ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Σύνθεση Φωνής από Κείμενο με Βάση Κρυφά Μαρκοβιανά Μοντέλα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Χρήστου Α. Μιναρετζή

Επιβλέπων: Πέτρος Α. Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2011

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Σύνθεση Φωνής από Κείμενο με Βάση Κρυφά Μαρκοβιανά Μοντέλα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Χρήστου Α. Μιναρετζή

Επιβλέπων: Πέτρος Α. Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 22 Ιουλίου 2011.

.....	.....	.....
Πέτρος Μαραγκός Καθηγητής Ε.Μ.Π.	Κώστας Τζαφέστας Επίκουρος Καθηγητής Ε.Μ.Π.	Γεράσιμος Ποταμιάνος Διευθυντής Ερευνών, Δημόκριτος

Αθήνα, Ιούλιος 2011

Περίληψη

Η συγκεκριμένη διπλωματική πραγματεύεται το θέμα της Σύνθεσης Φωνής από Κείμενο με στατιστική μοντελοποίηση με Κρυφών Μαρκοβιανών Μοντέλων. Το συγκεκριμένο θέμα εντάσσεται στη συνδυαστική ερευνητική περιοχή της επεξεργασίας φωνής και της αναγνώρισης προτύπων. Σταδιακά στα πρώτα Κεφάλαια αναλύεται όλο το απαιτούμενο θεωρητικό υπόβαθρο που αφορά το συγκεκριμένο πρόβλημα. Ιδιαίτερη έμφαση δίνεται στη μελέτη και υλοποίηση διαφορετικών Vocoders για την επιλογή των κατάλληλων χαρακτηριστικών. Στο τελευταίο Κεφάλαιο περιγράφονται αναλυτικά τα στάδια που ακολουθήθηκαν για την υλοποίηση του συνθέτη φωνής από κείμενο περιλαμβάνοντας και τα πειραματικά αποτελέσματα, τα οποία και είναι ιδιαίτερα ενθαρρυντικά.

Abstract

The goal of this Thesis is the study and the implementation of an HMM Text to Speech Synthesizer. This applications is part of the combinational research area of Speech Processing and Pattern Recognition. In the first part the thesis focuses on the theoretical framework of dealing with the problem of HMM Text to Speech Synthesis. Additionally, there is a special research on implementing different Vocoders, in order to figure out which are the best characteristics. Finally, this thesis describes the main steps of the HMM TTS system implementation, including very encouraging experimental results.

Ευχαριστίες

Αυτή η Διπλωματική είναι αποτέλεσμα των ερευνητικών ερεθισμάτων που προέκυψαν μέσα από τη διδασκαλία των μαθημάτων της Αναγνώρισης Προτύπων και της Ψηφιακής Επεξεργασίας Σήματος του επιβλέποντα καθηγητή μου κυριου Πέτρου Μαραγκού. Τον ευχαριστώ, λοιπόν, ιδιαίτερα για την πολύτιμή βοήθειά του.

Παράλληλα θα ήθελα να ευχαριστήσω το Σταύρο Θεοδωράκη, που φάνηκε σημαντικός αρωγός της συνολικής μου προσπάθειας και ιδιαίτερα στις πιο δύσκολες στιγμές. Ακόμη, ευχαριστώ και όλα τα υπόλοιπα μέλη του εργαστηρίου, που κατά καιρούς με βοήθησαν, καθώς και τους Πύρρο Τσιάκουλη και Σωτήρη Καραμπέτσο από το Ι.Ε.Λ., για τις πολύτιμες συμβουλές τους.

Επιπρόσθετα, θα ήθελα να ευχαριστήσω τους συμφοιτητές μου, που ήταν σε όλη αυτήν την προσπάθεια κοντά μου και κυρίως το φίλο και συνεγάτη μου Δημήτρη Θεοδωράκη, με τον οποίο πραγματοποιήσαμε ένα μεγάλο μέρος της αρχικής έρευνας της διπλωματικής, η οποία δημοσιεύτηκε στα πλαίσια του Συνεδρίου Φοιτητών Ηλεκτρολόγων, ΣΦΗΜΜΥ 2009.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και κυρίως τη μητέρα μου, η οποία μου εμφύσησε την αγάπη και το σεβασμό στη γνώση.

Περιεχόμενα

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

Στη σημερινή κοινωνία της πληροφόρησης, η επιστήμη εισάγει νέες και καινοτόμες εφαρμογές με στόχο τη βελτίωση της ποιότητας ζωής του σύγχρονου ανθρωπου. Ειδικότερα, έχει δημιουργηθεί ένας ολόκληρος επιστημονικός κλάδος, ο οποίος στοχεύει στην ταχύτερη διαχείριση των πληροφοριών μέσα από τα διαθέσιμα υπολογιστικά συστήματα. Κάτι τέτοιο θα τελειοποιηθεί τη στιγμή που άνθρωπος και υπολογιστής θα "μιλάνε την ίδια γλώσσα". Σε αυτήν την κατεύθυνση αναπτύσσονται τόσο συστήματα αναγνώρισης φωνής όσο και συστήματα σύνθεσης φωνής από κείμενο.

Αυτά τα συστήματα πέρα από τη διευκόλυνση της επικοινωνίας του ανθρώπου με τον υπολογιστή, μπορούν να εξυπηρετήσουν και άτομα με ειδικές ανάγκες ακοής όσο και όρασης. Ακόμη, μπορούν να διευκολύνουν τη διαχείριση μεγάλου όγκου ανθρώπων που χρησιμοποιούν τηλεφωνικές υπηρεσίες. Τέλος, η συνεχής εξέλιξη των πολυμέσων και των υπολογιστικών συσκευών που χρησιμοποιούνται για ψυχαγωγία, εισάγει ένα τεράστιο εύρος εφαρμογών.

1.1 Ιστορική Αναδρομή

Την τελευταία πεντηκονταετία πραγματοποιείται έντονη επιστημονική έρευνα όσον αφορά την ανάπτυξη και τη συνεχή βελτίωση συστημάτων σύνθεσης φωνής από κείμενο [36]. Μέχρι τη δεκαετία του 1980 τα συστήματα που αναπτύσσονταν βασίζονταν σε **Πρώτης Γενιάς** Τεχνικές. Αυτές οι τεχνικές είχαν βάση τη ρύθμιση του συνθέτη φωνής με βάση κάποιους γενικούς κανόνες μοντελοποίησης αυτής. Τέτοιες συσκευές ήταν οι Formant Synthesizers του Klaat και του Holmes [3, 19, 12]. Παράλληλα, από τα πρώτα συστήματα ήταν ο Articulatory Synthesizer του von Kempelen, ο οποίος και βασιζόταν στη φυσική μοντελοποίηση της παραγωγής της φωνής στο φωνητικό σωλήνα

μέσα από τη ροή του αέρα από ειδικά διαμορφωμένους σωλήνες. Τέλος οι LP Synthesizers βασίζονταν στη θεωρία γραμμικής πρόβλεψης και αυτοί αποτελούν το μεταβατικό στάδιο για το πέρασμα στη **Δεύτερη Γενιά** συστημάτων.

Στη Δεύτερη γενιά δόθηκε έμφαση στην πραγματική φωνή και σε ηχογραφημένα δεδομένα και όχι σε βασικούς νόμους παραγωγής τη φωνής. Συγκεκριμένα, από μία βάση ηχογραφημένων και απομονωμένων στοιχείων φωνής γινόταν η σύνθεση της τελικής πρότασης, η οποία προσαρμόζονταν στα απαιτούμενα δεδομένα προσωδίας και διάρκειας με μεθόδους μετατροπής του pitch και εξομάλυνσης των ασυνεχειών. Οι πιο διάσημες υλοποιήσεις αυτών των συνθετών φωνής, χρησιμοποιούν δίφωνα, δηλαδή ζεύγη φωνημάτων, ενώ οι πιο γνωστοί αλγόριθμοι που χρησιμοποιούν για περαιτέρω φωνητική επεξεργασία είναι οι PSOLA, MBROLA, Sinusoidal Analysis [25, 7, 8].

Στην **Τρίτη Γενιά** συστημάτων ανήκουν τα λεγόμενα Unit-Selection συστήματα. Αυτά λειτουργούν με τη χρήση μεγάλου όγκου ηχογραφημένων βάσεων εκφωνήσεων, οι οποίες διαθέτουν πολλαπλές εκδοχές από απομονωμένα στοιχεία φωνητικών μονάδων (π.χ. φωνήματα, δίφωνα κτλ.). Έτσι, κατά τη διδικασία δημιουργίας της συνθετικής φωνής επιλέγεται η βέλτιστη ακολουθία επιλογών φωνημάτων, που καθορίζεται από κριτήρια απόστασης χαρακτηριστικών [45]. Το πρώτο Unit Selection σύστημα εμφανίστηκε με την εργασια από τους Hunt και Black το 1996 [13].

Τα τελευταία δέκα χρόνια εμφανίστηκε ένας νέος τρόπος υλοποίησης αυτόματων συστημάτων σύνθεσης φωνής από κείμενο, που βασιζεται σε **Κυρφά Μαρκοβιανά Μοντέλα**. Συγκεκριμένα, ένας συνθέτης φωνής που βασίζεται σε στατιστική μοντελοποίηση της φωνής δεν έχει κάποια φωνητική βάση σαν πληροφορία, αλλά μία στατιστική βάση μοντελοποιημένων παραμέτρων. Αυτές οι παραμετροι αποτελούν χαρακτηριστικά της φωνής [42]. Η υλοποίηση του Συνθέτη φωνής από κείμενο παρουσιάζει πολλά πλεονεκτήματα σε σχέση με τις άλλες τεχνικές, όπως την πολύ εύκολη προσαρμογή σε νέους ομιλητές [46, 18, 35], την αρκετά περιορισμένη βάση δεδομένων, λόγω της μοντελοποίησης καθώς και της σταθερής απόδοσης ποιότητας ήχου και την πολύ εύκολη προσαρμογή σε άλλες γλώσσες, αφού ανάλογα συστήματα έχουν υλοποιηθεί στα Αγγλικά [42], στα Ιαπωνικά και στα Κινέζικα [52, 29], στα Κορεάτικα [17], στα Γερμανικά [44], στα Πορτογαλικά [22, 4], στα Σουηδικά [21], στα Φιλανδικά [28], στα Σλοβένικα [43], στα Κροάτικα [23], στα Αραβικά [2] και στα Ελληνικά [15]. Σε σύγκριση με συνθέτες υψηλής ποιότητας Unit Selection, οι HMMs TTS δεν μπορούν να τους ανταγωνιστούν σε φράσεις που ανήκουν στη φωνητική βάση των πρώτων, αλλά για όλες τις άλλες φράσεις η απόδοση τους είναι υψηλότερη και πιο σταθερή [5].

1.2 Σκοπός της Διπλωματικής

Η συγκεκριμένη Διπλωματική Εργασία αφορά τη μελέτη και την ανάπτυξη ενός ολοκληρωμένου συστήματος Σύνθεσης φωνής από κείμενο που βασίζεται στη στατιστική μοντελοποίηση με Κρυφά Μαρκοβιανά Μοντέλα. Με βαση αυτή την υλοποίηση σκοπός είναι να μελετηθούν και να βελτιστοποιηθούν οι παράμετροι που καθορίζουν την ποιότητα του αποτελέσματος. Μιας και στη Συνθετική φωνή δεν υπάρχουν αντικειμενικά μετρικά της ποιότητας του ήχου και δεδομένου ότι στην προκειμένη περίπτωση δεν υπάρχει ούτε ήχος αναφοράς, η αξιολόγηση των αποτελεσμάτων γίνεται με το ανθρώπινο υποκειμενικό κριτήριο.

1.3 Συνεισφορές της Διπλωματικής

Αυτή η διπλωματική ερασία, αποτυπώνει την έρευνα που έγινε για την βέλτιστη υλοποίηση ενός Συνθέτη φωνής από κείμενο στην ελληνική γλώσσα. Συγκεκριμένα, οι επιστημονικές συνεισφορές της συνοψίζονται στα ακόλουθα βασικά σημεία:

1. **Μελέτη καταλληλότητας των εξαγόμενων χαρακτηριστικών για την υλοποίηση του Συνθέτη:** Συγκεκριμένα, υλοποιήθηκαν 6 διαφορετικοί Vocoders, καθένας από τους οποίους βασίζεται σε διαφορετική τεχνική ανάλυσης των φωνητικών σημάτων. Συνεπώς, εξήχθησαν 6 διαφορετικά ηχητικά χαρακτηριστικά, τα οποία μελετήθηκαν και αξιολογήθηκαν ως προς την ακουστική ποιότητα της συνθετικής φωνής, όσο και ως προς την καταλληλότητά τους για τη στατιστική μοντελοποίηση με Κρυφά Μαρκοβιανά Μοντέλα.

2. **Εφαρμογή ενός Ολοκληρωμένου Συνθέτη Φωνής από Κείμενο στα Ελληνικά:** Με τη χρήση συγκεκριμένων υπολογιστικών εργαλείων επεξεργασίας φωνητικών σημάτων [1] και μοντελοποίησης Κρυφών Μαρκοβιανών Μοντέλων [50, 51], προγραμματίστηκε ένα άρτιο σύστημα παραγωγής συνθετικής φωνής από κείμενο στην ελληνική γλώσσα.

3. **Βελτίωση των ηχητικών αποτελεσμάτων της Σύνθεσης:** Κατόπιν της παραγωγής της συνθετικής φωνής από το Συνθέτη, ακολούθησε επεξεργασία των ήχων για τη βελτίωσή τους. Συγκεκριμένα, εφαρμόστηκαν τεχνικές ομαλοποίησής τους, ώστε να απομακρυνθούν όλα τα συνθετικά στοιχεία, που καθιστούσαν τα ηχητικά αποτελέσματα μη φυσικά.

1.4 Οργάνωση της Διπλωματικής

Η οργάνωση της συγκεκριμένης διπλωματικής ακολουθεί κατά βάση τα τρία στάδια της ερευνητικής προσπάθειας, όπως περιγράφηκαν πιο πάνω. Συνεπώς στο **Κεφάλαιο 2** παρουσιάζεται όλη η διαδικασια της υλοποίησης των 6 διαφορετικών Vocoders, καθώς και η αξιολόγηση των συγκεκριμένων χαρακτηριστικών. Ακολούθως, στο **Κεφάλαιο 3**, αναλύεται θεωρητικά όλη η διαδικασία εκπαίδευσης και ρύθμισης των παραμέτρων των μοντέλων, που απαιτείται για την υλοποίηση του Συνθέτη φωνής. Στο **Κεφάλαιο 4** ακολουθεί η θεωρητική ανάλυση της διαδικασίας παραγωγής των χαρακτηριστικών της συνθετικής φωνής, από τις Κρυφές Μαρκοβιανές Αλυσίδες που προκύπτουν από την ενοποίηση των μοντέλων των φωνητικών μονάδων, κατόπιν ανάλυσης του κειμένου που εισάγεται στο σύστημα. Η υλοποίηση ενός ολοκληρωμένου συστήματος σύνθεσης φωνής από κείμενο βασισμένο σε Κρυφά Μαρκοβιανά Μοντέλα, αναλύεται σε θεωρητικό επίπεδο στο **Κεφάλαιο 5**. Τέλος, το **Κεφάλαιο 6** περιγράφει αναλυτικά όλες τις πειραματικές διαδικασίες που ακολουθήθηκαν για την υλοποίηση του συνθέτη, καθώς και για τη μετεπεξεργασία των παραγόμενων ήχων. Στο **Κεφάλαιο 7** πραγματοποιείται μία λεπτομερής σύνοψη όλης της ερευνητικής προσπάθειας, καθώς και μία αναλυτική περιγραφή των ουσιαστικών συνεισφορών και των ερευνητικών προεκτάσεων της συγκεκριμένης διπλωματικής.

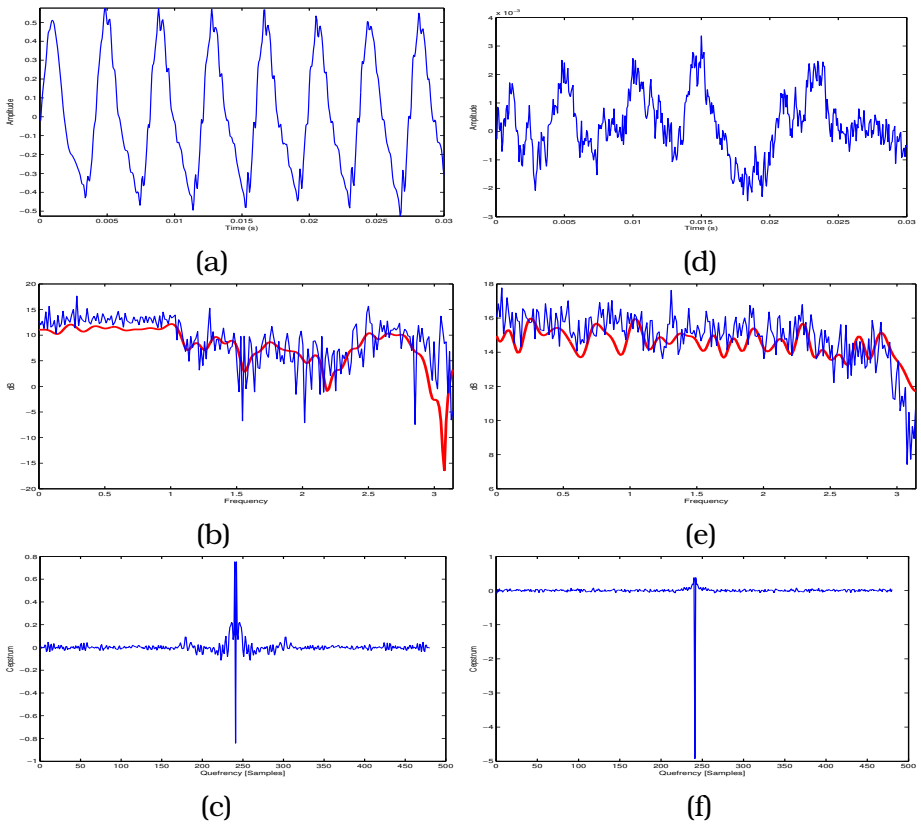
Κεφάλαιο 2

Εξαγωγή Χαρακτηριστικών

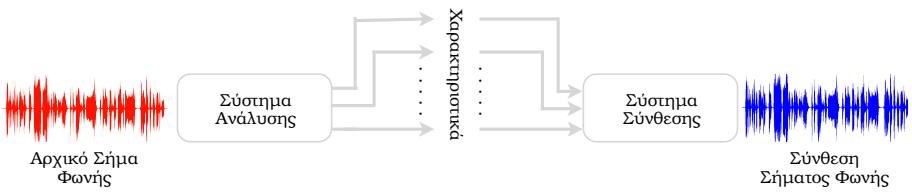
2.1 Εξαγωγή του Pitch από ένα Σήμα Φωνής

Κάθε έμφωνο τμήμα ενός ήχου έχει μία βασική συχνότητα pitch και κάποιες διακριτές δευτερεύουσες συχνότητες τα formants. Κατά την παραγωγή της φωνής η βασική συχνότητα pitch παράγεται από την ταλάντωση των φωνητικών χορδών του ανθρώπου και εν συνεχεία ο φωνητικός σωλήνας διαμορφώνει αυτό το αρχικό σήμα εισάγωντας τις συχνότητες των formants. Αυτός ο διαχωρισμός είναι εμφανής κατά την φασματική απεικόνιση ενός τμήματος έμφωνου ήχου . Συγκεκριμένα, όπως φαίνεται και στο Σχήμα 2.1 η βασική συχνότητα pitch αποτελεί τη συχνότητα ταλάντωσης του φάσματος ενώ οι συχνότητες των formants διαμορφώνουν τη φασματική περιβάλλουσα του σήματος. Σε έναν έμφωνο ήχο ο ρυθμός μεταβολής της ακολουθίας του pitch δίνει την προσωδία του ήχου, δηλαδή το ηχόχρωμα μιας ομιλίας. Χάριν στην προσωδία μία πρόταση μπορεί να ειπωθεί ως ερώτηση, ως κατάφαση ή ως δήλωση κάποιου συναισθήματος. Στη σύνθεση φωνής η πληροφορία του pitch είναι αναγκαία.

Υπάρχουν πολλοί τρόποι εξαγωγής του pitch από ένα φωνητικό σήμα. Ο πιο δημοφιλής τρόπος βασίζεται στην ομομορφική ανάλυση ενός σήματος φωνής και στην αποσυνέλιξη (deconvolution) της διέγερσης από το φάσμα του φωνητικού σωλήνα [33, 34]. Μελετώντας τα Σχήματα 2.1(c) και (f), φαίνεται ότι η ακολουθία των cepstrum ενός έμφωνου και ενός άφωνου τμήματος φωνής διαφέρουν ως προς την εμφάνιση κάποιων παράπλευρων κορυφών εκτός από την κεντρική. Συγκεκριμένα η συχνότητα εμφάνισης αυτών των κορυφών στο πεδίο των cepstrum ενός έμφωνου ήχου καθορίζει και τη βασική συχνότητα pitch. Αντίστοιχα, η απουσία των παράπλευρων κορυφών δηλώνει την ύπαρξη άφωνου τμήματος ήχου.



Σχήμα 2.1: Απεικόνιση (a) έμφωνου τμήματος ήχου μετά από παραθύρωση με Hamming, (b) της ομαλής περιβάλουσας του ήχου μετά από liftering και (c) της ακολουθίας των cepstrum. Απεικόνιση (d) άφωνου τμήματος ήχου μετά από παραθύρωση με Hamming, (e) της ομαλής περιβάλουσας του ήχου μετά από liftering και (f) της ακολουθίας των cepstrum.



Σχήμα 2.2: Βασική απεικόνιση ενός Vocoder

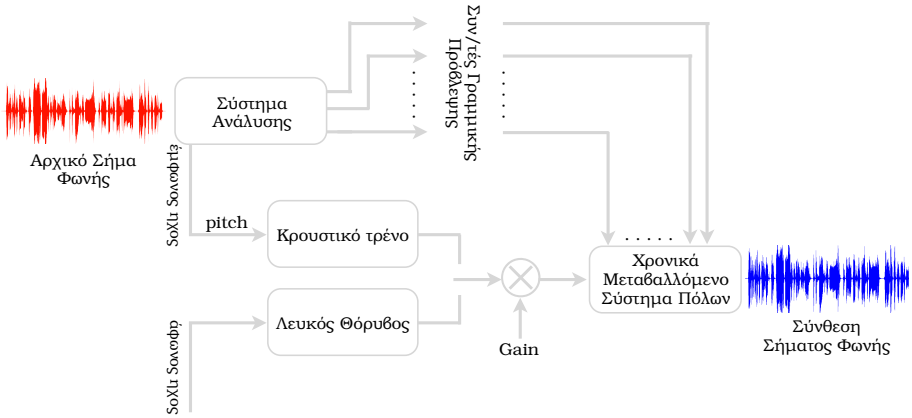
2.2 Μελέτη Διαφορετικών Vocoders

Vocoder είναι ένα σύστημα Ανάλυσης και Σύνθεσης ήχου. Συγκεκριμένα, ένας Vocoder στο τμήμα της Ανάλυσης εξάγει από τον ήχο που παίρνει ως είσοδο κάποια χαρακτηριστικά, τα οποία εισάγονται στο δεύτερο τμήμα του συνθέτοντας ένα ήχο που προσεγγίζει τον αρχικό. Οι Vocoders χρησιμοποιούνται κυρίως στην Ανάλυση και Σύνθεση ανθρώπινης φωνής και επιτυγχάνουν μεγάλα ποσοστά συμπίεσης δεδομένων.

Ως πρώτο στάδιο υλοποίησης ενός HMM Text To Speech συστήματος, απαιτείται η χρήση ενός Vocoder. Συγκεκριμένα, τα χαρακτηριστικά που εξάγονται από το τμήμα της Ανάλυσής του, θα χρησιμοποιηθούν για την εκπαίδευση των Κρυφών Μαρκοβιανών Μοντέλων. Παράλληλα, το τμήμα της Σύνθεσης φωνής του Vocoder θα συνθέσει τα εκτιμώμενα χαρακτηριστικά σε συνθετική φωνή. Έτσι στα πλαίσια της θεωρητικής μελέτης και για την επιλογή των πιο κατάλληλων χαρακτηριστικών, υλοποιήθηκαν διαφορετικοί vocoders. Τα αποτελέσματα που προέκυψαν κρίθηκαν ως προς την ποιότητά τους και την κατάλληλότητα της εφαρμογής τους για την υλοποίηση του συγκεκριμένου συστήματος.

2.2.1 LPC Vocoder

Ο πιο απλός τρόπος να προσεγγιστεί το πρόβλημα της μοντελοποίησης της παραγωγής φωνής από τον άνθρωπο είναι με τη μέθοδο της Γραμμικής Πρόβλεψης. Βάσει αυτής της θεωρίας υλοποιείται και ένας LPC Vocoder (Linear Prediction Coefficients). Στο τμήμα της Ανάλυσης ένας LPC Vocoder εξάγει τους συντελεστές γραμμικής πρόβλεψης και τη βασική συχνότητα της φωνής(pitch). Η εξαγωγή των χαρακτηριστικών γίνεται με τη μέθοδο της παραθύρωσης διαδοχικά επικαλυπτόμενων τμημάτων φωνής στο πεδίο του χρόνου (OverLapAdd, OLA Method). Σε κάθε χρονικό παράθυρο οι συντελεστές γραμμικής πρόβλεψης συνθέτοντας ένα σύστημα που έχει μόνο πόλους(all-pole), προσεγγίζουν τη συνάρτηση μεταφοράς του φωνητικού σωλήνα, ενώ η



Σχήμα 2.3: Βασική απεικόνιση ενός LPC Vocoder

βασική συχνότητα pitch εκτιμά τη συχνότητα που πάλλονται οι φωνητικές χορδές του ατόμου. Έτσι, στο στάδιο της σύνθεσης, ένας LPC-Vocoder σε κάθε χρονικό παράθυρο δημιουργεί ένα σύστημα πόλων βάσει των συντελεστών της γραμμικής πρόβλεψης και δέχεται ως είσοδο ένα σήμα κρουστικού τρένου με συχνότητα pitch στα τμήματα του έμφωνου ήχου ή εναλλακτικά λευκό θόρυβο στα άφωνα τμήματα. Το λειτουργικό διάγραμμα ενός LPC Vocoder απεικονίζεται στο Σχήμα 2.3[30].

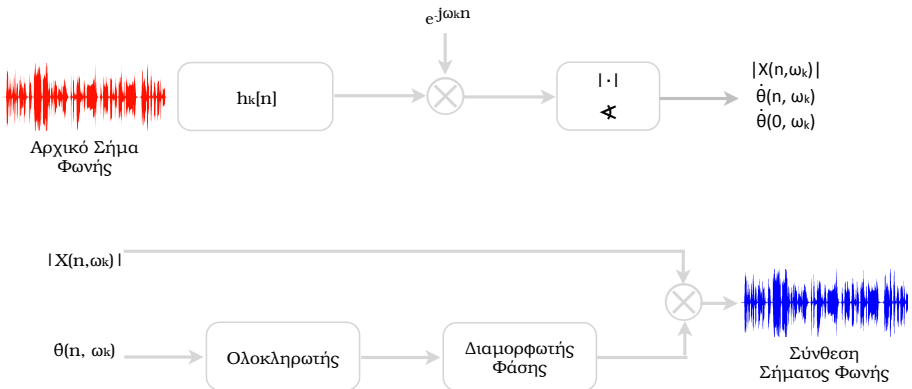
Συγκεριμένα, η συνάρτηση μεταφοράς $H(z)$ είναι σταθερή σε κάθε παραθυρωμένο τμήμα του λόγου. Οι πόλοι της συνάρτησης είναι οι ρίζες του πολυωνύμου με συντελεστές τους συντελεστές γραμμικής πρόβλεψης (Εξ. 2.1). Για τον υπολογισμό αυτών των συντελεστών υπάρχουν διάφοροι αλγόριθμοι όπως ο αναδρομικός αλγόριθμος του Levinson ή η μέθοδος της αυτοσυσχέτισης[30]. Όλοι στοχεύουν στην ελαχιστοποίηση του μέσου τετραγωνικού λάθους(Εξ. 2.3) ως προς τους γραμμικούς συντελεστές a_k.

$$H(z) = \frac{A}{\sum_{k=1}^{p} a_k z^{-k}} \tag{2.1}$$

$$E = \sum_{m=n-M}^{n+M} (s[m] - \tilde{s}[m])^2 = \sum_{m=n-M}^{n+M} e^2[m] \tag{2.2}$$

όπου,

$$\tilde{s}[m] = \sum_{k=1}^{p} a_k s[m-k] \tag{2.3}$$



Σχήμα 2.4: Βασική απεικόνιση ενός Phase Vocoder

2.2.2 Phase Vocoder

Ξεφεύγοντας από την παραθύρωση στο πεδίο του χρόνου, ένας Vocoder μπορεί να στηριχθεί και στην μέθοδο ανάλυσης σε φασματικές ζώνες(Filter Bank-FBS Method). Αυτή η μέθοδος ανάλυσης ενός σήματος φωνής είναι πολλές φορές προτιμητέα σε σχέση με τη μέθοδο της απλής παραθύρωσης στο χρόνο. Ένας Phase Vocoder, βασίζεται σε αυτή τη λογική. Συγκεκριμένα, στο στάδιο της Ανάλυσης το σήμα φωνής κατακερματίζεται από την τράπεζα φίλτρων στο πεδίο της συχνότητας και στη συνέχεια υπολογίζονται κατά προσέγγιση το μέτρο και ο ρυθμός μεταβολής της φάσης του παραθυρωμένου σήματος φωνής συναρτήσει του χρόνου. Ο ρυθμός μεταβολής της φάσης απεικονίζει στην ουσία τη στιγμιαία συχνότητα του παραθυρωμένου σήματος[30]. Το λειτουργικό διάγραμμα ενός phase Vocoder παρουσιάζεται στο Σχήμα 2.4.

Οι συναρτήσεις παραθύρωσης του αρχικού σήματος φωνής δίνονται από την Εξ. 2.4

$$h_k[n] = w[n]e^{j\frac{2\pi}{N}kn} \quad , 0 \leqslant n < N_w \tag{2.4}$$

$$y_k[n] = a_k[n]e^{j\phi_k[n]} \tag{2.5}$$

Η ανακατασκευή του αρχικού σήματος προκύπτει από την Εξ. 2.6

$$y[n] = \frac{1}{Nw[0]} \sum_{k=0}^{N-1} a_k[n]e^{j\phi_k[n]} \tag{2.6}$$

2.2.3 Ημιτονοειδής Ανάλυση/Σύνθεση Φωνής

Ο Vocoder αυτός, που προέκυψε από την έρευνα των McAulay και Quatieri [25, 30], βασίζεται στην αναπαράσταση με ημιτονοειδή σε πολύ βασικές συχνότητες του σήματος φωνής. Συγκεκριμένα, γνωρίζοντας ότι το σήμα φωνής προκύπτει ως απόκριση του χρονικά μεταβαλλόμενου φίλτρου της φωνητικής οδού στη διέγερση του σήματος που παράγουν οι φωνητικές χορδές, μπορούμε να το αναπαραστήσουμε στη μορφή της Εξ. 2.7.

$$s(t) = \int_0^t h(t-\tau;t)e(t)d\tau \tag{2.7}$$

Στη συνέχεια, υποθέτουμε ότι η διέγερση μπορεί να γραφεί σαν άθροισμα ημιτονοειδών με αυθαίρετα πλάτη, συχνότητες και αρχικές φάσεις. Έτσι γράφοντας τη χρονικά μεταβαλλόμενη συνάρτηση μεταφοράς, που μοντελοποιεί το φωνητικό σωλήνα, σε πολική μορφή(Εξ. 2.8), προκύπτει ότι η κυματομορφή συνολικά γράφεται όπως παρουσιάζεται στην Εξ. 2.9.

$$H(\omega;t) = M(\omega;t)e^{j\Phi(\omega;t)} \tag{2.8}$$

$$s(t) = \sum_{l=1}^{L(t)} A_l(t)e^{j\psi(t)} \tag{2.9}$$

όπου

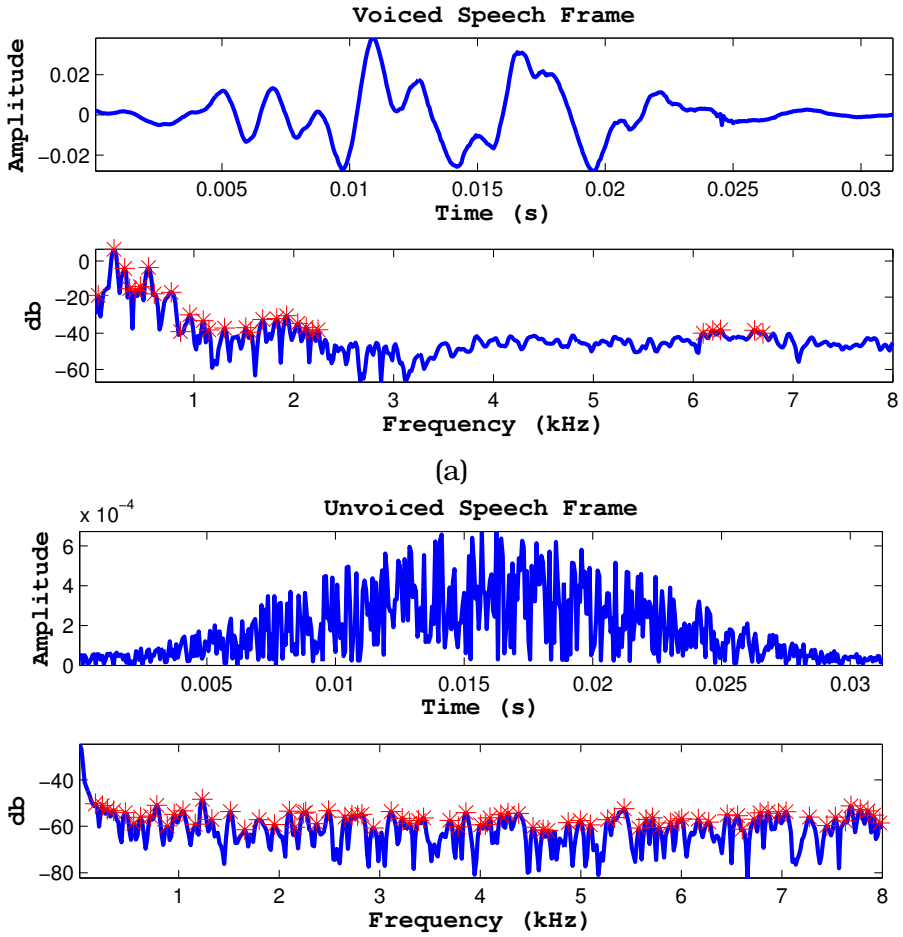
$$A_l(t) = a_l(t)M[\omega_l(t);t] \tag{2.10}$$

$$\psi_l(t) = \int_0^t \omega_l(\sigma)d\sigma + \Phi[\omega_l(t),t] + \phi_l \tag{2.11}$$

Μ' αυτές τις παραδοχές στη συνέχεια αναπτύσσεται μία σθεναρή διαδικασία εξαγωγής των σημαντικών συχνοτήτων του σήματος φωνής. Στις συγκεκριμένες συχνότητες υπολογίζονται και τα πλάτη και οι αρχικές φάσεις των ημιτονοειδών. Λαμβάνεται, λοιπόν, το σήμα στο k-th τμήμα του(Εξ. 2.12), χρησιμοποιείται η εκτιμήτρια συνάρτηση(Εξ. 2.13) ώστε να ελαχιστοποιηθεί το μέσο τετραγωνικό σφάλμα.

$$s(n) = \sum_{l=1}^{L^k} \gamma_l^k e^{jnl\omega_o^k} \tag{2.12}$$

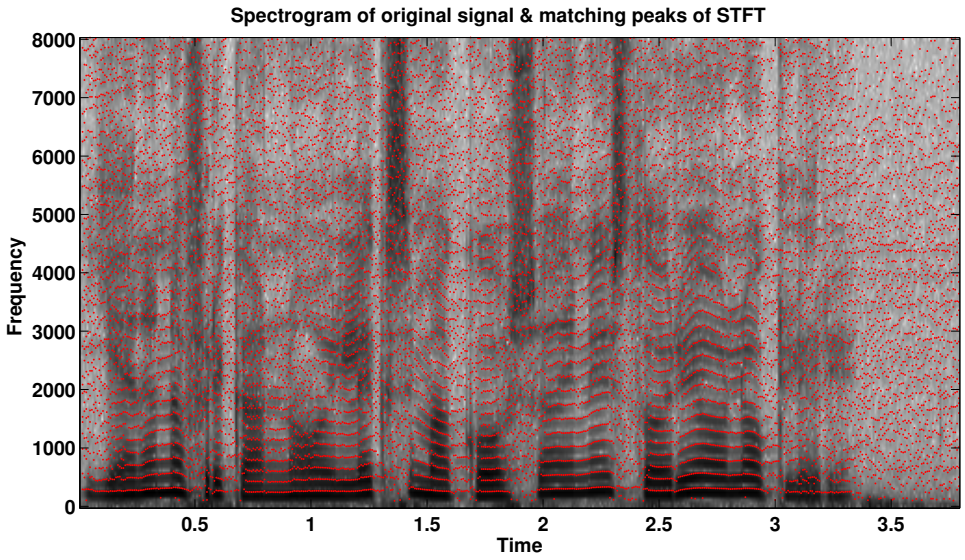
$$|Y(\omega)|^2 \approx \sum_{l=1}^{L^k} |\gamma_l^k|^2 sinc^2(\omega_l^k - \omega) \tag{2.13}$$



Σχήμα 2.5: Εντοπισμός Κορυφών στο φάσμα (a) του έμφωνου και (b) του άφωνου τμήματος ήχου

Στη συνέχεια εξάγονται οι κορυφές του περιοδογράμματος του σήματος φωνής που έχει παραθυρωθεί στο χρόνο με παράθυρο Hamming. Στο Σχήμα 2.6 είναι εμφανές ότι ένα έμφωνο τμήμα ήχου από ένα άφωνο έχει περισσότερο διακριτές κορυφές, οπότε και πιο σημαντικές. Έτσι, στη συνέχεια εφαρμόζεται ένας αλγόριθμος που ονομάζεται Frame-To-Frame Peak Matching, ο οποίος συσχετίζει της συχνοτικές κορυφές με αυτές των γειτονικών παραθύρων και έτσι προκύπτει η σημαντικότητα ή μη της κάθε κορυφής. Το αποτέλεσμα στο περιοδόγραμμα της δικής μας αρχικής φωνής φαίνεται στο Σχήμα 2.5.

Στη συνέχεια η ανασύνθεση του σήματος φωνής από τις συχνότητες των ημιτονοειδών, τα πλάτη και τις φάσεις γίνεται με την παρεμβολή των παρα-



Σχήμα 2.6: Η Επιλογή Κορυφών Απεικονισμένη Πάνω στο Περιοδόγραμμα του Σήματος Φωνής

μέτρων κατά μήκος των γειτονικών frames. Η παρεμβολή στα πλάτη είναι απλή και παρουσιάζεται στην Εξ. 2.14, ενώ η παρεμβολή στη φάση λόγω του ότι πρέπει να είναι συνεχής από frame σε frame γίνεται με ένα κυβικό πολυώνυμο(Εξ. 2.15) . Από τις συνθήκες συνέχειας (Εξ. 2.16, 2.17, 2.18, 2.19) προκύπτει ο υπολογισμός των παραμέτρων. Στη συνέχεια ελαχιστοποιώντας τη συνάρτηση ομαλότητας(Εξ. 2.20) προκύπτει η ανασύνθεση του σήματος φωνής(Εξ. 2.21).

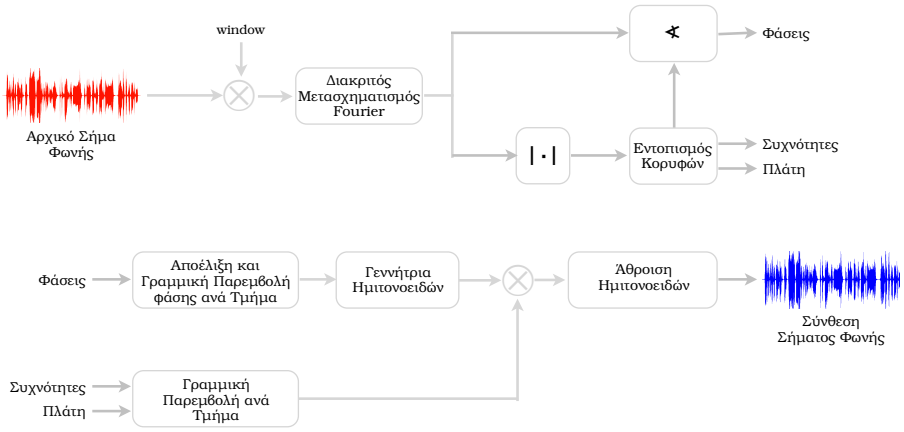
$$\tilde{A}(n) = \hat{A}^k + \frac{\hat{A}^{k+1} - \hat{A}^k}{S} n \tag{2.14}$$

$$\tilde{\theta}(t) = \zeta + \gamma t + \alpha t^2 + \beta t^3 \tag{2.15}$$

$$\tilde{\theta}(0) = \zeta = \hat{\theta}^k \tag{2.16}$$

$$\dot{\tilde{\theta}}(0) = \gamma = \hat{\omega}^k \tag{2.17}$$

$$\tilde{\theta}(T) = \tilde{\theta}^k + \tilde{\omega}^k T + \alpha T^2 + \beta T^3 = \tilde{\theta}^{k+1} + 2\pi M \tag{2.18}$$



Σχήμα 2.7: Βασική απεικόνιση ενός Vocoder που βασίζεται στην Ημιτονοειδή Ανάλυση/Σύνθεση της φωνής

$$\dot{\tilde{\theta}}(T) = \tilde{\omega}^k + \alpha T + 3\beta T^2 = \hat{\omega}^{k+1} \tag{2.19}$$

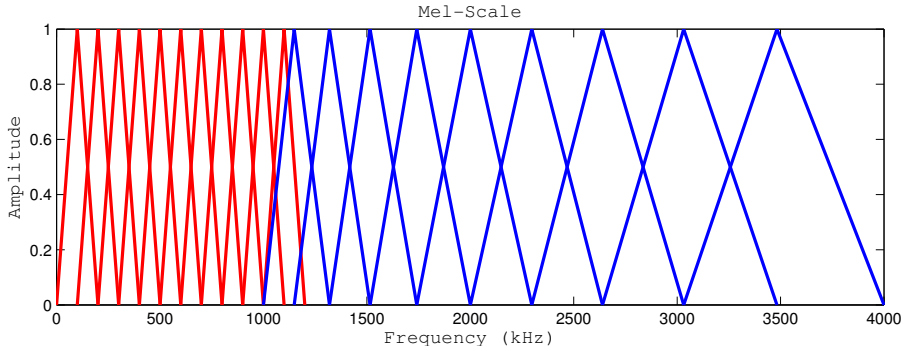
$$f(M) = \int_0^T [\ddot{\theta}(t; M)]^2 dt \tag{2.20}$$

$$\tilde{s}(n) = \sum_{l=1}^{L^k} \tilde{A}_l(n) cos[\tilde{\theta}_l(n)] \tag{2.21}$$

Το λειτουργικό διάγραμμα του Vocoder που βασίζεται στην ημιτονοειδή Ανάλυση και Σύνθεση παρουσιάζεται στο Σχήμα 2.7.

2.2.4 Vocoders στην κλίμακα Mel

Όπως προαναφέρθηκε, ένας Vocoder μπορεί να αναλύει το αρχικό σήμα φωνής τόσο στο πεδίο του χρόνου, κατακερματίζοντας το χρονικό άξονα σε παράθυρα, όσο και στο πεδίο της συχνότητας, όπου το κάθε χρονικό κομμάτι του σήματος φωνής παραθυρώνεται από μία τράπεζα φίλτρων που καλύπτουν διαφορετικά εύρη συχνοτήτων. Μετά από έρευνες της ψυχοακουστικής, έχει αποδειχθεί ότι το ανθρώπινο αυτί αντιλαμβάνεται τον ήχο αναλύοντάς τον σε συχνοτικές μπάντες. Είναι, λοιπόν, χρήσιμο να αναλύουμε το σήμα φωνής στο πεδίο της συχνότητας σε κλίμακες που προσεγγίζουν τον τρόπο αντίληψης των ήχων από το ανθρώπινο αυτί. Συγκεκριμένα τέτοιες κλίμακες είναι η Mel και η Bark. Η Mel κλίμακα αποτελεί αναδίπλωση του συχνοτικού άξονα σε



Σχήμα 2.8: Απεικόνιση της τράπεζας φίλτρων που προτάθηκε από τους Davis και Mermelstein για την ανάλυση του σήματος φωνής σε Mel κλίμακα. [6]

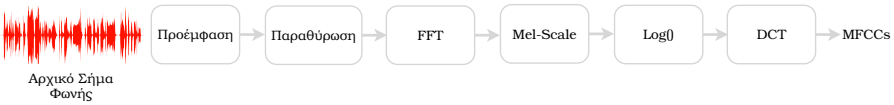
λογαριθμική κλίμακα. Μια τράπεζα φίλτρων σε Mel κλίμακα πρότειναν το 1980 οι Davis και Mermelstein [6], η οποία φαίνεται και στο Σχήμα 2.8. Όπως είναι εμφανές, τα φίλτρα είναι τριγωνικά με κεντρικές συχνότητες σε γραμμική κλίμακα από τα 100Hz έως τα 1 kHz, και σε λογαριθμική κλίμακα πάνω από το 1kHz.

Όσον αφορά την εξαγωγή χαρακτηριστικών από το τμήμα Ανάλυσης της φωνής ενός Vocoder, η χρησιμοποίηση της κλίμακας Mel είναι πολύ σημαντική για το λόγο ότι έχει παρατηρηθεί ότι σε μία μεγάλη περίοδο χρόνου η ενέργειες του κάθε καναλιού τείνουν να ακολουθήσουν κανονική κατανομή, αποτέλεσμα που συνδυάζεται άμεσα και με την επιλογή εκπαίδευσης των Κρυφών Μαρκοβιανών Μοντέλων του τελικού συστήματος.

MFCC Vocoder

Τα MFCC (Mel Frequency Cepstrum Coefficients) αποτελούν χαρακτηριστικά ανάλυσης της φωνής σε Mel κλίμακα. Η διαδικασία εξαγωγής των MFCC ξεκινάει με τον υπολογισμό του μετασχηματισμού Fourier των παραθυρωμένων τμημάτων του σήματος φωνής (Εξ. 2.23). Στη συνέχεια υπολογίζονται οι ενέργειες του φάσματος σε κάθε τριγωνικό παράθυρο στην κλίμακα Mel (Εξ. 2.24), όπως παρουσιάζεται στο Σχήμα 2.8. Τέλος, εφαρμόζεται ο διακριτός μετασχηματισμός συνημιτόνου DCT (Εξ. 2.25) στους λογαρίθμους των ενεργειών αυτών (Σχήμα 2.9).

$$y[n] = s[n]h[n] \tag{2.22}$$



Σχήμα 2.9: Λειτουργικό Διάγραμμα Εξαγωγής των MFCC. [26]

$$|Y(f)| = |\sum_{n=0}^{N-1} y[n] e^{\frac{-j2\pi fn}{N}}| \tag{2.23}$$

$$Y_k = \sum_{f=0}^{\frac{N}{2}-1} |Y(f)| w_k(f) \tag{2.24}$$

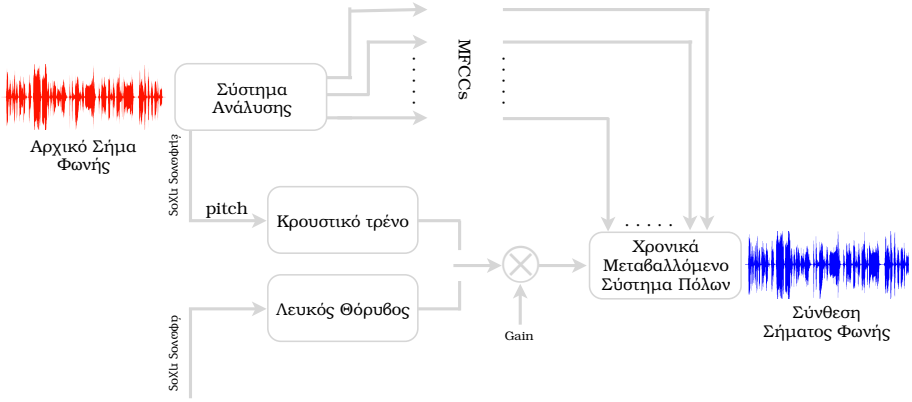
$$c_y[n] = u_n \sum_{k=0}^{K-1} (logY_k) cos\left(\frac{(2k+1)n\pi}{2K}\right) \tag{2.25}$$

Η σύνθεση του σήματος φωνής από τα MFCCs δεν μπορεί να γίνει με κάποια ανάστροφη διαδικασία. Έχουν προταθεί διάφοροι τρόποι σύνθεσης φωνής από ένα MFCC Vocoder, άλλοι βασίζονται στη φυσική μοντελοποίηση του φωνητικού σωλήνα [26] και άλλοι στην ανάλυση του ήχου με ημιτονοειδή [27]. Σύμφωνα με την πρώτη μοντελοποίηση, έχοντας τα MFCCs υπολογίζουμε με τον ανάστροφο διακριτό μετασχηματισμό συνημιτόνου, το λογάριθμο της εκτιμώμενης φασματικής ισχύος σε κλίμακα Mel (Εξ. 2.26). Παράλληλα, για να φτάσουμε στο αρχικό σήμα φωνής υπολογίζουμε μέχρι και το τελικό στάδιο των MFCCs την είσοδο της προέμφασης και της παραθύρωσης στη Mel κλίμακα. Εφόσον τα MFCC βρίσκονται στο πεδίο του cepstrum αφαιρούμε από τα τελικά MFCCs αυτά που μόλις υπολογίσαμε (Εξ. 2.27). Τέλος, υπολογίζοντας τους συντελεστές αυτοσυσχέτισης του μοντέλου του φωνητικού σωλήνα μέσα από την Εξίσωση 2.28, κατασκευάζουμε το τμήμα Σύνθεσης του vocoder, όπως και στην περίπτωση του LPC-Vocoder. Το λειτουργικό διάγραμμα φαίνεται στο Σχήμα 2.10.

$$log\hat{Y}_k = \sum_{n=0}^{K-1} u_n c_y(n) cos\left(\frac{(2k+1)n\pi}{2K}\right) \tag{2.26}$$

$$c_x = c_y - c_w - c_p \tag{2.27}$$

$$\hat{r}_i = \frac{1}{N} \sum_{f=0}^{N-1} |\hat{X}(f)|^2 e^{\frac{j2\pi fi}{N}} \tag{2.28}$$



Σχήμα 2.10: Λειτουργικό Διάγραμμα ενός MFCC Vocoder. [26]

Mel-Cepstral Vocoder

Στην mel-cepstral Ανάλυση/Σύνθεση, ακολουθείται το μοντέλο της φυσικής μοντελοποίησης του φωνητικού σωλήνα με ένα χρονικά μεταβαλλόμενο φίλτρο$H(e^{j\omega})$. Η συνάρτηση μεταφοράς είναι τάξης M και γράφεται συναρτήσει των Mel-cepstral συντελεστών $\tilde{c}(m)$ ως εξής:

$$H(z) = exp(\sum_{m=o}^{M} \tilde{c}(m)\tilde{z}^{-m}) \tag{2.29}$$

όπου

$$\tilde{z}^{-1} = \frac{z^{-1} - a}{1 - az^{-1}} \quad |a| < 1 \tag{2.30}$$

Αντικαθιστώντας τη μιγαδική μεταβλητή $\tilde{z} = e^{j\tilde{\omega}}$, προκύπτει η χαρακτηριστική της φάσης στην Εξ. 2.31.

$$\tilde{\omega} = tan^{-1}\frac{(1-a^2)sin\omega}{(1+a^2)cos\omega - 2a} \tag{2.31}$$

Για συχνότητα δειγματοληψίας 16kHz, το $\tilde{\omega}$ αποτελεί μια κάλη προσέγγιση της κλίμακας mel με την παράμετρο $a = 0.42$. Για να εξάγουμε μια αμερόληπτη εκτιμήτρια, χρησιμοποιούμε την πιο κάτω συνάρτηση(Εξ. 2.32) ελαχιστοποιώντάς την ως προς τους mel-cepstral συντελεστές [11].

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} expR(\omega) - R(\omega) - 1\ d\omega \tag{2.32}$$

όπου

$$R(\omega) = logI_N(\omega) - log|H(e^{j\omega})|^2 \tag{2.33}$$

και το $I_N(\omega)$ είναι το τροποποιημένο περιοδόγραμμα του παραθυρωμένου τμήματος φωνής μήκους N. Εξάγοντας τον παράγοντα κέρδους της $H(z)$, ξαναγράφουμε τη συνάρτηση μεταφοράς στη μορφή της Εξ. 2.34.

$$H(z) = exp\sum_{m=0}^{M} b(m)\Phi_m(z) = K \cdot D(z) \tag{2.34}$$

όπου

$$K = e^{b(0)} \tag{2.35}$$

$$D(z) = exp(\sum_{m=0}^{M} b(m)\Phi_m(z)) \tag{2.36}$$

και

$$b(m) = \begin{cases} c(m) & m = M \\ c(m) - ab(m+1) & 0 \leqslant m < M \end{cases} \tag{2.37}$$

$$\Phi_m(z) = \begin{cases} 1 & m = M \\ \frac{(1-a^2)z^{-1}}{1-az^{-1}}\tilde{z}^{-(m-1)} & 0 \leqslant m < M \end{cases} \tag{2.38}$$

Εφόσον η $H(z)$ είναι συνάρτηση μεταφοράς ελάχιστης φάσης (Minimum Phase), μπορούμε να δείξουμε ότι η ελαχιστοπόιηση του E (Εξ. 2.32) ως προς τους συντελεστές $\tilde{c}(m)_{m=0}^{M}$ είναι ισοδύναμο πρόβλημα με την ελαχιστοπόιηση του ϵ , όπως φαίνεται στην Εξ. 2.39.

$$\epsilon = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{I_N(\omega)}{|D(e^{j\omega}|} d\omega \tag{2.39}$$

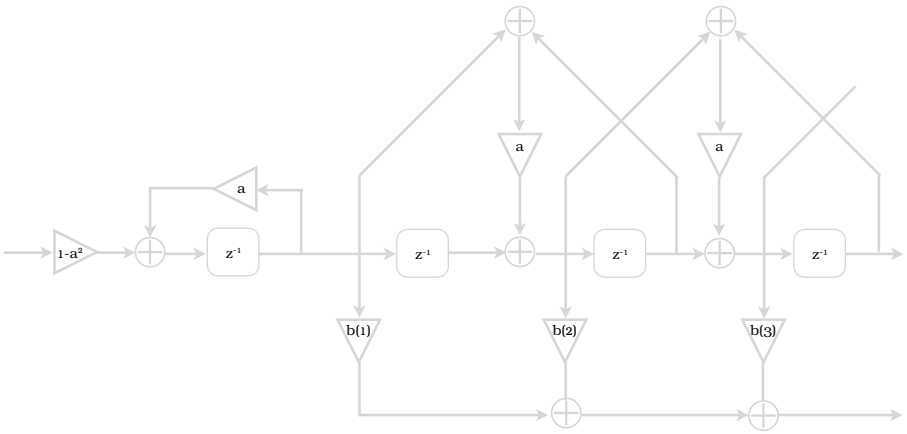
ως προς το

$$\mathbf{b} = [b(1), b(2), ..., b(M)]^T \tag{2.40}$$

Όπου το κέρδος K προσδιορίζεται στην Εξ. 2.41 από την ελαχιστοποίηση του E θέτοντας το $\frac{\partial E}{\partial K} = 0$, όπου ϵ_{min} είναι η ελάχιστη τιμή του ϵ.

$$K = \sqrt{\epsilon_{min}} \tag{2.41}$$

Δεδομένου ότι η χρονικά μεταβαλλόμενη συνάρτηση μεταφοράς της σύνθεσης (Εξ. 2.36) $D(z)$ δεν είναι ρητή συνάρτηση των mel-cepstrum συντελεστών, όπως στην περίπτωση του LPC Vocoder, χρησιμοποιείται το Mel



Σχήμα 2.11: Λειτουργικό Διάγραμμα βασικού φίλτρου F(z) για $L = 4$.

Log Spectrum Aproximation (MLSA) φίλτρο. Με αυτόν τον τρόπο η συνάρτηση μεταφοράς προσεγγίζεται με ιδιαίτερη ακρίβεια και μετατρέπεται σε ένα ελάχιστης φάσης IIR φίλτρο. Η μιγαδική εκθετική συνάρτηση $exp\omega$ προσεγγίζεται από τη ρητή συνάρτηση της εξίσωσης 2.42.

$$exp\omega \simeq R_L(F(z)) = \frac{1 + \sum_{l=1}^{L} A_{L,l}\omega^l}{1 + \sum_{l=1}^{L} A_{L,l}(-\omega)^l} \tag{2.42}$$

οπότε και η $D(z)$ προσεγγίζεται:

$$R_L(F(z)) \simeq exp(F(z)) = D(z) \tag{2.43}$$

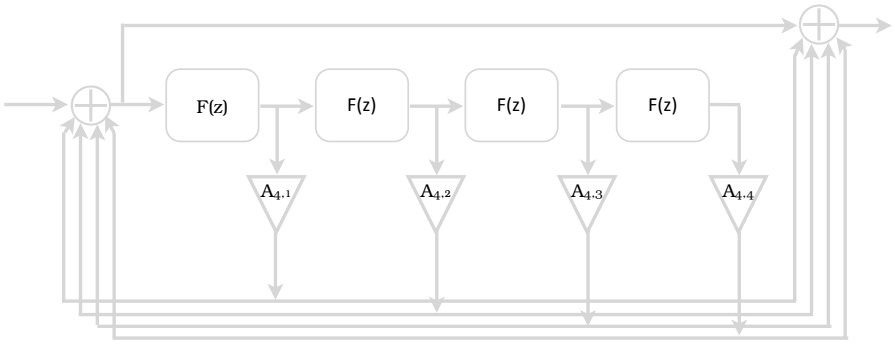
όπου

$$F(z) = \sum_{m=1}^{M} b(m)\Phi_m(z) \tag{2.44}$$

Το λειτουργικό διάγραμμα του MLSA φίλτρου φαίνεται στα Σχήματα 2.11, 2.12 για την περίπτωση που $L = 4$.

Mel-Generalized-Cepstral Vocoder

Η Mel-Generalized Cepstral Ανάλυση αποτελεί μία γενίκευση της Mel-Cepstral Ανάλυσης της προηγούμενης υποενότητας. Συγκεκριμένα, σε αυτή



Σχήμα 2.12: Λειτουργικό Διάγραμμα Προσέγγισης $R_L(F(z)) \simeq D(z)$ για $L = 4$.

Σχήμα 2.13: Απεικόνιση γενικευμένων χαρακτηριστικών

την ανάλυση του φωνητικού σήματος παρεισφρύει ένας συντελεστής γ, ο οποίος λαμβάνοντας τις ακραίες του τιμές -1 και 0 μετατρέπει τα εξαγόμενα χαρακτηριστικά σε LPC και Cepstral συντελεστές αντίστοιχα [38]. Στις ενδιάμεσες τιμές του διαμορφώνει ένα συνεχές πεδίο γενικευμένων χαρακτηριστικών όπου πραγματοποιείται η μετάβαση από τα LPC στα Cepstral. Παράλληλα, με την εισαγωγή και του συντελεστή a στα χαρακτηριστικά, ο οποίος αναδιαμορφώνει τον άξονα των συχνοτήτων, επιτυγχάνεται και η εξαγωγή χαρακτηριστικών μεταξύ LPC σε κλίμακα Mel και Mel-Cepstral χαρακτηριστικά. Για τις κατάλληλες τιμές του a και του γ λαμβάνουμε τα γενικευμένα Mel-cepstral χαρακτηριστικά ανάλυσης του σήματος φωνής, όπως φαίνεται και στο Σχήμα 2.13. Η γενικευμένη λογαριθμική συνάρτηση φαίνεται στην Εξ. 2.45

$$s_\gamma(w) = \begin{cases} (w^\gamma - 1)/\gamma, & 0 < |\gamma| \leqslant 1 \\ logw, & 0 \leqslant m < M \end{cases} \tag{2.45}$$

Χρησιμοποιώντας την πιο πάνω συνάρτηση υπολογίζουμε το γενικευμένο cepstrum ενός πραγματικού σήματος $x[n]$, και προκύπτει το αποτέλεσμα στην Εξ. 2.46.

$$s_\gamma(X(e^{j\omega})) = \sum_{m=-\infty}^{+\infty} c_{a,\gamma}(m) e^{-j\beta_a(\omega)m} \tag{2.46}$$

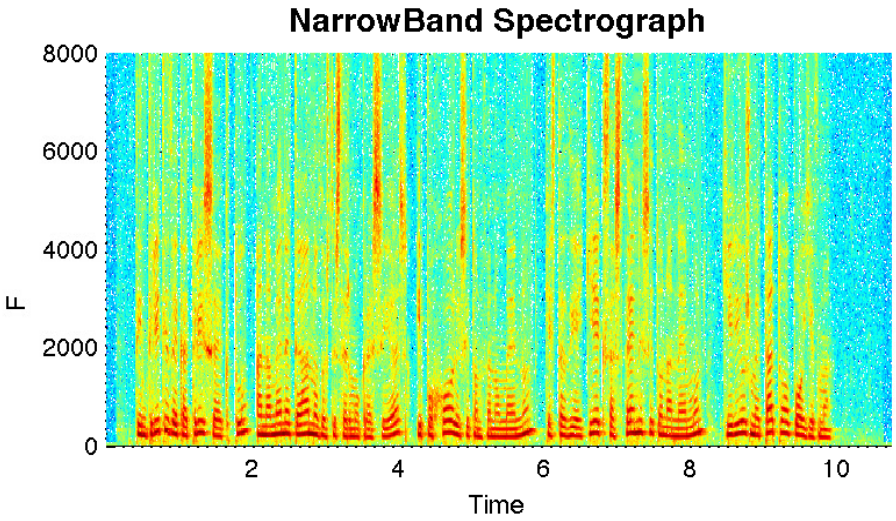
Συνεπώς η συνάρτηση μεταφοράς του συστήματος Σύνθεσης από γενικευμένα Mel-Cepstral χαρακτηριστικά προσεγγίζεται όπως στην Εξ. 2.47.

$$\begin{aligned} H(z) &= s_\gamma^{-1}\Big(\sum_{m=0}^{M} c_{a,\gamma}(m)\Psi_a^m(z)\Big)^{1/\gamma} = \\ &= \begin{cases} (1+\gamma\sum_{m=0}^{M} c_{a,\gamma}(m)\Psi_a^m(z))^{1/\gamma}, & 0 < |\gamma| \leqslant 1 \\ exp(\sum_{m=o}^{M} c_{a,\gamma}(m)\Psi_a^m(z)), & 0 \leqslant m < M \end{cases} \end{aligned} \tag{2.47}$$

Η διαδικασία ανάλυσης και σύνθεσης με τους γενικευμένους συντελεστές είναι ακριβώς αντίστοιχη με την προηγούμενη.

2.3 Υλοποίηση Vocoders - Πειραματικά Αποτελέσματα

Όλοι οι Vocoders υλοποιήθηκαν σε πρότυπο σήμα φωνής (βλ. στη βάση αρχείο με όνομα 1.wav) που ανήκει στη βάση εκφωνήσεων, με την οποία



Σχήμα 2.14: NarrowBand Spectrograph του φωνητικού σήματος 1.wav

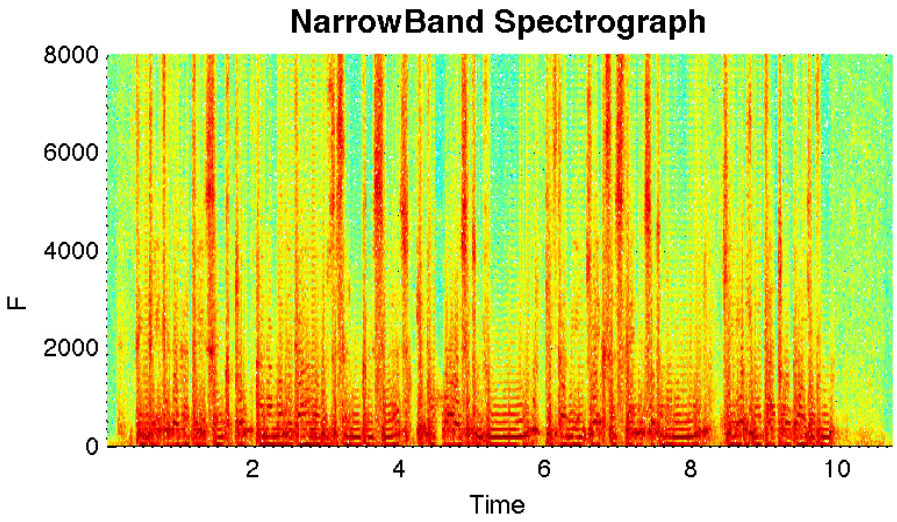
εκπαιδεύτηκε το συνολικό σύστημα. Η συχνότητα δειγματοληψίας ήταν ίση με 16kHz. Στο Σχήμα 2.14 φαίνεται το NarrowBand Spectrograph του αρχικού σήματος.

LPC-Vocoder

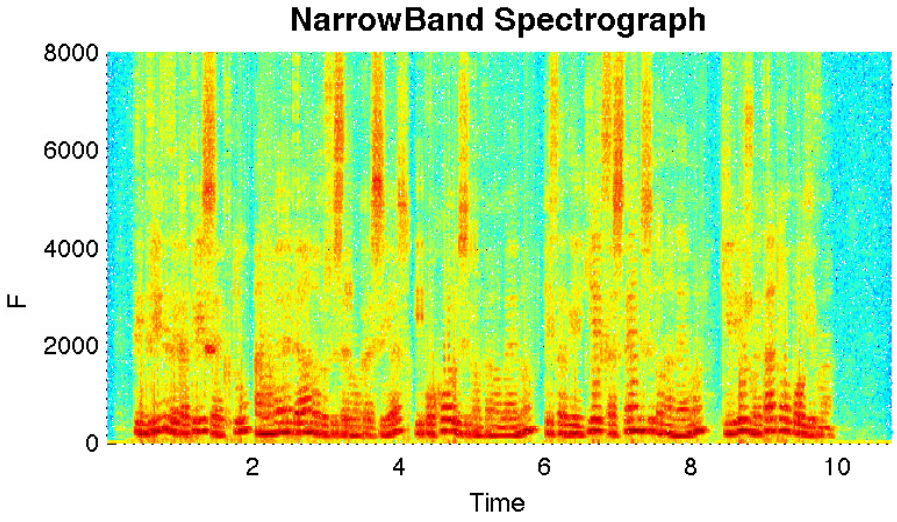
Ο LPC-Vocoder υλοποιήθηκε σε MATLAB. Συγκεκριμένα, η παραθύρωση του σήματος έγινε με παράθυρο Hamming 512 δειγμάτων με βήμα επικάλυψης κάθε 100 δείγματα. Το χρονικά μεταβαλλόμενο σύστημα μοντελοποίησης του φωνητικού σωλήνα, προσεγγίστηκε με 19 συντελεστές γραμμικής πρόβλεψης και 1 συντελεστή κέρδους. Η ακολουθία των pitch του σήματος φωνής υπολογίστηκε με το πρόγραμμα επεξεργασίας φωνής SPTK. Στο Σχήμα 2.15 παρουσιάζεται το NarrowBand Spectrograph της συνθετικής φωνής.

Phase Vocoder

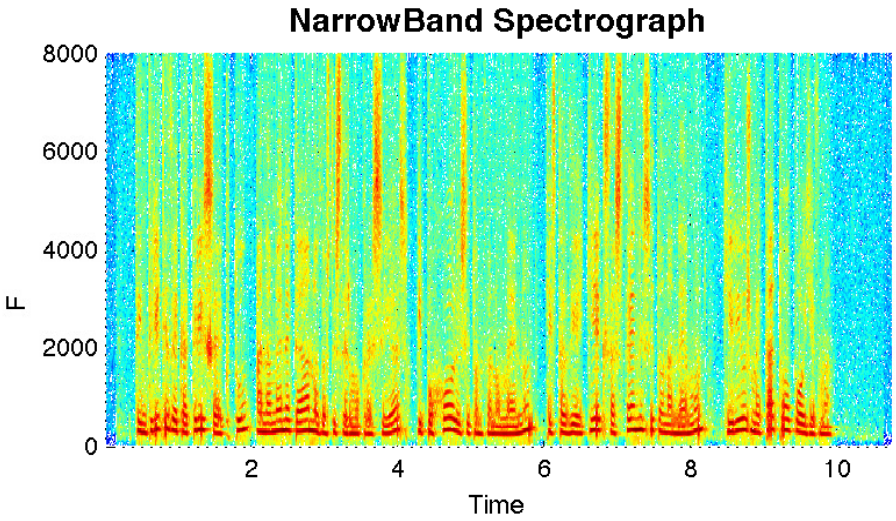
Ο Phase Vocoder υλοποιήθηκε σε MATLAB. Η εφαρμογή της Filter Bank για τον υπολογισμό του Μετασχηματισμού Fourier βραχέως χρόνου, έγινε με παραθύρωση Hanning ανά 1024 δείγματα φωνής. Το ακουστικό αποτέλεσμα βρίσκεται στη βάση της διπλωματικής σε αρχείο με όνομα s_voice_Phase.wav. Στο Σχήμα 2.16, φαίνεται το Σπεκτρογράφημα της συνθετικής φωνής.



Σχήμα 2.15: NarrowBand Spectrograph της Συνθετικής Φωνής με LPC Vocoder (βλ. s_voice_LPC.wav)



Σχήμα 2.16: NarrowBand Spectrograph της Συνθετικής Φωνής με Phase Vocoder (βλ. s_voice_Phase.wav)



Σχήμα 2.17: NarrowBand Spectrograph της Συνθετικής Φωνής με Vocoder που βασίζεται στην ημιτονοειδή Ανάλυση/Σύνθεση (βλ. s_voice_Sinusoidal.wav)

Vocoder βασισμένος στην ημιτονοειδή Ανάλυση/Σύνθεση

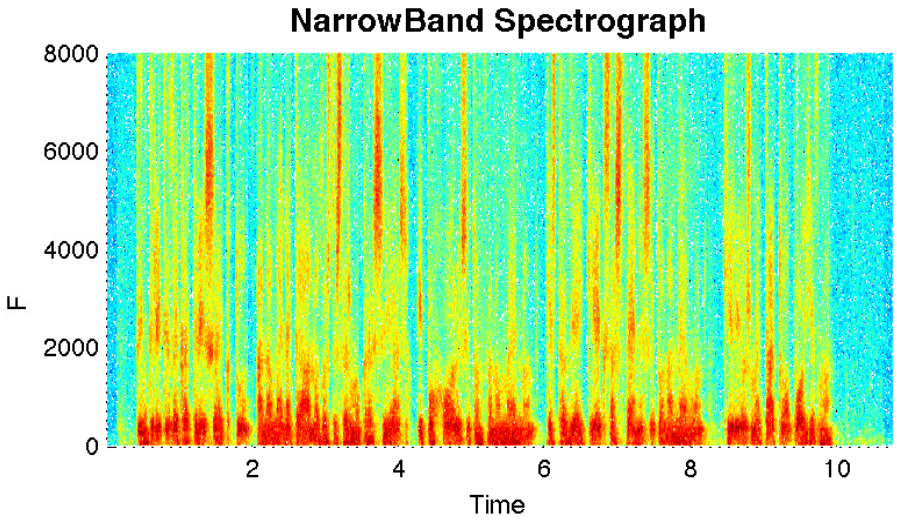
Ο Vocoder αυτός υλοποιήθηκε σε MATLAB. Συγκεκριμένα, εφαρμόστηκε παραθύρωση του αρχικού σήματος 256 δειγμάτων. Για την ανίχνευση των τοπικών μεγίστων και γαι τη γραμμική παρεμβολή χρησιμοποιήθηκαν οι συναρτήσεις quadmaxloc.m και slinterp.m [9]. Το αρχείο με όνομα s_voice_Sinusoidal.wav αποτελεί τη συνθετική φωνή με το συγκεκριμένο Vocoder. Στο Σχήμα 2.17 απεικονίζεται το Σπεκτρογράφημα του συνθετικού ήχου.

MFCC Vocoder

Ο Vocoder αυτός υλοποιήθηκε σε MATLAB. Συγκεκριμένα, χρησιμοποιήθηκαν οι συναρτήσεις melfcc.m και invmelfcc.m [10]. Ο συνθετικός ήχος του συγκεκριμένου Vocoder είναι αποθηκευμένος με όνομα s_voice_MFCC.wav. Στο Σχήμα 2.18 παρουσιάζεται το αντίστοιχο Σπεκτρογράφημα.

MCep Vocoder

Ο MCep Vocoder αυτός υλοποιήθηκε σε SPTK με μήκος απρθύρωσης 512 δειγμάτων, με βήμα μετατόπισης παραθύρου ίσο με 100 δείγματα, με υπολογισμό 20 συντελεστών MCep και ένα συντελεστή κέρδους καθώς και



Σχήμα 2.18: NarrowBand Spectrograph της Συνθετικής Φωνής με MFCC Vocoder (βλ. s_voice_MFCC.wav)

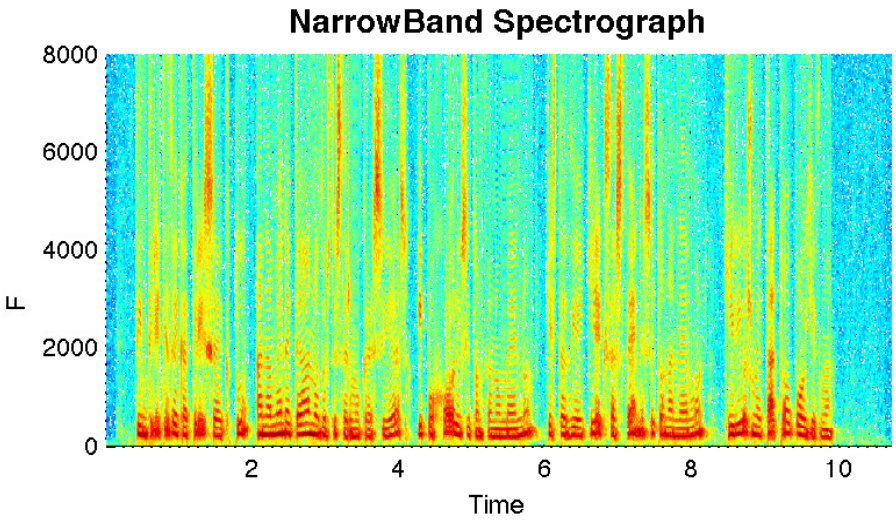
με συντελεστή $a = 0.42$. Συγκεκριμένα, το SPTK υλοποιεί το MLSA φίλτρο ανακατασκευής της φωνής από τους Mel Cepstrum συντελεστές. Πιο κάτω παρουσιάζεται το Σπεκτρογράφημα της συνθετικής φωνής(Σχήμα 2.19).

MGCep Vocoder

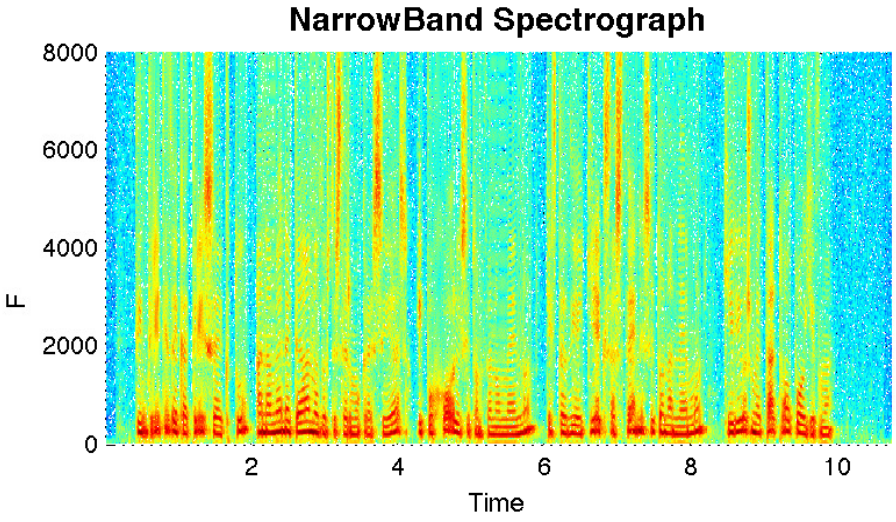
Ο MGCep Vocoder αυτός υλοποιήθηκε σε SPTK με τις παραμέτρους που υλοποιήθηκε και ο προηγούμενος Vocoder και με συντελεστή γενίκευσης $\gamma = -\frac{1}{7}$. Παράλληλα, ο συγκεκριμένος Vocoder δοκιμάστηκε και στη σύνθεση φωνής με 4 διαφορετικές εκδοχές παραμέτρων. Τα αποτελέσματα αυτών των πειραμάτων βρίσκονται στο φάκελο MGCep της βάσης της Διπλωματικής. Πιο κάτω παρουσιάζεται το Σπεκτρογράφημα της συνθετικής φωνής(Σχήμα 2.20).

2.4 Σύγκριση των Εξαγομενων Χαρακτηριστικών

Μετά την υλοποίηση των πιο πάνω Vocoders τα κριτήρια επιλογής των χαρακτηριστικών που χρησιμοποιήθηκαν στη συγκεκριμένη διπλωματική αφορούν τόσο την ποιότητα της συνθετικής φωνής που προκύπτει από τους Vocoders, όσο και το πόσο εύρωστα είναι τα αντίστοιχα χαρακτηριστικά ώστε να χρησιμοποιηθούν στη μοντελοποίηση του συστήματος σύνθεσης φωνής με Κρυφά Μαρκοβιανά Μοντέλα.



Σχήμα 2.19: NarrowBand Spectrograph της Συνθετικής Φωνής με MCep Vocoder (βλ. s_voice_MCep.wav)



Σχήμα 2.20: NarrowBand Spectrograph της Συνθετικής Φωνής με MGCep Vocoder (βλ. s_voice_MGCep.wav)

2.4.1 Ποιότητα Ακουστικών Αποτελεσμάτων

Ξεκινώντας από τον LPC Vocoder η προκύπτουσα συνθετική φωνή είναι κατανοητή αλλά δεν έχει καλή ακουστική ποιότητα, εφόσον εμπεριέχει αρκετά στοιχεία ανεπιθύμητου θορύβου. Αυτό συμβαίνει λόγω του ότι η διέγερση του χρονικά μεταβαλλόμενου συστήματος είναι ένα απλό κρουστικό τρένο με περίοδο pitch. Προχωρώντας στον Phase Vocoder η σύνθεση έχει καλύτερα αποτελέσματα. Παρόλα αυτά, ο Phase Vocoder στο στάδιο της ανάλυσης απλά επιτυγχάνει το διαχωρισμό της φάσης στο πεδίο της συχνότητας, χωρίς δηλαδή να εξάγει συγκεκριμένα χαρακτηριστικά τα οποία θα μπορούσαν να χρησιμοποιηθούν στη μοντελοποίηση του Συνθέτη φωνής από Κείμενο. Σχεδόν τέλεια απόδοση φωνής έχει ο Vocoder της Ημιτονοειδούς Ανάλυσης/Σύνθεσης. Παράλληλα επιτυγχάνει μεγάλη συμπίεση δεδομένων, εξάγοντας χαρακτηριστικά που πιθανόν να χρησιμοποιηθούν στη συγκεκριμένη μοντελοποίηση. Πιο κάτω παρατίθεται ένας πίνακας σύγκρισης της ακουστικής ποιότητας των αποτελεσμάτων των αλγορίθμων, με βάση το Mean Opinion Score[1], έτσι όπως προέκυψε από την βαθμολόγησή τους από 73 άτομα σε κλίμακα από το 1 έως το 5. Τα αποτελέσματα της συγκεκριμένης έρευνας δημοσιεύτηκαν στο ΣΦΗΜΜΥ 2009 μετά από εργασία, που έγινε στα πλαίσια του μαθήματος "Ψηφιακής Επεξεργασίας Σήματος". Η συγκεκριμένη εργασία αποτελεί μία πρώτη έρευνα της διπλωματικής [14].

Μέθοδος	**Βαθμολογία**	
	Σήμα Φωνής	Σήμα Μουσικής
LPC error-excited Vocoder	2.7167	1,7205
Phase Vocoder	2.2478	1.6754
Ημιτονοειδής Ανάλυση/Σύνθεση	3.1103	3.0556

Παράλληλα, συνεχίζοντας από τον MFCC Vocoder προς τους Mel-Cepstrum Vocoders τα αποτελέσματα βελτιώνονται κατά πολύ. Ο πρώτος έχει εξίσου χαμηλής ποιότητας απόδοση με τον LPC ενώ οι MCep και MGCep Vocoders δίνουν αποτελέσματα σχεδόν φυσικά.

2.4.2 Ευρωστία Χρήσης των Εξαγόμενων Χαρακτηριστικών

Μία πολύ σημαντική παράμετρος επιλογής των χαρακτηριστικών είναι συμπίεση δεδομένων, ώστε να μειωθεί η πολυπλοκότητα του συστηματος και

[1]Το χρονικά μεταβαλλόμενο σύστημα με LPC μοντελοποίηση, σε αυτήν την εργασία διεγείρεται από το σφάλμα Πρόβλεψης και όχι από κάποιο κρουστικό τρένο με περίοδο pitch, όπως στη δική μας υλοποίηση, και γι' αυτό έχει καλύτερη επίδοση από τον Phase Vocoder.

να γίνει επεξεργασία λιγότερων δεδομένων, τόσο στο στάδιο της εκπαίδευσης όσο και στο στάδιο της παραγωγής των χαρακτηριστικών. Οι πιο πολλοί Vocoders χρησιμοποιούν την ακολουθία του pitch και μέχρι 20 χαρακτηριστικά ανά χρονικό πλαίσιο των 100 δειγμάτων περίπου. Συνεπώς επιτυγχάνουν σχεδόν συμπίεση της τάξης του 80%. Μία άλλη σημαντική παράμετρος επιλογής είναι ότι τα χαρακτηριστικά των τριών τελευταίων Vocoders είναι ανεπτυγμένα στην κλίμακα Mel. Αξίζει να συνυπολογιστεί ότι το σύστημα μετατροπής του γραπτού λόγου σε συνθετική φωνή που βασίζεται στη μοντελοποίηση με Κρυφά Μαρκοβιανά Μοντέλα, έχει ακριβώς ανάστροφη δομή από αυτήν ενός συστήματος αναγνώρισης φωνής με Κρυφά Μαρκοβιανά Μοντέλα. Συνεπώς, είναι γνωστό ότι χαρακτηριστικά σε συχνοτικές κλίμακες κοντά στην αντίληψη του ανθρώπου (όπως η κλίμακα Mel και η Bark) βελτιώνουν τα ποσοστά επιτυχίας των συστημάτων Αναγνώρισης Φωνής. Αυτό, βέβαια δεν είναι άσχετο από το ότι τα χαρακτηριστικά σε μεγάλα χρονικά τμήματα φωνητικών σημάτων, όταν υπολογίζονται σε Mel κλίμακα ακολουθούν Γκαουσιανή Κατανομή, αφού με τέτοιου είδους κατανομές γίνεται και η εκπαίδευση των Κρυφών Μαρκοβιανών Μοντέλων του συνολικού συστήματος.

Κεφάλαιο 3

Εκπαίδευση των HMMs του Συστήματος

3.1 Μοντελοποίηση των Mel-Cepstrum Χαρακτηριστικών

Το βασικό τμήμα εκπαίδευσης των Κρυφών Μαρκοβιανών Μοντέλων, δε διαφέρει καθόλου από αυτό ενός συστήματος αναγνώρισης φωνής που βασίζεται σε HMMs. Συνεπώς, όσον αφορά τη μοντελοποίηση των χαρακτηριστικών που προσεγγίζουν τη χρονικά μεταβαλλόμενη συνάρτηση μεταφοράς του φωνητικού σωλήνα, χρησιμοποιούνται Κρυφά Μαρκοβιανά Μοντέλα συνεχούς συνάρτησης πυκνότητας πιθανότητας (Continuous Density HMMs).

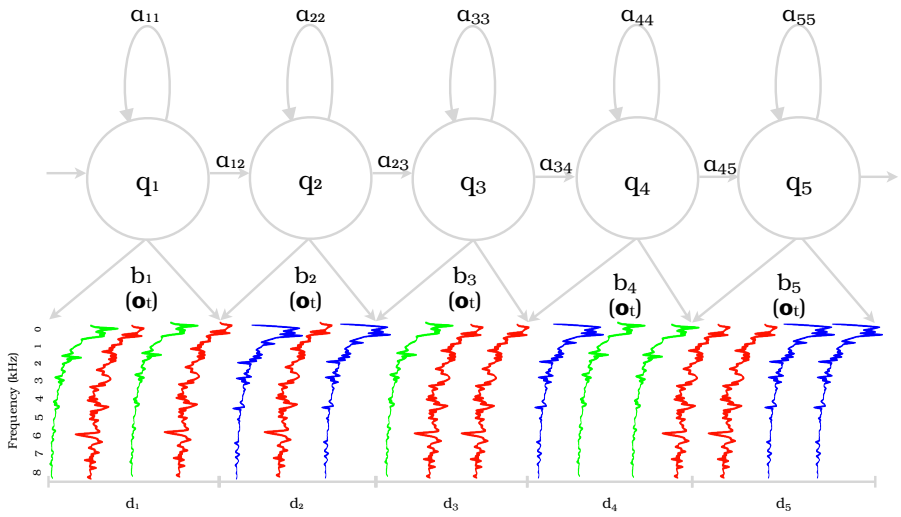
3.1.1 Βασική Περιγραφή ενός HMM

Ένα Κρυφό Μαρκοβιανό Μοντέλο αποτελεί μία μηχανή πεπερασμένων καταστάσεων, η οποία καθορίζεται από συγκεκριμένες παραμέτρους [31, 32]:

1. Έχει N πεπερασμένες κρυφές καταστάσεις. Σε μία τυχαία χρονική στιγμή t το μοντέλο βρίσκεται στην κατάσταση q_t, όπου η $q_t = \{1, 2, ..., N\}$

2. Από κάθε κατάσταση εξάγεται μία παρατήρηση $\mathbf{o}_t$, η οποία ανήκει στο χώρο των στοιχείων, τα οποία μοντελοποιούνται από το Κρυφό Μαρκοβιανό Μοντέλο.

3. Σε κάθε κατάσταση αντιστοιχεί μία πιθανότητα μετάβασης από την κατάσταση i στην κατάσταση j, η $A = \{a_{ij}\}$, όπου



Σχήμα 3.1: Αναπαράσταση ενός HMM 5 καταστάσεων που μοντελοποιεί ένα φωνητικό σήμα

$$a_{ij} = P[q_{t+1} = j | q_t = i], \quad 1 \leq i, j \leq N \tag{3.1}$$

4. Κάθε κατάσταση του HMM έχει τη δική της συνάρτηση κατανομής εξαγωγής της παρατήρησης. Αυτή μοντελοποιείται για παρατηρήσεις συνεχείς και όχι διακριτές, με μία συνάρτηση από M Γκαουσιανές συναρτήσεις πυκνότητας πιαθανότητας(βλ. Εξ. 3.2)

$$b_j(\mathbf{o}_t) = P[\mathbf{o}_t | q_t = j] \tag{3.2}$$

5. Στην αρχική κατάσταση αντιστοιχίζεται μία κατανομή πιθανότητας $\pi = \{\pi_i\}$, η οποία δίνει την πιθανότητα να ξεκινάει το HMM από τη συγκεκριμένη κατάσταση i.

$$\pi_i = P[q_1 = i], \quad 1 \leq i \leq N \tag{3.3}$$

Συνοπτικά, λοιπόν, μπορούμε να αναφερόμαστε σε ένα HMM με το συμβολισμό:

$$\lambda = (A, B, \pi) \tag{3.4}$$

Όσον αφορά τη μοντελοποίηση της ακολουθίας φωνητικών χαρακτηριστικών, τα HMMs είναι της μορφής όπως στο Σχήμα 3.1. Συγκεκριμένα, η ροή των καταστάσεων είναι δεξιόστροφη, χωρίς να υπάρχει η δυνατότητα ανάστροφης ροής. Παράλληλα, δεν μπορεί να γίνει μετάβαση προσπερνώντας κάποια κατάσταση. Αυτή, η λογική ταιριάζει στο φυσικό μοντέλο ροής της φωνής με το χρόνο. Μαθηματικά, αυτές οι συνθήκες αποτυπώνονται στην Εξ. 3.5. Παράλληλα, η πιθανότητα της κάθε κατάστασης να εξάγει μία παρατήρηση δίνεται από την υπέρθεση M Γκαουσιανών κατανομών όπως παρουσιάζεται στην Εξ. 3.6.

$$a_{ii} + a_{ij} = 1, \quad j = 1 + 1 \tag{3.5}$$

$$\begin{aligned} b_j(\mathbf{o}_t) &= \sum_{k=1}^{M} c_{jk} \mathcal{N}(\mathbf{o}_t, \mu_{jk}, \Sigma_{jk}) = \\ &= \sum_{k=1}^{M} c_{jk} \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_{jk}|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\mathbf{o}_t - \mu_{jk})^T \Sigma_{jk}^{-1} (\mathbf{o}_t - \mu_{jk})\} \end{aligned} \tag{3.6}$$

Όπου c_{jk}, μ_{jk} και Σ_{jk} είναι οι συντελεστές των Γκαουσιανών, το διάνυσμα των μέσων όρων και ο τετραγωνικός πίνακας συμμεταβλητότητας για την k-οστή κατανομή της κατάστασης j, αντίστοιχα. Ο πίνακας συμμεταβλητότητας λαμβάνεται ως διαγώνιος εάν τα στοιχεία θεωρούνται ανεξάρτητα. Προφανώς, οι συντελεστές των κατανομών ικανοποιούν τη συνθήκη της Εξ. 3.7.

$$\begin{aligned} \sum_{k=1}^{M} c_{jk} = 1, \quad 1 \leq j \leq N \\ c_{jk} \geq 0, \quad 1 \leq j \leq N, 1 \leq k \leq M \end{aligned} \tag{3.7}$$

Τέλος, δεδομένου ότι υιοθετείται το δεξιόστροφο μοντέλο, η πιθανότητα ξεκινήματος ισούται με 1 για το ξεκίνημα από την 1η κατάσταση, όπως φαίνεται και στην Εξ. 3.8.

$$\pi = \begin{cases} 0 & i \neq 1 \\ 1 & i = 1 \end{cases} \tag{3.8}$$

3.1.2 Υπολογισμός των Πιθανοτήτων

Είναι γνωστό ότι τα βασικά προβλήματα προς επίλυση δεδομένου ενός HMM είναι τρία, έτσι όπως διατυπώνονται στη βιβλιογραφία [31, 32].

Το πρώτο και πιο απλό πρόβλημα διατυπώνεται ως εξής:

Πρόβλημα 1

Δοθείσης της ακολουθίας παρατηρήσεων $\mathbf{O} = (\mathbf{o}_1\mathbf{o}_2...\mathbf{o}_T)$ και ενός μοντέλου HMM με παραμέτρους $\lambda = (A, B, \pi)$, πώς μπορούμε να υπολογίσουμε την πιθανότητα $P(\mathbf{O}|\lambda)$ της ακολουθίας παρατηρήσεων δεδομένου του μοντέλου;

Για την επίλυση του παραπάνω προβλήματος υπάρχει ο απευθείας τρόπος υπολογισμού της πιθανότητας, ο οποίος αλγοριθμικά χρειάζεται περίπου $2T \cdot N^2$ υπολογισμοί. Για την αποφυγή, λοιπόν, του απευθείας υπολογισμού, χρησιμοποιείται ο forward-backward αλγόριθμος, ο οποίος περιγράφεται πιο κάτω.

Forward Algorithm

Έστω η μεταβλητή $\alpha_t(i)$ η οποία ορίζεται όπως στην Εξ. 3.9

$$\alpha_t(i) = P(\mathbf{o}_1\mathbf{o}_2...\mathbf{o}_t, q_t = i|\lambda) \tag{3.9}$$

Συνεπώς, πρόκειται για την πιθανότητα να συμβεί μία ακολουθία παρατηρήσεων μέχρι τη χρονική στιγμή t στην κατάσταση i, δεδομένου του μοντέλου λ. Με αναδρομή διατυπώνεται πιο κάτω ο αλγόριθμος σε βήματα:

1. Αρχικοποίηση

$$\alpha_1(i) = \pi_i b_i(\mathbf{o}_1), \quad 1 \leq i \leq N \tag{3.10}$$

2. Αναδρομή

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i)a_{ij}]b_j(\mathbf{o}_{t+1}), \quad \begin{cases} 1 \leq t \leq T-1 \\ 1 \leq j \leq N \end{cases} \tag{3.11}$$

3. Τερματισμός

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{3.12}$$

Backward Algorithm

Ακριβώς αντίστοιχα με πριν, έστω η μεταβλητή $\beta_t(i)$, η οποία ορίζεται όπως στην Εξ. 3.13

$$\beta_t(i) = P(\mathbf{o}_{t+1}\mathbf{o}_{t+2}...\mathbf{o}_T|q_t = i, \lambda) \tag{3.13}$$

Συνεπώς, πρόκειται για την πιθανότητα να συμβεί μία ακολουθία παρατηρήσεων από τη χρονική στιγμή $t+1$ μέχρι το τέλος, δεδομένου ότι βρίσκεται στην κατάσταση i και δεδομένου του μοντέλου λ. Με αναδρομή διατυπώνεται πιο κάτω ο αλγόριθμος σε βήματα:

1. Αρχικοποίηση

$$\beta_T(i) = 1, \quad 1 \leq i \leq N \tag{3.14}$$

2. Αναδρομή

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{o}_{t+1}\beta_{t+1}(j)), \quad \begin{cases} t = T-1, .., 1 \\ 1 \leq i \leq N \end{cases} \tag{3.15}$$

3. Τερματισμός

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \beta_1(i) \tag{3.16}$$

Η λογική των δύο παραπάνω αλγορίθμων, βασίζεται στη δομή των διαγραμμάτων trellis [32]. Η πολυπλοκότητα αυτών των αλγορίθμων μειώνεται κατά πολύ σε σχέση με τον απευθείας υπολογισμό, αφού χρειάζονται μόνο $N(N-1)(T-1)$ υπολογισμοί.

3.1.3 Ρύθμιση των Παραμέτρων του HMM

Το τρίτο και πιο πολύπλοκο πρόβλημα των HMMs διατυπώνεται ως εξής:

Πρόβλημα 3

Πώς ρυθμίζονται οι παράμετροι του μοντέλου $\lambda = (A, B, \pi)$, ώστε να μεγιστοποιείται η πιθανότητα $P(\mathbf{O}|\lambda)$;

Αυτό που καθιστά δύσκολη την επίλυση αυτού του προβλήματος είναι ότι δεν υπάρχει αναλυτικός τρόπος επίλυσής του. Παρόλ' αυτά υπάρχει τρόπος να υπολογίσουμε τις παραμέτρους του $\lambda = (A, B, \pi)$ έτσι ώστε η πιθανότητα $P(\mathbf{O}|\lambda)$ να μεγιστοποιείται τοπικά. Αυτό γίνεται μέσω του αλγορίθμου του Baum-Welch, του γνωστού EM (Expectation Maximization) αλγορίθμου.

Ξεκινώντας τη διαδικασία του προσδιορισμού των παραμέτρων, καθορίζουμε την πιθανότητα $\xi_t(i, j)$ να βρίσκεται στην κατάσταση i τη χρονική στιγμή t, και στην κατάσταση j τη χρονική στιγμή $t + 1$, δοθέντος του μοντέλου και της ακολουθίας παρατηρήσεων(βλ. Εξ. 3.17).

$$\xi_t(i, j) = P(q_t = i, q_{t+1} = j|\mathbf{O}, \lambda) \tag{3.17}$$

Από τον ορισμό των παραμέτρων των αλγορίθμων forward και backward, η παραπάνω πιθανότητα μπορεί να γραφεί όπως στην Εξ. 3.18.

$$\begin{aligned} \xi_t(i, j) &= \frac{P(q_t = i, q_{t+1} = j, \mathbf{O}|\lambda)}{P(\mathbf{O}|\lambda)} = \\ &= \frac{\alpha_t(i)a_{ij}b_j(\mathbf{o}_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(\mathbf{o}_{t+1})\beta_{t+1}(j)} \end{aligned} \tag{3.18}$$

Ορίζοντας την πιθανότητα $\gamma_t(i)$ ως την πιθανότητα του να βρίσκεται στην κατάσταση i τη χρονική στιγμή t, υπολογίζεται ότι

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j) \tag{3.19}$$

Ο αλγόριθμος EM μπορεί να εξαχθεί από τον ορισμό της βοηθητικής συνάρτησης του Baum(βλ. Εξ. 3.20) ως προς τις παραμέτρους του λ.

$$Q(\lambda', \lambda) = \sum_{\mathbf{q}} P(\mathbf{O}, \mathbf{q}|\lambda') \log P(\mathbf{O}, \mathbf{q}|\lambda) \tag{3.20}$$

$$Q(\lambda', \lambda) \geq Q(\lambda', \lambda') \Rightarrow P(\mathbf{O}, \mathbf{q}|\lambda) \geq P(\mathbf{O}, \mathbf{q}|\lambda') \tag{3.21}$$

Η βασική λογική είναι η μεγιστοποίηση της $Q(\lambda', \lambda)$ ως προς το λ ώστε να βελτιωθεί η παράμετρος λ' και να αυξηθεί η πιθανοφάνεια $P(\mathbf{O}, \mathbf{q}|\lambda)$.

Μεγιστοποίηση της βοηθητικής συνάρτησης Q

Για μία δοθείσα ακολουθία παρατηρήσεων $\mathbf{O}$ και ένα μοντέλο λ', εξάγονται οι παράμετροι του λ, οι οποίοι μεγιστοποιούν την $Q(\lambda', \lambda)$. Η πιθανοφάνεια γράφεται στην Εξ. 3.22

$$P(\mathbf{O}, \mathbf{q}|\lambda) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\mathbf{o}_t) \tag{3.22}$$

$$\log P(\mathbf{O}, \mathbf{q}|\lambda) = \log \pi_{q_0} + \sum_{t=1}^{T} \log a_{q_{t-1}q_t} + \sum_{t=1}^{T} \log b_{q_t}(\mathbf{o}_t) \tag{3.23}$$

Η βοηθητική συνάρτηση Q γράφεται όπως στην Εξ. 3.24

$$\begin{aligned} Q(\lambda', \lambda) &= Q_\pi(\lambda', \pi) + \\ &+ \sum_{t=1}^{T} Q_{a_i}(\lambda', \mathbf{a}_i) + \\ &+ \sum_{t=1}^{T} Q_{b_i}(\lambda', \mathbf{b}_i) = \\ &= \sum_{i=1}^{N} P(\mathbf{O}, q_0 = i|\lambda') \log \pi_i + \\ &+ \sum_{j=1}^{N} \sum_{t=1}^{T} P(\mathbf{O}, q_{t-1} = i, q_t = j|\lambda') \log a_{ij} + \\ &+ \sum_{i=1}^{N} \sum_{t=1}^{T} P(\mathbf{O}, q_t = i|\lambda') \log b_i(\mathbf{o}_t) \end{aligned} \tag{3.24}$$

όπου τηρούνται όλοι οι στοχαστικοί περιορισμοί. Έτσι προκύπτουν τα αποτελέσματα των παραμέτρων μοντελοποίησης από τον αλγόριθμο ΕΜ, όπως φαίνονται στις Εξ. 3.25,3.26,3.27.

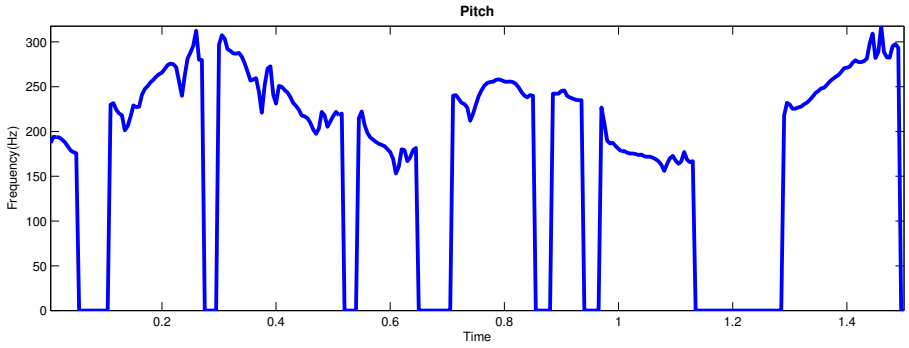
$$c_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \tag{3.25}$$

$$\mu_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot \mathbf{o}_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3.26}$$

$$\mathbf{\Sigma}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot (\mathbf{o}_t - \mu_{jk})(\mathbf{o}_t - \mu_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3.27}$$

όπου

$$\gamma_t(j,k) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \cdot \frac{c_{jk}\mathcal{N}(\mathbf{o}_t, \mathbf{o}_{jk}, \mathbf{\Sigma}_{jk})}{\sum_{m=1}^{M} \mathcal{N}(\mathbf{o}_t, \mathbf{o}_{jk}, \mathbf{\Sigma}_{jk})} \tag{3.28}$$



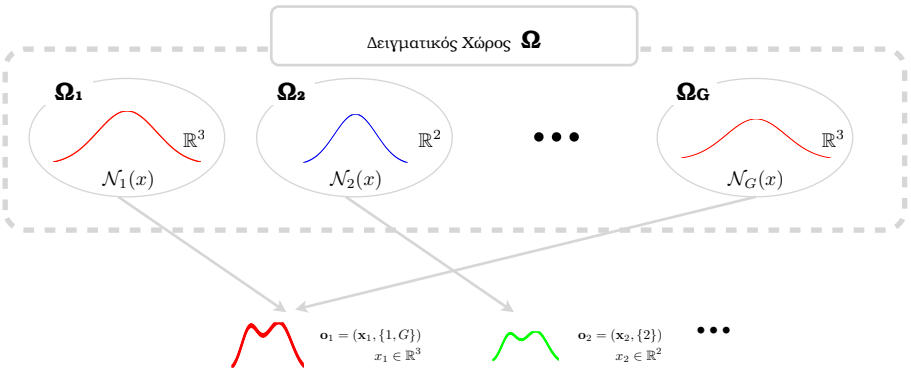
Σχήμα 3.2: Απεικόνιση μιας εξαγόμενης ακολουθίας pitch.

3.2 Μοντελοποίηση της Ακολουθίας Pitch

Κατά την εξαγωγή της ακολουθίας του pitch από ένα σήμα φωνής, είναι αναμενόμενο, τα τμήματα έμφωνου ήχου να παρουσιάζουν κάποιες τιμές pitch και τα άφωνα τμήματα να μην παρουσιάζουν τιμές pitch. Κάτι τέτοιο είναι εμφανές στο Σχήμα 3.2. Αυτό έχει ως αποτέλεσμα, να εμφανίζεται κάποια δυσκολία όσον αφορά τη μοντελοποίηση των ακολουθιών του pitch με Κρυφά Μαρκοβιανά Μοντέλα συνεχούς τυχαίας μεταβλητής, όπως τα προηγούμενα. Αυτή η ιδιαιτερότητα, λοιπόν, αντιμετωπίζεται με τη χρήση ενός συνδυαστικού Κρυφού Μαρκοβιανού μοντέλου, όπου η συνάρτηση κατανομής των παρατηρήσεων δεν είναι ούτε διακριτή, ούτε συνεχής. Συγκεκριμένα, υιοθετείται ένα πιθανοτικό μοντέλο σε πολλαπλούς χώρους.

3.2.1 Συνάρτηση Κατανομής Πιθανότητας σε Πολλαπλούς Χώρους

Για τον ορισμό αυτής της συνάρτησης κατανομής πιθανότητας, υποθέτουμε ότι έχουμε ένα δειγματικό χώρο $\mathbf{\Omega}$ ο οποίος αποτελείται από επιμέρους δειγματικούς χώρους $\mathbf{\Omega}_g$, όπου $\Omega = \bigcup_{g=1}^{G} \Omega_g$. διαφορετικών διαστάσεων n_g [40, 39]. Σε κάθε χώρο αντιστοιχίζεται μία πιθανότητα $P(\Omega_g) = w_g$. Οι πιθανότητες των χώρων του συνολικού δειγματικού χώρου, επαληθεύουν το κριτήριο της στοχαστικότητας(βλ. Εξ. 3.29). Στην περίπτωση που ο επιμέρους χώρος έχει διάσταση $n_g = 0$, θεωρούμε ότι περιέχει μόνο ένα δείγμα. Σε κάθε άλλη περίπτωση όπου $n_g > 0$, ο χώρος έχει συνάρτηση πυκνότητας πιθανότητας μία Γκαουσιανή ιδίων διαστάσεων(βλ. Εξ. 3.30, 3.31).



Σχήμα 3.3: Κατανομή Πιθανότητας σε Πολλαπλών Διαστάσεων Δειγματικούς χώρους [40].

$$\sum_{g=1}^{G} w_g = 1 \tag{3.29}$$

$$n_g > 0 \Rightarrow P(\Omega_g) = \mathcal{N}_g(x), \quad x \in R^{n_g} \tag{3.30}$$

$$n_g = 0 \Rightarrow P(\Omega_g) = 1 \tag{3.31}$$

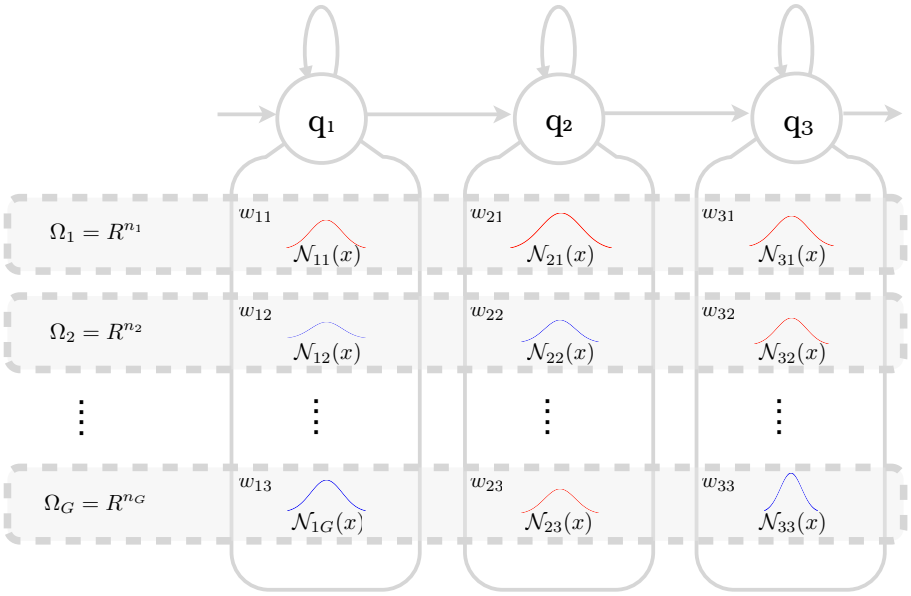
Κάθε παρατήρηση του δειγματικού χώρου περιγράφεται από ένα ζευγάρι παραμέτρων $\mathbf{o}(X, x)$, όπου το X καθορίζει τους επιμέρους χώρους στους οποίους αντιστοιχίζεται η παρατήρηση και το x είναι η τυχαία μεταβλητή με διαστάσεις όμοιες με των επιμέρους χώρων που ανήκει(βλ. Εξ. 3.32 3.33). Στο Σχήμα 3.3 απεικονίζεται η βασική λογική της κατανομής πιθανότητας πολλαπλών χώρων.

$$b(o) = \sum_{g \in S(o)} w_g \mathcal{N}_g(V(o)) \tag{3.32}$$

$$S(o) = X, \quad V(o) = x \tag{3.33}$$

3.2.2 Ορισμός HMM σε Πολλαπλών Διαστάσεων Χώρους

Ένα HMM σε Πολλαπλών Διαστάσεων Χώρους (MSD-HMM) N καταστάσεων λ ορίζεται με πιθανότητες αρχικής κατάστασης $\pi = \{\pi_j\}_{j=1}^{N}$, με πιθανότητες μετάβασης καταστάσεων $A = \{a_{ij}\}_{i,j=1}^{N}$ και με πιθανότητες εξόδου κάθε κατάστασης $B = \{b_i(\mathbf{o})\}_{i=1}^{N}$(βλ. Εξ. 3.34). [40, 39]



Σχήμα 3.4: Απεικόνιση ενός HMM σε Πολλαπλών Διαστάσεων Χώρους [40, 39].

$$b_i(\mathbf{o}) = \sum_{g \in S(\mathbf{o}} w_{ig} \mathcal{N}_{ig}(V(\mathbf{o})) \tag{3.34}$$

Κάθε κατάσταση i έχει G συναρτήσεις κατανομής πιθανότητας και τα αντίστοιχα βάρη αυτών, όπως φαίνεται στο Σχήμα 3.4. Η Πιθανότητα μιας ακολουθίας παρατηρήσεων $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$ δεδομένου του μοντέλου λ γράφεται όπως στην Εξ. 3.35.

$$P(\mathbf{O}|\lambda) = \sum_{all\mathbf{q}} \prod_{t=1}^{T} a_{q_{t-1}q_t} w_{q_t l_t} \mathcal{N}_{q_t l_t}(V(\mathbf{o}_t)) \tag{3.35}$$

Υπολογίζοντας εν συνεχεία με ακριβώς αντίστοιχη διαδικασία τις μεταβλητές των forward και backward αλγορίθμων $\alpha_t(i)$ και $\beta_t(i)$, προκύπτει αντίστοιχα το αποτέλεσμα της Εξ. 3.36.

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} \beta_1(i) \tag{3.36}$$

3.2.3 Ρύθμιση των Παραμέτρων ενός HMM σε Πολλαπλών Διαστάσεων Χώρους

Η διαδικασία της ρύθμισης των παραμέτρων είναι ακριβώς αντίστοιχη με αυτή του απλού HMM, από την άποψη της χρήσης της βοηθητικής συνάρτησης Q (βλ. Εξ. 3.37). Η μεγιστοποίηση της συνάρτησης Q βασίζεται πάνω σε τρία βασικά θεωρήματα τα οποία διατυπώνονται πιο κάτω.

$$Q = (\lambda', \lambda) = \sum_{all\ \mathbf{q},\mathbf{l}} P(\mathbf{O}, \mathbf{q}, \mathbf{l}|\lambda') \log P(\mathbf{O}, \mathbf{q}, \mathbf{l}|\lambda) \tag{3.37}$$

Θεώρημα 1

$Q(\lambda', \lambda) \geq Q(\lambda', \lambda') \rightarrow P(\mathbf{O}, \lambda) \geq P(\mathbf{O}, \lambda')$

Θεώρημα 2

Εάν για κάθε χώρο Ω_g, μεταξύ των $V(\mathbf{o}_1), V(\mathbf{o}_2), ..., V(\mathbf{o}_T)$, υπάρχουν $n_g + 1$ παρατηρήσεις, όπου $g \in S(\mathbf{o}_t)$, κάθε n_g από τις οποίες είναι γραμμικά ανεξάρτητη, η $Q = (\lambda', \lambda)$ έχει μοναδικό ολικό μέγιστο σαν συνάρτηση του λ, όπου και αυτό είναι το μοναδικό κρίσιμο σημείο.

Θεώρημα 3

Ένα σύνολο παραμέτρων λ αποτελεί κρίσιμο σημείο της πιθανοφάνειας $P(\mathbf{O}|\lambda)$ εάν και μόνο εάν είναι κρίσιμο σημείο της συνάρτησης Q.

Έτσι για μία δοθείσα ακολουθία παρατηρήσεων $\mathbf{O}$ και ένα μοντέλο λ', εξάγονται οι παράμετροι του λ οι οποίες μεγιστοποιούν τη συνάρτηση Q. Αναλύοντας το λογάριθμο του γινομένου σε αθροίσματα, η βοηθητική συνάρτηση Q μπορεί να γραφει όπως στην Εξ. 3.38 .

$$
\begin{aligned}
Q(\lambda', \lambda) = &\sum_{i=1}^{N} P(\mathbf{O}, q_1 = i|\lambda') \log \pi_i \\
&+ \sum_{i,j=1}^{N} \sum_{t=1}^{T-1} P(\mathbf{O}, q_t = i, q_{t+1} = j|\lambda') \log a_{ij} \\
&+ \sum_{i=1}^{N} \sum_{g=1}^{G} \sum_{t \in T(\mathbf{O},g)} P(\mathbf{O}, q_t = i, l_t = g|\lambda') \log w_{ij} \\
&+ \sum_{i=1}^{N} \sum_{g=1}^{G} \sum_{t \in T(\mathbf{O},g)} P(\mathbf{O}, q_t = i, l_t = g|\lambda') \log \mathcal{N}_{ij}(V(\mathbf{o}_t))
\end{aligned}
\tag{3.38}
$$

όπου,

$$T(\mathbf{O}, g) = \{t | g \in S(\mathbf{o}_t)\} \tag{3.39}$$

Προκύπτει, λοιπόν το σύνολο των παραμέτρων από τις Εξ. 3.40, 3.41, 3.42, 3.43, 3.44

$$\pi_i = \sum_{g \in S(\mathbf{o}_1} \gamma'_1(i, g) \tag{3.40}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi'_t(i, j)}{\sum_{t=1}^{T-1} \sum_{g \in S(\mathbf{o}_1)} \gamma'_t(i, g)} \tag{3.41}$$

$$w_{ij} = \frac{\sum_{t \in T(\mathbf{o},g)} \gamma'_t(i, g)}{\sum_{h=1}^{G} \sum_{t \in T(\mathbf{o},h)} \gamma'_t(i, h)} \tag{3.42}$$

$$\mu_i g = \frac{\sum_{t \in T(\mathbf{o},g)} \gamma'_t(i, g) V(\mathbf{o}_t)}{\sum_{g \in S(\mathbf{o}_1)} \gamma'_t(i, g)}, \quad n_g > 0 \tag{3.43}$$

$$\Sigma_{ig} = \frac{\sum_{t \in T(\mathbf{o},g)} \gamma'_t(i, g)(V(\mathbf{o}_t) - \mu_{ig}) \cdot (V(\mathbf{o}_t) - \mu_{ig})^T}{\sum_{t \in T(\mathbf{o},g)} \gamma'_t(i, g)} \tag{3.44}$$

Όπου οι μεταβλητές $\gamma_t(i, h)$ και $\xi_t(i, j)$ υπολογίζονται χρησιμοποιώντας τις μεταβλητές των αλγορίθμων forward και backward, όπως παρουσιάζεται στις Εξ. 3.45, 3.46

$$\gamma_t(i, h) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \cdot \frac{w_{ih}\mathcal{N}_{ih}(V(\mathbf{o}_t))}{\sum_{g \in S(\mathbf{o}_t)} w_{ig}\mathcal{N}_{ig}(V(\mathbf{o}_t))} \tag{3.45}$$

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(\mathbf{o}_{t+1})\beta_{t+1}(j)}{\sum_{h,k=1}^{N}\alpha_t(h)a_{hk}b_k(\mathbf{o}_{t+1})\beta_{t+1}(k)} \tag{3.46}$$

3.3 Μοντελοποίηση Διάρκειας Καταστάσεων των HMMs

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. Με τη μέχρι τώρα μορφή των Κρυφών Μαρκοβιανών Μοντέλων, η πυκνότητα πιθανότητας διάρκειας συμπλέκεται στη μοντελοποίηση της φωνής και αποτελεί εκθετική συνάρτηση της πιθανότητας παραμονής στην ίδια κατάσταση όπως φαίνεται στην Εξ. 3.47. Η εκθετική αυτή μοντελοποίηση της διάρκειας καταστάσεων δεν είναι αποτελεσματική.

$$p_i(d) = (a_{ii})^{d-1}(1-a_{ii}) \tag{3.47}$$

Απλά Hidden semi-Markov Models

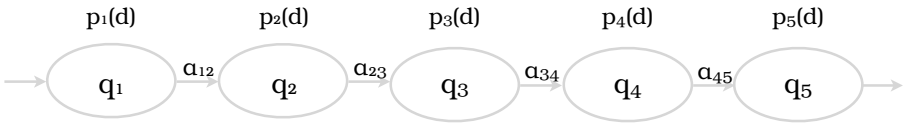
Για την πιο εύρωστη μοντελοποίηση, παρεισφρύει μέσα στο Μαρκοβιανό Μοντέλο μία συνάρτηση πυκνότητας πιθανότητας, που αφορά τη διάρκεια παραμονής σε μία κατάσταση [32]. Έτσι δημιουργούνται τα Hidden semi-Markov Models, τα οποία και παρουσιάζονται στο Σχήμα 3.5.

Συγκεκριμένα, ένα semi-Markov Model:

1. Ξεκινά από την κατάσταση q_1.
2. Η διάρκεια d_1 παραμονής στην κατάσταση q_1 εξάγεται από τη συνάρτηση πυκνότητας πιθανότητας $p_{q_1}(d_1)$.
3. Παρατηρήσεις $\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_{d_1}$ εξάγονται από την πιθανότητα $b_{q_1}(\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_{d_1}) = \prod_{t=1}^{d_1} b_{q_1}(\mathbf{o}_t)$.
4. Η επόμενη κατάσταση $q_2 = j$, προκύπτει από την πιθανότητα $a_{q_1q_2}$, με τον περιορισμό ότι $a_{q_1q_1} = 0$.

Όπως και στα απλά HMMs έτσι και στα HsMMs ορίζονται οι forward/backward μεταβλητές(βλ. Εξ. 3.48, 3.49).

$$\alpha_t(i) = P(\mathbf{o}_1\ \mathbf{o}_2\ ...\mathbf{o}_t, stay\ in\ q_i\ ends\ at\ t|\lambda) \tag{3.48}$$



Σχήμα 3.5: Απεικόνιση ενός semi-HMM [32].

$$\beta_t(i) = P(\mathbf{o}_{t+1}\ \mathbf{o}_{t+2}\ ...\mathbf{o}_T|stay\ in\ q_i\ ends\ at\ t, \lambda) \tag{3.49}$$

Η μεταβλητή $\alpha_t(i)$ γράφεται αναλυτικά όπως στην Εξ. 3.50.

$$\alpha_t(i) = \sum_{i=1}^{N}\sum_{d=1}^{D}\alpha_{t-d}(i)a_{ij}p_j(d)\prod_{s=t-d+1}^{t}b_j(\mathbf{o}_s) \tag{3.50}$$

οπότε και προκύπτει,

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N}\alpha_t(i)a_{ij} \tag{3.51}$$

Παράλληλα με τις μεταβλητές των Εξ. 3.48 και 3.49, ορίζονται και άλλες δύο βοηθητικές μεταβλητές, όπως παρουσιάζονται στις Εξ. 3.52, 3.53.

$$\alpha_t^*(i) = P(\mathbf{o}_1\ \mathbf{o}_2\ ...\mathbf{o}_t, stay\ at\ q_i\ starts\ at\ t+1|\lambda) \tag{3.52}$$

$$\beta_t^*(i) = P(\mathbf{o}_{t+1}\ \mathbf{o}_{t+2}\ ...\mathbf{o}_T|stay\ at\ q_i\ starts\ at\ t+1, \lambda) \tag{3.53}$$

όπου,

$$\alpha_t^*(i) = \sum_{i=1}^{N}\alpha_t(i)a_{ij} \tag{3.54}$$

$$\beta_t(i) = \sum_{j=1}^{N}a_{ij}\beta_t^*(j) \tag{3.55}$$

$$\alpha_t(i) = \sum_{d=1}^{D}\alpha_{t-d}^*(i)p_i(d)\prod_{s=t-d+1}^{t}b_i(\mathbf{o}_s) \tag{3.56}$$

$$\beta_t^*(i) = \sum_{d=1}^{D}\beta_{t+d}(i)p_i(d)\prod_{s=t+1}^{t+d}b_i(\mathbf{o}_s) \tag{3.57}$$

Με τον αλγόριθμο ρύθμισης προκύπτουν οι παράμετροι όπως φαίνεται και στις Εξ. 3.58, 3.59, 3.60, 3.61.

$$\bar{\pi}_i = \frac{\pi_i \beta_0^*(i)}{P(\mathbf{O}|\lambda)} \tag{3.58}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T} \alpha_t(i) a_{ij} \beta_t^*(j)}{\sum_{j=1}^{N} \sum_{t=1}^{T} \alpha_t(i) a_{ij} \beta_t^*(j)} \tag{3.59}$$

$$\bar{b}_i(k) = \frac{\sum_{t=1}^{T} [\sum_{\tau<t} \alpha_\tau^*(i) \cdot \beta_\tau^*(i) - \sum_{\tau<t} \alpha_\tau(i) \cdot \beta_\tau(i)]}{\sum_{k=1}^{M} \sum_{t=1}^{T} [\sum_{\tau<t} \alpha_\tau^*(i) \cdot \beta_\tau^*(i) - \sum_{\tau<t} \alpha_\tau(i) \cdot \beta_\tau(i)]} \tag{3.60}$$

$$\bar{p}_i(d) = \frac{\sum_{t=1}^{T} \alpha_t^*(i) p_i(d) \beta_{t+d}(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{o}_s)}{\sum_{d=1}^{D} \sum_{t=1}^{T} \alpha_t^*(i) p_i(d) \beta_{t+d}(i) \prod_{s=t+1}^{t+d} b_i(\mathbf{o}_s)} \tag{3.61}$$

Για την μείωση της πολυπλοκότητας των υπολογισμών που έχει αυτό το μοντέλο, έχουν προταθεί και μοντέλα για τη διάρκεια της κατάστασης του HMM, με συνάρτηση πυκνότητας πιθανότητας που ακολουθεί κάποια κατανομή παραμετροποιημένη. Ένα τέτοιο μοντέλο προτάθηκε από τον Levinson [20]. Σύμφωνα με αυτό το μοντέλο, η διάρκεια παραμονής σε κάθε κατάσταση ακολουθεί την κατανομή γάμα, όπως παρουσιάζεται στην Εξ. 3.62. Με τη συγκεκριμένη μοντελοποίηση, ο αλγόριθμος ρύθμισης των παραμέτρων έχει πολύ μικρότερο υπολογιστικό κόστος από τον προηγούμενο, μιας και η κατανομή της διάρκειας καθορίζεται από τις παραμέτρους της κατανομής γάμα.

$$d_i(\tau) = \frac{\eta_i^{v_i}}{\Gamma(v_i)} \tau^{v_i - 1} e^{-\eta_i \tau}, \quad \tau > 0 \tag{3.62}$$

Μοντελοποίηση Συστήματος Παραγωγής Φωνής από Κείμενο με Hidden semi-Markov Models

Όσον αφορά την σύνθεση φωνής με Κρυφά Μαρκοβιανά Μοντέλα, έχει προταθεί η μοντελοποίηση της διάρκειας καταστάσεων ταυτόχρονα με τη μοντελοποίηση των Μοντέλων των χαρακτηριστικών [48, 53]. Η διάρκεια παραμονής σε κάθε κατάσταση μοντελοποιείται με κατανομή απλής Γκαουσιανής. Ο υπολογισμός των παραμέτρων αυτής της Γκαουσιανής γίνεται όπως φαίνεται και στις Εξ. 3.63, 3.64, από τις forward/backward μεταβλητές κατά την τελευταία επανάληψη του αντίστοιχου αλγορίθμου, μέσω του υπολογισμού της πιθανότητας παραμονής στην κατάσταση j από τη χρονική στιγμή t_0 μέχρι την t_1 $\chi_{t_0,t_1}(j)$(Εξ. 3.65).

$$\xi_j = \frac{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(j)\cdot(t_1-t_0+1)}{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(j)} \tag{3.63}$$

$$\sigma_j = \frac{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(j)\cdot(t_1-t_0+1)^2}{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(j)} - \xi_j^2 \tag{3.64}$$

$$\chi_{t_0,t_1}(j) = \frac{[\{\sum_{i\neq j}\alpha_{t_0-1}(i)a_{ij}\}\cdot\prod_{s=t_0}^{t_1}b_j(\mathbf{o}_s)\cdot a_{jj}^{t_1-t_0}\{\sum_{k\neq j}a_{jk}b_k(\mathbf{o}_{t_1+1})\beta_{t_1+1}(k)\}]}{P(\mathbf{o}|\lambda)} \tag{3.65}$$

Μοντελοποίηση με Hidden semi-Markov Models σε Πολλαπλών Διαστάσεων Χώρους

Για τη μοντελοποίηση της ακολουθίας του pitch, όπως έχει προαναφερθεί, γίνεται χρήση των Κρυφών Μαρκοβιανών Μοντέλων σε Πολλαπλούς Χώρους (MSD-HMMs). Συνεπώς η προηγούμενη θεωρία της μοντελοποίησης της διάρκειας καταστάσεων γενικεύεται και επεκτείνεται και στα Κρυφά Μαρκοβιανά Μοντέλα σε χώρους πολλαπλών διαστάσεων [53]. Δεδομένου ότι ισχύει η Εξ. 3.34, από τη γενίκευση του αλγορίθμου ρύθμισης των παραμέτρων προκύπτουν οι Εξ. 3.66, 3.67, 3.68, 3.69, 3.70, 3.71.

$$\bar{w}_{jg} = \frac{\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(j,g)}{\sum_{t=h}^{G}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(j,h)} \tag{3.66}$$

$$\bar{\mu}_{jg} = \frac{\sum_{t=1}^{T}\sum_{d=1}^{t}\zeta_t^d(j,g)}{\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(j,g)}, \quad n_g > 0 \tag{3.67}$$

$$\bar{\mathbf{\Sigma}}_{jg} = \frac{\sum_{t=1}^{T}\sum_{d=1}^{t}\eta_t^d(j,g)}{\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(j,g)}, \quad n_g > 0 \tag{3.68}$$

όπου,

$$\begin{aligned}\gamma_t^d(j,g) = &\sum_{i=1,i\neq j}^{N'}\alpha'_{t-d}(i)a'_{ij}p'_j(d)\beta_t(j)\\ \cdot &\sum_{s=t-d+1}^{t}w_{jg}\mathcal{N}(\mathbf{V}(\mathbf{o}_s)|\mu_{jg},\mathbf{\Sigma}_j g)\prod_{k=t-d+1,k\neq s}^{t}b'_j(\mathbf{o}_k)\end{aligned} \tag{3.69}$$

$$\zeta_t^d(j,g) = \sum_{i=1,i\neq j}^{N'} \alpha'_{t-d}(i)a'_{ij}p'_j(d)\beta_t(j) \cdot \sum_{s=t-d+1}^{t} w_{jg}\mathcal{N}(\mathbf{V}(\mathbf{o}_s)|\mu_{jg},\mathbf{\Sigma}_j g) \prod_{k=t-d+1,k\neq s}^{t} b'_j(\mathbf{o}_k)\mathbf{V}(\mathbf{o}_s) \tag{3.70}$$

$$\eta_t^d(j,g) = \sum_{i=1,i\neq j}^{N'} \alpha'_{t-d}(i)a'_{ij}p'_j(d)\beta_t(j) \cdot \sum_{s=t-d+1}^{t} w_{jg}\mathcal{N}(V(\mathbf{o}_s)|\mu_{jg},\mathbf{\Sigma}_j g) \prod_{k=t-d+1,k\neq s}^{t} b'_j(\mathbf{o}_k) [\mathbf{V}(\mathbf{o}_s)-\mu_{jg}][\mathbf{V}(\mathbf{o}_s)-\mu_{jg}]^{\mathsf{T}} \tag{3.71}$$

Κεφάλαιο 4

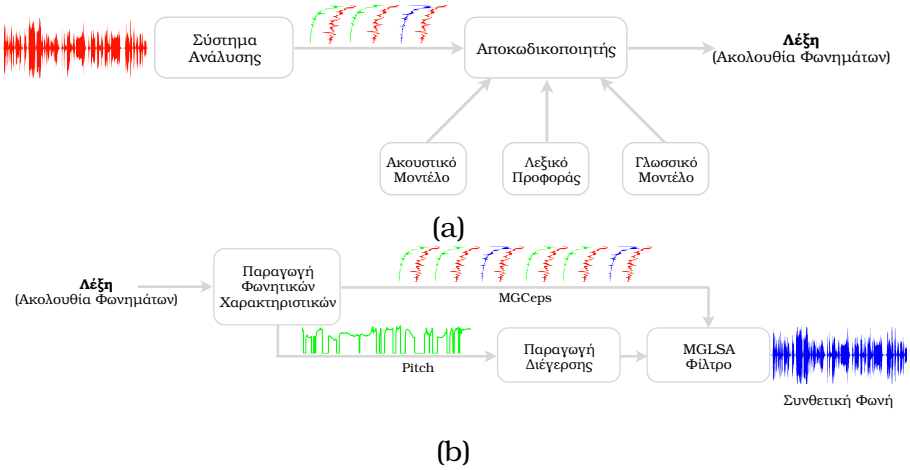
Παραγωγή των Χαρακτηριστικών από τα HMMs

Όπως έχει προαναφερθεί, η λογική ενός συνθέτη φωνής από κείμενο, που βασίζεται στη μοντελοποίηση με Κρυφά Μαρκοβιανά Μοντέλα, αποτελεί κατά βάση μία αντιστροφή του Συστήματος Αναγνώρισης Φωνής. Συνεπώς, είναι εμφανές ότι το σύστημα των HMMs στην περίπτωση της αναγνώρισης δέχεται ως είσοδο ένα φωνητικό σήμα που στη συνέχεια μετατρέπεται σε μία ακολουθία διανυσμάτων χαρακτηριστικών, ώστε να εξαχθεί η ακολουθία των φωνημάτων που αναγνωρίζονται, ενώ στο συγκεκριμένο σύστημα είσοδος είναι μία ακολουθία φωνημάτων και έξοδος μία ακολουθία φωνητικών χαρακτηριστικών(βλ. Σχήμα 4.1). Για την ολοκληρωμένη λειτουργία του συστήματος παραγωγής φωνής από κείμενο, θα πρέπει να χρησιμοποιηθεί κάποιος αλγόριθμος που να μοντελοποιεί το σύστημα μετατροπής της ακολουθίας παρατηρήσεων σε ακολουθία φωνητικών χαρακτηριστικών.

4.1 Παραγωγή Ακολουθίας Φωνητικών Χαρακτηριστικών

Σε αυτό το σημείο, χρησιμοποιείται ο αλγόριθμος παραγωγής φωνητικών χαρακτηριστικών [41, 37], με βάση το κριτήριο μέγιστης πιθανοφάνειας. Συγκεκριμένα, δίνεται ένα HMM παραμέτρων λ, ώστε να εξαχθεί μία ακολουθία διανυσμάτων φωνητικών χαρακτηριστικών $\mathbf{O}$, από τη μεγιστοποίηση της πιθανοφάνειας $P(\mathbf{O}, \lambda)$ ως προς το διάνυσμα παρατηρήσεων.Όπου,

$$\mathbf{O} = [\mathbf{o}_1^\mathsf{T}, \mathbf{o}_2^\mathsf{T}, ..., \mathbf{o}_T^\mathsf{T}]^\mathsf{T} \tag{4.1}$$



Σχήμα 4.1: Λειτουργικό Διάγραμμα (a) ενός απλού συστήματος αναγνώρισης φωνής σε σύγκριση με (b) ενός συστήματος παραγωγής φωνής από κείμενο.

$$P(\mathbf{O}, \lambda) = \sum_{all\mathbf{Q}} P(\mathbf{O}, \mathbf{Q}, \lambda) \tag{4.2}$$

όπου

$$\mathbf{Q} = \{(q_1, i_1), (q_2, i_2), ..., (q_T, i_T)\} \tag{4.3}$$

αποτελεί την ακολουθία καταστάσεων και μίγματος Γκαουσιανών. Μία παρατήρηση $\mathbf{o}_t$ αποτελείται από τα διανύσματα των στατικών και των δυναμικών χαρακτηριστικών, όπως φαίνεται και στην Εξ. 4.4.

$$\mathbf{o}_T = [\mathbf{c}_t^{\mathsf{T}}, \Delta\mathbf{c}_t^{\mathsf{T}}, \Delta^2\mathbf{c}_t^{\mathsf{T}}]^{\mathsf{T}} \tag{4.4}$$

όπου,

$$\Delta\mathbf{c}_t = \sum_{\tau=-L^{(1)-}}^{L^{(1)+}} w^{(1)}(\tau)\, \mathbf{c}_{t+\tau} \tag{4.5}$$

$$\Delta^2\mathbf{c}_t = \sum_{\tau=-L^{(2)-}}^{L^{(2)+}} w^{(2)}(\tau)\, \mathbf{c}_{t+\tau} \tag{4.6}$$

Τρία είναι τα Προβλήματα που διατυπώνονται σε αυτήν την περίπτωση:

Πρόβλημα 1. Για δοθέντα HMM παραμέτρων λ και ακολουθίας $\mathbf{Q}$, μεγιστοποιείται η $P(\mathbf{O}|\mathbf{Q}, \lambda)$ ως προς την ακολουθία παρατηρήσεων $\mathbf{O}$ και ταυτόχρονα επαληθεύοντας τις Εξ. 4.5, 4.6.

Πρόβλημα 2. Για δοθέν HMM παραμέτρων λ, μεγιστοποιείται η $P(\mathbf{O}|\mathbf{Q},\lambda)$ ως προς την ακολουθία παρατηρήσεων $\mathbf{O}$ και ακολουθίας $\mathbf{Q}$ και ταυτόχρονα επαληθεύοντας τις Εξ. 4.5, 4.6.

Πρόβλημα 3. Για δοθέν HMM παραμέτρων λ, μεγιστοποιείται η $P(\mathbf{O}|\lambda)$ ως προς την ακολουθία παρατηρήσεων $\mathbf{O}$ και ταυτόχρονα επαληθεύοντας τις Εξ. 4.5, 4.6.

4.1.1 Πρόβλημα 1 - Μεγιστοποίηση του $P(\mathbf{O}|\mathbf{Q},\lambda)$ ως προς την Ακολουθία Παρατηρήσεων O

Αρχικά, μεγιστοποιείται η πιθανοφάνεια $P(\mathbf{O}|\mathbf{Q},\lambda)$ ως προς την ακολουθία παρατηρήσεων $\mathbf{O}$, θεωρώντας την ακολουθία $\mathbf{Q}$ σταθερή. Ο λογάριθμος της πιθανοφάνειας, λοιπόν, γράφεται στην Εξ. 4.7.

$$\log P(\mathbf{O}|,\mathbf{Q},\lambda) = -\frac{1}{2}\mathbf{O}^{\mathsf{T}}\mathbf{U}^{-1}\mathbf{O} + \mathbf{O}^{\mathsf{T}}\mathbf{U}^{-1}\mathbf{M} + K \tag{4.7}$$

όπου,

$$\mathbf{U}^{-1} = \mathrm{diag}[\mathbf{U}^{-1}_{q_1,i^1}, \mathbf{U}^{-1}_{q_2,i^2}, ..., \mathbf{U}^{-1}_{q_T,i^T}] \tag{4.8}$$

$$\mathbf{M} = [\mu^{\mathsf{T}}_{q_1,i_1}, \mu^{\mathsf{T}}_{q_2,i_2}, ..., \mu^{\mathsf{T}}_{q_T,i_T}]^{\mathsf{T}} \tag{4.9}$$

όπου μ_{q_t,i_t} είναι το διάνυσμα $3M \times 1$ της μέσης τιμής και U_{q_t,i_t} είναι η μήτρα συμμεταβλητότητας διάστασης $3M \times 3M$, που σχετίζονται με το $i-th$ μίγμα Γκαουσιανών της κατάστασης q_t. Η σταθερά K είναι ανεξάρτητη της ακολουθίας παρατηρήσεων $\mathbf{O}$. Είναι εμφανές ότι η $P(\mathbf{O}|\mathbf{Q},\lambda)$ μεγιστοποιείται όταν $\mathbf{O} = \mathbf{M}$. Οι Εξ. 4.5 και 4.6, μπορούν να γραφούν διανυσματικά όπως στην Εξ. 4.10:

$$\mathbf{O} = \mathbf{WC} \tag{4.10}$$

όπου

$$\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_T]^{\mathsf{T}} \tag{4.11}$$

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_T]^{\mathsf{T}} \tag{4.12}$$

$$\mathbf{w}_t = [\mathbf{w}^{(0)}_t, \mathbf{w}^{(1)}_t, \mathbf{w}^{(2)}_t] \tag{4.13}$$

$$\begin{aligned}\mathbf{w}_t^{(n)} = [&\mathbf{0}_{M\times M}, ..., \mathbf{0}_{M\times M}, w^{(n)}(-L_-^{(n)})\mathbf{I}_{M\times M},\\ &..., w^{(n)}(0)\mathbf{I}_{M\times M}, ..., w^{(n)}(L_+^{(n)})\mathbf{I}_{M\times M},\\ &\mathbf{0}_{M\times M}, ..., \mathbf{0}_{M\times M}]^{\mathsf{T}}, \quad n = 0, 1, 2\end{aligned} \tag{4.14}$$

όπου $L_-^{(0)} = L_+^{(0)} = 0$ και $w^{(0)}(0) = 1$. Έτσι με βάση την Εξ. 4.10, η μεγιστοποίηση της πιθανοφάνειας $P(\mathbf{O}|\mathbf{Q}, \lambda)$ ως προς το $\mathbf{O}$ ισοδυναμεί με τη μεγιστοποίηση ως προς το $\mathbf{C}$. Συνεπώς θέτοντας

$$\frac{\partial \log P(\mathbf{WC}|\mathbf{Q}, \lambda)}{\partial \mathbf{C}} = \mathbf{0} \tag{4.15}$$

προκύπτει το σύνολο των εξισώσεων,

$$\mathbf{W}^{\mathsf{T}}\mathbf{U}^{-1}\mathbf{WC} = \mathbf{W}^{\mathsf{T}}\mathbf{U}^{-1}\mathbf{M}^{\mathsf{T}} \tag{4.16}$$

Για την άμεση επίλυση της Εξ. 4.16, χρειάζονται $O(T^3M^3)$ υπολογισμοί, διότι η μήτρα $\mathbf{W}^{\mathsf{T}}\mathbf{U}^{-1}\mathbf{W}$ είναι διάστασης $TM \times TM$. Έτσι χρησιμοποιώντας την ειδική δομή της $\mathbf{W}^{\mathsf{T}}\mathbf{U}^{-1}\mathbf{W}$, η Εξ. 4.16 μπορεί να επιλυθεί με την Μέθοδο Αποσύνθεσης του Cholesky ή με την αποσύνθεση QR με υπολογιστικό κόστος ίσο με $O(TM^3L^2)$, όπου $L = max_{n\in\{1,2\},s\in\{-,+\}}L_s^{(n)}$. Οι Tokuda et. al. [37] προτείνουν ένα αλγόριθμο επίλυσης του πιο πάνω προβλήματος με υπολογιστικό κόστος $O(T^2M^3)$.

4.1.2 Πρόβλημα 2 - Μεγιστοποίηση του $P(\mathrm{O}, \mathrm{Q}|\lambda)$ ως προς την Ακολουθία Παρατηρήσεων O και την Ακολουθία καταστάσεων Q

Το συγκεκριμένο πρόβλημα μπορεί να επιλυθεί υπολογίζοντας τη μέγιστη πιθανότητα $max_C P(\mathbf{O}, \mathbf{Q}|\lambda) = max_C P(\mathbf{O}|\mathbf{Q}, \lambda)P(\mathbf{Q}|\lambda)$ για όλες τις πιθανές ακολουθίες $\mathbf{Q}$. Παρόλα αυτά είναι μη πρακτικό να υπολογιστεί άμεσα μιας και οι πιθανές ακολουθίες $\mathbf{Q}$ είναι πάρα πολλές. Για αυτό το λόγο χρησιμοποιείται ένας αρκετά γρήγορος αλγόριθμος που αναζητά τη βέλτιστη ακολουθία καταστάσεων διατηρώντας βέλτιστο το $\mathbf{C}$, υπό την έννοια ότι η $P(\mathbf{O}|\mathbf{Q}, \lambda)$ μεγιστοποιείται ως προς το $\mathbf{C}$ [41, 37].

Εισαγωγή των Μοντέλων της Διάρκειας Καταστάσεων

Για να ελεγχθεί η χρονική εξέλιξη και η διάρκεια των παραγόμενων παραμέτρων του ήχου, θα πρέπει να συνυπολογιστούν και οι κατανομές της διάρκειας καταστάσεων. Η $P(\mathbf{O}, \mathbf{Q}|\lambda)$ μπορεί να γραφεί ως εξής:

$$P(\mathbf{O}, \mathbf{Q}|\lambda) = P(\mathbf{O}, \mathbf{i}|\mathbf{q}, \lambda)P(\mathbf{q}|\lambda) \tag{4.17}$$

όπου $\mathbf{q} = \{q_1, q_2, ..., q_T\}, \mathbf{i} = \{i_1, i_2, ..., i_T\}$, όπου η πιθανότητα διάρκειας μιας κατάστασης γράφεται όπως στην Εξ. 4.18

$$\log P(\mathbf{q}|\lambda) = \sum_{n=1}^{N} P \log p_{q_n}(d_{q_n}) \tag{4.18}$$

όπου σε T χρονικά πλαίσια έχει επισκεφτεί N καταστάσεις, και $p_{q_n}(d_{q_n})$ αποτελεί την πιθανότητα να προκύψουν d_{q_n} παρατηρήσεις στην κατάσταση q_n. Εάν καθορίσουμε την ακολουθία καταστάσεων $\mathbf{q}$ από την $P(\mathbf{q}|\lambda)$ ανεξάρτητα από την $\mathbf{O}$, η μεγιστοποίηση της Εξ. 4.17 ως προς την ακολουθία $\mathbf{O}$ και $\mathbf{Q}$ είναι ισοδύναμη με τη μεγιστοποίηση της $P(\mathbf{O}, \mathbf{i}|\mathbf{q}, \lambda)$ ως προς $\mathbf{O}$ και $\mathbf{i}$. Η λύση της εξίσωσης προκύπτει από της επίλυση της Εξ. 4.16 με την προηγούμενη μέθοδο.

4.1.3 Πρόβλημα 3 - Μεγιστοποίηση του $P(\mathrm{O}|\lambda)$ ως προς την Ακολουθία Παρατηρήσεων O

Σε αυτό το πρόβλημα χρησιμοποιείται ένας αλγόριθμος που βασίζεται στον ΕΜ, ο οποίος εντοπίζει ένα κρίσιμο σημείο της συνάρτησης πιθανοφάνειας $P(\mathbf{O}|\lambda)$. Χρησιμοποιούμε μία βοηθητική συνάρτηση Q της τρέχουσας ακολουθίας παρατηρήσεων $\mathbf{O}$ και της νέας $\mathbf{O}'$ που παρουσιάζεται στην Εξ. 4.19 [41] .

$$Q(\mathbf{O}, \mathbf{O}') = \sum_{all\mathbf{Q}} P(\mathbf{O}, \mathbf{Q}|\lambda) \log P(\mathbf{O}', \mathbf{Q}|\lambda)) \tag{4.19}$$

Μπορεί να αποδειχτεί ότι αντικαθιστώντας το $\mathbf{O}$ με το $\mathbf{O}'$ που μεγιστοποιεί το $Q(\mathbf{O}, \mathbf{O}')$, η πιθανοφάνεια αυξάνεται εκτός εάν το $\mathbf{O}'$ είναι κρίσιμο σημείο αυτής. Η Εξ. 4.19 μπορεί να γραφτεί ως εξής,

$$Q(\mathbf{O}, \mathbf{O}') = P(\mathbf{O}, \lambda)\{-\frac{1}{2}\mathbf{O}'^{\mathsf{T}}\overline{\mathbf{U}^{-1}}\mathbf{O}' + \mathbf{O}'^{\mathsf{T}}\overline{\mathbf{U}^{-1}\mathbf{M}} + \overline{K}\} \tag{4.20}$$

όπου

$$\overline{\mathbf{U}^{-1}} = \mathrm{diag}[\overline{\mathbf{U}_1^{-1}}, \overline{\mathbf{U}_2^{-1}}, ..., \overline{\mathbf{U}_T^{-1}}] \tag{4.21}$$

$$\overline{\mathbf{U}_t^{-1}} = \sum_{q,i} \gamma_t(q, i) U_{q,i}^{-1} \tag{4.22}$$

$$\overline{\mathbf{U}^{-1}M} = [\overline{\mathbf{U}_1^{-1}\mu_1}^{\mathsf{T}}, \overline{\mathbf{U}_2^{-1}\mu_2}^{\mathsf{T}}, ..., \overline{\mathbf{U}_T^{-1}\mu_T}^{\mathsf{T}}]^{\mathsf{T}} \tag{4.23}$$

$$\overline{\mathbf{U}_t^{-1}\mu_t} = \sum_{q,i} \gamma_t(q,i) U_{q,i}^{-1} \mu_{q,i} \tag{4.24}$$

και η σταθερά $\overline{K}$ είναι ανεξάρτητη του $\mathbf{O}'$. Η πιθανότητα παραμονής $\gamma_t(q,i)$ δίνεται από τον τύπο της Εξ. 4.25 και μπορεί να υπολογιστεί από τον αλγόριθμο forward/backward.

$$\gamma_t(q,i) = P(q_t = (q,i)|\mathbf{O}, \lambda) \tag{4.25}$$

Πάντοτε ισχύει η $\mathbf{O}' = \mathbf{W}\mathbf{C}'$, όπου η $\mathbf{C}'$ που μεγιστοποιεί την $Q(\mathbf{O}, \mathbf{O}')$ δίνεται από το σύνολο εξισώσεων που απεικονίζεται στην Εξ. 4.26.

$$\mathbf{W}^{\mathsf{T}}\overline{\mathbf{U}^{-1}}\mathbf{W}\mathbf{C}' = \mathbf{W}^{\mathsf{T}}\overline{\mathbf{U}^{-1}\mathbf{M}} \tag{4.26}$$

Το παραπάνω σύστημα εξισώσεων έχει τη μορφή της Εξ. 4.16, οπότε λύνεται και με τον αντίστοιχο αλγόριθμο. Συνοπτικά, όλη η αλγοριθμική διαδικασία ρύθμισης των παραμέτρων παρουσιάζεται σε 4 βήματα πιο κάτω.

Βήμα1 Επιλέγεται ένα αρχικό διάνυσμα $\mathbf{C}$

Βήμα2 Υπολογίζεται η $\gamma_t(q,i)$ με τον αλγόριθμο forward/backward

Βήμα3 Υπολογίζονται οι μητρικές μορφές $\overline{\mathbf{U}^{-1}}$ και $\overline{\mathbf{U}^{-1}M}$ από τις 4.21-4.24, ώστε να επιλυθεί η Εξ. 4.26.

Βήμα4 Τίθεται το $\mathbf{C} = \mathbf{C}'$. Εάν κάποιο κριτήριο σύγκλισης επαληθεύεται η διαδικασία τελειώνει, αλλιώς δρομολογείται στο **Βήμα 2.**

Η εισαγωγή των μοντέλων διάρκειας καταστάσεων παρεισφρύουν στις εξισώσεις με την ίδια λογική, όπως και στην περίπτωση της επίλυσης του Προβλήματος 2.

Κεφάλαιο 5

Συνθέτης Φωνής από Κείμενο με HMMs

Με βάσει τις θεωρητικές αναλύσεις των επιμέρους θεμάτων όλων των προηγούμενων κεφαλαίων, μπορεί να υλοποιηθεί ένας Συνθέτης Φωνής από Κείμενο με Κρυφά Μαρκοβιανά Μοντέλα. Το λειτουργικό διάγραμμα ενός τέτοιου συστήματος παρουσιάζεται στο Σχήμα 5. Είναι εμφανές ότι το συνολικό σύστημα αποτελείται από 4 επιμέρους πολύ βασικά τμήματα:

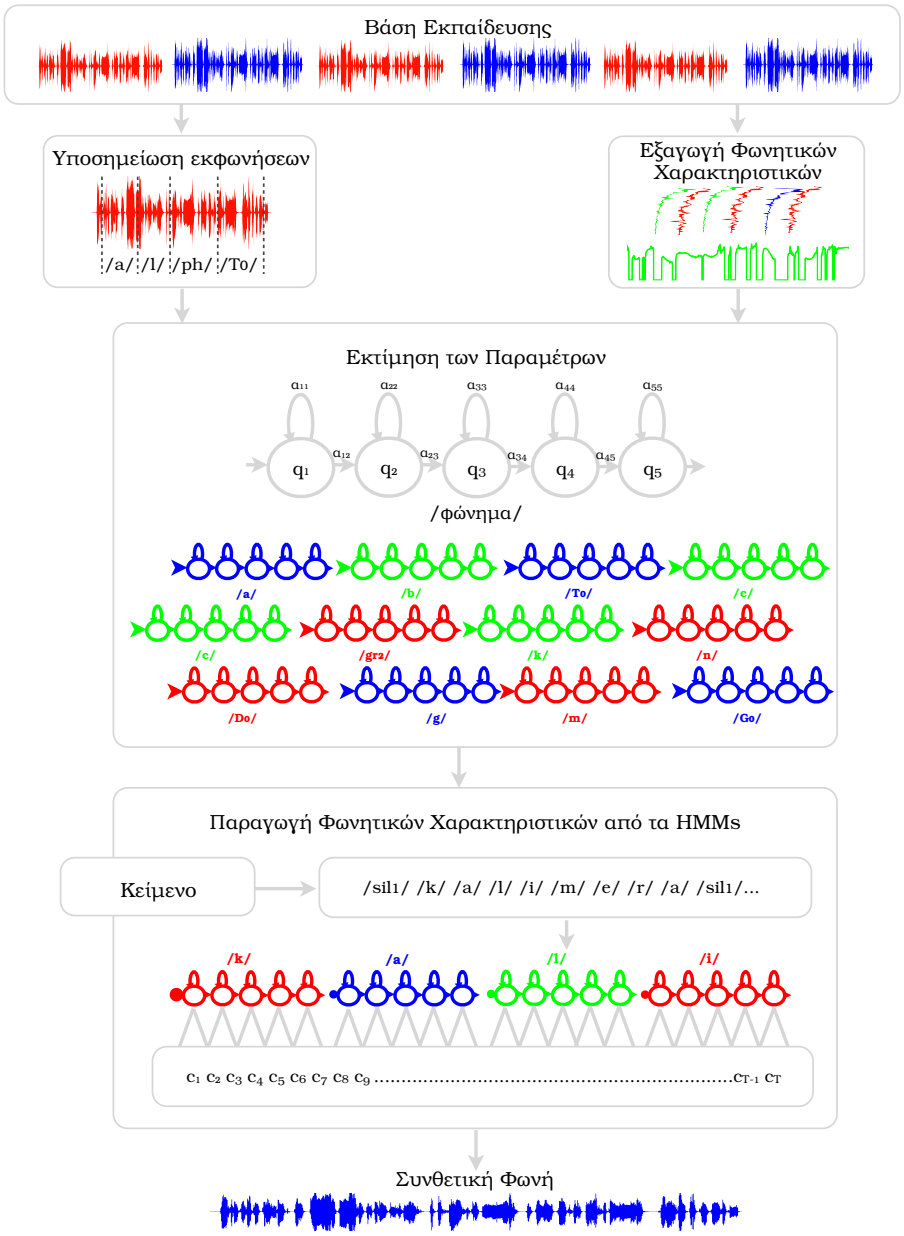
1. Την Ανάλυση της Εκφωνήσεων της Βάσης Εκπαίδευσης
2. Την Εκτίμηση και Εκπαίδευση των Παραμέτρων των Κρυφών Μαρκοβιανών Μοντέλων
3. Την παραγωγή των φωνητικών χαρακτηριστικών από τα HMMs
4. Τη Σύνθεση της Φωνής από το Διάνυσμα των Χαρακτηριστικών

5.1 Ανάλυση της Βάσης Εκπαίδευσης του Συστήματος

Για την εκπαίδευση του συστήματος θα πρέπει να εκτελεστούν δύο βασικές διαδικασίες. Αυτές αφορούν την επεξεργασία της βάσης δεδομένων τόσο σε λεκτικό όσο και σε επίπεδο φωνητικής επεξεργασίας.

Υποσημείωση Εκφωνήσεων

Η βάση εκφωνήσεων αποτελείται από ένα σύνολο ηχογραφημένων προτάσεων. Για την καλύτερη επεξεργασία αυτών, είναι σημαντικό να δημιουργηθούν και τα αρχεία που περιέχουν την ακριβή λεκτική αντιστοίχιση των



Σχήμα 5.1: Λειτουργικό Διάγραμμα ενός Συνθέτη Φωνής από Κείμενο με Κρυφά Μαρκοβιανά Μοντέλα

φωνητικών σημάτων σε επίπεδο φωνημάτων. Κάθε πειραματική βάση εκφωνήσεων συνοδεύεται και από την αντίστοιχη βάση των ετικέτων (Transcription files).

Εξαγωγή Φωνητικών Χαρακτηριστικών

Όπως έχει περιγραφεί και στο Κεφάλαιο 2, ένα από τα πιο βασικά στοιχεία υλοποίησης του Συνθέτη, είναι η επιλογή των κατάλληλων φωνητικών χαρακτηριστικών. Αυτά τα φωνητικά χαρακτηριστικά θα χρησιμοποιηθούν στη συνέχεια για τη μοντελοποίηση του συστήματος με Κρυφά Μαρκοβιανά Μοντέλα. Για αυτό το λόγο η επιλογή αυτών είναι πάρα πολύ κρίσιμη όσον αφορά την απόδοση του Συνθέτη.

5.2 Μοντελοποίηση των HMMs

Για τη δημιουργία ενός συνθέτη φωνής από κείμενο με HMMs, είναι απαραίτητη η εκπαίδευση του συστήματος ώστε να δημιουργηθούν τα μοντέλα των φωνημάτων. Το τμήμα αυτό του συνθέτη είναι ακριβώς αντίστοιχο με εκείνο της εκπαίδευσης ενός συστήματος αναγνώρισης φωνής. Όλο το θεωρητικό υπόβαθρο για αυτή τη μοντελοποίηση περιγράφεται αναλυτικά στο Κεφάλαιο 3.

5.3 Παραγωγή των Φωνητικών Χαρακτηριστικών από τα HMMs

Σε αυτό το τμήμα του Συνθέτη, χρησιμοποιούνται τα HMMs, ώστε να παραχθούν τα χαρακτηριστικά της συνθετικής φωνής. Όπως και το πρώτο τμήμα έτσι και αυτό διαχωρίζεται στη φωνηματική και στη φωνητική επεξεργασία.

Εισαγωγή Κειμένου

Όπως είναι φανερό, το σύστημα δέχεται ως είσοδο κείμενο και εξάγει ήχο. Συνεπώς, το κείμενο που δέχεται ως είσοδο το μετατρέπει σε ακολουθία φωνημάτων και κατ' επέκταση σε ακολουθία τριφώνων. Έτσι δημιουρείται μία λεκτική πρόταση αποτελούμενη από τρίφωνα.

Υπολογισμός των Χαρακτηριστικών

Στη συνέχεια αυτή η λεκτική πρόταση τριφώνων μετατρέπεται σε ενιαία αλυσίδα των Κρυφών Μαρκοβιανών Μοντέλων των επιμέρους τριφώνων. Για

αυτή την αλυσίδα εφαρμόζεται όλη η θεωρία που παρουσιάζεται στο Κεφάλαιο 4, οπότε και προκύπτει ηα ακολουθία των αντίστοιχων φωνητικών χαρακτηριστικών.

5.4 Εξαγωγή Συνθετικής Φωνής

Στο σημείο αυτό, εφαρμόζεται ολοκληρωτικά πλέον η θεωρία της Ανάλυσης/Σύνθεσης, που παρουσιάστηκε στο Κεφάλαιο 2. Συγκεκριμένα, εφαρμόζεται η θεωρία του τμήματος Σύνθεσης των Vocoders, για τα αντίστοιχα χαρακτηριστικά με τα οποία έχει μοντελοποιηθεί το σύστημα εξαρχής.

Κεφάλαιο 6

Υλοποίηση Συνθέτη Φωνής από Κείμενο

Μετά από αυτή τη θεωρητική μελέτη και ως αντικείμενο της συγκεκριμένης διπλωματικής, μελετήθηκε και υλοποιήθηκε εξ' αρχής ένας Συνθέτης φωνής από κείμενο στην ελληνική γλώσσα.

6.1 Επεξεργασία Βάσης Εκφωνήσεων

Για τη συγκεκριμένη διπλωματική χρησιμοποιήθηκε μία βάση που παραχωρήθηκε από το ΙΕΛ, η οποία περιέχει 117 εκφωνήσεις δελτίων καιρού. Πιο συγκεκριμένα, η βάση έχει διάρκεια εκφωνήσεων 1303 δευτερολέπτων, δηλαδή περίπου 21 λεπτών της ώρας. Η φωνή είναι γυναικεία με συχνότητα pitch που κυμαίνεται από τα 100Hz μέχρι και τα 350Hz.

6.1.1 Γλωσσολογική Ανάλυση της Βάσης

Η ελληνική γλώσσα αποτελείται από 38 φωνήματα. Η συγκεκριμένη βάση εκφωνήσεων συνοδεύεται από αρχεία υποσημείωσης, στα οποία δηλώνεται η χρονική αρχή και το τέλος του κάθε φωνήματος σε κάθε εκφώνηση. Το σύνολο των 38 φωνημάτων, η γλωσσική τους αντιστοίχιση καθώς και η συχνότητα εμφάνισής τους στη βάση εκφωνήσεων που χρησιμοποιήθηκε παρουσιάζονται στον παρακάτω Πίνακα.

Φώνημα	Γλωσσική Αντιστοίχιση	Συχνότητα Εμφάνισης στη Βάση
a	α	1288
e	ε	1206
i	ι	1538
o	ο	969
u	ου	216
gr1	Τονισμένο α	349
gr2	Τονισμένο ε	413
gr3	Τονισμένο ι	537
gr4	Τονισμένο ο	512
gr5	Τονισμένο ου	78
g, G0	Ουρανικό γκ και πριν από ι	2, 2
h	Ένρινο ν πριν από χ	3
b	Χειλικό μπ	17
f	Χειλικό φ	135
p	Χειλικό π	495
v	Χειλικό β	223
k, c	Ουρανικό κ και πριν από ι	400, 372
J0, j	Ουρανικό γ και πριν από ι	60,121
x, X0	Ουρανικό χ και πριν από ι	113,168
D0	Οδοντικό δ	276
d	Οδοντικό ντ	372
T0	Οδοντικό θ	301
t	Οδοντικό τ	1126
M0	Λανθάνον Ένρινο μ	18
m	Ένρινο μ	503
N0	Λανθάνον Ένρινο ν	10
n	Ένρινο ν	916
l, L0	Υγρό λ και πριν από ι	290, 4
r	Υγρό ρ	872
s	Συριστικό ς	1424
z	Συριστικό ζ	21
sil1, sli2	Άφωνα	407, 240

6.1.2 Δημιουργία Διανύσματος Χαρακτηριστικών

Συνυπολογίζοντας όλα τα πειραματικά αποτελέσματα που παρουσιάστηκαν στο **Κεφάλαιο 2.4** επιλέχθηκαν τα Generalized Mel Cepstrum για τη μοντελοποίηση του συστήματος. Η χρήση αυτών είναι βέλτιστη τόσο από την

άποψη της ποιότητας απόδοσης του Vocoder όσο και από την άποψη της ανάλυσης σε κλίμακα Mel.

Για την εκπαίδευση του συστήματος της συγκεκριμένης διπλωματικής εξήχθησαν 24 Generalized Mel Cepstrum χαρακτηριστικά σε χρονικά παραθυρωμένα τμήματα φωνής διάρκειας $25ms$ με επικάλυψη 20% με συντελεστή αναδίπλωσης του συχνοτικού άξονα $a = 0.42$, όπου και εξασφαλίζει τον υπολογισμό των χαρακτηριστικών σε κλίμακα mel. Ο συντελεστής γενίκευσης των χαρακτηριστικών τέθηκε ίσος με $\gamma = -\frac{1}{7}$.

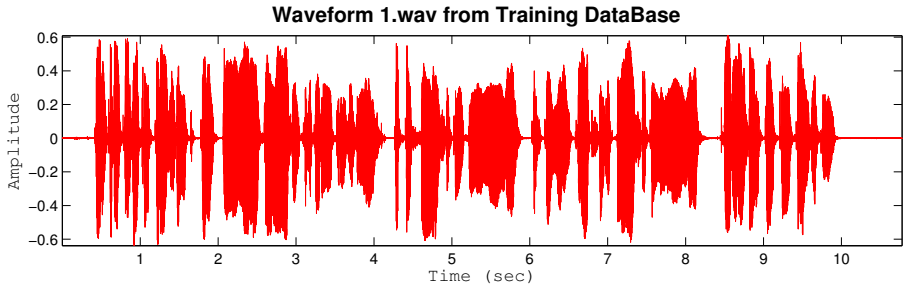
Πιο κάτω στο Σχήμα 6.1, φαίνονται τα αποτελέσματα της εξαγωγής των χαρακτηριστικών αυτών πάνω σε ένα σήμα φωνής. Παράλληλα, απεικονίζεται και το ανακατασκευασμένο σήμα φωνής από τον mgcep-Vocoder.

Το διάνυσμα χαρακτηριστικών που χρησιμοποιήθηκε στην εκπαίδευση του συστήματος δημιουργήθηκε από 78 χαρακτηριστικά. Συγκεκριμένα, αποτελείται από 3 blocks των 24 mgceps και 1 συντελεστή κέρδους, όπου το πρώτο αποτελείται από τα στατικά χαρακτηριστικά και τα άλλα δύο από τις διαφορές των πρώτων ως προς τα γειτονικά τους κατα μήκος διαδοχικών χρονικών παραθύρων και το τρίτο από τις διαφορές των διαφορών αυτών. Τα υπόλοιπα 3 χαρακτηριστικά αποτελούνται από το λογάριθμο του pitch και τις πρώτες και τις δεύτερες διαφορές αυτών αντίστοιχα. Η δομή του διανύσματος χαρακτηριστικών απεικονίζεται στο Σχήμα 6.2. Τα δυναμικά χαρακτηριστικά των mgceps και της ακολουθίας του pitch δίνονται στις Εξ. 6.1, 6.2.

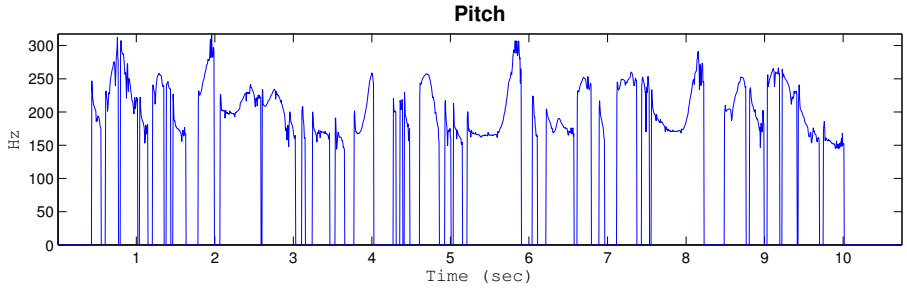
$$\begin{aligned} \delta c_t^{(i)} &= \frac{1}{2} \cdot c_{t-1}^{(i)} + \frac{1}{2} \cdot c_{t+1}^{(i)} \\ \delta^2 c_t^{(i)} &= \frac{1}{4} \cdot c_{t-1}^{(i)} - \frac{1}{2} \cdot c_t^{(i)} + \frac{1}{4} \cdot c_{t+1}^{(i)} \end{aligned} \tag{6.1}$$

$$\begin{aligned} \delta \log p_t &= \frac{1}{2} \cdot \log p_{t-1} + \frac{1}{2} \cdot \log p_{t+1} \\ \delta^2 \log p_t &= \frac{1}{4} \cdot \log p_{t-1} - \frac{1}{2} \cdot \log p_t + \frac{1}{4} \cdot \log p_{t+1} \end{aligned} \tag{6.2}$$

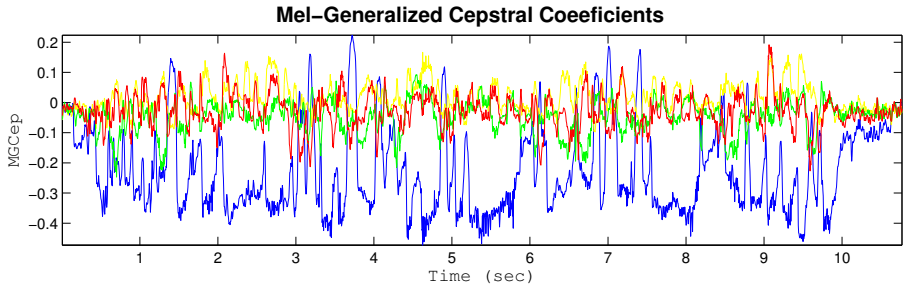
Για την εξαγωγή των χαρακτηριστικών καθώς και για την ομαδοποίησή τους χρησιμοποιήθηκαν τα υπολογιστικά εργαλεία επεξεργασίας ήχου SPTK [1].



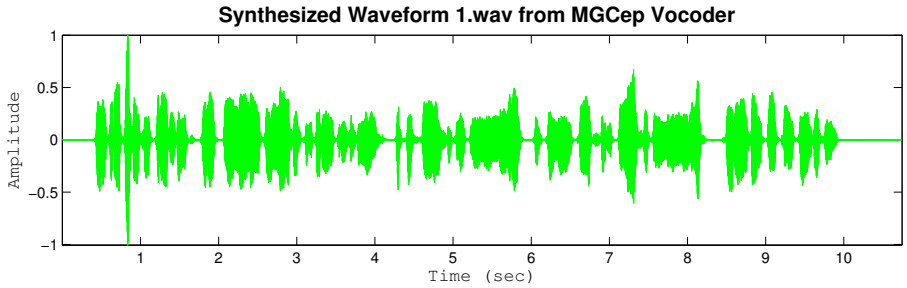
(a)



(b)



(c)



(d)

Σχήμα 6.1: Απεικόνιση (a) του αρχικού σήματος φωνής, (b) της εξαγόμενης ακολουθίας pitch που ανιχνεύθηκε, (c) της εξαγόμενης ακολουθίας των 4 πρώτων mgcep και (d) του συνθετικού τμήματος φωνής.

Σχήμα 6.2: Δημιουργία Διανύσματος Χαρακτηριστικών

6.2 Εκπαίδευση του Βασικού Τμήματος του Συνθέτη

6.2.1 Βασικοί Παράμετροι των Μοντέλων

Όπως αναφέρθηκε και στα προηγούμενα κεφάλαια, ο συνθέτης φωνής από κείμενο, που βασίζεται σε Κρυφά Μαρκοβιανά Μοντέλα, αποτελεί το αντίστροφο ενός συστήματος αναγνώρισης φωνής. Συγκεκριμένα, η εκπαίδευση των μαρκοβιανών μοντέλων είναι ακριβώς η ίδια διαδικασία με την εκπαίδευση των μοντέλων για ένα σύστημα αναγνώρισης. Για την μοντελοποίηση της βάσης του συνθέτη, λοιπόν, χρησιμοποιήθηκαν κρυφά μαρκοβιανά μοντέλα 5 καταστάσεων, με 4 streams. Το πρώτο stream περιέχει τα 75 χαρακτηριστικά των mel-generalized cepstrum και των δυναμικών χαρακτηριστικών αυτών. Τα υπόλοιπα 3 streams περιλαμβάνουν τους στατικούς συντελεστές του λογάριθμου του pitch και τους αντίστοιχους δυναμικούς πρώτης και δεύτερης τάξης. Οι πίνακες συμμεταβλητότητας αυτών είναι διαγώνιοι, οπότε και η εκπαίδευση των χαρακτηριστικών είναι ανεξάρτητη.

Όσον αφορά τη μοντελοποίηση του pitch, χρησιμοποιείται η εφαρμογή των Κρυφών Μαρκοβιανών Μοντέλων σε πολλαπλούς χώρους παρατηρήσεων (Multi-Space Distribution HMMs). Κάτι τέτοιο υλοποιείται με την αντιστοίχιση δύο κατανομών σε κάθε ένα από τα τρία τελευταία streams των μοντέλων. Συγκεκριμένα, η πρώτη κατανομή είναι απλή κατανομή μονοδιάστατης συνεχούς μεταβλητής, ενώ η δεύτερη μοντελοποιεί το σύμβολο του άφωνου ήχου. Η χρήση των κρυφών μαρκοβιανών μοντέλων σε πολλαπλούς χώρους είναι απαραίτητη, από την άποψη ότι οι τιμές που εξάγονται για την εκτίμηση του pitch στα έμφωνα τμήματα του ήχου, κυμαίνονται μεταξύ 100Hz και 350Hz, ενώ για τα άφωνα τμήματα, οι τιμές είναι απλά μηδενικές. Συνεπώς, διαχωρίζουμε από τη μία τη διακριτή τιμή των 0Hz που αντιπροσωπεύουν τα άφωνα τμήματα του ήχου, και από την άλλη τις συνεχείς τιμές των έμφωνων ήχων που είναι πολύ μακριά από τα 0Hz. Έτσι, δημιουργείται η ανάγκη της συνδυαστικής μοντελοποίησης των δύο αντιπροσωπευτικών δειγμάτων του pitch ταυτόχρονα, χωρίς οι εκτιμήσεις για τα έμφωνα τμήματα να αλλοιώνονται από τα μηδενικά των άφωνων, ούτε οι άφωνοι ήχοι να προκύπτουν από παρατηρήσεις κοντά στα 0Hz. Ο τρόπος, λοιπόν, της μοντελοποίησης με κρυφά

μαρκοβιανά μοντέλα σε πολλαπλούς χώρους, αποτελεί ιδανικό τρόπο αντιμετώπισης του συγκεκριμένου προβλήματος.

Σύμφωνα, με όλες τις παραπάνω περιγραφές, παρουσιάζεται πιο κάτω το πρότυπο HMM, βάσει του οποίου έγινε η μοντελοποίηση του συστήματος. Η συμβολική απεικόνιση του παρακάτω μοντέλου, έχει γίνει σύμφωνα με τα υπολογιστικά εργαλεία για HMMs και σύνθεση με HMMs, το HTK και το HTS αντίστοιχα. [50, 51]

Πρότυπο HMM

```
~o <VecSize>78<USER><DIAGC>
<MSDInfo>4 0 1 1 1 <StreamInfo> 4 75 1 1 1
~h "proto"
<BeginHMM>
    <NumStates>7
    <State>2
        <SWeights> 4 1.0 1.0 1.0 1.0
        <Stream> 1
            <Mean> 75
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
            <Variance> 75
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
.....
.....
```

```
.....
.....
          <Stream> 2
               <NumMixes> 2
                    <Mixture>1 0.5000
                         <Mean> 1
                                   0.0
                         <Variance> 1
                                   1.0
                    <Mixture>2 0.5000
                         <Mean> 0
                         <Variance> 0

 <TransP>7
     0.0e+0 1.0e+0 0.0e+0 0.0e+0 0.0e+0 0.0e+0 0.0e+0
     0.0e+0 6.0e -1 4.0e -1 0.0e+0 0.0e+0 0.0e+0 0.0e+0
     0.0e+0 0.0e+0 6.0e -1 4.0e -1 0.0e+0 0.0e+0 0.0e+0
     0.0e+0 0.0e+0 0.0e+0 6.0e -1 4.0e -1 0.0e+0 0.0e+0
     0.0e+0 0.0e+0 0.0e+0 0.0e+0 6.0e -1 4.0e -1 0.0e+0
     0.0e+0 0.0e+0 0.0e+0 0.0e+0 0.0e+0 6.0e -1 4.0e -1
     0.0e+0 0.0e+0 0.0e+0 0.0e+0 0.0e+0 0.0e+0 0.0e+0
<EndHMM>
```

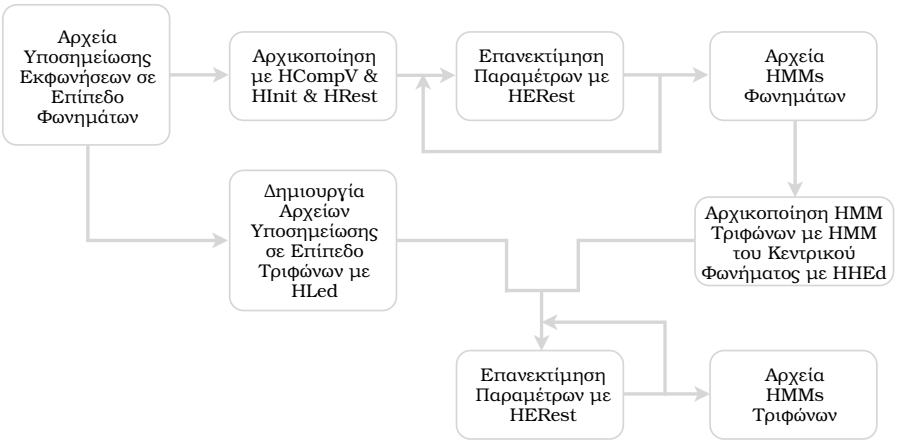
6.2.2 Αρχικοποίηση και Βασική Εκπαίδευση των Μοντέλων

Επίπεδο Φωνημάτων

Τα μοντέλα αρχικοποιήθηκαν από τη βάση εκφωνήσεων και στη συνέχεια εκπαιδεύτηκαν για κάθε φώνημα που εμφανίζεται. Χρησιμοποιώντας τα εργαλεία των HTK και HTS [50, 51], έγινε η εκπαίδευση 38 διαφορετικών μοντέλων, καθένα από τα οποία μοντελοποιούν κάθε ένα φώνημα. Πιο συγκεκριμένα, έγινε χρήση των συναρτήσεων HCompV, HInit, HRest, HERest κ.α. Παράλληλα εξήχθησαν και τα μοντέλα για τη διάρκεια των φωνημάτων.

Επίπεδο Τριφώνων

Για τη βελτιστοποίηση της ποιότητας των αποτελεσμάτων και για να δοθεί έμφαση στα συμφραζόμενα του κάθε φωνήματος, διαχωρίζονται τα τρίφωνα



Σχήμα 6.3: Λειτουργικό Διάγραμμα του Τμήματος Εκπαίδευσης του Συνθέτη Φωνής με τη Χρήση του Εργαλείου HTS

και εκπαιδεύονται εκ νέου τα μοντέλα. Συγκεκριμένα, με αρχικοποίηση που προκύπτει από τα προηγούμενα φωνήματα, κάθε τρίφωνο επανεκπαιδεύεται, τώρα πια με τα χαρακτηριστικά που αντιστοιχούν στα στιγμιότυπα εμφάνισης των τριφώνων. Χρησιμοποιώντας, λοιπόν, την HERest, εκπαιδεύονται όλα τα νέα μοντέλα τριφώνων. Το λειτουργικό διάγραμμα του τμήματος της εκπαίδευσης του συνθέτη φωνής παρουσιάζεται πιο κάτω στο Σχήμα 6.3.

6.3 Παραγωγή Συνθετικής Φωνής

6.3.1 Παραγωγή Φωνητικών Χαρακτηριστικών από τα Κρυφά Μαρκοβιανά Μοντέλα

Στο τμήμα αυτό του Σύνθέτη δημιουργείται ένα σύστημα, το οποίο δέχεται ως είσοδο το κείμενο το οποίο θα μετατραπεί σε φωνή. Το κείμενο αυτό εισάγεται στο σύστημα μας με τη μορφή ακολουθίας φωνημάτων. Στη συνέχεια, χρησιμοποιώντας τη συνάρτηση HMGenS του HTS [51] προκύπτουν τα εκτιμώμενα φωνητικά χαρακτηριστικά. Έτσι εξάγεται τόσο η διάρκεια κάθε κατάστασης, όσο και οι συντελεστές των mgceps και του pitch, που αντιστοιχούν σε κάθε χρονικό παράθυρο. Ως όρισμα, αυτή η συνάρτηση λαμβάνει και το ποιο από τα τρία βασικά προβλήματα, που παρουσιάζονται στο **Κεφάλαιο 4.1**, πρέπει να επιλυθεί. Παράλληλα, στην πιο πάνω συνάρτηση εισάγεται και η πληροφορία του ποια χαρακτηριστικά από τα εκτιμώμενα είναι στατικά και ποια δυναμικά, καθώς και τα αντίστοιχα παράθυρα υπολογισμού τους. Με

αυτόν τον τρόπο παρεισφρύει η έννοια της δυναμικότητας και της χρονικής συνέχειας στα εκτιμώμενα φωνητικά χαρακτηριστικά.

Αρχείο Εκτίμησης της διάρκειας καταστάσεων

t-i+n.state[2]: duration=1 (frame), mean=1.298813e+00
t-i+n.state[3]: duration=1 (frame), mean=1.137526e+00
t-i+n.state[4]: duration=2 (frame), mean=1.610275e+00
t-i+n.state[5]: duration=3 (frame), mean=2.799829e+00
t-i+n.state[6]: duration=2 (frame), mean=1.908825e+00
t-i+n: duration=9 (frame), mean=8.755268e+00
i-n+t.state[2]: duration=1 (frame), mean=1.695340e+00
i-n+t.state[3]: duration=2 (frame), mean=1.545794e+00
i-n+t.state[4]: duration=5 (frame), mean=5.181313e+00
i-n+t.state[5]: duration=2 (frame), mean=1.909217e+00
i-n+t.state[6]: duration=2 (frame), mean=1.734045e+00
i-n+t: duration=12 (frame), mean=1.206571e+01
n-t+r.state[2]: duration=4 (frame), mean=3.821033e+00
n-t+r.state[3]: duration=1 (frame), mean=1.166671e+00
n-t+r.state[4]: duration=2 (frame), mean=2.500008e+00
n-t+r.state[5]: duration=3 (frame), mean=3.000000e+00
n-t+r.state[6]: duration=3 (frame), mean=2.333394e+00
n-t+r: duration=13 (frame), mean=1.282111e+01
t-r+gr3.state[2]: duration=3 (frame), mean=3.200440e+00
t-r+gr3.state[3]: duration=2 (frame), mean=2.078582e+00
...

6.3.2 Τελική Σύνθεση Φωνής

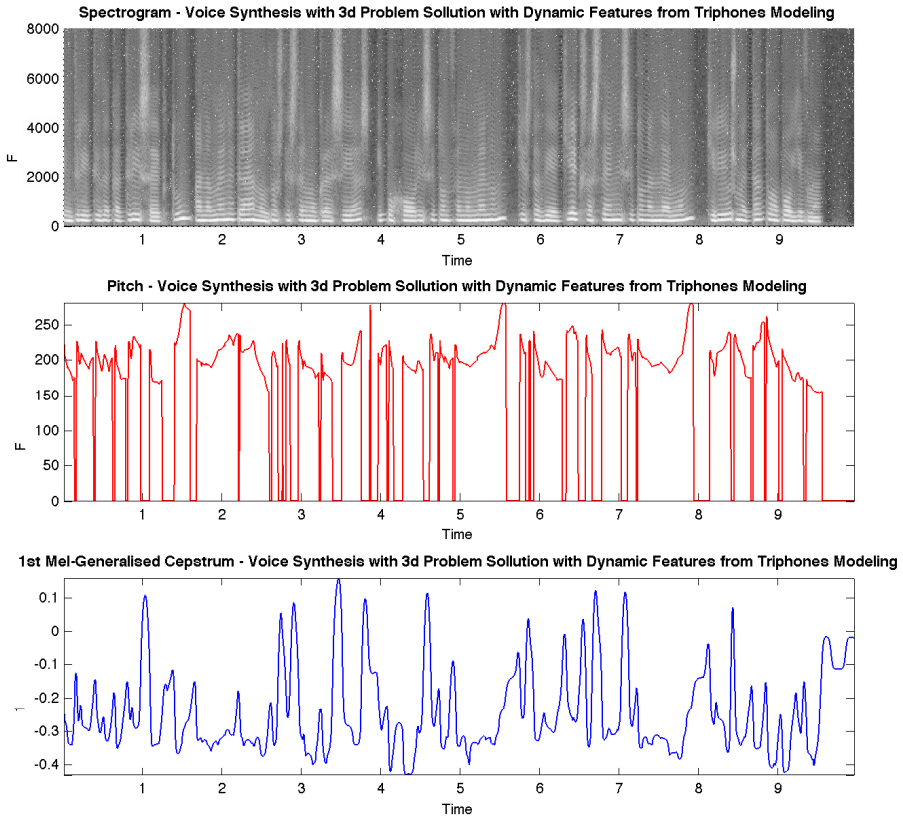
Στο τελευταίο αυτό τμήμα του συνθέτη, εφαρμόζεται όλη η θεωρία που παρουσιάστηκε στο **Κεφάλαιο 2.2.4**. Συγκεκριμένα, εφαρμόζεται το MLSA φίλτρο για την αντιστροφή των χαρακτηριστικών των mgceps και του pitch. Η εφαρμογή της σύνθεσης έγινε με τη χρήση της κατάλληλης συνάρτησης του SPTK [1]. Οι παράμετροι παραθύρωσης του ήχου, η σταθερά γενίκευσης των χαρακτηριστικών καθώς και η σταθερά αναδίπλωσης του συχνοτικού άξονα είναι ακριβώς ίδιες με αυτές που χρησιμοποιήθηκαν και στην αρχική εξαγωγή των παραμέτρων.

6.4 Πειραματικά Αποτελέσματα

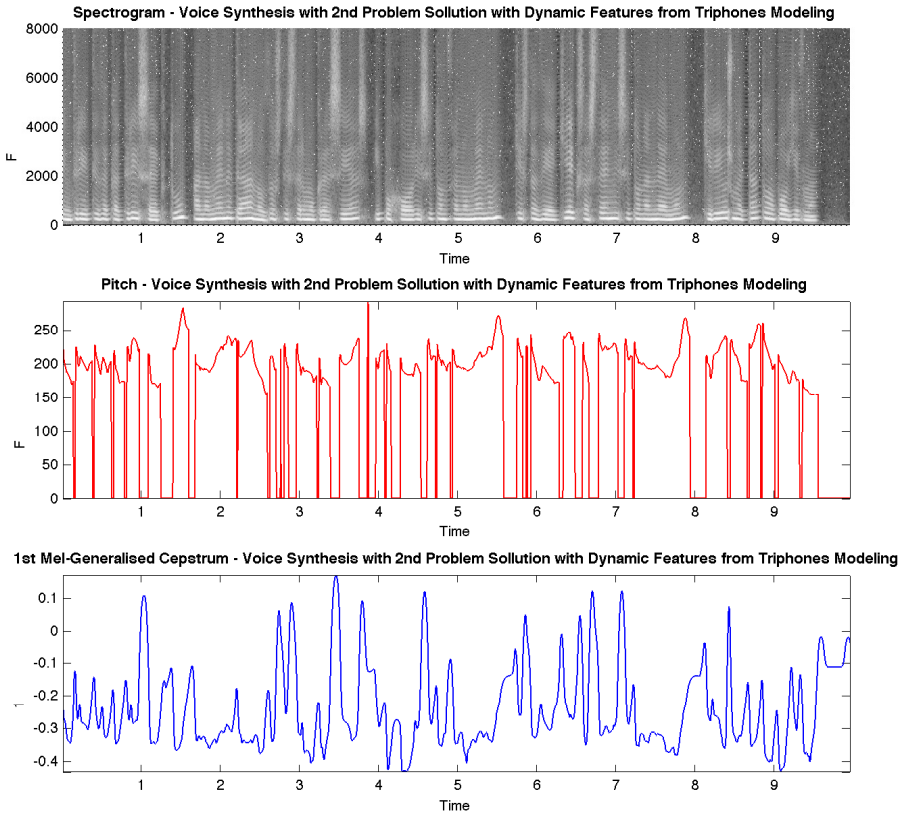
6.4.1 Παραγωγή Πρότασης από τη Βάση Εκπαίδευσης

Αρχικά, η πρώτη δοκιμή έγινε με την παραγωγή της πρώτης πρότασης της βάσης εκπαίδευσης. Τα ηχητικά αρχεία της συγκεκριμένης πρότασης βρίσκονται αποθηκευμένα στη βάση της διπλωματικής στο φάκελο Database_first_voice_extraction_wavs. Στο Σχήμα 6.4 εμφανίζεται το αποτέλεσμα της εκτίμησης των παραμέτρων της συνθετικής φωνής καθώς και το σπεκτρογράφημα αυτής. Συγκεκριμένα, στο σχήμα αυτό παρατηρούμε την πολύ ομαλή ως προς το χρόνο, εκτίμηση των παραμέτρων. Το ηχητικό αποτέλεσμα είναι σχεδόν τέλειο χωρίς να εμφανίζεται κάποια παρουσία επαναλαμβανόμενου ήχου, που θα μπορούσε να προκύπτει από τη χρονική ασυνέχεια των εκτιμώμενων παραμέτρων. Είναι πολύ σημαντικό, να τονιστεί στο συγκεκριμένο σημείο, ότι η παραγωγή μίας πρότασης από τη βάση δε σημαίνει ότι το αποτέλεσμα είναι πλήρως μεροληπτικό, από την άποψη ότι δεν υπάρχει αποθηκευμένος ήχος. Παράλληλα, κάθε τρίφωνο που ζητείται να παραχθεί έχει μοντελοποιηθεί με ένα HMM, το οποίο έχει εκπαιδευτεί από περισσότερες από μία εμφανίσεις στη βάση και όχι μόνο από την εμφάνιση του στη συγκεκριμένη εκφώνηση. Αυτό αποτελεί και τη βασική διαφοροποίηση του Συνθέτη που βασίζεται σε Κρυφά Μαρκοβιανά Μοντέλα σε σχέση με το Συνθέτη τεχνολογίας Unit Selection. Συγκεκριμένα, εάν από τον τελευταίο ζητηθεί να εκφωνηθεί μία φράση που υπάρχει στη βάση εκπαίδευσης η εκφώνηση θα είναι τέλεια και ίδια με την αρχική, μιας και θα προκύψει από τη συνένωση των τμημάτων του αποθηκευμένου ήχου της βάσης, ενώ αντίθετα σε ένα Συνθέτη σαν το συγκεκριμένο, η παραγωγή μίας πρότασης από τη βάση εκφωνήσεων θα είναι αρκετά καλή, αλλά σε καμία περίπτωση ίδια με αυτή της βάσης [5].

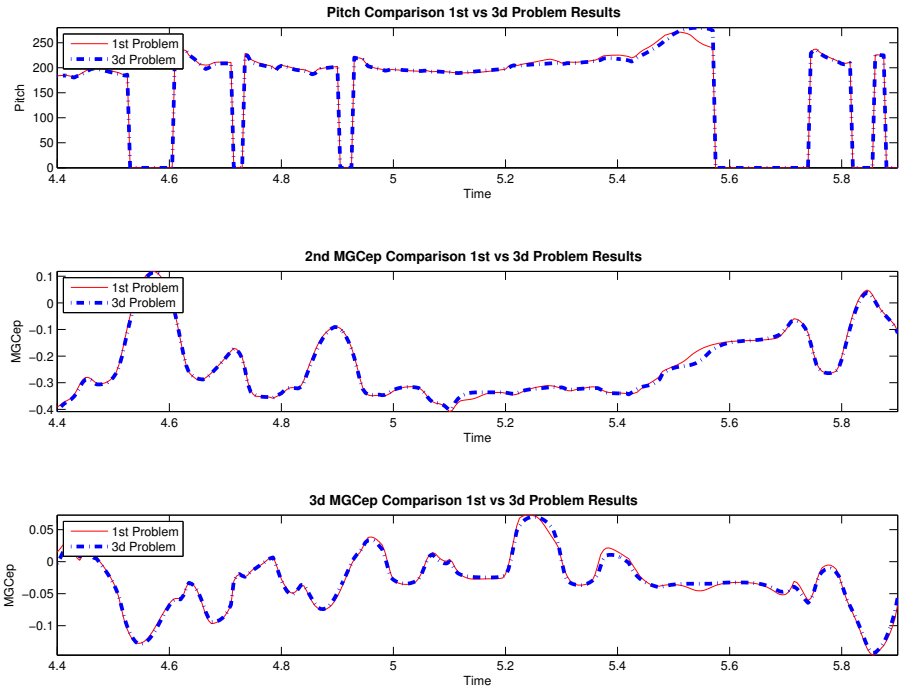
Παράλληλα, η συγκεκριμένη πρόταση συντέθηκε και με τους τρεις τρόπους παραγωγής, οι οποίοι αντιστοιχούν στην επίλυση των τριών προβλημάτων του **Κεφαλαίου 4.1**. Στο Σχήμα 6.5 απεικονίζονται τα ίδια αποτελέσματα όπως και στο προηγούμενο, αλλά για την περίπτωση της επίλυσης του 2ου προβλήματος, ενώ στο Σχήμα 6.4, τα αποτελέσματα αφορούν την επίλυση του 3ου προβλήματος. Αν και ακουστικά δεν εντοπίζονται διαφορές στα δύο πειραματικά αποτελέσματα, σε επίπεδο χαρακτηριστικών είναι εμφανές ότι οι καμπύλες του χρόνου έχουν ομαλοποιηθεί αρκετά στην περίπτωση της επίλυσης του προβλήματος 3, όπως είναι εμφανές και στο Σχήμα 6.6.



Σχήμα 6.4: Απεικόνιση του Σπεκτρογραφήματος της Συνθετικής Φωνής και της Εκτιμώμενης Ακολουθίας τοPitch και του 1ου Συντελεστή Mgcep, μετά την Επίλυση του 3ου Προβλήματος



Σχήμα 6.5: Απεικόνιση του Σπεκτρογραφήματος της Συνθετικής Φωνής και της Εκτιμώμενης Ακολουθίας τοPitch και του 1ου Συντελεστή Mgcep, μετά την Επίλυση του 1ου Προβλήματος



Σχήμα 6.6: Απεικόνιση Αποκλίσεων Εκτίμησης των Φωνητικών Χαρακτηριστικών από την Επίλυσης των Προβλημάτων 1 και 3

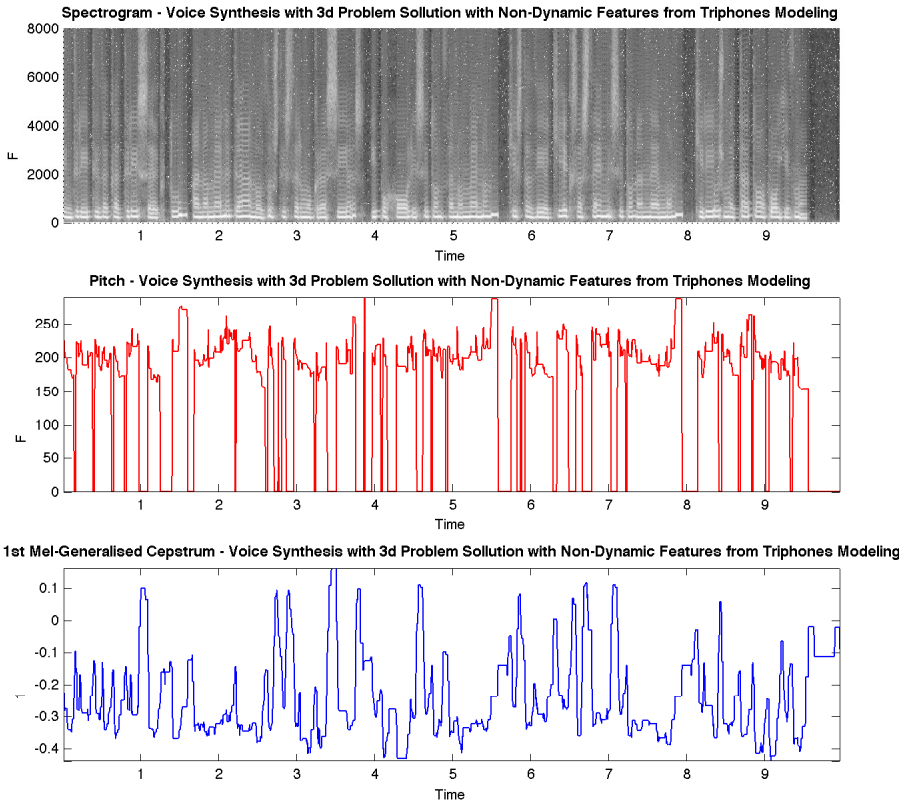
Παραγωγή Χωρίς να Συνυπολογιστούν τα Δυναμικά Χαρακτηριστικά

Στη συνέχεια πραγματοποιήθηκε παραγωγή της συνθετικής φωνής, χωρίς να συνυπολογιστεί η ύπαρξη των δυναμικών χαρακτηριστικών. Συγκεκριμένα, στην περίπτωση αυτή, οι ήχοι που προέκυψαν παρουσίασαν μία έντονη ακουστική διαφορά. Ενώ το μήνυμα της πρότασης, που ζητήθηκε να εκφωνηθεί, αποδίδεται ξεκάθαρα, παράλληλα εντοπίζεται ένας περιοδικός ήχος, ο οποίος υποβιβάζει την ποιότητα της συνθετικής φωνής. Η παρείσφρυση αυτού του περιοδικού ήχου, οφείλεται στην ασυνέχεια των εκτιμώμενων παραμέτρων από το σύστημά μας. Αυτό είναι εμφανές στο Σχήμα 6.7. Συγκεκριμένα, παρατηρούνται ξεκάθαρα τόσο οι ασυνέχειες στην εκτίμηση των mgceps, όσο και στην εκτίμηση του pitch. Στο Σχήμα 6.8 παρατηρούνται οι ασυνέχειες που παρουσιάζονται στην εκτίμηση των παραμέτρων του pitch και των mgceps στις περιπτώσεις που συνυπολογίζονται και στις περιπτώσεις που δε συνυπολογίζονται τα δυναμικά χαρακτηριστικά αντίστοιχα. Μία πολύ καλή απεικόνιση των ασυνεχειών που παρουσιάζονται απεικονίζεται στο Σχήμα 6.9, όπου στα πλαίσια της μεγένθυσης φαίνονται ξεκάθαρα τα κάθετα τμήματα του φάσματος που εμφανίζονται και συνενώνουν τις οριζόντιες τροχιές των συχνοτήτων.

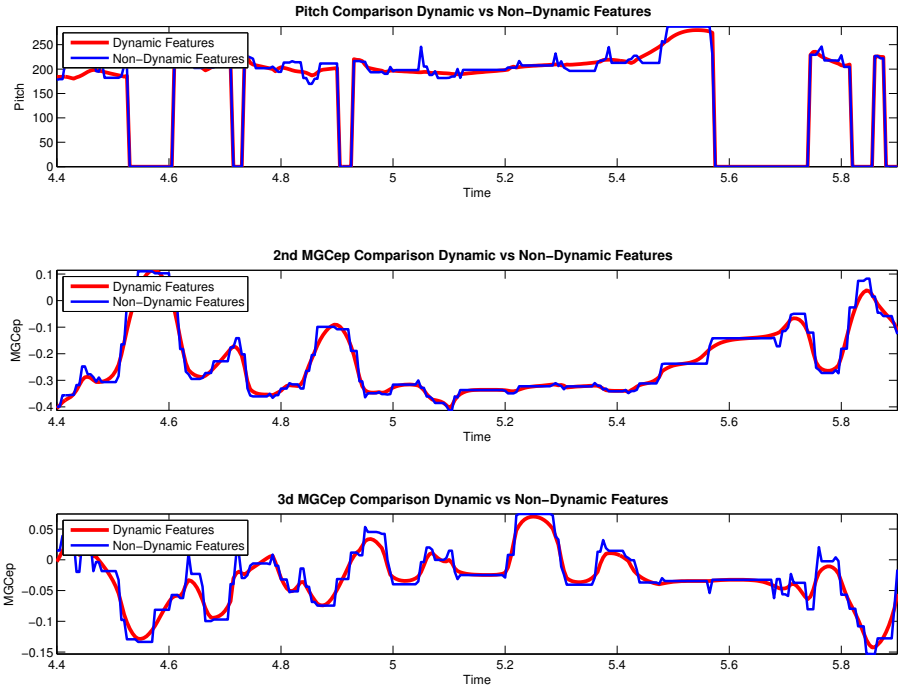
Για την αντιμετώπιση αυτών των ασυνεχειών που εμφανίζονται στο σήμα του ήχου, εφαρμόστηκαν αρκετές τεχνικές, ομαλοποίησης. Αρχικά εφαρμόστηκε ομαλοποίηση απευθείας του σήματος της συνθετικής φωνής. Αυτή η διαδικασία μπορεί γραφικά να βελτίωσε ελάχιστα το αποτέλεσμα, αλλά ακουστικά πέτυχε απλά την αποβολή του αθροιστικού θορύβου που προέκυπτε, χωρίς να αποβάλει τα ηχητικά αποτελέσματα των ασυνεχειών της σύνθεσης. Στο Σχήμα 6.10 εμφανίζεται το αποτέλεσμα στο σπεκτρογράφημα, στο οποίο παρατηρούμε και γραφικά τη μη αποβολή των ασυνεχειών αλλά την ομαλοποίηση της φασματικής πληροφορίας σε συχνότητες μεγαλύτερες των 3.5kHz.

Στη συνέχεια, εφαρμόστηκε ομαλοποίηση των εξαγόμενων χαρακτηριστικών από το Σύστημα. Με την απλή ομαλοποίηση, επιτεύχθηκε η αποκοπή των ασυνεχειών στα χαρακτηριστικά, αλλά με την ομαλοποίηση των ακμών, ειδικά στην περίπτωση του pitch, μεταβλήθηκε η προσωδία της παραγόμενης φωνής. Συγκεκριμένα, ο ήχος αυτός βρίσκεται στη βάση της διπλωματικής με όνομα αρχείου fc_02_ndf_smooth_pitch_and_mgc.wav. Κατά το άκουσμα αυτού αλλάζει το ηχόχρωμα της φωνής, επειδή μεταβάλλεται ο ρυθμός μεταβολής της βασικής συχνότητας.

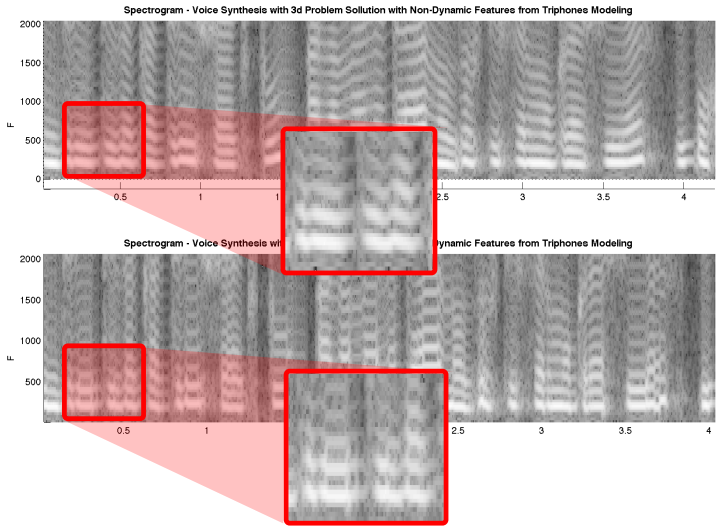
Τέλος, για την αποφυγή της συνέπειας της αλλαγής της προσωδίας στο ηχητικό αποτέλεσμα, πραγματοποιήθηκε η εξομάλυνση των χαρακτηριστικών εφαρμόζωντας δύο διαδοχικά φίλτρα, της διαμέσου για κάθε 9 γειτονικά σημεία και στη συνέχεια του μέσου όρου για τα 3 γειτονικά σημεία και τέλος την περικοπή των ακράιων τιμών. Γραφικά, το αποτέλεσμα στο επίπεδο



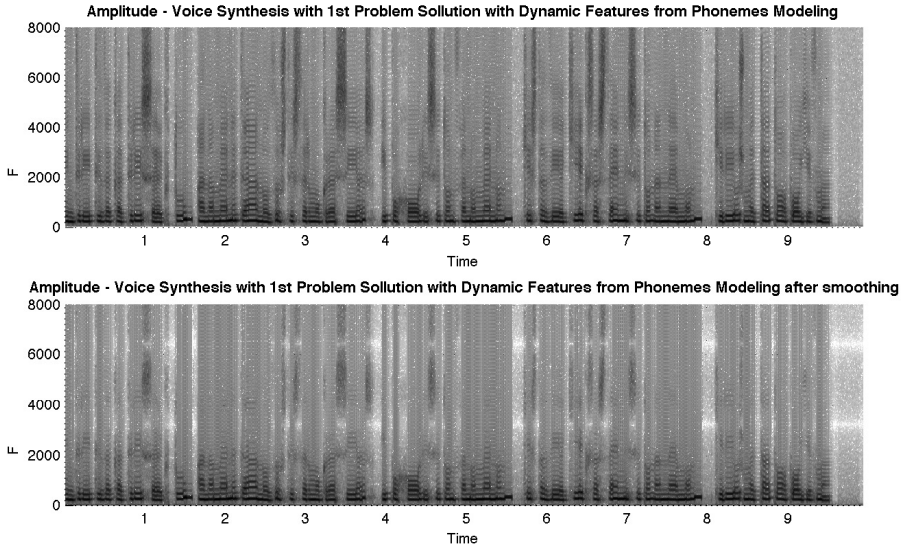
Σχήμα 6.7: Απεικόνιση του Σπεκτρογραφήματος της Συνθετικής Φωνής και της Εκτιμώμενης Ακολουθίας τοPitch και του 1ου Συντελεστή Mgcep, μετά την Επίλυση του 3ου Προβλήματος, Χωρίς να Συνυπολογιστούν τα Δυναμικά Χαρακτηριστικά



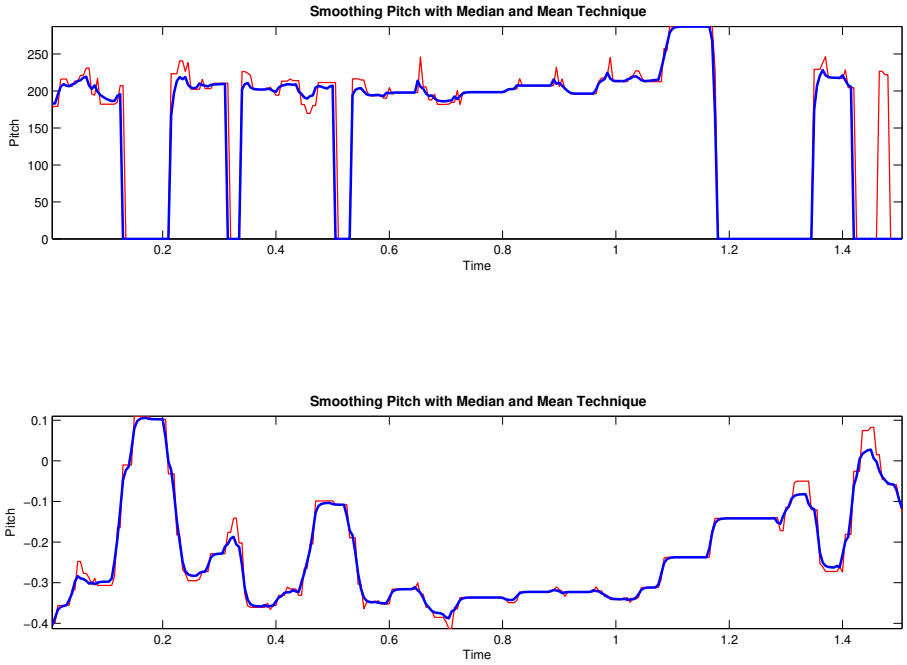
Σχήμα 6.8: Συγκριτική Απεικόνιση των Εκτιμώμενων Παραμέτρων στις Περιπτώσεις Συνυπολογισμού και του μη-Συνυπολογισμού των Δυναμικών Χαρακτηριστικών.



Σχήμα 6.9: Συγκριτική Απεικόνιση των Σπεκτρογραφήματος στις Περιπτώσεις Συνυπολογισμού και του μη-Συνυπολογισμού των Δυναμικών Χαρακτηριστικών.



Σχήμα 6.10: Απεικόνιση των Σπεκτρογραφημάτων του Συνθετικού Σήματος και το Ομαλοποιημένου στο Χρόνο Σήματος

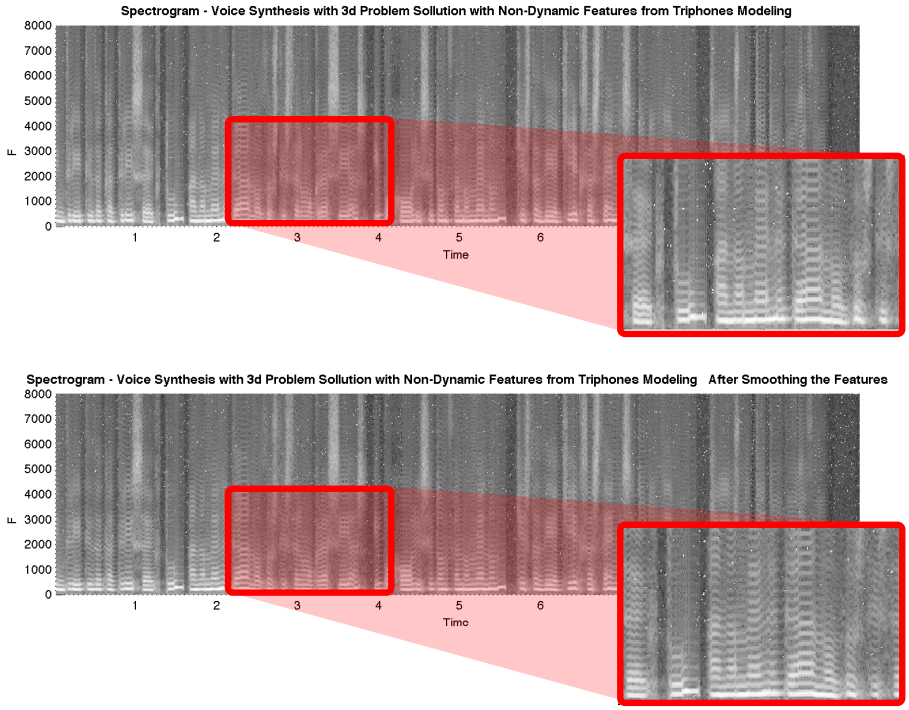


Σχήμα 6.11: Απεικόνιση των Ομαλοποιημένων Χαρακτηριστικών σε Αντιπαράθεση με τα Αρχικά.

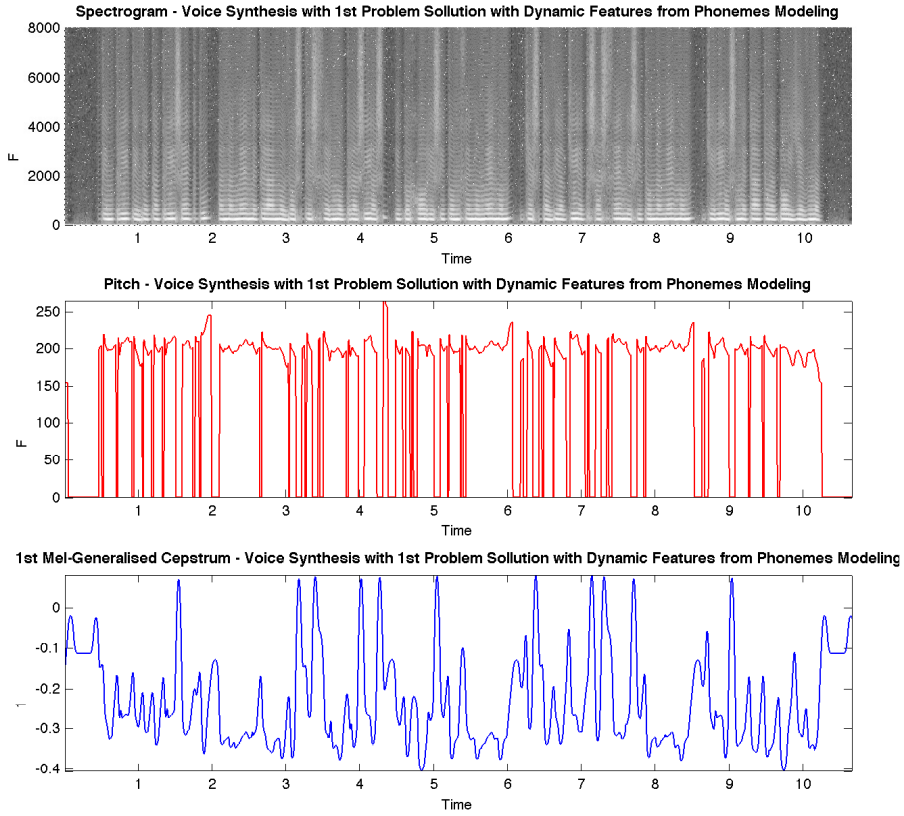
των χαρακτηριστικών απεικονίζεται στο Σχήμα 6.11. Είναι φανερό ότι με τη συγκεκριμένη τεχνική πετύχαμε τέλεια ομαλοποίηση χαρακτηριστικών όπως είναι εμφανές. Το ηχητικό αποτέλεσμα το οποίο προέκυψε ήταν αρκετά πιο ομαλοποιημένο ακουστικά. Κάτι τέτοιο είναι εμφανές και στο Σχήμα 6.12, όπου στο εστιασμένο κομμάτι παρατηρούμε τη μερική απομάκρυνση των κάθετων τμημάτων.

Παραγωγή από τα HMMs των Φωνημάτων

Στη συνέχεια, χρησιμοποιώντας τη μοντελοποίηση σε επίπεδο φωνημάτων συντέθηκε η πρώτη πρόταση της βάσης εκπαίδευσης. Τα αποτελέσματα αυτής της σύνθεσης παρουσιάζονται στο Σχήμα 6.13. Είναι εμφανής η συνέχεια των παραγόμενων χαρακτηριστικών, που αποδεικνύει την σύνθεση φωνής μετά από τον συνυπολογισμό των δυναμικών χαρακτηριστικών. Το συγκεκριμένο ακουστικό αποτέλεσμα περιλαμβάνεται στη βάση της διπλωματικής με όνομα αρχείου ph_00_df.wav. Ακουστικά το αποτέλεσμα είναι πάρα πολύ καλό, στοιχείο που είναι εμφανές και στο καθαρό σπεκτρογράφημα της εκφώνησης. Η διαφορά του από αυτό με τη μοντελοποίηση τριφώνων εντοπίζεται στο



Σχήμα 6.12: Συγκριτική Απεικόνιση των Σπεκτρογραφημάτων μετά από Ομαλοποίηση των Χαρακτηριστικών.



Σχήμα 6.13: Απεικόνιση του Σπεκτρογραφήματος της Συνθετικής Φωνής και της Εκτιμώμενης Ακολουθίας το Pitch και του 1ου Συντελεστή Mgcep

σημείο του ότι κάθε φώνημα ακούγεται σχεδόν ανεξάρτητα, χωρίς όμως να δίνεται η αίσθηση της χρονικής ασυνέχειας, λόγω της μοντελοποίησης με το συνυπολογισμό των δυναμικών χαρακτηριστικών. Αυτό συμβαίνει λόγω του ότι το κάθε φώνημα είναι εκπαιδευμένο ανεξάρτητα.

6.4.2 Παραγωγή Πρότασης ανεξάρτητη από τη Βάση Εκπαίδευσης

Το επόμενο και πιο βασικό πειραματικό αποτέλεσμα ήταν η παραγωγή τυχαίας εκφώνησης που δεν περιλαμβάνονταν στην αρχική βάση εκφωνήσεων. Συγκεκριμένα, η παραγωγή ήχων έγινε και με τους τρεις τρόπους όπως και στο προηγούμενο υποκεφάλαιο. Δηλαδή, έγινε παραγωγή βάσει της μοντελοποίησης τριφώνων με το συνυπολογισμό των δυναμικών φωνητικών χαρακτηριστικών, χωρίς τον συνυπολογισμό αυτών και με βάση τη μοντελοποίηση

απλών φωνημάτων.

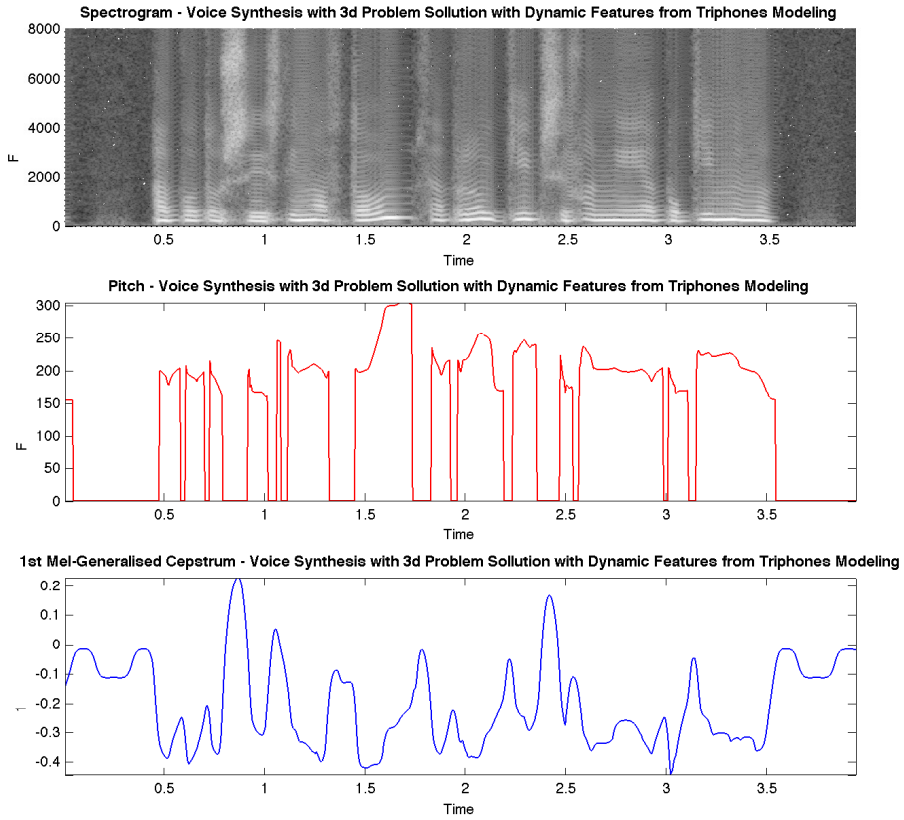
Παραγωγή από Μοντέλα Τριφώνων με Δυναμικά Χαρακτηριστικά

Στο σημείο αυτό παραγάγαμε μια σειρά από εκφωνήσεις τυχαίες και εκτός βάσης. Μία από τις εκφωνήσεις που παρήχθησαν ήταν και η πρόταση: "Από το σύστημα αυτό θα προκύψει φωνή από κείμενο". Στο πλαίσιο που ακολουθεί παρουσιάζεται η ακολουθία τριφώνων από την οποία προέκυψε η εκφώνηση αυτή. Στο Σχήμα 6.14 εμφανίζονται το Σπεκτρογράφημα του ήχου που συντέθηκε καθώς και η εκτιμώμενη ακολουθία pitch και ο Πρώτος συντελεστής mgcep. Το αποτέλεσμα ηχητικά προκύπτει αρκετά ομαλό, πράγμα το οποίο είναι εμφανές και στο σπεκτρογράφημα του Σχήματος. Ιδιαίτερα ομαλά είναι και τα εκτιμώμενα χαρακτηριστικά του pitch και των mgceps. Οι αποκλίσεις από την επιθυμητή ποιότητα στο συγκεκριμένο αποτέλεσμα μπορεί να προέρχονται από δύο πηγές, αφενός από τη μη ύπαρξη συγκεκριμένου τριφώνου, καθώς και από το λάθος τονισμό και τη μη επιθυμητή προκύπτουσα προσωδία.

•sil1-sil1+sil1 •sil1-a+p •a-p+o •p-o+t •o-t+o •t-o+s •o-s+gr3 •s-gr3+s •gr3-s+t •s-t+i •t-i+m •i-m+a •m-a+p •a-f+t •f-t+gr4 •t-gr4+sil2 •gr4-sil2+s •sil2-T0+a •T0-a+p •a-p+r •p-r+o •r-o+c •o-c+gr3 •c-gr3+p •gr3-p+s •p-s+i •s-i+f •i-f+o •f-o+n •o-n+gr3 •n-gr3+a •gr3-a+p •a-p+o •p-o+k •o-c+gr3 •c-gr3+m •gr3-m+e •m-e+n •e-n+o •n-o+s •sil1-sil1+sil1

Στην πρώτη περίπτωση, το πρόβλημα διορθώνεται με την εκπαίδευση και των μοντέλων των τριφώνων που δεν υπάρχουν στη βάση εκπαίδευσης. Αυτό μπορεί να επιτευχθεί με δενδρικό διαχωρισμό των κατηγοριών των τριφώνων, βάσει γλωσσολογικών κριτηρίων [49]. Δεδομένου, ότι σε αυτή τη διπλωματική τα κριτήρια αυτά δεν ήταν διαθέσιμα δεν έγινε η εκπαίδευση των μη υπαρχόντων τριφώνων. Ελλείψει βέβαια κάποιων, χρησιμοποιήθηκαν διαισθητικά κάποια μοντέλα, τα οποία προσέγγιζαν το επιθυμητό. Όπως είναι εμφανές, χρήση τέτοιων μοντέλων έγινε και στην εκφώνηση της συγκεκριμένης πρότασης αφού έλλειπε το μοντέλο m-a+f και στη θέση του χρησιμοποιήθηκε το m-a+p. Το αποτέλεσμα δεν παρουσίασε κάποιο πρόβλημα στην απόδοση της συνθετικής φωνής στο συγκεκριμένο σημείο.

Στη δεύτερη περίπτωση, όπου μπορεί να έχουμε λάθος τονισμό ή διαφορετική προσωδία από την επιθυμητή, οι τεχνικές αντιμετώπισης αφορούν την επεξεργασία του ήχου μετά την παραγωγή από το σύστημα. Μία πολύ δημοφιλής τεχνική προς αυτήν την κατεύθυνση έχει αναπτυχθεί από τους



Σχήμα 6.14: Απεικόνιση του Σπεκτρογραφήματος της Συνθετικής Φωνής και της Εκτιμώμενης Ακολουθίας τοPitch και του 1ου Συντελεστή Mgcep, μετά την Επίλυση του 3ου Προβλήματος

Kawahara et al. [16], και αποτελεί τεχνική ομαλοποίησης, προσαρμοζόμενη στο στιγμιαίο pitch της φωνής, τόσο στο πεδίο του χρόνου όσο και στο πεδίο της συχνότητας. Με αυτήν την τεχνική ανάλυσης σύνθεσης, η οποία έχει συνδυαστεί άμεσα με όλες τις ερευνητικές προσπάθειες της σύνθεσης φωνής από κείμενο, που βασίζονται στα Κρυφά Μαρκοβιανά Μοντέλα, μπορεί σε μία πρόταση να αλλαχθεί η προσωδία και να μετατραπεί από κατάφαση σε ερώτηση. Μία τέτοια τεχνική είναι πολύ χρήσιμη σε έναν συνθέτη φωνής από κείμενο, αφού αυτή η πληροφορία στο γραπτό λόγο εισέρχεται μέσω των σημείων στίξης.

Παραγωγή από Μοντέλα Τριφώνων χωρίς να συνυπολογιστούν τα Δυναμικά Χαρακτηριστικά

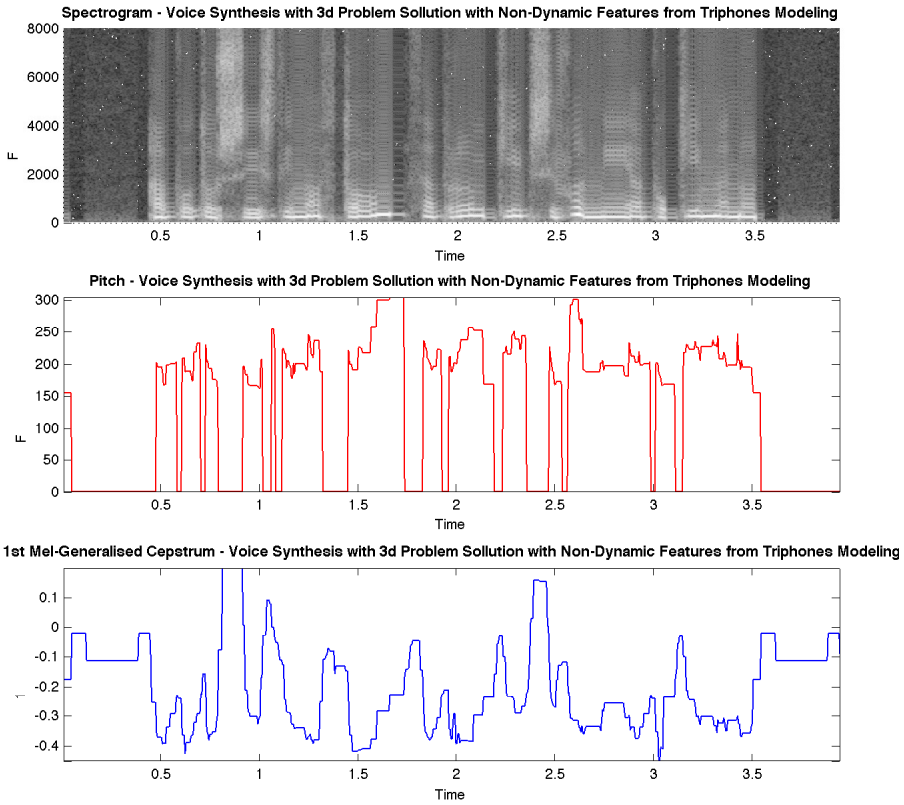
Όπως και στο προηγούμενο υποκεφάλαιο έτσι και τώρα συντέθηκε η ζητούμενη πρόταση χωρίς να συνυπολογιστούν τα δυναμικά χαρακτηριστικά του pitch και των mgceps. Γραφικά το αποτέλεσμα παρουσιάζεται στο Σχήμα 6.15. Είναι εμφανές ότι η εκτίμηση των χαρακτηριστικών και σε αυτή την περίπτωση παρουσιάζει ασυνέχειες, συνεπώς με παρόμοιες τεχνικές ομαλοποίησης αυτών θα προκύψει καλύτερο ηχητικό αποτέλεσμα.

Παραγωγή από Μοντέλα Απλών Φωνημάτων

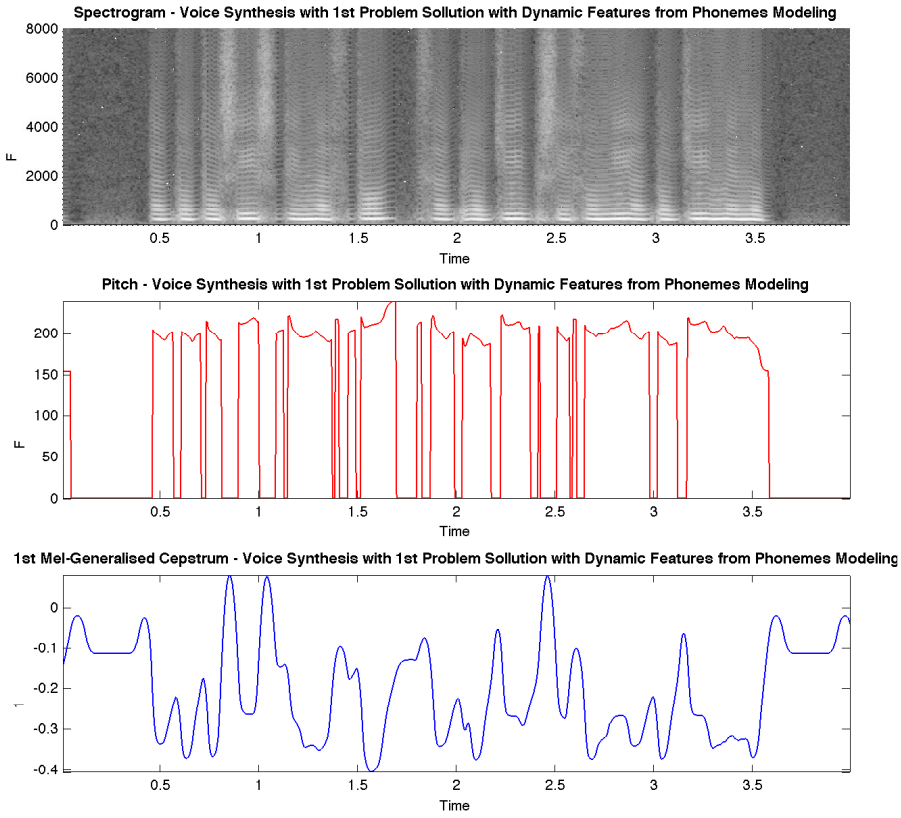
Στο στάδιο αυτό πραγματοποιείται και πάλι η σύνθεση της επιθυμητής πρότασης από τα μοντέλα των απλών φωνημάτων. Το αποτέλεσμα είναι πάρα πολύ καλό όσον αφορά τη συνέχεια των εκτιμώμενων χαρακτηριστικών. Κάτι τέτοιο είναι αντιληπτό και στο Σχήμα 6.16. Στην περίπτωση της μοντελοποίησης των φωνημάτων παρακάμπτεται το πρόβλημα της μη ύπαρξης μοντέλου που δημιουργείται με τη μοντελοποίηση των τριφώνων. Σίγουρα η απόδοση είναι πιο χαμηλή σε ποιότητα από αυτή της μοντελοποίησης των τριφώνων, αλλά δεν παρουσιάζει και τεράστια διαφορά.

6.5 Σύγκριση Αποτελεσμάτων

Τα πειραματικά αποτελέσματα που προέκυψαν καταδεικνύουν την προσαρμοστική δομή ενός τέτοιου συνθέτη φωνής. Συγκεκριμένα, παρατηρήθηκαν τα εξίσου καλά αποτελέσματα τόσο με τη μοντελοποίηση τριφώνων όσο και με τη μοντελοποίηση φωνημάτων. Ένα πάρα πολύ σημαντικό στοιχείο είναι το πολύ μικρό μέγεθος του συστήματος. Συγκεκριμένα, χρησιμοποιήθηκε μία βάση εκφωνήσεων συνολικού μεγέθους 41.9MB. Μετά από όλα τα στάδια του απαιτούμενου προγραμματισμού και χωρίς καμία συμπίεση, τα αποθηκευμένα μοντέλα καταλαμβάνουν 39.9MB σε επίπεδο τριφώνων ή μόνο



Σχήμα 6.15: Απεικόνιση του Σπεκτρογραφήματος της Συνθετικής Φωνής και της Εκτιμώμενης Ακολουθίας τοPitch και του 1ου Συντελεστή Mgcep, μετά την Επίλυση του 3ου Προβλήματος Χωρίς να Συνυπολογιστούν τα Δυναμικά Χαρακτηριστικά



Σχήμα 6.16: Απεικόνιση του Σπεκτρογραφήματος της Συνθετικής Φωνής και της Εκτιμώμενης Ακολουθίας τοPitch και του 1ου Συντελεστή Mgcep, μετά την Εκτίμησή τους από Μοντέλα Απλών Φωνημάτων

553kB σε επίπεδο φωνημάτων. Έτσι κατανοούμε ότι ένας τέτοιος συνθέτης φωνής είναι αρκετά εύρωστος τόσο από την άποψη καλού αποτελέσματος, όσο και από τη μνήμη που καταλαμβάνει.

Εστιάζοντας στη σύγκριση των αποτελεσμάτων μεταξύ μοντέλων τριφώνων και φωνημάτων, κατανοούμε ότι ποιοτικά με την πρώτη μοντελοποίηση τα αποτελέσματα είναι καλύτερα τόσο από τη άποψη της εκτίμησης των παραμέτρων όσο και από την άποψη της καλύτερης εκτίμησης της διάρκειας καταστάσεων. Παρόλα αυτά θα πρέπει να συνυπολογίσουμε και το μέγεθος του δεύτερου μοντέλου, το οποίο είναι γύρω στις 80 φορές μικρότερο, καθώς και την αποφυγή της έλλειψης των μη υπαρκτών τριφώνων.

Όσον αφορά την εκτίμηση των παραμέτρων χωρίς να συνυπολογίζονται τα δυναμικά χαρακτηριστικά, παρατηρείται ότι το πρόβλημα που προκύπτει είναι απλά η ασυνέχεια των χαρακτηριστικών. Αυτή μπορεί να επιλυθεί και με την μετεπεξεργασία του ήχου, μειώνοντας έτσι στο 1/3 το μέγεθος του μοντέλου μιας και τα δυναμικά χαρακτηριστικά δε χρειάζεται να υπολογίζονται και να αποθηκεύονται.

Κεφάλαιο 7

Συμπεράσματα και Μελλοντική Έρευνα

7.1 Ανακεφαλαίωση

Στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας, η έρευνα που έγινε αφορά τη μελέτη και υλοποίηση ενός συνθέτη φωνής από κείμενο, που βασίζεται σε Κρυφά Μαροκοβιανά Μοντέλα. Αρχικά, μελετήθηκαν 6 εναλλακτικά φωνητικά χαρακτηριστικά, τα οποία θα μπορούσαν να χρησιμοποιηθούν στη συγκεκριμένη έρευνα. Μετά από την αξιολόγηση αυτών ακολούθησε η μελέτη σε θεωρητικό επίπεδο όλων των βασικών τμημάτων ενός τέτοιου συνθέτη, έτσι ώστε να ακολουθήσει η υλοποίησή του. Για τον προγραμματισμό του τελικού συστήματος στην ελληνική γλώσσα, χρησιμοποιήθηκαν υπολογιστικά εργαλεία επεξεργασίας ήχου και φωνής, όπως το SPTK και το Praat καθώς και εργαλεία εφαρμογής και διαχείρισης Κρυφών Μαρκοβιανών Μοντέλων όπως το HTK και το HTS. Τέλος, για την μετεπεξεργασία του ήχου καθώς και για την απεικόνιση των αποτελεσμάτων χρησιμοποιήθηκε το MATLAB. Από αυτή τη διπλωματική προέκυψαν ηχητικά αποτελέσματα και αποτελέσματα εικόνων τα οποία περιλαμβάνονται στη βάση της μαζί με τον κώδικα.

7.2 Ερευνητικές Συνεισφορές

Οι συνεισφορές της συγκεκριμένης μελέτης εστιάζονται σε τρία σημεία, τα οποία περιγράφηκαν αναλυτικά στο αντίστοιχο υποκεφάλαιο της Εισαγωγής.

Όσον αφορά τη μελέτη και την υλοποίηση των διαφορετικών Vocoders, αναπτύχθηκε ένα σύνολο προγραμματιστικών υλοποιήσεων σε MATLAB και σε SPTK, το οποίο εφαρμόστηκε σε συγκεκριμένη εκφώνηση της βάσης εκ-

παίδευσης, ώστε να μπορεί να κριθεί όσο γίνεται πιο αντικειμενικά η ακουστική τους απόδοση. Συγκεκριμένα, εφαρμόστηκε όλη η θεωρητική ανάλυση των 6 ὅσοδερς (LPC-Vocoder[30], Phase-Vocoder[30], Sinusoidal Analysis/Synthesis [25], MFCC-Vocoder [26], MCep Vocoder [11] και MGCep-Vocoder[38]). Κατόπιν, της υπολογιστικής υλοποίησης των ξεχωριστών Vocoders, ακολούθησε η ρύθμιση των παραμέτρων για την καλύτερη ακουστική απόδοση των αποτελεσμάτων. Συνδυάζοντας το ανθρώπινο υποκειμενικό ακουστικό κριτήριο, τη μελέτη των προκύπτοντων σπεκτρογραφημάτων σε σύγκριση με αυτό του αρχικού ήχου, καθώς και την απαίτηση της κάθε διαφορετικής διαδικασίας ανάλυσης σε υπολογιστική μνήμη, έγινε η καταλληλότερη επιλογή χαρακτηριστικών για τη μοντελοποίηση του αποτελέσματος.

Παράλληλα, όσον αφορά την υλοποίηση του ολοκληρωμένου συστήματος, η μελέτη και ο προγραμματισμός έγινε εκ του μηδενός με βάση όλο το θεωρητικό υπόβαθρο [47, 24], όπως περιγράφηκε στα **Κεφάλαια 3,4** και **5**. Συγκεκριμένα, μετά την υλοποίηση της βασικής αρχιτεκτονικής του συστήματος με τη χρήση του HTK και του HTS, η οποία έγινε με χρήση της γλώσσας προγραμματισμού perl, ακολούθησε η διαδικασία ρύθμισης των παραμέτρων του συστήματος. Μέσα από το τελευταίο αυτό ερευνητικό στάδιο, προέκυψαν οι παραμετροποιήσεις όπως ο αριθμός των καταστάσεων των Κρυφών Μαρκοβιανών Μοντέλων, η χρήση της μοντελοποίησης σε πολλαπλούς χώρους, όσον αφορά το pitch κ.α.. Τέλος, για την παραγωγή ζητούμενων ήχων έγινε μελέτη όσον αφορά τις παραμέτρους, που εισάγονται στο τελικό στάδιο του συνθέτη για την εκτίμηση των φωνητικών χαρακτηριστικών.

Τέλος, οι συνθετικοί ήχοι καθώς και οι ακολουθίες των εκτιμώμενων χαρακτηριστικών που προέκυψαν από το σύστημα, αναλύθηκαν και ομαλοποιήθηκαν ώστε να βελτιωθεί αισθητά το ηχητικό αποτέλεσμα. Η ηχητική επεξεργασία που εφαρμόστηκε αφορά τόσο τα τελικά ηχητικά σήματα όσο και τα εκτιμώμενα φωνητικά χαρακτηριστικά (mgceps, pitch), πριν να συντεθούν από το MLSA φίλτρο του Vocoder.

7.3 Προεκτάσεις για Μελλοντική Έρευνα

Ο συνθέτης φωνής από κείμενο που υλοποιήθηκε στα πλαίσια της συγκεκριμένης διπλωματικής αποτελεί μία πρώτη εφαρμογή ενός μεγάλου και αρκετά σύγχρονου επιστημονικού κλάδου. Συνεπώς, υπάρχουν ποικίλλες προεκτάσεις για μελλοντική έρευνα.

7.3.1 Σε επίπεδο Μοντελοποίησης

Όσον αφορά το στάδιο της μοντελοποίησης, στο μέλλον μπορούν να εφαρμοστούν και άλλες παραμετροποιήσεις, με τη διαφοροποίηση του αριθμού των καταστάσεων, με την αλλαγή των χρησιμοποιούμενων χαρακτηριστικών καθώς και με τη χρήση ολοκληρωμένων πινάκων συμμεταβλητότητας και όχι απλά διαγώνιων, ώστε στατιστικά τα χαρακτηριστικά να μη μοντελοποιούνται ανεξάρτητα. Παράλληλα, πολύ βασικό στοιχείο θα ήταν και η χρήση μιας πολύ πιο ανελυμένης γλωσσολογικά βάσης, η οποία θα μπορούσε να βελτιώσει το αποτέλεσμα του συνθέτη, τόσο με την πιο συγκεκριμένη μοντελοποίηση, όσο και με την αξιοποίηση της πληροφορίας για την οργάνωση των τριφώνων-μοντέλων σε δέντρα, ώστε να μπορούν να εκτιμηθούν και αυτά τα οποία δεν εμφανίζονται στη βάση εκπαίδευσης.

7.3.2 Σε επίπεδο Μετεπεξεργασίας της φωνής

Τέτοια συστήματα παρουσιάζουν πολύ σημαντικά πλεονεκτήματα όσον αφορά την προσαρμογή τους σε άλλους ομιλητές, χωρίς την εξ' αρχής επανεκπαίδευσή τους [46, 18, 35]. Παράλληλα, μια πληθώρα τεχνικών επεξεργασίας ήχου μπορεί να εφαρμοστεί για τη βελτίωση των αποτελεσμάτων, αλλά και για τη διαφοροποίηση της προσωδίας όταν το επιβάλλει το κείμενο. Έτσι, ως μελλοντική έρευνα στο πεδίο της σύνθεσης φωνής από κείμενο στα πλαίσια της ελληνικής γλώσσας, περιλαμβάνονται τα πεδία της σθεναρής προσαρμογής σε άλλους ομιλητές, καθώς και της επεξεργασίας των συνθετικών ήχων τόσο για τη βελτίωση της ποιότητας όσο και για τη διαφοροποίηση των χαρακτηριστικών, όπως τη διαφοροποίηση της βασικής συχνότητας, του ρυθμού μεταβολής αυτής, άρα της προσωδίας, του καλύτερου ελέγχου της απαιτούμενης διάρκειας κάθε φωνήματος και πολλών άλλων εφαρμογών.

Βιβλιογραφία

[1] *REFERENCE MANUAL for Speech Signal Processing Toolkit Ver. 3.2*, Νοέμβριος 2008.

[2] O. Abdel-Hamid, S. Abdou και M. Raswan. Improving Arabic HMM Based Speech Synthesis Quality. Στο *Interspeech*, σελίδες 1332–1335, 2006.

[3] J. Allen, S. Hunnicut και D. Klaat. *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge, 1987.

[4] M. Barros, R. Maia, K. Tokuda, D. Freitas και F. Resende Jr. HMM-based European Portuguese Speech Synthesis. Στο *Interspeech*, σελίδες 2581–2584, 2005.

[5] A. Black, H. Zen και K. Tokuda. Statistical Parametric Speech Synthesis. Στο *ICASSP*, σελίδες 1229–1232, 2007.

[6] S. Davis και P. Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:357–366, 1980.

[7] Th. Dutoit και H. Leich. MBR-PSOLA : Text-to-Speech Synthesis Based on an MBE re-Synthesis of the Segments Database. *Speech Communication*, 13(34):167–184, Νοέμβριος 1993.

[8] Th. Dutoit, V. Pagel, F. Bataille και O. Vreken. The MBROLA Project: Towards a High-Quality Speech Synthesizers Free of Use for non-Commercial Purposes. Στο *ICLSP*, τόμος 3, σελίδες 1393–1396, New York, 1996.

[9] D. P. W. Ellis. Sinewave and Sinusoid+Noise Analysis/Synthesis in Matlab, 2003. online web resource.

[10] D. P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab. online web resource, 2005.

[11] T. Fukada, K. Tokuda, T. Kobayashi και S. Imai. An Adaptive Algorithm for Mel-Cepstral Analysis of Speech. Στο *ICASSP-92*, σελίδες I–137–I–140. IEEE, 1992.

[12] J. Holmes, I. Mattingly και J. Shearme. Speech Synthesis by Rule. *Lunguage and Speech*, 7:127–143, 1964.

[13] A. Hunt και A. Black. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. Στο *ICASSP*, σελίδες 373–376, 1996.

[14] Δ. Θεοδωράκης, Χ. Μιναρετζής και Π. Μαραγκός. Συγκριτική Μελέτη και Υλοποίηση Τρόπων Μεταβολής της Βασικής Συχνότητας Pitch σε Ηχητικά Σήματα. Στο *ΣΦΗΜΜΥ*, Απρίλιος 2009.

[15] S. Karabetsos, S. Tsiakoulis, A. Chalamandaris και S. Raptis. *HMM-based Speech Synthesis for the Greek Language*, σελίδες 349–356. Springer, 2008.

[16] H. Kawahara, I. Masuda-Katsuse και A. 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. Στο *Speech Communication*, σελίδες 187–207, 1999.

[17] S. Kim, J. Kim και S. Hahn. Implementation and Evaluation of an HMM-based Korean Speech Synthesis System. Στο *IEICE*, τόμος E89-D, σελίδες 1116–1119, 2006.

[18] S. King, K. Tokuda, H. Zen και J. Yamagishi. Unsupervised Adaptation for HMM-Based Speech Synthesis. Στο *Interspeech*, Σεπτέμβριος 2008.

[19] D. Klaat. Software for a Cascade/Parallel Formant Synthesizer. *Journal of the Acoustical Society of America*, 67:971–995, 1980.

[20] S. Levinson. Continuously Variable Duration Hidden Markov Models for Speech Analysis. Στο *ICASSP*, σελίδες 1241–1244, 1986.

[21] A. Lundgren. An HMM-based Text to Speech Synthesis System Applied to Swedish. Διπλωματική εργασία Master, Royal Institute of Technology (KTH), 2005.

[22] R. Maia, H. Zen, K. Tokuda, T. Kitamura και F. Resende Jr. Towards the Development of a Brazilian Portuguese Text-to-Speech System Based on HMM. Στο *Eurospeech*, σελίδες 2465–2468, 2003.

[23] S. Martinic-Ipsic και I. Ipsic. Croatian HMM-based Speech Synthesis. *Journal of Computing and Information Technology*, 14(4):307–313, 2006.

[24] T. Masuko. *HMM-based Speech Synthesis and Its Applications*. Διδακτορική Διατριβή , Department of Electrical and Computer Engineering Nagoya Institute of Technology, Νοέμβριος 2002.

[25] R. McAulay και T. Quatieri. Speech Analysis/Synthesis Based on a Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-34(4):744–754, Αύγουστος 1986.

[26] B. Milner και X. Shao. Speech Reconstruction from Mel-Frequency Cepstral Coefficients using a Source-Filter Model. Στο *ICLSP*, σελίδες 2421–2424, Denver, CO, USA, 2002.

[27] B. Milner και X. Shao. Clean Speech Reconstruction from MFCC Vectors and Fundamental Frequency using an Integrated Front-End. *Speech Communication*, 48:697–715, Οκτώβριος 2005.

[28] T. Ojala. Auditory Quality Evaluation of Present Finish Text-to-Speech Systems. Διπλωματική εργασία Master, Helsinky University of Technology, 2006.

[29] Y. Qian, F. Soong, Y. Chen και M. Chu. An HMM-based Mandarin Chinese Text-to-Speech System. Στο *ISCSLP*, 2006.

[30] T. Quatieri. *Speech Signal Processing*. Prentice Hall, Upper Saddle River, NJ, 2002.

[31] L. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEE*, 77(2):257–286, Φεβρουάριος 1989.

[32] L. Rabiner και B. H. Juang. *Fundamental of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[33] L. Rabiner και R. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, Englewood Cliffs, NJ, 1978.

[34] L. Rabiner και R. Schafer. *Theory and Applications of Digital Speech Processing*. Pearson, Upper Saddle River, NJ, φιρστη έκδοση, 2011.

[35] M. Tamura, T. Masuko, K. Tokuda και T. Kobayashi. Adaptation of Pitch and Spectrum for HMM-based Speech Synthesis Using MLLR. Στο *ICASSP*, σελίδες 805–808, 2001.

[36] P. Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[37] K. Tokuda, T. Kobayashi και S. Imai. Speech Parameter Generation from HMM Using Dynamic Features. Στο *ICASSP*, σελίδες 660–663, 1995.

[38] K. Tokuda, T. Kobayashi, T. Masuko και S. Imai. Mel-Generalized Cepstral Analysis- A Unified Apprach to Speech Spectral Estimation. Στο *ICLSP*, σελίδες 1043–1046, Σεπτέμβριος 1994.

[39] K. Tokuda, T. Masuko, N. Miyazaki και T. Kobayashi. Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling. Στο *ICASSP*, τόμος 1, σελίδες 229–232, Μάρτιος 1999.

[40] K. Tokuda, T. Masuko, N. Miyazaki και T. Kobayashi. Multi-Space Probability Distribution. *IEICE Transactions Information and Systems*, E85-D(3):455–464, Μάρτιος 2002.

[41] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi και T. Kitamura. Speech Parameter Generation Algorithms for HMM-based Speech Synthesis. Στο *ICASSP*, τόμος 3, σελίδες 1315–1318, Ιούνιος 2000.

[42] K. Tokuda, H. Zen και A. Black. An HMM-based Speech Synthesis System Applied to English. Στο *SSW*. IEEE, Σεπτέμβριος 2002.

[43] B. Vesnicer και F. Mihelic. Evaluation of the Slovenian HMM-based Speech Synthesis System. Στο *TSD*, σελίδες 513–520, 2004.

[44] C. Weiss, R. Maia, K. Tokuda και W. Hess. Low Resource HMM-based Speech Synthesis Applied to German. Στο *ESSP*, 2005.

[45] J. Wouters και M. Macon. A Perceptual Evaluation of Distance Measures for Concatenative Speech Synthesis. Στο *ICLSP*, σελίδες 2747–2750, 1998.

[46] J. Yamagishi, T. Nose, H. Zen, Z. H. Ling, T. Toda, K. Tokuda, S. King και S. Renals. Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 17(6):1208–1230, Αύγουστος 2009.

[47] T. Yoshimura. *Simultaneous Modeling of Phonetic and Prosodic Parameters, and Characteristic Conversion for HMM-based Text-to-Speech Systems*. Διδακτορική Διατριβή , Department of Electrical and Computer Engineering Nagoya Institute of Technology, Ιανουάριος 2002.

[48] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi και T. Kitamura. Duration Modeling for HMM-based Speech Synthesis. Στο *ICLSP*, σελίδες 29–32, Δεκέμβριος 1998.

[49] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi και T. Kitamura. Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-based Speech Synthesis. Στο *Eurospeech*, τόμος 5, σελίδες 2347–2350, 1999.

[50] S. Young, G. Evermann, M. Gale, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtachev και P. Woodland. *The HTK Book (for HTK Version 3.4)*, 9η έκδοση, 2009.

[51] Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black και K. Tokuda. The HMM-based Speech Synthesis System (HTS) Version 2.0. Στο *6th ISCA Workshop on Speech Synthesis*, σελίδες 294–299, Bonn, Germany, Αύγουστος 2007.

[52] H. Zen, J. Lu, J. Ni, K. Tokuda και H. Kawai. HMM Based Prosody Modeling and Synthesis for Japanese and Chinese Speech Synthesis. Τεχνική Αναφορά υπ. αρίθμ. TP-ΣΛΤ-0032, ATR-SLT, 2003.

[53] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi και T. Kitamura. Hidden semi-Markov Models Based Speech Synthesis. Στο *ICLSP*, σελίδες 1180–1185, 2004.