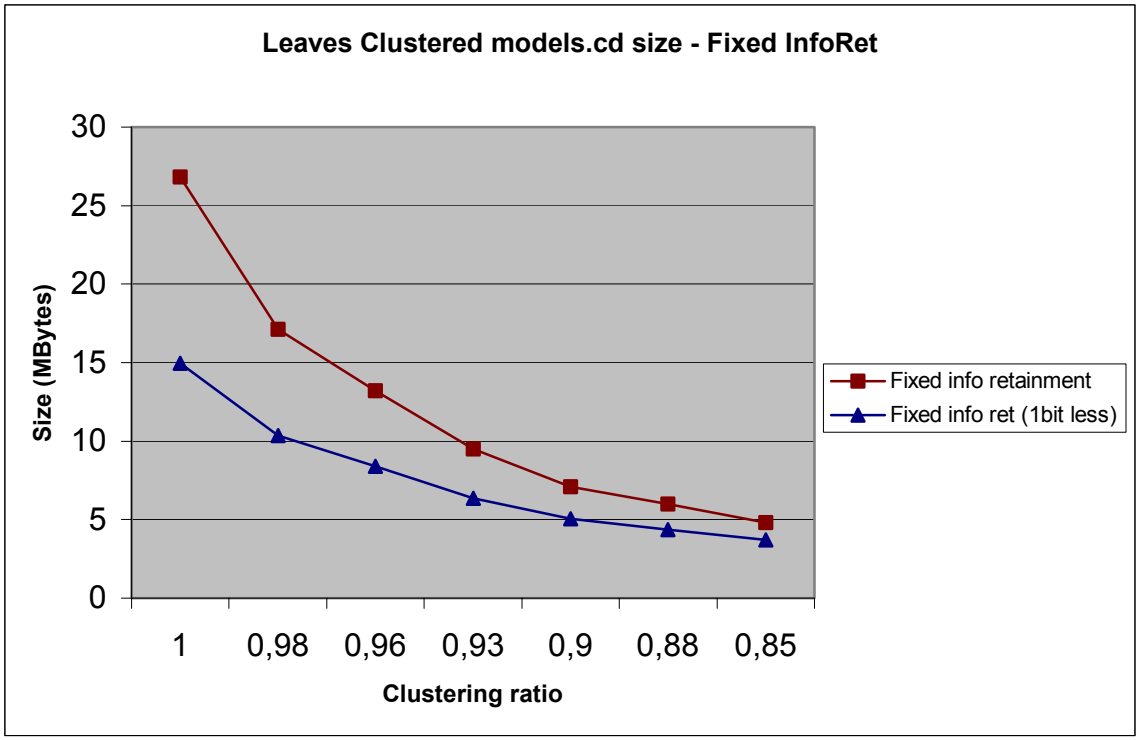
άρα δεν είναι απαραίτητα ίδιο το απαιτούμενο ποσό μνήμης στη διάρκεια της διαδικασίας της αναγνώρισης.

Όπως αναφέρθηκε και παραπάνω, αρχικά έγιναν πειράματα για διαφορετικό βαθμό ομαδοποίησης σε συγκεκριμένο ακουστικό μοντέλο που προκύπτει από το στάδιο ***mvq1-1***.



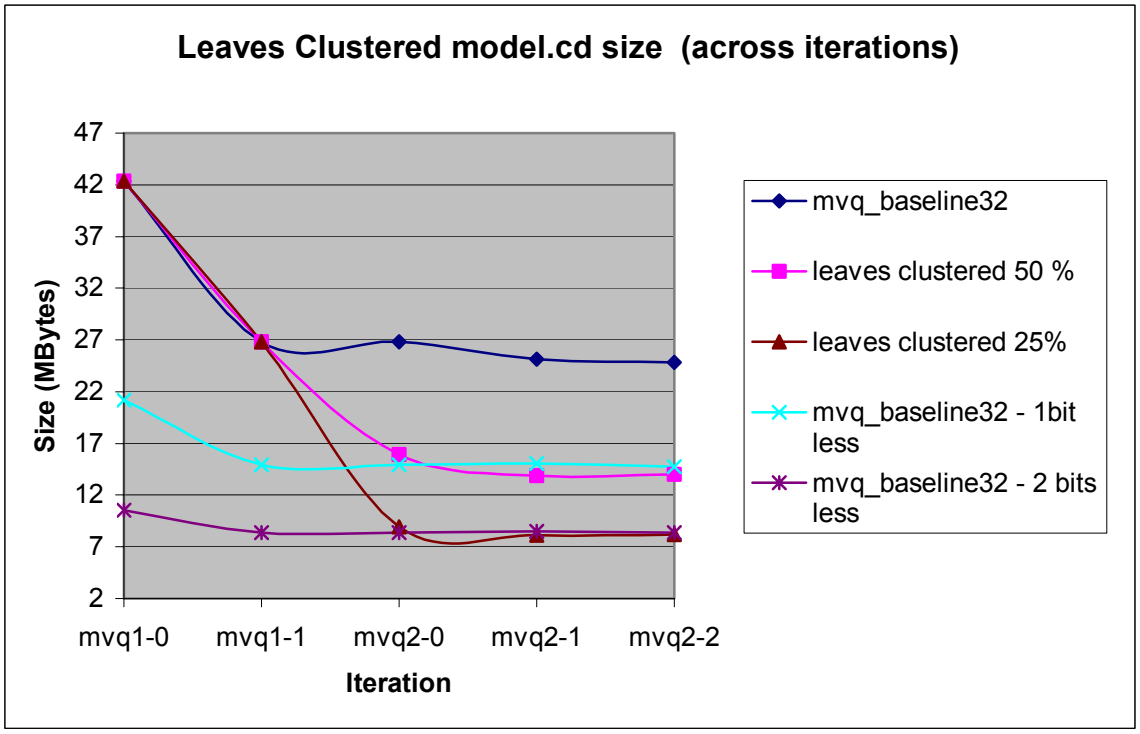
Σχήμα 5-21

Και στην περίπτωση της διατήρησης του ποσοστού ελάττωσης της πληροφορίας σε σταθερά επίπεδα για τα διάφορα υποδιανύσματα, παρατηρήθηκε αντίστοιχη συμπεριφορά ως προς τη μείωση του μεγέθους του αρχείου των παραμέτρων:


Leaves Clustered models.cd size - Fixed InfoRet
Size (MBytes)
30
25
20
15
10
5
0
1
0,98
0,96
0,93
0,9
0,88
0,85
Clustering ratio
Fixed info retainment
Fixed info ret (1bit less)

Σχήμα 5-22

Η μείωση στο μέγεθος του αρχείου των παραμέτρων των διανυσματικών κατανομών είναι εμφανής και κατά την εφαρμογή της εναλλακτικής διαδικασίας εκπαίδευσης στα διάφορα iterations που ακολουθούν τη διακριτοποίηση:



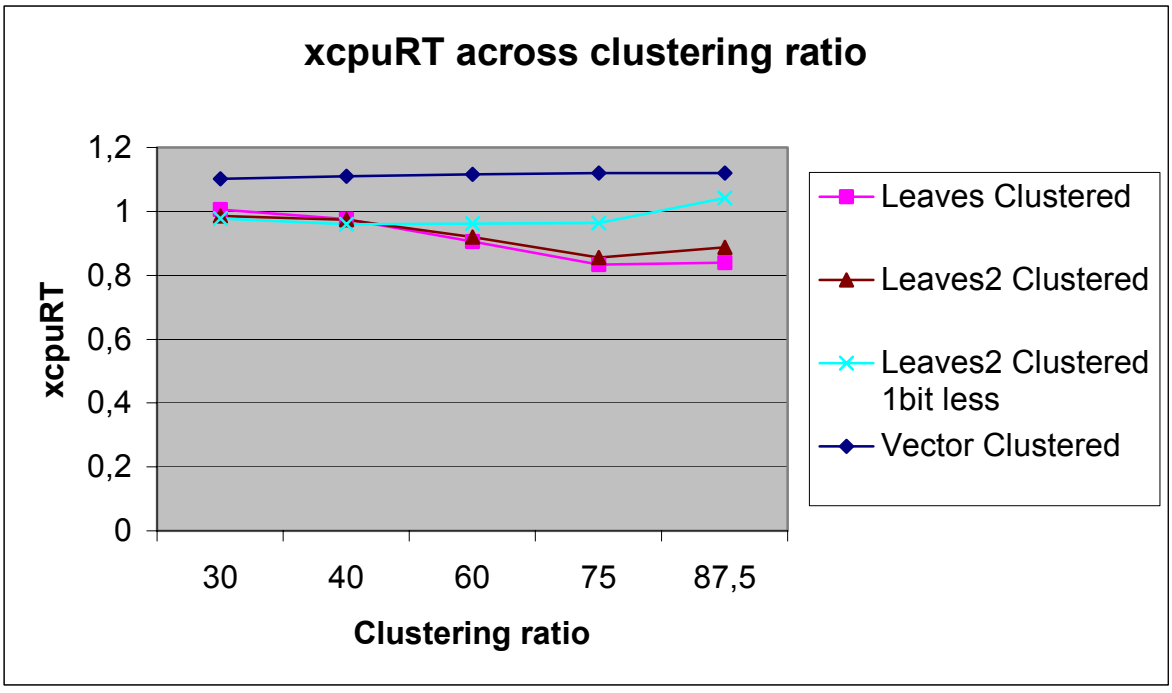
Σχήμα 5-23

Στην περίπτωση του διανυσματικής ομαδοποίησης (vector clustering) το τελικό μέγεθος παραμένει στα ίδια επίπεδα μιας και, όπως τονίστηκε και παραπάνω, πρακτικά υπήρχε επανάληψη (replication) των κατανομών σε διάφορα σημεία του τελικού αρχείου κατανομών για να επιτευχθεί συμβατότητα με την τρέχουσα υλοποίηση.

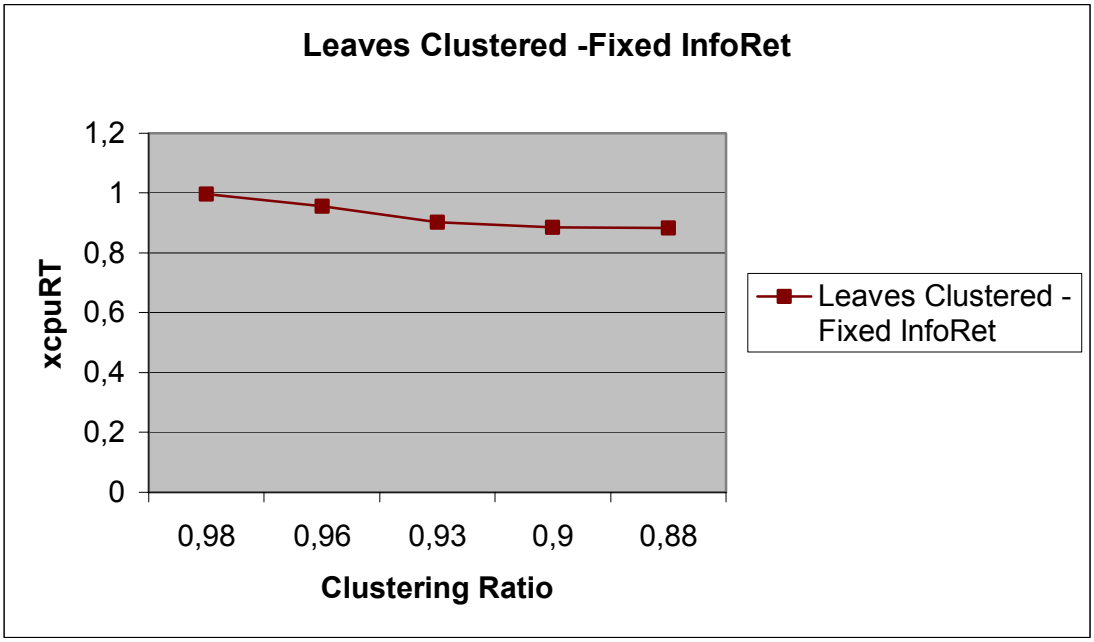
5.8 Αποτελέσματα ως προς την ταχύτητα αναγνώρισης

Στα παρακάτω σχήματα εμφανίζονται οι απαιτήσεις ως προς την υπολογιστική ισχύ (ταχύτητα αναγνώρισης) για ένα από τα testsets που χρησιμοποιήθηκαν για τον έλεγχο της επίδοσης των εξεταζόμενων τεχνικών. Τα αποτελέσματα για τα baseline πειράματα μοντέλων συνεχών κατανομών, που περιλαμβάνονται στο γράφημα, αντιστοιχούν στα genonic (genX-X) iterations. Οι τιμές του χρόνου είναι κανονικοποιημένες ως προς τη διάρκεια των προτάσεων του testset αλλά και ως προς τις υπολογιστικές δυνατότητες του μηχανήματος όπου έγιναν τα πειράματα και αναφέρονται ως μονάδες πραγματικού χρόνου (xcpuRT).

Αρχικά μελετήθηκε η επίπτωση των τεχνικών ομαδοποίησης στη συμπεριφορά ως προς την ταχύτητα αναγνώρισης ως προς τον βαθμό ομαδοποίησης:



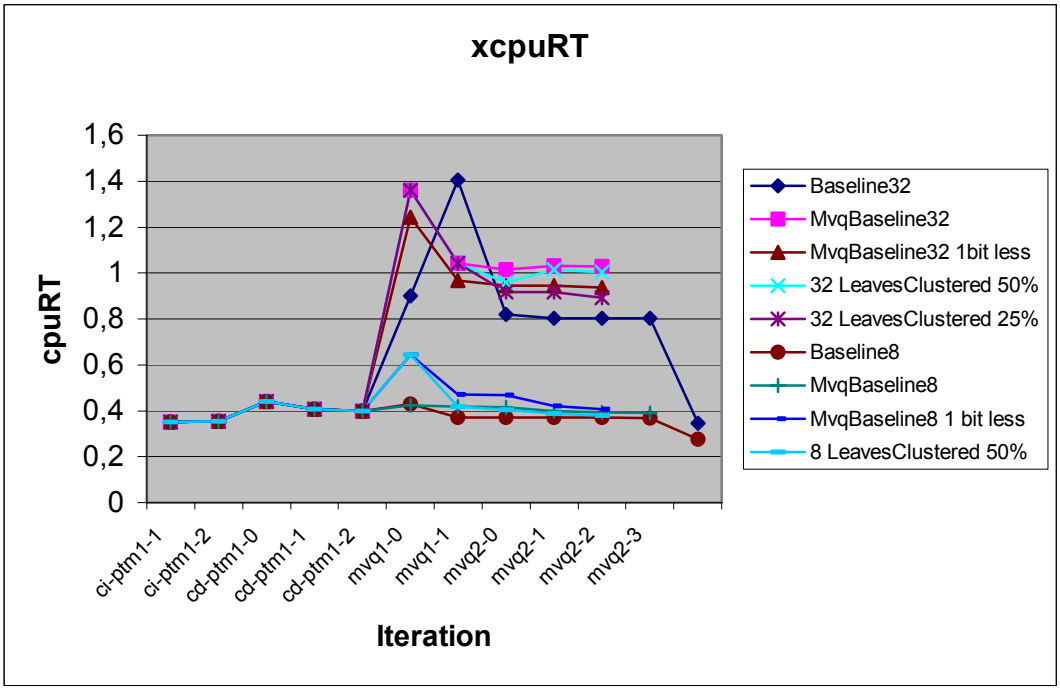
Σχήμα 5-24



Σχήμα 5-25

Στην περίπτωση της διανυσματικής ομαδοποίησης (vector clustering), δεν έχει ληφθεί υπόψη το επιπλέον κόστος του indirection που εμπλέκεται κατά την ανάκτηση της (προϋπολογισμένης) τιμής μιας κατανομής.

Στο ακόλουθο γράφημα συμπεριλαμβάνονται και αποτελέσματα πειραμάτων με 8 components ανά μείγμα για καλύτερη σύγκριση των τιμών.



Σχήμα 5-26 Ενδεικτικές υπολογιστικές απαιτήσεις κατά την αναγνώριση

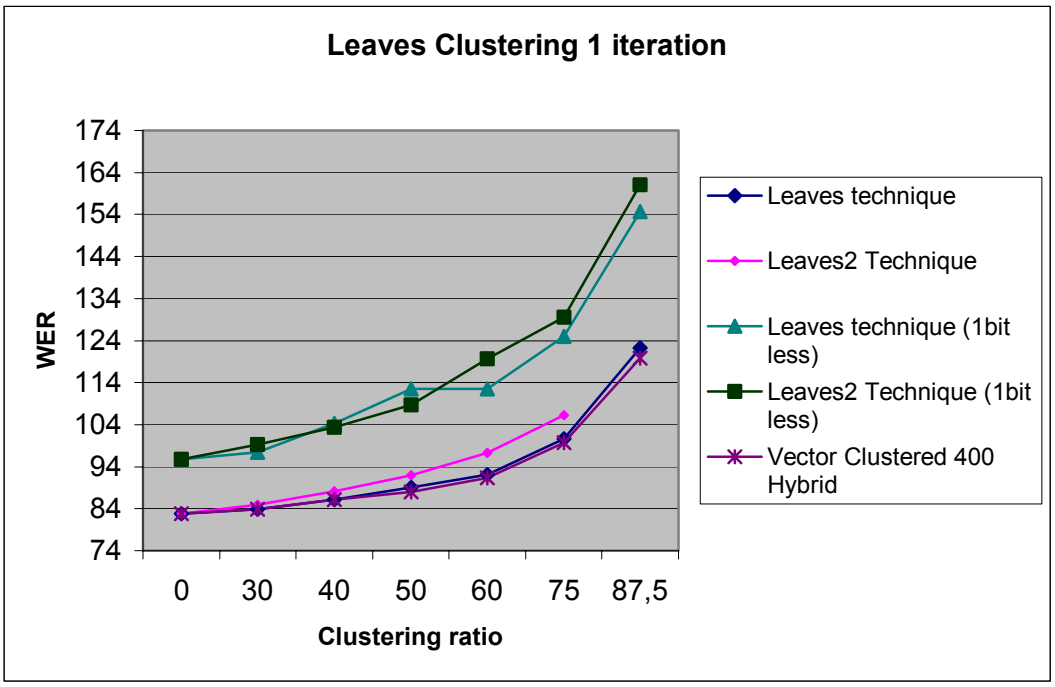
Όπως αναμένεται σημαντικό ρόλο στην τελική ταχύτητα αναγνώρισης παίζει ο αριθμός των components ανά mixture ή genone. Γενικά παρατηρούμε ότι όσο μεγαλύτερος είναι ο αριθμός των Gaussians σε κάθε genone στην περίπτωση των genonic HMMs, τόσο μεγαλύτερες είναι οι απαιτήσεις πραγματικού χρόνου κατά τη διάρκεια της αναγνώρισης. Το peak που παρατηρείται στα στάδια ***gen1-0*** και ***gen1-1***, οφείλεται στο ότι εκεί ο αριθμός των συστατικών πολυδιάστατων κατανομών (components) είναι 64 αντί για 32.

Παρατηρούμε επίσης ότι στη συγκεκριμένη υλοποίηση δεν είναι εμφανές το αναζητούμενο κέρδος στην ταχύτητα με τη χρήση μοντέλων μειγμάτων διακριτών κατανομών έναντι των αντίστοιχων με συνεχείς (κανονικές) κατανομές. Βέβαια στην περίπτωση της «ενσωματωμένης» (embedded) διαδικασίας αναγνώρισης, σε πολλές περιπτώσεις δεν υπάρχει καν η επιλογή της προσέγγισης των συνεχών κατανομών μιας και απουσιάζει (για λόγους εξοικονόμησης ενέργειας και αυτονομίας) μονάδα επεξεργασίας αριθμών κινητής υποδιαστολής (floating point unit).

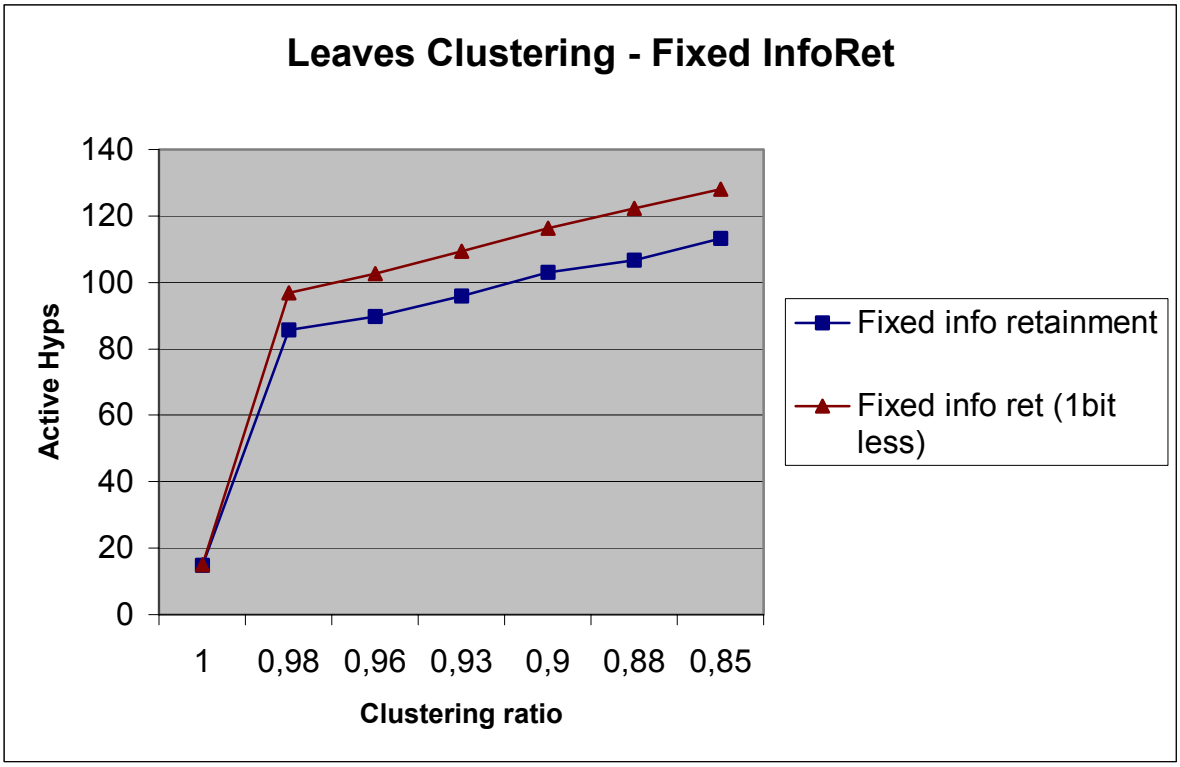
Η υπεροχή των μοντέλων μειγμάτων συνεχών κατανομών στη συγκεκριμένη περίπτωση οφείλεται στον αυξημένο αριθμό προσβάσεων στη μνήμη (για να

ανακτηθούν οι προϋπολογισμένες τιμές) για τα μοντέλα μειγμάτων διακριτών κατανομών. Στην περίπτωση των διακριτών μοντέλων *δεν* έχουν εφαρμοστεί τεχνικές βελτιστοποίησης αυτής της διαδικασίας (π.χ. η τεχνική των mixturelists αντίστοιχη της τεχνικής των Gaussian Shortlists [1] στο πεδίο των συνεχών κατανομών).

Οι απαιτήσεις παραμένουν στα ίδια επίπεδα παρά την βελτιστοποίηση του μεγέθους μιας και παρατηρείται μια αύξηση των ενεργών υποθέσεων στην εφαρμογή της τεχνικής του beam-search κατά την αναγνώριση.



Σχήμα 5-27


Leaves Clustering - Fixed InfoRet
Active Hyps
140
120
100
80
60
40
20
0
1
0,98
0,96
0,93
0,9
0,88
0,85
Clustering ratio
Fixed info retainment
Fixed info ret (1bit less)

Σχήμα 5-28

5.9 Συμπεράσματα

Ο αριθμός των παραμέτρων του μοντέλου πριν την ομαδοποίηση κατανομών ή φύλλων είναι αρκετά μεγάλος. Έτσι, παρά το μεγάλο όγκο δεδομένων εκπαίδευσης, για πολλές από τις παραμέτρους δεν συγκεντρώνεται ο κατάλληλος αριθμός παρατηρήσεων με αποτέλεσμα να μην εκπαιδεύονται σωστά.

Με την εφαρμογή των τεχνικών ομαδοποίησης, ο αριθμός των ελεύθερων παραμέτρων μειώνεται σημαντικά, με αποτέλεσμα τα δεδομένα εκπαίδευσης να επαρκούν σε μεγάλο βαθμό για την εκτίμηση των τιμών τους και αυτό να αντικατοπτρίζεται στην μη μείωση και (σε μερικές περιπτώσεις καλυτέρευση) της επίδοσης του μοντέλου.

Με την υπερβολική μείωση του αριθμού των παραμέτρων, οι αντίστοιχες κατανομές τείνουν να «υπερεκπαιδευτούν» (overfitting phenomenon) στα δεδομένα εκπαίδευσης με αποτέλεσμα τη χαμηλή ανάλυση του ακουστικού χώρου και τη χαμηλή επίδοση σε δεδομένα διαφορετικά των δεδομένων εκπαίδευσης.

Την ίδια στιγμή επιτυγχάνεται και ο στόχος της μείωσης των απαιτήσεων σε μνήμη κατά την αναγνώριση ή αποθήκευση του μοντέλου μιας και ο «βαθμός συμπίεσης» είναι αρκετά μεγάλος.

6 Βιβλιογραφία

[1] V.Digalakis,P.Monaco,H.Murveit: **"Genones: Generalized Mixture Tying in Continuous Hidden Markov Model Based Speech Recognizers"**, IEEE Transactions Speech and Audio Processing, July 1996, pp. 281-289.

[2] V.Digalakis,S.Tsakalidis,L.Neumeyer: **"Efficient Speech Recognition Using Subvector Quantization and Discrete-Mixture HMMs"**, Proceedings ICASSP '99.

[3] V.Digalakis, L.Neumeyer, M.Perakakis: **"Quantization of Cepstral Parameters for Speech Recognition over the World Wide Web"**, IEEE Journal on Selected Areas of Communications, 1999.

[4] T. M. Cover and J. A. Thomas: **Elements of Information Theory**, John Wiley & Sons, 1991.

[5] J. Lin: **"Divergence measures based on the Shannon entropy"**, IEEE Trans. Inform. Theory, 37(1), 1991

[6] R. Duda, P. Hart, D. Stork: **Pattern Classification**, 2nd Edition, John Wiley & Sons, Inc, Chapter 10, Unsupervised Learning and Clustering, pp 517-601

[7] Γαβαλάκης Πέτρος: **"Συγκριτική μελέτη μεθόδων ομαλοποίησης διακριτών Μαρκοβιανών μοντέλων"**, Διπλωματική εργασία, Πολυτεχνείο Κρήτης, Τμήμα ΗΜΜΥ, Φεβρουάριος 2000

[8] L.Rabiner, B. Juang: **Fundamentals of Speech Recognition,** Prentice-Hall Signal Processing Series

[9] N. Slonim, N. Tishby: **Agglomerative Information Bottleneck**

[10] V.Digalakis, S.Tsakalidis, C. Harizakis, L. Neumeyer: **"Efficient speech recognition using subvector quantization and discrete-mixture HMMs",** Computer Speech and Language (1999) **00**, 1–14

[11] H. Murveit, V. Digalakis: **"Genones: Optimizing the degree of tying in a large vocabulary HMM-based speech recognizer"**, 1994 IEEE ICASSP, pp. I-537—I-540.