Γραμμικά Δυναμικά Μοντέλα και η Εφαρμογή τους στην Αναγνώριση Φωνής

Γιώργος Τσόντζος
Ηλεκτρονικών Μηχανικών & Μηχανικών Υπολογιστών

Επιβλέπων καθηγητής:
Βασίλης Διγαλάκης

Επιτροπή:
Βασίλης Διγαλάκης
Αλέξανδρος Ποταμιάνος
Αθανάσιος Λιάβας

Πολυτεχνείο Κρήτης
Χανιά - Ελλάδα

12 Οκτωβρίου 2006

Περιεχόμενα

Κατάλογος Σχημάτων

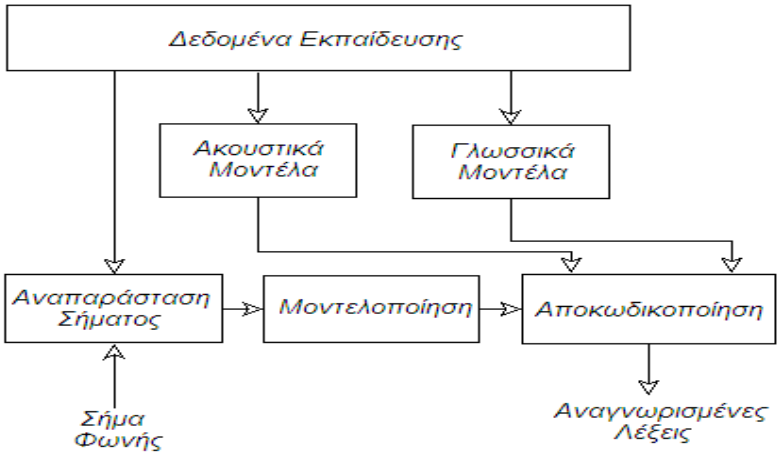
Κεφάλαιο 1

Αυτόματα Συστήματα Αναγνώρισης Φωνής

Η ταχεία εξέλιξη της τεχνολογίας έχει δημιουργήσει μεγάλες δυνατότητες στον έλεγχο, στην αποθήκευση και στη μετάδοση των πληροφοριών. Η επικοινωνία όμως μεταξύ ανθρώπου και μηχανής απαιτεί κάποιο μέσω ανάθεσης των εντολών για την αποκομιδή των πληροφοριών. Μεγάλο επίτευγμα θα ήταν αν η μηχανή μπορέσει να αποκτήσει την ιδιότητα κατανόησης της γλώσσας των ανθρώπων, με σκοπό την ανάθεση των εντολών χωρίς την μεσολάβηση κάποιου μέσου. Ο κλάδος της Αναγνώρισης Φωνής (Speech Recognition) έχει κάνει αρκετά βήματα για την επίτευξη του στόχου αυτού. Πραγματικές εφαρμογές έχουν ήδη αρχίσει να εμφανίζονται δείχνοντας την σπουδαιότητα ως κλάδος στην υπηρεσία του ανθρώπου.

Σ′ αυτό το κεφάλαιο θα αναλύσουμε τη δομή και τον τρόπο λειτουργίας, ενός αυτόματου συστήματος αναγνώρισης φωνής. Ένα γενικό αυτόματο σύστημα αναγνώρισης φωνής αποτελείται από τέσσερα κύρια τμήματα : η επεξεργασία του ακουστικού σήματος (Front-End), το ακουστικό μοντέλο (Acoustic Model), το γλωσσικό μοντέλο (Language Model) και η αποκωδικοποίηση (Decoding). Στο Σχήμα (1.1) φαίνετε πως συνδυάζονται τα τέσσερα αυτά τμήματα για την υλοποίηση ενός μοντέρνου αυτόματου συστήματος αναγνώρισης φωνής.

Το κεφάλαιο οργανώνεται ως ακολούθως: στο εδάφιο (1.1) εξηγούμε πως γίνεται η επεξεργασία του ακουστικού σήματος, στο εδάφιο (1.2) αναφέρονται οι λόγοι που κάνουν το ακουστικό μοντέλο απαραίτητο στην αναγνώριση της φωνής, καθώς επίσης και τον ορισμό του, στο εδάφιο (1.3) αναλύεται το γλωσσικό μοντέλο και στο εδάφιο (1.4) η σημαντικότητα της



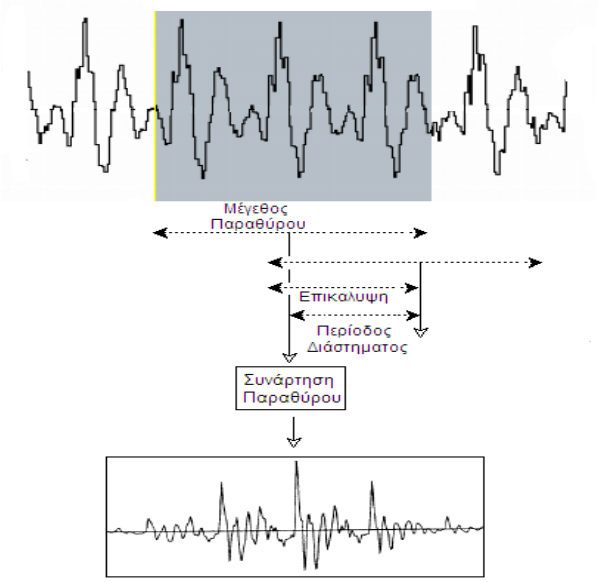
Σχήμα 1.1: Αυτόματοο Σύστημα Αναγνώρισης Φωνής

αποκωδικοποίησης.

1.1 Επεξεργασία του ακουστικού σήματος

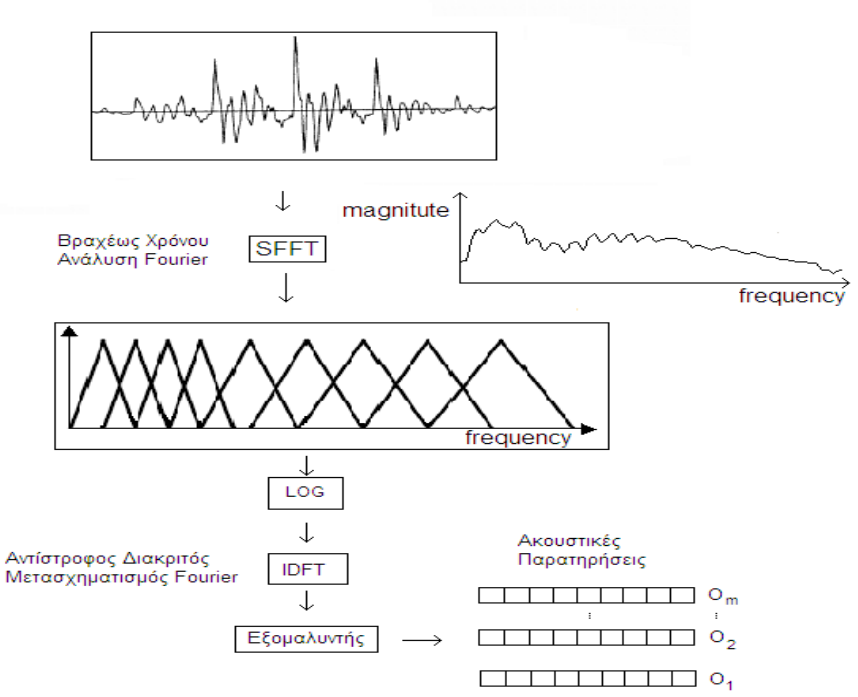
Ως πρώτο κύριο μέρος ενός συστήματος Αναγνώρισης Φωνής είναι η επεξεργασία του ακουστικού σήματος (Front-End). Η διαδικασία αναγνώρισης της φωνής ξεκινά απο τα σήματα της φωνής, που έχουν υποστεί δειγματοληψία. Η δειγματοληψία πραγματοποιείται έχοντας όλη την απαραίτητη πληροφορία του σήματος της φωνής σε ένα μικρό αριθμό συντελεστών, συμβατό με τα στοχαστικά μοντέλα που πρόκειται να την χρησιμοποιήσουν. (automatic speech recogniser - ASR).

Αναλυτικότερα, το σήμα της φωνής είναι τοπικά-στάσιμο (quasi-stationary) σήμα, όμως για μικρά χρονικά διαστήματα μπορεί να θεωρηθεί ως στάσιμο. Για τον λόγο αυτό κατα την επεξεργασία το σήμα διαιρείται σε διαστήματα (frames) και από κάθε διάστημα εξάγεται ένα εξομαλυσμένο φάσμα. Εφαρμόζοντας σε κάθε διάστημα μια συνάρτηση επικάλυψης παραθύρου με απόσβεση στα άκρα (έχει επικρατήσει Hamming window) μπορούμε να διατηρήσουμε τα τοπικά χαρακτηριστικά του σήματος. Συνήθως τα χρονικά διαστήματα είναι διάρκειας $10msec$, ενώ το παράθυρο που χρησιμοποιείται είναι συνήθως της τάξεως των $20msec$ με $25msec$. Στο σχήμα (1.2) φαίνεται η διαδικασία παραθυροποίησης ενός σήματος φωνής.



Σχήμα 1.2: Δειγματοληψία ακουστικού σήματος εφαρμόζοντας την τεχνική της επικάλυψης παραθύρου.

Επειδή το σήμα της φωνής είναι στάσιμο σε μικρά χρονικά διαστήματα, χρησιμοποιείται ο βραχέως χρόνου μετασχηματισμός Fourier (Short Time Fourier Transform - STFT). Ο βραχέως χρόνου μετασχηματισμός συνδέει τις μεταβλητές του μήκους του παραθύρου με τη συχνότητα. Άν θέσουμε μεγάλες τιμές στο μήκος του παραθύρου θα υπάρξει καλή ανάλυση στο πεδίο της συχνότητας ενώ αντίθετα μικρές τιμές στο μήκους του παραθύρου δίνουν χειρότερη ανάλυση στο πεδίο της συχνότητας, αλλά έτσι οι χρονικές μεταβλητές παρακολουθούνται καλύτερα. Το αποτέλεσμα του μετασχηματισμού, εισάγεται σε μια προκαθορισμένη ομάδα φίλτρων. Η διαδικασία συνεχίζεται λογαριθμίζοντας και εφαρμόζοντας αντίστροφο διακριτό μετασχηματισμό Fourier. Το διάνυσμα χαρακτηριστικών που προκύπτει είναι το cepstrum του σήματος της φωνής. Η ακουστική πληροφορία μπορεί να απεικονισθεί με μια ακολουθία τέτοιων χαρακτηριστικών διανυσμάτων εφαρμόζοντας κατάλληλο υπολογισμό εξομάλυνσης στα χαρακτηριστικά των φίλτρων. Τα χαρακτηριστικά αυτά διανύσματα ονομάζονται ακουστικές παρατηρήσεις και συμβολίζονται με $Y = y_1, y_2, ..., y_m$, όπου m ειναι ο αριθμός των παρατηρήσεων. Στο σχήμα (1.3) φαίνεται η διαδικασία εξαγωγής των ακουστικών παρατηρήσεων.



Σχήμα 1.3: *Η* διαδικασία εξαγωγής των ακουστικών παρατηρήσεων.

1.2 Ακουστικό στατιστικό μοντέλο

Το ακουστικό στατιστικό μοντέλο χρησιμοποιείται για να μοντελοποιήσει το ακουστικό σήμα. Θα πρέπει να είναι σε θέση να χειριστεί την μεταβλητότητα που υπάρχει στο σήμα της φωνής. Οι σημαντικότερες αιτίες στις οποίες οφείλεται η ύπαρξη της μεταβλητότητας είναι:

- συναφείς εκφράσεις (context variability),
- μεταβλητότητα στο ύφος (style variability),
- μεταβλητότητα στους ομιλητές (speakers variability), και
- μεταβλητότητα του περιβάλλοντος (environment variability).

Όλες οι παραπάνω αιτίες είναι σημαντικές και δημιουργούν προβλήματα, που το ακουστικό μοντέλο καλείται να επιλύσει. Το πρόβλημα των συναφών εκφράσεων προκύπτει από τις λέξεις οι οποίες είναι όμοιες ηχητικά, αλλά έχουν διαφορετική σημασία έχοντας κάποιο διαφορετικό

σημείο έμφασης - στίξης η μία από την άλλη. Η μεταβλητότητα στο ύφος των λέξεων οφείλεται στην ύπαρξη του συναισθήματος στην φωνή. Η μεταβλητότητα στους ομιλητές είναι προφανής λόγος αφού η διάκριση και η διαφορετικότητα της φωνής ανάμεσα στα φύλα, αλλά και λόγο της ηλικίας. Τέλος και ποιο σημαντικό, είναι η μεταβλητότητα του περιβάλλοντος που προσθέτει στοιχεία θορύβου στο σήμα της φωνής.

Ο σκοπός του ακουστικού στατιστικού μοντέλου είναι να παρέχει τη μέθοδο υπολογισμού της κατανομής πιθανότητας για την ακολουθία των διανυσμάτων της ακουστικής παρατήρησης, δοθέντος μιας λέξης. Στην πράξη η επιθυμητή κατανομή πιθανότητας καθορίζεται με την συμμετοχή πλήθους περιπτώσεων εμφάνισης της κάθε λέξης και συγκεντρώνοντας κατ΄ εξακολούθηση τα στατιστικά της στοιχεία. Στην περίπτωση που σχεδιάζουμε σύστημα αναγνώρισης φωνής με μεγάλο λεξικό, δεν είναι εφικτό να το διατηρήσουμε σε επίπεδο λέξεων. Τότε είναι απαραίτητο να αναλύσουμε της λέξεις σε πεπερασμένα μικρότερα τμήματα, σε επίπεδο φωνημάτων. Αν όμως έχουμε ως στόχο να σχεδιάσουμε το σύστημα αναγνώρισης για κάποιο συγκεκριμένο αριθμό λέξεων, τότε μπορούμε να υπολογίσουμε την κατανομή της πιθανότητας σε επίπεδο λέξεων.

Για να ορίσουμε το ακουστικό στατιστικό μοντέλο, παριστάνουμε το σήμα φωνής καθ΄ ενός μοντέλου, με μία ακολουθία διανυσμάτων με βάση τα γνωρίσματα (features) του $Y = [y_i, i = 1, ..., N]$, όπου y_i είναι μεταβλητού μήκους τυχαίο διάνυσμα και ισχύει $y_i = [y_{k_i}, ..., y_{k_i+N-1}]$.

Επιπλέον, το στατιστικό ακουστικό μοντέλο βασίζεται σε ακολουθία καταστάσεων και κατανομές εξόδου. Η ακολουθία καταστάσεων είναι η κρυφή παράμετρος του μοντέλου η οποία μοντελοποιεί την χρονική μεταβλητότητα του σήματος. Οι κατανομές εξόδου εκφράζουν την πιθανότητα μία παρατήρηση να έχει προκύψει από την συγκεκριμένη κατάσταση.

Ορισμός 1.2.1 *(Ακουστικού μοντέλου) Ακουστικό μοντέλο για ακολουθία γνωρισμάτων Y είναι η τετράδα (Q, B, Π, δ), όπου κάθε όρος ορίζεται ξεχωριστά ως εξής [1]:*

- *Q είναι το πεπερασμένο και διακριτό σύνολο καταστάσεων του ακουστικού μοντέλου με ακολουθία καταστάσεων $Q = [q_i, i = 1, ..., n]$, όπου κάθε q_i είναι τυχαία μεταβλητή παίρνοντας τιμές από το Q, συνεπάγοντας κατάτμηση όμοια με αυτή των παρατηρήσεων $Y = [y_i, i = 1, ..., n]$.*

- *$B = \{p_{y_t|y_1,...,y_{t-1},q_t}(\cdot|\cdot), q_t \epsilon Q\}$ είναι η συλλογή των μετρικών των πιθανοτήτων, και υποθέτουμε πως η κατανομή y_t δεδομένου των $q_i, i = 1, ..., n$ εξαρτάται μόνο από την εκάστοτε κατάσταση q_t,*

- *Π είναι ντετερμινιστική ή στοχαστική γραμματική η οποία περιγράφει το δυναμικό των καταστάσεων,*

- *δ είναι συνάρτηση αποκωδικοποίησης, που μας παρέχει την πιθανότερη ακολουθία καταστάσεων του μηνύματος, $\delta : Q^n \rightarrow M$*

1.3 Γλωσσικό μοντέλο

Άλλο ένα πρόβλημα της αναγνώρισης φωνής είναι η επιλογή του σωστού γλωσσικού μοντέλου και ο βέλτιστος υπολογισμός αυτού. Η γλώσσα υπόκειται σε κανόνες γραμματικής, σύνταξης και σημασιολογίας. Για να είναι μία φράση σωστή πρέπει να υπακούει σ' αυτούς τους κανόνες για να μπορέσουμε να εξάγουμε το νόημα της φράσης. Το γεγονός πως το μήκος της κάθε φράσης προς αναγνώριση είναι άγνωστο εκ των προτέρων, περιπλέκει σημαντικά την διαδικασία. Υπάρχουν δύο κύριοι μέθοδοι μοντελοποίησης της γλώσσας:

- η γλωσσική μοντελοποίηση κατά τον *Chomsky* [27], και

- η στοχαστική γλωσσική μοντελοποίηση.

Η γλωσσική μοντελοποίηση κατά τον *Chomsky* χρησιμοποιείται στον καθορισμό επιτρεπτών ακολουθιών λέξεων, με αποτέλεσμα να παράγει μόνο την ίδια ακολουθία λέξεων η

οποία έχει προκαθοριστεί. Κατά την στοχαστική γλωσσική μοντελοποίηση μπορούμε να υπολογίσουμε την πιθανότητα της λέξη να υπάρχει δοθέντος των λέξεων της πρότασης. Έστω μια ακολουθία λέξεων $W = \{w_1, ..., w_n\}$. Η πιθανότητα:

$$P(w) = \prod_{i=1}^{n} P(w_i|w_1, ..., w_n) \tag{1.1}$$

Ο πιο απλός, αποδοτικός και ευρέως διαδεδομένος τρόπος υπολογισμού του γλωσσικού μοντέλου είναι η χρήση των Μαρκοβιανών γραμματικών 3ης τάξης (trigram). Κατά τον οποίο ενδιαφε- ρόμαστε για της επόμενες δύο λέξεις που προηγούνται της επιθυμητής.

$$P(w) = \prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2}) \tag{1.2}$$

1.4 Αποκωδικοποίηση

Η αποκωδικοποίηση είναι το τελικό στάδιο στο σύστημα της αναγνώρισης της φωνής. Στο σημείο αυτό γίνεται ο συνδυασμός των υπολογισμών του ακουστικού και του γλωσσικού μοντέλου. Η αποκωδικοποίηση εγγυάται την εύρεση της πιθανότερης αναγνωρισμένης ακολουθίας. Ο ρόλος της αποκωδικοποίησης είναι να χειρίζεται τις προτάσεις που εμφανίζονται να έχουν μικρές πιθανότητες να αποκλείονται από το υπόλοιπο της επεξεργασίας. Ο σχεδιασμός αποδοτικών αποκωδικοποιητών είναι ο κρισιμότερος παράγοντας στη λειτουργία ενός πρακτικού συστήματος που έχει στόχο την ταχύτητα και την ακρίβεια του αναγνωριστή.

1.5 Σύνοψη

Στο κεφάλαιο αυτό περιγράψαμε τα διάφορα τμήματα που απαρτίζουν ένα σύγχρονο αυτόματο σύστημα Αναγνώρισης Φωνής, ώστε η εφαρμογή τους να ανταποκρίνεται αποτελεσματικά σε μεγάλο αριθμό από διαφορετικούς χρήστες και για μεγάλο επιτρεπτό λεξιλόγιο. Η επεξεργασία του ακουστικού σήματος είναι απαραίτητη για την εξαγωγή των χαρακτηριστικών του σήματος της φωνής, σε μορφή δεδομένων για επεξεργασία. Το γλωσσικό μοντέλο να καθορίζει το επιτρεπτό λεξιλόγιο. Το ακουστικό μοντέλο προσπαθεί να εντοπίσει τα πιθανά λεγόμενα και τέλος ο αποκωδικοποιητής να συνδυάζει γλωσσικό μοντέλο και ακουστικό μοντέλο, δίνοντας

το αποτέλεσμα αναγνώρισης. Το ακουστικό μοντέλο είναι αυτό που καθορίζει σε μεγάλο βαθμό την απόδοση, ενός συστήματος αναγνώρισης. Για τον λόγο αυτό υπάρχουν διάφοροι μέθοδοι υλοποίησης του ακουστικού μοντέλου, οι οποίοι θα αναλυθούν στο επόμενο κεφάλαιο.

Κεφάλαιο 2

Στατιστικά Ακουστικά Μοντέλα

Θεωρούμε πως μία άγνωστη κυματομορφή σήματος φωνής παραμετροποιείται μέσω μιας διαδικασίας επεξεργασίας βραχέως χρόνου, παράγοντας μια ακολουθία διανυσμάτων, έστω $Y = y_1, y_2, ..., y_n$ οι ακουστικές παρατηρήσεις. Θεωρούμε επίσης ότι το σήμα αυτό της φωνής εμπεριέχει κάποιο λεξιλόγιο που υπόκειται σε κανόνες γραμματικής, σύνταξης και σημασιολογίας. Έστω $W = w_1, w_2, ..., w_n$ η ακολουθία των λέξεων. Για να υπολογίσουμε την πιο πιθανή ακολουθία λέξεων, γίνεται στο τμήμα του γλωσσικού μοντέλου, αρκεί να εκτιμηθεί η ποσότητα $P(w_i|w_{i-1}w_{i-2}...w_{i-n})$ για τις w_i λέξεις. Ο κλάδος της αναγνώρισης φωνής ενδιαφέρεται να συνδυάσει τον παραπάνω υπολογισμό δοθέντων των ακουστικών παρατηρήσεων και να επιλέξει στατιστικά την βέλτιστη ακολουθία. Το πιο διαδεδομένο κριτήριο είναι του υπολογισμού της Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimation), το οποίο θα αναλυθεί εκτενέστερα στο Κεφάλαιο 3. Η μέγιστη πιθανότητα της ακολουθίας των λέξεων δοθέντος των παρατηρήσεων είναι:

$$\hat{W} = \arg_{\hat{W}} \max P(W|O) \tag{2.1}$$

Η ιδέα του κριτήριου αυτού είναι πως κατορθώνουμε μέγιστη πιθανότητα, όταν το μοντέλο έχει την ίδια κατανομή με αυτήν των παρατηρήσεων.
Εφαρμόζοντας στη συνέχεια τον κανόνα του *Bayes* η εξίσωση παίρνει την μορφή,

$$\hat{W} = \arg_{\hat{W}} \max P(W|O) = \arg_{\phi} \max \frac{P(W)P(O|W)}{P(O)} \Leftrightarrow$$

$$\hat{W} = \arg_{\hat{W}} \max P(W)P(O|W) \tag{2.2}$$

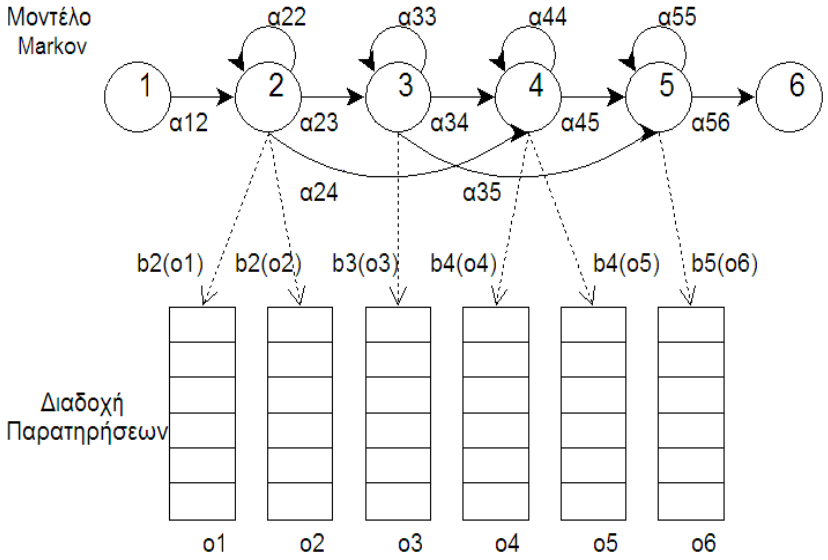
Για τον υπολογισμό του όρου $P(W)$ χρησιμοποιείται το γλωσσικό μοντέλο, δίνοντας την a-priori πιθανότητα της ακολουθίας ανεξάρτητα απο το σήμα που παρατηρήθηκε. Ο όρος $P(O|W)$ εκτιμάται από το ακουστικό μοντέλο.

Στη συνέχεια του κεφαλαίου θα αναλυθούν οι μέχρι τώρα μέθοδοι υλοποίησης του ακουστικού μοντέλου, στο εδάφιο (2.1) περιγράφεται η μέθοδος της ακουστικής μοντελοποίησης των Κρυφών Μαρκοβιανών Μοντέλων (HMM), στο εδάφιο (2.2) τα Βασισμένα στην Τμηματοποίηση Μοντέλα (Segment - Based Models), τέλος στο εδάφιο (2.2) την μέθοδο των Γραμμικών Δυναμικών Μοντέλων.

2.1 Κρυφά Μαρκοβιανά Μοντέλα (HMM)

Τα HMM είναι ως σήμερα η πιο φημισμένη και αποδοτική προσέγγιση στην ακουστική μοντελοποίηση, αφού μοντελοποιούν αρκετά καλά τα σήματα της φωνής και την χρονικά μεταβαλλόμενη φύση της. Τα μοντέλα αυτά επιτρέπουν μεταβάσεις μόνο από αριστερά προς τα δεξιά και μόνο για μία ή δύο καταστάσεις, με αποτέλεσμα να επιβάλει περιορισμό, κατά την εξαγωγή των ακουστικών παρατηρήσεων σε frame-based παρατηρήσεις. Επίσης το μοντέλο δεν μπορεί να επιστρέψει σε προηγούμενη κατάσταση (βλέπε Σχημ 2.1).

Θα προσπαθήσουμε να ορίσουμε τα HMM σύμφωνα με τον γενικότερο ορισμό του ακουστικού μοντέλου που δόθηκε στο 1ο Κεφάλαιο. Ένα HMM είναι συνδυασμός δύο στοχαστικών διαδικασιών, μιας κρυφής αλυσίδας $Markov$ Q που διατηρεί την στατιστική πληροφορία της χρονικής μεταβλητότητας και μιας φανερής Y η οποία περιγράφει την φασματική μεταβλητότητα. Οι κατανομές εξόδου B μπορούν να είναι είτε συνεχείς είτε διακριτές, και εξαρτώνται από την κατάστασης που βρισκόμαστε την δεδομένη χρονική στιγμή. Στις κατανομές εξόδου υποθέτουμε πως τα παρατηρήσιμα διανύσματα είναι υπο-συνθήκη ανεξάρτητα δεδομένου της επικρατούσας ακολουθίας. Η γραμματική Π είναι είτε στοχαστική είτε ντετερμινιστική, περιέχει τις πιθανότητες μετάβασης και τις αρχικές πιθανότητες του μοντέλου. Τελευταία παράμετρος για να ολοκληρωθεί ο ορισμός ενός HMM ως ακουστικό μοντέλο είναι η συνάρτηση αποκωδικοποίησης δ. Η συνάρτηση αυτή ποικίλει, και εξαρτάται από το είδος της αναγνώρισης που ενδιαφερόμαστε να πραγματοποιήσουμε.



Σχήμα 2.1: *HMM* 5 καταστασεων.

Ορισμός 2.1.1 *(HMM μοντέλου) Η προσαρμογή του Κρυφού Μαρκοβιανού μοντέλου ως ακουστικό μοντέλο πραγματώνεται με τις παραμέτρους [1]:*

- *Η επικρατούσα ακολουθία Q μπορεί να είναι είτε διακριτή είτε συνεχής,*
- *Οι παρατηρήσεις θεωρούνται υπο-συνθήκη ανεξάρτητες δοθέντος επικρατούσας ακολουθίας $B = \{p_{y_t|y_1,\ldots,y_{t-1},q_t}(\cdot|\cdot), q_t \epsilon Q\} = \{p_{y_t|q_t}(\cdot|\cdot), q_t \epsilon Q\}$.*
- *Η ακολουθία καταστάσεων μοντελοποιείται ως πρώτης τάξης αλυσίδα Markou, έτσι η γραμματική Π περιέχει τις πιθανότητες μετάβασης και τις πιθανότητες αρχικοποίησης.*
- *η συνάρτηση αποκωδικοποίησης δ, εξαρτάται από την ιδιαίτερη κάθε φορά, τοπολογία των HMM.*

Από τον ορισμό του HMM ως ακουστικού μοντέλου προκύπτουν κάποια μειονεκτήματα στην ακρίβεια της υλοποίησή τους. Πιο συγκεκριμένα αυτοί είναι:

- δεν παρέχουν ικανοποιητική αναπαράσταση της χρονικής διάρθρωσης της φωνής
- η πιθανότητα μετάβασης εξαρτάται μόνο από την αρχή και τον προορισμό, και
- η υπο-συνθήκη ανεξαρτησία των παρατηρήσεων

Αναλυτικότερα κατά την αναπαράσταση της χρονικής διάθρωσης, η πιθανότητα της καταστάσεως (*state*) ελαττώνεται εκθετικά με τον χρόνο. Αν δούμε το σχήμα 2.1, η πιθανότητα από t συνεχόμενες παρατηρήσεις, για μια απο τις καταστάσεις i είναι η πιθανότητα να πάρουμε την ίδια κατάσταση i για t φορές, μαθηματικά εκφράζεται:

$$d_i(t) = a_{ii}^t(1 - a_{ii}) \tag{2.3}$$

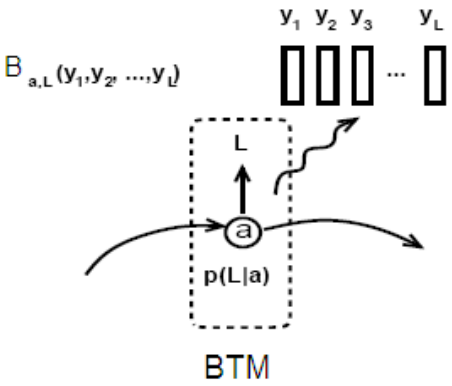
όπου φαίνεται η εκθετική ελάττωση [16]. Όπου a_{ii} πιθανότητα μετάβασης, δηλαδή $0 < a_{ii} \leq 1$.

Από την στιγμή που η ακολουθία καταστάσεων στα HMM΄ς είναι μια αλυσίδα *Markov*, είναι πιθανό να έχουμε κάποια τμήματα παρατεταμένα και κάποια συμπιεσμένα, κάτι που είναι ανακόλουθο μ΄ αυτά που γνωρίζουμε για το σήμα της φωνής. Η υπο-συνθήκη ανεξαρτησία είναι άλλος ένας σημαντικός περιορισμός, κατά τον οποίο τα παρατηρήσιμα διαστήματα εξαρτώνται μόνο από την κατάσταση που τα δημιούργησε και όχι από τα γειτονικά παρατηρήσιμα διαστήματα. Η υπόθεση αυτή προσθέτει δυσκολίες στον αποτελεσματικό χειρισμό των μη-στάσιμων διαστημάτων, τα οποία είναι έντονα συσχετισμένα.

Τα μειονεκτήματα αυτά στάθηκαν αφορμή για τη εισήγηση διαφορετικών υλοποιήσεων με βάση των HMM. Αυτές όμως δεν μπόρεσαν να επιλύσουν τα μειονεκτήματα αυτά. Για την επίλυση κάποιων μειονεκτημάτων των HMM, εισήχθηκε μια νέα μέθοδος ακουστικής μοντελοποίησης, τα βασισμένα στην τμηματοποίηση μοντέλα (Segment-based models).

2.2 Βασισμένα στην Τμηματοποίηση Μοντέλα

Ένα ακουστικό μοντέλο θα πρέπει να μπορεί να αντιμετωπίσει υψηλού βαθμού φαινόμενα και να αξιοποιεί τα χαρακτηριστικά που θα εξάγονται από μεγαλύτερα παράθυρα επεξεργασίας του σήματος της φωνής. Τα Βασισμένα στην Τμηματοποίηση Μοντέλα (ΒΤΜ) βασίζονται στην διαίρεση ενός σήματος της φωνής σε τμήματα, χωρίς να υπάρχουν κενά μεταξύ των



Σχήμα 2.2: Σχηματική περιγραφή ενός Βασισμένου στην Τμηματοποίηση Μοντέλου [7].

τμημάτων. Αν έχουμε σήμα φωνής, με N αριθμό διαστημάτων θεωρούμε ότι από το διάστημα 1 έως το διάστημα p, με $p-1>0$, είναι το l_1 τμήμα, από $p+1$ έως r με $r-p+1>0$ είναι το l_2 τμήμα, κλπ. Το μήκος l των τμημάτων δεν επιλέγεται τυχαία αλλά καθορίζεται από μια συλλογή με σταθερές τιμές μήκους τμήματος. Έστω L η συλλογή των αποδεκτών μηκών που κάθε φορά είναι ίσο με το άθροισμα όλων των επιμέρους μηκών.

$$L = l_1 + l_2 + ... + l_N \tag{2.4}$$

Για κάθε μήκος θα πρέπει να ορίζουμε επίσης ποία γλωσσική μονάδα μοντελοποιεί και με βάση αυτό γίνεται η κατάτμηση του διανύσματος των παρατηρήσεων Y. Θεωρούμε ότι έχουμε μια άγνωστη κυματομορφή σήματος φωνής την οποία χωρίζουμε σε ακολουθία τμημάτων, έστω $A = [a_1, a_2, ...a_n]$, τα μήκη των τμημάτων έχουν διάρκεια που καθορίζεται από το σύνολο L. Η πιθανότητα της επικρατούσας ακολουθίας μπορεί να γραφεί:

$$P(Q) = P(L, A) = P(L|A)P(A) = \prod_{i=1}^{n} p(l_i|a_i)p(a_i|a_1, ..., a_{i-1}) \tag{2.5}$$

από ΄που φαίνεται πως η διάρκεια ενός τμήματος είναι υπο-συνθήκη ανεξάρτητη δοθέντος η ακολουθία του μοντέλου. Στα βασισμένα στην τμηματοποίηση μοντέλα παρατηρούμε πως δεν υπάρχει περιορισμός στο τύπο της διάρκειας των κατανομών, όπως συμβαίνει στη περίπτωση των HMM. Ανακεφαλαιώνοντας ορίζουμε το BTM ως ακουστικό μοντέλο.

Ορισμός 2.2.1 *(Βασισμένα στην Τμηματοποίηση Μοντέλα) Τα βασισμένα στην τμηματοποίηση μοντέλα παραμετροποίουν το ακουστικό μοντέλο [1]:*

- *Η επικρατούσα ακολουθία $Q = (a, l) \epsilon A \times L$, όπου A είναι το σύνολο των γλωσσικών μονάδων και L η συλλογή των διαρκειών των τμημάτων.*
- *Η συλλογή των υπολογισμών των πιθανοτήτων των παρατηρήσεων για ένα τμήμα είναι $p = (Y(p+1, r)|q_i)$ όπου $p+1$ αρχή του τμήματος και r το τέλος του.*
- *Η γραμματική έχει τον περιορισμό (2.4) και τα στοχαστικά συστατικά της εξίσωσης (2.5).*
- *Η συνάρτηση αποκωδικοποίησης καθορίζεται από $\delta(Q) = \delta'(A, L)$.*

Το γεγονός πως τα Βασισμένα στην Τμηματοποίηση Μοντέλα χαρακτηρίζονται απο τμήματα που η διάρκεια τους είναι προκαθορισμένη από τη συλλογή L, αναγκάζει την ακολουθία των παρακολουθήσεων να εξομαλύνεται με την χρονική διάρκεια τους, πριν την εφαρμογή του ακουστικού μοντέλου. Αυτό έχει ως αποτέλεσμα να περιπλέκει την διαδικασία αποκωδικοποιησης, αφού οι υποψήφιες υποθετικές περιπτώσεις τμηματοποίησης για τον υπολογισμό των πιθανοτήτων είναι αρκετές για όλες τις περιπτώσεις των διαφορετικών μηκών. Το πρόβλημα αυτό ώθησε στην εισαγωγή των Στοχαστικών Τμηματοποιημένων Μοντέλων.

Στοχαστικά Τμηματοποιημένα Μοντέλα

Τα Στοχαστικά Τμηματοποιημένα Μοντέλα διαφέρουν ως πρός τα Βασισμένα στην Τμηματοποίηση Μοντέλα μόνο στο καθορισμό της διάρκειας ενός τμήματος l, τη θεωρούν ως τυχαία μεταβλητή που ακολουθεί *Gaussian* κατανομή. Έστω ένα τμήμα παρατηρήσεων διάρκειας l, $y = \{y_1, ..., y_l\}$, ενός μοντελου m. Το μοντέλο m παράγει τμήματα παρατηρήσεων από την εξίσωση:

$$p(y, l|m) = p(y|l, m)p(l|m) \tag{2.6}$$

όπου το ακουστικό μοντέλο υπολογίζει τον όρο $p(y|l, m)$ ενώ το μοντέλο τμηματοποίησης m τον $p(l|m)$.

Τα Βασισμένα στην Τμηματοποίηση Μοντέλα και ιδιαίτερα τα Στοχαστικά Τμηματοποιημένα Μοντέλα πρότειναν τον τρόπο να χειριστούμε τη μεταβλητότητα της διάρκειας των μοντέλων. Σύμφωνα με την έρευνα όμως του Βασιλείου Διγαλάκη [1], δεν έλυσαν το πρόβλημα των εξαρτήσεων μεταξύ των χαρακτηριστικών συντελεστών της φωνής, που υπήρχε και κατά την υλοποίηση των HMM. Το πρόβλημα μετατέθηκε τόσο στα εσωτερικά πλαίσια των τμημάτων όσο στα σύνορα μεταξύ των πλαισίων των τμημάτων. Η νέα πρόταση ήταν η εισαγωγή των Γραμμικών Δυναμικών Μοντέλων στην αναγνώριση της φωνής.

Γραμμικά Δυναμικά Μοντέλα

Τα Γραμμικά Δυναμικά Μοντέλα, που είναι και το αντικείμενο της εργασίας αυτής, βασίζονται στα Βασισμένα στην Τμηματοποίηση Μοντέλα, έχουν δηλαδή την δυνατότητα να διαιρούν τις παρατηρήσεις σε τμήματα, τόσο σε σταθερά μήκη όσο και σε μεταβαλλόμενα, ορίζονται όπως και τα Βασισμένα στην Τμηματοποίηση Μοντέλα, αλλά διαφοροποιούνται στον τρόπο υλοποίησης, έχοντας το πλεονέκτημα να διατηρούν τις εξαρτήσεις στα εσωτερικά πλαίσια ενός τμήματος.

Τα Γραμμικά Δυναμικά Μοντέλα εισαγάγανε δύο μεθόδους υλοποίησης:

- της αμεταβλητότητας της τροχιάς και
- της αμεταβλητότητας της συσχέτισης.

Κατά τον τρόπο της αμεταβλητότητας της τροχιάς υποθέτουμε πως υπάρχει μοναδική, σταθερή τροχιά για κάθε βασικό φωνητικό τμήμα (ως βασικό φωνητικό τμήμα δηλώνουμε το φώνημα). Ο τρόπος της αμεταβλητότητας της συσχέτισης υποθέτει πως εντός των τμημάτων υπάρχουν περιοχές στις οποίες η συσχέτιση μεταξύ των πλαισίων είναι στατική. Σύμφωνα με την εργασία [2], η ταξινόμηση των δύο μεθόδων έδειξε παρόμοια συμπεριφορά για μεγάλου μήκους βασικά φωνητικά τμήματα, όμως μικρή επίδοση προέκυψε κατά την εφαρμογή της αμεταβλητότητας της τροχιάς όταν το μήκος της ακολουθίας των παρατηρήσεων ήταν μικρότερη από την αρχική υπόθεση. Το πρόβλημα αποδόθηκε στο γεγονός ότι η συσχέτιση σε περίπτωση σύντομου τμήματος ελαττώνεται κατά την κανονικοποίηση της διάρκειας των τμημάτων. Έτσι μεταξύ

των δύο, υιοθετούμε την μέθοδο της αμεταβλητότητας της συσχέτισης και για την υλοποίηση του μοντέλου μας.

2.3 Σύνοψη

Είδαμε στο Κεφάλαιο αυτό τις διάφορες μεθόδους προσέγγισης του ακουστικού μοντέλου, παρουσιάζοντας τα μειονεκτήματα και πλεονεκτήματα, της κάθε μεθόδου. Τα στατιστικά μοντέλα, στην ακουστική μοντελοποίηση είναι το μεγαλύτερο αντικείμενο έρευνας στην Αναγνώριση Φωνής. Τα HMM ως παλαιότερη μέθοδος έχει αναλυθεί εκτενώς, όντας η πιο διαδεδομένη μέθοδος. Τα Βασισμένα στην Τμηματοποίηση μοντέλα και ιδιαίτερα με τη μέθοδο των γραμμικών συστημάτων είναι καινούργια προσέγγιση, έτσι τυγχάνει μεγαλύτερης προσοχής και έρευνας. Πριν όμως δούμε την εφαρμογή τους, θεωρείται απαραίτητο να αναλύσουμε την θεωρία που περικλείεται γύρω απ΄ αυτά. Στο επόμενο κεφάλαιο θα δοθεί η θεωρία υπό ένα γενικότερο βλέμμα.

Κεφάλαιο 3

Γραμμικά Δυναμικά Μοντέλα

Το γραμμικό δυναμικό μοντέλο είναι υποσχόμενη μέθοδος υλοποίησης του ακουστικού μοντέλου. Η θεωρία για να καταλήξουμε στην εφαρμογή του, ως ακουστικό μοντέλο είναι εκτενής, αφού συνδυάζει θέματα μελέτης διαφορετικών επιστημονικών πεδίων, με πολλές έννοιες, μεθόδους και αλγορίθμους. Η προσπάθειά μας είναι να δείξουμε, όσο πιο απλα γίνεται, πως συνδυάζονται υπό μια ενιαία οπτική. Θα δείξουμε πως από το γενικό γραμμικό σύστημα μέσω των ιδιοτήτων του μπορεί να πάρει μορφή ικανή για την εφαρμογή του σε δεδομένα φωνής και συνδυάζοντας τους υπολογισμούς του, με μεθόδους πιθανοτικής συλλογιστικής και να εξάγουμε ασφαλή συμπεράσματα.

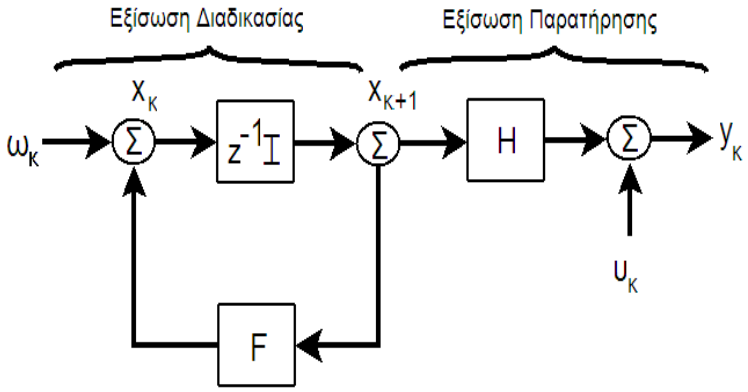
Το κεφάλαιο οργανώνεται ως ακολούθως: στο εδάφιο (3.1) περιγράφεται το γραμμικό δυναμικό σύστημα, στο εδάφιο (3.2) εξηγείται η έννοια του κοντρόλ για το γραμμικό δυναμικό σύστημα, στο εδάφιο (3.3) αναλύεται ο τρόπος εφαρμογής του γραμμικού δυναμικού συστήματος στην μοντελοποίηση.

3.1 Γραμμικό δυναμικό σύστημα

Το γενικό γραμμικό δυναμικό σύστημα διακριτού χρόνου, *state-space* πολλών μεταβλητών, με επιπλέον εισόδους θορύβου σε είσοδο και έξοδο περιγράφεται από τις εξισώσεις (3.1) και (3.2), ενώ σχηματικά φαίνεται στο Σχημ. 3.1:

$$x_{k+1} = Fx_k + \omega_k \qquad \omega_k \sim N(q, Q) \tag{3.1}$$

$$y_k = Hx_k + \upsilon_k \qquad \upsilon_k \sim N(r, R) \tag{3.2}$$



Σχήμα 3.1: *Μοντέλο παραγόμενο από γραμμικό δυναμικό σύστημα*

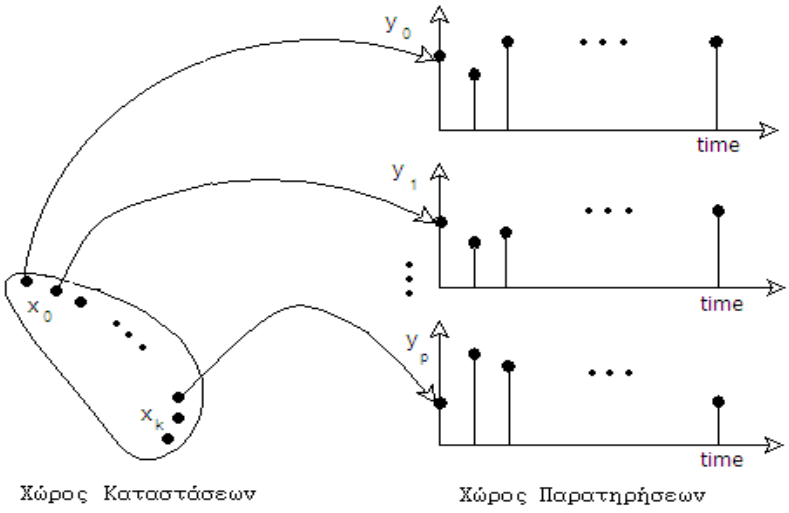
Στο παραπάνω σύστημα με x_k συμβολίζεται το διάνυσμα κατάστασης (*state*), διάστασης $(n\times 1)$, το οποίο είναι έμμεσα παρατηρήσιμο, με y_k το διάνυσμα παρατήρησης διάστασης $(p\times 1)$ που μας παρέχει την άμεσα παρατηρήσιμη έξοδο του συστήματος. Ο πίνακας F είναι ο πίνακας μετάβασης καταστάσεων, διάστασης $(n\times n)$, και ο H είναι ο πίνακας παρατήρησης, διάστασης $(p\times p)$. Οι όροι ω_k και υ_k είναι επιπλέον ασυσχέτιστοι θόρυβοι με μέσους q και r και με συνδιακύμανση Q και R αντίστοιχα, για τους οποίους ισχύει:

$$E\{w_n w_k^T\} = \begin{cases} Q & \text{αν } n = k \\ 0 & \text{αν } n \neq k \end{cases}$$

και

$$E\{\upsilon_n \upsilon_k^T\} = \begin{cases} R & \text{αν } n = k \\ 0 & \text{αν } n \neq k \end{cases}$$

Το διάνυσμα x_k σε *state space* συστήματα καλείται διάνυσμα κατάστασης, επειδή δίνει την πλήρη περιγραφή για την κατάσταση του συστήματος την χρονική στιγμή k, και σε κάθε περίπτωση καθορίζει την εξέλιξη του συστήματος στο μέλλον. Ακόμη το διάνυσμα κατάστασης έχει την ιδιότητα να συλλέγει πληροφορίες για την αρχικοποίηση του συστήματος. Έχοντας ως γνώση τις συνθήκες αυτές σε μια δεδομένη χρονική στιγμή μαζί με τον προσδιορισμό των μελλοντικών εισόδων, είναι ότι χρειαζόμαστε ώστε να μπορέσουμε να καθορίσουμε την μελλοντική συμπεριφορά του συστήματος. Όταν κάποιος αναφέρεται σε *state space* συστήματα μεγέθους χώρου n, καθορίζει έτσι το πεδίο τιμών του διανύσματος κατάστασης (βλέπε Σχήμ. 3.2). Επομένως κάποιος μπορεί να φανταστεί την εφαρμογή του γραμμικού δυναμικού



Σχήμα 3.2: *Οι* παρατηρήσεις y_p συσχετίζονται με τις καταστάσεις x_k μέσω μιας συνάρτησης μετατροπής

συστήματος στην περιγραφή ενός φαινομένου, παρατηρώντας μόνο το αποτέλεσμα του, δηλαδή την έξοδο του, μοντελοποιώντας την συμπεριφορά του δίνοντας κατάλληλες τιμές στα δομικά του στοιχεία. Τα δομικά του στοιχεία τα συμβολίζουμε με το σύνολο μεταβλητών

$$\Theta = \{F, H, Q, R, \mu_0, \Sigma_0\}$$

και τα ονομάζουμε παραμέτρους του μοντέλου, όπου μ_0, Σ_0 είναι η αρχικοποίηση του μοντέλου.

3.2 Έλεγχος, Παρακολούθηση, Κανονικοποίηση

Η *state space* περιγραφή του γραμμικού δυναμικού συστήματος επικεντρώνεται στην συμπεριφορά του συστήματος, δηλαδή στην εξέλιξη των καταστάσεων. Από την οπτική του κοντρόλ, η δομή των εισόδων και των εξόδων είναι ολοκληρωμένος τρόπος περιγραφής του δυναμικού συστήματος. Η δομή της εισόδου καθορίζει τον βαθμό που μπορεί να διαφοροποιηθεί η συμπεριφορά του συστήματος, ενώ η δομή της εξόδου εμπεριέχει το είδος της πληροφορίας ικανό για τον έλεγχο του συστήματος. Τα δύο αυτά στοιχεία αλληλεπιδρώντας δίνουν τη βάση για την επίτευξη του κοντρόλ του συστήματος.

Έλεγχος

Για τον έλεγχο του συστήματος αρκεί να παρατηρήσουμε την είσοδο του συστήματος. Αν στην εξίσωση (3.1) θέσουμε την πρόσθετη είσοδο ολοκληρωμένα, δηλαδή όπου $w_k = Qu_k$ προκύπτει.

$$x_{k+1} = Fx_k + Qu_k \tag{3.3}$$

με x_k διάνυσμα κατάστασης, διάστασης $(n \times 1)$, ο πίνακας F είναι ο πίνακας καταστάσεων μετάβασης, διάστασης $(n \times n)$ και όπου $w_k = Qu_k$ με πίνακα κατανομής Q, διάστασης $(n \times m)$ και u_k είσοδος, διάστασης $(m \times 1)$.

Θεώρημα 3.2.1 *Ένα διακριτό σύστημα είναι πλήρως ελεγχόμενο αν και μόνο αν ο πίνακας ελέγχου M, διάστασης $(n \times nm)$*

$$M = [Q, FQ, ..., F^{n-1}Q] \tag{3.4}$$

έχει rank n [10].

Από το θεώρημα προκύπτει πως αν ένα σύστημα είναι πλήρως ελεγχόμενο, μπορούμε να δώσουμε κατάλληλη είσοδο με την οποία από μια κατάσταση σε αυθαίρετη θέση, θα μεταβεί σε μια άλλη κατάσταση επίσης αυθαίρετης θέσης, μετά από πεπερασμένο αριθμό βημάτων. Δηλαδή θεωρούμε τις καταστάσεις x_a και x_b, σε αυθαίρετες θέσεις a και b αντίστοιχα, με $b - a = n > 0$, με την εισοδο u_k μηδέν το σύστημα θα μεταβεί στην κατάσταση $F^n x_a$ σε n βήματα, ή αλλιώς με επιθυμητή είσοδος το σύστημα θα μεταφερθεί απο την αρχική κατάσταση x_a στην κατάσταση $x_b - F^n x_a$ σε n βήματα.

Παρακολούθηση

Για την παρακολούθηση του συστήματος αρκεί να παρατηρήσουμε την έξοδο του συστήματος. Από τις εξισώσεις (3.1),(3.2) θεωρούμε τις εισόδους ω_k και υ_k να είναι ίσοι με το μηδέν. Τότε έχουμε τις εξισώσεις

$$x_{k+1} = Fx_k \tag{3.5}$$

$$y_k = Hx_k \tag{3.6}$$

με x_k συμβολίζεται το διάνυσμα κατάστασης ($state$), διάστασης $(n \times 1)$, με y_k το διάνυσμα παρατήρησης διάστασης $(p \times 1)$. Ο πίνακας F είναι ο πίνακας καταστάσεων μετάβασης, διάστασης $(n \times n)$,και ο H είναι ο πίνακας παρατηρήσεων, διάστασης $(p \times p)$.

Θεώρημα 3.2.1 *Ένα διακριτό σύστημα είναι πλήρως παρατηρήσιμο αν και μόνο αν ο πίνακας παρακολούθησης S, διάστασης $(pn \times n)$*

$$S = [H, HF, ..., HF^{n-1}]^T \tag{3.7}$$

έχει rank n [10].

Η έννοια της παρακολούθησης είναι πολύ σημαντική στην οπτική του κοντρόλ. Αν η πληροφορία της εξόδου είναι ελλιπής τότε θα έχουμε ελλιπή πληροφόρηση και για την κατάσταση του συστήματος, κάτι που περιορίζει την σχεδίαση του κοντρόλ.

Κανονικοποίηση

Ο έλεγχος και η παρακολούθηση μας παρέχουν την σχέση μεταξύ της εισόδου και της εξόδου ώστε να γνωρίζουμε αυτό που συμβαίνει εσωτερικά του συστήματος. Αυτό μπορεί να επιτευχθεί μετατρέποντας το διάνυσμα κατάστασης με μερικές πιθανές μορφές που λέγονται κανονικές. Οι μορφές αυτές μετατρέπουν και τους πίνακες του συστήματος, δίνοντας τους συγκεκριμένη μορφή. Ο λόγος που κάνει την κανονικοποίηση απαραίτητη είναι επειδή τα στοιχεία του συστήματος υπολογίζονται με μεγαλύτερη ευκολία.

3.3 Εκτίμηση Παραμέτρων

Για την εκτίμηση των παραμέτρων του γραμμικού δυναμικού μοντέλου, θεωρούμε γραμμικό δυναμικό σύστημα με $Gaussian$ θορύβους ως είσόδο. Το γραμμικό δυναμικό μοντέλο περιγράφεται από τις εξισώσεις:

$$x_{k+1} = Fx_k + \omega_k \qquad \omega_k \sim N(0, Q) \tag{3.8}$$

$$y_k = Hx_k + \upsilon_k \qquad \upsilon_k \sim N(0, R) \tag{3.9}$$

Για την κατάσταση υποθέτουμε ότι εξελίσσεται σαν πρώτης τάξης αλυσίδα $Markov$, η έξοδος ειναι απλή γραμμική, αποτέλεσμα της εκάστοτε κατάστασης. Και στις δύο εφαρμόζεται πρόσθετος $Gaussian$ θόρυβος, ω_k και υ_k αντίστοιχα, με μηδενική μέση τιμή. Για τους οποίους ισχύει:

$$E\{w_n w_k^T\} = \begin{cases} Q & \text{αν } n = k \\ 0 & \text{αν } n \neq k \end{cases}$$

και

$$E\{\upsilon_n \upsilon_k^T\} = \begin{cases} R & \text{αν } n = k \\ 0 & \text{αν } n \neq k \end{cases}$$

Οι θόρυβοι είναι ανεξάρτητοι μεταξύ τους, επομένως $E[\omega_k \upsilon_k] = 0$, και αλλάζουν τιμή σε κάθε χρονική στιγμή, αλλά για την περίπτωσή μας τους θεωρούμε σταθερούς. Η γενική ιδέα του γραμμικού συστήματος ως βασικό μοντέλο είναι πως η ακολουθία της κατάστασης θα πρέπει να είναι ερμηνεία στην πολυπλοκότητα της ακολουθίας εξόδου και ότι θα είναι πιο περιεκτική σε πληροφορία από την ίδια την έξοδο.

Υπολογισμοί

Έστω οτι μας δίνονται οι παράμετροι του μοντέλου των εξισώσεων (3.8) και (3.9), $\Theta = \{F, H, Q, R, \mu_0, \Sigma_0\}$, με μια ακολουθία διανυσμάτων από παρατηρήσεις $Y = \{y_1, ..., y_n\}$. Η βασική ποσότητα που θέλουμε να υπολογίσουμε είναι η ολική πιθανότητα της ακολουθίας των παρατηρήσεων με βάση τις παραμέτρους του μοντέλου, την πιθανή ακολουθία καταστάσεων $x_k = \{x_1, ..., x_n\}$. Η ολική πιθανότητα υπολογίζεται από την εξίσωση:

$$P(\{y_1, ..., y_n\}|\Theta) = \int_{all\{x_1,...,x_n\}} P(\{x_1, ..., x_n\}, \{y_1, ..., y_n\}|\Theta) d\{x_1, ..., x_n\} \tag{3.10}$$

Ο υπολογισμός της συνδυασμένης πιθανότητας γίνεται μέσω της εξίσωσης.

$$P(\{x_1, ..., x_n\}, \{y_1, ..., y_n\}|\Theta) = P(x_0) \prod_{k=1}^{n-1} P(x_{k+1}|x_k) \prod_{k=1}^{n-1} P(y_k|x_k) \tag{3.11}$$

Όπου x_0 η αρχικοποίηση του συστήματος. Εφαρμόζοντας τον κανόνα του $Bayes$ υπολογίζουμε την υπο-συνθήκη κατανομή της πιθανότητας.

$$P(\{x_1, ..., x_k\}|\{y_1, ..., y_k\}, \Theta) = \frac{P(\{x_1, ..., x_k\}, \{y_1, ..., y_k\}|\Theta)}{P(\{y_1, ..., y_k\}|\Theta)} \tag{3.12}$$

Συνήθως η κατανομή της πιθανότητας της κατάστασης είναι χρήσιμη για κάποια συγκεκριμένη χρονική στιγμή, δηλαδή συνήθως χρειάζεται ο υπολογισμός των πιθανοτήτων $P(x_k|\{y_1, ..., y_n\})$. Ο υπολογισμός των πιθανοτήτων αυτών, επιτυγχάνεται με την εφαρμογή του αλγόριθμου Εμπρός - Πισω ($Forward - Backward\ Algorithm$)

Εμπρός - Πισω Αλγόριθμος

Ο αλγόριθμος Εμπρός - Πίσω είναι μία επαναληπτική μέθοδος για τον υπολογισμό της πιθανότητας να βρισκόμαστε σε μια συγκεκριμένη κατάσταση, σε μια συγκεκριμένη χρονική στιγμή. Το όνομα του αλγορίθμου είναι αντιπροσωπευτικό για τον τρόπο που λειτουργεί στον υπολογισμό των πιθανότητων $P(x_k|\{y_1, ..., y_n\})$. Η γενική μορφή του αλγορίθμου έχει ως εξής:
Αν συμβολίσουμε την συνδυασμένη πιθανότητα της κατάστασης x_k με την ακολουθία παρατηρήσεων $\{y_1, ..., y_k\}$ ως:

$$a_k(x_k) = P(x_k, \{y_1, ..., y_k\}) \tag{3.13}$$

Ο επαναληπτικός υπολογισμός για το Εμπρός βήμα του αλγορίθμου έχει ως εξής:

$$a_k(x_k) = \sum_{x_{k-1}} P(x_k, x_{k-1}, \{y_1, ..., y_k\}) \Leftrightarrow$$

$$a_k(x_k) = \sum_{x_{k-1}} P(x_{k-1}, \{y_1 ... y_{k-1}\}) P(x_k|x_{k-1}) P(y_k|x_k, \{y_1, ..., y_{k-1}\}) \Leftrightarrow$$

$$a_k(x_k) = \Big[\sum_{x_{k-1}} a_{k-1}(x_{k-1}) P(x_k|x_{k-1}) \Big] P(y_k|x_k, \{y_1, ..., y_{k-1}\}) \tag{3.14}$$

Στην δεύτερη εξίσωση χρησιμοποιήθηκε η ταυτότητα

$$P(x_k|x_{k-1}, \{y_1, ..., y_{k-1}\}) = P(x_k|x_{k-1})$$

που μπορεί να εξαχθεί από την ιδιότητα του Markov της *mode* ακολουθία και απο βασική ιδιότητα της υπο-συνθήκης ανεξαρτησίας του ακουστικού μοντέλου, για την κατανομή των πιθανοτήτων των παρατηρήσεων, για τις επόμενες και τις προηγούμενες επικρατούσες ακολουθίες (mode sequences) δοθέντος της εκάστοτε κατάστασης.

Κατά τον ίδιο τρόπο αν συμβολίσουμε:

$$b_k(x_k) = P(\{y_{k+1}, ..., y_n\}|x_k, \{y_1, ..., y_k\}) \tag{3.15}$$

Ο επαναληπτικός υπολογισμός για το Πίσω βήμα του αλγορίθμου έχει ως εξής:

$$b_k(x_k) = \sum_{x_{k+1}} P(\{y_{k+1}, ..., y_n\}|x_k, \{y_1, ..., y_k\}) \Leftrightarrow$$

$$b_k(x_k) = \sum_{x_{k+1}} P(y_{k+1}|y_k)P(y_{k+1}|x_{k+1}, \{y_1, ..., y_k\})P(\{y_{k+2}, ..., y_k\}|x_{k+1}, \{y_1, ..., y_{k+1}\}) \Leftrightarrow$$

$$b_k(x_k) = \sum_{x_{k+1}} P(y_{k+1}|y_k)P(y_{k+1}|x_{k+1}, \{y_1, ..., y_k\})b_{k+1}(x_{k+1}) \tag{3.16}$$

όπου και εδώ χρησιμοποιήθηκε η ίδια ταυτότητα που χρησιμοποιήθηκε στο Εμπρός βήμα κατά τον ίδιο τρόπο για $k+1$. Για την περίπτωση των τμηματοποιημένων, για την οποία υποθέτουμε η υπο-συνθήκη ανεξαρτησία μεταξύ των τμημάτων δοθέντος της επικρατούσας διαδικασίας (mode sequence), προκύπτει η σχέση:

$$P(y_{k+1}|x_{k+1}, \{y_1, ..., y_k\}) = P(y_{k+1}|x_{k+1}) \tag{3.17}$$

Από την παραπάνω υπόθεση οι εξισώσεις (3.14) και (3.16) παίρνουν την μορφή:

$$a_k(x_k) = \Big[\sum_{x_{k-1}} a_{k-1}(x_{k-1})P(x_k|x_{k-1})\Big]P(y_k|x_k) \tag{3.18}$$

$$b_k(x_k) = \sum_{x_{k+1}} P(y_{k+1}|y_k)P(y_{k+1}|x_{k+1})b_{k+1}(x_{k+1}) \tag{3.19}$$

Ο υπολογισμός της επιθυμητής πιθανότητας $P(x_k|\{y_1, ..., y_n\})$ προκύπτει απο τον συνδυασμό των βημάτων και είναι:

$$P(x_k|\{y_1, ..., y_n\}) = \frac{P(x_k, \{y_1, ..., y_n\})}{P(\{y_1, ..., y_n)\}} = \frac{a_k(x_k)b_k(x_k)}{\sum_{x_k} a_k(x_k)b_k(x_k)} \tag{3.20}$$

Εκτίμηση Μέγιστης Πιθανοφάνειας

Η βασική ιδέα της εκτίμησης της μέγιστης πιθανοφάνειας είναι πως ο καλύτερος υπολογισμός προκύπτει απο εκείνον που μεγιστοποιεί την πιθανοφάνεια, περιλαμβανομένου των πραγματικών παρατηρήσεων. Εφόσον αναφερόμαστε σε *state space* μοντέλα όπου η κατάσταση

είναι ακριβέστερη περιγραφή, των παρατηρήσεων, η μέγιστη πιθανοοφάνεια προκύπτει όταν το μοντέλο ακολουθεί κατανομή παρόμοια αυτής των παρατηρήσεων. Η εκτίμηση της μέγιστης πιθανοφάνειας έχει τις ακόλουθες ιδιότητες, όσο περισσότερα δεδομένα εκπαίδευσης έχουμε τόσο καλύτερα συγκλίνει, τις περισσότερες φορές είναι αρκετά απλή στην υλοποίηση της, αλλά βεβαίως δεν έχει πάντα αναλυτική λύση.

Θεωρούμε ένα απλοποιημένο παράδειγμα του υπολογισμού της μέγιστης πιθανοφάνειας, όπου θέλουμε να υπολογίσουμε την μέγιστη πιθανότητας μιας ακολουθίας ν δειγμάτων $x_1, ..., x_n$ δεδομένων των παραμέτρων θ. Η υπο-συνθήκη κατανομή πιθανότητας δίνεται από την εξίσωση:

$$P(\{x_1...x_n\}|\Theta) = \prod_{k=1}^{n} P(x_k|\theta) \tag{3.21}$$

Λογαριθμίζοντας την παραπάνω σχέση ο παράγοντας του γινομένου μετατρέπεται σε αθροιστικό παράγοντα, κάνοντας έτσι τους υπολογισμούς ταχύτερους. Ορίζοντας την λογαριθμική πιθανότητα:

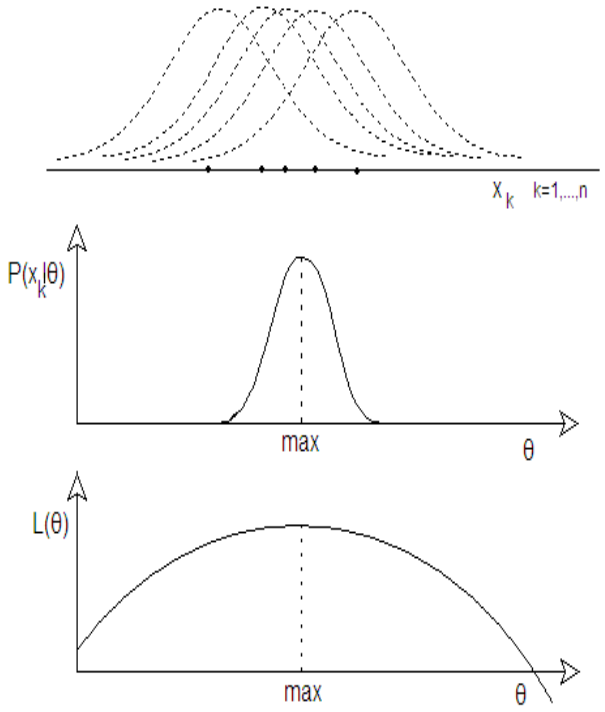
$$L(\theta) = lnP(x_k|\theta) \tag{3.22}$$

Τότε οι παράμετροι του μοντέλου προκύπτουν ως εξής:

$$\hat{\theta} = arg_\theta max L(\theta) \tag{3.23}$$

Στο σχήμα (3.3) φαίνεται ο τρόπος υπολογισμού της μέγιστης πιθανότητας.

Ο περιορισμός της μεθόδου είναι πως στην περίπτωση που τα δεδομένα έχουν κενά σε ενδιάμεσα χρονικά σημεία (sparse data), όπως και στην περίπτωσή μας, και όταν η ακολουθία της κατάστασης δεν είναι άμεσα παρατηρήσιμη, η μέθοδος δεν μπορεί να παράξει σωστό αποτέλεσμα. Για αυτές τις περιπτώσεις χρησιμοποιείται ο αλγόριθμος αναμονής μεγιστοποίησης (Expectation Maximization Algorithm - EM Algorithm).



η

Σχήμα 3.3: Στο επάνω σχήμα δείχνουμε την κατανομή που ακολουθούν μερικά δεδομένα εκπαίδευσης, στη μέση φαίνεται ο υπολογισμός με βάση την εξίσωση (3.21) και το πότε γίνεται μεγιστοποίηση της ποσότητας και στην τρίτη ο υπολογισμός με βάση την εξίσωση (3.22).

Αλγόριθμος ΕΜ

Ο αλγόριθμος ΕΜ αντιμετωπίζει το πρόβλημα μετατίθοντας τον υπολο- γισμό της μέγιστης πιθανότητας. Αντί να μεγιστοποιήσουμε την ολική πιθανοφάνεια, (εξίσωση 3.11), μπορούμε να μεγιστοποιήσουμε την παρακάτω ποσότητα επαναληπτικά.

$$Q(\theta_k|\theta_{k-1}) = E\{P(\{x_1...x_n\},\{y_1...y_n\}|\Theta)|\{y_1...y_n\},\theta_{k-1}\} \tag{3.24}$$

Η επανάληψη πραγματοποιείται σε δύο βήματα, πρώτα γίνεται το βήμα της αναμονής (Expectation) όπου υπολογίζεται η ποσότητα (3.24), δοθέντος του υπολογισμού των παραμέτρων της προηγούμενης επανάληψης θ_{k-1}

$$Q(\theta_k|\theta_{k-1}) = E\{P(\{x_1...x_n\},\{y_1...y_n\}|\Theta)|\{y_1...y_n\}\} \tag{3.25}$$

και στο δεύτερο βήμα γίνεται η μεγιστοποίηση των παραμέτρων του μοντέλου ($Maximization$) από το πρώτο βήμα της αναμονής

$$Q(\theta_k) = arg_\theta max Q(\theta_k|\theta_{k-1}) \tag{3.26}$$

Ο αλγόριθμος μας υπόσχεται αύξηση της μέγιστης πιθανότητας σε κάθε επανάληψη [1]. Η εφαρμογή του ΕΜ αλγορίθμου σαν διαδικασία εκπαίδευσης στηρίζεται στην ιδέα, ο υπολογισμός της κρυφής μεταβλητής (κατάστασης) δοθέντος των παρατηρήσεων, μέσω της διαδικασίας Εμπρός - Πίσω, εφαρμόζεται στις υπάρχουσες παραμέτρους του μοντέλου. Έχοντας τους υπολογισμούς της διαδικασίας Εμπρός - Πίσω, τους χρησιμοποιούμε για την ανανέωση των παραμέτρων του μοντέλου. Η διαδικασία συνεχίζει χρησιμοποιώντας τους ανανεωμένους παραμέτρους επανεκτιμώντας με βάση αυτούς τις τιμές της κατάστασης. Η επανάληψη της διαδικασίας γίνεται μέχρις ότου οι τιμές των παραμέτρων συγκλίνουν. Το κριτήριο σύγκλισης είναι η Μέγιστη Πιθανότητα να αποκτήσει τιμή σχεδόν σταθερή.

3.4 Σύνοψη

Στο κεφάλαιο αυτό αναλύθηκε όλη η θεωρία που χρησιμοποιείται για την εφαρμογή των γραμμικών συστημάτων στην αναγνώριση φωνής. Επιδίωξη ήταν να δοθεί η θεωρία, όσο το

δυνατόν, στη γενικότερη μορφή της, περιλαμβάνοντας τις γενικές ιδέες της εκάστοτε έννοιας. Στο επόμενο κεφάλαιο θα δοθεί η ακριβής εφαρμογή και ο πως ο συνδυασμός τον εννοιών οδηγεί στην εξαγωγή ασφαλών συμπερασμάτων.

Κεφάλαιο 4

Γραμμικά Δυναμικά Συστηματα ως Ακουστικό Μοντέλο

Στο κεφάλαιο αυτό ερευνούμε μία νέα δομή παραμετροποίησης του γραμμικού συστήματος και εισάγουμε έναν νέο τρόπο υπολογισμού (*estimation*) των γραμμικών δυναμικών μοντέλων στην Αναγνώριση Φωνής. Η δομή παραμετροποίησης προτείνεται Lennart Ljung [2], είναι όμως η πρώτη φορά που χρησιμοποιείται για την δόμηση γραμμικών δυναμικών μοντέλων στην Αναγνώριση Φωνής. Στην επεξεργασία σήματος ο υπολογισμός της συνδιακύμανσης είναι απαραίτητος σε κάθε περίπτωση. Ο λόγος αυτός κάνει απαραίτητη την χρησιμοποίηση της θεωρίας υπολογισμού των πινάκων στην επεξεργασία σήματος. Ο νέος τρόπος υπολογισμού είναι η εφαρμογή $element - wise$ για τον υπολογισμό των παραμέτρων των γραμμικών δυναμικών μοντέλων. Κατά την τεχνική αυτή η εκτίμηση γίνεται στοιχείο-στοιχείο.

Στο κεφάλαιο αυτό αρχικά παρουσιάζουμε την κανονική μορφή του γραμμικού συστήματος και για τους λόγους που επιλέχθηκε, εδάφιο (4.1), στο εδάφιο (4.2) παρουσιάζουμε τον νέο τρόπο υπολογισμού των παραμέτρων των γραμμικών δυναμικών μοντέλων, στο εδάφιο (4.3) εξηγούμε την εφαρμογή των παραπάνω στην διαδικασία εκπαίδευσης με τεχνητά δεδομένα και πραγματικά δεδομένα φωνής, και τέλος στο εδάφιο (4.4) αποτιμούμε τη νέα μέθοδος.

4.1 Γενικευμένη Κανονικοποιημένη Φόρμα

Περιορισμοί προηγούμενων υλοποιήσεων

Οι υπάρχουσες υλοποιήσεις των γραμμικών δυναμικών συστημάτων, εισήγαγαν αρκετούς

περιορισμούς στις παραμέτρους του συστήματος. Κατά την εισαγωγή τους [1] οι πίνακες συνδιακύμανσης των θορύβων, τέθηκαν διαγώνιοι για τον λόγο ότι αναπαριστούν πιο εξομαλυμένες κατανομές, σε σχέση με το πλήρη πίνακα συνδιακύμανσης. Ο πίνακας F κανονικοποιείται σύμφωνα με αυτή του παραρτήματος Γ′ ενώ ο πίνακας H θεωρείται γνωστός και είναι ίσος με τον μοναδιαίο πίνακα. Άλλες υλοποιήσεις εισήγαγαν διαφορετικούς περιορισμούς, χρησιμοποιώντας διάφορες τεχνικές μοντελοποίησης. Η factor analysis (FA) τεχνική θέτει τον πίνακα F να είναι ίσος με το μηδεν, $F = 0$, και περιορίζει την συμμεταβλητότητα του θορύβου της παρατήρησης να είναι διαγώνιος [9]. Είναι όμως γνωστό πως αν η συμμεταβλητότητα του θορύβου της εξίσωσης της κατάστασης, είναι είτε διαγώνιος είτε μοναδιαίος, το γεγονός αυτό δεν επηρεάζει την γενικότητα. Ακόμη η συμμεταβλητότητα του θορύβου της εξίσωσης της παρατήρησης ομοίως θεωρήθηκε διαγώνιος, με την θεώρηση αυτή υπάρχουν απώλειες στη γενικότητα, αλλά έχει το πλεονέκτημα ότι δίνει λιγότερους ελεύθερους παραμέτρους. Για τον πίνακα F ο περιορισμός να είναι θετικός αλλά και μικρότερος της μονάδας, $0 < |F| < 1$ επειδή αν είναι μεγαλύτερος της μονάδας, η εξέλιξη της κατάσταση θα ακολουθούσε εκθετική αύξηση, συμπεριφορά που δημιουργεί προβλήματα σε τμήματα με λίγα διαστήματα [3].

Γραμμικό δυναμικό σύστημα χωρίς περιορισμούς

Έστω γραμμικό δυναμικό σύστημα που περιγράφεται από τις εξισώσεις (4.1) και (4.2).

$$x_{k+1} = Fx_k + B\varrho_k + \omega_k \qquad \omega_k \sim N(0, Q) \tag{4.1}$$

$$y_k = Hx_k + \upsilon_k \qquad \upsilon_k \sim N(0, R) \tag{4.2}$$

όπου B είναι πίνακας γεμάτος παραμέτρους, διάστασης $(n \times l)$, με n την διάσταση του πίνακα F και ϱ_k ντετερμινιστικό διάνυσμα, διάστασης $(l \times l)$.

Για ένα state space μοντέλο πολλών μεταβλητων ο Lennart Ljung [2] (σελίδα 119) συνιστά παραμετροποιημένη δομή για τον πίνακα του συστήματος (system matrix) F, διάστασης $n \times n$.

Για παράδειγμα με n=9 προκύπτει ο πίνακας.

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \times & \times & \times & \times & \times & \times \end{bmatrix}$$

όπου $\times$ συμβολίζονται οι ελεύθεροι παράμετροι του πίνακα του συστήματος.

Ο πίνακας F καθορίζει τόσο την διάσταση του διανύσματος κατάστασης, όσο και την διάσταση του διανύσματος των παρατηρήσεων, το ίδιο ισχύει και αντιστρόφως. Όπως είναι προφανές από το σύστημα εξισώσεων του γραμμικού συστήματος (4.1) και (4.2), η διάσταση του διανύσματος κατάστασης x_k είναι $n \times 1$. Η διάσταση του διανύσματος των παρατηρήσεων y_k είναι ίση με τον αριθμό των γραμμών του πίνακα, που τα στοιχεία τους είναι ελεύθεροι παράμετροι. Έστω m ο αριθμός αυτών των γραμμών, για το παράδειγμα μας $m = 3$. Οι ελεύθεροι παράμετροι μπορούν να τοποθετηθούν σε τυχαία επιλεγμένες γραμμές. Όπως θα δούμε στην συνέχεια, η θέση τους επηρεάζει μόνο την δομή του πίνακα των παρατηρήσεων του συστήματος H, η οποία καθορίζεται με την εφαρμογή αλγορίθμου, που επίσης δόθηκε στο βιβλίο του Lennart Ljung. Ο πίνακας F περιορίζεται μόνο σε τιμή θετική.

Ο κύριος περιορισμός στην παραμετροποίηση του συστήματος είναι στον πίνακα παρατήρησης H του συστήματος. Η διάσταση του καθορίζεται από τις διαστάσεις των διανυσμάτων κατάστασης και παρατήρησης. Άν x_k και y_k, έχουν $n \times 1$ και $m \times 1$ διαστάσεις αντίστοιχα, τότε ο πίνακας παρατήρησης θα έχει διάσταση $m \times n$. Για το παράδειγμά, όπου $n = 9$ και $m = 3$, ο πίνακας θα έχει διάσταση 3×9.

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Οι θέσεις των μονάδων εντός του πίνακα δεν είναι τυχαία επιλεγμένες, προκύπτουν από τις θέσεις των γραμμών με ελεύθερους παραμέτρους, του πίνακα του συστήματος F. Κάθε γραμμή επιτρέπεται να έχει ένα μόνο στοιχείο της ίσο με την μονάδα. Για τον εντοπισμό τον κατάλληλων θέσεων η διαδικασία είναι η εξής:

Θεωρούμε ένα πίνακα F, διάστασης $n \times n$, με m αριθμό γραμμών ελευθέρων παραμέτρων. Για την πρώτη γραμμή r_i του πίνακα F με ελεύθερους παραμέτρους, θέτουμε $row_1 = i$, για την δεύτερη γραμμή r_i του πίνακα F με ελεύθερους παραμέτρους θέτουμε $row_2 = i$, ομοίως για τις υπόλοιπες γραμμές. Τότε ο πίνακας H, στη γραμμή j, θα έχει μονάδα στην στήλη $c = row_{j-1} + 1$, με $row_0 = 0$. Η υπόθεση για το r_0 δηλώνει πως στην πρώτη γραμμή μονάδα θα έχει πάντοτε η πρώτη στήλη.

Στο παράδειγμα ο F έχει ελεύθερους παραμέτρους στις γραμμές r_3, r_5, r_9. Θέτουμε $row_1 = 3$, $row_2 = 5$ και $row_3 = 9$. Ο πίνακας H έχει άσσους στις θέσεις,

$$j = 1 \Rightarrow c = row_{1-1} + 1 = row_0 + 1 = 1 \rightarrow H(j, c) = H(1, 1) = 1$$

$$j = 2 \Rightarrow c = row_{2-1} + 1 = row_1 + 1 = 4 \rightarrow H(j, c) = H(2, 4) = 1$$

$$j = 3 \Rightarrow c = row_{3-1} + 1 = row_2 + 1 = 6 \rightarrow H(j, c) = H(3, 6) = 1$$

Ο πίνακας B θεωρείται πρόσθετο συστατικό για τον έλεγχο του συστήματος. Η μορφή και τα στοιχεία του είναι ελεύθερα στον καθορισμό τους. Η εισαγωγή ενός τέτοιου πίνακα δίνει μεγάλες δυνατότητες, αφού επιτρέπει την εισαγωγή πρόσθετων φαινομένων στο παρακολουθούμενο σύστημα.

$$\mathbf{B} = \begin{bmatrix} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{bmatrix}$$

Τέλος οι συνδιακυμάνσεις των θορύβων μπορούν να είναι πλήρης, αλλά περιορίζονται ως θετικοί και συμμετρικοί. Για την κανονικοποιημένη παραμετροποίηση αυτή, ο Lennart Ljung [2] (σελίδα 121) αποδεικνύεται ότι είναι πλήρως ταυτοποιημένη (global identifiability), προυπόθεση απαραίτητη στην μοντελοποίηση.

4.2 Συνολική εικόνα υπολογισμού

Στο κεφάλαιο 3 αναλύθηκε η θεωρητική βάση των εννοιών, που χρησιμοποιούνται για τον υπολογισμό των παραμέτρων, με στόχο την προτυποποίηση. Αφού καθορίσαμε το δυναμικό μοντέλο στο προηγούμενο εδάφιο, με την νέα κανονικοποιημένη δομή, μπορούμε να αναλύσουμε την διαδικασία υπολογισμού των γραμμικών δυναμικών μοντέλων, ως πραγματική εφαρμογή με κάθε λεπτομέρεια, δίνοντας τόσο τεχνητά δεδομένα όσο με δεδομένα φωνής. Η διαδικασία που αναλύεται κυρίως ακολουθεί την εργασία του Βασίλη Διγαλάκη [1].

Θεωρώντας το σύστημα των εξισώσεων (4.1) και (4.2) διακρίνουμε δύο περιπτώσεις υλοποίησης, ανάλογα με την τιμή του πίνακα B. Η πρώτη είναι να τον θεωρήσουμε μηδενικό πίνακα και την άλλη μή - μηδενικό. Στην εργασία αυτή, εφαρμόστηκε μόνο η περίπτωση για B μηδενικό πίνακα, σε πραγματικά δεδομένα φωνής. Οι επιπλέον υπολογισμοί στην περίπτωση μη-μηδενικού B υπάρχουν στο Παράρτημα Δ.

Μηδενικός Πίνακας B

Αν θεωρήσουμε μηδενικό B προκύπτει το γραμμικό σύστημα των εξισώσεων:

$$x_{k+1} = Fx_k + \omega_k \qquad \omega_k \sim N(0, Q) \tag{4.3}$$

$$y_k = Hx_k + \upsilon_k \qquad \upsilon_k \sim N(0, R) \tag{4.4}$$

Ο υπολογισμός της Μεγιστης Πιθανότητας (Maximun Likelihood) για τους παραμέτρους θ του συστήματος, δοθέντος το διάνυσμα των παρατηρήσεων $Y = [y_1, ..., y_N]$ και με διάνυσμα κατάστασης $X = [x_1, ..., x_N]$, με βάση τις εξισώσεις (B-1) και (B-2) είναι:

$$P(x_k|x_{k-1}, \theta) = \frac{1}{\sqrt{(2\pi)^a|Q|}} exp\Big\{ -\frac{1}{2}(x_k - Fx_{k-1})^T Q^{-1}(x_k - Fx_{k-1}) \Big\} \tag{4.5}$$

$$P(y_k|x_k, \theta) = \frac{1}{\sqrt{(2\pi)^a|R|}} exp\Big\{ -\frac{1}{2}(y_k - Hx_k)^T R^{-1}(y_k - Hx_k) \Big\} \tag{4.6}$$

αντίστοιχα, όπου a η διάσταση των πινάκων Q και R.

Ο υπολογισμός της συνδυασμένης πιθανότητας για τα διανύσματα Y και X γίνεται μέσω

της εξίσωσης, όπως είδαμε και στο κεφάλαιο 3 (σελ. ;;;).

$$P(X,Y|\theta) = P(x_0)\prod_{k=1}^{N-1} P(x_{k+1}|x_k,\theta)\prod_{k=1}^{N} P(y_k|x_k,\theta) \tag{4.7}$$

όπου $P(x_0)$ είναι η αρχικοποίηση, για την οποία θα αναφερθούμε στην συνέχεια για τον τρόπο υπολογισμού της.

Αντικαθιστώντας τις εξισώσεις (4.5) και (4.6) στην (4.7), και λογαριθμίζοντας.

$$L(\mathbf{X},\mathbf{Y},\theta) = -\sum_{k=1}^{N}\Big\{\log|Q| + (x_k - Fx_{k-1})^T Q^{-1}(x_k - Fx_{k-1})\Big\}$$

$$-\sum_{k=0}^{N}\Big\{\log|R| + (y_k - Hx_k)^T R^{-1}(y_k - Hx_k)\Big\} + constant \tag{4.8}$$

Για τον υπολογισμό των παραμέτρων του μοντέλου μέχρι τώρα στην Αναγνώριση Φωνής πραγματοποιούνταν μεγιστοποιώντας την εξίσωση της Μέγιστης Πιθανότητας για κάθε παράμετρο του μοντέλου διαδοχικά. Μια νέα προσέγγιση είναι η χρησιμοποίηση της element - wise τεχνικής υπολογισμού των πινάκων [13]. Έτσι ο υπολογισμός των παραμέτρων του μοντέλου πραγματοποιείται μεγιστοποιώντας κάθε στοιχείο των παραμέτρων του μοντέλου ξεχωριστά. Η εξαγωγή τον εξισώσεων προέκυψαν από την προσπάθεια του Χρίστου Κόνιαρη.

$$\hat{F}_{ij} = \frac{\sum_{c=1}^{M}\Big\{(cof(\hat{P}_{ic}))(S_{cj}^{(4)})\Big\}}{(cof(\hat{P}_{ii}))(S_{jj}^{(3)})} - \frac{\sum_{c=1,c\neq i}^{M}\Big\{(cof(\hat{P}_{ic}))(\hat{F}_{cj})(S_{jj}^{(3)})\Big\}}{(cof(\hat{P}_{ii}))(S_{jj}^{(3)})} - \frac{\sum_{c=1}^{M}\Big\{(cof(\hat{P}_{ic}))\sum_{r=1,r\neq j}^{M}\Big\{(\hat{F}_{cr})(S_{rj}^{(3)})\Big\}\Big\}}{(cof(\hat{P}_{ii}))(S_{jj}^{(3)})} \tag{4.9}$$

$$\hat{P}_{ij} = (S_{ij}^{(2)}) - \sum_{r=1}^{M}(\hat{F}_{ir})(S_{jr}^{(4)}) - \sum_{r=1}^{M}(\hat{F}_{jr})(S_{ir}^{(4)}) + \sum_{c=1}^{M}\sum_{r=1}^{M}(\hat{F}_{ic})(\hat{F}_{jr})(S_{cr}^{(3)}) \tag{4.10}$$

Για την συνδιακύμανση του θορύβου των παρατηρήσεων, R, εκτιμάτε ως εξής.

$$\hat{R} = S^{(5)} - S^{(6)}(S^{(1)})^{-1}(S^{(6)})^T \tag{4.11}$$

Η εφαρμογή του EM αλγόριθμου, όπως των αναλύσαμε και στο κεφάλαιο 3, πραγματοποιείται σε δύο στάδια, ένα της μεγιστοποίησης και ένα της αναμονής. Κατά το στάδιο της μεγιστοποίησης ο υπολογισμός των στατιστικών $S^{(1)}$, $S^{(2)}$, $S^{(3)}$, $S^{(4)}$, $S^{(5)}$ και $S^{(6)}$ γίνεται ως εξής:

$$S^{(1)} = \frac{1}{N+1}\sum_{k=0}^{N} x_k x_k^T \tag{4.12}$$

$$S^{(2)} = \frac{1}{N}\sum_{k=1}^{N} x_k x_k^T \tag{4.13}$$

$$S^{(3)} = \frac{1}{N}\sum_{k=1}^{N} x_{k-1} x_{k-1}^T \tag{4.14}$$

$$S^{(4)} = \frac{1}{N}\sum_{k=1}^{N} x_k x_{k-1}^T \tag{4.15}$$

$$S^{(5)} = \frac{1}{N+1}\sum_{k=0}^{N} y_k y_k^T \tag{4.16}$$

$$S^{(6)} = \frac{1}{N+1}\sum_{k=0}^{N} y_k x_k^T. \tag{4.17}$$

Από τις εξισώσεις των στατιστικών κατά το στάδιο της μεγιστοποίησης φαίνεται πως είναι απαραίτητος ο υπολογισμός της αναμενόμενης τιμής των ποσοτήτων $y_k x_k^T, y_k y_k^T, x_k x_{k-1}^T, x_k x_k^T$ (στάδιο αναμονής). Από την στιγμή που είσοδοι του συστήματος είναι *Gaussian* τότε και η κατάσταση είναι *Gaussian*, έτσι η υπο-συνθήκη κατανομή της κατάστασης δοθέντος των παρατηρήσεων καθορισμένου διαστήματος, είναι:

$$P(x_k|Y) \sim N(\hat{x}_{k|N}, \Sigma_{k|N})$$

Έτσι τα στατιστικά, για το στάδιο της αναμονής, δίνονται από τις εξισώσεις:

$$E_{\theta^{(p)}}\{y_k x_k^T|\mathbf{Y}\} = y_k \hat{x}_{k|N} \tag{4.18}$$

$$E_{\theta^{(p)}}\{y_k y_k^T|\mathbf{Y}\} = y_k y_k^T \tag{4.19}$$

$$E_{\theta^{(p)}}\{x_k x_{k-1}^T|\mathbf{Y}\} = \Sigma_{k,k-1|N} + \hat{x}_{k|N}\hat{x}_{k-1|N}^T \quad (4.20)$$

$$E_{\theta^{(p)}}\{x_k x_k^T|\mathbf{Y}\} = \Sigma_{k|N} + \hat{x}_{k|N}\hat{x}_{k|N}^T. \quad (4.21)$$

Οι ποσότητες $\hat{x}_{k|N}$, $\hat{x}_{k-1|N}$, $\Sigma_{k|N}$ και $\Sigma_{k,k-1|N}$ προκύπτουν από την εφαρμογή του αλγορίθμου Εμπρός - Πίσω (Forward - Backward Algorithm). Για το εμπρός βήμα χρησιμοποιείται ο επαναληπτικός υπολογισμός του φίλτρου Kalman (Παράρτημα Β). Οι επιθυμητές ποσότητες προκύπτουν από το Πίσω βήμα, για το οποίο χρησιμοποιείται ο Rauch-Tung-Striebel εξομαλυντής. Οι εξισώσεις υπολογισμού φάινονται παρακάτω. Οι επιπλέον cross covariance ποσότητες, αποδεικνύονται στην εργασία του Βασίλη Διγαλάκη [1].

Εμπρός Επαναληπτικές εξισώσεις

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k e_k \quad (4.22)$$

$$\hat{x}_{k+1|k} = F\hat{x}_{k|k} \quad (4.23)$$

$$e_k = y_k - H\hat{x}_{k|k-1} \quad (4.24)$$

$$K_k = \Sigma_{k|k-1}H^T\Sigma_{e_k}^{-1} \quad (4.25)$$

$$\Sigma_{e_k} = H\Sigma_{k|k-1}H^T + R \quad (4.26)$$

$$\Sigma_{k|k} = \Sigma_{k|k-1} - K_k\Sigma_{e_k}K_k^T \quad (4.27)$$

$$\Sigma_{k,k-1|k} = (I - K_k H)F\Sigma_{k-1|k-1} \quad (4.28)$$

$$\Sigma_{k+1|k} = F\Sigma_{k|k}F^T + P \quad (4.29)$$

Πίσω Επαναληπτικές εξισώσεις

$$\hat{x}_{k-1|N} = \hat{x}_{k-1|k-1} + A_k[\hat{x}_{k|N} - \hat{x}_{k|k-1}] \quad (4.30)$$

$$\Sigma_{k-1|N} = \Sigma_{k-1|k-1} + A_k[\Sigma_{k|N} - \Sigma_{k|k-1}]A_k^T \quad (4.31)$$

$$A_k = \Sigma_{k-1|k-1}F^T\Sigma_{k|k-1}^{-1} \quad (4.32)$$

$$\Sigma_{k,k-1|N} = \Sigma_{k,k-1|k} + [\Sigma_{k|N} - \Sigma_{k|k}]\Sigma_{k|k}^{-1}\Sigma_{k,k-1|k}$$

4.3 Εκπαιδευτική Διαδικασία

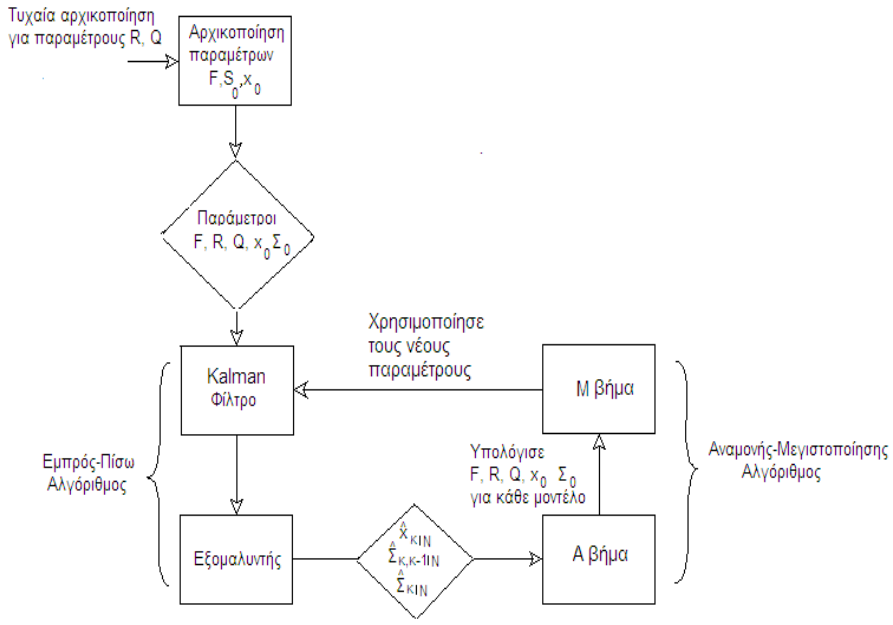
Η εφαρμογή του γραμμικού δυναμικού μοντέλου στην ακουστική μοντελοποίηση, γίνεται εκπαιδεύοντας τμήματα φωνής, που περιέχουν συγκεκριμένη μεταβλητότητα. Για να το πετύχουμε αυτό, θεωρούμε τα τμήματα της φωνής προς μοντελοποίηση, τμήματα μεταβλητού μήκους διαστημάτων. Θέτοντας τμήματα μεταβλητού μήκους διαστημάτων αναπαρίστανται καλύτερα τα χαρακτηριστικά του τμήματος της φωνής. Ανάμεσα στα διαστήματα υποθέτουμε ότι η συσχέτιση παραμένει αμετάβλητη.

Κάθε τμήμα φωνής που επιλέγεται για μοντελοποίηση, του αποδίδουμε από ένα σετ παραμέτρων συστήματος.

$$\theta = \{F, H, Q, R, \Sigma_0\} \tag{4.33}$$

όπου Σ_0 η αρχική συμμεταβλητότητα για το εκάστοτε τμήμα φωνής.

Η διαδικασία εκπαίδευσης φαίνεται στο παρακάτω σχήμα.



Σχήμα 4.1: *Εκπαιδευτική διαδικασία*

Ο υπολογισμός της αρχικής συμμεταβλητότητας και της αρχικής μέσης τιμής δίνεται απο την εξίσωση:

$$\hat{\Sigma}_0 = \frac{1}{N}\sum_{k=1}^{N} E\{x_0 x_0^T | Y\} \tag{4.34}$$

$$\hat{x}_0 = \frac{1}{N}\sum_{k=1}^{N} E\{x_0 | Y\} \tag{4.35}$$

όπου N το πλήθος εμφανίσεων ως αρχικά τμήματα.

Η αρχικοποίηση του πίνακα κατάστασης F γίνεται χρησιμοποιώντας την έξοδο του συστήματος, δηλαδή την ακολουθία $Y = [y_1, y_2, ..., y_N]$. Θεωρούμε την εξίσωση της καταστάσεως και θέτουμε όπου κατάσταση x την παρατήρηση y, χωρίς να συμπεριλαμβάνουμε τον θόρυβο, έτσι έχουμε:

$$y_{k+1} = F y_k \tag{4.36}$$

Κατ΄ αντιστοιχία με την διαδικασία υπολογισμού, εφαρμόζοντας τον αλγόριθμο Αναμονής Μεγιστοποίησης εύκολα βρίσκουμε:

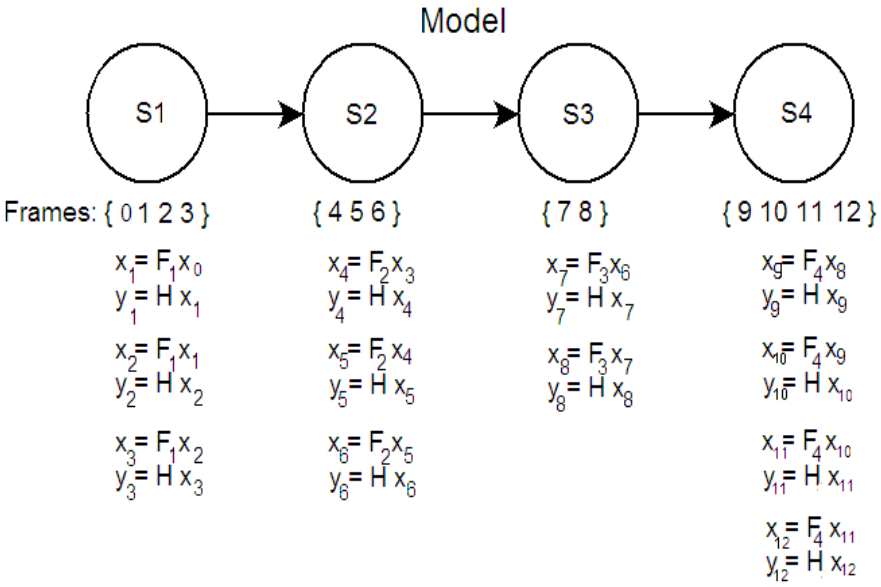
$$F = \left(\frac{1}{N}\sum_{k=1}^{N} E\{y_k y_{k-1}^T | Y\}\right)\left(\frac{1}{N}\sum_{k=1}^{N} E\{y_{k-1} y_{k-1}^T | Y\}\right)^{-1} \tag{4.37}$$

όπου N το πλήθος των διαστημάτων που ανήκουν στο εκάστοτε μοντέλο.

Κατά την εκπαίδευση υιοθετήσαμε την παραδοχή, η τιμή του πίνακα R να καθορίζεται από την πρώτη μόλις επανάληψη του αλγορίθμου Μεγιστοποίησης Αναμονής, ένας άλλος τρόπος είναι να πολλαπλασιάσουμε με μία μικρή σταθερά. Ο λόγος για τον οποίο έγινε η παραδοχή αυτή είναι επειδή παρατηρήθηκε το φαινόμενο της υπερ-εκπαίδευσης (over-train). Η υπερ-εκπαίδευση επιδρούσε στις τιμές του πίνακα κατά τον υπολογισμό, δίνοντας πολύ μικρές τιμές. Αυτό είχε σαν αποτέλεσμα να μην διατηρούνται τα τοπικά χαρακτηριστικά.

Ένα άλλο θέμα ήταν αυτό της αρχικοποίησης. Το πρόβλημα της διατήρησης των εξαρτήσεων ανάμεσα στα μοντέλα (inter-segmental) δίνει τη δυνατότητα δύο υλοποιήσεων,

1. κάθε μοντέλο αρχικοποιείται κάθε φορά που εμφανίζεται, από τον υπολογισμό της εξίσωσης (4.34), και



Σχήμα 4.2: Στο παράδειγμα, η επεξεργασία γίνεται χρησιμοποιώντας την αρχικοποίηση του μοντέλου που περιγράφουμε.

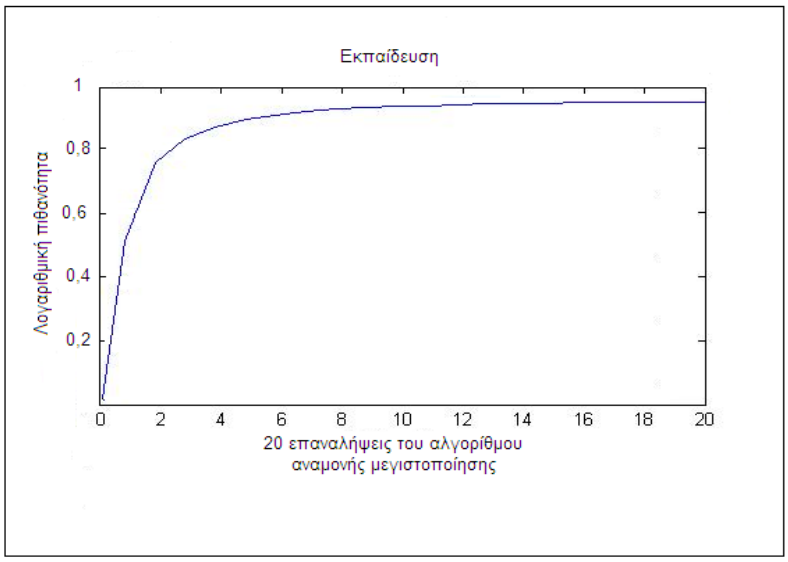
2. να χρησιμοποιήσουμε ως αρχικοποίηση, των τελευταίο υπολογισμό, του προηγούμενου μοντέλου.

Κάθε μοντέλο είναι ανεξάρτητο των περιεχομένων του (context-indepentent). Η επεξεργασία ενός μοντέλου που αποτελείται από τέσσερα τμήματα, ο τρόπος επεξεργασίας του γίνεται όπως φαίνεται στο (4.2).

Προετοιμασία δεδομένων

Η φωνητική βάση που χρησιμοποιήθηκε κατά την εκπαίδευση και την αναγνώριση, όπως θα δούμε στο επόμενο κεφάλαιο, είναι η AURORA 2. Η αιτία που διαλέξαμε την AURORA2 ήταν για τα διάφορα επίπεδα πραγματικού θορύβου που προσφέρει (20db, 15db, 10db, 5db, 0db, -5db). Η βάση αυτή έχει ως λεξιλόγιο τα νούμερα (one, two, three, four, five, six, seven, eight, nine, zero, oh), έτσι η μοντελοποίηση έγινε σε επίπεδο λέξεων. Για την εξαγωγή των Mel frequency cespstal coefficients (MFC) και για τα time-alignment χρησιμοποιήθηκε το HTK.

Κάθε λέξη χωρίστηκε σε τμήματα, δύο τμήματα ανά φώνημα.



Σχήμα 4.3: *H* Μέγιστη πιθανότητα για δεδομένα εκπαίδευσης μετά από 20 επαναλήψεις του αλγορίθμου Μεγιστοποίησης Αναμονής

one	two	three	four	five	six	seven	eight	nine	oh	zero
6	4	6	6	6	4	8	4	6	2	6

Κάθε τμήμα παραμετροποιήθηκε με σετ παραμέτρων, που καθορίστηκε στην εξίσωση (4.33). Τα δεδομένα εκπαίδευσης της AURORA2 που χρησιμοποιήσαμε είναι οι 8400 προτάσεις, με πολλαπλές λέξεις, χωρίς θόρυβο. Η συνολική Μέγιστη Πιθανότητα μετά από 20 επαναλήψεις του αλγορίθμου Μεγιστοποίησης Αναμονής, όπως φαίνεται στην γραφική παράσταση (4.3), η τιμή της φαίνετε να σταθεροποιείται μετά από μόνο 12 επαναλήψεις.

4.4 Σύνοψη

Συνοψίζοντας, με μια νέα παραμετροποίηση και ένας νέος τρόπος υπολογισμού των παραμέτρων του συστήματος, διατηρήθηκαν τα πλεονεκτήματα της εφαρμογής των γραμμικών δυναμικών μοντέλων. Η εφαρμογή του Αναμονή Μεγιστοποίηση αλγόριθμου διατήρησε την ιδιότητα σύγκλισης, για τα χρονικά μεταβαλλόμενα τμήματα διατηρήθηκαν οι εξαρτήσεις

μεταξύ των διαστημάτων τους. Ο στόχος όμως, κάθε νέας εφαρμογής είναι η αξιολόγηση της . Στο επόμενο κεφάλαιο θα δούμε πως γίνεται η αξιολόγηση, και θα συγκρίνουμε τη μέθοδο των γραμμικών μοντέλων, με την μέθοδο των HMM.

Κεφάλαιο 5

Ταξινόμηση

Ένα σύστημα αναγνώρισης φωνής πρέπει να αξιολογηθεί ως προς την απόδοσή του. Ένας τρόπος είναι η ταξινόμηση. Εφαρμόσαμε την μέθοδο αυτή με δύο τρόπους, μία για κάθε ξεχωριστή γλωσσική μονάδα, και μία στην οποία υπάρχουν πολλές γλωσσικές μονάδες στην σειρά. Για να αξιολογήσουμε το μοντέλο μας το θέσαμε αντιμέτωπο στην πιο διαδεδομένη μέθοδο αναγνώρισης, αυτή των HMM, χρησιμοποιώντας της ίδιες συνθήκες και δεδομένα.

Το κεφάλαιο οργανώνεται ως ακολούθως, στο εδάφιο (5.1) αναλύουμε τα δεδομένα που χρησιμοποιήθηκαν, στο εδάφιο (5.2) δίνεται ο τρόπος ταξινόμησης κατά λέξη, και στο (5.3) τα αποτελέσματα της. Στο εδάφιο (5.4) αναλύουμε την διαδικασία της ταξινόμησης κατά πρόταση και στο (5.5) τα αποτελέσματα της.

5.1 Πρετοιμασία Δεδομένων

Ένα κομμάτι από τα δεδομένα ελέγχου της AURORA 2 χρησιμοποιήθηκαν για την αποτίμηση του μοντέλου. Χρησιμοποιήθηκε από το σετ A το κομμάτι του αυτοκινητοδρόμου (ο τεχνητός προσθετικός θόρυβος στα δεδομένα φωνής είχε τα χαρακτηριστικά πραγματικού αυτοκινητο- δρόμου). Για την εξαγωγή των Mel frequency cespstal coefficients (MFCC) και για τα time-alignment χρησιμοποιήθηκε το HTK.

Τα time-alignment χρειάστηκαν μια επιπρόσθετη διαδικασία σε σχέση με αυτά της εκπαίδευσης. Επειδή κατά την αποτίμηση δεν πρέπει να γνωρίζουμε ποία είναι πρόταση, περιμένουμε από το σύστημα να μας δώσει την πρόταση υπολογίζοντας όλους τους πιθανούς συνδυασμούς

των μοντέλων και να επιλέξει την πιο πιθανή. Τα time-alignments χρειάστηκαν να εξαχθούν καλύπτοντας όλους αυτούς τους πιθανούς συνδιασμούς. Η αναντιστοιχία του περιεχομένου όμως μεταξύ των μοντέλων δημιούργησε προβλήματα στην αποτίμηση του συστήματος. Το οποίο παρακάμφηκε θεωρώντας τα σύνορα των πραγματικών time-alignments.

5.2 Ταξινόμηση κατά λέξη

Η διαδικασία της ταξινόμησης κατά λέξη βασίζεται στην πιθανοφάνεια, για τον υπολογισμό της πιθανοφάνειας χρησιμοποιούμε την σχέση (5.1). Έτσι αν θεωρήσουμε ένα το γραμμικό σύστημα, το ίδιο που χρησιμοποιήσαμε στην διαδικασία εκπαίδευσης, παράγει την ακολουθία παρατηρήσεων $Y = [y_1, y_2, ..., y_N]$, γνωρίζοντας τους παραμέτρους του συστήματος θ, βρίσκοντας τον μέγιστο υπολογισμό της Μέγιστης Πιθανότητας της εξίσωσης (5.1), τα μοντέλα μπορούν να ταξινομηθούν.

$$L(\mathbf{Y}, \theta) = \sum_{k=0}^{N} \Big\{ \log |\Sigma_{e_k}(\theta)| + e_k^T(\theta) \Sigma_{e_k}^{-1}(\theta) e_k^T(\theta) \Big\} + \text{constant}$$

όπου $e_k(\theta)$ και $\Sigma_{e_k}(\theta)$ είναι το σφάλμα πρόβλεψης και τη συμμεταβλητότητα του, που μπορούν να υπολογιστούν από το φίλτρο Kalman, δες Παράρτημα Β.

5.3 Αποτελέσματα Ταξινόμησης κατά λέξη

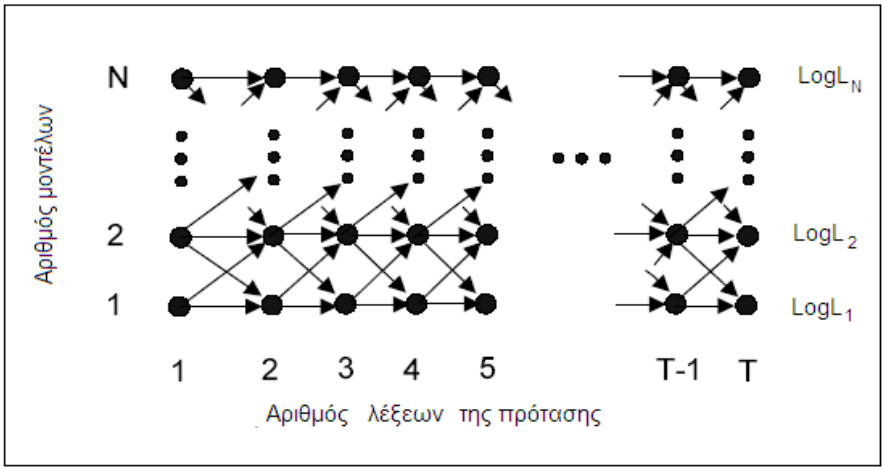
Τα αποτελέσματα της ταξινόμησης του γραμμικού δυναμικού μοντέλου, θα τα συγκρινουμε με τα αποτελέσματα των HMM που προέκυψαν απο το HTK. (LDS - linear dynamic models).

AURORA 2/ Subway	HMM	LDS
	mfc+E	mfc+E
clean	97.88%	98.47%
SNR20	93.24%	94.25%
SNR15	84.64%	88.35%
SNR10	60.13%	77.68%
SNR5	40.44%	64.72%

5.4 Ταξινόμηση κατά πρόταση

Για την ταξινόμηση κατά πρόταση, χρειάζεται να βρούμε ποια είναι, η πιο πιθανή ακολουθία γλωσσικών μονάδων. Η διαφορά στην υλοποίηση, σε σχέση με αυτή της ταξινόμησης κατά λέξης, είναι πως λαμβάνουμε υπόψη την πιθανοφάνεια μιας λέξης αφού γνωρίζουμε την πιθανοφάνεια προηγούμενη της. Πιθανοφάνεια της προηγούμενης είναι όλοι οι πιθανοί συνδυασμοί των μοντέλων. Ο υπολογισμός της ποσότητας Φ εστιάζεται στο πρόβλημα της αναζήτησης (search problem).

Το κριτήριο είναι η εύρεση της πιο πιθανής ακολουθίας, αυτό το πρόβλημα μπορεί να λυθεί χρησιμοποιώντας το διάγραμμα Trellis, αφού μας ενδιαφέρει μόνο το μονοπάτι με την μεγαλύτερη πιθανότητα. Στο σχήμα (5.1) φαίνεται το διάγραμμα Trellis. Ο υπολογισμό



Σχήμα 5.1: Διάγραμμα Trellis

των πιθανοτήτων για κάθε κόμβο στο διάγραμμα, έγινε χρησιμοποιώντας την μέθοδο που εφαρμόστηκε και κατά την ταξινόμηση, με την διαφορά ότι σε επόμενο χρονικά κόμβο, οι πιθανότητες προκύπτουν προσθέτοντας όλες τις πιθανότητες των προηγούμενων κόμβων. Τελικά, την τελευταία χρονική θέση κόμβων, επιλέγεται η μέγιστη πιθανότητα και κάνοντας back-tracking μας παρέχει την πιο πιθανή ακολουθία. Η διαδικασία αυτή είναι γνωστή και ως Viterbi decoding.

Λόγω του γεγονότος χρησιμοποίησης φωνητικής βάσης σε επίπεδο λέξης, βρέθηκε πως

δεν είναι απαραίτητος ο υπολογισμός όλων των μεταβάσεων στο διάγραμμα. Παρατηρήθηκε πως όταν μεταβαίνουμε από προηγούμενους κόμβους σε έναν επόμενο, είχαμε μεταβολή της πιθανότητας μόνο για τα 8-9 διαστήματα. Με βάση αυτή την παρατήρηση θεωρήσαμε διαδικασία pruning, θεωρώντας πως τα 3 μοντέλα με την μεγαλύτερη πιθανότητα στο πρώτο τους τμήμα, είναι αυτά που θα δώσουν τελικά την μεγαλύτερη πιθανότητα. Από σειρά πειραμάτων φάνηκε η αναγνώριση να μην επηρεάζεται από την βελτιστοποίηση αυτή.

5.5 Αποτελέσματα Ταξινόμησης κατά πρόταση

Τα αποτελέσματα της ταξινόμησης του γραμμικού δυναμικού μοντέλου, θα τα συγκρίνουμε με τα αποτελέσματα των HMM που προέκυψαν απο το HTK. (LDS - linear dynamic models).

Accuracy				
AURORA 2/ Subway	HMM		LDS	
	mfcc,E	$+\delta+\delta\delta$	mfcc,E	$+\delta+\delta\delta$
clean	97.19%	97.57%	97.53%	97.61%
SNR20	90.91%	95.71%	93.23%	95.12%
SNR15	80.09%	91.76%	87.91%	91.13%
SNR10	57.68%	81.93%	76.29%	82.69%
SNR5	36.01%	64.24%	54.87%	63.56%

5.6 Σύνοψη

Στο κεφάλαιο αυτό είδαμε πως γίνεται η αποτίμηση, της μεθόδου των γραμμικού δυναμικών συστημάτων. Φαίνεται πως η ταξινόμηση κατά λέξη δίνει καλύτερα αποτελέσματα σε σχέση με την ταξινόμηση κατά πρόταση. Στην πραγματικότητα όμως, η ταξινόμηση κατά πρόταση είναι αυτή που δίνει την εικόνα, για πραγματικές συνθήκες αναγνώρισης. Επίσης το κεφάλαιο αυτό απέδειξε την σπουδαιότητα των γραμμικών δυναμικών μοντέλων στην αναγνώριση φωνής. Τα αποτελέσματα δείχνουν πως μία πρώτη προσέγγιση, σε μια νέα μέθοδο εφαρμογής τους, έχει αποτελέσματα σε επίπεδα θορύβου πολύ καλύτερα από αυτά των HMM.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντική Εργασία

6.1 Αποτίμηση της εργασίας

Στην εργασία αυτή είδαμε την εφαρμογή των γραμμικών δυναμικών μοντέλων στην Αναγνώριση Φωνής. Καθορίσαμε τις ιδιότητες που απαιτούνται να έχει ένα γραμμικό μοντέλο για την εφαρμογή του ως ακουστικό μοντέλο. Προτείναμε μία νέα γενικότερη παραμετροποίηση διατηρώντας τις απαραίτητες αυτές ιδιότητες. Δείξαμε την μέθοδο εκπαίδευσης και τον τρόπο αποτίμησης των γραμμικών μοντέλων ως ακουστικό μοντέλο. Εγινε φανερό πως το πλεονέκτημα διατήρησης των εξαρτήσεων, εντός των τμημάτων, παρέχει βελτιστοποίηση στα διάφορα επίπεδα θορύβου σε σχέση με την παραδοχή της υπο-συνθήκης ανεξαρτησίας της μεθόδου των HMM, διατηρώντας παράλληλα συναφή απόδοση στα καθαρά δεδομένα. Όμως η αδυναμία διατήρησης των εξαρτήσεων ανάμεσα στα τμήματα δεν ολοκληρώνει την υλοποίηση.

6.2 Προτάσεις για Μελλοντική Εργασία

Η λύση του προβλήματος της διατήρησης των εξαρτήσεων ανάμεσα στα τμήματα είναι το κυριότερο αντικείμενο για περαιτέρω έρευνα. Η διαδικασία αναγνώρισης είναι πολλή αργή για πιθανή πραγματική εφαρμογή, κατάλληλοι μέθοδοι για ταχύτερη επεξεργασία είναι αναγκαίοι. Η εισαγωγή του πρόσθετου κοντρόλ υπόσχεται ακόμα καλύτερη προοπτική στην αναγνώριση σε περιβάλλον θορύβου. Τέλος η δόμηση των παραμέτρων του συστήματος

δίνει την δυνατότητα μεταβολλής του διανύσματος κατάστασης σε σχέση με το διάνυσμα παρατήρησης γεγονός ενδιαφέρον.

Παράρτημα Α΄

Στατιστική

Πριν αναλυθεί και εξαχθεί το φίλτρο $Kalman$ είναι απαραίτητο να επαναληφθούν οι βασικές ιδέες του βέλτιστου υπολογισμού. Για απλοποίηση θεωρείται η παρατήρηση δίνεται απο την εξίσωση:

$$y_k = x_k + v_k$$

όπου x_k η άγνωστη - κρυφή μεταβλητή και v_k προσθετικός θόρυβος. Ορίζουμε ως $\hat{x}_k$ τον μεταγενέστερη εκτίμηση για την μεταβλητή x_k δοθέντος των παρατηρήσεων $y_1, ..., y_k$. Για να εκτιμηθεί η ποσότητα $\hat{x}_k$ με βέλτιστο τρόπο, είναι αναγκαία μια συνάρτηση που να περιγράφει το μέγεθος του λάθους της εκτίμησης. Η συνάρτηση θα πρέπει να ικανοποιεί δύο προυποθέσεις:

1. η συνάρτηση να είναι θετική ή μηδέν και

2. να είναι συνάρτηση αύξησης του σφάλματος εκτίμησης $\tilde{x}_k$, το οποίο ισούται

$$\tilde{x}_k = x_k - \hat{x}_k \tag{A-1}$$

Οι δύο αυτές προυποθέσεις ικανοποιούνται από μέσο - τετραγωνικό σφάλμα ($mean - square\ error$) που περιγράφεται από την εξίσωση:

$$J_k = E\{(x_k - \hat{x}_k)^2\} = E\{\tilde{x}_k^2\}$$

όπου Ε ειναι η τελεστής της αναμονής (*expectation*).
Για την εξαγωγή της βέλτιστης εκτίμησης της ποσότητας $\hat{x}_k$, θα συμπεριλάβουμε δύο θεωρήματα από την θεωρία της στατιστικής [5],[11].

Θεώρημα Α΄.0.1 *Άν οι στοχαστικές διαδικασίες y_k και x_k είναι συνδυασμένες (jointly) Gaussian τότε η βέλτιστη εκτίμηση του $\hat{x}_k$ η οποία ελαχιστοποιεί το μέσο - τετραγωνικό σφάλμα J_k είναι η υπο - συνθήκη μέση εκτίμηση:*

$$\hat{x}_k = E\{x_k|y_1, ..., y_k\}. \tag{A-2}$$

Θεώρημα Α΄.0.1 *Έστω οι στοχαστικές διαδικασίες y_k και x_k έχουν μηδενικό μέσο $E\{x_k\} = E\{y_k\} = 0$ για κάθε k. Άν:*

- *οι διαδικασίες y_k και x_k είναι συνδυασμένες (jointly) Gaussian,ή*
- *η βέλτιστη εκτίμηση $\hat{x}_k$ είναι γραμμική συνάρτηση των παρατηρήσεων και η συνάρτηση J_k είναι του μέσου-τετραγωνικού σφάλματος.*

Τότε η βέλτιστη εκτίμηση του $\hat{x}_k$ δοθέντος των παρατηρήσεων $y_1, ..., y_k$, είναι η ορθογώνια προβολή του x_k στο χώρο ύπαρξης των παρατηρήσεων.

Τα θεωρήματα αυτά θα βοηθήσουν στην εξαγωγή των εξισώσεων του φίλτρου *Kalman* Παράρτημα Β.

Παράρτημα Β′

Φίλτρο ΚΑΛΜΑΝ

Το *Kalman* φίλτρο είναι ο βέλτιστος επαναληπτικός αλγόριθμος στην επεξεργασία των δεδομένων. Η αξία του οφείλεται στην προσαρμοστικότητα του για όλες της πληροφορίες που μπορούν να προέρχονται από μη ακριβές μετρήσεις. Υπολογίζει τις εκάστοτε μεταβλητές που ενδιαφέρουν χρησιμοποιώντας την γνώση που έχουμε για το σύστημα, την στατιστική περιγραφή των θορύβων του συστήματος και τις διαθέσιμες πληροφορίες για τις αρχικές συνθήκες των μεταβλητών που θέλουμε. Αφού είναι επαναληπτικός δεν χρειάζεται όλες τις προηγούμενες πληροφορίες να διατηρούνται αλλά μόνο αυτή που υπολογίζεται την εκάστοτε στιγμή. Στην ουσία δεν είναι φίλτρο με την έννοια ότι περιέχει ηλεκτρονικά μέρη ώστε προσαρμόζεται στην έξοδο ενος συστήματος, αλλά ένας αλγόριθμος που υλοποιείται προγραμματιστικά, αυτό υποδηλώνει πως δομικά εντάσσεται κυρίως σε διακριτού χρόνου μετρήσεις, από αυτές του συνεχές χρόνου.

Για να καθοριστεί ο τρόπο εκτίμησης της κατάστασης στη θέση k δοθέντος των προηγουμένων, $\hat{x}_{k|k}$, δηλαδή το γραμμικό ελάχιστο τετραγωνικό υπολογισμό της x δοθέντος της ακολουθίας παρατηρήσεων $\{y_1, ..., y_n\}$. Ορίζουμε την εξίσωση της καινοτομίας ως:

$$\epsilon_k = y_k - \hat{y}_{k|k-1} \tag{B-1}$$

όπου $\hat{y}_{k|k-1}$ είναι ο γραμμικά ελάχιστος τετραγωνικός υπολογισμός της παρατήρησης y_k δοθέντος της ακολουθίας των παρατηρήσεων $\{y_1, ..., y_{n-1}\}$. Η ποσότητα της εξίσωσης είναι ασυσχέτιστη με όλες τις προηγούμενες παρατηρήσεις, επειδή η y_k δεν μας δίνει ολοκληρωμένη την νέα πληροφορία από την στιγμή που η άλλη ποσότητα είναι ήδη καθορισμένη πλήρως.

$$\hat{x}_{k|k} = \text{l.l.s.e. of x given } \{y_1, ..., y_k\}$$

$$\hat{x}_{k|k} = \text{l.l.s.e. of x given } \{\epsilon_1, ..., \epsilon_k\}$$

αφού τα ϵ_k ειναι ασυσχέτιστα μεταξύ τους

$$\hat{x}_{k|k} = \sum_{i=1}^{k} E\{x\epsilon_i^T\}(E\{\epsilon_i\epsilon_i^T\})^{-1}\epsilon_i \tag{B-2}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + E\{x\epsilon_k^T\}(E\{\epsilon_k\epsilon_k^T\})^{-1}\epsilon_k \tag{B-3}$$

Η εξίσωση (B-2) είναι η βάση των εξισώσεων του $Kalman$ φίλτρου. Ας πάρουμε πρώτα ο μέρος της παρατήρησης:

$$y_k = Hx_k + v_k \qquad\qquad v_k \sim N(0, C) \tag{B-4}$$

Η καινοτομία ειναι

$$\epsilon_k = y_k - \hat{y}_{k|k-1} = y_k + H\hat{x}_{k|k-1} = H\tilde{x}_{k|k-1} + v_k \tag{B-5}$$

όπου $\tilde{x}_{k|k-1} = x_k - \hat{x}_{k|k-1}$ το σφάλμα πρόβλεψης, από την B-3 μπορούμε να υπολογίσουμε την συνδιακύμανση της καινοτομίας, τον δεύτερο όρο του γινομένου της εξίσωσης (B-3).

$$\begin{aligned} E\{\epsilon_k\epsilon_k^T\} = E\{H\tilde{x}_{k|k-1} + v_k\}\{H\tilde{x}_{k|k-1}^T + v_k\} \Leftrightarrow \\ E\{\epsilon_k\epsilon_k^T\} = HE\{\tilde{x}_{k|k-1}\tilde{x}_{k|k-1}^T\}H^T + E\{v_k v_k^T\} + HE\{\tilde{x}_{k|k-1}v_k^T\} + \\ + E\{v_k\tilde{x}_{k|k-1}^T\}H^T \end{aligned}$$

όμως $\tilde{x}_{k|k-1}$ και v_k είναι ασυσχέτιστες ποσότητες άρα οι δύο τελευταίοι όροι είναι μηδέν, και η συνδιακύμανση του σφάλματος πρόβλεψης είναι $\Sigma_{k|k-1} = \tilde{x}_{k|k-1}\tilde{x}_{k|k-1}^T$, έχουμε:

$$R_k = E\{\epsilon_k\epsilon_k^T\} = H\Sigma_{k|k-1}H^T + C \tag{B-6}$$

Για δεύτερη ποσότητα του γινομένου της εξίσωσης (B-3) έχουμε:

$$\begin{aligned} E\{x_k\epsilon_k^T\} = E\{x_k\tilde{x}_{k|k-1}^T + H^T + v_k^T\} \Leftrightarrow \\ E\{x_k\epsilon_k^T\} = E\{x_k\tilde{x}_{k|k-1}^T H^T\} \Leftrightarrow \\ E\{x_k\epsilon_k^T\} = E\Big\{\{\tilde{x}_{k|k-1} + \hat{x}_{k|k-1}\}\tilde{x}_{k|k-1}^T\Big\}H' \end{aligned}$$

επειδή όμως $\tilde{x}_{k|k-1}$ και $\hat{x}_{k|k-1}$ είναι ορθογώνια, προκύπτει η εξίσωση (B-7).

$$E\{x_k \epsilon_k^T\} = \Sigma_{k|k-1} H^T \tag{B-7}$$

Η ποσότητα $E\{x\epsilon_k^T\}(E\{\epsilon_k \epsilon_k^T\})^{-1}$ λέγεται το κέρδος $Kalman$ και ισούται:

$$K_k = \Sigma_{k|k-1} H^T R_k^{-1} \tag{B-8}$$

αντικαθιστώντας τις εξισώσεις (B-7) και (B-6) στην εξίσωση (B-3) βρίσκουμε την επαναληπτική φόρμουλα υπολογισμού για την ακολουθία καταστάσεων.

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(y_k - H\hat{x}_{k|k-1}) \tag{B-9}$$

Μπορούμε επίσης να εκφράσουμε και την συνδιακύμανση $\Sigma_{k|k}$ ως συνέπεια της προηγούμενης $\Sigma_{k|k-1}$.

$\Sigma_{k|k} = E\{\tilde{x}_{k|k}\tilde{x}_{k|k}\} \Leftrightarrow$

$\Sigma_{k|k} = E\{\tilde{x}_{k|k-1}\tilde{x}_{k|k-1}^T\} + K_k E\{\epsilon_k \epsilon_k^T\} K_k^T - K_k E\{\epsilon_k \tilde{x}_{k|k-1}^T\} - E\{\tilde{x}_{k|k-1}\epsilon_k^T\} K_k^T$

Έτσι προκύπτει η εξίσωση:

$$\Sigma_{k|k} = \Sigma_{k|k-1} - K_k R_k K_k^T \tag{B-10}$$

Από τις εξισώσεις (B-9) και (B-8) έχουμε τον επαναληπτικό υπολογισμό της κατάστασης και της μεταβλητότητάς της στη χρονική στιγμή k, με βάση τις τιμές τις προηγούμενης ακριβώς χρονικής στιγμής $k-1$.

Για να υπολογίσουμε την πρόβλεψη για την επόμενη χρονική στιγμή χρησιμοποιούμε την εξίσωση υπολογισμού της κατάστασης:

$$\hat{x}_{k+1} = F\hat{x}_k + w_k \qquad\qquad w_k \sim N(0, Q) \tag{B-11}$$

είναι ευκολο να δούμε πως $\hat{x}_{k+1|k} = F\hat{x}_{k|k} + w_{k|k}$

και απο τη στιγμή που ο θόρυβος ειναι ασυσχέτιστος με οποιαδήποτε αλλη τιμή στο χρόνο, $w_{k|k} = 0$ προκύπτει:

$$\hat{x}_{k+1|k} = F\hat{x}_{k|k} \tag{B-12}$$

Από την εξίσωση (B-10) και την (B-11) μπορούμε να εκτιμήσουμε την απριόρι εκτίμηση λάθους με έναν ακόμη τρόπο:

$$\tilde{x}_{k+1|k} = x_{k+1|k} - \hat{x}_{k+1|k} \Leftrightarrow$$
$$\tilde{x}_{k+1|k} = Fx_{k|k} - F\hat{x}_{k|k} \Leftrightarrow$$
$$\tilde{x}_{k+1|k} = F(x_{k|k} - \hat{x}_{k|k})$$

Τελικά προκύπτει:

$$\tilde{x}_{k+1|k} = F\tilde{x}_{k|k} \tag{B-13}$$

Έτσι η εξίσωση πρόβλεψης της συνδιακύμανσης για την επόμενη χρονική στιγμή προκύπτει:

$$\Sigma_{k+1|k} = FE\{\tilde{x}_{k|k}\tilde{x}_{k|k}^T\}F^T + E\{w_k w_k^T\}$$

δηλαδη,

$$\Sigma_{k+1|k} = F\Sigma_{k|k}F^T + Q \tag{B-14}$$

Συνοψίζοντας οι εξισώσεις του $Kalman$ φίλτρου.

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k e_k$$

$$\hat{x}_{k+1|k} = F\hat{x}_{k|k}$$

$$e_k = y_k - H\hat{x}_{k|k-1}$$

$$K_k = \Sigma_{k|k-1}H^T\Sigma_{e_k}^{-1}$$

$$\Sigma_{e_k} = H\Sigma_{k|k-1}H^T + R$$

$$\Sigma_{k|k} = \Sigma_{k|k-1} - K_k\Sigma_{e_k}K_k^T$$

$$\Sigma_{k+1|k} = F\Sigma_{k|k}F^T + P$$

Πίνακας Β΄.1: Εξισώσεις φιλτραρίσματος

Παράρτημα Γ′

Κανονική Φόρμα

Σύστημα που περιγράφεται από της εξισώσεις:

$$x_{k+1} = Fx_k + \omega_k \qquad \omega_k \sim N(0, Q) \tag{B-1}$$

$$y_k = Hx_k + \upsilon_k \qquad \upsilon_k \sim N(0, R) \tag{B-2}$$

Παραμετροποίηση των πινάκων F και H, με Q και R διαγώνιους.

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & . & . & . & 0 & 0 & 0 \\ 0 & 0 & 1 & . & . & . & 0 & 0 & 0 \\ . & & & & & & & & . \\ . & & & & & & & & . \\ . & & & & & & & & . \\ 0 & 0 & 0 & . & . & . & 0 & 0 & 1 \\ \times & \times & \times & \times & \times & \times & \times & \times & \times \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Παράρτημα Δ΄

Πρόσθετο κοντρόλ Β

Αν έχουμε τον παράγοντα Bu_k ο υπολογισμός της μέγιστης πιθανότητας είναι:

$$J(\mathbf{X},\mathbf{Y},\theta) = -L(\mathbf{X},\mathbf{Y},\theta) = \sum_{k=1}^{N}\Big\{\log|P| + (x_k - Fx_{k-1} - Bu_{k-1})^T P^{-1}(x_k - Fx_{k-1} - Bu_{k-1})\Big\} + \sum_{k=0}^{N}\Big\{\log|R| + (y_k - Hx_k)^T R^{-1}(y_k - Hx_k)\Big\} + \text{ςονστ.} \tag{B-1}$$

Εφαρμόζοντας $element - wise$ τεχνική για τον υπολογισμό τον παραμέτρων προκύπτει:

$$\hat{F}^{i,j} = \frac{\sum_{c=1}^{M}\Big\{(cof(\hat{P}^{i,c}))(\Gamma_4^{c,j})\Big\}}{(cof(\hat{P}^{i,i}))(\Gamma_3^{j,j})} - \frac{\sum_{c=1}^{M}\Big\{(cof(\hat{P}^{i,c}))\sum_{q=1}^{T}\Big\{(\hat{B}^{c,q})(\Gamma_8^{q,j})\Big\}\Big\}}{(cof(\hat{P}^{i,i}))(\Gamma_3^{j,j})} - \frac{\sum_{c=1,c\neq i}^{M}\Big\{(cof(\hat{P}^{i,c}))(\hat{F}^{c,j})(\Gamma_3^{j,j})\Big\}}{(cof(\hat{P}^{i,i}))(\Gamma_3^{j,j})} - \frac{\sum_{c=1}^{M}\Big\{(cof(\hat{P}^{i,c}))\sum_{r=1,r\neq j}^{M}\Big\{(\hat{F}^{c,r})(\Gamma_3^{r,j})\Big\}\Big\}}{(cof(\hat{P}^{i,i}))(\Gamma_3^{j,j})} \tag{B-2}$$

$$\hat{B}^{i,j} = \frac{\sum_{c=1}^{M}\Big\{(cof(\hat{P}^{i,c}))(\Gamma_{10}^{c,j})\Big\}}{(cof(\hat{P}^{i,i}))(\Gamma_9^{j,j})}$$

$$
\begin{aligned}
&-\frac{\sum_{c=1}^{M}\left\{(cof(\hat{P}^{i,c}))\sum_{s=1}^{M}\left\{(\hat{F}^{c,s})(\Gamma_7^{s,j})\right\}\right\}}{(cof(\hat{P}^{i,i}))(\Gamma_9^{j,j})} \\
&-\frac{\sum_{c=1,c\neq i}^{M}\left\{(cof(\hat{P}^{i,c}))(\hat{B}^{c,j})(\Gamma_9^{j,j})\right\}}{(cof(\hat{P}^{i,i}))(\Gamma_9^{j,j})} \\
&-\frac{\sum_{c=1}^{M}\left\{(cof(\hat{P}^{i,c}))\sum_{q=1,q\neq j}^{T}\left\{(\hat{B}^{c,q})(\Gamma_9^{q,j})\right\}\right\}}{(cof(\hat{P}^{i,i}))(\Gamma_9^{j,j})}
\end{aligned}
\tag{B-3}
$$

$$
\begin{aligned}
\hat{P}^{i,j} = (\Gamma_2^{i,j}) &- \sum_{r=1}^{M}(\hat{F}^{i,r})(\Gamma_4^{j,r}) - \sum_{r=1}^{M}(\hat{F}^{j,r})(\Gamma_4^{i,r}) \\
&+ \sum_{c=1}^{M}\sum_{r=1}^{M}(\hat{F}^{i,c})(\hat{F}^{j,r})(\Gamma_3^{c,r}) - \sum_{q=1}^{T}(\hat{B}^{i,q})(\Gamma_{10}^{j,q}) \\
&+ \sum_{q=1}^{T}\sum_{r=1}^{M}(\hat{B}^{i,q})(\hat{F}^{j,r})(\Gamma_8^{q,r}) - \sum_{q=1}^{T}(\hat{B}^{j,q})(\Gamma_{10}^{i,q}) \\
+\sum_{r=1}^{M}\sum_{q=1}^{T}(\hat{F}^{i,r})(\hat{B}^{j,q})(\Gamma_7^{r,q}) &+ \sum_{q=1}^{T}\sum_{p=1}^{T}(\hat{B}^{i,q})(\hat{B}^{j,p})(\Gamma_9^{q,p})
\end{aligned}
\tag{B-4}
$$

$$
\hat{R} = \Gamma_5 - \Gamma_6\Gamma_1^{-1}\Gamma_6^T \tag{B-5}
$$

Χρειαζόμαστε και κάποια πρόσθετα στατιστικά.

$$
\Gamma_7 = \frac{1}{N}\sum_{k=1}^{N} x_{k-1}u_{k-1}^T \tag{B-6}
$$

$$
\Gamma_8 = \frac{1}{N}\sum_{k=1}^{N} u_{k-1}x_{k-1}^T \tag{B-7}
$$

$$
\Gamma_9 = \frac{1}{N}\sum_{k=1}^{N} u_{k-1}u_{k-1}^T \tag{B-8}
$$

$$
\Gamma_{10} = \frac{1}{N}\sum_{k=1}^{N} x_k u_{k-1}^T \tag{B-9}
$$

Τα στατιστικά για τον υπολογισμό τους για κάποια επανάληψη p χρειάζονται τους υπολογισμούς:

$$
E_{\theta^{(p)}}\{x_{k-1}u_{k-1}^T|\mathbf{Y}\} = \hat{x}_{k-1|N}u_{k-1}^T \tag{B-10}
$$

$$E_{\theta^{(p)}}\{u_{k-1}x_{k-1}^T|\mathbf{Y}\} = u_{k-1}\hat{x}_{k-1|N}^T \quad \text{(B-11)}$$

$$E_{\theta^{(p)}}\{u_{k-1}u_{k-1}^T|\mathbf{Y}\} = u_{k-1}u_{k-1}^T \quad \text{(B-12)}$$

$$E_{\theta^{(p)}}\{x_k u_{k-1}^T|\mathbf{Y}\} = \hat{x}_{k|N}u_{k-1}^T. \quad \text{(B-13)}$$

Η Εμπρός - Πίσω διαδικασία είναι η ίδια με την μόνη διαφορά στον υπολογισμό του $\hat{x}_{k+1|k} = J$ όπου $J = [F\ B]$. Η εξαγωγή τον εξισώσεων προέκυψαν από την προσπάθεια του Χριστου Κόνιαρη.

Βιβλιογραφία

[1] V. Digalakis, "Segment-based stochastic models of spectral dynamics for continuous speech recognition", *Ph.D. Thesis*, Boston University, Jan. 1992.

[2] L. Ljung, "System Identification: Theory for the User (2nd Edition)" *Prentice Hall PTR* ISBN: 0136566952, 2nd edition December, 1998.

[3] J. Frankel, "Linear dynamic models for automatic speech recognition", *Ph.D. Thesis*, The Centre for Speech Technology Research, Edinburgh University, 2003.

[4] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "ML Estimation of a Stochastic Linear System with the EM Algorithm and Its Application to Speech Recognition", *IEEE Transactions on Speech and Audio Processing* , vol. 1, no. 4, Oct. 1993.

[5] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems", *Trans. ASME, Series D, J. Basic Eng.*, Vol. 82, pp. 35–45, March 1960.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin , "Maximum Likelihood Estimation from Incomplete Data", *Journal of the Royal Statistical Society (B)*, Vol. 39, No.1, pp. 1-38, 1977.

[7] M. Ostendorf, V. Digalakis, O.A. Kimball, "From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 4, No. 5, pp.360-378, 1996.

[8] H. E. Rauch, F. Tung and C. T. Striebel, "Maximum Likelihood Estimates of Linear Dynamic Systems", *AIAA Journal*, Vol. 3, no. 8, pp. 1445–1450, August 1965.

[9] S. Roweis and Z. Ghahramani, "A unified review of the linear Gaussian models", *Neural Computation*, vol. 11, No. 2, 1999.

[10] P. E. Caines, "Linear Stochastic Systems", *John Wiley Sons*, 1998

[11] H.L. Van Trees, "Detection, Estimation, and Modulation Theory, Part I.", *NewYork: Wiley*, 1968.

[12] Βασίλης Διγαλάκης, "Σημειώσεις του Μαθήματος: Εισαγωγή στην Επεξεργασία Φωνής "

[13] C. Koniaris, "Estimation of General Identifiale State-Space Models", *M.Sc. Thesis*, Department of ECE, Technical University of Crete, Aug. 2006.

[14] H.G. Hirsch and D. Pearce, "The aurora experimentalframework for the performance evaluation of speech recognition systems under noise conditions", in Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, 2000.

[15] A. Rosti and M. Gales, "Generalised linear gaussian models", Tech. Rep., Engineering, Cambridge University, 2001.

[16] Huang, Xuedong, "Spoken language processing: a guide to theory, algorithm, and system development", *Prentice Hall PTR* , ISBN: 0130226165, 2001.

[17] Steven M. Kay, "Fundamentals of Statistical Signal Processing: Estimation Theory", *Prentice Hall PTR Upper Saddle River, NJ 07458.*

[18] Athanasios Papoulis, "Propability, Random Variables, and Stochastic Processes",*McGraw-Hill, Inc*, ISBN: 960-7219-34-1, 1994.

[19] Stuart Russell, Peter Norvig, "Artificial Intelligence: A Modern Approach, Second Edition", *Prentice Hall by Pearson Education, Inc*, ISBN: 960-209-774-4, 2003

[20] Simon Haykin, "Kalman Filtering and Neural Networks", *John Wiley & Sons, Inc*, ISBN: 0-471-22154-6, 2001.

[21] Greg Welch, Gary Bishop "An Introduction to the Kalman Filter", Department of Computer Science University of North Carolina at Chapel Hill, 2004.

[22] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, Vol. 77, No. 2, 1989.

[23] Kaare Brandt Petersen, Michael Syskind Pedersen, "The Matrix Cookbook", Petersen & Pedersen, 2005.

[24] Max Welling, "EM-algorithm", California Institute of Technology 136-93 Pasadena, CA 91125.

[25] "http://www.cs.berkley.edu/ murphyk/Bayes/bayes.html".

[26] R. Duda, D. Stork, P. Hart, "Pattern Classification", John Wiley & Sons, 2000.

[27] X. D. Huang, Y. Ariki, Mervyn A. Jack "Hidden Markov Models for Speech Recognition (Edinburgh Information Technology Series, 7)", *Edinburgh Univ Pr*, ISBN: 0748601627, February, 1991.

ΕΙΣΑΓΩΓΗ

Η ταχεία εξέλιξη της τεχνολογίας έχει δημιουργήσει μεγάλες δυνατότητες στον έλεγχο, στην αποθήκευση και στη μετάδοση των πληροφοριών. Η επικοινωνία όμως μεταξύ ανθρώπου και μηχανής απαιτεί κάποιο μέσω ανάθεσης των εντολών για την αποκομιδή των πληροφοριών. Μεγάλο επίτευγμα θα ήταν αν η μηχανή μπορέσει να αποκτήσει την ιδιότητα κατανόησης της γλώσσας των ανθρώπων, με σκοπό την ανάθεση των εντολών χωρίς την μεσολάβηση κάποιου μέσου. Ο κλάδος της Αναγνώρισης Φωνής (Speech Recognition) έχει κάνει αρκετά βήματα για την επίτευξη του στόχου αυτού. Πραγματικές εφαρμογές έχουν ήδη αρχίσει να εμφανίζονται δείχνοντας την σπουδαιότητα ως κλάδος στην υπηρεσία του ανθρώπου.

Μεγάλο αντικείμενο έρευνας στον κλάδο της Αναγνώρισης Φωνής είναι η ακουστική μοντελοποίηση. Η ακουστική μοντελοποίηση προσπαθεί να χειριστεί το πρόβλημα της μεταβλητότητας που υπάρχει στο σήμα της φωνής. Για την ακουστική μοντελοποίηση υπάρχουν διάφοροι μέθοδοι υλοποίησης, όπως τα Κρυφά Μαρκοβιανά μοντέλα, τα νευρωνικά δίκτυα και τα γραμμικά δυναμικά μοντέλα. Η εργασία αυτή ασχολείται με το πρόβλημα αυτό, χρησιμοποιώντας την μέθοδο των γραμμικών δυναμικών μοντέλων.

Στην εργασία αυτή προτείνεται μια νέα παραμετροποίηση των δομικών στοιχείων του δυναμικού μοντέλου, και ένας νέος τρόπος υπολογισμού τους. Οι νέες αυτές προτάσεις έδωσαν καλά αποτελέσματα, με αποτέλεσμα τη δημιουργία εργασίας για τη συμμετοχή στο 32ο διεθνές συνέδριο ακουστικής επεξεργασίας, επεξεργασίας φωνής και επεξεργασίας σήματος της υπο την αιγίδα της IEEE.

Τέλος, η εργασία περιέχει τα ακόλουθα. Στο 1ο Κεφάλαιο αναλύουμε από τι αποτελείται ένα σύγχρονο αυτόματο σύστημα αναγνώρισης της φωνής. Στο 2ο Κεφάλαιο παρουσιάζουμε τις μεθόδους υλοποίησης του ακουστικού μοντέλου, στο 3ο Κεφάλαιο αναλύεται το γραμμικό δυναμικό σύστημα και

οι διάφορες έννοιες που θα χρησιμοποιήσουμε κατά την εφαρμογή του ως ακουστικό μοντέλο. Στο 4ο Κεφάλαιο η εφαρμογή των γραμμικών δυναμικών μοντέλων και η διαδικασία εκπαίδευσης, στο 5ο Κεφάλαιο γίνεται η αποτίμηση των μοντέλων με την μέθοδο της ταξινόμησης. Τέλος στο 6ο Κεφάλαιο δίνονται τα συμπεράσματά μας από την εργασία και τον προτάσεις για μελλοντική εργασία.

ESTIMATION OF GENERAL IDENTIFIABLE LINEAR DYNAMIC MODELS WITH AN APPLICATION IN SPEECH RECOGNITION

G. Tsontzos, V. Diakoloukas, Ch. Koniaris and V. Digalakis

Dept. of Electronics & Computer Engineering
Technical University of Crete, GR-73100 Chania, Greece
{gtsntzs,vdiak,chkoniaris,vas}@telecom.tuc.gr

ABSTRACT

Although Hidden Markov Models (HMMs) provide a relatively efficient modeling framework for speech recognition, they suffer from several shortcomings which set upper bounds in the performance that can be achieved. Alternatively, linear dynamic models (LDM) can be used to model speech segments. Several implementations of LDM have been proposed in the literature. However, all had a restricted structure to satisfy identifiability constraints. In this paper, we relax all these constraints and use a general, canonical form for a linear state-space system that guarantees identifiability for arbitrary state and observation vector dimensions. For this system, we present a novel, element-wise Maximum Likelihood (ML) estimation method. Classification experiments on the AURORA2 speech database show performance gains compared to HMMs, particularly on highly noisy conditions.

Index Terms— Speech Recognition, Modeling, Identification

1. INTRODUCTION

Hidden Markov Models (HMMs) dominate in today's speech recognition engines. This is primarily attributed to their ability to efficiently model the time varying statistical characteristics of the speech signal through a set of discrete states. However, they still posses many modelling inadequacies that derive from the numerous assumptions that are made to simplify the speech recognition problem. For instance, dynamic information in HMMs is included through the time-derivatives in the observation vector under the false frame-independence assumption and the spatial correlation of the observation vector is ignored when diagonal covariance matrices are considered.

This work is motivated from our belief that these assumptions set upper limits in the progress that can be made when using HMMs in speech recognition. In an effort to improve robustness, particularly under noisy conditions, we examine new modeling schemes that can explicitly model time and spatial correlations such as the linear dynamical models (LDM). LDMs were first proposed to be used for speech recognition in [1]. They characterize complete speech segments such as words, phonemes or sub-phoneme units with a linear state evolution process and a linear observation process. Thus, they can be seen as a variation of segment-based modeling which, in turn, can be considered as a generalization of the HMMs with a continuous state-space instead of a discrete one[2].

There are several variations of the LDMs that can be found in the literature. In [1] LDMs were used to obtain a smoothed realization of a Gauss-Markov model. In [3] and [4] several statistical modeling techniques such as factor analysis (FA) and principle component analysis (PCA) are presented as special cases of a general LDM. Other variations are also discussed in [5]. In all cases, several modeling constraints were applied in an effort to obtain good system convergence, stability and identifiability. However, these constraints alter the properties of the model, and diminish the benefits of the general system architecture.

In this paper, we introduce a generalized linear dynamic system in an identifiable canonical form. The system is a multivariate state-space linear dynamic model which follows the identifiable form that was proposed by Ljung [6]. We begin by introducing the linear system and its parametric structure. We describe our novel element-wise estimation method based on the Expectation-Maximization (EM) algorithm. Finally, we present classification results on the AURORA2 speech database.

2. THE LINEAR DYNAMIC SYSTEM

The LDM is described from the following pair of equations

$$x_{k+1} = Fx_k + w_k \tag{1}$$

$$y_k = Hx_k + v_k \tag{2}$$

where the state x_k at time k is a $(n \times 1)$ vector, the observation y_k is $(m \times 1)$ and w_k, v_k are uncorrelated, zero-mean Gaussian vectors with covariances

$$E\{w_k w_l^T\} = P\delta_{kl} \tag{3}$$

$$E\{v_k v_l^T\} = R\delta_{kl} \tag{4}$$

This work was partially supported by the EU-IST FP6 research project HIWIRE.

In the above equation δ_{kl} denotes the Kronecker delta and T denotes the transpose of a matrix. The initial state x_0 is Gaussian with known mean and covariance μ_0, Σ_0. Equation (1) describes the state dynamics, while (2) shows a prediction of the observation based on the state estimation.

The parametric structure of our multivariate state-space model has the following identifiable canonical form for the case in which x_k is a 5×1 vector and the observation vector y_k is a 3×1 vector.

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ 0 & 0 & 0 & 0 & 1 \\ \times & \times & \times & \times & \times \end{bmatrix} \tag{5}$$

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{6}$$

The number of rows with $\times$' s in F represents the free parameters of the matrix and equals the size of the output vector m. The ones in matrix H are equal to the number of the rows in F that are filled with free parameters, and their position is related to the location of these rows in F.

To construct the form of the state transition matrix F we follow the process described in [6]. First, we set its elements along the superdiagonal equal to one and the remaining elements are zeroed. Then, we choose arbitrarily the m row numbers r_i to be filled with free parameters, where $i = 1, ..., m$. There is only one constraint, that $r_m = n$, where m denotes the dimension of the observation and n the dimension of the state vector. In addition, we set $r_0 = 0$.

The observation matrix H is then constructed as follows. First, we define H to be $m \times n$ in size and filled with zeros. Then we set each row $i = 1, ..., m$ of the H matrix to have a one in column $c_i = r_{i-1} + 1$. For instance, for the example shown in (5) and (6) we get:
$r_1 = 2 \Rightarrow c_1 = r_{1-1} + 1 = r_0 + 1 = 1$
$r_2 = 3 \Rightarrow c_2 = r_{2-1} + 1 = r_1 + 1 = 3$
$r_3 = 5 \Rightarrow c_3 = r_{3-1} + 1 = r_2 + 1 = 4$.
Hence, the observation matrix H will have ones in columns 1, 3 and 4 for its rows 1, 2 and 3, respectively.

Ljung [6] proves that the above canonical form is identifiable if and only if it is also controllable. Furthermore, this canonical form does not impose any loss of generality in the LDM, that is, any state-space system described by equations 1 and 2 can be transformed to have the structure of equations 5 and 6.

3. ELEMENT-WISE ESTIMATION WITH EM

The matrices of the LDM presented in section 2 contain, by-construction, free parameters at very specific positions. An estimation algorithm for linear state-space systems that is based on the Expectation-Maximization (EM) algorithm was introduced in [1]. This algorithm assumed that all matrices $\theta = F, H, P, R$ are filled with free parameters. In our case, however, the free parameters of the system are located in specific position, hence the estimation must be performed in an element-wise fashion. Given the observations $\mathbf{Y} = [y_0 \ \dots \ y_N]$ and the state vectors $\mathbf{X} = [x_0 \ \dots \ x_N]$, the ML estimates of θ are obtained by minimizing the quantity:

$$\begin{aligned} L(\mathbf{X}, \mathbf{Y}, \theta) = \\ -\sum_{k=1}^{N} \Big\{ \log|P| + (x_k - Fx_{k-1})^T P^{-1} (x_k - Fx_{k-1}) \Big\} \\ -\sum_{k=0}^{N} \Big\{ \log|R| + (y_k - Hx_k)^T R^{-1} (y_k - Hx_k) \Big\} \end{aligned}$$

It can be shown that the estimates of the system's parameters are given by:

$$\begin{aligned} \hat{F}_{ij} = & \frac{\sum_{c=1}^{M} \Big\{ (cof(\hat{P}_{ic}))(S_{cj}^{(4)}) \Big\}}{(cof(\hat{P}_{ii}))(S_{jj}^{(3)})} \\ & - \frac{\sum_{c=1, c\neq i}^{M} \Big\{ (cof(\hat{P}_{ic}))(\hat{F}_{cj})(S_{jj}^{(3)}) \Big\}}{(cof(\hat{P}_{ii}))(S_{jj}^{(3)})} \\ & - \frac{\sum_{c=1}^{M} \Big\{ (cof(\hat{P}_{ic})) \sum_{r=1, r\neq j}^{M} \Big\{ (\hat{F}_{cr})(S_{rj}^{(3)}) \Big\} \Big\}}{(cof(\hat{P}_{ii}))(S_{jj}^{(3)})} \end{aligned} \tag{7}$$

$$\begin{aligned} \hat{P}_{ij} = (S_{ij}^{(2)}) - \sum_{r=1}^{M} (\hat{F}_{ir})(S_{jr}^{(4)}) - \sum_{r=1}^{M} (\hat{F}_{jr})(S_{ir}^{(4)}) \\ + \sum_{c=1}^{M} \sum_{r=1}^{M} (\hat{F}_{ic})(\hat{F}_{jr})(S_{cr}^{(3)}) \end{aligned} \tag{8}$$

$$\hat{R} = S^{(5)} - S^{(6)} (S^{(1)})^{-1} (S^{(6)})^T \tag{9}$$

where $cof(\hat{P}_{ic})$ is the cofactor of the element $\hat{P}_{ic}$ of the covariance $\hat{P}$. Index i denotes the $i-th$ row of a matrix, and j denotes the $j-th$ column. The sufficient statistics that in-

volved in the previous equations are given by[1]

$$S^{(1)} = \frac{1}{N+1}\sum_{k=0}^{N} x_k x_k^T \quad (10)$$

$$S^{(2)} = \frac{1}{N}\sum_{k=1}^{N} x_k x_k^T \quad (11)$$

$$S^{(3)} = \frac{1}{N}\sum_{k=1}^{N} x_{k-1} x_{k-1}^T \quad (12)$$

$$S^{(4)} = \frac{1}{N}\sum_{k=1}^{N} x_k x_{k-1}^T \quad (13)$$

$$S^{(5)} = \frac{1}{N+1}\sum_{k=0}^{N} y_k y_k^T \quad (14)$$

$$S^{(6)} = \frac{1}{N+1}\sum_{k=0}^{N} y_k x_k^T. \quad (15)$$

The statistics shown above require the following quantities at each iteration p:

$$E_{\theta^{(p)}}\{y_k x_k^T|\mathbf{Y}\} = y_k \hat{x}_{k|N} \quad (16)$$

$$E_{\theta^{(p)}}\{y_k y_k^T|\mathbf{Y}\} = y_k y_k^T \quad (17)$$

$$E_{\theta^{(p)}}\{x_k x_{k-1}^T|\mathbf{Y}\} = \Sigma_{k,k-1|N} + \hat{x}_{k|N}\hat{x}_{k-1|N}^T \quad (18)$$

$$E_{\theta^{(p)}}\{x_k x_k^T|\mathbf{Y}\} = \Sigma_{k|N} + \hat{x}_{k|N}\hat{x}_{k|N}^T. \quad (19)$$

Equations (7) through (9) form the Maximization step of the EM algorithm. For the Expectation step of the EM algorithm we need to compute the required statistics, and we use the fixed interval smoothing form of the Kalman filter (RTS smoother) [7]. It consists of a backward pass that follows the standard Kalman filter forward recursions [8]. In addition, we compute also the cross-covariances proposed by Digalakis [1] in both the forward and the backward pass.

4. APPLICATION TO SPEECH RECOGNITION

A straightforward way to model speech units using LDMs is to train separate segment-specific models, each one corresponding to a sub-word or sub-phoneme unit. The correlation between consecutive frames within the same segment is modelled with the same set of parameters. Furthermore, the inter-segment correlation is also captured since the initial state estimate of a segment derives from the last state estimate of the previous segment. The process is also illustrated in Figure 1 for a 4 segment example.

During classification, each model segment is classified based on the log-likelihood computed by:

$$L(\mathbf{Y},\theta) = -\sum_{k=0}^{N}\left\{\log|\Sigma_{e_k}(\theta)| + e_k^T(\theta)\Sigma_{e_k}^{-1}(\theta)e_k^T(\theta)\right\} + C$$

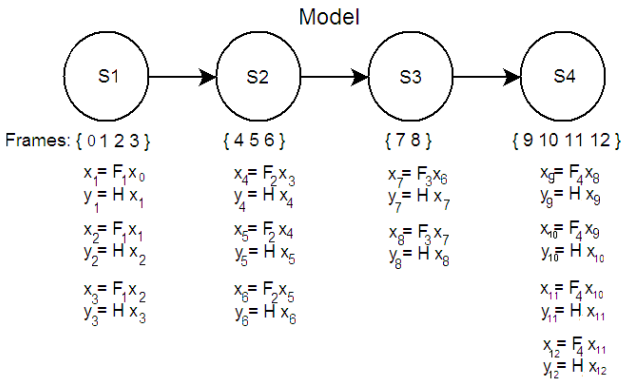


Fig. 1. Example of an LDM with 4 segments.

where $e_k^T(\theta)$, $\Sigma_{e_k}(\theta)$ is the prediction error and its covariance obtained from the Kalman filter equations and C is a constant.

5. EXPERIMENTS

We have performed a series of word-classification experiments in order to validate our LDM system for speech recognition and evaluate the estimation algorithm. In specific, we used the AURORA2 speech database[9], which is a connected digit corpus based on TIDIGITS, downsampled to 8KHz and with several types of noise artificially added at several SNRs. The front-end uses a total of 13 Mel-warped cepstral coefficients plus energy. In some experiments we also augmented the observation vector with the first (δ) and second order derivatives ($\delta\delta$).

We used 11 word-models corresponding to the words in the AURORA2 corpus (digits 1 to 9, zero and oh). Each word-model has a number of time-invariant regions (segments) ranging from 2 to 8, depending on the phonetic transcription of each word. Table 1 shows the number of regions for each word-model that we considered.

one	two	three	four	five	six
6	4	6	6	6	4

seven	eight	nine	oh	zero
8	4	6	2	6

Table 1. Number of regions for each word-model

The first issue in implementing the dynamical system is the dimensionality of the state-space. Based on the general canonical forms of the LDM that we examine, the size of the state-vector can be equal or larger than the size of the observation vector. When the state and observation vectors are at equal size, the observation matrix becomes the identity matrix and the observation vector is just a noisy version of the state vector. Even in this case, our scheme relaxes the constraints

of other approaches (i.e. in [1]).

Another important issue is the initialization of system parameters. The noise covariance matrices are initialized randomly, while the initial state-transition matrices, and the covariance of the initial state x_0 are directly estimated from the observations.

As far as the classification is concerned, at this moment we do not perform any search over all possible segmentations, but we keep the true word-boundaries produced by an HMM fixed. We do search, however, over all possible word histories given the segmentation. To speed-up the classification process we apply a suboptimum search and pruning algorithm which keeps the 11 most probable word-histories for each word in the sentence.

For our experiments, we used a clean training set consisting of 104 gender-balanced speakers and 8444 sentences. The evaluation was done on a separate test set defined as the AURORA2-A test set, with subway additive noise at several SNRs, which consisted of 1000 sentences from the training speakers. Table 2 summarizes the classification performance of the LDM for several SNR values. As can be seen, appending the derivatives in the MFCCs results in performance gains which increase as the noise level rises.

AURORA 2/ Subway	LDM	
	Mfcc,energy	$+\delta+\delta\delta$
clean	97.53%	97.61%
SNR20	93.23%	95.12%
SNR15	87.91%	91.13%
SNR10	76.29%	82.69%
SNR5	54.87%	63.56%

Table 2. Word-classification performance of the LDM system

AURORA 2/ Subway	HMM	
	Mfcc,energy	$+\delta$ $+\delta\delta$
clean	97.19%	97.57%
SNR20	90.91%	95.71%
SNR15	80.09%	91.76%
SNR10	57.68%	81.93%
SNR5	36.01%	64.24%

Table 3. Performance of an HMM system

To compare the performance of our system to HMMs, we also performed a set of classification experiments using the standard HTK configuration. Each word was modelled as a 16-state continuous density HMM with a mixture of 3 Gaussian components associate in each state. The front-end configuration and the word-boundaries were the same as with the LDM. The recognition accuracy of the HMM is shown in Table 3. Without derivatives, the LDM outperforms significantly the HMM, especially as the SNR level decreases. When derivatives are used for both models, their performance is similar.

6. CONCLUSIONS

In this paper, we presented the application of linear dynamic models with general, canonical forms for their parameters in speech recognition and we showed a novel and efficient element-wise Maximum Likelihood estimation. We evaluated our scheme with a series of classification experiments on the AURORA2 speech database. Since we have now introduced a methodology to use general identifiable forms of state-space systems in speech recognition, we plan to investigate in the future several combinations of state and observation vector dimensions.

7. REFERENCES

[1] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the em algorithm and its application to speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, Oct. 1993.

[2] M. Ostendorf, V. Digalakis, and O.A. Kimball, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 360–378, 1996.

[3] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural Computation*, vol. 11, 1999.

[4] A. Rosti and M. Gales, "Generalised linear gaussian models," Tech. Rep., Engineering, Cambridge University, 2001.

[5] J. Frankel and S. King, "Speech recognition using linear dynamic models," *IEEE Transactions on Speech and Audio Processing*, January 2007.

[6] L. Ljung, *System Identification: Theory for the User (2nd Edition)*, Prentice Hall PTR, 1998.

[7] H.E. Rauch, F. Tung, and C.T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, pp. 1445–1450, August 1965.

[8] R.E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, Series D, J. Basic Eng.*, vol. 82, pp. 35–45, March 1960.

[9] H.G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noise conditions," in *Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, 2000.

Ευχαριστίες

Πρώτα απ΄ όλα ευχαριστώ τους γονείς μου

Ακόμη θα ήθελα να ευχαριστήσω τον κ. Βασίλη Διγαλακη για της γνώσεις που μου πρόσφερε, το Βασίλη Διακολουκά για τις ατέλειωτες ώρες συζήτησης και την τεράστια βοήθεια που μου έδωσε, και τέλος τον Χρίστο Κόνιαρη για την συνεργασία που είχαμε, αποτέλεσμα της οποίας είναι αυτή η διπλωματική.