ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Σύνθεση Φωνής με Στατιστικά Μοντέλα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Σιδέρη Παναγιώτη

Επιβλέπων: Βασίλειος Διγαλάκης
Καθηγητής Πολυτεχνείου Κρήτης

Χανιά Μάρτιος 2016

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την εξεταστική μου επιτροπή και πιο συγκεκριμένα τον κ.Διγαλάκη Βασίλη, τον κ. Λαγουδάκη Μιχαήλ και τον κ. Ιωάννη Στυλιανου. Επίσης τον κ.Διακολουκά Βασίλη και Τσιάρα Βασίλη για την άμεση συνεργασία που είχαμε αλλά και την βοήθεια που μου προσέφεραν. Τέλος ιδιαίτερες ευχαριστίες στην οικογένειά μου για την στήριξη όλα αυτά τα χρόνια αλλά και στους φίλους μου Δελή Γεώργιο , Βαρσαμίδη Χαρίση και Σουρίλα Αλκιβιάδη.

Περίληψη

Στην εργασία αυτή διερευνούμε τη χρήση των Γραμμικών Δυναμικών Μοντέλων (LDMs) στη σύνθεση φωνής. Υπάρχουν διάφορες οικογένειες τεχνικών πάνω στη σύνθεση φωνής, εκ των οποίων, μία από τις πιο δημοφιλείς είναι τα στατιστικά παραμετρικά μοντέλα (SPSS) τα οποία και περιγράφουμε. Στην οικογένεια των SPSS ανήκουν και τα Γραμμικά Δυναμικά Μοντέλα στα οποία και επικεντρωθήκαμε λόγω αξιοπιστίας και δυνατοτήτων. Καταφέραμε να αποδείξουμε ότι η σύνθεση φωνής με αυτό το μοντέλο είναι εφικτή και αξιολογήσαμε την ποιότητα της συνθετικής φωνής που παράξαμε, μέσω αντικειμενικών μετρικών. Επιπλέον για την υλοποίηση της εργασίας αυτής χρησιμοποιήσαμε το Straight vocoder, ο οποίος αποτελεί τεχνολογία αιχμής πάνω στο πεδίο της μελέτης μας. Τέλος παραθέσαμε τα συμπεράσματα σχετικά με την ποιότητα του μοντέλου αλλά και τα προβλήματα που αντιμετωπίσαμε στην διαδικασία αυτή.

Περιεχόμενα

Αντιστοίχιση Όρων – Ακρωνύμια

TTS	Text-to-Speech
POS	Part-of-Speech
G2P	Graphic-to-Phoneme
NLP	Natural Language Processing
SPSS	Statistical Parametric Speech Synthess
ML	Maximum Likelihood
HMM	Hidden Markov Model
EM	Expectation Maximization
LDM	Linear Dynamic Model
ARHMM	Autoregressive Hidden Markov Model
MGC	Mel Generalized Cepstrum
GV	Global Variance
C/N	Carrier to Noise

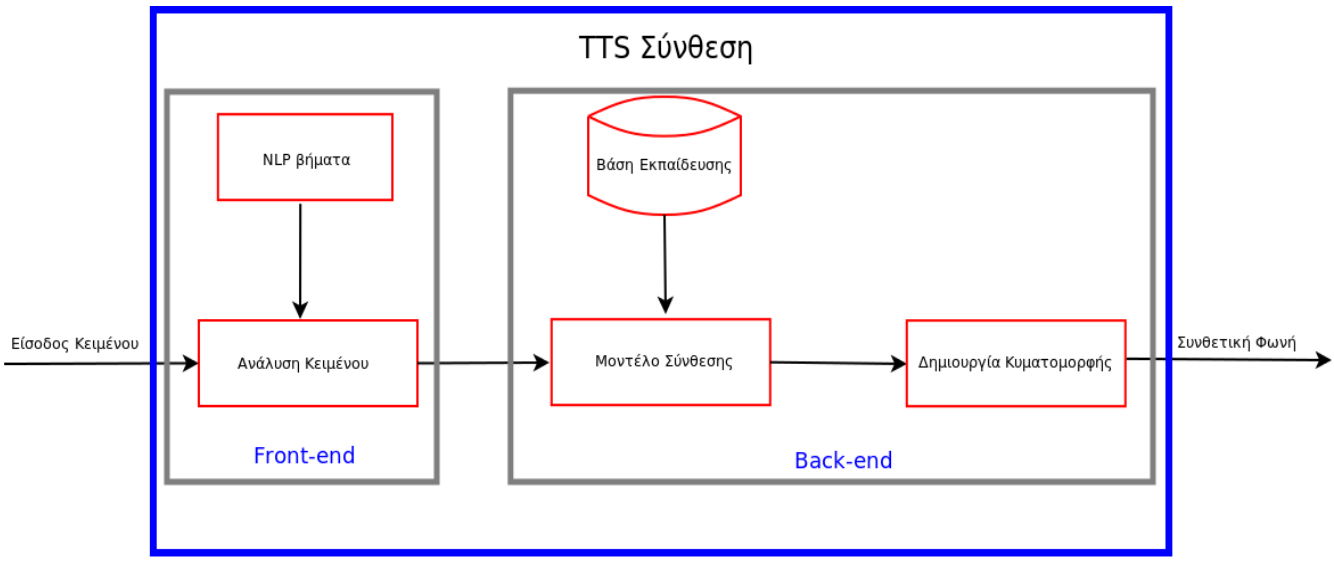
Κεφάλαιο 1

Εισαγωγή

1.1 Text-to-speech σύνθεση (TTS Synthesis)

Ένα σύστημα TTS έχει σαν στόχο να μετατρέψει ένα κείμενο σε φυσική ομιλία. Αυτό επιτυγχάνεται προσομοιώνοντας την διαδικασία παραγωγής ανθρώπινης ομιλίας. Μπορούμε να πούμε ότι η σύνθεση φωνής με το παραπάνω σύστημα αποτελεί, ουσιαστικά, την αντιστοίχιση μίας ακολουθίας διακριτών συμβόλων (το κείμενο που θέλουμε να συνθέσουμε) σε μία συνεχή ακολουθία (κυματομορφή που παράγεται).

Τα TTS συστήματα έχουν χρησιμοποιηθεί σε διάφορες εφαρμογές στις μέρες μας. Μπορούν να αποτελέσουν σπουδαίο εργαλείο για άτομα με ειδικές ανάγκες τόσο σε επίπεδο εκμάθησης όσο και επικοινωνίας. Επίσης τα τελευταία χρόνια έχουν ενσωματωθεί σε κινητά τηλέφωνα αλλά και σε υπολογιστές διευκολύνοντας την επικοινωνία ανθρώπου-μηχανής καθώς εισάγεται φυσικότητα τόσο στον χειρισμό των συσκευών από τους χρήστες όσο και στην απόκριση των μηχανών πίσω σε αυτούς. Στα TTS συστήματα ανήκουν τα στατιστικά παραμετρικά μοντέλα σύνθεσης φωνής αλλά και η συνδετική σύνθεση φωνής (concatenative speech synthesis).



Γράφημα 1: Διάγραμμα TTS σύνθεσης

Ένα TTS σύστημα απαρτίζεται από δύο μέρη, την ανάλυση του κειμένου και την σύνθεση ομιλίας. Η ανάλυση κειμένου καλείται front-end του συστήματος και είναι αντιστοίχιση μίας διακριτής ακολουθίας, που είναι το κείμενο που θέλουμε να συνθέσουμε, σε μία άλλη διακριτή ακολουθία, που είναι το ίδιο κείμενο έπειτα από γλωσσική ανάλυση. Η γλωσσική ανάλυση περιλαμβάνει μια σειρά από NLP (natural language processing) βήματα, όπως τμηματοποίηση λέξης (word segmentation), κανονικοποίηση κειμένου (text normalization), POS σημείωση (part-of-speech tagging) και G2P μετατροπή (graphic to phoneme conversion).

Όσο αφορά τη σύνθεση ομιλίας, αυτή περιλαμβάνει το μοντέλο σύνθεσης, στο οποίο γίνεται η πρόβλεψη της προσωδίας (prosodic prediction) και η δημιουργία της κυματομορφής (speech waveform generation). Στο βήμα αυτό το κείμενο που έχει προκύψει από γλωσσική ανάλυση μετατρέπεται σε ομιλία, δηλαδή η διακριτή ακολουθία συμβόλων αντιστοιχίζεται σε μία συνεχή, και καλείται back-end του συστήματος. Είναι βασικό ότι και το front-end αλλά και το back-end είναι εξίσου σημαντικά για την παραγωγή ομιλίας υψηλής ποιότητας.

1.2 Συνδετική Σύνθεση Φωνής (Concatenative Speech Synthesis)

Στην ενότητα αυτή περιγράφουμε το βασικό μοντέλο πάνω στην συνδετική σύνθεση φωνής που είναι το unit-selection. Το unit-selection αποτελεί τεχνολογία αιχμής καθώς η ποιότητα της συνθετικής φωνής είναι ιδιαίτερα υψηλή. Βασίζεται στην ιδέα ότι μπορούμε να συνθέσουμε φωνή μέσω της κατάλληλης επιλογής τμημάτων της (υπο)λέξης (sub-word units) από μια βάση δεδομένων που περιλαμβάνει ηχογραφήσεις φυσικής ομιλίας. Υπάρχουν δύο βασικές τεχνικές στη σύνθεση με unit-selection οι οποίες βασίζονται στον υπολογισμό του *target cost* $C^{(t)}$ και του *concatenation cost* $C^{(c)}$. Το *target cost* αντιπροσωπεύει το πόσο καλά ένα υποψήφιο τμήμα ταιριάζει με το απαιτούμενο ενώ το *concatenation cost* αντιπροσωπεύει το πόσο καλά δύο διαδοχικά επιλεγμένα τμήματα συνδυάζονται. Ο υπολογισμός αυτών γίνεται από τις παρακάτω σχέσεις:

$$C^{(t)}(t_i, u_i) = \sum_{j=1}^{p} w_j^{(t)} C_j^{(t)}(t_i, u_i) \qquad \text{(A1)}$$

$$C^{(c)}(u_{i-1}, u_i) = \sum_{k=1}^{q} w_k^{(c)} C_k^{(c)}(u_{i-1}, u_i) \qquad \text{(A2)}$$

όπου u_i το υποψήφιο τμήμα, t_i το απαιτούμενο τμήμα, j τα φωνητικά και προσωδιακά χαρακτηριστικά αντίστοιχα και k τα φασματικά και ακουστικά χαρακτηριστικά (spectral and acoustic features).

Στόχος είναι η εύρεση της κατάλληλης συμβολοσειράς $u_{1:n}$

$$u_{1:n} = \{u_1, \ldots, u_n\} \quad \text{(A3)}$$

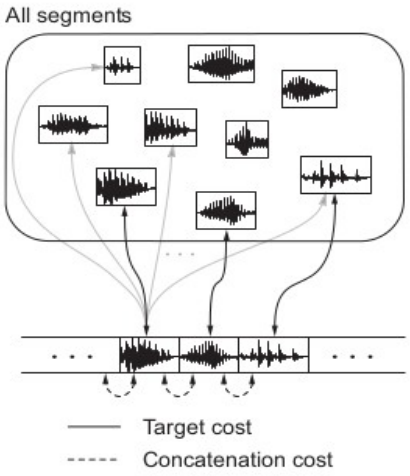
από τη βάση δεδομένων η οποία θα ελαχιστοποιεί το συνολικό κόστος $C(t_{1:n}, u_{1:n})$.

$$\hat{u}_{1:n} = \arg\min_{u_{1:n}} \{C(t_{1:n}, u_{1:n})\} \quad \text{(A4)}$$

Όπου:

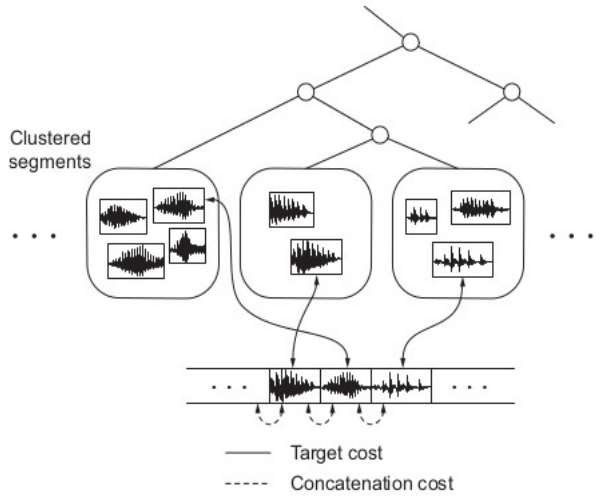
$$C(t_{1:n}, u_{1:n}) = \sum_{i=1}^{n} C^{(t)}(t_i, u_i) + \sum_{i=2}^{n} C^{(c)}(u_{i-1}, u_i) \quad \text{(A5)}$$

Στο γράφημα 2 απεικονίζεται η unit-selection σύνθεση χωρίς ομαδοποίηση. Σκοπός είναι η δημιουργία μιας ακολουθίας (η οποία αποτελείται από τμήματα της βάσης δεδομένων), που ανταποκρίνεται στην αυθεντική κυματομορφή της πρότασης που θέλουμε να συνθέσουμε. Για κάθε απαιτούμενο τμήμα (unit) επιλέγονται από μία βάση δεδομένων ορισμένα υποψήφια τμήματα, αυτά με το χαμηλότερο *target-cost* (ενιαίες γραμμές) ενώ παράλληλα εξετάζεται το *concatenation-cost* μεταξύ 2 διαδοχικών υποψήφιων τμημάτων (διακεκομμένες γραμμές). Τελικά επιλέγεται η ακολουθία που ελαχιστοποιεί το συνολικό κόστος.



Γράφημα 2 : Unit-selection χωρίς ομαδοποίηση

Στο γράφημα 3 απεικονίζεται η unit-selection σύνθεση με ομαδοποίηση. Η διαφορά της πρώτης με τη δεύτερη τεχνική είναι ότι στην τελευταία γίνεται ομαδοποίηση των παραπλήσιων τμημάτων. Αυτό μας βοηθάει στον εκ των προτέρων υπολογισμό του target cost για καθένα από τα ομαδοποιημένα τμήματα. Για να επιτευχθεί η διαδικασία αυτή εξετάζονται με τη βοήθεια ερωτήσεων ποια χαρακτηριστικά παίζουν ρόλο στη συγκεκριμένη χρονική στιγμή της σύνθεσης . Ομοίως με την πρώτη τεχνική με ενιαία γραμμή συμβολίζεται το *target-cost* και με διακεκομμένη το *concatenation-cost* .



Γράφημα 3: Unit-selection με ομαδοποίηση

Οι βασικοί παράμετροι που επηρεάζουν άμεσα την ποιότητα της unit-selection σύνθεσης είναι τα χαρακτηριστικά που πρέπει να χρησιμοποιηθούν και πως να υπολογιστεί βέλτιστα το βάρος τους, το ιδανικό μέγεθος της μονάδας (unit) που επιλέγουμε να αντιστοιχίσουμε (μεγαλύτερες μονάδες απαιτούν μεγαλύτερη βάση δεδομένων), ο περιορισμός του πεδίου σύνθεσης (synthesis domain) αλλά και το μέγεθος της βάσης δεδομένων.

Τέλος το unit-selection μπορεί να έχει εντυπωσιακά αποτελέσματα όσον αφορά την ποιότητα της σύνθεσης ωστόσο παρουσιάζει κάποια βασικά μειονεκτήματα σε σχέση με τα στατιστικά παραμετρικά μοντέλα. Τα σημαντικότερα μειονεκτήματα είναι:

- Η άμεση εξάρτηση από την ποιότητα των ηχογραφήσεων της βάσης δεδομένων.
- Η περίπτωση όπου κατά τη σύνθεση μία πρότασης απαιτούνται κάποια χαρακτηριστικά (πχ προσωδιακά) και αυτά δεν αντιπροσωπεύονται σε επαρκή βαθμό από τη βάση δεδομένων τότε η ποιότητα της σύνθεσης υστερεί σε σημαντικό βαθμό.

- Ο περιορισμός στην τροποποίηση των επιλεγμένων τμημάτων από τη βάση δεδομένων. Αυτό έχει σαν αποτέλεσμα η φωνή που παράγεται να έχει το ίδιο στυλ με αυτή των ηχογραφήσεων της βάσης μας. Επομένως προκύπτει η ανάγκη χρήσης μεγαλύτερης βάσης δεδομένων πράγμα δύσκολο και κοστοβόρο.
- Η δυσκολία προσαρμογής σε άλλες γλώσσες.

Κεφάλαιο 2

Στατιστική Παραμετρική Σύνθεση Φωνής (SPSS)

Παρόλο που η σύνθεση φωνής με unit-selection έχει εκπληκτικά αποτελέσματα ως προς την ποιότητα της φωνής που παράγεται (να σημειώσουμε ότι τα καλύτερα παραδείγματα σύνθεσης φωνής με unit-selection υπερέχουν αυτών του SPSS), οι δυνατότητες των στατιστικών παραμετρικών μοντέλων είναι τεράστιες πράγμα που οδήγησε τους ερευνητές στην μελέτη τους τα τελευταία χρόνια. Τα βασικότερα πλεονέκτημά των στατιστικών παραμετρικών μοντέλων (SPSS) σε σχέση με τα unit-selection μοντέλα είναι:

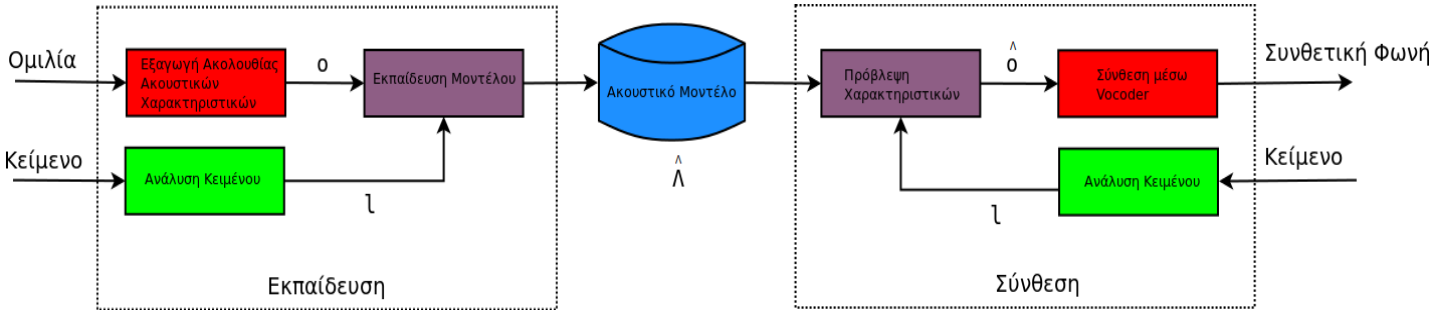
- Δυνατότητα εύκολης προσαρμογής σε νέους ομιλητές.
- Δυνατότητα επεξεργασίας της προσωδίας.
- Αρκετά περιορισμένη βάση δεδομένων σε σχέση με το unit-selection.
- Σταθερή απόδοση ποιότητας ήχου.
- Δυνατότητα προσαρμογής σε άλλες γλώσσες με την αλλαγή των γλωσσικών χαρακτηριστικών και των δέντρων απόφασης.

2.1 Περιγραφή Στατιστικού Παραμετρικού Μοντέλου

Το SPSS σύστημα συνδυάζει ένα ακουστικό μοντέλο (acoustic model) και ένα vocoder για την παραγωγή φωνής. Στο πυρήνα ενός στατιστικού παραμετρικού μοντέλου σύνθεσης φωνής βρίσκεται ένα στατιστικό μοντέλο (πχ HMM), το οποίο αρχικά εκπαιδεύεται (training). Στη συνέχεια τα εκπαιδευμένα μοντέλα που προέκυψαν σε συνδιασμό με το vocoder πραγματοποιούν την σύνθεση φωνής (synthesis). Η ανάλυση και η σύνθεση της φωνής γίνεται με βάση τις φασματικές παραμέτρους και τις παραμέτρους διέγερσης όπως η θεμελιώδης συχνότητα και το φάσμα (F0, spectrum).

Οι παράγοντες που επηρεάζουν την ποιότητα των SPSS μοντέλων είναι η ακρίβεια του ακουστικού μοντέλου, η ακρίβεια των ηχογραφήσεων της βάσης εκπαίδευσης, η ποιότητα του vocoder και η επίδραση του oversmoothing. Το oversmoothing είναι η υπερβολική εξομάλυνση των

παραμέτρων που προκέκυψαν κατά την ανάλυση της φωνής. Αυτή η υπερβολική εξομάλυνση οδηγεί σε απώλεια πληροφορίας πράγμα που επηρεάζει την φυσικότητα της συνθετικής φωνής.



Γράφημα 4: Βασική δομή ενος SPSS συστήματος

Στο γράφημα 4 μπορούμε να παρατηρήσουμε τη δομή ενός SPSS συστήματος. Αρχικά στην εκπαίδευση από μία βάση δεδομένων η οποία περιλαμβάνει ηχογραφήσεις εξάγεται μία ακολουθία ακουστικών χαρακτηριστικών o (sequence of acoustic feature vectors). Από το αντίστοιχο κείμενο εξάγεται μία ακολουθία γλωσσικών χαρακτηριστικών l (linguistic feature sequence). Η πρώτη περιλαμβάνει φασματικές παραμέτρους και παραμέτρους διέγερσης. Στη συνέχεια, το ακουστικό μοντέλο εκπαιδεύεται και με τη βοήθεια του ML (maximum likelihood) κριτηρίου υπολογίζονται οι παράμετροι του μοντέλου,

$$\hat{\Lambda} = \arg\max_{\Lambda} \{ p(o|l, \Lambda) \} \qquad \text{(B1)}$$

όπου o και l τα ακουστικά και γλωσσικά χαρακτηριστικά αντίστοιχα και Λ το ακουστικό μοντέλο (ένα σετ από παραμέτρους που αφορούν το μοντέλο).

Στο πεδίο της σύνθεσης αρχικά εξάγουμε από το κείμενο που θέλουμε να συνθέσουμε τα γλωσσικά χαρακτηριστικά. Έπειτα τα μοντέλα τα οποία προέκυψαν κατά την εκπαίδευση (context dependent phoneme models) ενώνονται σε μία ενιαία αλυσίδα από την οποία παράγουμε την πιο πιθανή ακολουθία ακουστικών χαρακτηριστικών $\hat{o}$ που αντιστοιχεί σε αυτά

$$\hat{o} = \arg\max_{o} \{ p(o|l, \hat{\Lambda}) \} \qquad \text{(B2)}$$

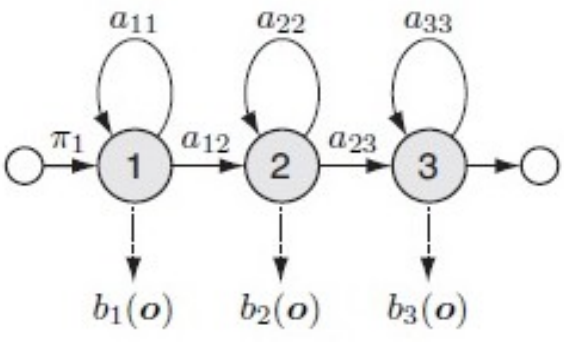
Στο τελευταίο στάδιο γίνεται η σύνθεση φωνής με τη χρήση του vocoder και της ακολουθίας των ακουστικών χαρακτηριστικών που έχει παραχθεί.

2.2.1 Ακουστικά Μοντέλα

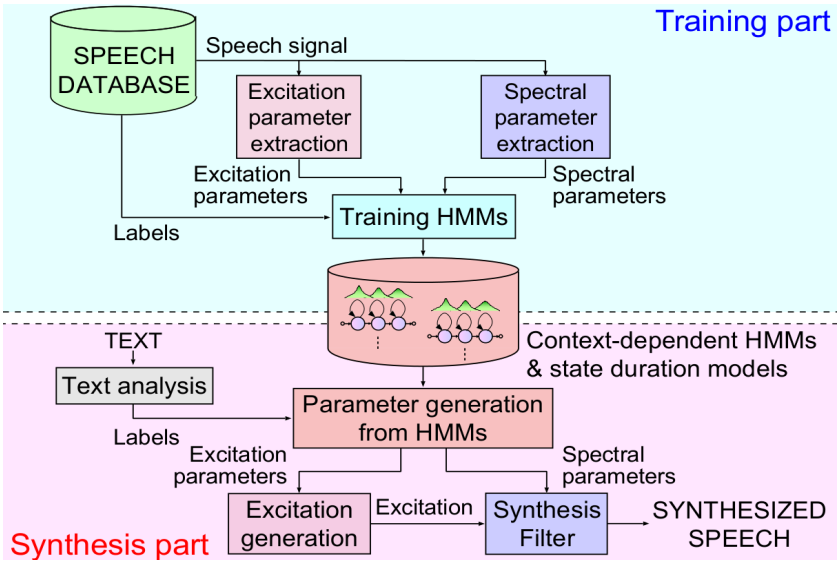
Στη σύνθεση φωνής μπορούν να χρησιμοποιηθούν διάφορα στατιστικά μοντέλα από τα οποία το πιο διαδεδομένο είναι το κρυφό μοντέλο Markov (HMM). Στο σημείο αυτό θα περιγράψουμε επίσης κάποια μοντέλα, τα οποία είναι παραλλαγές του HMM και θα εντοπίσουμε τα πλεονεκτήματα αλλά και τα μειονεκτήματά τους (ARHMMs, Trajectory-HMMs). Να επισημάνουμε ότι εστιάσαμε στο HMM καθώς ήταν το πρώτο στατιστικό παραμετρικό μοντέλο που μελετήθηκε.

HMMs

Το HMM είναι ένα στατιστικό μοντέλο το οποίο παράγει μία ακολουθία από παρατηρήσεις μέσω μίας διακριτής ακολουθίας κρυφών καταστάσεων και απεικονίζεται στο γράφημα 5. Οι συναρτήσεις πυκνότητας πιθανότητας των καταστάσεων εξόδου περιγράφουν την κατανομή των παρατηρήσεων που ανήκουν στις αντίστοιχες καταστάσεις ενώ η αλλαγή των καταστάσεων εξαρτάται από την πιθανότητα μετάβασης.



Γράφημα 5: Κρυφό Μοντέλο Markov (3 state left-to-right HMM)



Γράφημα 6: Σύνθεση φωνής με HMM

Στο γράφημα 6 απεικονίζεται το διάγραμμα ενός HMM συστήματος σύνθεσης φωνής. Όπως αναφέραμε και παραπάνω αρχικά υλοποιείται η εκπαίδευση και έπειτα η σύνθεση. Στο στάδιο της εκπαίδευση χρησιμοποιούμε μία βάση εκπαίδευσης η οποία περιλαμβάνει ηχογραφήσεις αλλά και τα αντίστοιχα κείμενά τους. Αρχικά εξάγωνται από τις ηχογραφήσεις τα ακουστικά χαρακτηριστικά (συνήθως μέσω του vocoder), όπως οι παράμετροι φάσματος και διέγερσης (mel-cepstral coefficients, log F0 και δυναμικά στοιχεία αυτών) ενώ από τα κείμενα εξάγωνται τα γλωσσικά χαρακτηριστικά. Σκοπός της διαδικασίας είναι να γίνει η αντιστοίχιση των ακουστικών με τα γλωσσικά χαρακτηριστικά (linguistic-acoustic mapping) και να δημιουργηθούν τα context-dependent HMMs (για την ακρίβεια να υπολογιστούν οι παράμετροι των μοντέλων). Η δημιουργία των context-dependent HMMs πραγματοποιείται με τη βοήθεια του EM (expectation maximization) αλγορίθμου.

Κάθε HMM έχει επίσης μία κατανομή διάρκειας καταστάσεων (συνήθως Gaussian) για την μοντελοποίηση της χρονικής δομής της ομιλίας, η οποία υπολογίζεται με τη βοήθεια του forward-backword αλγορίθμου. Η διάρκεια παραμονής σε κάθε κατάσταση αποτελεί πολύ σημαντική παράμετρο για την υλοποίηση του συστήματος σύνθεσης φωνής από κείμενο. Η μοντελοποίηση της διάρκειας θα ρυθμίσει στην πορεία τη διάρκεια παραγωγής της συνθετικής ομιλίας καθώς και το ρυθμό αυτής. Να σημειώσουμε ότι τα μοντέλα διάρκειας δεν θα εξεταστούν στα πλαίσια αυτής της εργασίας.

Στη συνέχεια καθένα από τα τρία παραπάνω στοιχεία (mel cepstral coefficients, logF0 ,

duration) ομαδοποιείται ανεξάρτητα με τη βοήθεια δέντρων τα οποία σχηματίζονται βάση ερωτήσεων ως προς τα contexts καθώς διαφορετικά contexts λαμβάνονται υπόψη για καθένα από αυτά. Σκοπός της ομαδοποίησης αυτής είναι να μειώσει τον αριθμό των context-dependent HMMs. Να επισημάνουμε ότι για κάθε φώνημα εξετάζονται όλοι οι πιθανοί συνδυασμοί ως προς τα contexts και για καθένα από αυτούς τους συνδυασμούς προκύπτει ένα context-dependent HMM. Ένα τέτοιο σύνολο από μοντέλα θα προκαλούσε πρόβλημα καθώς δεν είναι δυνατόν μια βάση να περιλαμβάνει αρκετά παραδείγματα για καθένα από αυτά επομένως κάποιοι πιθανοί συνδυασμοί θα αντιπροσωπεύονταν από περιορισμένο αριθμό πράγμα που θα οδηγήσει σε λάθος αντιστοίχιση (συνήθως overfitting). Επομένως είναι αναγκαίο να περιορίσουμε τον αριθμό τους πράγμα που επιτυγχάνεται μέσω της ομαδοποίησης που αναφέραμε.

Στο πεδίο της σύνθεσης πραγματοποιείται ο υπολογισμός της πιθανότερης ακολουθίας ακουστικών χαρακτηριστικών δεδομένου ότι έχουμε τα γλωσσικά χαρακτηριστικά της πρότασης που θέλουμε να συνθέσουμε και τα context-dependent HMMs. Αρχικά γίνεται ανάλυση κειμένου όπου μία πρόταση μετατρέπεται σε context-dependent ακολουθία φωνημάτων. Με τη βοήθεια των μοντέλων που εκπαιδεύσαμε η ακολουθία των φωνημάτων μετατρέπεται σε ακολουθία context-dependent phoneme HMMs. Δηλαδή από τα επιμέρους HMMs προκύπτει ένα ενιαίο. Σκοπός είναι από αυτή την ακολουθία να εξαχθούν τα ακουστικά χαρακτηριστικά τα οποία έπειτα θα χρησιμοποιηθούν για τη φωνητική σύνθεση της πρότασης αυτής. Για τον υπολογισμό της ακολουθίας των ακουστικών χαρακτηριστικών είναι αναγκαία η εύρεση της βέλτιστης ακολουθίας καταστάσεων.

Το πρόβλημα που εντοπίζουμε κατά την σύνθεση είναι ότι τα ακουστικά χαρακτηριστικά ισούνται με τη μέση τιμή της κατανομής που περιγράφει την κατάσταση στην οποία βρίσκονται. Αυτό αποτελεί μειονέκτημα για την συνθετική φωνή καθώς δεν ανταποκρίνεται στην πραγματικότητα και σαν αποτέλεσμα εντοπίζονται ασυνέχειες κατά την μετάβαση από μία κατάστασης στην επόμενη. Το πρόβλημα αυτό αντιμετωπίζεται με την εισαγωγή δυναμικών στατιστικών χαρακτηριστικών (π.χ. Δc) στο διάνυσμα των παρατηρήσεων ως περιορισμός για τον αλγόριθμο παραγωγής παραμέτρων φωνής (parameter generation algorithm). Το αποτέλεσμα της παραπάνω διαδικασίας είναι ότι η καμπύλη των ακουστικών χαρακτηριστικών που προκύπτει δεν είναι πλέον στάσιμη όπως προηγουμένως εφόσον εισάγαμε δυναμικά χαρακτηριστικά. Καταφέραμε με αυτόν τον τρόπο να προσδώσουμε φυσικότητα στην ομιλία λόγω ομαλών μεταβάσεων. Στο τελευταίο βήμα, εφόσον έχουμε εξάγει την ακολουθία των ακουστικών χαρακτηριστικών, με τη βοήθεια του vocoder συνθέτουμε τελικά τη φωνή.

Η σύνθεση φωνής με χρήση HMM μοντέλων έχει κάποια βασικά πλεονεκτήματα. Ορισμένα από αυτά είναι η αποτελεσματική ομαδοποίηση λόγω ύπαρξης αλγορίθμου, η γρήγορη σύνθεση

καθώς το υπολογιστικό κόστος είναι μικρό και η εκπαίδευση που ελέγχεται εύκολα λόγω εφαρμογής του ΕΜ αλγορίθμου. Τα μειονεκτήματά της είναι η απουσία συνέπειας καθώς δυναμικά χαρακτηριστικά χρησιμοποιήθηκαν μόνο στο κομμάτι της σύνθεσης και όχι της εκπαίδευσης, η υψηλή αδράνεια (high latency) που είναι O(T) και η πολυπλοκότητα ως προς το debugging καθώς σε περίπτωση σφάλματος εξετάζονται οι κόμβοι των δέντρων και ελέγχονται τα στατιστικά που έχουν προκύψει από αυτά. Επίσης τα στατιστικά στοιχεία αλλάζουν μόνο αν μεταβεί από μία κατάσταση σε άλλη συνεπώς είναι στάσιμα εντός της κάθε κατάστασης. Τέλος το μοντέλο που περιγράφει τη διάρκεια των καταστάσεων θεωρείται αρκετά ασθενές.

ARHMMs

Τα Autoregressive HMMs (ARHMMs) ανήκουν και αυτά στην κατηγορία των μοντέλων χώρου-καταστάσεων. Διαφέρουν ως προς τα HMMs καθώς κάθε ακολουθία παρατηρήσεων έχει εξάρτηση με την προηγούμενή της αν το μοντέλο είναι πρώτης τάξης ή και με περισσότερες για μοντέλα μεγαλύτερων τάξεων. Εφόσον τα γειτονικά πλαίσια (frames) δεν είναι ανεξάρτητα η ακολουθία ακουστικών χαρακτηριστικών που προκύπτει είναι πιο ομαλή. Στα ARHMM που οι καταστάσεις του συστήματος είναι παρατηρήσιμες οι παράμετροι του μοντέλου υπολογίζονται μέσω της επίλυσης αλγεβρικών εξισώσεων κλειστής μορφής όπου οι άγνωστες μεταβλητές είναι οι παράμετροι του μοντέλου και οι γνωστές τα επαρκή στατιστικά στοιχεία (sufficient statistics). Οι εξισώσεις που περιγράφουν το ARHMM είναι:

$$x_1 \sim N(g_{1,} Q_1) \qquad \text{(B3)}$$

$$x_t = F x_{t-1} + g + w^{(x)} , \quad w^{(x)} \sim N(0, Q) \qquad \text{(B4)}$$

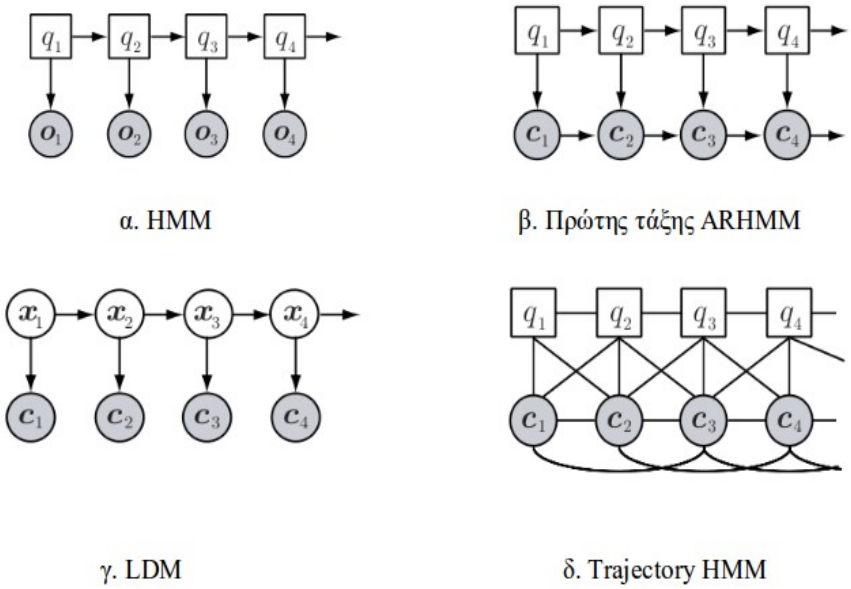
Βασικό πλεονέκτημα των ARHMM είναι ότι δεν είναι απαραίτητη η ενσωμάτωση δυναμικών στατιστικών στοιχείων καθώς η εξάρτηση αυτή υπάρχει ήδη στο μοντέλο πράγμα που το καθιστά συνεπές (consistent). Επιπλέον είναι δυνατόν να εφαρμοστεί ο ΕΜ αλγόριθμος έτσι η εκπαίδευση του μοντέλου ελέγχεται σχετικά εύκολα και η ύπαρξη αλγορίθμου για σχηματισμό δέντρων παρέχει αποτελεσματική ομαδοποίηση. Το latency είναι χαμηλό O(1) καθώς για τον υπολογισμό του $\widehat{c}_1$ (γράφημα 7) απαιτούνται στατιστικά στοιχεία μόνο για το πρώτο πλαίσιο (frame) και η σύνθεση είναι πιο γρήγορη σε σχέση με τα HMMs. Ένα βασικό μειονέκτημα όπως

και σχεδόν σε κάθε μοντέλο του SPSS είναι ότι το debugging είναι αρκετά δύσκολο.

Trajectory HMMs

Το Trajectory HMM αποτελεί το μοναδικό μη κατευθυνόμενο μοντέλο σε σχέση με τα παραπάνω. Αντιμετωπίζει τη βασική αδυναμία του HMM διότι η ακολουθία των παρατηρήσεων αλλάζει δυναμικά μέσα σε μία κατάσταση αλλά και το διάνυσμα παρατηρήσεων κάθε κατάστασης δεν είναι ανεξάρτητο από των άλλων καταστάσεων του μοντέλου. Λόγω των παραπάνω η ακολουθία ακουστικών χαρακτηριστικών που εξάγεται είναι πιό ομαλή οπότε και η συνθετική φωνή υψηλότερης ποιότητας.

Το Trajectory-HMM χαρακτηρίζεται από συνέπεια λόγω της χρήσης δυναμικών στατιστικών τόσο στη φάση της εκπαίδευσης όσο και της σύνθεσης. Επίσης η σύνθεση επιτυγχάνεται γρήγορα (βέλτιστο σε σχέση με HMM και ARHMM). Κάποια από τα μειονεκτήματά του είναι η έλλειψη αποδοτικού αλγορίθμου ομαδοποίησης, το υψηλό latency O(T) και το πολύπλοκο debugging.



Γράφημα 7: Ακουστικά μοντέλα του SPSS συστήματος

Στο γράφημα 7 απεικονίζονται τα δυναμικά μοντέλα που εξετάσαμε (με εξαίρεση το LDM που θα το μελετήσουμε αναλυτικά). Να επισημάνουμε ότι το ARHMM είναι πρώτης τάξης και στο Trajectory HMM τα δυναμικά χαρακτηριστικά έχουν υπολογιστεί από το αντίστοιχο, το προηγούμενο και το επόμενο πλαίσιο (frame).

Κεφάλαιο 3

Γραμμικό Δυναμικό Μοντέλο (LDM)

Το κύριο ακουστικό μοντέλο που μελετήθηκε στη σύνθεση φωνής είναι το HMM, το οποίο είναι αποτελεσματικό ως προς την ποιότητα σύνθεσης αλλά και την τροποποίηση της συνθετικής φωνής. Παρόλα αυτά παρουσιάζει ορισμένα βασικά μειονεκτήματα όπως αναφέραμε και παραπάνω καθώς οι παρατηρήσεις είναι υπό συνθήκη ανεξάρτητες δεδομένου της ακολουθίας καταστάσεων και τα στατιστικά στοιχεία κάθε κατάστασης είναι στάσιμα δηλαδή δεν αλλάζουν δυναμικά. Αυτό μας οδήγησε στη μελέτη ενός άλλου μοντέλου που καλείται γραμμικό δυναμικό μοντέλό (LDM) καθώς έχει μικρή απαίτηση σε υπολογιστική ισχύ και χαμηλή καθυστέρηση (latency) ενώ μπορεί να χρησιμοποιηθεί σε εφαρμογές που έχουν απαιτήσεις πραγματικού χρόνου (real-time requirements).

Ένα γραμμικό δυναμικό μοντέλο έχει καταστάσεις που περιγράφονται από συνεχείς μεταβλητές και αποτελεί μοντέλο χώρου-καταστάσεων. Στα μοντέλα αυτά η μελλοντική κατάσταση εξαρτάται μόνο από την τωρινή οπότε υπακούν στην Μαρκοβιανή υπόθεση. Επομένως οι αλγόριθμοι για τον υπολογισμό των καταστάσεων είναι ουσιαστικά ίδιοι με αυτούς των HMMs. Επίσης η ακολουθία των παρατηρήσεων εξαρτάται μόνο από την κατάσταση του μοντέλου εκείνη τη χρονική στιγμή. Ο Minka υποστήριξε ότι οι ιδιότητες που έχει ένα μοντέλο εισάγωνται μόνο στο τελικό στάδιο και κατά συνέπεια υπολόγισε τις καταστάσεις του γραμμικού δυναμικού μοντέλου βάση των HMM. Όταν οι καταστάσεις δεν μπορουν να παρατηρηθούν όπως στα LDMs τότε οι παράμετροι μπορούν να υπολογιστούν από κοινού με τις καταστάσεις του μοντέλου μέσω του EM αλγορίθμου.

Τα βασικά εμπόδια που έχουμε να αντιμετωπίσουμε στη σύνθεση φωνής με γραμμικά δυναμικά μοντέλα είναι η αντιστοίχιση των γλωσσικών χαρακτηριστικών που προέκυψαν από την ανάλυση κειμένου με τα ακουστικά μοντέλα καθώς και το oversmoothing των χαρακτηριστικών που εξάγονται. Το πρώτο πρόβλημα επιλύεται με τη χρήση δέντρων απόφασης και στην εργασία αυτή χρησιμοποιήθηκε top-down greedy spliting αλγόριθμός βασισμένος σε ερωτήσεις που αφορούν το γλωσσικά χαρακτηριστικά. Το oversmoothing αντιμετωπίζεται με τη χρήση του GV (global variance). Να αναφέρουμε ότι στην εργασία αυτή δεν εισάγαμε GV για τη σύνθεση της φωνής.

Σημαντικό πλεονέκτημά των γραμμικών δυναμικών μοντέλων (LDMs) είναι η ομαλή εξαγωγή ακουστικών χαρακτηριστικών (acoustic features) αλλά και η συνέπειά τους εφόσον δεν χρησιμοποιούνται δυναμικά στατιστικά. Ως προς την εκπαίδευσή τους υπάρχει έλεγχος σε

σημαντικό βαθμό αφού και εδώ μπορεί να εφαρμοστεί ο ΕΜ αλγόριθμος ενώ η σύνθεση επιτυγχάνεται πιο γρήγορα σε σχέση με τα HMM. Τα μειονέκτημά τους αφορούν την ομαδοποίηση καθώς ο ΕΜ αλγόριθμος πρέπει να τρέχει για κάθε διακλάδωση του δέντρου αλλά και το δύσκολο debugging.

3.1 Περιγραφή Γραμμικού Δυναμικού Μοντέλου

Όπως αναφέραμε και παραπάνω τα διανύσματα των καταστάσεων του γραμμικού δυναμικού μοντέλου είναι συνεχή και η διαδικασία εξέλιξης των καταστάσεων αποτελεί μία γραμμική πρώτης τάξης Gauss-Markov τυχαία διαδικασία. Το μοντέλο αυτό περιγράφεται από τις εξής σχέσεις:

$$x_1 \sim N(g_{1,}, Q_1) \qquad (\Gamma 1)$$

$$x_t = F x_{t-1} + g + w^{(x)}, \quad w^{(x)} \sim N(0, Q) \qquad (\Gamma 2)$$

$$y_t = H x_t + \mu + w^{(y)}, \quad w^{(y)} \sim N(0, R) \qquad (\Gamma 3)$$

όπου (Γ1) η κατανομή της αρχικής κατάστασης του συστήματος και (Γ2),(Γ3) η σχέσεις που περιγράφουν τις καταστάσεις και τις παρατηρήσεις του μοντέλου αντίστοιχα.

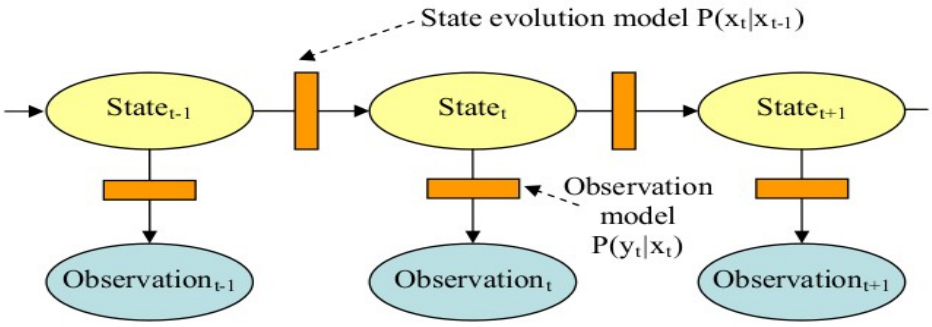
Το F και H αποτελούν πίνακες διαστάσεων *[n x n]* και *[m x n]* που σχετίζονται με την αλλαγή των καταστάσεων και των παρατηρήσεων αντίστοιχα.

Το x είναι διάνυσμα μεγέθους n που είναι η κατάσταση του μοντέλου και διαμορφώνεται βάση της σχέσης (Γ2) και x_1 η αρχική κατάσταση του μοντέλου της οποίας η κατανομή είναι γνωστή. Το y είναι το διάνυσμα παρατηρήσεων μεγέθους m που εξάγεται με δειγματοληψία ανά διακριτά χρονικά διαστήματα και περιγράφεται από τη σχέση (Γ3) .

Τα $w^{(x)}$ και $w^{(y)}$ είναι τα διανύσματα που περιγράφουν την εξέλιξη της κατάστασης και του θορύβου αντίστοιχα και είναι ανεξάρτητα μεταξύ τους.

Στο γράφημα 8 παρατηρούμε τη δομή ενός γραμμικού δυναμικού μοντέλου (LDM). Μπορούμε να το περιγράψουμε ως μία ακολουθία η οποία έχει χωριστεί σε τμήματα ανά χρονικά

διαστήματα (καταστάσεις του μοντέλου). Κάθε κατάσταση επικοινωνεί με την επόμενή της και σε κάθε κατάσταση αντιστοιχίζεται μία ακολουθία παρατηρήσεων.



Γράφημα 8: Δομή ενός LDM μοντέλου

3.2 Από Κοινού Πιθανότητα σε Γραμμικά Δυναμικά Μοντέλα

Η από κοινού πιθανότητα των καταστάσεων Y και των παρατηρήσεων X δεδομένου των παραμέτρων θ του μοντέλου σε ένα γραμμικό δυναμικό μοντέλο αναλύεται σε επιμέρους τμήματα. Μπορούμε να παρατηρήσουμε ότι η κατανομή αυτή είναι Gaussian καθώς καθένα από τα επιμέρους τμήματα έχει και αυτό Gaussian κατανομή. Θα χρησιμοποιήσουμε αυτή τη σχέση παρακάτω.

$$P(X, Y|\theta) = P(x_1|\theta)\prod_{t=2}^{T} P(x_t|x_{t-1}, \theta)\prod_{t=1}^{T} P(y_t|x_t, \theta) \qquad (\Gamma 16)$$

3.3 Forward-Backward Αναδρομές

Στην υποενότητα αυτή αναλύουμε τον Forward-Backward αλγόριθμο με σκοπό να εξηγήσουμε τις σχέσεις που περιγράφουν τα φίλτρα Kalman (Kalman Filter) και τους εξομαλυντές Kalman (Kalman Smoother).

Η δομή των Μαρκοβιανών αλυσίδων επιτρέπει τον αναδρομικό υπολογισμό των $p(Y)$, $p(x_t|Y)$ και $p(x_{t-1}, x_t|Y)$ τα οποία θα χρησιμοποιηθούν στα δύο πρώτα βήματα του LDM όπως θα εξηγήσουμε και παρακάτω. Πρέπει να σημειώσουμε ότι για τον υπολογισμό των παραπάνω πιθανοτήτων αρκεί να υπολογίσουμε την από κοινού πιθανότητα $p(x_\tau, Y)$.

$$p(x_t, Y) = p(x_t, y_{1:t}, y_{t+1:T}) = p(x_t, y_{1:t})\, p(y_{t+1:T}|x_t, y_{1:t})$$

Επειδή $y_{t+1:T}$ δεδομένου του x_t είναι ανεξάρτητο του $y_{1:t}$ $\Rightarrow$

$$p(x_t, y_{1:t}, y_{t+1:T}) = p(x_t, y_{1:t})\, p(y_{t+1:T}|x_t)$$

Ο Rabiner εισήγαγε τις σχέσεις (Γ17)(Γ18)

$$a_t(x_t) = p(x_t, y_{1:t}) \qquad (\Gamma 17)$$

$$b_t(x_t) = p(y_{t+1:T}|x_t) \qquad (\Gamma 18)$$

Επομένως η από κοινού πιθανότητα της κατάστασης την χρονική στιγμή t και των παρατηρήσεων Y ορίζεται ως:

$$p(x_t, Y) = a_t(x_t)\, b_t(x_t) \qquad (\Gamma 19)$$

Οπότε κατασκευάζουμε την αναδρομή για το $a_t(x_t)$.

$$a_t(x_t) = p(x_t, y_{1:t}) = p(x_t, y_{1:t-1}, y_t) = p(y_t|x_t, y_{t-1})\, p(x_t, y_{1:t-1}) =$$

$$p(y_t|x_t) \int p(x_t, x_{t-1} = z, y_{1:t-1})\, dz \;=$$

$$p(y_t|x_t) \int p(x_t|x_{t-1} = z, y_{1:t-1})\, p(x_{t-1} = z, y_{1:t-1})\, dz$$

Η αναδρομή του $a_t(x_t)$ είναι

$$a_t(x_t) = p(y_t|x_t) \int p(x_t|x_{t-1} = z)\, a_{t-1}(z)\, dz \qquad (\Gamma 20)$$

Ομοίως ορίζουμε την αναδρομική συνάρτηση του $b_t(x_t)$.

$$b_{t-1}(x_{t-1}) = p(y_{t:T} | x_{t-1}) = \int p(y_{t:T}, x_t = z | x_{t-1}) dz$$

$$\int p(y_t, y_{t+1:T}, x_t = z | x_{t-1}) dz \;=$$

$$\int p(y_t | y_{t+1:T}, x_t = z, x_{t-1}) p(y_{t+1:T}, x_t = z | x_{t-1}) dz \;=$$

$$\int p(y_t | x_t = z) p(x_t = z | x_{t-1}) p(y_{t+1:T} | x_t = z) dz$$

Οπότε η σχέση που ορίζει την αναδρομή της συνάρτησης $b_t(x_t)$ είναι :

$$b_{t_1}(x_{t-1}) = \int p(x_t = z | x_{t-1}) p(y_t | x_t = z) b_t(z) dz \quad (\Gamma 21)$$

Για να ξεκινήσουν οι αναδρομές θεωρούμε:

$$a_1(x_1) = p(x_1) p(y_1 | x_1) \quad (\Gamma 22)$$

$$b_T(x_T) = 1 \quad (\Gamma 23)$$

Επομένως υπολογίζεται το $p(x_t, Y)$ και μέσω αυτού προκύπτουν οι σχέσεις (Γ24)(Γ25)(Γ26) .

$$p(Y) = \int p(x_t, Y) dx_t \quad (\Gamma 24)$$

$$p(x_t | Y) = \frac{p(x_t, Y)}{p(Y)} \quad (\Gamma 25)$$

$$p(x_{t-1}, x_t | Y) = \frac{p(x_{t-1}, x_t, Y)}{p(Y)} = \frac{p(x_{t-1}, x_t, y_{1:t-1}, y_t, y_{t+1:T})}{p(Y)} \;=$$

$$\frac{p(x_{t-1}, y_{1:t-1})\, p(y_t|x_t)\, p(x_t|x_{t-1})\, p(y_{t+1:T}|x_t)}{p(Y)} \Rightarrow$$

$$p(x_{t-1}, x_t|Y) = \frac{a_{t-1}(x_{t-1})\, p(y_t|x_t)\, p(x_t|x_{t-1})\, b_t(x_t)}{p(Y)} \quad (\Gamma 26)$$

3.3.1 Scaling factors

Στην υποενότητα αυτή ορίζουμε τις τιμές των $\hat{a}_t(\cdot)$ και $\hat{b}_t(\cdot)$.

Ορίζουμε

$$c_t = p(y_t|t_{1:t-1}) \quad (\Gamma 27)$$

τότε η πιθανότητα της ακολουθίας των παρατηρήσεων μπορεί να οριστεί ως:

$$p(Y) = \prod_{\tau=1}^{t} c_\tau \quad (\Gamma 28)$$

Χρησιμοποιώντας τη σχέση (Γ28) ορίζουμε το $a_t \, \omega\varsigma \, \pi\rho o\varsigma \, το \, \hat{a}_t$

$$a_t(x_t) = p(x_t, y_{1:t}) = p(y_t|x_t, y_{t-1})\, p(x_t, y_{1:t-1}) \Rightarrow$$

$$a_t(x_t) = (\prod_{\tau=1}^{t} c_\tau)\, \hat{a}_t(x_t) \Rightarrow \quad (\Gamma 29)$$

$$\hat{a}_t(x_t) = p(x_t|y_{1:t}) \quad (\Gamma 30)$$

Από τις σχέσεις (Γ20),(Γ29) έχουμε:

$$(\prod_{\tau=1}^{t} c_\tau)\hat{a}_t(x_t) = p(y_t|x_t)\int p(x_t|x_{t-1}=z)(\prod_{\tau=1}^{t-1} c_\tau)\hat{a}_{t-1}(z)\,dz \quad \Rightarrow \quad (\Gamma 31)$$

$$\hat{a}_t(x_t) = \frac{1}{c_\tau} p(y_t|x_t)\int p(x_t|x_{t-1}=z)\hat{a}_{t-1}(z)\,dz \qquad (\Gamma 32)$$

Ομοίως ορίζουμε $b_t \, \omega\varsigma \, \pi\rho o\varsigma \, \tau o \, \hat{b}_t$

$$b_t(x_t) = (\prod_{\tau=1}^{t} c_\tau)\hat{b}_t(x_t) \qquad \Gamma(33)$$

Από τις σχέσεις (Γ21),(Γ33) έχουμε:

$$\hat{b}_{t_1}(x_{t-1}) = \frac{1}{c_\tau}\int p(x_t=z|x_{t-1})\,p(y_t|x_t=z)\hat{b}_t(z)\,dz \qquad (\Gamma 34)$$

Οπότε μπορούμε να ορίσουμε ως προς $\hat{a}_t(\cdot)$ και $\hat{b}_t(\cdot)$ τις παρακάτω πιθανότητες.

$$p(x_t|Y) = \hat{a}_t(x_t)\hat{b}_t(x_t) \qquad (\Gamma 35)$$

$$p(x_{t-1}, x_t|Y) = \frac{1}{c_\tau}\hat{a}_{t-1}(x_{t-1})\,p(y_t|x_t)\,p(x_t|x_{t-1})\hat{b}_t(x_t) \quad (\Gamma 36)$$

3.3.2 Sequential Recursions

Οι αναδρομές για τα $\hat{a}_t(\cdot)$ και $\hat{b}_t(\cdot)$ υλοποιούνται ανεξάρτητα η καθεμία. Στόχος μας είναι να παράξουμε μία ακολουθιακή αναδρομή έτσι ώστε το $\hat{b}_t(\cdot)$ να υπολογιστεί μέσω του $\hat{a}_t(\cdot)$ και όχι απευθείας από τα δεδομένα. Αυτού του τύπου την αναδρομή θα χρησιμοποιήσουμε για την κατασκευή του LDM.

Πολλαπλασιάζοντας την εξίσωση (Γ34) με $\hat{a}_{t-1}(x_{t-1})$ έχουμε:

$$\hat{a}_{t-1}(x_{t-1})\hat{b}_{t_1}(x_{t-1})=\frac{\hat{a}_{t-1}(x_{t-1})}{c_\tau}\int p(x_t=z|x_{t-1})\,p(y_t|x_t=z)\,\hat{b}_t(z)\,dz$$

$$= \frac{\int \hat{a}_{t-1}(x_{t-1})\,p(x_t=z|x_{t-1})\,p(y_t|x_t=z)\,\hat{a}_t(z)\hat{b}_t(z)}{c_\tau \hat{a}_t(z)}dz$$

$$= \frac{\int p(x_{t-1}|y_{1:t-1})\,p(x_t=z|x_{t-1},y_{1:t-1})\,p(y_t|x_t=z)\,p(x_t|Y)}{c_\tau \hat{a}_t(z)}dz \quad (\Gamma 37)$$

Όμως ο όρος $c_\tau \hat{a}_t(z)$ μπορεί να γραφτεί ως εξής:

$$c_\tau \hat{a}_t(z)=p(y_t|y_{1:t-1})\,p(x_t|y_{1:t})=p(y_t|y_{1:t-1})\,p(x_t|y_t,y_{1:t-1}) =$$

$$p(y_t|x_t,y_{1:t-1})\,p(x_t|y_{1:t-1}) = p(y_t|x_t)\,p(x_t|y_{1:t-1}) \Rightarrow$$

$$c_\tau \hat{a}_t(x_t)=p(y_t|x_t)\,p(x_t|y_{1:t-1}) \qquad (\Gamma 38)$$

Από τις σχέσεις (Γ37),(Γ38) έχουμε:

$$\hat{a}_{t-1}(x_{t-1})\hat{b}_{t_1}(x_{t-1})= \frac{\int p(x_{t-1}|y_{1:t-1})\,p(x_t=z|x_{t-1},y_{1:t-1})\,p(y_t|x_t=z)\,p(x_t|Y)}{p(y_t|x_t)\,p(x_t|y_{1:t-1})}dz \Rightarrow$$

$$\hat{a}_{t-1}(x_{t-1})\hat{b}_{t-1}(x_{t-1}) = \int p(x_{t-1}|x_t=z, y_{1:t-1})\hat{a}_t(z)\hat{b}_t(z)dz \quad (\Gamma 39)$$

3.3.3 Φίλτρο Kalman (Kalman Filter)

Στα γραμμικά δυναμικά μοντέλα οι μεταβλητές που περιγράφουν τις καταστάσεις του μοντέλου είναι συνεχείς. Επίσης θεωρούμε ότι οι πιθανότητες της αρχικής μετάβασης και των αρχικών παρατηρήσεων είναι Gaussian επομένως οι από κοινού (joint) και οι περιθωριακές (marginal) πυκνότητες πιθανότητας είναι και αυτές Gaussian. Αυτό μας επιτρέπει τον ακριβή υπολογισμό των συναρτήσεων $\hat{a}_t(\cdot)$ και $\hat{b}_t(\cdot)$.

Σε αυτό το σημείο ορίζουμε την μέση τιμή και συνδιακύμανση των $p(x_t|y_{1:t})$ και $p(x_t|y_{1:t-1})$ ως $\hat{x}_{t|t}$, $\hat{\Sigma}_{t|t}$ και $\hat{x}_{t|t-1}$, $\hat{\Sigma}_{t|t-1}$ αντίστοιχα.

Γνωρίζουμε ότι:

$$\hat{a}_t(x_t) = p(x_t|y_{1:t}) = N(x_t; \hat{x}_{t|t}, \hat{\Sigma}_{t|t}) \quad (\Gamma 40)$$

και

$$p(x_t|y_{1:t-1}) = N(x_t; \hat{x}_{t|t-1}, \hat{\Sigma}_{t|t-1}) \quad (\Gamma 41)$$

Θα υπολογίσουμε τη μέση τιμή και συνδιακύμανση του $p(x_t|y_{1:t-1})$ ως προς τη μέση τιμή και συνδιακύμανση του $p(x_{t-1}|y_{1:t-1})$.

$$\hat{x}_{t|t} = E\{x_t|y_{1:t}\} = E\{Fx_{t-1}+g+w|y_{1:t-1}\} = F\hat{x}_{t-1|t-1}+g \quad (\Gamma 42)$$

$$\hat{\Sigma}_{t|t-1} = cov\{x_t|y_{1:t-1}\} = cov\{Fx_{t-1}+g+w|y_{1:t-1}\} = F\hat{\Sigma}_{t-1|t-1}F^T+Q \quad (\Gamma 43)$$

Στη συνέχεια θα υπολογίσουμε τη μέση τιμή και συνδιακύμανση του $p(x_t|y_{1:t})$ ως προς τη

μέση τιμή και συνδιακύμανση του $p(x_t|y_{1:t-1})$.

Από την σχέση (Γ38) έχουμε:

$$c_\tau \hat{a}_t(x_t) = p(y_t|y_{1:t-1})\, p(x_t|y_{1:t}) = p(y_t|x_t)\, p(x_t|y_{1:t-1}) \quad \Rightarrow$$

$$p(y_t|y_{1:t-1})\, N(x_t; \hat{x}_{t|t}, \hat{\Sigma}_{t|t}) \;=\; N(y_t; H x_t + \mu, R)\, N(x_t; x_{t|t-1}, \Sigma_{t|t-1})$$

Οπότε για την από κοινού κατανομή $[x_t^T, y_t^T]^T$ δεδομένου του $y_{1:t-1}$ ισχύει:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} \sim N\left(\begin{bmatrix} \hat{x}_{t|t-1} \\ H\hat{x}_{t|t-1} + \mu \end{bmatrix}, \begin{bmatrix} \hat{\Sigma}_{t|t-1} & \hat{\Sigma}_{t|t-1} H^T \\ H\hat{\Sigma}_{t|t-1} & H\hat{\Sigma}_{t|t-1} H^T + R \end{bmatrix} \right) \qquad (\Gamma 44)$$

Επομένως οι παράμετροι του $p(x_t|y_t, y_{1:t-1}) = p(x_t|y_{1:t})$ είναι οι εξής:

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t e_t \qquad (\Gamma 45)$$

$$\hat{\Sigma}_{t|t} = \hat{\Sigma}_{t|t-1} - K_t H \hat{\Sigma}_{t|t-1} \qquad (\Gamma 46)$$

Όπου

$$K_t = \hat{\Sigma}_{t|t-1} H^T (H \hat{\Sigma}_{t|t-1} H^T + R)^{-1} \qquad (\Gamma 47)$$

$$e_t = y_t - H\hat{x}_{t|t-1} - \mu \qquad (\Gamma 48)$$

Και

$$c_t = p(y_t | y_{1:t-1}) = N(y_t ; H\,\hat{x}_{t|t-1} + \mu \ , \ H\,\hat{\Sigma}_{t|t-1}\,H^T + R) \quad \Rightarrow$$

$$c_t = N(e_t ; 0, H\,\hat{\Sigma}_{t|t-1}\,H^T + R) \qquad (\Gamma 49)$$

3.3.4 Εξομαλυντής Kalman (Kalman Smoother)

Η συνάρτηση πυκνότητας πιθανότητας $p(x_t | Y)$ έχει Gaussian κατανομή με μέση τιμή και συνδιακύμανση $\hat{x}_{t|T}$, $\hat{\Sigma}_{t|T}$ αντίστοιχα. Έχουμε αποδείξει παραπάνω ότι μπορεί να γραφτεί ως γινόμενο των συναρτήσεων $\hat{a}_t(\cdot)$ και $\hat{b}_t(\cdot)$.

$$\hat{a}_t(x_t)\hat{b}_t(x_t) = p(x_t | Y) = N(x_t ; \hat{x}_{t|T} \ , \hat{\Sigma}_{t|T}) \qquad (\Gamma 50)$$

Θα δημιουργήσουμε μία αναδρομή μέσω της οποίας θα υπολογίσουμε τις παραμέτρους του $p(x_{t-1} | Y)$ ως προς τις παραμέτρους των $p(x_t | Y)$, $p(x_t | y_{1:t-1})$ και $p(x_{t-1} | y_{1:t-1})$.

Η από κοινού κατανομή του $[x_{t-1}^T, x_t^T]^T$ δεδομένου του $y_{1:t-1}$ είναι η εξής:

$$\begin{bmatrix} x_{t-1} \\ x_t \end{bmatrix} \sim N\left(\begin{bmatrix} \hat{x}_{t-1|t-1} \\ \hat{x}_{t|t-1} \end{bmatrix}, \begin{bmatrix} \hat{\Sigma}_{t-1|t-1} & \hat{\Sigma}_{t-1|t-1}F^T \\ F\,\hat{\Sigma}_{t-1|t-1} & F\,\hat{\Sigma}_{t-1|t-1}F^T + Q \end{bmatrix} \right)$$

Επομένως

$$p(x_{t-1}|x_t=z, y_{1:t-1}) = N(x_{t-1};\ \hat{x}_{t-1}+J_t(z-\hat{x}_{t|t-1}),(I-J_tF)\hat{\Sigma}_{t-1|t-1}) \qquad (\Gamma 51)$$

Όπου

$$J_t=\hat{\Sigma}_{t-1|t-1}F^T(F\hat{\Sigma}_{t-1|t-1}F^T+Q)^{-1} = \hat{\Sigma}_{t-1|t-1}F^T\hat{\Sigma}_{t|t-1}^{-1} \qquad (\Gamma 52)$$

Από τις σχέσεις (Γ39),(Γ50),(Γ51),Γ(52) προκύπτει:

$$N(x_{t-1};\hat{x}_{t-1|T},\hat{\Sigma}_{t-1|T}) = \hat{a}_{t-1}(x_{t-1})\hat{b}_{t-1}(x_{t-1})$$

$$=\int p(x_{t-1}|x_t=z, y_{1:t-1})\hat{a}_t(z)\hat{b}_t(z)dz$$

$$=\int N(x_{t-1};\ \hat{x}_{t-1|t-1}+J_t(z-\hat{x}_{t|t-1})\ ,\ (I-J_tF)\hat{\Sigma}_{t-1|t-1})\ \ N(z;\hat{x}_{t|T},\hat{\Sigma}_{t|T})dz$$

$$= N(x_{t-1};\ \hat{x}_{t-1|t-1}+J_t(\hat{x}_{t|T}-\hat{x}_{t|t-1})\ ,\ J_t\hat{\Sigma}_{t|T}J_t+(I-J_tF)\hat{\Sigma}_{t-1|t-1}) \qquad (\Gamma 53)$$

Επομένως

$$\hat{x}_{t-1|T} = \hat{x}_{t-1|t-1}+J_t(\hat{x}_{t|T}-\hat{x}_{t|t-1}) \qquad (\Gamma 54)$$

$$\hat{\Sigma}_{t-1|T} = J_t\hat{\Sigma}_{t|T}J_t+(I-J_tF)\hat{\Sigma}_{t-1|t-1} \qquad (\Gamma 55)$$

Τέλος θα υπολογίσουμε την διασυνδιασπορά (cross-covariance) $\hat{\Sigma}_{t-1,t|T}$ ως προς $\hat{\Sigma}_{t|T}$
Αντικαθιστώντας τον όρο $c_t\hat{a}_t(x_t)$ από τη σχέση (Γ38) στην (Γ36) και με τη χρήση της (Γ52)

προκύπτει:

$$p(x_{t-1}, x_t|Y) = \frac{\hat{a}_{t-1}(x_{t-1})\, p(x_t|x_{t-1})\, p(y_t|x_t)\, \hat{a}_t(x_t)\, \hat{b}_t(x_t)}{c_t \hat{a}_t(x_t)}$$

$$= \frac{p(x_{t-1}|y_{1:t-1})\, p(x_t|x_{t-1}, y_{1:t-1})\, p(y_t|x_t)\, p(x_t|Y)}{p(y_t|x_t)\, p(x_t|y_{1:t-1})}$$

$$= \frac{p(x_t|y_{1:t-1})\, p(x_{t-1}|x_t, y_{1:t-1})\, p(x_t|Y)}{p(x_t|y_{1:t-1})}$$

$$= p(x_{t-1}|x_t, y_{1:t-1})\, p(x_t|Y)$$

$$= N(x_{t-1}; \hat{x}_{t-1|t-1} + J_t(x_t - \hat{x}_{t|t-1}),\ (I - J_t F)\hat{\Sigma}_{t-1|t-1})\ \ N(x_t; \hat{x}_{t|T}, \hat{\Sigma}_{t|T}) \quad (\Gamma 56)$$

Επομένως οριζουμε

$$p(x_{t-1}, x_t|Y) = N\left(\begin{bmatrix} \hat{x}_{t-1|T} \\ \hat{x}_{t|T} \end{bmatrix},\ \begin{bmatrix} \hat{\Sigma}_{t-1|T} & J_t \hat{\Sigma}_{t|T} \\ \hat{\Sigma}_{t|T} J_t^T & \hat{\Sigma}_{t|T} \end{bmatrix}\right) \qquad (\Gamma 57)$$

και

$$\hat{\Sigma}_{t-1,t|T} = J_t \hat{\Sigma}_{t|T} \quad .$$

3.4 Βασικές Ενέργειες στα Γραμμικά Δυναμικά Μοντέλα

Δεδομένου ότι έχουμε μία ή και παραπάνω ακολουθίες παρατηρήσεων και ένα μοντέλο, τρία είναι τα βασικά βήματα που έχουμε να υλοποιήσουμε. Αυτά είναι η εκτίμηση (evaluation) , η συμπερασματολογία (inference) και η εκμάθηση (learning). Θα αναλύσουμε καθένα από αυτά ξεχωριστά και θα παραθέσουμε τους αλγορίθμους που χρησιμοποιούνται.

3.4.1 Εκτίμηση (Evaluation)

Στο βήμα αυτό υπολογίζουμε την πιθανότητα μίας ακολουθίας παρατηρήσεων $Y(y_1 \dots y_T)$ να προήλθε από αυτό το μοντέλο. Η εκτίμηση και η συμπερασματολογία πραγματοποιούνται μέσω Forward-Backward αναδρομών. Πίο συγκεκριμένα για την εκτίμηση χρησιμοποιούμε Forward αναδρομή η οποία αποτελεί το φίλτρο Kalman.

Θεωρώντας γνωστές τις παραμέτρους του μοντέλου μπορούμε μέσω του αλγορίθμου 1 (Kalman filter) να υπολογίσουμε τη μέση τιμή και συνδιακύμανση των $p(x_t | y_{1:t-1})$ και $p(x_t | y_{1:t})$ καθώς και τη συνάρτηση πυκνότητας πιθανότητας $p(Y)$. Για την ακρίβεια υπολογίζουμε το $\log(p(Y))$. Ο λογάριθμος αυτός μπορεί να ερμηνευτεί ως log-likelihood των παραμέτρων του μοντέλου δεδομένου των παρατηρήσεων δηλαδή $\log(L(\theta|Y)) = \log(p(Y|\theta))$. Η μέση τιμή και συνδιακύμανση της κατάστασης τη χρονική στιγμή t δεδομένου των παρατηρήσεων $y_{1:t-1}$ είναι:

$$\hat{x}_{t|t-1} = E[x_t | y_{1:t-1}, \theta] \qquad (\Gamma 58)$$

$$\hat{\Sigma}_{t|t-1} = E[(x_t - \hat{x}_{t|t-1})(x_t - \hat{x}_{t|t-1})^T | y_{1:t-1}, \theta] \qquad (\Gamma 59)$$

Ενώ οι παράμετροι τη χρονική στιγμή t δεδομένου των παρατηρήσεων $y_{1:t}$ είναι:

$$\hat{x}_{t|t} = E[x_t | y_{1:t}, \theta] \qquad (\Gamma 60)$$

$$\hat{\Sigma}_{t|t} = E[(x_t - \hat{x}_{t|t})(x_t - \hat{x}_{t|t})^T | y_{1:t}, \theta] \qquad (\Gamma 61)$$

Στο γράφημα 9 απεικονίζεται ο αλγόριθμος 1 που είναι το φίλτρο Kamlan. Μπορούμε να παρατηρήσουμε ότι για την εκκίνηση της αναδρομής αρχικοποιούνται τα $\hat{x}_{t|t-1}$ και $\hat{\Sigma}_{t|t-1}$. Επίσης οι τιμές των $\hat{x}_{t|t-1}$, $\hat{\Sigma}_{t|t-1}$, $\hat{x}_{t|t}$ και $\hat{\Sigma}_{t|t}$ αποθηκεύονται καθώς θα χρησιμοποιηθούν στην Backward αναδρομή.

Data: Observations, $y_{1:T}$, and model parameters: $F, g, Q, H, \mu, R, g_1, Q_1$
Result: $\log\mathrm{L} = \log(p(y_{1:T}))$ and statistics $\hat{x}_{t|t}$, $\hat{\Sigma}_{t|t}$, $t \in \{1, \ldots, T\}$,
$\hat{x}_{t|t-1}$, $\hat{\Sigma}_{t|t-1}$, $t \in \{2, \ldots, T\}$

/* Initialization */
$\hat{x}_{t|t-1} = g_1; \quad \hat{\Sigma}_{t|t-1} = Q_1; \quad \log\mathrm{L} = 0$

for $t = 1{:}T$ **do**
  /* Prediction */
  if $t > 1$ **then**
    $\hat{x}_{t|t-1} = F\hat{x}_{t-1|t-1} + g$
    $\hat{\Sigma}_{t|t-1} = F\hat{\Sigma}_{t-1|t-1}F^T + Q$
  /* Update */
  $e_t = y_t - (H\hat{x}_{t|t-1} + \mu)$
  $\hat{\Sigma}_{e_t} = H\hat{\Sigma}_{t|t-1}H^T + R$
  $K_t = \hat{\Sigma}_{t|t-1}H^T\hat{\Sigma}_{e_t}^{-1}$
  $\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t e_t$
  $\hat{\Sigma}_{t|t} = \hat{\Sigma}_{t|t-1} - K_t H\hat{\Sigma}_{t|t-1}$
  $\log\mathrm{L} = \log\mathrm{L} + \log(\mathcal{N}(e_t; 0, \hat{\Sigma}_{e_t}))$ /* $c_t = \mathcal{N}(e_t; 0, \hat{\Sigma}_{e_t})$ */

Γράφημα 9: Αλγόριθμος 1 (Kalman Filter)

3.4.2 Εξαγωγή Συμπεράσματος (Inference)

Στο βήμα αυτό ουσιαστικά υπολογίζουμε την πιθανότητα το σύστημα να βρίσκεται σε μία κατάσταση z την χρονική στιγμή t δεδομένου των παρατηρήσεων δηλαδή $P(x_t = z | Y)$. Εφόσον ολοκληρώθηκε η Forward αναδρομή , πραγματοποιώντας Backward αναδρομή μπορούμε να υπολογίσουμε τα απαραίτητα στατιστικά στοιχεία των κρυφών καταστάσεων. Ο αλγόριθμος που χρησιμοποιείται είναι ο Kalman smoother ο οποίος επιστρέφει τις ροπές δεύτερης τάξης $\hat{R}_{t|T}$ και $\hat{R}_{t,t-1|T}$ και όχι τις συνδιακυμάνσεις καθώς αυτές θα χρησιμοποιηθούν παρακάτω για το βήμα της εκμάθησης. Η μέση τιμή και συνδιακύμανση της κατάστασης τη χρονική στιγμή t δεδομένου όλων των παρατηρήσεων Y είναι:

$$\hat{x}_{t|T} = E[x_t | Y, \theta] \qquad (\Gamma 62)$$

$$\hat{\Sigma}_{t|T} = E[(x_t - \hat{x}_{t|T})(x_t - \hat{x}_{t|T})^T | Y, \theta] \qquad (\Gamma 63)$$

Επίσης

$$\hat{\Sigma}_{t,t-1|T} = E[(x_t - \hat{x}_{t|T})(x_{t-1} - \hat{x}_{t-1|T})^T | Y, \theta] \qquad (\Gamma 64)$$

$$\hat{R}_{t|T} = E[x_t x_t^T | Y, \theta] \qquad (\Gamma 65)$$

$$\hat{R}_{t,t-1|T} = E[x_t x_{t-1}^T | Y, \theta] \qquad (\Gamma 66)$$

Οι σχέσεις μεταξύ $\hat{R}$ και $\hat{\Sigma}$ είναι:

$$\hat{R}_{t|T} = \hat{\Sigma}_{t|T} + \hat{x}_{t|T} \hat{x}_{t|T}^T$$

$$\hat{R}_{t,t-1|T} = \hat{\Sigma}_{t,t-1|T} + \hat{x}_{t|T} \hat{x}_{t-1|T}^T$$

Στο γράφημα 10 απεικονίζεται ο αλγόριθμος 2 που είναι ο Kalman Smoother. Δέχεται σαν είσοδο τα στατιστικά στοιχεία που υπολογίστηκαν από τον αλγόριθμο 1 και εξάγει τα $\hat{x}_{t|T}$, $\hat{R}_{t|T}$ και $\hat{R}_{t,t-1|T}$.

Data: Statistics $\hat{x}_{t|t}$, $\hat{\Sigma}_{t|t}$, $\hat{x}_{t|t-1}$, $\hat{\Sigma}_{t|t-1}$ calculated from Kalman filter, and model parameter F
Result: Statistics $\hat{x}_{t|T}$, $\hat{R}_{t|T}$, $t \in \{1, \ldots, T\}$ and $\hat{R}_{t,t-1|T}$, $t \in \{2, \ldots, T\}$

$\hat{R}_{T|T} = \hat{\Sigma}_{T|T} + \hat{x}_{T|T}\hat{x}_{T|T}^T$

for t = *T:-1:2* **do**

$J_t = \hat{\Sigma}_{t-1|t-1}F^T\hat{\Sigma}_{t|t-1}^{-1}$
$\hat{x}_{t-1|T} = \hat{x}_{t-1|t-1} + J_t(\hat{x}_{t|T} - \hat{x}_{t|t-1})$
$\hat{\Sigma}_{t-1|T} = \hat{\Sigma}_{t-1|t-1} + J_t(\hat{\Sigma}_{t|T} - \hat{\Sigma}_{t|t-1})J_t^T$
$\hat{\Sigma}_{t,t-1|T} = J_t\hat{\Sigma}_{t|T}$

$\hat{R}_{t-1|T} = \hat{\Sigma}_{t-1|T} + \hat{x}_{t-1|T}\hat{x}_{t-1|T}^T$
$\hat{R}_{t,t-1|T} = \hat{\Sigma}_{t,t-1|T} + \hat{x}_{t|T}\hat{x}_{t-1|T}^T$

Γράφημα 10: Αλγόριθμος 2 (Kalman Smoother)

Με την ολοκλήρωση των παραπάνω αλγορίθμων υπολογίζονται τα απαραίτητα στατιστικά στοιχεία. Τα στοιχεία αυτά εξάγονται για μία ακολουθία αλλά και για ένα σύνολο ακολουθιών και είναι απαραίτητα για το βήμα της εκμάθησης.

Τα στατιστικά στοιχεία πάνω σε μία ακολουθία παρατηρήσεων και μία ακολουθία κρυφών καταστάσεων και σε ένα σύνολο από ακολουθίες παρατηρήσεων D_y και ακολουθίες κρυφών καταστάσεων D_x απεικονίζονται στις στήλες ένα και δύο του πίνακα 1 αντίστοιχα.

Στατιστικά στοιχεία για 1 ακολουθία παρατηρήσεων και 1 ακολουθία καταστάσεων	**Στατιστικά στοιχεία για N ακολουθίες παρατηρήσεων και N ακολουθίες καταστάσεων**
$\zeta_1 = \sum_{t=1}^{T-1} \hat{x}_{t\|T}$ (Γ67)	$\zeta_0 = \sum_{l=1}^{N} \hat{x}_{l,1\|T_l}$ (Γ77)
$\zeta_2 = \sum_{t=2}^{T} \hat{x}_{t\|T}$ (Γ68)	$\zeta_1 = \sum_{l=1}^{N} \sum_{t=1}^{T_l-1} \hat{x}_{l,t\|T_l}$ (Γ78)
$\zeta_3 = \sum_{t=1}^{T} \hat{x}_{t\|T}$ (Γ69)	$\zeta_2 = \sum_{l=1}^{N} \sum_{t=2}^{T_l} \hat{x}_{l,t\|T_l}$ (Γ79)
$\zeta_4 = \sum_{t=1}^{T} y_t$ (Γ70)	$\zeta_3 = \sum_{l=1}^{N} \sum_{t=1}^{T_l} \hat{x}_{l,t\|T_l}$ (Γ80)
$\Gamma_1 = \sum_{t=1}^{T-1} \hat{R}_{t\|T}$ (Γ71)	$\zeta_4 = \sum_{l=1}^{N} \sum_{t=1}^{T_l} y_{l,t}$ (Γ81)
$\Gamma_2 = \sum_{t=1}^{T-1} \hat{R}_{t\|T}$ (Γ72)	$\Gamma_0 = \sum_{l=1}^{N} \hat{R}_{l,1\|T_l}$ (Γ82)
$\Gamma_3 = \sum_{t=1}^{T-1} \hat{R}_{t\|T}$ (Γ73)	$\Gamma_1 = \sum_{l=1}^{N} \sum_{t=1}^{T_l-1} \hat{R}_{l,t\|T_l}$ (Γ83)
$\Gamma_4 = \sum_{t=2}^{T} \hat{R}_{t,t-1\|T}$ (Γ74)	$\Gamma_2 = \sum_{l=1}^{N} \sum_{t=2}^{T_l} \hat{R}_{l,t\|T_l}$ (Γ84)
$\Gamma_5 = \sum_{t=1}^{T} y_t \hat{x}_{t\|T}^T$ (Γ75)	$\Gamma_3 = \sum_{l=1}^{N} \sum_{t=1}^{T_l} \hat{R}_{l,t\|T_l}$ (885)
$\Gamma_6 = \sum_{t=1}^{T} y_t y_t^T$ (Γ76)	$\Gamma_4 = \sum_{l=1}^{N} \sum_{t=2}^{T_l} \hat{R}_{l,t,t-1\|T_l}$ (Γ86)
	$\Gamma_5 = \sum_{l=1}^{N} \sum_{t=1}^{T_l} y_t \hat{x}_{l,t\|T_l}^T$ (Γ87)
	$\Gamma_6 = \sum_{l=1}^{N} \sum_{t=1}^{T_l} y_{l,t} y_{l,t}^T$ (Γ88)

Πίνακας 1: Στατιστικά Στοιχεία

3.4.3 Εκμάθηση (Learning)

Το βήμα της εκπαίδευσης στα LDM μπορεί να υλοποιηθεί με τη χρήση του EM αλγορίθμου (Expectation Maximization). Για να βρούμε τις ML (maximum likelihood) εκτιμήσεις όσον αφορά τις παραμέτρους του μοντέλου πρέπει να μεγιστοποιήσουμε την από κοινού log-likelihood των δεδομένων $L(\theta)=\log p(Y|\theta)$ και των κρυφών μεταβλητών X .

$$L(\theta)=\log p(Y|\theta)=\log\left(\int_X p(X,Y|\theta)dX\right) \qquad (\Gamma 89)$$

Χρησιμοποιώντας οποιοδήποτε κατανομή $q(X)$ που αφορά τις κρυφές μεταβλητές μπορούμε να ορίσουμε ένα κάτω φράγμα για το L . Οπότε προκύπτει η σχέση (Γ90).

$$L(\theta)=\log\left(\int_X q(X)\frac{p(X,Y|\theta)}{q(X)}dX\right) \qquad (\Gamma 90)$$

Από την ανισότητα του Jensen έχουμε:

$$L(\theta)\geq\int_X q(X)\log\left(\frac{p(X,Y|\theta)}{q(X)}\right)dX$$

$$L(\theta)\geq\int_X q(X)\log p(X,Y|\theta)-\int_X q(X)\log q(X)dX \qquad (\Gamma 91)$$

Ορίζουμε την συνάρτηση F .

$$F(q,\theta)=\int_X q(X)\log p(X,Y|\theta)-\int_X q(X)\log q(X)dX \qquad (\Gamma 92)$$

Ο ΕΜ αλγόριθμος αποτελείται από δύο μέρη. Το πρώτο μέρος καλείται *E-step* και το δεύτερο *M-step* . Στο *E-step* μεγιστοποιούμε την κατανομή q διατηρώντας σταθερές τις παραμέτρους του μοντέλου θ (Γ93). Για να το πετύχουμε αυτό επιλέγουμε ως δεσμευμένη κατανομή την posterior της ακολουθίας των καταστάσεων δεδομένου των παρατηρήσεων αλλά και των παραμέτρων του μοντέλου (Γ94). Πρέπει να διευκρινίσουμε ότι κατά την εκκίνηση του ΕΜ αλγορίθμου αρχικοποιούμε τις παραμέτρους του μοντέλου. Στο *M-step* μεγιστοποιούμε τις παραμέτρους θ διατηρώντας την κατανομή q σταθερή (Γ95). Εφόσον μόνο ο πρώτος όρος της σχέσης (Γ92) εξαρτάται από τις παραμέτρους μας αρκεί να μελετήσουμε αυτόν (Γ96). Στη συνέχεια ορίζουμε μία βοηθητική συνάρτηση η οποία θα χρησιμοποιηθεί για να βρούμε τις νέες παραμέτρους του μοντέλου.

E-step:

$$q_{i+1} \leftarrow \arg\max_{q} F(q_{i+1}, \theta) \qquad (\Gamma 93)$$

$$q*_{i+1}(X) = p(X \mid Y, \theta_i) \qquad (\Gamma 94)$$

M-step:

$$\theta_{i+1} \leftarrow \arg\max_{\theta} F(q_{i+1}, \theta) \qquad (\Gamma 95)$$

$$\Rightarrow \theta*_{i+1} \leftarrow \arg\max_{\theta} \int_X p(X \mid Y, \theta_i) \log p(X, Y \mid \theta)\, dX \qquad (\Gamma 96)$$

Η βοηθητική συνάρτηση που θα χρησιμοποιήσουμε για την εύρεση των νέων παραμέτρων του μοντέλου είναι η *Q*.

$$Q(\theta_i, \theta) = E[\log p(X, Y \mid \theta) \mid Y, \theta_i] = \int_X p(X \mid Y, \theta_i) \log p(X, Y \mid \theta)\, dX \quad (\Gamma 97)$$

λόγω της σχέσης (Γ16) προκύπτει:

$$\log p(X,Y|\theta) = \log p(x_1|\theta) + \sum_{t=1}^{T-1} \log p(x_{t+1}|x_t) + \sum_{t=1}^{T} \log p(y_t|x_t) \qquad (\Gamma 98)$$

Από τις σχέσεις (Γ97),(Γ98) προκύπτει:

$$\begin{aligned} Q(\theta_i,\theta) = E[\log p(x_1|\theta)|Y,\theta_i] + \sum_{t=1}^{T-1} E[\log p(x_{t+1}|x_t,\theta)|Y,\theta_i] \\ + \sum_{t=1}^{T} E[\log p(y_t|x_t,\theta)|Y,\theta_i] \end{aligned} \qquad (\Gamma 90)$$

Στην περίπτωση που μελετάμε γραμμικά δυναμικά μοντέλα η βοηθητική συνάρτηση (Γ90) ανάγεται στη σχέση (Γ91)

$$\begin{aligned} Q(\theta_i,\theta) = const - \frac{1}{2}|Q_1| - \frac{1}{2}E[(x_1-g_1)^T Q_1^{-1}(x_1-g_1)|Y,\theta_i] - \frac{T-1}{2}\log|Q| \\ - \frac{1}{2}\sum_{t=2}^{T} E[(x_t - F x_{t-1} - g)^T Q_1^{-1}(x_t - F x_{t-1} - g)|Y,\theta_i] \\ - \frac{T}{2}\log|R| - \frac{1}{2}\sum_{t=1}^{T} E[(y_t - H x_t - \mu)^T R_1^{-1}(y_t - H x_t - \mu)|Y,\theta_i] \end{aligned} \qquad (\Gamma 91)$$

Με βάση την παραπάνω βοηθητική συνάρτηση παραγωγίζοντας ως προς την κάθε παράμετρο του μοντέλου και εξισώνοντας με 0 βρίσκουμε τις καινούριες σχέσεις τους. Ακολουθούν οι σχέσεις που περιγράφουν πώς ενημερώνεται κάθε παραμέτρος για μία ακολουθία παρατηρήσεων.

Σχέση για ενημέρωση παραμέτρου g_1:

$$\frac{\partial Q(\theta_i,\theta)}{\partial g_1} = -\frac{1}{2} 2 Q_1^{-1} E([x_1|Y,\theta_i] - g1) = 0$$

$$g_1 = \hat{x}_{1|T} \qquad (92\alpha)$$

Σχέση για ενημέρωση παραμέτρου Q_1:

$$\frac{\partial Q(\theta_i,\theta)}{\partial Q_1^{-1}} = \frac{1}{2}Q_1 - \frac{1}{2}E\left[(x_1-g_1)(x_1-g_1)^T \mid Y,\theta_i\right] = 0$$

$$Q_1 = \hat{R}_{1|T} - g_1 g_1^T \qquad (\Gamma 93a)$$

Σχέση για ενημέρωση παραμέτρου g:

$$\frac{\partial Q(\theta_i,\theta)}{\partial g} = -\frac{1}{2}(-2)Q^{-1}\sum_{t=2}^{T} E\left[x_t - F\,x_{t-1} - g \mid Y,\theta_i\right] = 0 \;\Rightarrow$$

$$\Rightarrow \;\; (T-1)\,g = \sum_{t=2}^{T} E\left[x_t \mid Y,\theta_i\right] - F\sum_{t=2}^{T} E\left[x_{t-1} \mid Y,\theta_i\right]$$

$$g = \frac{1}{(T-1)}\left(\zeta_2 - F\zeta_1\right) \qquad (\Gamma 94\alpha)$$

Σχέση για ενημέρωση παραμέτρου F:

$$\frac{\partial Q(\theta_i,\theta)}{\partial F} = -\frac{1}{2}(-2)Q^{-1}\sum_{t=2}^{T} E\left[\left(x_t - F\,x_{t-1} - g\right)x_{t-1}^T \mid Y,\theta_i\right] = 0$$

$$\Rightarrow \sum_{t=2}^{T} E\left[x_t x_{t-1}^T \mid Y,\theta_i\right] - F\sum_{t=2}^{T} E\left[x_{t-1}x_{t-1}^T \mid Y,\theta_i\right] - g\sum_{t=2}^{T} E\left[x_{t-1}^T \mid Y,\theta_i\right] = 0$$

$$\Rightarrow \sum_{t=2}^{T} \hat{R}_{t,t-1|T} - F\sum_{t=1}^{T-1} \hat{R}_{t|T} - g\sum_{t=1}^{T-1} \hat{x}_{t|T} = 0$$

$$\Rightarrow \;\; \Gamma_4 - F\Gamma_1 - \frac{1}{T-1}\zeta_2\zeta_1^T + F\frac{1}{T-1}\zeta_1\zeta_1^T = 0$$

$$F = (\Gamma_4 - \frac{1}{T-1}\zeta_2\zeta_1^T)(\Gamma_1 - \frac{1}{T-1}\zeta_1\zeta_1^T)^{-1} \qquad \Gamma(95\alpha)$$

Σχέση για ενημέρωση παραμέτρου *Q*:

$$\frac{\partial Q(\theta_i,\theta)}{\partial Q^{-1}} = -\frac{1}{2}(-2)Q^{-1}\sum_{t=2}^{T} E[(x_t - F\,x_{t-1} - g)x_{t-1}^T | Y,\theta_i] = 0 \quad \Rightarrow$$

$$\sum_{t=2}^{T} E[x_t x_{t-1}^T | Y,\theta_i] - F\sum_{t=2}^{T} E[x_{t-1}x_{t-1}^T | Y,\theta_i] - g\sum_{t=2}^{T} E[x_{t-1}^T | Y,\theta_i] = 0 \quad \Rightarrow$$

$$\sum_{t=2}^{T} \hat{R}_{t,t-1|T} - F\sum_{t=1}^{T-1} \hat{R}_{t|T} - g\sum_{t=1}^{T-1} \hat{x}_{t|T} = 0 \quad \Rightarrow$$

$$Q = \frac{1}{T-1}(\Gamma_2 - F\,\Gamma_4^T - g\,\zeta_2^T) \qquad \Gamma(96\alpha)$$

Σχέση για ενημέρωση παραμέτρου *μ*:

$$\frac{\partial Q(\theta_i,\theta)}{\partial \mu} = -\frac{1}{2}(-2)Q^{-1}\sum_{t=2}^{T} E[(x_t - F\,x_{t-1} - g)x_{t-1}^T | Y,\theta_i] = 0$$

$$\mu = \frac{1}{T}\left(\sum_{t=1}^{T} y_t - H\sum_{t=1}^{T} \hat{x}_{t|T}\right) \quad \Rightarrow$$

$$\mu = \frac{1}{T}(\zeta_4 - H\,\zeta_3) \qquad (\Gamma 97\alpha)$$

Σχέση για ενημέρωση παραμέτρου *H*:

$$\frac{\partial Q(\theta_i,\theta)}{\partial H}=-\frac{1}{2}(-2)R^{-1}\sum_{t=1}^{T}E[(y_t-H\,x_t-\mu)x_t^T|Y,\theta_i]=0 \quad \Rightarrow$$

$$\sum_{t=1}^{T}y_t E[x_t|Y,\theta_i]-\sum_{t=1}^{T}E[x_t x_t^T|Y,\theta_i]-\mu\sum_{t=1}^{T}E[x_t^T|Y,\theta_i]=0 \quad \Rightarrow$$

$$H=(\Gamma_5-\frac{1}{T}\zeta_4\zeta_3^T)(\Gamma_3-\frac{1}{T}\zeta_3\zeta_3^T)^{-1} \qquad (\Gamma 98\alpha)$$

Σχέση για ενημέρωση παραμέτρου *R*:

$$\frac{\partial Q(\theta_i,\theta)}{\partial R^{-1}}=\frac{T}{2}R-\frac{1}{2}\sum_{t=1}^{T}E[(y_t-H\,x_t-\mu)(y_t-H\,x_t-\mu)^T|Y,\theta_i]=0 \quad \Rightarrow$$

$$R=\frac{1}{T}\left(\sum_{t=1}^{T}y_t y_t^T-H\sum_{t=1}^{T}\hat{x}_{t|T}y_t^T-\mu\sum_{t=1}^{T}y_t^T\right) \quad \Rightarrow$$

$$R=\frac{1}{T}(\Gamma_6-H\,\Gamma_5^T\mu\,\zeta_4^T) \qquad (\Gamma 99\alpha)$$

Στο πίνακα 3 παραθέτονται οι συναρτήσεις για την εύρεση των νέων παραμέτρων του γραμμικού δυναμικού μοντέλου για μία ακολουθία παρατηρήσεων και για ένα σύνολο από Ν ακολουθίες.

Παράμε/τρος	**Σχέση ενημέρωσης για ακολουθία παρατηρήσεων μήκους Τ**	**Σχέση ενημέρωσης για Ν ακολουθίες παρατηρήσεων**
g_1	$g_1 = \hat{x}_{1\|T}$ (Γ92α)	$g_1 = \frac{1}{N}\zeta_0$ (Γ92β)
Q_1	$Q_1 = \hat{R}_{1\|T} - g_1 g_1^T$ (Γ93α)	$Q_1 = \frac{1}{N}\Gamma_0 - g_1 g_1^T$ (Γ93β)
g	$g = \frac{1}{(T-1)}(\zeta_2 - F\zeta_1)$ (Γ94α)	$g = \frac{1}{\sum_{l=1}^{N} T_l - N}(\zeta_2 - F\zeta_1)$ (Γ94β)
F	$F = (\Gamma_4 - \frac{1}{T-1}\zeta_2\zeta_1^T)(\Gamma_1 - \frac{1}{T-1}\zeta_1\zeta_1^T)^{-1}$ (Γ95α)	$F = (\Gamma_4 - \frac{1}{\sum_{l=1}^{N} T_l - N}\zeta_2\zeta_1^T)(\Gamma_1 - \frac{1}{\sum_{l=1}^{N} T_l - N}\zeta_1\zeta_1^T)^{-1}$ (Γ95β)
Q	$Q = \frac{1}{T-1}(\Gamma_2 - F\Gamma_4^T - g\zeta_2^T)$ (Γ96α)	$Q = \frac{1}{\sum_{l=1}^{N} T_l - N}(\Gamma_2 - F\Gamma_4^T - g\zeta_2^T)$ (Γ96β)
μ	$\mu = \frac{1}{T}(\zeta_4 - H\zeta_3)$ (Γ97α)	$\mu = \frac{1}{\sum_{l=1}^{N} T_l}(\zeta_4 - H\zeta_3)$ (Γ97β)
H	$H = (\Gamma_5 - \frac{1}{T}\zeta_4\zeta_3^T)(\Gamma_3 - \frac{1}{T}\zeta_3\zeta_3^T)^{-1}$ (Γ98α)	$H = (\Gamma_5 - \frac{1}{\sum_{l=1}^{N} T_l}\zeta_4\zeta_3^T)(\Gamma_3 - \frac{1}{\sum_{l=1}^{N} T_l}\zeta_3\zeta_3^T)^{-1}$ (Γ98β)
R	$R = \frac{1}{T}(\Gamma_6 - H\Gamma_5^T\mu\zeta_4^T)$ (Γ99α)	$R = \frac{1}{\sum_{l=1}^{N} T_l}(\Gamma_6 - H\Gamma_5^T\mu\zeta_4^T)$ (Γ99β)

Πίνακας 3: Σχέσεις για την ενημέρωση των παραμέτρων στο γραμμικό δυναμικό μοντέλο

Κεφάλαιο 4

Σύνθεση Φωνής με Γραμμικά Δυναμικά Μοντέλα

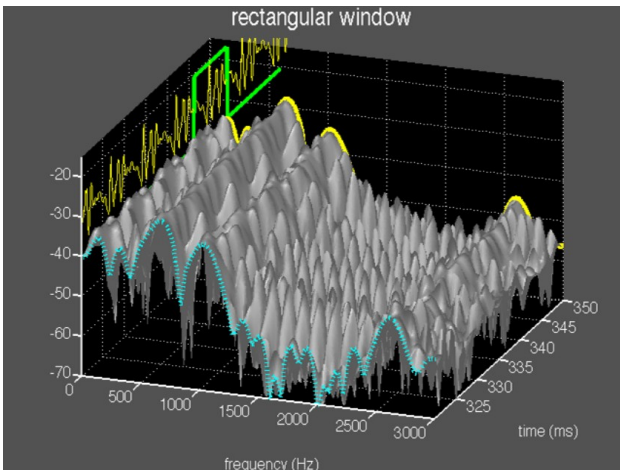
4.1 Εξαγωγή Ακουστικών Χαρακτηριστικών

Η βάση δεδομένων που χρησιμοποιήθηκε για τη σύνθεση φωνής είναι η Cmu_us_arctic_slt (female), η οποία περιέχει 1131 ηχογραφημένα μηνύματα δειγματοληπτημένα στα 16kHz συνολικής διάρκειας περίπου μίας ώρας. Η φωνή είναι γυναικεία και το F0 κυμαίνεται από τα 100Hz μέχρι και τα 350Hz . Χρησιμοποιώντας το Straight εξάγαμε τα ακουστικά χαρακτηριστικά τα οποία είναι το F0 και το aperiodicity (το spectrum δεν χρειάζεται να το χρησιμοποιήσουμε καθώς θα το κατασκευάσουμε εμείς). Αυτά σε συνδυασμό με το spectrum που θα παράξουμε με την διαδικασία που θα αναλύσουμε παρακάτω χρησιμοποιούνται από τον vocoder για την παραγωγή της συνθετικής φωνής.

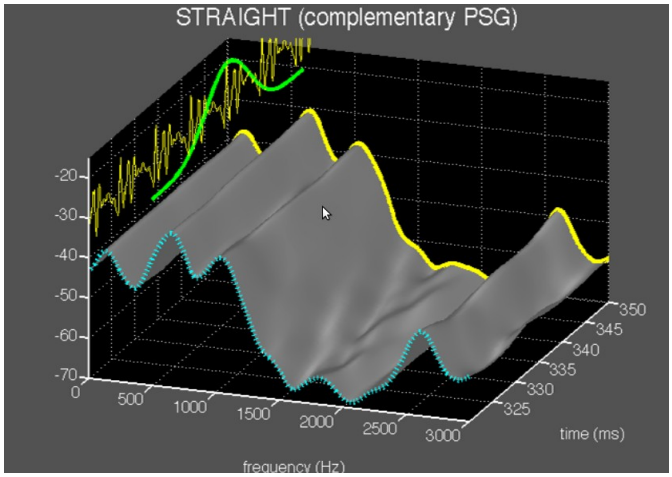
4.2 Straight Vocoder

Το Straight είναι ένας Vocoder υψηλής ποιότητας ο οποίος αποτελείται από τρία βασικά βήματα. Αυτά είναι ο υπολογισμός φασματικού φακέλου (φάσματος), η εξαγωγή θεμελιώδους συχνότητας από στιγμιαία συχνότητα και η πρόσδοση φυσικότητας μέσω του group delay.

Στο πρώτο βήμα αρχικά εξάγεται το φάσμα το οποίο παρουσιάζει ασυνέχειες στην συχνότητα και στο χρόνο. Με την εφαρμογή κατάλληλων παραθύρων πρώτα εξομαλύνονται οι ασυνέχειες στο πεδίο της συχνότητας και στη συνέχεια στο πεδίο του χρόνου. Το πρόβλημα όμως που προκύπτει είναι το oversmoothing λόγω των παραθύρων που εφαρμόστηκαν, οπότε με τη χρήση κατάλληλου αντισταθμιστικού παραθύρου αντιμετωπίζεται το πρόβλημα αυτό. Στα γραφήματα 11α και 11β μπορούμε να παρατηρήσουμε το αποτέλεσμα της εξομάλυνσης πάνω στο αρχικό φάσμα που εξάγεται.



Γράφημα 11α: Αρχικό Φάσμα



Γράφημα 11β: Εξομαλυμένο Φάσμα

Στο δεύτερο βήμα γίνεται ο υπολογισμός του F0 (pitch) μέσω της στιγμιαίας συχνότητας και fixed-point αλγορίθμων. Είναι πολύ σημαντικό να εξάγουμε μία ακριβή και αξιόπιστη θεμελιώδη συχνότητα F0 ώστε να μπορέσουμε να παράξουμε φυσική ομιλία. Αρχικά με τη χρήση ζωνοπερατών φίλτρων και την εφαρμογή τους στον άξονα των συχνοτήτων μπορούμε να περιορίσουμε τις παρεμβολές από άλλες αρμονικές συχνότητες. Τα φίλτρα αυτά είναι αποτέλεσμα συνέλιξης ενός Gabor φίλτρου με ένα cardinal B spline φίλτρο που έχει οριστεί ως προς ένα υποθετικό F0. Στη συνέχεια με χρήση fixed-point αλγορίθμων εξάγουμε ένα σετ από υποψήφια σημεία τα οποία σχετίζονται με την θεμελιώδη συχνότητα. Εφόσον έχουμε εντοπίσει τα σημεία αυτά υπολογίζουμε τη C/N αναλογία (carrier-to-noise ratio) και επιλέγουμε τα fixed-points με τη μεγιστοποιούν. Το πλεονέκτημα της χρήσης αυτής της μεθόδου είναι ότι παράλληλα με τη θεμελιώδη συχνότητα εξάγουμε και το aperiodicity καθώς τα γειτονικά fixed-points μπορούν να υποδείξουν την ύπαρξη έμφωνων και άφωνων περιοχών. Ουσιαστικά το aperiodicity είναι η ενέργεια στις μη αρμονικές συχνότητες ομαλοποιημένη ως προς την συνολική ενέργεια.

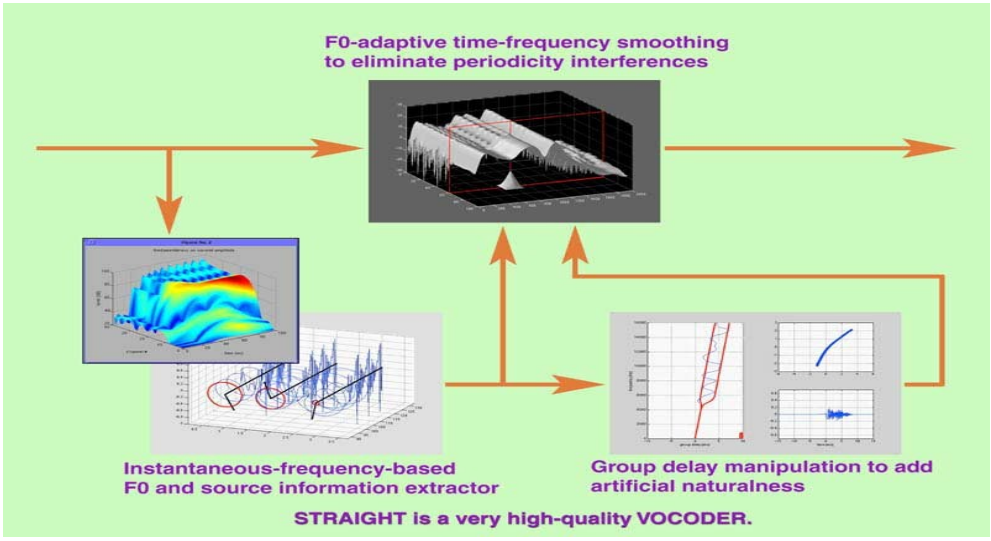
Εκτός όμως από την εφαρμογή fixed-point αλγορίθμων στη συχνότητα έχουμε και εφαρμογή στο χρόνο. Ο λόγος που χρειάζεται αυτή η διαδικασία είναι προκειμένου να οριστεί η χρονική στιγμή που συμβαίνει η κάθε διέγερση δηλαδή να βρούμε τα στιγμιότυπα. Στο βήμα αυτό πρέπει να αναφέρουμε κάποιες βασικές αρχές πάνω στα σήματα ομιλίας. Τα σήματα ομιλίας διεγείρονται από ποικίλους παράγοντες ορισμένοι από τους οποίους είναι οι περιοδικές διακυμάνσεις στη ροή του αέρα λόγω δονήσεων των φωνητικών χορδών, ο στροβιλισμός του αέρα που παράγει τυχαίους θορύβους και η απότομη ροή αέρα στη φωνητική οδό. Για παράδειγμα στις

υψηλές συχνότητες ο παράγοντας που οδηγεί στο σχηματισμό των φωνηέντων είναι η ασυνέχεια της ροή του αέρα από το κλείσιμο των φωνητικών χορδών. Σαν αποτέλεσμα διαφορετικά φάσματα προκύπτουν ανάλογα με τον ήχο (πχ η περιβάλλουσα ενός φωνήεντος μοιάζει με ενός στιγμιαίου συμφώνου αλλά διαφέρει από ενός εξακολουθητικού καθώς το φάσμα του τελευταίου μεταβάλλεται τυχαία αφού το σήμα αυτό προσεγγίζεται από ένα συνεχόμενο σταθερό θόρυβο). Ένας τρόπος για να αναπαραστήσουμε αυτές τις καταστάσεις είναι ερμηνεύοντας τον ήχο ως μία συλλογή στιγμιοτύπων (events collection). Ορίζουμε ως στιγμιότυπο την συγκέντρωση ενέργειας στο χρόνο. Επειδή όμως η ομιλία αποτελείται από πολλαπλά στιγμιότυπα είναι απαραίτητο να απομονώσουμε καθένα από αυτά εφαρμόζοντας παράθυρα στο πεδίο του χρόνου. Ομοίως με πριν εφαρμόζουμε κατάλληλα παράθυρα και στη συνέχεια επιλέγουμε τα υποψήφια σημεία. Να σημειώσουμε ότι εκτός από τον εντοπισμό των στιγμιοτύπων διέγερσης παίρνουμε πληροφορία σχετικά με τη διάρκεια που έχει η κάθε διέγερση.

Το τελευταίο βήμα είναι η εφαρμογή του group delay. Είναι σημαντικό να κατανοήσουμε ότι υπάρχει διαφορά μεταξύ της χρονικής στιγμής που ξεκίνησε η διέγερση με αυτήν που την αντιλαμβάνεται ο άνθρωπος. Οπότε εισάγοντας μία καθυστέρηση που ανταποκρίνεται σε αυτήν την διαφορά μπορούμε να προσδώσουμε ακρίβεια στην διαδικασία αυτή.

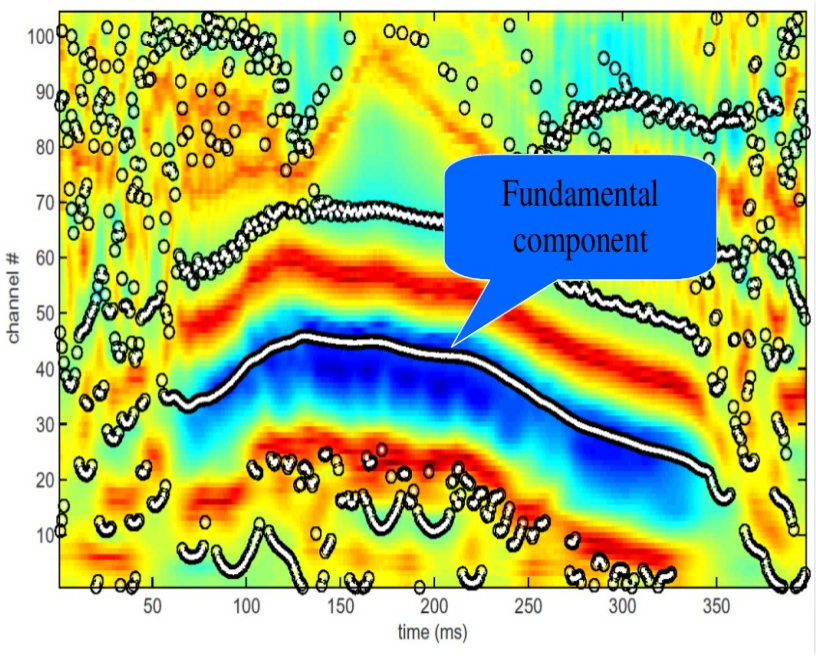
Τελικά έχουμε καταφέρει να εξάγουμε το F0, το φάσμα (spectrum) και το aperiodicity τα οποία μπορούμε να τα τροποποιήσουμε και στη συνέχεια να τα χρησιμοποιήσουμε για την σύνθεση της φωνής.

Στα παρακάτω γραφήματα παρουσιάζεται η βασική δομή του Straight καθώς και η εξαγωγή των βασικών χαρακτηριστικών.

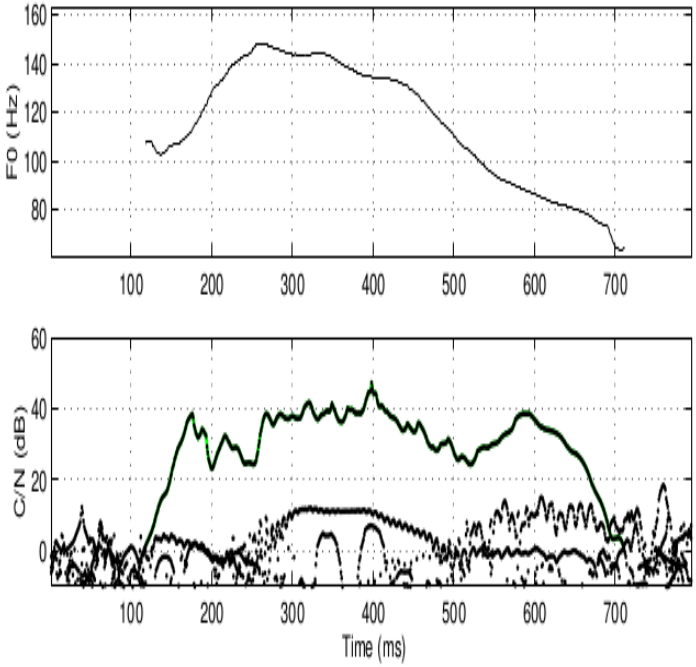


Γράφημα 12: Σχηματικό Διάγραμμα του Straight

Τα διαγράμματα 13α και 13β προκύπτουν από το demo του Straight το οποίο χρησιμοποιεί την πρόταση /aiueo/ για την εξαγωγή των χαρακτηριστικών.



Γράφημα 13α : Εξαγωγή F0 με βάση την αναλογία C/N
(με μπλε συμβολίζονται οι μεγάλες τιμές C/N)



Γράφημα 13β: F0 trajectory (top) , Aperiodicity trajectory (bottom/green line)

4.3 Labels of Dataset

Το δεύτερο βήμα είναι η αντιστοίχηση των φωνητικών χαρακτηριστικών με τα γλωσσικά χαρακτηριστικά (linguistic-acoustic mapping). Να αναφέρουμε ότι η δομή αυτή δημιουργήθηκε από προσωπικό του εργαστηρίου και περιέχει στοιχεία για τα context-dependent phonemes (decision-tree clustering) αλλά και για την εκπαίδευση του μοντέλου. Οπότε δεν θα αναφέρουμε επιπλέον στοιχεία για το βήμα αυτό.

4.4 Μοντελοποίηση του Γραμμικού Δυναμικού Μοντέλου

Για την μοντελοποίηση του LDM ακολουθάμε την μεθοδολογία που αναγράφεται αναλυτικά στο κεφάλαιο 3 δηλαδή υλοποιούμε τα τρία βασικά βήματα (evaluation , inference και learning). Αφού κατασκευάσαμε το μοντέλο το εκπαιδεύσαμε με τη δομή που μας δόθηκε οπότε δημιουργήσαμε τα context-dependent phoneme LDMs. Για την εκπαίδευση χρησιμοποιήθηκαν και οι 1131 προτάσεις που υπάρχουν στη βάση εκπαίδευσης Cmu_us_arctic_slt (female). Τα παραπάνω βήματα αφορούν το πεδίο της εκπαίδευσης.

4.5 Δημιουργία Διανύσματος Ακουστικών Χαρακτηριστικών

Στο πεδίο της σύνθεσης τα context-dependent μοντέλα σε συνδυασμό με τα αντίστοιχα labels που εξάγονται από κάθε πρόταση που θέλουμε να συνθέσουμε παράγουν τα φωνητικά χαρακτηριστικά (mgc – Generalized Mel Cepstrum) .

Στην εργασία αυτή δημιουργήσαμε 25 συνθετικές προτάσεις οπότε αντίστοιχα χρειαστήκαμε 25 labels. Μία πρόταση που θέλουμε να συνθέσουμε μπορεί να αποτυπωθεί σαν μία ακολουθία από φωνήματα τα οποία αντιστοιχίζονται σε context-dependent φωνήματα. Κάθε label περιλαμβάνει πληροφορία σχετικά με αυτήν την ακολουθία των context-dependent φωνημάτων δηλαδή το μοντέλο (από αυτά που εκπαιδεύσαμε) που αντιστοιχεί σε κάθε φώνημα καθώς και τη διάρκειά του φωνήματος. Επομένως από μία ακολουθία context-dependent φωνημάτων δημιουργήσαμε μία ακολουθία από context-dependent μοντέλων τα οποία θα χρησιμοποιηθούν για την εξαγωγή των φωνητικών χαρακτηριστικών (mgc-generalized mel cepstrum).

Έχουμε δημιουργήσει ένα γραμμικό δυναμικό μοντέλο το οποίο αποτελείται από

διαφορετικά context-dependent μοντέλα οπότε μένει να εξάγουμε την ακολουθία χαρακτηριστικών. Για να επιτευχθεί αυτό εφαρμόζουμε τον κανόνα ML εφόσον το μοντέλο είναι γνωστό δηλαδή γνωρίζουμε τις παραμέτρους του. Η ακολουθία τον παρατηρήσεων δίνεται από την παρακάτω σχέση:

$$p(Y|\theta)=\int_X p(X,Y|\theta)dX=\int_X p(Y|X,\theta)\,p(X|\theta)dX \qquad (\Delta 1)$$

Μπορούμε βρίσκοντας την βέλτιστη ακολουθία καταστάσεων $\hat{X}=[\hat{x}_1,\hat{x}_2,\ldots,\hat{x}_T]$ να μειώσουμε το υπολογιστικό κόστος. Οπότε το πρόβλημα μας ανάγεται σε δύο βήματα, το πρώτο είναι η εύρεση της βέλτιστης ακολουθίας καταστάσεων και το δεύτερο είναι βάση της ακολουθίας αυτής η εύρεση των παρατηρήσεων. Αυτό αποτυπώνεται στις παρακάτω σχέσεις.

$$\hat{X}=\arg\max_X p(X|\theta)=\arg\max_X p(x_1|\theta)\prod_{t=2}^{T} p(x_1|x_{t-1},\theta)$$

$$\hat{X}=\arg\max_X p(X|\theta)=\arg\max_X N(x_1;g_1,Q_1)\prod_{t=2}^{T} N(x_t;F\,x_{t-1}+g,Q) \qquad (\Delta 2)$$

Γνωρίζουμε ότι η ML εκτίμηση μίας Gaussian κατανομής είναι η μέση τιμή της οπότε οι σχέσεις που περιγράφουν την ακολουθία των καταστάσεων είναι οι εξής:

$$\hat{x}=g_1 \qquad (\Delta 3)$$

$$\hat{x}_t=F\,x_{t-1}+g \quad , \quad t\in\{2,\ldots,T\} \qquad (\Delta 4)$$

Εφόσον υπολογίσαμε την ακολουθία καταστάσεων εξάγουμε τις παρατηρήσεις.

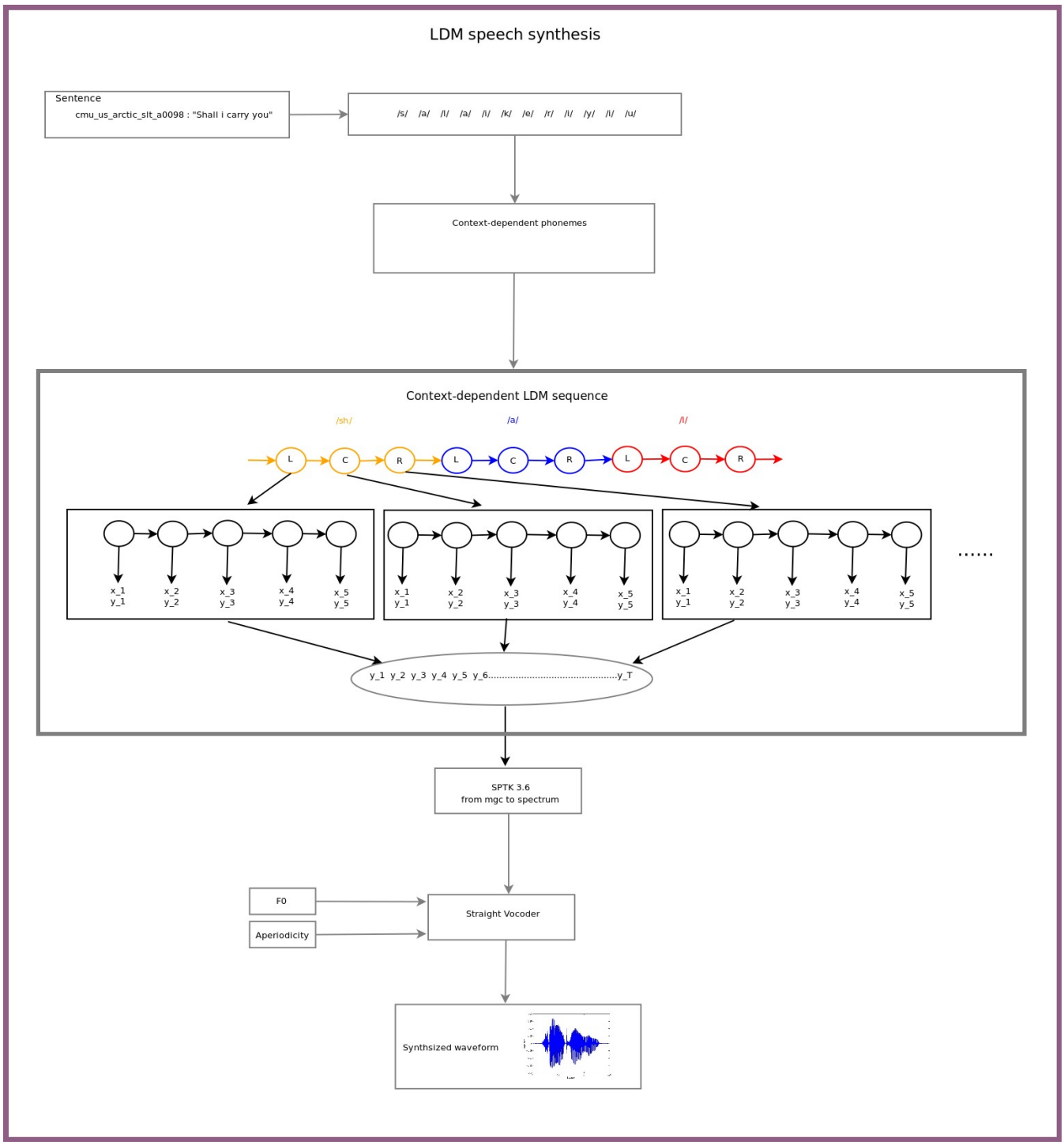
$$p(Y|\hat{X},\theta)=\prod_{t=1}^{T} p(y_t|x_t,\theta)$$

$$p(Y \mid \hat{X}, \theta) = \prod_{t=1}^{T} N(y_t ; H x_t + \mu, R) \qquad (\Delta 5)$$

Επομένως για την ακολουθία των παρατηρήσεων έχουμε την σχέση:

$$y_t = H \hat{x}_t + \mu \quad , \quad t \in \{ 1 , \ldots , T \} \qquad (\Delta 6)$$

Το διάνυσμα των παρατηρήσεων που έχουμε εξάγει αντιστοιχεί στο διάνυσμα ακουστικών χαρακτηριστικών (generalized mel cepstrum) που θα χρησιμοποιήσουμε παρακάτω για τη σύνθεση τη φωνής. Να αναφέρουμε ότι για κάθε πρόταση παράχθηκαν 50 generalized mel cepstrum coefficients, δηλαδή το μέγεθος του διανύσματος *y* είναι 50. Να σημειώσουμε ότι όσο μεγαλύτερο είναι το μέγεθος του *y* τόσο καλύτερα κωδικοποιείται η φωνή με την βοήθεια των mcep. Ωστόσο μεγάλο y δημιουργεί προβλήματα στην εκμάθηση των παραμέτρων των στατιστικών μοντέλων (LDMs , HMMs, κτλ) επομένως επιλέγουμε $40 < y < 50$ για την σύνθεση φωνής. Στο γράφημα 14 απεικονίζεται η διαδικασία της σύνθεσης για μία από τις προτάσεις που συνθέσαμε.

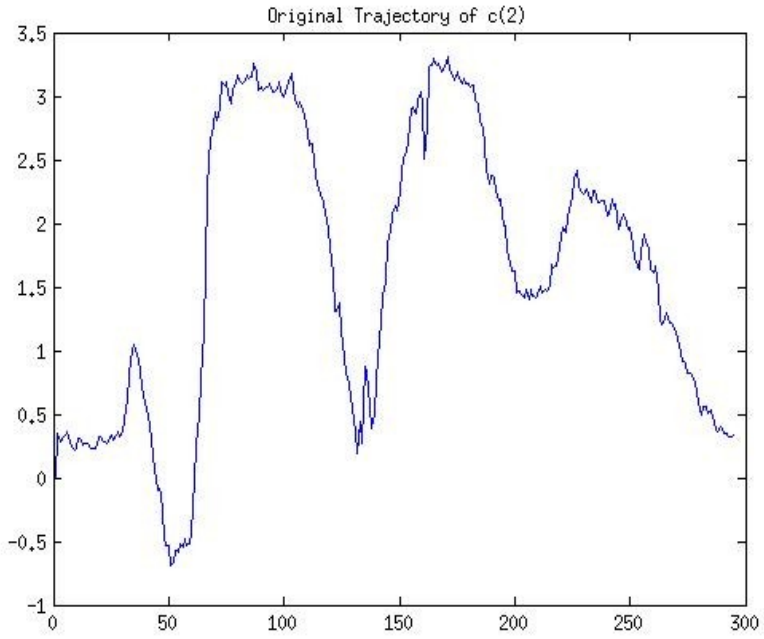


Γράφημα 14: Σχηματικό Διάγραμμα σύνθεσης φωνής με γραμμικά δυναμικά μοντέλα (μόνο για το τμήμα της σύνθεσης)

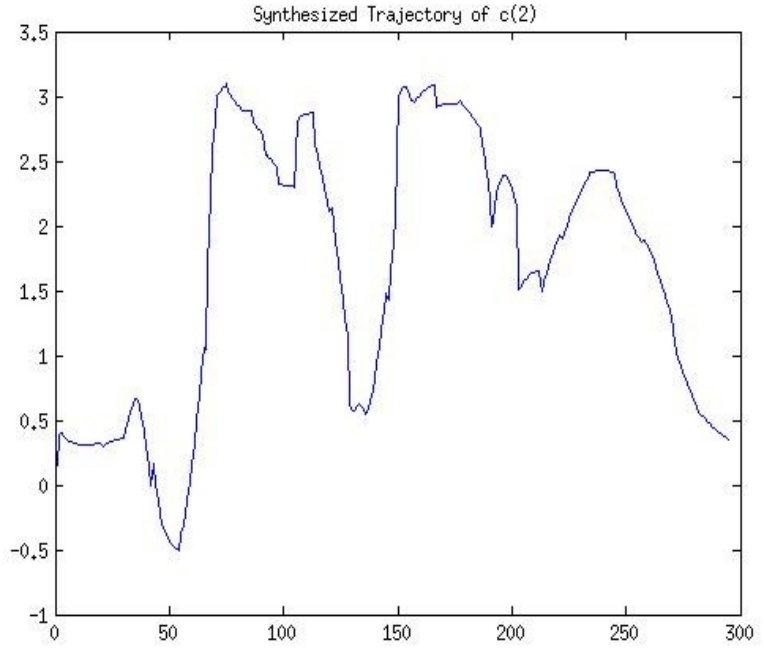
4.6 Σύνθεση Φωνής

Ο vocoder Straight για την παραγωγή συνθετικής φωνής χρησιμοποιεί το *F0* , *aperiodicity* και *spectrum* . Πρώτα μετατρέψαμε τα generalized mel cepstrums σε spectrums για την διαδικασία αυτή χρησιμοποιήσαμε το **SPTK 3.6** με παραμέτρους α =0,55 (συντελεστής αναδίπλωσης) και γ=0 (συντελεστής γενίκευσης). Στη συνέχεια εισάγοντας τα τρία αυτά στοιχεία στο Straight παράξαμε την φωνή. Για τη σύνθεση επιλέξαμε τυχαία 25 προτάσεις από τις 1131 του training set. Να επισημάνουμε ότι δεν θα μπορούσαμε να επιλέξουμε προτάσεις από το test set διότι συνθέσμαμε μόνο τα φάσματα (mcep) και όχι τις παραμέτρους F0 και aperiodicity καθώς αυτές τις εξάγαμε από το Straight. Σε περίπτωση που θέλαμε να χρησιμοποιήσουμε προτάσεις από το test set θα έπρεπε να συνθέσουμε όλα τα παραπάνω στοιχεία (spectrum, F0, aperiodicity) αλλά και τη διάρκεια των φωνημάτων.

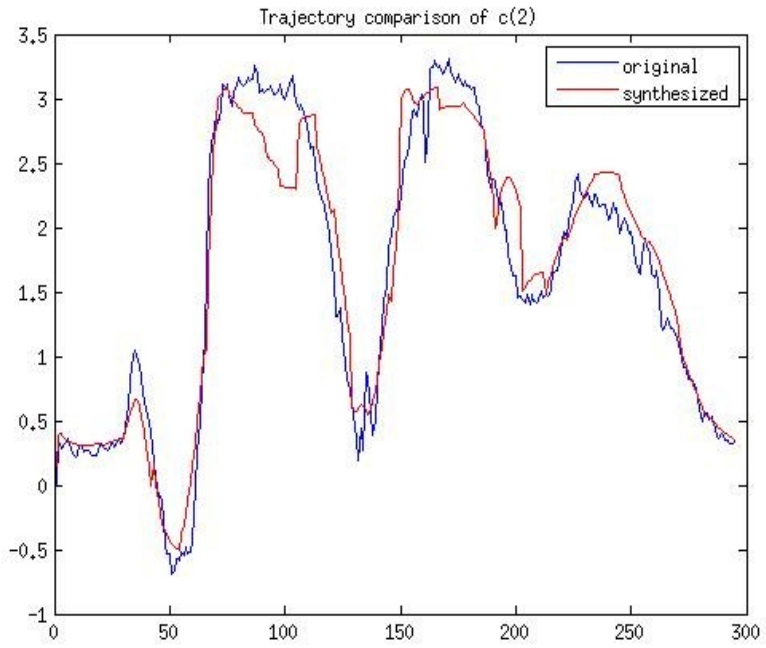
Στα γραφήματα 15α, 15β απεικονίζονται οι καμπύλες των αυθεντικών και συνθετικών mel cepstrum coefficients (επιλέξαμε τυχαία 1 από τα 50 που προκύπτουν, στην πρώτη περίπτωση είναι το c(2) από το cmu_us_arctic_slt_a0098) αντίστοιχα και στο γράφημα 15γ είναι η σύγκρισή τους. Ομοίως για τα γραφήματα 16α, 16β, 16γ και 17α, 17β, 17γ (επιλέχθηκε το c(4) από το cmu_us_arctic_slt_a0421 και το το c(3) από το cmu_us_arctic_slt_b0021). Στη συνέχεια υπολογίσαμε την ευκλείδεια απόσταση μεταξύ αυθεντικών και συνθετικών καμπυλών και για το πρώτο παράδειγμα είναι 5,0047 ενώ για το δεύτερο 3,3976 πράγμα που αποδεικνύει την υψηλή ποιότητα εξαγωγής ακουστικών χαρακτηριστικών.



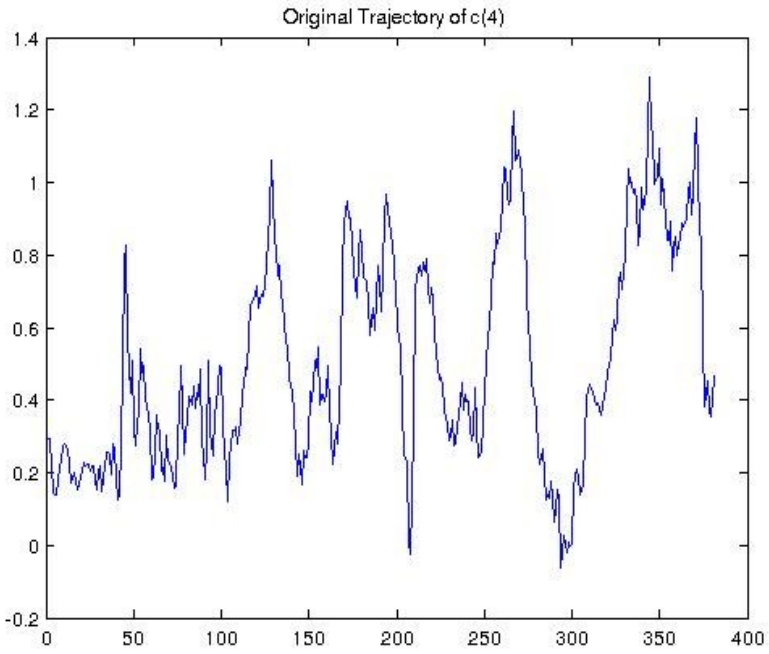
Γράφημα 15α: Original Trajectory of c(2) from cmu_us_arctic_slt_a0098



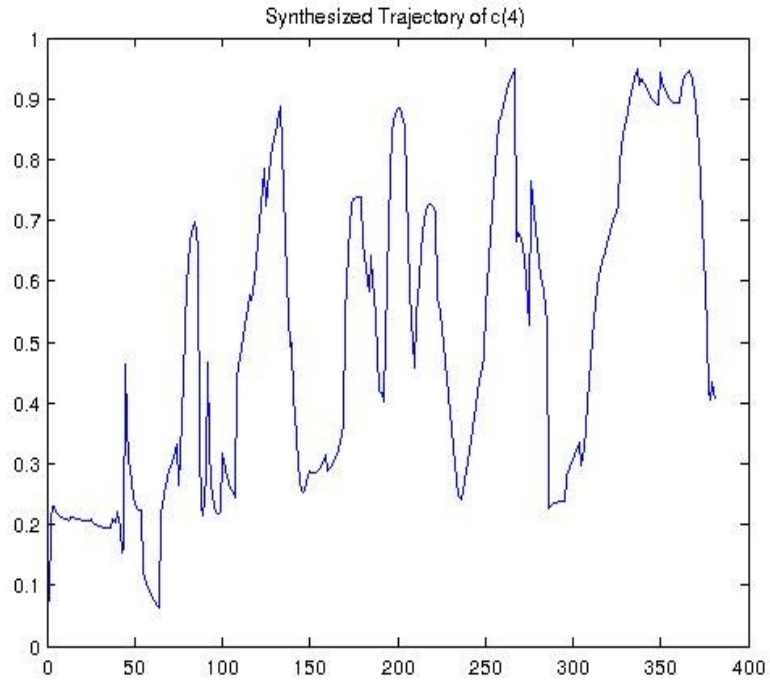
Γράφημα 15β: Synthesized Trajectory of c(2) from cmu_us_arctic_slt_a0098



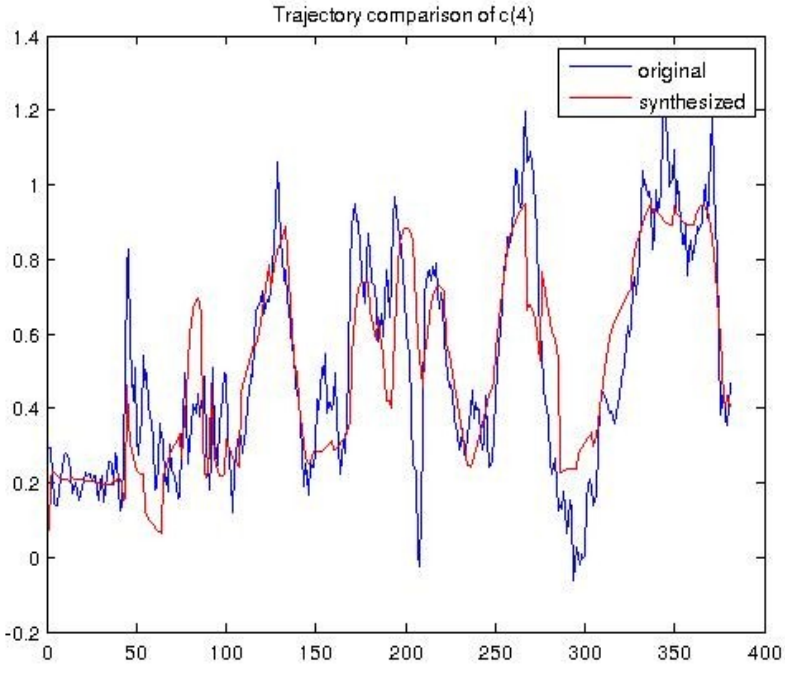
Γράφημα 15γ: Trajectory comparison of c(2) from cmu_us_arctic_slt_a0098



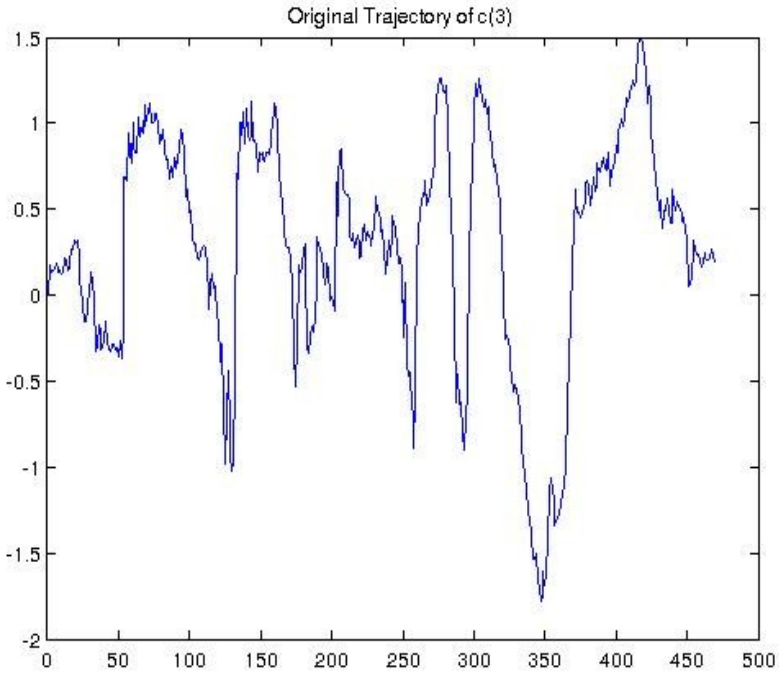
Γράφημα 16α: Original Trajectory of c(4) from cmu_us_arctic_slt_a0421



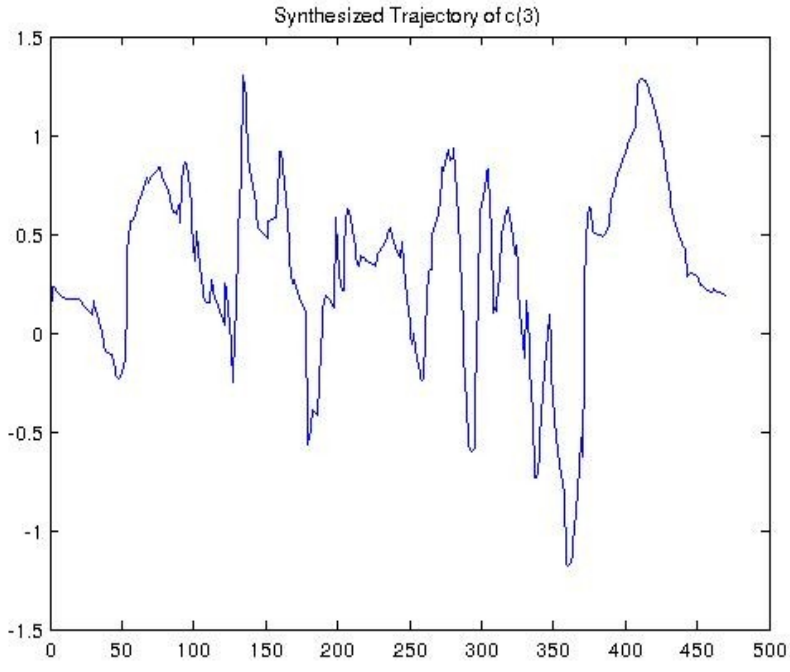
Γράφημα 16β: Synthesized Trajectory of c(4) from cmu_us_arctic_slt_a0421



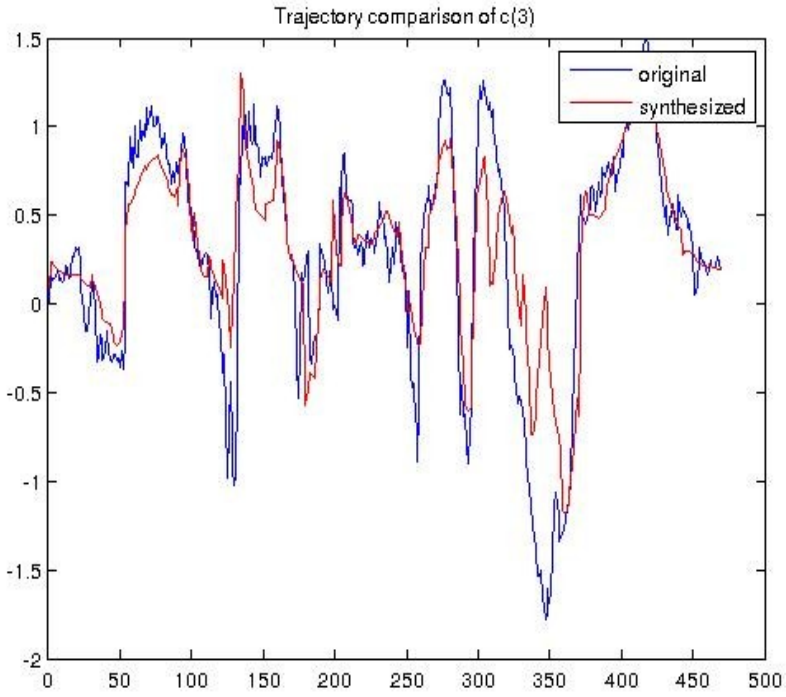
Γράφημα 16γ: Trajectory comparison of c(4) from cmu_us_arctic_slt_a0421



Γράφημα 17α: Original Trajectory of c(3) from cmu_us_arctic_slt_b0021



Γράφημα 17β: Synthesized Trajectory of c(3) from cmu_us_arctic_slt_b0021



Γράφημα 17γ: Trajectory comparison of c(3) from cmu_us_arctic_slt_b0021

4.7 Αξιολόγηση Συνθετικής Φωνής

Με τη χρήση του λογισμικού **Pesq** αξιολογήσαμε την ποιότητα τον συνθετικών προτάσεων. Τα αποτελέσματα για 25 συνθετικές φωνές απεικονίζονται στον Πίνακα 4. Η βαθμολογία κυμαίνεται από 1 έως 5 με 1 να είναι το bad και 5 το excellent. Η καλύτερη συνθετική φωνή προήλθε από το *cmu_us_arctic_slt_a0421* και η χειρότερη από το *cmu_us_arctic_slt_b0188.* Παρατηρούμε ότι καμία συνθετική πρόταση δεν βαθμολογήθηκε με *poor* (2) ή *bad* (1) αντιθέτως οι 9 από τις 25 θεωρήθηκαν καλής ποιότητας. Επίσης καμία πρόταση δεν βαθμολογήθηκε με excellent (5). Επομένως μπορούμε να θεωρήσουμε αυτό το μοντέλο αρκετά αποδοτικό ως προς την ποιότητα της σύνθεσης. Ακολουθεί ο πίνακας 4.

Wav file	**Overall Quality**	**Background Distortion**	**Signal Distortion**	**Score**
cmu_us_arctic_slt_a0098	3.6156	1.8160	2.7874	Good
cmu_us_arctic_slt_a0110	3.2531	1.8221	2.5316	Fair
cmu_us_arctic_slt_a0133	3.5417	1.6849	2.6259	Good
cmu_us_arctic_slt_a0139	3.4056	1.7216	2.5666	Fair
cmu_us_arctic_slt_a0154	3.4719	1.7232	2.6023	Fair
cmu_us_arctic_slt_a0268	3.4303	1.6919	2.5618	Fair
cmu_us_arctic_slt_a0295	3.3100	**1.5973**	2.4219	Fair
cmu_us_arctic_slt_a0296	3.6313	1.8715	2.8438	Good
cmu_us_arctic_slt_a0351	3.5877	1.8128	2.7356	Good
cmu_us_arctic_slt_a0413	3.3296	1.6285	2.4572	Fair
cmu_us_arctic_slt_a0420	3.4402	1.7211	2.5752	Fair
cmu_us_arctic_slt_a0421	**3.7564**	1.8476	**2.9153**	Good
cmu_us_arctic_slt_a0455	3.4192	1.7416	2.5337	Fair
cmu_us_arctic_slt_a0471	3.4486	1.7559	2.6329	Fair
cmu_us_arctic_slt_a0501	3.5081	1.7170	2.6213	Good
cmu_us_arctic_slt_b0021	3.3541	1.8685	2.6160	Fair
cmu_us_arctic_slt_b0165	3.3739	1.9764	2.6777	Fair
cmu_us_arctic_slt_b0188	**3.2501**	1.6301	**2.2783**	Fair
cmu_us_arctic_slt_b0286	3.4995	1.9325	2.6636	Fair
cmu_us_arctic_slt_b0294	3.6726	**2.0199**	2.8574	Good
cmu_us_arctic_slt_b0297	3.4067	1.9715	2.6225	Fair
cmu_us_arctic_slt_b0339	3.4787	2.0181	2.7528	Fair
cmu_us_arctic_slt_b0499	3.2881	1.8847	2.5057	Fair
cmu_us_arctic_slt_b0516	3.6840	2.0676	2.8689	Good
cmu_us_arctic_slt_b0521	3.5803	1.9930	2.7984	Good
Best	3.7564	2.0199	2.9153	Good
Worst	3.2501	1.5973	2.2783	Fair
Average	3.4695	1.8206	2.6421	Fair

Πίνακας 4: Βαθμολογίες από λογισμικό Pesq

Κεφάλαιο 5

Συμπεράσματα

Στόχος της εργασίας αυτής είναι η κατασκευή και εκπαίδευση ενός στατιστικού ακουστικού μοντέλου το οποίο θα εφαρμοστεί στον τομέα της σύνθεσης φωνής. Το γραμμικό δυναμικό μοντέλο που δημιουργήσαμε αποδείχτηκε ένα αξιόπιστο μοντέλο. Οι βαθμολογίες που σημείωσε στο Pesq ήταν εντυπωσιακές. Να σημειώσουμε ότι παρόλο που η ποιότητα των συνθετικών φωνών που παράξαμε δεν ήταν βέλτιστη όλες οι φωνές που συνθέσαμε ήταν κατανοητές. Αυτή η σταθερότητα αποτελεί ισχυρό πλεονέκτημα. Οι δυνατότητες αυτού του μοντέλου είναι τεράστιες και το καθιστούν ένα από τα πιο αποδοτικά μοντέλα καθώς τα πλεονεκτήματα σε σχέση με τα μειονεκτήματα υπερτερούν σε σημαντικό βαθμό. Αρχικά η ποιότητα της σύνθεσης είναι υψηλή και με περιθώρια βελτίωσης. Η ταχύτητά του είναι εντυπωσιακή καθώς για τη σύνθεση μίας πρότασης διάρκειας 3 δευτερολέπτων χρειάζεται μόλις λίγα δευτερόλεπτα (μιλάμε καθαρά για το κομμάτι της σύνθεσης και όχι της εκπαίδευσης). Επίσης επειδή ανήκει στην κατηγορία των στατιστικών παραμετρικών μοντέλων του δίνεται η δυνατότητα να τροποποιηθεί ως προς τη γλώσσα, τους ομιλητές και την προσωδία επομένως απευθύνεται σε ευρύ κοινό. Παράλληλα για τους παραπάνω λόγους θα μπορούσε εύκολα να χρησιμοποιηθεί σε εφαρμογές και ειδικά σε πραγματικού χρόνου. Να σημειώσουμε ότι δεν απαιτείται μεγάλη βάση δεδομένων για την εκπαίδευσή του ωστόσο μία μεγαλύτερη βάση θα βοηθούσε στα αποτελέσματα της σύνθεσης καθώς η αντιστοίχηση των γλωσσικών με τα ακουστικά χαρακτηριστικά (linguistic-acoustic mapping) θα ήταν καλύτερη. Ορισμένα από τα προβλήματα που παρατηρήσαμε είναι η διάρκεια της εκπαίδευσης η οποία διαρκεί αρκετές ώρες. Επίσης το debug είναι αρκετά δύσκολο αλλά και χρονοβόρο καθώς συνήθως απαιτείται επανεκπαίδευση του μοντέλου. Η εξάρτησή του από το vocoder είναι άμεση επομένως η εύρεση ενός πιο αποδοτικού vocoder παραμένει μία πρόκληση που θα ευνοήσει σημαντικά την ποιότητα σύνθεσης του γραμμικού δυναμικού μοντέλου όπως και η δημιουργία καλύτερων και λιγότερο πολύπλοκων δέντρων απόφασης για την ομαδοποίηση των context-dependent φωνημάτων. Θα επισημάνουμε ότι ήδη υπάρχουν διάφοροι μέθοδοι για την βελτίωσή του όπως μέθοδοι για την απομάκρυνση του oversmoothing. Τα αποτελέσματα έπειτα από εφαρμογή των παραπάνω σίγουρα θα έχουν εντυπωσιακά αποτελέσματα στη σύνθεση φωνής. Ολοκληρώνοντας θα πούμε ότι το γραμμικό δυναμικό μοντέλο είναι σαφώς ένα μοντέλο που αξίζει να μελετηθεί αλλά και να εφαρμοστεί στον τομέα της σύνθεσης φωνής.

Βιβλιογραφία

1. Alan W. Black, Heiga Zen, Keiichi Takuda “Statistical parametric speech synthesis”, Speech Communication, 51(11), pp 1039-1064, November 2009.

2. Alexis Roche “EM algorithm and variants: an informal tutorial” , September 7, 2012.

3. Annamaria Mesaros, Toni Heittola, Antti Eronen and Tuomas Virtanen “Acoustic event detection in real life recordings”, Aalborg, Denmark, August 23-27, 2010.

4. Antony W. Rix, John G. Beerens, Michael P. Hollier and Andries P. Hekstra “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs”, Acoustics, Speech, and Signal Processing, 2001.

5. A. P. Dempster, N. M. Laird and D. B. Rubin “Maximum likelihood from incomplete data via the EM algorithm” , Vol. 39, No. 1. (1977), pp 1-38.

6. Carl Quillen “Kalman filter based speech synthesis” , Acoustics, Speech, and Signal Processing, 2010.

7. Heiga Zen “Statistical parametric speech synthesis synthesis: From HMM to LSTM-RNN” , July 29, 2015.

8. Hideki Kawahara, Ikuyo-Katsuse and Alain de Cheveigne “Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds ” , September 22, 1998.

9. Hideki Kawahara, Haruhiro Katayose, Alain de Cheveigne and Roy D. Patterson “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity” , September 5, 1999.

10. Hideki Kawahara, Yoshinori Atake and Parham Zolfaghari “Accurate vocal event detection

method based on a fixed-point analysis of mapping from time to weighted average group delay", ICSLP-2000, Beijing, pp.664-667 2000.

11. Hideki Kawahara and Parham Zolfaghari "Systematic F0 glitches around nasal-vowel transition".

12. Hideki Kawahara, Jo Estill and Osamu Fujimura "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system Straight" , September 13-15, 2001.

13. Hideki Kawahara, Alain de Cheveigne, Hideki Banno, Toru Takahasi and Toshio Irino "Nearly detect-free F0 trajectory extracion for expressive speech modification based on Straight" , Proc. Interspeech2005, Lisboa, pp.537-540, September 2005.

14. Hideki Kawahara "Getting started with Straight in command mode" , September 22, 2005.

15. Hideki Kawahara "Fixed-point representations for very high quality speech and sound modification system" , August 9, 2002.

16. Hideki Kawahara "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revised", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP '97), vol.2, pp.1303-1306 (1997.4).

17. Jithendra Vepa and Simon King "Kalman-filter based on cost for unit-selection speech synthesis", Eurospeech 2003.

18. Keiichi Tokuda, Heiga Zen "Fundamentals and recent advances in HMM-based speech synthesis" , October 15, 2011.

19. Keiichi Tokuda, Heiga Zen and Alan W. Black "An HMM-based speech synthesis system applied to english", 2002.

20. Keiichi Tokuda, Heiga Zen, Alan W. Black, Shinji Sako, Takashi Nose and Takashi Masuko

"The HMM-based speech synthesis (HTS) version 2.0", 2008.

21. Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki and Takao Kobayashi "Multi-space probability distribution HMM" , March 3, 2002.

22. Keiichi Tokuda, Heiga Zen and Tadashi Kitamura "Reformulating the HMM as a trajectory model" , Computer Speech and Language, Vol. 21, pp. 153-173, 2007.

23. Lawrence R. Rabiner, "Readings in speech recognition," chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pp. 267–296. Morgan Kaufmann Publishers Inc., 1990.

24. Paul Taylor "Text-to-Speech synthesis", Cambridge University Press, Cambridge, 2009.

25. Pavlos Papadopoulos and Vasilios Digalakis "Identification of systems in canonical form with EM algorithm", April 2010.

26. Thierry Dutoit "A short introduction to text-to-speech synthesis" , http://tcts.fpms.ac.be/synthesis/introtts_old.html.

27. Thomas P. Minka "From Hidden Markov models to linear dynamical systems" , 18 July, 1999.

28. Vassilis Tsiaras "Linear Dynamic Models in Speech Synthesis" , October, 2015.