Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Οπτικοακουστική Σύνθεση Φωνής με Χρήση Κρυφών Μαρκοβιανών Μοντέλων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΦΙΛΝΤΙΣΗΣ ΠΑΝΑΓΙΩΤΗΣ ΠΑΡΑΣΚΕΥΑΣ

Επιβλέπων : Μαραγκός Πέτρος
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2015

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Οπτικοακουστική Σύνθεση Φωνής με Χρήση Κρυφών Μαρκοβιανών Μοντέλων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΦΙΛΝΤΙΣΗΣ ΠΑΝΑΓΙΩΤΗΣ ΠΑΡΑΣΚΕΥΑΣ

Επιβλέπων : Μαραγκός Πέτρος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30η Οκτωβρίου 2015.

Μαραγκός Πέτρος Καθηγητής Ε.Μ.Π.	Ποταμιάνος Αλέξανδρος Αναπ. Καθηγητής Ε.Μ.Π.	Πρωτόπαπας Αθανάσιος Αναπ. Καθηγητής Ε.Κ.Π.Α.

Αθήνα, Οκτώβριος 2015

Φιλντίσης Παναγιώτης Παρασκευάς

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στην παρούσα διπλωματική εργασία παρουσιάζεται ένα πλήρες οπτικοακουστικό σύστημα σύνθεσης φωνής για την Ελληνική γλώσσα. Κατά την υλοποίηση ενός τέτοιου συστήματος αντλούνται τεχνικές από διάφορους επιστημονικούς τομείς όπως η Μηχανική Μάθηση, η Επεξεργασία Σημάτων, και η Όραση Υπολογιστών. Εκκινώντας με την εισαγωγή, παρουσιάζουμε την ιστορική αναδρομή και τις σημαντικότερες μεθόδους για την υλοποίηση ενός οπτικοακουστικού συνθέτη φωνής. Εν συνεχεία, στα επόμενα κεφάλαια παρουσιάζεται η απαραίτητη θεωρητική ανάλυση για την υλοποίηση του οπτικοακουστικού συστήματος σύνθεσης φωνής, παράλληλα με τα πειραματικά αποτελέσματα που λήφθηκαν κατά την υλοποίηση και αξιολόγηση του συστήματος. Η αξιολόγηση του συστήματος είναι ιδιαίτερα ενθαρρυντική τόσο για την παραγόμενη ομιλία, όσο και για την παραγόμενη εικονοσειρά, ανοίγοντας διάπλατα τον δρόμο για την μετέπειτα εξέλιξη του συστήματος σε εφαρμογές όπως η συναισθηματική οπτικοακουστική σύνθεσης φωνής, μια πρώτη προσέγγιση και αξιολόγηση της οποίας κάνουμε στο τελευταίο Κεφάλαιο.

Λέξεις κλειδιά

aam, active appearance model, hidden markov model, mel generalized cepstrum, hmm-based speech synthesis, οπτικοακουστική σύνθεση φωνής, παραμετρική μοντελοποίηση χαρακτηριστικών προσώπου, σύνθεση φωνής, συναισθηματική οπτικοακουστική σύνθεση φωνής,

Abstract

In the present diploma thesis, we present a complete audiovisual text-to-speech synthesis system for the Greek language. During the implementation of such a system, we draw tools from a variety of scientific fields, such as Machine Learning, Signal Processing and Computer Vision. Starting with the introduction, we present the history and most important methods for the implementation of an audiovisual text-to-speech synthesis system. In the next chapters we present the necessary theoretical analysis for the implementation of the system, and at the same time we present our experimental results and evaluation. The evaluation of the system appears especially encouraging both for the synthetic speech and video, opening the way for the evolution of our system for applications such as emotional and expressive speech synthesis, on which we do a first approach and evaluation in the last Chapter.

Key words

aam, active appearance model, hidden markov model, mel generalized cepstrum, expressive audio-visual speech synthesis,hmm-based speech synthesis, parametric modelling of facial features speech synthesis, audiovisual speech synthesis

Ευχαριστίες

Στην παρούσα σελίδα, θα ήθελα να ευχαριστήσω τον καθηγητή Κ. Μαραγκό Πέτρο, για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα εξαιρετικά ανερχόμενο, και ενδιαφέρον αντικείμενο, το οποίο αντλεί την ισχύ του από διάφορους ερευνητικούς τομείς, το ενδιαφέρον μου για τους οποίους δημιουργήθηκε κατά την παρακολούθηση των μαθημάτων που διδάσκει στο Εθνικό Μετσόβιο Πολυτεχνείο.

Παράλληλα, θα ήθελα να ευχαριστήσω το Νάσο Κατσαμάνη, χωρίς την αμέριστη βοήθεια του οποίου η παρούσα διπλωματική δεν θα μπορούσε να είχε τελειώσει. Ήταν εκεί για κάθε εμπόδιο που συναντούσα, έτοιμος όχι να μου δώσει έτοιμες λύσεις, αλλά κυρίως να μου δώσει τα εναύσματα αυτά και την καθοδήγηση που θα καθιστούσαν την υπέρβαση του εμποδίου μια ευχάριστη και ικανοποιητική διαδικασία.

Θα ήθελα επίσης να ευχαριστήσω εκ βαθέων καρδίας την Ιωάννα Πατσά, η οποία επί συνεχόμενες ημέρες ερχόταν στο studio στο ερευνητικό κέντρο Αθηνά, για να δημιουργήσουμε τη βάση δεδομένων, μιλώντας ακατάπαυστα έως ότου να κλείσει η φωνής της.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου, και κυρίως τον αδερφό μου Αναστάσιο, για την υποστήριξη του όλο αυτόν τον καιρό και την ανοχή που μου έχει δείξει και συνεχίζει να μου δείχνει.

Φιλντίσης Παναγιώτης Παρασκευάς,

Αθήνα, 30η Οκτωβρίου 2015

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Στο σύγχρονο κόσμο, οι ηλεκτρονικοί υπολογιστές έχουν γίνει αναπόσπαστο κομμάτι της καθημερινότητάς μας, αλλά ταυτόχρονα και το σημαντικότερο μέρος αφανών διαδικασιών που επηρεάζουν έμμεσα, αλλά σε μεγάλο βαθμό, τη ζωή μας. Η καθημερινή πρόοδος και οι νέες εφαρμογές στον τομέα αυτό, μας επιφυλάσσουν στο μέλλον ευρήματα, τα οποία όπως και τα σημερινά μέχρι πρότινος βρίσκονταν μόνο στη σφαίρα της φαντασίας.

Στο πλαίσιο αυτό η παρούσα διπλωματική εργασία στοχεύει να συμβάλλει στις καθημερινές αλληλεπιδράσεις του ανθρώπου με τους ηλεκτρονικούς υπολογιστές εστιάζοντας στη σύνθεση φωνής και εικονοσειράς, με απώτερο στόχο τη μεγιστοποίηση της φυσικότητας των αλληλεπιδράσεων. Η ανθρώπινη επικοινωνία αποτέλεσε το θεμέλιο λίθο για την οικοδόμηση των ανθρώπινων κοινωνιών και την πρόοδο η οποία αποτυπώνεται στο πέρασμα των αιώνων. Σήμερα, με τη μετάβαση στην ψηφιακή εποχή, ο φιλόδοξος πλην σπουδαίος στόχος της τελειοποίησης της επικοινωνίας ανθρώπου και υπολογιστή δύναται να οδηγήσει σε μια σειρά εφαρμογών που θα βελτιώσουν πολλές πλευρές του ανθρώπινου βίου [10, 66].

Ένα σύστημα οπτικοακουστικής σύνθεσης φωνής, δύναται να διευκολύνει τις διεπαφές (interfaces) ανάμεσα σε άνθρωπο και υπολογιστή, γεγονός ιδιαίτερα σημαντικό σε διάφορους τομείς, αναφέροντας ενδεικτικά τους κάτωθι:

- **Ιατρική**: Παρέχεται η δυνατότητα σε άτομα με πρόβλημα ακοής να αντιλαμβάνονται την ομιλία, μέσω των κινήσεων των χειλιών και του προσώπου, διευκολύνοντας τη χρήση ηλεκτρονικών συσκευών και την πλήρη ενσωμάτωσή τους στο σύγχρονο κόσμο της τεχνολογίας. Επιπλέον δύναται να αποτελέσει σημαντικό εργαλείο στα χέρια των λογοθεραπευτών, καθώς οι κινήσεις των χειλιών θα βελτιώσουν την ικανότητες εκμάθησης των ασθενών.
- **Οικονομία**: Λαμβάνοντας υπ' όψιν την ευρεία χρήση robots σε εργοστάσια και επιχειρήσεις η οποία επεκτείνεται με ιλιγγιώδη ρυθμό, η οπτικοακουστική σύνθεση φωνής δύναται να συμβάλλει στο επίπεδο διαχείρισης των απαιτούμενων εργασιών.
- **Καθημερινότητα**: Με την συνεχή ανάπτυξη έξυπνων συσκευών, η σύνθεση φωνής και εικονοσειράς μπορεί να βρει εφαρμογή στο γραφικό περιβάλλον χρήστη(GUI) εστιάζοντας στην ανθρώπινη προσωμοίωση για τη διαβίβαση εντολών σε αυτές.
- **Ψυχαγωγία**: Διευκολύνεται η ανάπτυξη ψηφιακών ηθοποιών και τη βελτίωση του computer animation αλλάζοντας το χάρτη της ψυχαγωγίας.

Στο παρόν πρώτο εισαγωγικό Κεφάλαιο, θα προχωρήσουμε σε μια εισαγωγή αρχικά στη Σύνθεση Φωνής, και στη συνέχεια στην Οπτικοακουστική Σύνθεση Φωνής, αναφέροντας την ιστορική τους αναδρομή, καθώς και αναλύοντας τις σημαντικότερες μεθόδους που χρησιμοποιούνται για την υλοποίηση τους. Στο τέλος, θα δηλώσουμε τους στόχους, τις συνεισφορές και την οργάνωση της διπλωματικής εργασίας, καθώς και θα περιγράψουμε την βάση δεδομένων που υλοποιήθηκε στα πλαίσια της εργασίας.

1.1 Σύνθεση Φωνής

Ο όρος *Σύνθεση Φωνής* αναφέρεται στην τεχνητή παραγωγή ανθρώπινης φωνής. Η ομιλία αποτελεί το κυριότερο μέσο επικοινωνίας μεταξύ των ανθρώπων και γι' αυτό η σύνθεση φωνής έχει αποτελέσει αντικείμενο μελέτης εδώ και αρκετούς αιώνες, με τις πιο ραγδαίες εξελίξεις να έχουν λάβει χώρα τις τελευταίες δεκαετίες. Ένα σύστημα σύνθεσης φωνής από κείμενο (Text-To-Speech-Synthesizer) υλοποιείται μέσω ηλεκτρονικών υπολογιστών και αναλαμβάνει τη μετατροπή της εισόδου, δηλαδή του κειμένου, σε λόγο. Σύμφωνα με τον Dutoit [32] ένα σύστημα σύνθεσης φωνής πρέπει να διαθέτει τα εξής χαρακτηριστικά:

- Αποτελεσματική επεξεργασία οποιουδήποτε κειμένου εισόδου,
- Παραγωγή κατανοητής ομιλίας, και
- Παραγωγή ομιλίας με φυσικότητα.

Ειδικά η αξιολόγηση του τελευταίου χαρακτηριστικού αποτελεί σχετικά δύσκολο εγχείρημα και βασίζεται κυρίως στον ανθρώπινο υποκειμενικό παράγοντα όπως θα δούμε κατά την αξιολόγηση του τελικού συστήματος στο Κεφάλαιο 4.

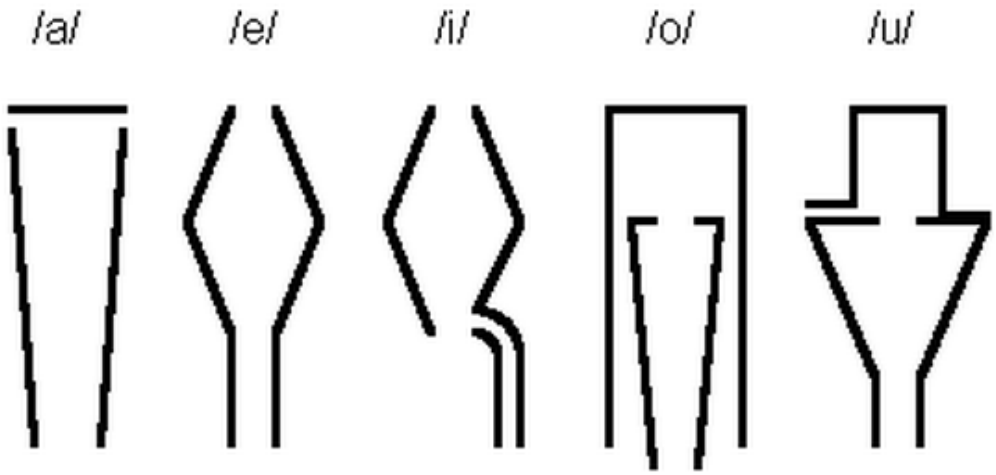
1.1.1 Ιστορική Αναδρομή

Παρ' όλο που ο κλάδος της τεχνητής παραγωγής ομιλίας αναπτύχθηκε ραγδαίως μόλις τις τελευταίες δεκαετίες, η συνθετική ομιλία υπήρξε στόχος των ανθρώπων ήδη από τις αρχές του 11ου αιώνα. Οι πρώτες προσπάθειες βεβαίως ήταν πλήρως μηχανικού χαρακτήρα μέχρι που κατασκευάστηκε ο πρώτος ηλεκτρονικός συνθέτης.

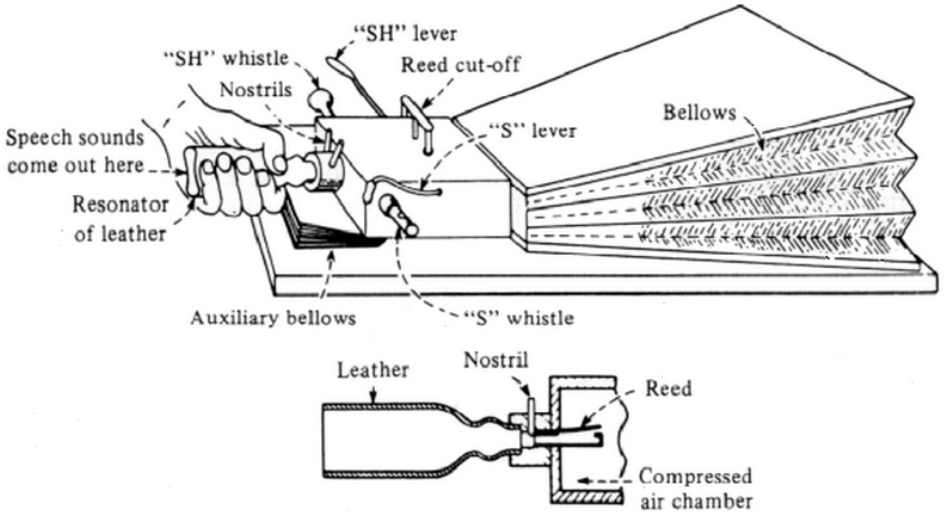
Οι πρώτες καταγεγραμμένες προσπάθειες έγιναν πριν από τουλάχιστον 2 αιώνες [36, 37, 73, 52], όταν ο Δανός καθηγητής Christian Kratzenstein το 1779, ενώ εργαζόταν στην Ρωσική Ακαδημία Επιστημών, εξήγησε τις φυσιολογικές (physiological) διαφορές μεταξύ των πέντε μακρών φωνηέντων (/a/, /e/, /i/, /o/, and /u/) και έχτισε μοντέλα για την ανθρώπινη φωνητική οδό. Κατασκεύασε ακουστικούς συντονιστές η δομή των οποίων φαίνεται στο Σχήμα 1.1. Οι ταλαντωτές αυτοί προσομοίαζαν την ανθρώπινη φωνητική οδό και ενεργοποιούνταν με τη χρήση δονούμενων ελασμάτων όπως στα μουσικά όργανα. Συγκεκριμένα το φωνήεν /i/ παραγόταν όχι με τη χρήση ελάσματος όπως τα άλλα φωνήεντα αλλά με φύσημα στην κάτω οπή του αντίστοιχου οργάνου, το οποίο ομοιάζει με φλογέρα.

Μερικά χρόνια αργότερα, το 1791, στη Βιέννη ο Wolfgan von Kempelen παρουσιάσε την Ακουστική-Μηχανική Μηχανή Ομιλίας ("Acoustic-Mechanical Speech Machine"), που ήταν ικανή να παράγει όχι μόνο ήχους αλλά και λέξεις και μικρές προτάσεις [47, 73]. Η κατασκευή της μηχανής είχε ξεκινήσει ήδη από το 1769. Τα κύρια μέρη της μηχανής αυτής ήταν ένα δωμάτιο πίεσης που προσομοιάζει τους πνεύμονες, ένα δονούμενο έλασμα για τις φωνητικές χορδές, και ένας δερμάτινος σωλήνας για την φωνητική οδό. Τα διαφορετικά φωνήεντα παράγονταν με το χειρισμό του σχήματος του δερμάτινου σωλήνα. Τα σύμφωνα παράγονταν με τη χρήση τεσσάρων ξεχωριστών περιορισμένων οδών των οποίων ο χειρισμός γινόταν με τα δάχτυλα. Με τη χρήση της μηχανής αυτής ο von Kempelen κατασκεύασε επίσης μία ομιλούσα μηχανή που έπαιζε σκάκι (The Turk).

Με βάση την μηχανή του von Kempelen, το 1837 ο Άγγλος Charles Wheatstone κατασκεύασε τη δικιά του μηχανή παραγωγής που φαίνεται στο Σχήμα 1.2. Η μηχανή αυτή ήταν πιο πολύπλοκη και ήταν ικανή να παράγει φωνήεντα, καθώς και τα περισσότερα σύμφωνα και επομένως ήταν δυνατή η παραγωγή μερικών λέξεων. Τα φωνήεντα παράγονταν με τη χρήση ενός δονούμενου ελάσματος με τις ξεχωριστές οδούς κλειστές, ενώ τα σύμφωνα, συμπεριλαμβανομένων των ένρινων, παράγονταν με τη χρήση στροβιλώδους ροής διαμέσου μιας από τις ξεχωριστές οδούς με το έλασμα ανενεργό.



Σχήμα 1.1: Οι συντονιστές του Kratzenstein ([73]).



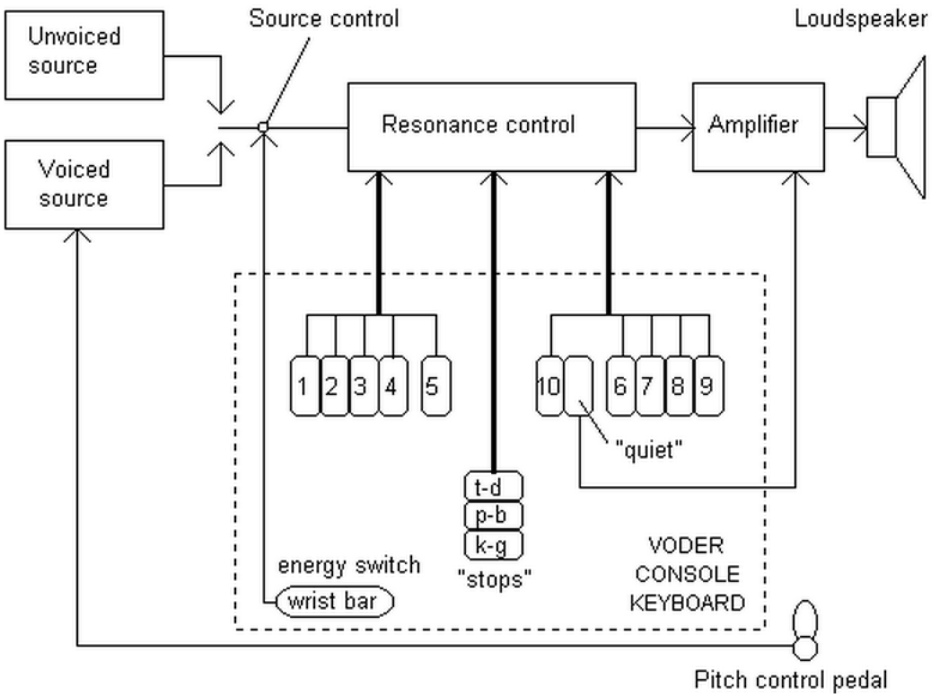
Σχήμα 1.2: Η μηχανή παραγωγής φωνής του Wheatstone ([36]).

Η μηχανή του Wheatstone επίσης χρησιμοποιήθηκε το 1923 από τον Paget ο οποίος κατάφερε να παράξει μερικές μεμονωμένες λέξεις [57].

Η πρώτη πλήρως ηλεκτρική συσκευή σύνθεσης παρουσιάστηκε από τον Stewart το 1922. Η μηχανή ήταν ικανή να παράγει φωνήεντα με τη χρήση των δύο κατώτερων formants, αλλά όχι σύμφωνα ή φράσεις.

Το μεγάλο ενδιάφερον του επιστημονικού κόσμου για την τεχνητή παραγωγή φωνής αναπτύχθηκε όταν το 1939 ο Homer Dudley παρουσίασε στη διεθνή έκθεση της Νέας Υόρκης τον VODER [52] (Voice Operator DEmonstratoR) - μία ηλεκτρονική μηχανή παραγωγής ομιλίας που βασιζόταν στο VOCODER (VOice CODeR) που αναπτύχθηκε στα μέσα της δεκαετίας του 1930 στα Bell Laboratories. Το VOCODER αποτελούσε μία συσκεύη που ήταν ικανή να αναλύει το λόγο σε διάφορες ακουστικές παραμέτρους που στη συνέχεια μπορούσαν να τροφοδοτηθούν σε ένα synthesizer ώστε να παράγουν προσεγγιστικά το αρχικό σήμα ομιλίας. Η ποιότητα και καταληπτότητα της ομιλίας ήταν πάρα πολύ χαμηλή, όμως επιδείχθηκαν οι μεγάλες δυνατότητες για την τεχνητή παραγωγή

λόγου. Η δομή του VODER φαίνεται στο Σχήμα 1.3.



Σχήμα 1.3: Δομή του VODER ([47])

Ακολουθώντας την μηχανική παραγωγής ομιλίας του Dudley, το 1951 ο Franklin Cooper και οι συνεργάτες τους ανέπτυξαν ένα synthesizer στα Εργαστήρια Haskins [47], [37], που μετέτρεπε ηχογραφημένα μοτίβα φασματογραφημάτων σε ήχους. Το πρώτο synthesizer με χρήση formants, PAT (Parametric Artificial Talker) υλοποιήθηκε από τον Walter Lawrence το 1953 ενώ το πρωτο synthesizer με χρήση αρθρωτών παρουσιάστηκε από τον George Rosen στο Massachusetts Institute of Technology (M.I.T.) [47].

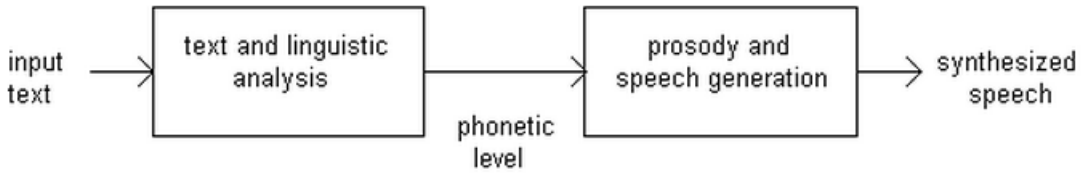
Το πρώτο ολοκληρωμένο σύστημα μετατροπής κειμένου σε ομιλία για την Αγγλική γλώσσα αναπτύχθηκε στο Electrotechical Laboratory, στην Ιαπωνία το 1968, από τον Noriko Umeda και τους συνεργάτες του [47], βασισμένο σε μοντέλο με αρθρωτές. Ο ήχος ήταν αρκετά κατανοητός αλλά μονότονος.

Στα τέλη της δεκαετίας των 1970 και την δεκαετία του 1980, άρχισαν να κυκλοφορούν τα πρώτα εμπορικά συστήματα σύνθεσης φωνής, με πρώτο το Votrax chip.

Από την δεκαετία του 1980 και μετά, με την αύξηση της υπολογιστικής ισχύος, αλλά και της δυνατότητας αποθήκευσης μεγάλου όγκου δεδομένων, η επιστημονική κοινότητα στράφηκε προς μεθόδους που χρησιμοποιούν μεγάλο όγκο δεδομένων (data-driven) [24], αλλά και σε παραμετρικές μεθόδους σύνθεσης όπως η σύνθεση με κρυφά μαρκοβιανά μοντέλα [81], που θα αναλύσουμε περισσότερο στη συνέχεια.

1.1.2 Δομή Συστήματος Σύνθεσης Φωνής

Στη γενική περίπτωση, ένα σύστημα σύνθεσης ομιλίας αποτελείται από δύο διακριτά υπόσυστήματα όπως βλέπουμε και στο Σχήμα 1.4. Το πρώτο σύστημα (front-end) αποτελεί το *Σύστημα Επεξεργασίας της Φυσικής Γλώσσας* και το δεύτερο αναλαμβάνει την *Ψηφιακή Επεξεργασία του Σήματος*.



Σχήμα 1.4: Αρχιτεκτονική Συστήματος Σύνθεσης Ομιλίας [52]

Το σύστημα στο πρώτο μέρος ενός συνθέτη ομιλίας αναλαμβάνει την λεξιλογική ανάλυση του κειμένου εισόδου, καθώς και τη μετατροπή του σε μια ενδιάμεση αναπαράσταση την οποία προωθεί στο σύστημα της Ψηφιακής Επεξεργασίας Σήματος. Το σύστημα Επεξεργασίας της Φυσικής Γλώσσας εξάγει σημαντικές πληροφορίες από την ανάλυση του κειμένου όπως τα προσωδιακά χαρακτηριστικά του κειμένου, τα φωνήματα από τα οποία αποτελείται κ.ά.

Οι πληροφορίες αυτές εν συνεχεία εισέρχονται στο σύστημα της Ψηφιακής Επεξεργασίας του Σήματος που αναλαμβάνει την επεξεργασία των πληροφοριών. Το σύστημα στο δεύτερο μέρος ενός συνθέτη ομιλίας αναλαμβάνει την επεξεργασία των πληροφοριών που λαμβάνει στην ενδιάμεση αναπαράσταση που προαναφέραμε από το πρώτο σύστημα έτσι ώστε να παράγει την συνθετική φωνή. Υπάρχουν αρκετές μέθοδοι για την επεξεργασία αυτών των πληροφοριών τις σημαντικότερες από τις οποίες και θα δούμε τώρα.

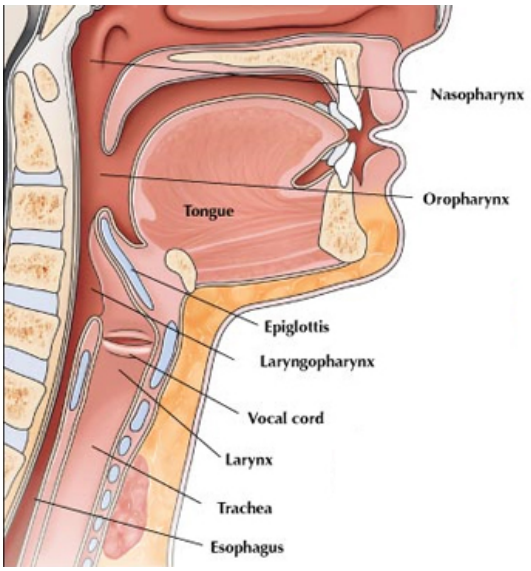
1.1.3 Μέθοδοι Παραγωγής Συνθετικής Φωνής

Παραμετρικές Μέθοδοι Σύνθεσης Φωνής

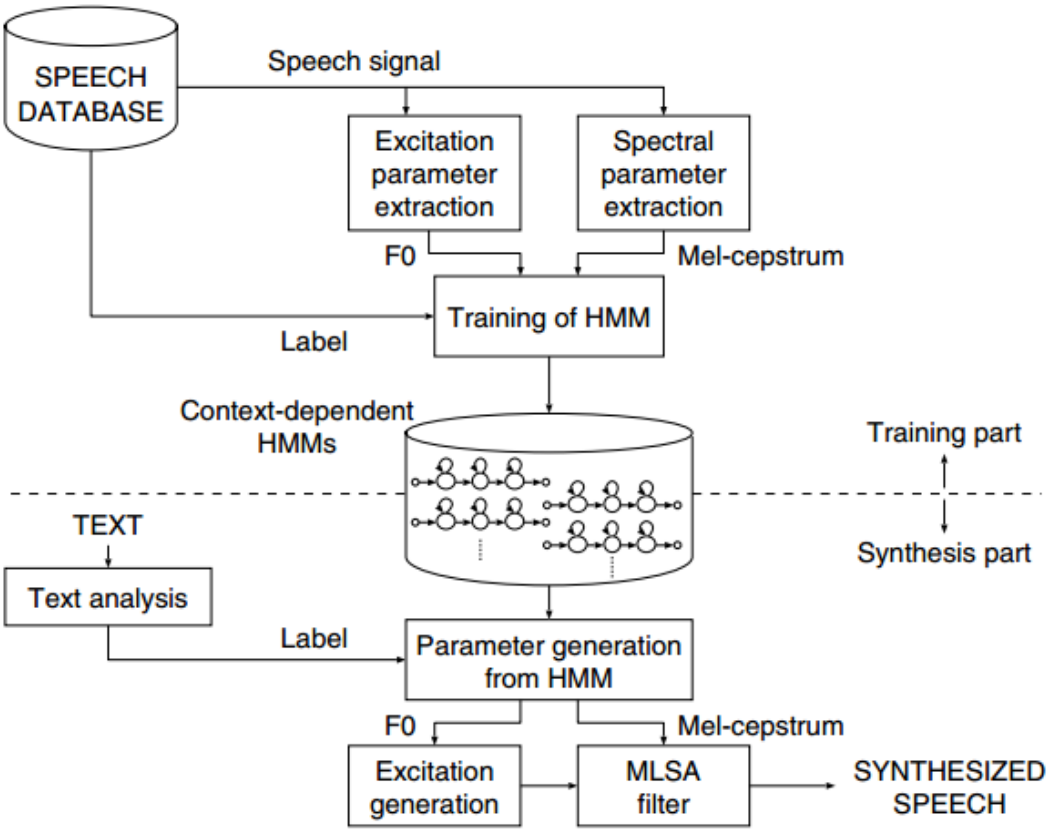
Σύνθεση με μοντελοποίηση αρθρωτών. Τα συστήματα που χρησιμοποιούν την μοντελοποίηση με αρθρωτές προσπαθούν να προσομοιώσουν την ανθρώπινη παραγωγή φωνής με τη χρήση υποσυστημάτων που προσομοιάζουν τα διάφορα όργανα του ανθρώπινου σώματος που λαμβάνουν μέρος στην παραγωγή φωνής (γλώσσα, στοματική κοιλότητα, φωνητικές χορδές κ.τ.λ.) (Σχήμα 1.5). Τέτοια συστήματα ήταν και τα πρώτα συστήματα που χρησιμοποιήθηκαν για την παραγωγή φωνής από τους Kempelen και Kratzenstein. Τα συστήματα αυτά παρουσιάζουν μεγάλη δυσκολία προς την υλοποίηση τους ενώ η υπολογιστική πολυπλοκότητα τους είναι σημαντικά μεγαλύτερη σε σχέση με τις υπόλοιπες μεθόδους παραγωγής ομιλίας [49].

Σύνθεση με κανόνες. Τα σύστηματα αυτά εμπεριέχουν την ανθρώπινη φωνητική οδό, ως ένα μαύρο κουτί του οποίου τα χαρακτηριστικά προσπαθούν να αναπαράγουν με χρήση συνήθως formant φίλτρων. Χωρίζονται σε δύο κατηγορίες, παράλληλης σύνδεσης και διαδοχικής σύνδεσης, αλλά συνήθως χρησιμοποιείται συνδυασμός των δύο. Ένα τέτοιο σύστημα χρησιμοποιεί ένα σύνολο κανόνων για την εύρεση των παραμέτρων για τη σύνθεση της επιθυμητής φωνής μέσω ενος formant συνθέτη [41].

Σύνθεση με χρήση κρυφών μαρκοβιανών μοντέλων. Τα συστήματα αυτά ανήκουν στην κατηγορία των παραμετρικών συστημάτων, όπου οι ακουστικές παράμετροι μοντελοποιούνται με χρήση στοχαστικών, παραγωγικών μοντέλων [81]. Η απόδοση της συγκεκριμένης τεχνικής έγινε εμφανής το 2005 στο Blizzard Challenge από το σύστημα σύνθεσης φωνής HTS που δημιουργήθηκε στο Nagoya Institute of Technology. Η δομή ενός συστήματος σύνθεσης φωνής με χρήση κρυφών μαρκοβιανών μοντέλων φαίνεται στο Σχήμα 1.6.



Σχήμα 1.5: Το ανθρώπινο σύστημα παραγωγής φωνής. *Πηγή:* `http://www.babelsdawn.com/babels_dawn/speech_organs/`.



Σχήμα 1.6: Δομή Συστήματος Σύνθεσης Φωνής με χρήση Κρυφών Μαρκοβιανών Μοντέλων [86].

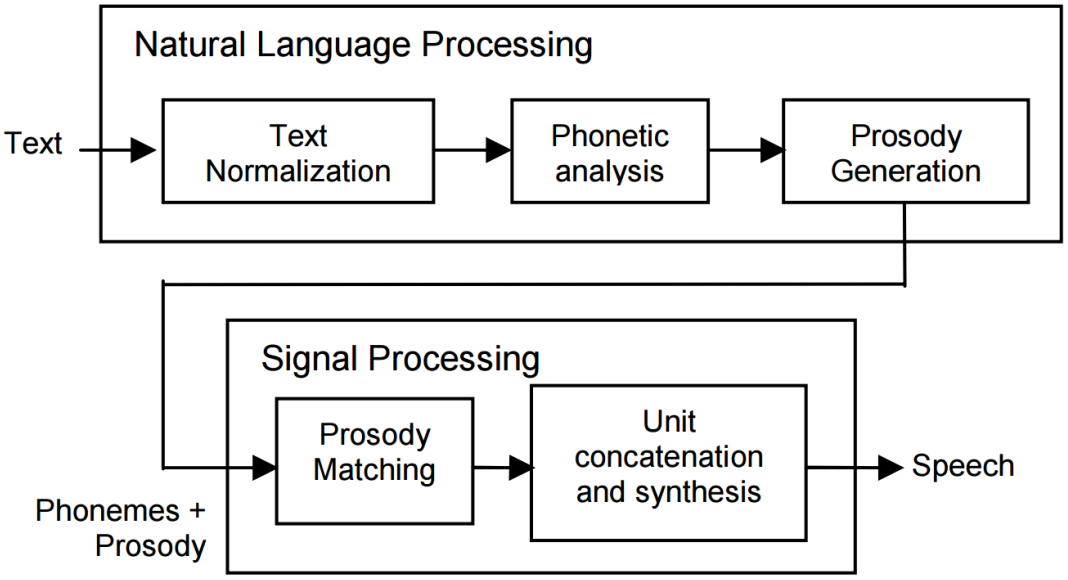
Όπως βλέπουμε στο Σχήμα, το σύστημα εκπαιδεύεται με τη φασματική και προσωδιακή πληροφορία από τη βάση δεδομένων, που είναι επισημειωμένη σε επίπεδο φωνήματος. Στα συγκεκριμένα συστήματα λαμβάνεται επίσης υπ' όψιν το περιβάλλον των φωνημάτων, για μεγαλύτερη ποιότητα της

συνθετικής φωνής. Στο στάδιο της σύνθεσης, μετά από την λεξιλογική ανάλυση της εισόδου, παράγονται οι παράμετροι του τελικού συστήματος σύνθεσης φωνής. Το συγκεκριμένο σύστημα σύνθεσης φωνής έχει τύχει μεγάλης αποδοχής στην επιστημονική κοινότητα, και έχει εφαρμοσθεί για πολλές γλώσσες όπως: Αγγλικα [82], Σουηδικά [54], Αραβικά [5], και Γερμανικά [50] καθώς και για την Ελληνική Γλώσσα στο [87]. Το σύστημα μας επίσης χρησιμοποιεί την συγκεκριμένη τεχνολογία και λεπτομερής ανάλυση του θα γίνει στα επόμενα κεφάλαια.

Data-driven Μέθοδοι Σύνθεσης Φωνής

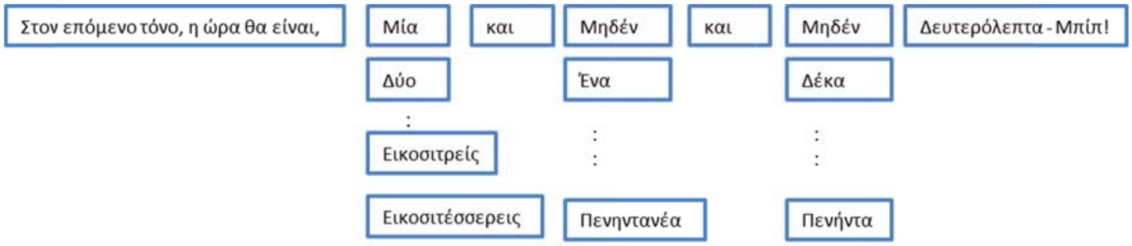
Σύνθεση με παράθεση ακουστικών μονάδων. Τα συστήματα παράθεσης ακουστικών μονάδων απαιτούν τη χρήση μιας μεγάλης βάσης δεδομένων προηχογραφημένης ομιλίας. Η σύνθεση φωνής σε αυτά τα συστήματα γίνεται με βάση την παράθεση και σύνδεση ακουστικών μονάδων για την δημιουργία της επιθυμητής πρότασης. Στα σύγχρονα συστήματα χρησιμοποιούνται ως ακουστικές μονάδες σύνήθως λέξεις, συλλαβές, φωνήματα, δι-φωνήματα κ.τ.λ. Συνήθως, όσο μεγαλύτερο είναι το μέγεθος της ακουστικής μονάδας, τόσο μεγαλύτερη είναι η φυσικότητα της παραγόμενης ομιλίας, και αντίστοιχα μεγαλύτερο είναι το μέγεθος των απαιτούμενων μονάδων και της βάσης δεδομένων. Μεγάλη απήχηση βρίσκουν επίσης τα συστήματα με σύνθεσης με την επιλογή της βέλτιστης ακουστικής μονάδας, στα οποία η επιλογή των μονάδων γίνεται με μεθόδους δυναμικού προγραμματισμού [88, 25].

Η δομή ενός συστήματος σύνθεσης με παράθεση ακουστικών μονάδων φαίνεται στο Σχήμα 1.7.



Σχήμα 1.7: Δομή Συστήματος Σύνθεσης Φωνής με παράθεση ακουστικών μονάδων [34].

Τα συστήματα σύνθεσης με παράθεση ακουστικών μονάδων έχουν την ικανότητα παραγωγής συνθετικής φωνής μεγάλης φυσικότητας, αλλά χαρακτηρίζονται από την μικρή τους ευελιξία, καθώς και τον μεγάλο όγκο δεδομένων που απαιτούν. Περιορίζονται επίσης στη συνθετική φωνή ενός ομιλητή, και στην ίδια έκφραση, καθώς γίνονται μικρές αλλαγές στα επιλεγμένα τμήματα προηχογραφημένης φωνής. Αν επιθυμείται η παραγωγή ομιλίας με διαφορετικούς τρόπους ομιλίας, ή διάφορα συναισθήματα, το μέγεθος των βάσεων αυξάνεται εκθετικά, και η ηχογράφηση των δεδομένων είναι σημαντικά επίπονη και χρονοβόρα διαδικασία [19].



Σχήμα 1.8: Παράδειγμα σύνθεσης φωνής με παράθεση ακουστικών μονάδων για την εκφώνηση ώρας [88].

1.2 Οπτικοακουστική Σύνθεση Φωνής

Είναι γεγονός ότι η επικοινωνία με ομιλία δεν εξαρτάται μόνο από τα ηχητικά αλλά και από τα οπτικά χαρακτηριστικά. Οι κινήσεις του προσώπου, όπως είναι οι διάφορες εκφράσεις, το κλείσιμο των ματιών κ.λ.π. δίνουν επιπλέον σημαντική πληροφορία για την συναισθηματική κατάσταση του ατόμου [17]. Επίσης, η οπτική αυτή πληροφορία και συγκεκριμένα οι κινήσεις του στόματος βοηθάνε επιπλέον στην πλήρη κατανόηση αλλά και στην στίξη του λόγου. Υπό την ύπαρξη θορύβου, ακουστικών προβλημάτων ή φιλτραρίσματος εύρους ζώνης, η οπτική αυτή πληροφορία αυξάνει δραματικά την κατανόηση της υποβαθμισμένης ηχητικής πληροφορίας [17]. Η οπτική πληροφορία βοηθάει επίσης στην διάκριση μεταξύ φωνημάτων τα οποία ομοιάζουν αρκετά στην ηχητική τους πληροφορία. Ένα τέτοιο παράδειγμα είναι τα αγγλικά σύμφωνα /b/ και /d/ όπου η οπτική πληροφορία αυξάνει σημαντικά την διάκριση τους [72]. Ίδιο παράδειγμα για την ελληνική γλώσσα μπορούν να αποτελέσουν τα σύμφωνα /μ/ και /ν/ που ενώ αποτελούν ρινικά σύμφωνα και τα δύο, το /μ/ επίσης ανήκει στα χειλικά και το /ν/ στα οδοντικά.

Είναι βέβαια πολύ σημαντικό να τονίσουμε ότι αν υπάρχει έλλειψη συνέπειας (π.χ. λάθος συγχρονισμός) μεταξύ της ηχητικής και της οπτικής πληροφορίας τότε η κατανόηση της ομιλίας μειώνεται δραματικά. Ένα τέτοιο παράδειγμα αποτελεί το φαινόμενο McGurk κατά το οποίο όταν ένα παρατηρητής λάβει την ηχητική πληροφορία ενός ήχου με την οπτική πληροφορία ενός άλλου ήχου, τότε οδηγείται στην αντίληψη ενός τρίτου ήχου [59]. Το φαινόμενο McGurk μελετήθηκε επίσης στα [26, 27] όπου παρατηρήθηκε ότι η ηχητική πληροφορία της συλλαβής /ba/ και η οπτική πληροφορία της συλλαβής /ga/ οδηγούν στην αντίληψη της συλλαβής /da/.

1.2.1 Ιστορική Αναδρομή

Η πρώτη μοντελοποίηση των εκφράσεων του ανθρώπινου προσώπου έγινε το 1972, όταν ο Parke παρουσίασε το πρώτο τρισδιάστατο πρόσωπο, και το 1974 παρουσίασε την πρώτη εκδοχή ενός παραμετρικού τρισδιάστατου μοντέλου [72]. Το μοντέλο αυτό αποτελείτο από 800 πολύγωνα που προσεγγίζουν το ανθρώπινο πρόσωπο που επηρεάζονται από 50 παραμέτρους [17]. Οι πιο σύγχρονοι συνθέτες εικόνας χρησιμοποιούν το παραμετρικό αυτό μοντέλο του Parke (του οποίου βελτιωμένη εκδοχή παρουσίασε το 1982), εφαρμόζοντας βεβαίως βελτιώσεις και αλλαγές για μεγαλύτερη ποιότητα και κατανόηση. Επιπλέον,

1.2.2 Δομή Συστήματος Οπτικοακουστικής Σύνθεσης Φωνής

Η γενική δομή ενός συστήματος οπτικοακουστικής σύνθεσης φωνής, με τη μορφή ενός ομιλούντος προσώπου, είναι πανομοιότυπη με την δομή ενός συστήματος σύνθεσης ομιλίας. Το σύστημα αποτελείται και πάλι από δύο μέρη: το *Σύστημα Επεξεργασίας της Φυσικής Γλώσσας* και το *Σύστημα Ψηφιακής Επεξεργασίας του Σήματος*. Το front-end του συστήματος, επιτελεί ακριβώς την ίδια λειτουργία

όπως και στο σύστημα σύνθεσης φωνής. Η αλλαγή στο σύστημα έγκειται στο δεύτερο κομμάτι, και πιο συγκεκριμένα στον τρόπο που χρησιμοποιεί το τμήμα τις πληροφορίες από το πρώτο σύστημα ώστε να παράγει το ομιλόν πρόσωπο μαζί με την συνθετική ομιλία.

1.2.3 Μέθοδοι Οπτικοακουστικής Σύνθεσης Φωνής

Σύνθεση με παράθεση. Τα συστήματα αυτού του τύπου, όπως ακριβώς και τα αντίστοιχα στη σύνθεση φωνής χρησιμοποιούν αποθηκευμένες εικόνες της οποίες συνδυάζουν με κάποια μορφοποίηση για την σύνθεση της ομιλίας [83, 53]. Όπως ακριβώς και στη σύνθεση φωνής, τα συστήματα αυτά τυγχάνουν χρήσης σε εφαρμογές περιορισμένου εύρους, καθώς είναι πλήρως ανελαστικά, και απαιτούν μεγάλο όγκο δεδομένων.

Σύνθεση με μοντελοποίηση των μυών. Τα συστήματα αυτά, ομοιάζουν με τα συστήματα μοντελοποίησης με αρθρωτές, καθώς προσπαθούν να μοντελοποιήσουν την κίνηση των μυών [6, 31]. Χαρακτηρίζονται επίσης από μεγάλη υπολογιστική πολυπλοκότητα και μεγάλη δυσκολία υλοποίησης και η κατασκευή ακριβών μοντέλων απαιτεί σημαντική προσπάθεια, ενώ τα αποτελέσματα δεν είναι ρεαλιστικά.

Παραμετρική Σύνθεση. Τα συστήματα που χρησιμοποιούν παραμετρικά μοντέλα τυγχάνουν της ευρύτερης αποδοχης λόγω των προβλημάτων που αναφέραμε στα συστήματα μοντελοποίησης της κίνησης των μυών [22, 58]. Στα συστήματα αυτά το πρόσωπο και οι κινήσεις του μοντελοποιούνται με μία γεωμετρική ερμηνεία της επιφάνειας του προσώπου χρησιμοποιώντας ένα σύνολο από παραμέτρους. Υπάρχουν πολλές μέθοδοι για την παραμετρική μοντελοποίηση του προσώπου και μία από αυτές παρουσιάζεται και χρησιμοποιείται στο παρόν κείμενο για τη σύνθεση εικονοσειράς όπως θα δούμε στο Κεφάλαιο 2.

1.3 Σκοπός και Συνεισφορές της Εργασίας

Σκοπός της παρούσας εργασίας είναι η θεωρητική ανάλυση, καθώς και η πρακτική υλοποίηση και αξιολόγηση ενός πλήρους συστήματος οπτικοακουστικής σύνθεσης φωνής με χρήση κρυφών μαρκοβιανών μοντέλων. Επίσης, μεγάλη βάση δόθηκε στην βελτιστοποίηση του ηχητικού και οπτικού αποτελέσματος. Τέλος, το σύστημα που δημιουργήθηκε, χρησιμοποιήθηκε για να μελετηθεί η απόδοση του σε 3 διαφορετικά συναισθήματα εκτός του ουδέτερου.

Πιο αναλυτικά, οι επιστημονικές συνεισφορές της παρούσας εργασίας συνοψίζονται στα ακόλουθα σημεία:

1. Δημιουργία ενός πλήρους συστήματος οπτικοακουστικής σύνθεσης φωνής για την Ελληνική γλώσσα, με χρήση active appearance μοντέλων και κρυφών μαρκοβιανών μοντέλων.
2. Δημιουργία πλήρους οπτικοακουστικής βάσης δεδομένων για την εκπαίδευση του οπτικοακουστικού συνθέτη φωνής.
3. Μελέτη για την βελτιστοποίηση του ηχητικού και οπτικού αποτελέσματος, μέσω των παραμέτρων και των τεχνικών που επηρεάζουν την απόδοση του συστήματος.
4. Μελέτη της απόδοσης του συστήματος για συναισθηματική οπτικοακουστική σύνθεση φωνής με χρήση κρυφών μαρκοβιανών μοντέλων, τόσο ως προς την ποιότητα της παραγόμενης ομιλίας και εικονοσειράς, όσο και ως προς το βαθμό αποτύπωσης του συναισθήματος.

1.4 Βάση Δεδομένων

Στα πλαίσια της διπλωματικής εργασίας δημιουργήθηκε η βάση δεδομένων CVSP-AudioVisual (CVSP-AV) για την εκπαίδευση του συνολικου οπτικοακουστικού συνθέτη φωνής.

Η συλλογή των πειραματικών δεδομένων, έγινε σε ειδικά διαμορφωμένο στούντιο, στο ερευνητικό κέντρο Αθηνά. Έγινε η καταγραφή συνολικά 3600 κατάλληλα επιλεγμένων προτάσεων (900 προτάσεις για κάθε μια από τέσσερις διαφορετικές συναισθηματικές καταστάσεις: κανονική, θυμωμένη, χαρούμενη, λυπημένη κατάσταση) ώστε να περιλαμβάνουν ένα μεγάλο εύρος των περιβαλλόντων των φωνημάτων της ελληνικής γλώσσας.

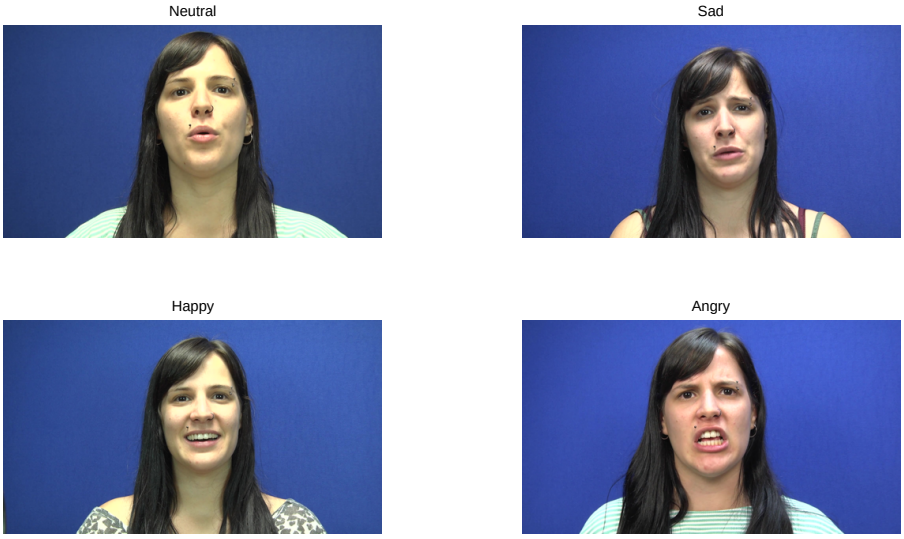
Κατά την καταγραφή, η ομιλήτρια (κα. Ιωάννα Πατσά) ήταν όσο το δυνατόν περισσότερο συναισθηματική, ανάλογα με το συναίσθημα κάθε φορά, ενώ τα φωνήματα αποτυπωνόντουσαν καθαρά.

Η αρχική καταγραφή του οπτικοακουστικού υλικού, έγινε σε 1080p (1920 x 1080 pxl) ανάλυση μέσω κάμερας, και 44100 Hz συχνότητα δειγματοληψίας μέσω του ειδικού μικροφώνου του στούντιο. Στη συνέχεια, για τον διαχωρισμό των προτάσεων αναπτύχθηκε η εφαρμογή, που ακολουθεί την εξής διαδικασία:

1. Υπολογισμός της χρονικής καθυστέρησης μέσω του οπτικού και του ακουστικού υλικού, υπολογίζοντας την ετεροσυσχέτιση μεταξύ του ήχου της κάμερας και του μικροφώνου και ενσωμάτωση του ήχου του μικροφώνου στα βίντεο.
2. Αντιστοιχία του κειμένου των προτάσεων με τα βίντεο, χρησιμοποιώντας αναγνώρισης φωνής με χρήση κρυφών μαρκοβιανών μοντέλων.
3. Χειροκίνητη επεξεργασία των αρχείων κειμένου στα οποία η ευθυγράμμιση δεν ήταν πλήρως σωστή. Η εύρεση των αρχείων αυτών έγινε αρχικά μέσω υπολογισμού της μέσης διάρκειας των φωνημάτων και στη συνέχεια με αναζήτηση των λανθασμένων φωνημάτων.
4. Κοπή των προτάσεων και αποθήκευση σε μορφή matroska multimedia container.

Στη συνέχεια, έγινε υποδειγματοληψία του ήχου των προτάσεων από 44100 Hz σε 16000 Hz, για μείωση του απαιτούμενου χώρου, και αύξηση της ταχύτητας των υπολογισμών του οπτικοακουστικού συνθέτη φωνής. Το μέγεθος της τελικής βάσης ανέρχεται στα 12 Gigabyte - με περίπου 3 Gigabyte να αντιστοιχούν σε κάθε συναίσθημα, ενώ συνολικά μαζεύτηκαν περί τις 120,000 εικόνες από τα video κάθε συναισθήματος. Στο τέλος επίσης, έγινε περαιτέρω επεξεργασία στις εικόνες, για να αντιμετωπιστούν όσο τον δυνατόν περισσότερο προβλήματα διαφορετικού φωτισμού που εμφανίστηκαν.

Τα ηχητικά και οπτικά φωνήματα καθώς και η συχνότητα εμφάνισης τους στη βάση δεδομένων, για ένα από τα τέσσερα διαφορετικά συναισθήματα εμφανίζονται στον Πίνακα 1.1, μαζί με ένα παράδειγμα. Στο Σχήμα 1.9 φαίνονται τέσσερις ενδεικτικές εικόνες - μία για κάθε συναίσθημα.



Σχήμα 1.9: Ενδεικτικές εικόνες από τη βάση δεδομένων CVSP-AV για κάθε ένα από τα τέσσερα διαφορετικά συναισθήματα.

Φώνημα (Συμβολισμός IPA [2])	Παράδειγμα	Συχνότητα Εμφάνισης
a	**ά**νδρας	4715
o	π**ό**νος	3796
e	**ε**λεύθερος	3784
u	**ου**ρανός	930
i	αντ**ί**ο	5813
b	**μπ**ασκέτα	98
c	**κ**ινώ	684
k	**κ**αλός	1094
d	**ντ**άμα	238
ð	**δ**άσος	829
f	**φ**άτνη	489
g	έ**γκ**αυμα	97
ɟ	**γκ**έμι	302
ɣ	**γ**αμπρός	486
l	**λ**άμπα	1069
ʎ	**λι**οτρίβι	48
m	**μ**άζα	1543
n	**ν**όμος	2566
ɲ	**νι**ότη	106
p	**π**όνος	1590
r	**ρ**ώμη	1767
s	**σ**ταθερός	3237
t	**τ**ομέας	2870
θ	**θ**ύμα	496
v	**β**άζω	377
x	**χ**άνω	316
ç	**χ**έρι	349
z	**ζ**ωή	304
sil	-	2072

Πίνακας 1.1: Συχνότητα εμφάνισης φωνημάτων στη βάση CVSP-AV για τις 900 προτάσεις κάθε συναισθήματος, συμπεριλαμβανομένης της παύσης. Οι προτάσεις της βάσης επιλέχθηκαν έτσι ώστε να συμπεριλαμβάνουν ένα μεγάλο αριθμό συνδυασμών φωνημάτων της ελληνικής γλώσσας.

1.5 Οργάνωση της Εργασίας

Το παρόν κείμενο χρησιμοποιεί την ακόλουθη δομή:

1. Στο Κεφάλαιο 2 γίνεται η απαραίτητη θεωρητική ανάλυση και πρακτική υλοποίηση της εξαγωγής των οπτικών και παραμέτρων για την εκπαίδευση του συστήματος οπτικοακουστικής σύνθεσης φωνής.

2. Στο Κεφάλαιο 3 γίνεται η θεωρητική ανάλυση ενός συστήματος οπτικοακουστικής σύνθεσης φωνής.

3. Στο Κεφάλαιο 4 παρουσιάζεται η πρακτική υλοποίηση του συστήματος οπτικοακουστικής σύνθεσης φωνής, και γίνεται αξιολόγηση του συστήματος.

4. Στο Κεφάλαιο 5 γίνεται μια πρώτη αξιολόγηση της συναισθηματικής σύνθεσης φωνής και της συναισθηματικής οπτικοακουστικής σύνθεσης φωνής για τρία διαφορετικά συναισθήματα εκτός του ουδέτερου.

5. Στο Κεφάλαιο 6 παρουσιάζονται τα συμπεράσματα της διπλωματικής εργασίας καθώς και οι μελλοντικές της κατευθύνσεις.

Κεφάλαιο 2

Παράμετροι Συστήματος Οπτικοακουστικής Σύνθεσης Φωνής με χρήση Κρυφών Μαρκοβιανών Μοντέλων

Στο Κεφάλαιο αυτό, θα μελετήσουμε τις παραμέτρους που περιέχονται στο διάνυσμα παρατήρησης $\boldsymbol{o}_t$ που χρησιμοποιείται για την εκπαίδευση των κρυφών μαρκοβιανών μοντέλων του συστήματος οπτικοακουστικής σύνθεσης φωνής και στη συνέχεια, αποτελεί την μερική έξοδο του συστήματος.

Για την φασματική ανάλυση των σημάτων φωνής από τα οποία απαρτίζεται η βάση δεδομένων CVSP-AudioVisual(CVSP-AV) και την εξαγωγή των φασματικών παραμέτρων, βασιζόμαστε κυρίως στην δουλειά των Tokuda et al. [80] που αφορά τους Mel Generalized Cepstral Coefficients, ενώ για την εξαγωγή της διέγερσης και των συντελεστών απεριοδικότητας χρησιμοποιούμε την ανάλυση των Kawahara et al. [45].

Για την εξαγωγή των οπτικών παραμέτρων των εικόνων από τις οποίες απαρτίζεται η βάση δεδομένων CVSP-AV, βασιζόμαστε κυρίως στην εργασία των Cootes et al. [30], για την δημιουργία ενός παραμετρικού μοντέλου που να επεξηγεί το πρόσωπο και τις κινήσεις του όπως αναφέραμε στην Υποενότητα 1.2.3 του 1ου Κεφαλαίου, μέσω του οποίου θα λάβουμε τις παραμέτρους που επεξηγούν το πρόσωπο σε κάθε διαφορετική εικόνα της βάσης δεδομένων.

Το Κεφάλαιο χωρίζεται σε δύο μέρη: το πρώτο μέρος αφορά την εξαγωγή των ακουστικών παραμέτρων του λόγου και την παραγωγή ενός σήματος φωνής με βάση αυτές, και το δεύτερο μέρος αφορά την εξαγωγή των οπτικών παραμέτρων και την παραγωγή ενός οπτικού σήματος (εικόνα) με βάση αυτές.

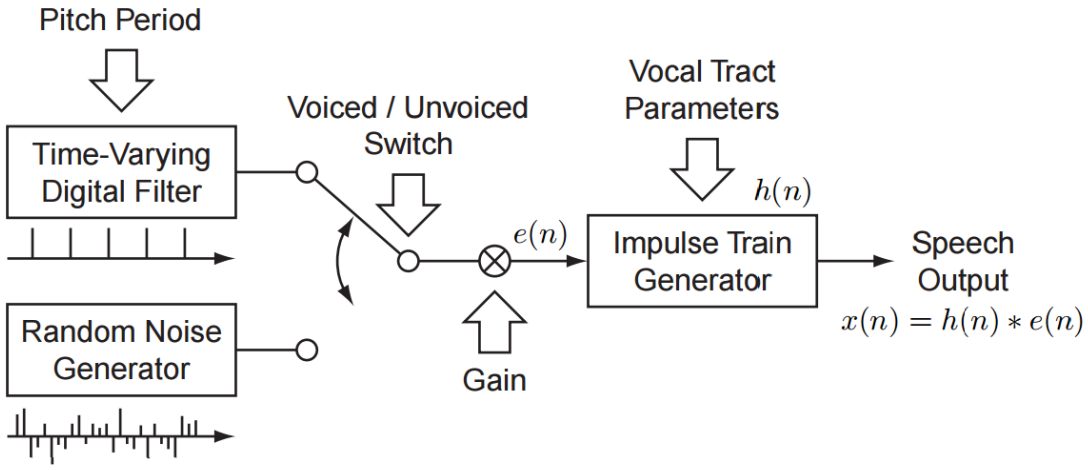
2.1 Ακουστικές Παράμετροι

2.1.1 Παραγωγή Φωνής

Η παραγωγή μιας διακριτής ακολουθίας που αναπαριστά ένα σήμα φωνής γίνεται με χρήση του Γραμμικού Χρονικά Αμετάβλητου (ΓΧΑ) φίλτρου του Σχήματος 2.1 [35]. Το φίλτρο $H(z)$ χρησιμοποιείται για την μοντελοποίηση της ανθρώπινης φωνητικής οδού. Η διέγερση, είναι είτε λευκός θόρυβος στην περίπτωση άφωνου σήματος (άφωνο λέγεται το σήμα που δεν περιλαμβάνει τη χρήση των φωνητικών χορδών), είτε μία ακολουθία παλμών με μεταβλητή περίοδο (quasi-periodic) στην περίπτωση φωνούμενου σήματος.

Κατά την παραγωγή ενός φωνητικού σήματος, οι παράμετροι του φίλτρου μεταβάλλονται, αλλά είναι εύλογο για μια μικρή χρονική περίοδο (5-10ms) να θεωρήσουμε το φίλτρο Χρονικά Αμετάβλητο. Επομένως, για μια δεδομένη διέγερση $e(n)$ η έξοδος του φίλτρου θα είναι η συνέλιξη της κρουστικής απόκρισης του με τη διέγερση [65], δηλαδή

$$y(n) = h(n) * e(n) \tag{2.1}$$



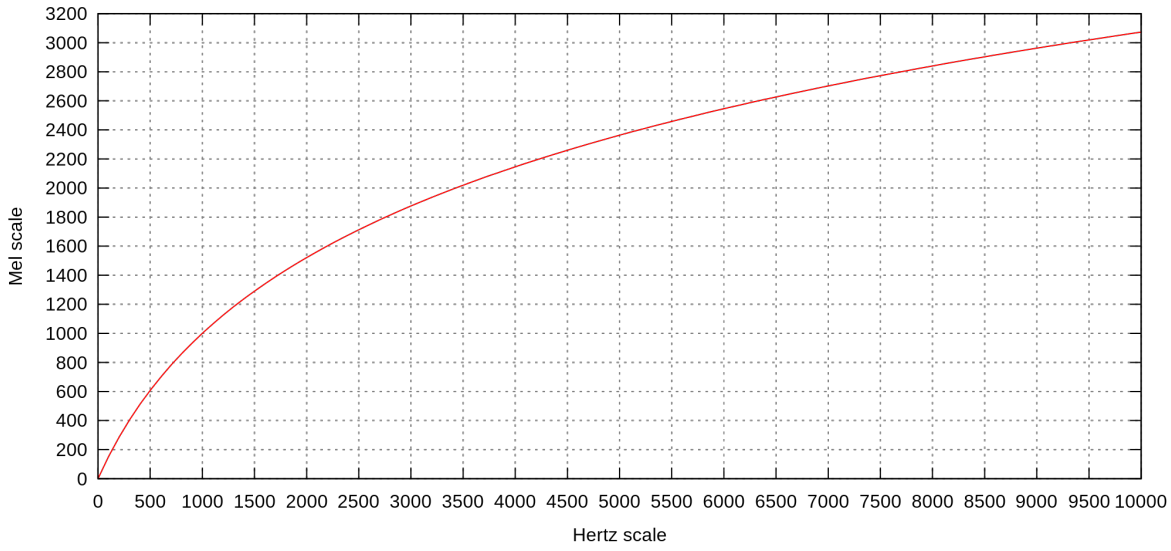
Σχήμα 2.1: Μοντελοποίηση ανθρώπινης παραγωγής ομιλίας [55].

2.1.2 Κλίμακα Συχνοτήτων Mel

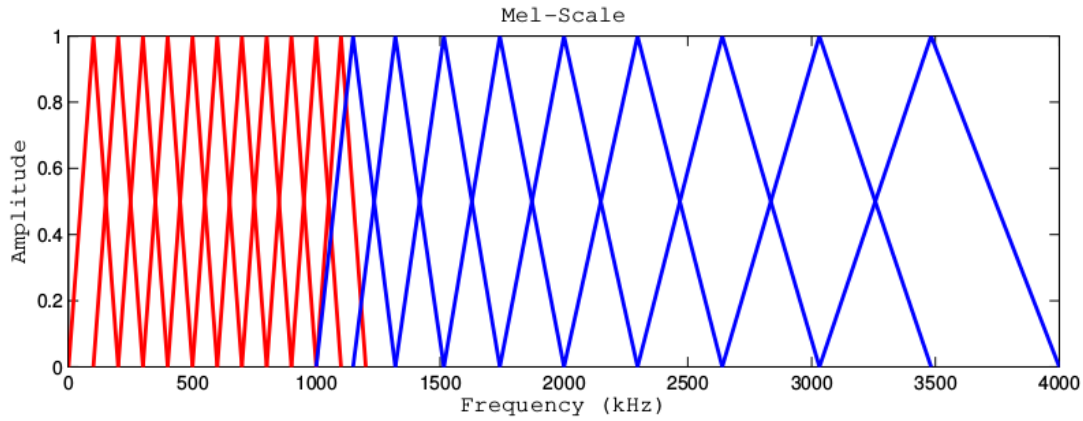
Το ανθρώπινο ακουστικό σύστημα δεν ακολουθεί γραμμική κλίμακα κατά την αντίληψη των ήχων, αλλά, εν γένει, αντιλαμβάνεται περισσότερο αλλαγές σε χαμηλές συχνότητες απ' ότι σε υψηλές συχνότητες. Για να λάβουμε χαρακτηριστικά που ομοιάζουν περισσότερο στα χαρακτηριστικά που αντιλαμβάνεται το ανθρώπινο ακουστικό σύστημα χρησιμοποιούμε την κλίμακα Mel [75], την οποία ενσωματώνουμε στα χαρακτηριστικά μας. Η μετατροπή από την γραμμική κλίμακα συχνοτήτων στην κλίμακα Mel γίνεται με χρήση της ακόλουθης εξίσωσης:

$$Mel(f) = 2595 \cdot \log_{10}(1 + \frac{f}{700}) \tag{2.2}$$

Το διάγραμμα της κλίμακας Mel συναρτήσει της συχνότητας φαίνεται στο Σχήμα 2.2. Η ανάλυση ενός σήματος φωνής στην κλίμακα Mel γίνεται με χρήση της τράπεζας τριγωνικών φίλτρων Mel που φαίνεται στο Σχήμα 2.3



Σχήμα 2.2: Η κλίμακα συχνοτήτων Mel. *Πηγή:* `https://en.wikipedia.org/wiki/Mel_scale`.



Σχήμα 2.3: Τράπεζα τριγωνικών φίλτρων Mel [87].

2.1.3 Φασματική Ανάλυση Φωνής με χρήση Mel Generalized Cepstral Συντελεστων

Mel Generalized Cepstral Coefficients. Η μοντελοποίηση της συνάρτησης μεταφοράς που προσομοιάζει τη λειτουργία της φωνητικής οδού $H(z)$ του Σχήματος 2.1 γίνεται με βάση τους M πρώτους mel generalised cepstral coefficients (mgc coefficients):

$$[c_{\alpha,\gamma}(0), c_{\alpha,\gamma}(1), c_{\alpha,\gamma}(2)...c_{\alpha,\gamma}(M-1)] \tag{2.3}$$

Για να καταλάβουμε πως ακριβώς γίνεται η μοντελοποίηση αυτή, θα αναπτύξουμε σταδιακά τον ορισμό των συντελεστών αυτών. Οι συντελεστές αυτοί εν γένει, συνιστούν μια ενοποίηση των μεθόδων της Cepstral Analysis και της Linear Prediction Analysis και προτάθηκαν από τους Tokuda et al [80].

Αρχικά, ορίζουμε τη *γενικευμένη λογαριθμική συνάρτηση* (generalized logarithmic function) [48] που αποτελεί γενίκευση της απλής λογαριθμικής συνάρτησης:

$$s_\gamma(w) = \begin{cases} \frac{(w^\gamma - 1)}{\gamma}, & 0 < |\gamma| \leq 1 \\ \log w, & \gamma = 0 \end{cases}, \qquad |\gamma| \leq 1, w \in (0, \infty) \tag{2.4}$$

Στο Σχήμα 2.4 βλέπουμε τη μορφή της συνάρτησης για διάφορες τιμές του γ.

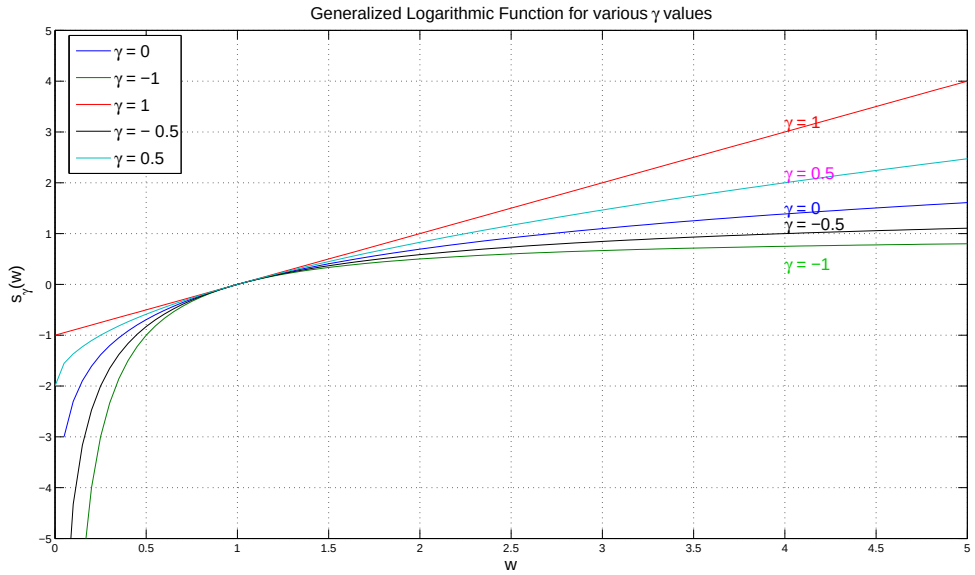
Η στρεβλωμένη (warped) κλίμακα συχνότητας $\beta_\alpha(\omega)$ ορίζεται ως η απόκριση φάσης του συστήματος διέλευσης (all-pass) [65]:

$$\Psi_\alpha(z) = \frac{z^{-1} - a}{1 - az^{-1}}|_{z=e^{j\omega}} = e^{-j\beta_\alpha(\omega)}, \quad |a| < 1 \tag{2.5}$$

όπου

$$\beta_\alpha(\omega) = \tan^{-1} \frac{(1-a^2)\sin\omega}{(1+a^2)\cos\omega - 2a} \tag{2.6}$$

Η απόκριση φάσης $\beta_\alpha(\omega)$ μας δίνει μια αρκετά καλή προσέγγιση στις ακουστικές κλίμακες συχνότητας για κατάλληλη επιλογή του α. Στον Πίνακα 2.1 φαίνεται η κατάλληλη επιλογή του α για την προσέγγιση της κλίμακας Mel, συναρτήσει της συχνότητας δειγματοληψίας.



Σχήμα 2.4: Διάγραμμα Γενικευμένης Λογαριθμικής Συνάρτησης για διάφορες τιμές του γ.

Sampling Frequency	8 kHz	10 kHz	12 kHz	16 kHz	48 kHz
Mel Scale	0.31	0.35	0.37	0.42	0.55

Πίνακας 2.1: Επιλογή του α για προσέγγιση της κλίμακας Mel συναρτήσει της συχνότητας δειγματοληψίας.

Με τη χρήση του παραπάνω ορισμού μπορούμε να ορίσουμε τώρα το mel-generalized cepstrum [80] $c_{\alpha,\gamma}(m)$ ως τον αντίστροφο μετασχηματισμό Fourier του γενικευμένου λογαριθμικού φάσματος, υπολογισμένο στη στρεβλωμένη κλίμακα συχνότητας $\beta_\alpha(\omega)$:

$$s_\gamma(X(e^{j\omega})) = \sum_{m=-\infty}^{\infty} c_{\alpha,\gamma}(m) \cdot \Psi_\alpha^m(z) \tag{2.7}$$

όπου $\Psi_\alpha^m(z) = e^{-j\beta_\alpha(\omega)m}$ και $X\left(e^{j\omega}\right)$ είναι ο μετασχηματισμός Fourier του σήματος $x(n)$.

Τώρα, το σύστημα για τη μοντελοποίηση της φωνητικής οδού $H(z)$ υλοποιείται με τη χρήση των πρώτων $M+1$ mgc συντελεστών, δηλαδή:

$$H(z) = s_\gamma^{-1}\left(\sum_{m=0}^{M} c_{\alpha,\gamma}(m)\Psi_\alpha^m(z)\right) = \begin{cases} \left(1+\gamma\sum_{m=0}^{M} c_{\alpha,\gamma}(m)\Psi_\alpha^m(z)\right)^{\frac{1}{\gamma}}, & 0<|\gamma|\leq 1 \\ \exp\sum_{m=0}^{M} c_{\alpha,\gamma}(m)\Psi_\alpha^m(z), & \gamma = 0 \end{cases} \tag{2.8}$$

Ανάλογα με την επιλογή του ζεύγους (α,γ), η συνάρτηση του φίλτρου λαμβάνει και διαφορετική μορφή:

1. Για $(\alpha,\gamma)=(0,0)$ παίρνουμε ένα απλό cepstral μοντέλο της συνάρτησης μεταφοράς:

$$H(z) = \exp\sum_{m=0}^{M} c_{\alpha,\gamma}(m) z^{-m} \tag{2.9}$$

2. Για $(\alpha, \gamma) = (0, -1)$ παίρνουμε ένα απλό autoregressive μοντέλο της συνάρτησης μεταφοράς:

$$H(z) = \frac{1}{1 - \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z^{-m}} \tag{2.10}$$

3. Για $(\alpha, \gamma) = (0.42, 0)$ παίρνουμε ένα cepstral μοντέλο σε στρεβλωμένη κλίμακα συχνότητας της συνάρτησης μεταφοράς:

$$H(z) = \exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m) \Psi_{\alpha}^{m}(z) \tag{2.11}$$

4. Για $(\alpha, \gamma) = (0.42, -1)$ παίρνουμε ένα autoregressive μοντέλο σε στρεβλωμένη κλίμακα συχνότητας της συνάρτησης μεταφοράς:

$$H(z) = \frac{1}{1 - \sum_{m=0}^{M} c_{\alpha,\gamma}(m) \Psi_{\alpha}^{m}(z)} \tag{2.12}$$

Εύρεση MGC Συντελεστών και Φίλτρο Σύνθεσης. Η εύρεση των πρώτων $M + 1$ mgc συντελεστών, γίνεται έτσι ώστε το ακόλουθο κριτήριο να ελαχιστοποιείται [43] :

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} [\exp R(\omega) - R(\omega) - 1] \, d\omega \tag{2.13}$$

όπου

$$R(\omega) = \log I_N(\omega) - \log |H(e^{j\omega})|^2 \tag{2.14}$$

και $I_N(\omega)$ είναι το τροποποιημένο περιοδοδιάγραμμα μιας ασθενούς στάσιμης διαδικασίας $x(n)$ με παράθυρο $w(n)$ μήκους N:

$$I_N(\omega) = \frac{|\sum_{n=0}^{N-1} w(n) x(n) \exp(-j\omega n)|^2}{\sum_{n=0}^{N-1} w^2(n)} \tag{2.15}$$

Για τη λύση αυτού του προβλήματος ελαχιστοποίησης, αρχικά, χρησιμοποιώντας τους μετασχηματισμούς:

$$c_{\alpha,\gamma}(m) = \begin{cases} b_{\alpha,\gamma}(m), & m = M \\ b_{\alpha,\gamma}(m) + \alpha b_{\alpha,\gamma}(m+1), & 0 \le m < M \end{cases} \tag{2.16}$$

$$b_{\alpha,\gamma}(m) = \begin{cases} c_{\alpha,\gamma}(m), & m = M \\ c_{\alpha,\gamma}(m) - \alpha c_{\alpha,\gamma}(m+1), & 0 \le m < M \end{cases} \tag{2.17}$$

$$b'_{\alpha,\gamma}(m) = \frac{b_{\alpha,\gamma}(m)}{1 + \gamma b_{\alpha,\gamma}(0)}, \quad m = 1, 2, ..., M \tag{2.18}$$

$$\Phi_a^m(z) = \begin{cases} 1, & m = 0 \\ \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}} (\Psi_a^m)^{-(m-1)}, & m \ge 1 \end{cases} \tag{2.19}$$

και παίρνοντας τον παράγοντα κέρδους $K = s_\gamma^{-1}\left(b'_{\alpha,\gamma}(0)\right)$ εκτός του φίλτρου σύνθεσης $H(z)$ που προκύπτει με τους ανωτέρω μετασχηματισμούς, έχουμε

$$H(z) = s_\gamma^{-1}\left(\sum_{m=0}^{M} c_{\alpha,\gamma}(m)\Phi_\alpha^m(z)\right) = K \cdot D(z) \tag{2.20}$$

όπου

$$K = s_\gamma^{-1}\left(b'_{\alpha,\gamma}(0)\right) \tag{2.21}$$

και

$$D(z) = s_\gamma^{-1}\left(\sum_{m=1}^{M} b'_{\alpha,\gamma}(m)\Phi_\alpha^m(z)\right) \tag{2.22}$$

Το πρόβλημα ελαχιστοποίησης της (2.13) ως προς το διάνυσμα $\mathbf{c}_{\alpha,\gamma} = [c_{\alpha,\gamma}(0), c_{\alpha,\gamma}(1), ..., c_{\alpha,\gamma}(M)]$ είναι ισοδύναμο με το πρόβλημα ελαχιστοποίησης ως προς το διάνυσμα $\mathbf{b}_{\alpha,\gamma} = [b_{\alpha,\gamma}(0), b_{\alpha,\gamma}(1), ..., b_{\alpha,\gamma}(M)]$ εφόσον συνδέονται με γραμμικό μετασχηματισμό, καθώς και με το πρόβλημα ελαχιστοποίησης των K και $\mathbf{b}'_{\alpha,\gamma} = [b'_{\alpha,\gamma}(1), ..., b'_{\alpha,\gamma}(M)]$ όπως βλέπουμε από τις εξισώσεις (2.21) και (2.22).

Καταλαβαίνουμε ότι εφόσον το φίλτρο $H(z)$ είναι φίλτρο σύνθεσης, αυτό πρέπει να είναι ευσταθές. Επομένως, αν θεωρήσουμε το φίλτρο $D(z)$ ως ελαχίστης φάσης, η ελαχιστοποίηση της (2.8) ως προς το $\mathbf{c}_{\alpha,\gamma}$ είναι ισοδύναμη με την ελαχιστοποίηση του

$$\epsilon = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{I_N(\omega)}{|D(e^{j\omega}|^2} \quad d\omega \tag{2.23}$$

ως προς το $\mathbf{b}'_{\alpha,\gamma}$ και ελαχιστοποίηση του E ως προς το K. Μάλιστα, όπως έχει αποδειχθεί στο [79], η ελαχιστοποίηση αυτού του προβλήματος, οδηγεί και στην ελαχιστοποίηση του $\epsilon = E[e^2(n))]$ και επομένως μπορεί να ερμηνευθεί ως μέθοδος μέσων ελαχίστων τετραγώνων του σφάλματος γραμμικής πρόβλεψης. Παίρνοντας την πρώτη παράγωγο του K και θέτοντας την ίση με το μηδέν, παίρνουμε:

$$K = \sqrt{\epsilon_{min}} \tag{2.24}$$

όπου ϵ_{min} είναι η ελάχιστη τιμή του ϵ.

Όταν $-1 \leq \gamma \leq 1$ το κριτήριο E είναι κυρτό ως προς το $c_{\alpha,\gamma}$ και επομένως υπάρχει μόνο ένα ελάχιστο σημείο του E. Μία λύση του προβλήματος, καθώς και η απόδειξη ότι η λύση του προβλήματος ελαχιστοποίησης είναι ευσταθής (για $-1 \leq \gamma \leq 1$), δίνεται με χρήση της μεθόδου Newton-Raphson στο [39].

Πρακτική Υλοποίηση Φίλτρου Σύνθεσης. Ένα άλλο πρόβλημα που θα πρέπει να αντιμετωπίσουμε είναι το γεγονός πως η $D(z)$ που χρησιμοποιείται για τη σύνθεση φωνής δεν είναι ρητή συνάρτηση και δεν γίνεται να υλοποιηθεί, επομένως είναι αναγκαστικό να καταφύγουμε σε κάποια ρητή προσέγγιση της συνάρτησης.

Μια καλή προσέγγιση στην εκθετική συνάρτηση γίνεται με χρήση της ρητής συνάρτησης

$$\exp w \approx R_L(w) = \frac{P_L(w)}{P_L(-w)} \tag{2.25}$$

όπου

$$P_L(w) = P_L^l(w) = 1 + \sum_{i=1}^{L} A_{L,l} w^l \tag{2.26}$$

Εδώ, επιλέγουμε τους συντελεστές

$$A_{L,l} = \frac{1 - \lambda_{L,l}}{l!} \binom{L}{l} / \binom{L}{l} \tag{2.27}$$

και επομένως η προσέγγιση μας αποτελεί προσέγγιση Padé $[L/L]$ τάξης. Ο υπολογισμός των συντελεστών $A_{L,l}$ της προσέγγισης Padé, μπορεί να γίνει με χρήση είτε του αλγορίθμου epsilon του Wynn, είτε με χρήση του εκτεταμένου Ευκλείδειου αλγορίθμου για τον υπολογισμό του μέγιστου κοινού διαιρέτη του πολυωνύμου. Περισσότερες πληροφορίες πάνω στην προσέγγιση Padé και στην εύρεση των συντελεστών περιλαμβάνονται στο Παράρτημα 1. Οι συντελεστές $\lambda_{L,l}$ επιλέγονται έτσι ώστε να ελαχιστοποιείται το σφάλμα προσέγγισης.

Με χρήση της ανωτέρω προσέγγισης, η συνάρτηση μεταφοράς $D(z)$ (για $\gamma = 0$ και χωρίς βλάβη της γενικότητας) προσεγγίζεται ως

$$D(z) = \exp F(z) \approx R_L(F(z)) \tag{2.28}$$

όπου

$$F(z) = \sum_{m=1}^{M} b'_{\alpha,\gamma}(m) \Phi_\alpha^m(z) \tag{2.29}$$

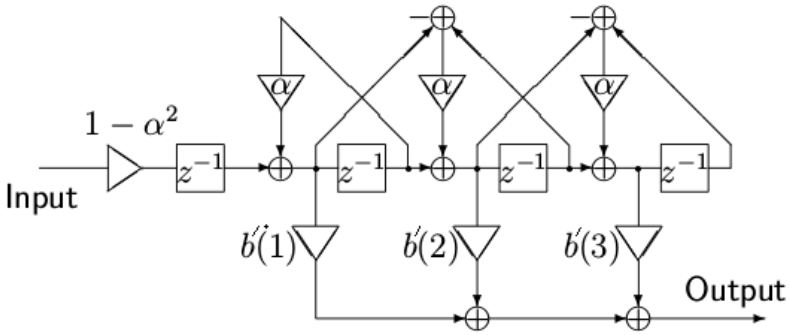
Είναι σημαντικό να τονίσουμε ότι η $F(z)$ περιλαμβάνει delay-free loop, δηλαδή βρόχο που δεν περιέχει καθυστέρηση. Η ύπαρξη ενός delay-free loop καθιστά τον υπολογισμό σε H/Y αδύνατη. Για την επίλυση του προβλήματος αυτού στο σύστημα μας εφαρμόζουμε έναν απλό μετασχηματισμό που μπορεί να βρεθεί στο [55].

Στο Σχήμα 2.5 απεικονίζεται το δομικό διάγραμμα της $F(z)$ για $M = 3$ και στο Σχήμα 2.6 το δομικό διάγραμμα της $R_L(F(z))$ για $L = 4$.

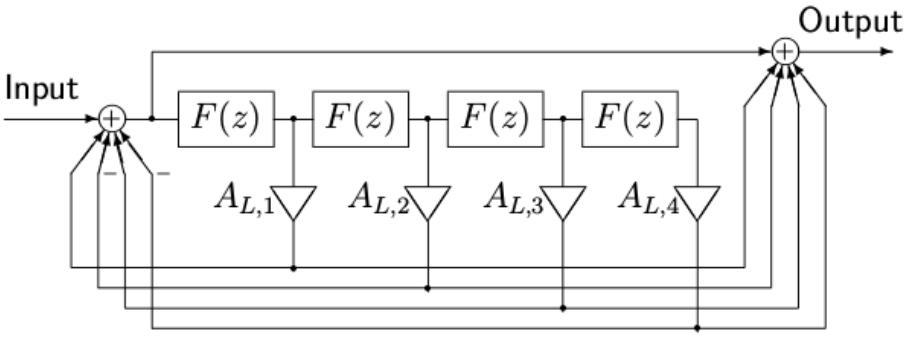
2.1.4 Διέγερση και Συντελεστές Απεριοδικότητας

Για την εξαγωγή της τροχιάς της θεμελιώδους συχνότητας της φωνής, από τα σήματα φωνής της βάσης δεδομένων, χρησιμοποιούμε το εργαλείο STRAIGHT το οποίο βασίζεται στην ανάλυση των Kawahara et al. [45]. Κατά την ανάλυση αυτή εκτός από την εξαγωγή της καμπύλης του pitch για τη διέγερση, γίνεται και εξαγωγή συντελεστών απεριοδικότητας οι οποίοι περιέχουν πληροφορίες για τα πιο ιδιαίτερα χαρακτηριστικά του σήματος φωνής. Η ανάλυση STRAIGHT υποθέτει ότι το σήμα φωνής έχει την ακόλουθη μορφή:

$$x(t) = \sum_{k=1}^{N} \alpha_k(t) \cos\left(\int_0^t \left(k\omega_0(\tau) + \omega_k(\tau) \, d\tau + \phi_k(0) \right) \right) \tag{2.30}$$



Σχήμα 2.5: Δομικό Διάγραμμα της $F(z)$ για $M = 3$ [55].



Σχήμα 2.6: Δομικό Διάγραμμα της $R_L(F(z))$ για $L = 4$ [55].

όπου $a_k\,(t)$ αναπαριστά το αργά μεταβαλλόμενο πλάτος, $\omega_k\,(\tau)$ αναπαριστά την αργά μεταβαλλόμενη διαταραχή της k-οστής αρμονικής συνιστώσας και $\phi_k\,(0)$ είναι η αρχική φάση.

Με βάση αυτή τη μοντελοποίηση το εργαλείο προχωράει σε εξαγωγή της θεμελιώδους συχνότητας με χρήση του αλγορίθμου TEMPO (Time-domain Excitation extractor using Minimum Pertubatin Operator) [46], και την εξαγωγή των συντελεστών απεριοδικότητας που υπολογίζονται με την διαίρεση του φάσματος που έχει υπολογιστεί με χρήση παρεμβολής ως προς τις κορυφές του αρχικού φάσματος ισχύος και του φάσματος που έχει υπολογιστεί με χρήση παρεμβολής ως προς τις κοιλάδες του αρχικού φάσματος ισχύος.

2.2 Οπτικές Παράμετροι

Για την εξαγωγή εικαστικών χαρακτηριστικών, χρησιμοποιούμε ως υπόβαθρο τα Active Appearance Models (aam) που προτάθηκαν από τους Cootes et al. [28]. Τα active appearance μοντέλα αποτελούν στατιστικά μοντέλα που περιγράφουν την μεταβολή του σχήματος και της υφής ενός αντικειμένου (χρώμα - ένταση) σε ένα σύνολο από διαφορετικές εικόνες.

2.2.1 Μοντελοποίηση Σχήματος

Για να μοντελοποιήσουμε το σχήμα ενός αντικειμένου, το αναπαριστούμε με ένα σύνολο σημείων που ορίζονται χειροκίνητα. Με την χειροκίνητη σημαδοποίηση των σημείων που περιγράφουν ένα αντικείμενο σε διαφορετικές εικόνες, επιθυμούμε να μοντελοποιήσουμε με το μοντέλο μας τον τρόπο με τον οποίο μεταβάλλονται τα σημεία του αντικειμένου σε διαφορετικές εικόνες.

Εν γένει, τα σημεία που μπορούν να χρησιμοποιηθούν για την αναπαράσταση αντικειμένων ταξινομούνται από τον Bookstein [21] ως:

1. Σημεία σε τμήματα του αντικειμένου με μεγάλη σημασία ανάλογα με την εφαρμογή του μοντέλου, όπως το κέντρο ενός ματιού ή τις γωνίες ενός σχήματος.
2. Σημεία μεγάλης σημασίας ανεξαρτήτως της εκάστοτε εφαρμογής, όπως το υψηλότερο σημείο σε έναν προσανατολισμό του σχήματος ή ακρότατα μιας καμπύλης.
3. Συνδυαστικά σημεία από τις δύο πρώτες κατηγορίες όπως τα σημεία που περιγράφουν τα όρια του σχήματος.

Για να μπορέσουμε να συγκρίνουμε τα αντίστοιχα σημεία που προέρχονται από διαφορετικές εικόνες, θα πρέπει αυτά να ευθυγραμμιστούν ως προς κάποιους άξονες καθώς το αντικείμενο σε διαφορετικές εικόνες είναι πιθανό να βρίσκεται σε διαφορετική θέση, ως προς ένα σταθερό σημείο της εικόνας. Επιτυγχάνουμε την ευθυγράμμιση αυτή εφαρμόζοντας γραμμικούς μετασχηματισμούς (συγκεκριμένα μετατόπιση, κλιμάκωση και περιστροφή) στα σχήματα εκπαίδευσης έτσι ώστε να βρίσκονται όσο το δυνατόν πλησιέστερα [29].

Έστω δύο διαφορετικά σύνολα σημείων

$$\boldsymbol{x}_i = (x_{i0}, y_{i0}, x_{i1}, y_{i1}, ..., x_{in-1}, y_{in-1})^T \tag{2.31}$$

$$\boldsymbol{x}_j = (x_{j0}, y_{j0}, x_{j1}, y_{j1}, ..., x_{jn-1}, y_{jn-1})^T \tag{2.32}$$

που περιγράφουν το ίδιο σχήμα σε δύο διαφορετικές εικόνες. Αν συμβολίσουμε ως $M(s, \theta)[\boldsymbol{x}])$ τη μήτρα που περιγράφει τους μετασχηματισμούς περιστροφής ως προς θ και κλιμακωσης ως προς s, όπου

$$M(s, \theta) \begin{bmatrix} x_{jk} \\ y_{jk} \end{bmatrix} = \begin{pmatrix} (s\cos\theta)x_{jk} - (s\sin\theta)y_{jk} \\ (s\sin\theta)x_{jk} + (s\cos\theta)y_{jk} \end{pmatrix} \tag{2.33}$$

και με το διάνυσμα $\boldsymbol{t}_j = (t_{xj}, t_{yj}, ..., t_{xj}, t_{yj})^T$ συμβολίσουμε το γραμμικό μετασχηματισμό της μετατόπισης, σκοπός μας είναι να ελαχιστοποιήσουμε το σταθμισμένο άθροισμα:

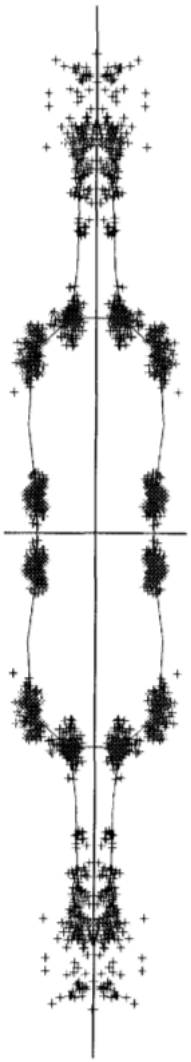
$$E_j = (\boldsymbol{x}_i - M(s_j, \theta_j)[\boldsymbol{x}_j] - \boldsymbol{t}_j)^T \boldsymbol{W} (\boldsymbol{x}_i - M(s_j, \theta_j)[\boldsymbol{x}_j] - \boldsymbol{t}_j) \tag{2.34}$$

όπου η μήτρα των βαρών $\boldsymbol{W}$ χρησιμοποιείται για να προσδώσουμε μεγαλύτερη σημασία στα σημεία που τείνουν να είναι πιο σταθερά, κατά την εναλλαγή των εικόνων. Με βάση αυτό τον ορισμό, αν συμβολίσουμε με R_{kl} την απόσταση μεταξύ των σημείων k και l και με $V_{R_{kl}}$ τη διακύμανση της απόστασης αυτής στο σύνολο των σχημάτων, τότε το βάρος που αποδίδεται σε ένα σημείο k είναι:

$$w_k = (\sum_{l=0}^{n-1} V_{R_{kl}})^{-1} \tag{2.35}$$

όπου βλέπουμε ότι όσο περισσότερο τείνει να μεταβάλλεται ένα σημείο, τόσο μικρότερο βάρος θα του αποδωθεί. Με βάση τις παραπάνω εξισώσεις και ορισμούς, για την ευθυγράμμιση των σχημάτων χρησιμοποιείται η ακόλουθη επαναληπτική διαδικασία:

1. Εφαρμογή μετασχηματισμών περιστροφής, κλιμάκωσης και μετατόπισης για την ευθυγράμμιση των σχημάτων των διαφορετικών εικόνων ως προς την πρώτη εικόνα.



Σχήμα 2.7: Ευθυγραμμισμένα Σχήματα Αντίστασης σε κοινό διάγραμμα [29].

2. Εύρεση του Μέσου Σχήματος Προσώπου μέσω της εξίσωσης (2.36).
3. Εκ νέου ευθυγράμμιση μέσω των μετασχηματισμών, ως προς το Μέσο Σχήμα Προσώπου.
4. Επανάληψη των βημάτων 2 και 3 έως τη σύγκλιση της διαδικασίας.

Αν μετά την ευθυγράμμιση όλων των σχημάτων, αναπαραστήσουμε τις θέσεις τους σε ένα κοινό διάγραμμα, θα λάβουμε μία εικόνα της μορφής του Σχήματος 2.7. Στο Σχήμα αυτό φαίνονται τα ευθυγραμμισμένα σχήματα μίας αντίστασης, όπως έχουν ληφθεί από 20 διαφορετικές εικόνες αντιστάσεων μεταβλητού μήκους και σχήματος. Όπως παρατηρούμε στο Σχήμα, μερικά σημεία τείνουν να είναι σταθερά, ενώ άλλα, όταν συνδυαστούν από διαφορετικές εικόνες, τείνουν να σχηματίζουν μικρά σηματικά σύνολα. Ο χώρος που μας δίνουν οι συνδυασμοί αυτοί τον σημείων καλείται Επιτρεπόμενος Χώρος του Σχήματος, και διαφορετικοί συνδυασμοί των σημείων από διαφορετικά σύνολα μας δίνουν την δυνατότητα δημιουργίας καινούργιων αναπαραστάσεων του αντικειμένου.

Στη συνέχεια, υπολογίζουμε το μέσο σχήμα ως

$$\bar{\boldsymbol{x}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \tag{2.36}$$

όπου $\mathbf{x}_i$ είναι το σύνολο των σημείων της εικόνας i, και N ο αριθμός των εικόνων εκπαίδευσης.

Οι κύριοι άξονες ως προς τους οποίους μεταβάλλεται το σχήμα μπορούν να βρεθούν χρησιμοποιώντας Ανάλυση σε Κύριες Συνιστώσες (Principal Component Analysis, PCA) στα δεδομένα. Κάθε άξονας θα μας δίνει και ένα τρόπο με τον οποίο κινούνται τα σημεία καθώς μεταβάλλεται το σχήμα.

Για να εφαρμόσουμε την Ανάλυση σε Κύριες Συνιστώσες αρχικά υπολογίζουμε την απόκλιση κάθε σχήματος ως προς το μέσο σχήμα:

$$d\boldsymbol{x}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}} \tag{2.37}$$

όπου N ο αριθμός των διαφορετικών εικόνων του σχήματος, και x_i το σύνολο σημείων του σχήματος για την εικόνα i. Στη συνέχεια υπολογίζουμε τη μήτρα συνδιακύμανσης $\boldsymbol{S}$ μεγέθους $2n \times 2n$, ως:

$$\boldsymbol{S} = \frac{1}{N}\sum_{i=1}^{N} d\boldsymbol{x}_i d\boldsymbol{x}_i^T \tag{2.38}$$

Τέλος, τα μοναδιαία ιδιοδιανύσματα της $\boldsymbol{S}$ που ικανοποιούν τις

$$\boldsymbol{S}\boldsymbol{p}_k = \lambda_k \boldsymbol{p}_k \tag{2.39}$$

$$\boldsymbol{p}_k^T \boldsymbol{p}_k = 1 \tag{2.40}$$

μας δίνουν τους κυριότερους τρόπους μεταβολής των σημείων του σχήματος. Συνήθως οι περισσότερες μεταβολές του σχήματος μπορούν να μοντελοποιηθούν με ένα μικρό αριθμό ιδιοδιανυσμάτων. Μία μέθοδος υπολογισμού του αριθμού των ιδιοδιανυσμάτων είναι η επιλογή τους, έτσι ώστε αυτά να επεξηγούν ένα επιθυμητό ποσοστό της συνολικής μεταβολής που ορίζει η μήτρα των ιδιοδιανυσμάτων. Το ποσοστό της μεταβολής που επεξηγεί ένα ιδιοδιάνυσμα $\mathbf{p}_i$ είναι ίσο με την αντίστοιχη ιδιοτιμή λ_i.

Τώρα, με βάση την ανάλυση σε κύριες συνιστώσες που κάναμε, οποιοδήποτε σημείο στον Επιτρεπόμενο Χώρο του Σχήματος (και επομένως οποιοδήποτε επιτρεπόμενο σχήμα) μπορεί να μοντελοποιηθεί με το άθροισμα του μέσου σχήματος, με ένα γραμμικό μετασχηματισμό των ιδιοδιανυσμάτων:

$$\boldsymbol{x} = \bar{\boldsymbol{x}} + \boldsymbol{P}\boldsymbol{b} \tag{2.41}$$

όπου $\mathbf{P}$ είναι η μήτρα των ιδιοδιανυσμάτων σχήματος (ιδιοσχήματα) $\mathbf{p}_i$ και $\mathbf{b}$ είναι ένα διάνυσμα βαρών.

Η εξίσωση 2.41 μας επιτρέπει επομένως να παράγουμε νέα παραδείγματα του σχήματος μεταβάλλοντας το διάνυσμα των βαρών $\mathbf{b}$. Το ποσοστό μεταβολής του διανύσματος συνήθως οριοθετείται με βάση τις ιδιοτιμές έτσι ώστε το τελικό αποτέλεσμα να ομοιάζει με το αρχικό. Συνήθη όρια που δίνουμε στη μεταβολή του διανύσματος είναι της τάξης του $|\sqrt{\lambda_k}|$ δηλαδή, τρεις φορές απόκλιση της μέσης τιμής.

2.2.2 Μοντελοποίηση Υφής

Ομοίως με την προηγούμενη υποενότητα, θα θέλαμε να εξάγουμε από το σύνολο των εικόνων και τα ιδιοδιανύσματα αυτά που μοντελοποιούν την μεταβολή της υφής [28]. Με τον όρο υφή, υποδηλωνουμε την ένταση ή το χρώμα του αντικειμένου στο σύνολο των pixel που βρίσκονται "μέσα" στο πλέγμα που δημιουργείται από τα σημεία του μέσου Σχήματος $\bar{\boldsymbol{x}}$ [56].

Η μοντελοποίηση της υφής του αντικειμένου γίνεται όπως πριν, εφαρμόζοντας Ανάλυση σε Κύριες Συνιστώσες στις διαφορετικές εικόνες που έχουν υποστεί όμως *κανονικοποίηση ως προς το σχήμα*. Κάθε μία από όλες τις εικόνες του συνόλου εκπαίδευσης υπόκειται σε στρέβλωση από το σχήμα $\boldsymbol{x}$ της εικόνας στο μέσο σχήμα $\bar{\boldsymbol{x}}$. Η μετατροπή αυτή γίνεται συνήθως με χρήση είτε ενός αφινικού μετασχηματισμού [28] είτε με thin plate splines [30].

Η Ανάλυση σε Κύριες Συνιστώσες τώρα μας δίνει:

$$\boldsymbol{A} = \bar{\boldsymbol{A}} + \boldsymbol{P}_t \boldsymbol{b}_t \tag{2.42}$$

όπου $\boldsymbol{A}$ είναι η υφή του αντικειμένου, $\bar{\boldsymbol{A}}$ είναι η μέση υφή, $\boldsymbol{P}_t$ είναι η μήτρα των ιδιοδιανυσμάτων υφής $\boldsymbol{p}_{t,i}$ και $\mathbf{b}_t$ είναι ένα διάνυσμα βαρών.

Ανακατασκευή Εικόνας. Ας υποθέσουμε ότι για μία εικόνα k που περιέχει το αντικείμενο μας, έχουμε εξάγει τα ιδιοδιανύσματα που μοντελοποιούν το σχήμα του αντικειμένου και την υφή. Τότε, μπορούμε να ανακατασκευάσουμε το σχήμα του αντικειμένου στην παρούσα εικόνα k με βάση την εξίσωση 2.41, και την υφή με βάση την εξίσωση 2.42. Πιο συγκεκριμένα, αφού υπολογίσουμε το σχήμα του αντικειμένου με βάση την εξίσωση 2.41 και την υφή με βάση την εξίσωση 2.42 - ως προς το μέσο Σχήμα $\boldsymbol{x}$, τότε με χρήση ενός μετασχηματισμού όπως πριν, αυτή τη φορά από το πλέγμα των σημείων του μέσου σχήματος, στο πλέγμα των σημείων του σχήματος που υπολογίστηκε $\boldsymbol{x}$ ανακατασκευάζουμε το αντικείμενο της συγκεκριμένης εικόνας.

Εφαρμογή ενός AAM σε καινούργια εικόνα. Τώρα, θα επικεντρωθούμε στο πρόβλημα: Έχοντας ένα εκπαιδευμένο active appearance μοντέλο, και μία καινούργια αναπαράσταση του αντίστοιχου αντικειμένου, πως μπορούμε να μεταβάλλουμε τις παραμέτρους του μοντέλου ώστε να επεξηγούν τη νέα αυτή αναπαράσταση;

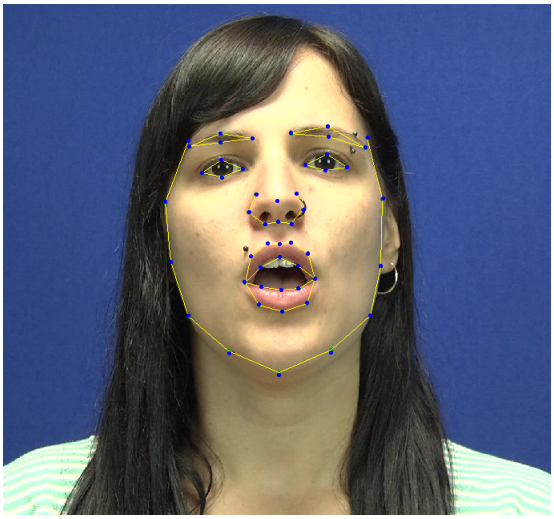
Το πρόβλημα αυτό αποτελεί ένα πρόβλημα βελτιστοποίησης υψηλής διαστατικότητας. Αν ορίσουμε την αρχική εικόνα του αντικειμένου ως $A_{\boldsymbol{w}}$ και την ανακατασκευασμένη εικόνα ως A_{λ} τότε το σφάλμα $E(\boldsymbol{q})$ είναι ίσο με

$$E(q) = A_{\boldsymbol{w}} - A_{\boldsymbol{\lambda}} \tag{2.43}$$

Σκοπός μας είναι η ελαχιστοποίηση της ευκλείδιας νόρμας $\Delta = \frac{1}{\sigma^2}||E((q))||_2^2$ της εικόνας σφάλματος όπου σ^2 είναι η διακύμανση του μοντέλου θορύβου. Έχουν δημιουργηθεί αρκετοί αλγόριθμοι για την ελαχιστοποίηση του σφάλματος. Στο παρόν κείμενο και σύστημα χρησιμοποιείται ο αλγόριθμος *Variable-Order Template-Update Inverse-Compositional* (VoTuIc) πλήρης ανάλυση του οποίου μπορεί να βρεθεί στο [67].

2.2.3 Πρακτική Δημιουργία Active Appearance Μοντέλου

Ανάλυση Σχήματος. Μετά την απαραίτητη θεωρητική ανάλυση, θα δούμε τώρα την δημιουργία ενός active appearance μοντέλου, το οποίο δημιουργήσαμε χρησιμοποιώντας 268 frames από την βάση



Σχήμα 2.8: Παράδειγμα Σημαδοποιημένης Εικόνας

δεδομένων CVSP-AV (χωρίς συναίσθημα). Αρχικά προχωρήσαμε σε σημαδοποίηση των περιοχών ενδιαφέροντος του προσώπου (περίγραμμα, μάτια, στόμα, φρύδια, μύτη) στις 268 διαφορετικές εικόνες, όπως φαίνεται στο Σχήμα 2.8. Ο αριθμός των σημείων που χρησιμοποιήθηκαν συνολικά ήταν 51. Η κατανομή τους ως προς τις περιοχές ενδιαφέροντος του προσώπου φαίνονται στον Πίνακα 2.2. Η σημαδοποίηση αυτή έγινε με χρήση του εργαλείου am-tools [7] υλοποιημένο από τον Edward Cootes.

Περιοχή	Έξ. περίγραμμα στόματος	Εσ. περ. στόματος	Φρύδια	Μύτη	Μάτια	Περίγραμμα προσώπου
Αριθμός Σημείων	10	6	8 (4 × 2)	8	8 (4 × 2)	11

Πίνακας 2.2: Αριθμός σημείων που χρησιμοποιήθηκαν για κάθε περιοχή ενδιαφέροντος του προσώπου.

Εν συνεχεία προχωρήσαμε στην εκπαίδευση του active appearance μοντέλου [3], ακολουθώντας την θεωρητική διαδικασία που αναφέραμε προηγουμένως. Η μεγάλη σημασία της ευθυγράμμισης των σχημάτων είναι εμφανής, όπως μπορούμε να δούμε στο Σχήμα 2.9. Στο Σχήμα 2.10 φαίνεται και το Μέσο Σχήμα του προσώπου.

Στη συνέχεια προχωρήσαμε σε Ανάλυση σε Κύριες Συνιστώσες (PCA) για να βρούμε τα ιδιοσχήματα που περιγράφουν τη μεταβολή του Σχήματος. Στο Σχήμα 2.11 φαίνεται το ποσοστό της ενέργειας της κίνησης την οποία αντιπροσωπεύουν τα πρώτα 20 ιδιοσχήματα.

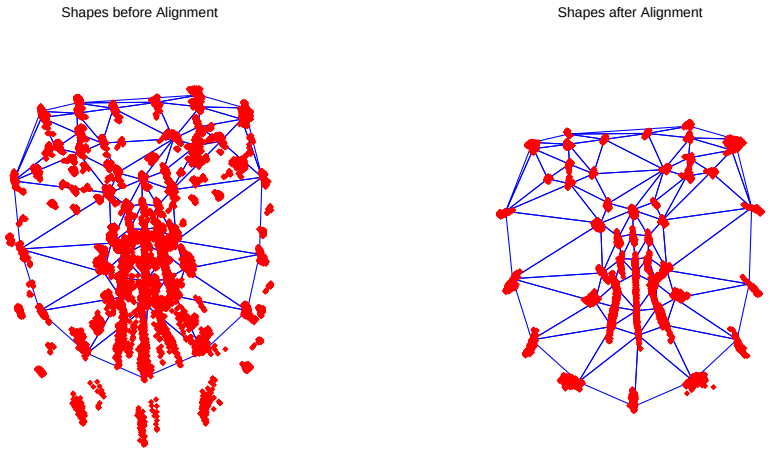
Από τα ιδιοσχήματα που βρέθηκαν με την Ανάλυση σε Κύριες Συνιστώσες στα επόμενα Σχήματα δείχνουμε τα 3 πρώτα (2.12, 2.13, 2.14), μαζί με την μεταβολή την οποία προκαλούν στο Μέσο Σχήμα.

Στο τελικό active appearance μοντέλο που εκπαιδεύτηκε κρατήθηκαν τα ιδιοσχήματα που επεξηγούν το 98% της μεταβολής του Σχήματος, $L = 10$ στον αριθμό.

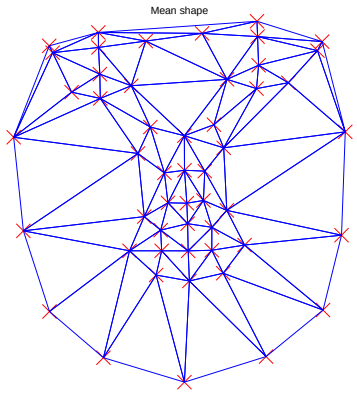
Ανάλυση Υφής. Αντιστοίχως, στα Σχήματα (2.15, 2.16, 2.17, 2.18, 2.19), φαίνονται τα αντίστοιχα διαγράμματα για την υφή και τα ιδιοδιανύσματα της υφής.

Στο τελικό active appearance μοντέλο που εκπαιδεύτηκε κρατήθηκαν τα ιδιοδιανύσματα υφής που επεξηγούν το 95% της μεταβολής του Σχήματος, $M = 32$ στον αριθμό.

Μετά την εκπαίδευση του active appearance μοντέλου έγινε η απαραίτητη εφαρμογή του μοντέλου στη βάση δεδομένων χωρίς συναίσθημα (περίπου 120, 000 εικόνες) για να λάβουμε τις παραμέτρους που θα χρησιμοποιηθούν για την εκπαίδευση του συνολικού συστήματος οπτικοακουστικής σύνθεσης φωνής. Για να αποφευχθεί η εσφαλμένη εύρεση των συντελεστών στις εικόνες που είχαν ακραίες



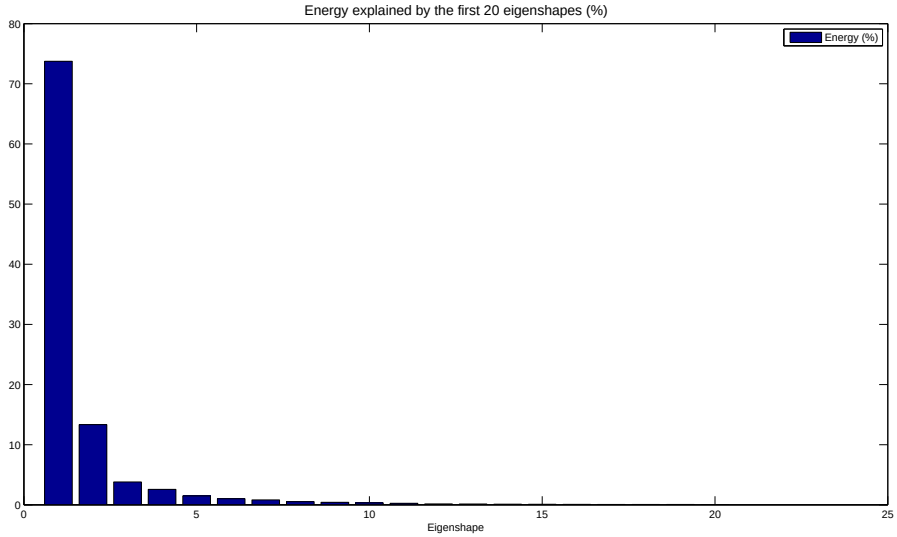
Σχήμα 2.9: Σχήματα Προσώπου πριν και μετά την ευθυγράμμιση (Επιτρεπόμενος Χώρος Σχήματος Προσώπου).



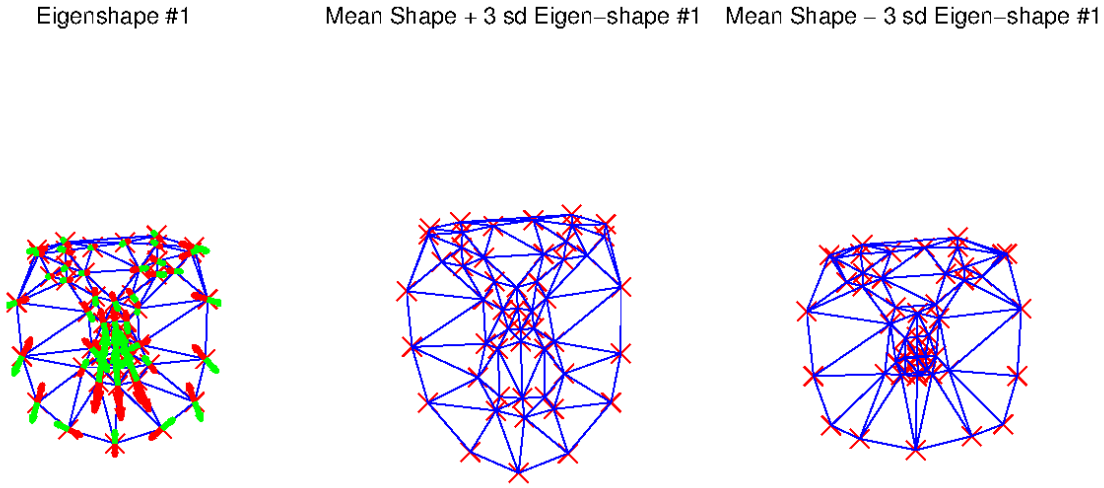
Σχήμα 2.10: Μέσο Σχήμα Προσώπου.

περιπτώσεις, στην περίπτωση που το εκτιμώμενο σφάλμα ξεπερνούσε ένα κατώφλι, τότε η συγκεκριμένη εικόνα έπερνε τις τιμές της προηγούμενης. Σε ακόμα πιο ακραίες περιπτώσεις όπου το σφάλμα συνέβαινε σε συνεχόμενες εικόνες (περίπου 3 προτάσεις) τότε δημιουργείται ασυνέχεια στο συνολικό video, που όμως ήταν προτιμώμενη από την εξαιρετική παραμόρφωση του προσώπου που θα αλλοίωνε την εκπαίδευση του συστήματος των κρυφών μαρκοβιανών μοντέλων.

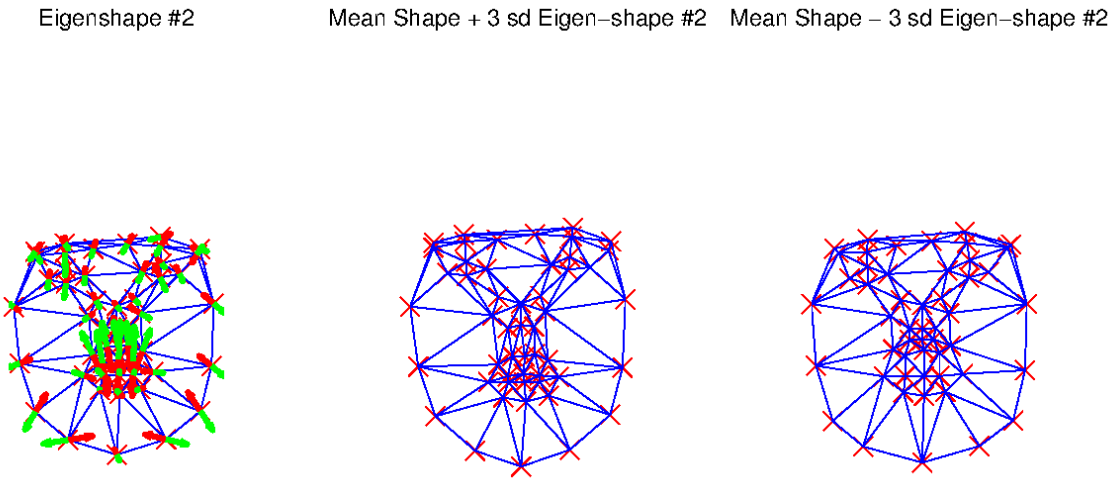
Η εφαρμογή αυτή έγινε με τη χρήση του αλγορίθμου VoTuIc που προαναφέραμε. Μετά το πέρας της διαδικασίας, οι παράμετροι που εξήχθησαν για κάθε μία από τις 900 διαφορετικές προτάσεις της βάσης δεδομένων αποθηκεύτηκαν σε δυαδική μορφή, κατάλληλη για την εκπαίδευση του τελικού συστήματος.



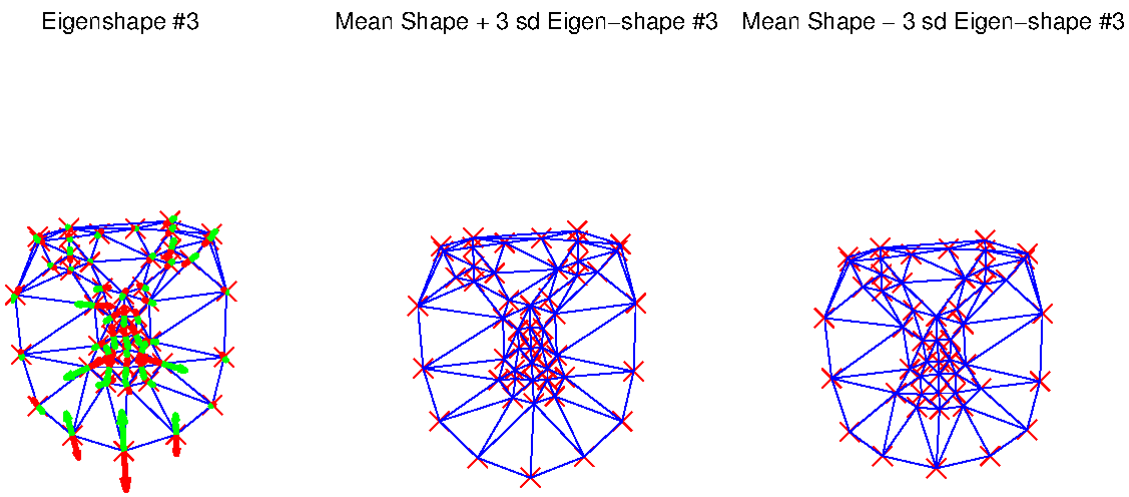
Σχήμα 2.11: Ενέργεια της μεταβολής που αντιπροσωπέυουν τα 20 πρώτα ιδιοσχήματα.



Σχήμα 2.12: Πρώτο ιδιοσχήμα και μεταβολή μέσου σχήματος.

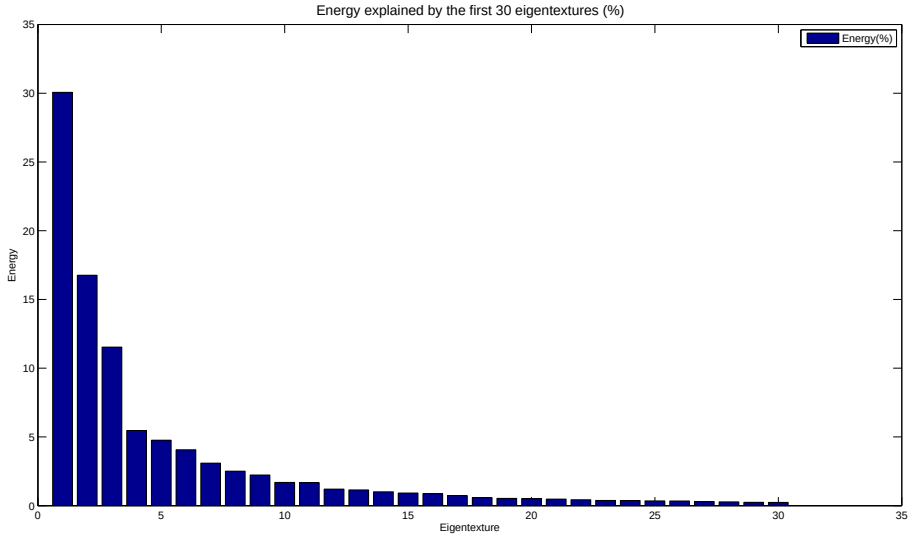


Σχήμα 2.13: Δεύτερο ιδιοσχήμα και μεταβολή μέσου σχήματος.

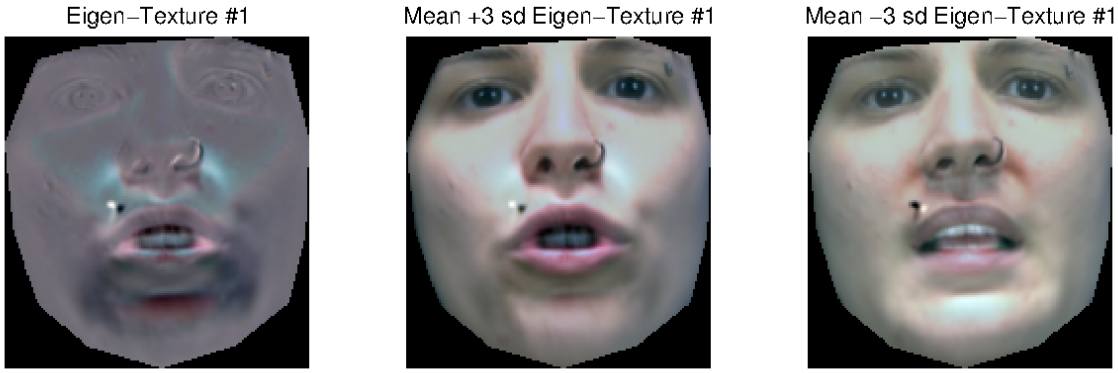


Σχήμα 2.14: Τρίτο ιδιοσχήμα και μεταβολή μέσου σχήματος.

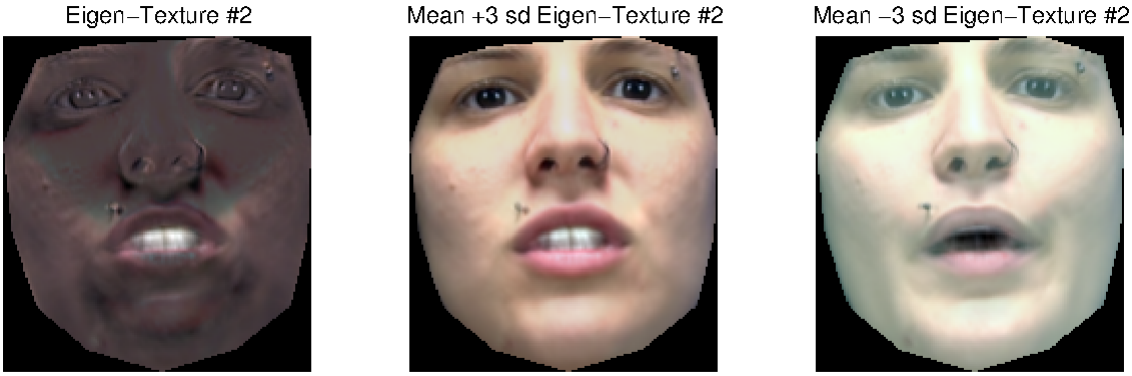
Σχήμα 2.15: Μέση υφή.



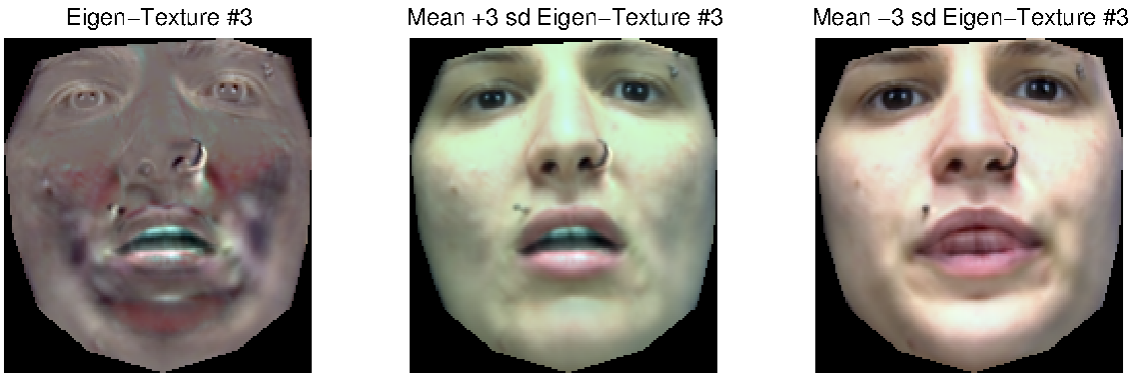
Σχήμα 2.16: Ενέργεια της μεταβολής που αντιπροσωπέυουν τα 30 πρώτα ιδιοδιανύσματα υφής.



Σχήμα 2.17: Πρώτο ιδιοδιάνυσμα υφής και μεταβολή μέσης υφής.



Σχήμα 2.18: Δεύτερο ιδιοδιάνυσμα υφής και μεταβολή μέσης υφής.



Σχήμα 2.19: Τρίτο ιδιοδιάνυσμα υφής και μεταβολή μέσης υφής.

Κεφάλαιο 3

Θεωρητική Ανάλυση Συστήματος Οπτικοακουστικής Σύνθεσης Φωνής με χρήση Κρυφών Μαρκοβιανών Μοντέλων

3.1 Εισαγωγή

Στο Κεφάλαιο αυτό θα προχωρήσουμε σε πλήρη θεωρητική ανάλυση του συνολικού συστήματος οπτικοακουστικής σύνθεσης φωνής με χρήση κρυφών μαρκοβιανών μοντέλων. Πριν όμως κάνουμε την θεωρητική αυτή ανάλυση θα παρουσιάσουμε συνοπτικά, αλλά και επαρκώς, τη λειτουργία ενός κρυφού μαρκοβιανού μοντέλου, όπως αυτό περιγράφεται από τους Rabiner & Huang [69], που αποτελεί και το κύριο δομικό συστατικό του οπτικοακουστικού συστήματος σύνθεσης φωνής που υλοποιούμε.

Στη συνέχεια, θα προχωρήσουμε στη θεωρητική ανάλυση του συνολικού συστήματος. Ως υπόβαθρο για την ανάλυση αυτή χρησιμοποιούμε την ανάλυση των Tokuda et al. [81] για ένα σύστημα *σύνθεσης* φωνής με χρήση κρυφών μαρκοβιανών μοντέλων, επεκτείνοντας την ταυτόχρονα για ένα σύστημα *οπτικοακουστικής σύνθεσης* φωνής.

3.2 Κρυφά Μαρκοβιανά Μοντέλα

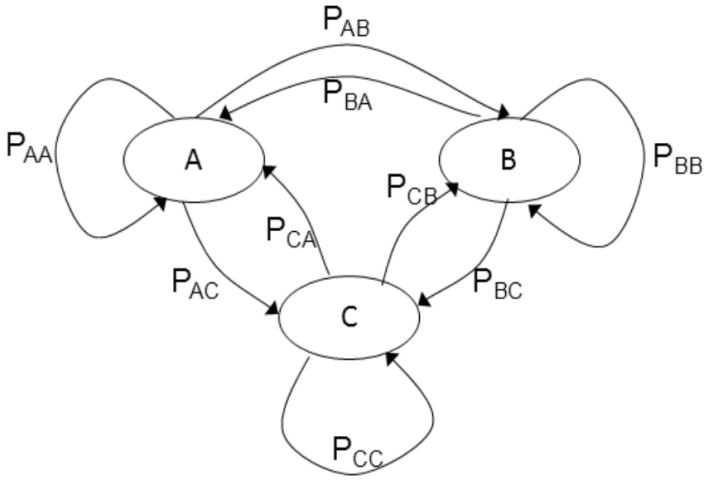
Με τον όρο κρυφά μαρκοβιανά μοντέλα (hidden markov models, hmm) αναφερόμαστε σε στατιστικές μαρκοβιανές διαδικασίες στις οποίες οι καταστάσεις είναι κρυφές στον παρατηρητή. Οι μαρκοβιανές διαδικασίες ικανοποιούν την μαρκοβιανή ιδιότητα, δηλαδή η επόμενη κατάσταση εξαρτάται μόνο από την τωρινή και όχι από τις προηγούμενες - δηλαδή η διαδικασία δεν έχει μνήμη. Η θεωρία των κρυφών μαρκοβιανών μοντέλων παρουσιάστηκε για πρώτη φορά από τον Αμερικανό Μαθηματικό Leonard E. Baum και τους συνεργάτες του σε μια σειρά από πέντε papers ([15, 13, 16, 14, 12]) από το 1966 έως το 1972. Από την παρουσίαση τους έως σήμερα έχουν αποδειχθεί πανίσχυρο εργαλείο σε πολλούς τομείς όπως

- Κρυπτανάλυση
- Αναγνώριση Φωνής
- Σύνθεση Φωνής
- Ανάλυση Χρονοσειρών
- Αναγνώριση Προτύπων
- Γονιδιακή Πρόβλεψη …

Πριν προχωρήσουμε στην παρουσίαση της γενικής θεωρίας των κρυφών μαρκοβιανών μοντέλων θα περιγράψουμε ένα απλούστερο μοντέλο - τη μαρκοβιανή αλυσίδα.

Ορισμός 1: *Μια μαρκοβιανή αλυσίδα περιγράφεται ως εξής: Έχουμε ένα σύνολο καταστάσεων* $S = \{1, 2, ..., N\}$. *Η διαδικασία εκκινά από μία κατάσταση σε κάποια χρονική στιγμή* t *και κινείται διαδοχικά από τη μία κατάσταση στην άλλη. Με* q_t *συμβολίζουμε την κατάσταση που βρίσκεται η μαρκοβιανή αλυσίδα τη χρονική στιγμή* t. *Όταν η διαδικασία βρίσκεται σε μία κατάσταση* i *τη χρονική στιγμή* t *τότε μεταβαίνει στην επόμενη κατάσταση* j *τη χρονική στιγμή* $t + 1$ *με πιθανότητα* p_{ij}. *Η πιθανότητα της μετάβασης εξαρτάται μόνο από την τωρινή κατάσταση και είναι ανεξάρτητη των προηγούμενων καταστάσεων από τις οποίες έχει κινηθεί η διαδικασία - μαρκοβιανή ιδιότητα. Σε μία μαρκοβιανή αλυσίδα οι καταστάσεις είναι παρατηρήσιμες και αποτελούν την έξοδο του συστήματος.*

Μία γραφική αναπαράσταση μιας μαρκοβιανής αλυσίδας μαζί με τις αντίστοιχες πιθανότητες σε μορφή μήτρας απεικονίζεται στο Σχήμα 3.1.



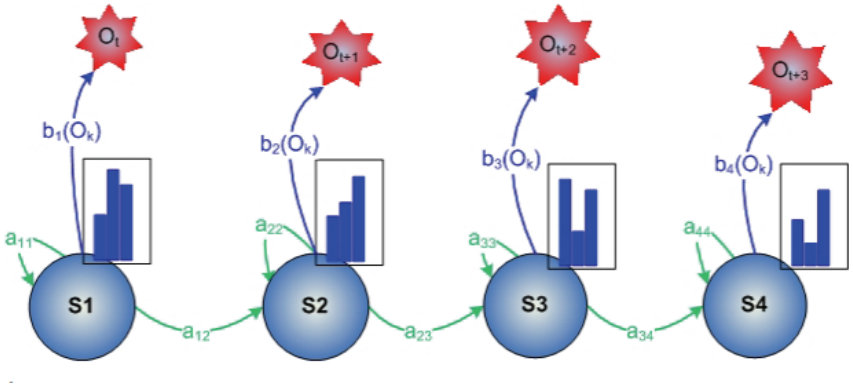
$$\mathbf{P}=\begin{bmatrix} p_{AA} & p_{AB} & p_{AC} \\ p_{BA} & p_{BB} & p_{BC} \\ p_{CA} & p_{CB} & p_{CC} \end{bmatrix}$$

Σχήμα 3.1: Μαρκοβιανή Αλυσίδα Τριών Καταστάσεων [71].

Με βάση τον ανώτερο ορισμό, μπορούμε τώρα να ορίσουμε το κρυφό μαρκοβιανό μοντέλο:

Ορισμός 2: *Ένα κρυφό μαρκοβιανό μοντέλο είναι μια διπλή στοχαστική διαδικασία όπου στο πρώτο επίπεδο έχουμε μια μαρκοβιανή αλυσίδα στην οποία οι μεταβάσεις ανάμεσα στις καταστάσεις γίνονται στοχαστικά μέσω του πίνακα μετάβασης καταστάσεων. Οι καταστάσεις από τις οποίες διέρχεται η αλυσίδα είναι κρυφές και δεν φαίνονται στην έξοδο του συστήματος. Σε δεύτερο επίπεδο, σε κάθε κατάσταση της αλυσίδας αντιστοιχεί μία κατανομή πιθανότητας με βάση την οποία επιλέγεται από την κατάσταση ως έξοδος ένα σύμβολο από ένα πεπερασμένο σύνολο* $V = \{v_1, v_2, ..., v_M\}$ *το οποίο ονομάζεται σύνολο συμβόλων παρατήρησης.*

Η δομή ενός κρυφού μαρκοβιανού μοντέλου φαίνεται στο Σχήμα 3.2.



Σχήμα 3.2: Παράδειγμα Κρυφού Μαρκοβιανού Μοντέλου [33]

3.3 Παράμετροι ενός Κρυφού Μαρκοβιανου Μοντέλου

Με βάση τον ανωτέρο ορισμό, ενα κρυφό μαρκοβιανό μοντέλο μπορεί να προσδιορισθεί πλήρως από τις εξής 5 παραμέτρους:

1. Τον αριθμό των διαφορετικών καταστάσεων N του μοντέλου που συμβολίζουμε ως $\{1, 2, ..., \}$.
2. Τον αριθμό των ξεχωριστών συμβόλων παρατήρησης που αντιστοιχούν στην έξοδο του συστήματος που μοντελοποιούμε. Συμβολίζουμε τα διαφορετικά σύμβολα ως $V = \{v_1, v_2, ..., v_M\}$.
3. Τον πίνακα μετάβασης καταστάσεων μεγέθους x με κάθε στοιχείο του πίνακα να ισούται

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

δηλαδή κάθε στοιχείο του πίνακα Aa_{ij} ισούται με την πιθανότητα η επόμενη κατάσταση να είναι η j δεδομένου ότι η τωρινή κατάσταση είναι i.

4. Την κατανομή πιθανότητας των συμβόλων παρατήρησης B

$$b_j(k) = P[o_t = v_k | q_t = j], \quad 1 <= k <= M$$

που ορίζει την κατανομή των συμβόλων στις καταστάσεις $j, j = 1, 2, ..., N$.

5. Την κατανομή πιθανότητας της αρχικής κατάστασης π

$$\pi_i = P[q_1 = i] \quad 1 <= i <= N$$

Από δω και στο εξής, όταν αναφερόμαστε σε ένα κρυφό μαρκοβιανό μοντέλο θα χρησιμοποιούμε την συμπαγή αναπαράσταση $\lambda = (A, B, \pi)$. Θεωρούμε εν γένει ότι οι πιθανότητες μετάβασης a_{ij} είναι χρονικά αμετάβλητες. Είναι εμφανές, με βάση τον ορισμό της πιθανότητας ότι ισχύει

$$\sum_{j=1}^{N} \alpha_{ij} = 1 \tag{3.1}$$

Η κατανομή πιθανότητας $b_j(k)$ εκφράζει την πιθανότητα (για διακριτά σύμβολα εξόδου) ή την πιθανοφάνεια (για συνεχείς κατανομές εξόδου) η κατάσταση j να παράγει την παρατήρηση $\boldsymbol{o}_t$. Προφανώς, και εδώ θα ισχύει

$$\sum_{\boldsymbol{o}} b_j(\boldsymbol{o}) = 1.0 \tag{3.2}$$

ή για συνεχή κατανομή

$$\int_{\boldsymbol{o}} b_j(\boldsymbol{o}) do = 1.0 \tag{3.3}$$

3.4 Προβλήματα Κρυφών Μαρκοβιανών Μοντέλων

Υπάρχουν τρία διαφορετικά προβλήματα που σχετίζονται με τα κρυφά μαρκοβιανά μοντέλα και που καλούμαστε να επιλύσουμε ώστε να τα χρησιμοποιήσουμε σε πραγματικές εφαρμογές [69]. Αυτά τα 3 προβλήματα είναι

1. **1ο Πρόβλημα** Δεδομένης της παρατηρούμενης ακολουθίας $O = (o_1 o_2 ... o_T)$ και ενός μοντέλου HMM $\lambda = (A, B, \pi)$ πως υπολογίζουμε αποδοτικά την πιθανότητα της παρατηρούμενης ακολουθίας $P(O|\lambda)$;
2. **2ο Πρόβλημα** Δεδομένης της παρατηρούμενης ακολουθίας $O = (o_1 o_2 ... o_T)$ και ενός μοντέλου HMM $\lambda = (A, B, \pi)$ πως επιλέγουμε την αντίστοιχη ακολουθία καταστάσεων $q = (q_1 q_2 ... q_T)$ που είναι βέλτιστη υπό μία έννοια (π.χ. εξηγεί καλύτερα της παρατηρήσεις);
3. **3ο Πρόβλημα** Πως προσαρμόζουμε τις παραμέτρους του μοντέλου $\lambda = (A, B, \pi)$ έτσι ώστε να μεγιστοποιήσουμε την $P(O|\lambda)$;

Στις επόμενες υποενότητες εξετάζουμε λεπτομερώς τη φύση καθενός από τα προβλήματα και επιδεικνύουμε αλγορίθμους για την επίλυση τους.

3.4.1 1ο Πρόβλημα-Πρόβλημα Αποτίμησης

Το πρώτο πρόβλημα αποτελεί ένα πρόβλημα αποτίμησης, το οποίο θα μπορούσαμε να το εκφράσουμε εναλλακτικά ως εξής: Δεδομένου ενός συνόλου μοντέλων λ καλούμαστε να επιλέξουμε το μοντέλο που ταιριάζει καλύτερα στην παρατηρούμενη ακολουθία.

Για μια δεδομένη ακολουθία καταστάσεων μήκους T $q = (q_1 q_2 ... q_T)$ και μια δεδομένη ακολουθία παρατηρήσεων $O = (o_1 o_2 ... o_T)$ η πιθανότητα $P(O, q|\lambda)$ δηλαδή η πιθανότητα να παρατηρήσουμε την ακολουθία με τη σταθερή ακολουθία καταστάσεων q δεδομένου του μοντέλου λ υπολογίζεται πολλαπλασιάζοντας τις πιθανότητες μετάβασης $\{\alpha_{1,2}, \alpha_{2,3}, ..., a_{T-1,T}\}$ με τις αντίστοιχες πιθανότητες εκπομπής των εισόδων παρατήρησης $\{b_1, b_2, ..., b_T\}$ καθώς και την πιθανότητα της αρχικής κατάστασης π_1, δηλαδή

$$P(O, q|\lambda) = \prod_{i=1}^{T} \alpha_{i-1,i} \dot{b}_i$$

όπου $\alpha_{0,1} = \pi_1$. Επομένως, αθροίζοντας τις πιθανότητες $P(O, q|\lambda)$ για κάθε πιθανή ακολουθία καταστάσεων παίρνουμε την επιθυμητή πιθανότητα

$$P(O, \lambda) = \sum_{\forall Q} P(P, q|\lambda)$$

Είναι προφανές ότι θα μπορούσαμε να υπολογίσουμε αυτή την ποσότητα με brute force, δηλαδή να υπολογίσουμε για όλες τις πιθανές ακολουθίες καταστάσεων την πιθανότητα να εκπέμψουν την ακολουθία συμβόλων μας και στη συνέχεια να τις προσθέσουμε. Όμως οι διαφορετικές αυτές ακολουθίες είναι N^T όπου ο αριθμός των διαφορετικών καταστάσεων επομένως μία τέτοια διαδικασία θα είχε πολυπλοκότητα της τάξης του .

Για την αποδοτική εύρεση της επιθυμητής πιθανότητας έχουν προταθεί δύο ισοδύναμοι αλγόριθμοι.

Ο forward αλγόριθμος ορίζει τη μεταβλήτη

$$\alpha_t(i) = P(o_1 o_2 ... o_t, q_t = i|\lambda)$$

που συμβολίζει την πιθανότητα να έχουμε την ακολουθία συμβόλων παρατήρησης $o_1 o_2 ... o_t$ τη χρονική στιγμή t και τη χρονική στιγμή t να βρισκόμαστε στην κατάσταση i δεδομένου του hmm λ. Αν λύσουμε επαγωγικά ως προς το $\alpha_t(i)$ από τη χρονική στιγμή 1 έχουμε:

1. **Initialization**

$$\alpha_1(i) = \pi_i b_i(\mathbf{o}_1), \qquad 1 \leq i \leq N \tag{3.4}$$

2. **Induction**

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\alpha_{ij}\right] b_j(\mathbf{o}_{t+1}), \qquad \begin{matrix} 1 \leq t \leq T-1 \\ 1 \leq j \leq N \end{matrix} \tag{3.5}$$

3. **Termination**

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{3.6}$$

Ο backward αλγόριθμος ορίζει τη μεταβλήτη

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

που συμβολίζει την πιθανότητα να έχουμε την ακολουθία συμβόλων παρατήρησης $\mathbf{o}_{t+1}\mathbf{o}_{t+2} ... \mathbf{o}_T$ δεδομένης της κατάστασης i τη χρονική στιγμή t και του μοντέλου λ. Αν λύσουμε και πάλι επαγωγικά ως προς το $\beta_t(i)$ παίρνουμε

1. **Initialization**

$$\beta_T(i) = 1, \qquad 1 \leq i \leq N \tag{3.7}$$

2. **Induction**

$$\beta_t(j) = \sum_{j=1}^{N} \alpha_(ij) b_j(\mathbf{o}_{t+1})\beta_{t+1}(j), \qquad \begin{matrix} t = T-1, T-2, ..., 1 \\ 1 \leq i \leq N \end{matrix} \tag{3.8}$$

Όπως μπορούμε να δούμε, και στους δύο αλγόριθμους, η υπολογιστική πολυπλοκότητα του πιο απαιτητικού τμήματος (της επαγωγής) είναι $O\left(N^2 T\right)$, άρα και του αλγορίθμου, είναι σημαντικά μικρότερη από την μέθοδο brute force με πολυπλοκότητα $O\left(N^T T\right)$.

3.4.2 2ο Πρόβλημα-Εύρεση Βέλτισης Ακολουθίας

Στο δεύτερο πρόβλημα προσπαθούμε να αποκαλύψουμε - αποκωδικοποιήσουμε το κρυφό μέρος του μοντέλου δηλαδή την ακολουθία των καταστάσεων που εξηγεί καλύτερα τις παρατηρήσεις και κανονικά είναι κρυφή στον παρατηρητή.

Η επίλυση του δεύτερου προβλήματος, δηλαδή της πιο πιθανής ακολουθίας καταστάσεων γίνεται με χρήση του αλγορίθμου δυναμικού προγραμματισμού Viterbi. Ο αλγόριθμος αυτός προτάθηκε από τον Andrew Viterbi το 1967 [38] και έκτοτε έχει χρησιμοποιηθεί σε πολλά προβλήματα μεγιστοποίησης που περιλαμβάνουν πιθανότητες, όπως π.χ. για την αποκωδικοποίηση των συνελικτικών κώδικων σε διάφορα συστήματα κινητής τηλεφωνίας.

Ο αλγόριθμος Viterbi χρησιμοποιεί την ποσότητα

$$\delta_t(i) = \max_{q_1,q_2,...,q_{t-1}} P[q_1 q_2 ... q_{t-1}, q_t = i, \mathbf{o}_1 \mathbf{o}_2 ... \mathbf{o}_t | \lambda]$$

. Όπως είναι προφανές, το $\delta_t(i)$ συμβολίζει τη μέγιστη πιθανότητα ενός μονοπατιού μήκους t που τελειώνει στην κατάσταση i, λαμβάνοντας υπ' όψιν τις πρώτες t παρατηρήσεις. Με βάση τον ορισμό αυτό με επαγωγή παίρνουμε

$$\delta_{t+1}(j) = [\max_i \delta_t(i)\alpha_{ij}]\dot{b}_j((o)_{t+1})$$

Τώρα επομένως μπορούμε να δηλώσουμε τη διαδικασία του αλγορίθμου Viterbi ως εξής:

1. **Initialization**

$$\delta_1 i = \pi_i b_i(\mathbf{o}_t), \qquad 1 \leq i \leq N \tag{3.9}$$

$$\psi_1(i) = 0, \qquad 1 \leq i \leq N \tag{3.10}$$

2. **Induction**

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i)\alpha_{ij}]\, b_j(\mathbf{o}_t), \qquad \begin{matrix} 2 \leq t \leq T \\ 1 \leq j \leq N \end{matrix} \tag{3.11}$$

$$\psi_t(j) = \arg\max_{1 \leq i \leq N} [\delta_{t-1}(i)\alpha_{ij}], \qquad \begin{matrix} 2 \leq t \leq T \\ 1 \leq j \leq N \end{matrix} \tag{3.12}$$

3. **Termination**

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)] \tag{3.13}$$

$$q_T^* = \arg\max_{1 \leq i \leq N} [\delta_T(i)] \tag{3.14}$$

4. **Path Backtracking**

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, ..., 1 \tag{3.15}$$

Είναι σημαντικό να σημειώσουμε ότι με τη εισαγωγή λογαρίθμων στον αλγόριθμο Viterbi απαλοίφεται η ανάγκη πολλαπλασιασμών και απαιτούνται αθροίσεις της τάξης N^2T:

1. **Preprocessing**

$$\tilde{\pi}_i = \log \pi_i, \qquad 1 \leq i \leq N \tag{3.16}$$

$$\tilde{b}_i(\mathbf{o}_t) = \log b_i(\mathbf{o}_t), \qquad 1 \leq i \leq N 1 \leq t \leq T \tag{3.17}$$

$$\tilde{\alpha}_{ij} = \log a_{ij}, \qquad 1 \leq i, j \leq N \tag{3.18}$$

2. **Initialization**

$$\tilde{\delta}_1 i = \tilde{\pi}_i + \tilde{b_i}(\mathbf{o}_t), \qquad 1 \leq i \leq N \tag{3.19}$$

$$\psi_1(i) = 0, \qquad 1 \leq i \leq N \tag{3.20}$$

3. **Induction**

$$\tilde{\delta}_t(j) = \max_{1 \leq i \leq N} \left[\tilde{\delta}_{t-1}(i) + \tilde{\alpha}_{ij} \right] + \tilde{b}_j(\mathbf{o}_t), \qquad \begin{matrix} 2 \leq t \leq T \\ 1 \leq j \leq N \end{matrix} \tag{3.21}$$

$$\psi_t(j) = \arg\max_{1 \leq i \leq N} \left[\tilde{\delta}_{t-1}(i) + \tilde{\alpha}_{ij} \right], \qquad \begin{matrix} 2 \leq t \leq T \\ 1 \leq j \leq N \end{matrix} \tag{3.22}$$

4. **Termination**

$$\tilde{P}^* = \max_{1 \leq i \leq N} \left[\tilde{\delta}_T(i) \right] \tag{3.23}$$

$$q_T^* = \arg\max_{1 \leq i \leq N} \left[\tilde{\delta}_T(i) \right] \tag{3.24}$$

5. **Path Backtracking**

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, ..., 1 \tag{3.25}$$

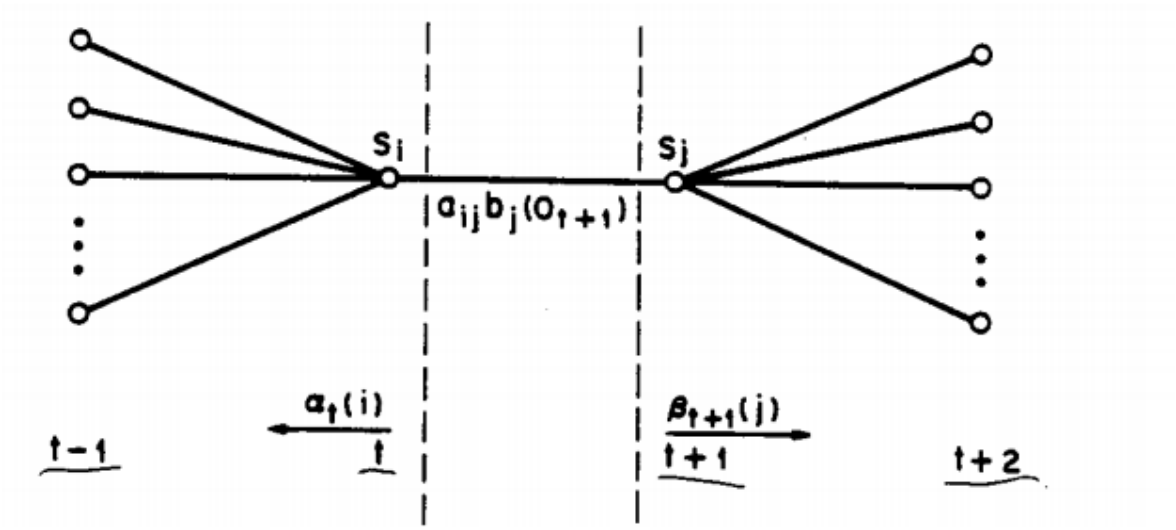
3.4.3 3ο Πρόβλημα-Εκτίμηση παραμέτρων

Στο τρίτο πρόβλημα προσπαθούμε να βελτιστοποιήσουμε τις παραμέτρους του μοντέλου για να εξηγήσουμε όσο το δυνατόν καλύτερα την ακολουθία σύμβολων παρατήρησης. Αυτή η ακολουθία χρησιμοποιείται για την εκπαίδευση του μοντέλου - ένα από τα σημαντικότερα χαρακτηριστικά των κρυφών μαρκοβιανών μοντέλων. Το πρόβλημα αυτό είναι μη γραμμικό, πολλών διαστάσεων, και δεν υπάρχει αναλυτική διαδικασία για την επίλυση του, δηλαδή την εύρεση των παραμέτρων λ έτσι ώστε να μεγιστοποιείται η πιθανότητα $P(\mathbf{O}\lambda)$ για μια δεδομένη ακολουθία παρατηρήσεων $\mathbf{O}$, καθώς υπάρχουν πολλά τοπικά μέγιστα που καθιστούν δύσκολη την εύρεση του ολικού. Για την εύρεση ενός μοντέλου που να μεγιστοποιεί τοπικά την επιθυμητή πιθανότητα έχει προταθεί ο επαναληπτικός αλγόριθμος Baum-Welch, γνωστός και ως EM (Expectation Maximization). Επίσης χρησιμοποιούνται και μέθοδοι gradient descent.

Για να περιγράψουμε τον αλγόριθμο EM, αρχικά ορίζουμε την μεταβλητή $\xi_t(i, j)$, ως την πιθανότητα να βρίσκεται το μοντέλο τη χρονική στιγμή t στην κατάσταση i και τη χρονική στιγμή $t + 1$ στην κατάσταση j:

$$\xi_t(i, j) = P(q_t = i, q_{t+1} = j | \mathbf{O}, \lambda) \tag{3.26}$$

Τα μονοπάτια που ικανοποιούν τις συνθήκες της εξίσωσης 3.26 απεικονίζονται στο Σχήμα 3.3.



Σχήμα 3.3: Μονοπάτι που ικανοποιεί τις συνθήκες της εξίσωσης 3.26 [69].

Με βάση τους ορισμούς των forward και backward μεταβλητών που δώσαμε για την επίλυση του Προβλήματος 1, μπορούμε να γράψουμε το $\xi_t(i,j)$ στη μορφή

$$\xi_t(i,j) = \frac{P(q_t = i, q_{t+1} = j, \mathbf{O}|\lambda)}{P(\mathbf{O}|\lambda)} = \frac{\alpha_t(i)\alpha_{ij}b_j\left(\mathbf{O}_{t+1}\right)\beta_{t+1}\left(j\right)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t\left(i\right)\alpha_{ij}b_j\left(\mathbf{o}_{t+1}\right)\beta_{i+1}\left(j\right)} \tag{3.27}$$

Αν τώρα ορίσουμε την $\gamma_t\left(i\right)$ ως την πιθανότητα να είναι το μοντέλο στην κατάσταση i τη χρονική στιγμή t, δεδομένης ολόκληρης της ακολουθίας παρατήρησης και του μοντέλου, μπορούμε να γράψουμε:

$$\gamma_t\left(i\right) = \sum_{i=1}^{N}\xi_t\left(i,h\right) \tag{3.28}$$

Τώρα, αθροίζοντας ως προς τον χρόνο t, παίρνουμε μια ποσότητα την οποία μπορούμε να ερμηνεύσουμε ως τον αναμενόμενο αριθμό μεταβάσεων από την κατάσταση i, αν εξαιρέσουμε τη χρονική στιγμή $t = T$. Ομοίως, αν αθροίσουμε το $\xi_t\left(i,j\right)$ ως προς το χρόνο από τη χρονική στιγμή $t = 1$ έως τη χρονική στιγμή $t = T-1$ παίρνουμε τον αναμενόμενο αριθμό μεταβάσεων από την κατάσταση i στην κατάσταση .

Με βάση τα ανωτέρω τώρα, ακολουθούν τώρα εξισώσεις για την επανεκτίμηση των παραμέτρων του κρυφού μαρκοβιανού μοντέλου:

$$\bar{\pi}_i = \textit{expected frequency (number of times) in state i at time}(t=1) = \gamma_1(i) \tag{3.29}$$

$$\bar{\alpha}_{ij} = \frac{\textit{expected number of transitions from state i to state j}}{\textit{expected number of transitions from state i}} = \frac{\sum_{t=1}^{T-1}\xi_t\left(i,j\right)}{\sum_{t=1}^{T-1}\gamma_t\left(i\right)} \tag{3.30}$$

$$\bar{b}_j\left(k\right) = \frac{\textit{expected number of states in state j and observing symbol } \mathbf{v}_k}{\textit{expected number of states in state i}} = \frac{\sum_{t=1,\mathbf{o}_t=\mathbf{v}_k}^{T-1}\gamma_t\left(j\right)}{\sum_{t=1}^{T-1}\gamma_t\left(i\right)} \tag{3.31}$$

Αν ορίσουμε το τρέχον μοντέλο ως $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$ και το ενημερωμένο μοντέλο μετά που παίρνουμε από τις ανωτέρω εξισώσεις ως $\bar{\lambda} = \left(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \bar{\pi}\right)$ τότε έχει αποδειχθεί από τον Baum και τους συνεργάτες του ότι ισχύει

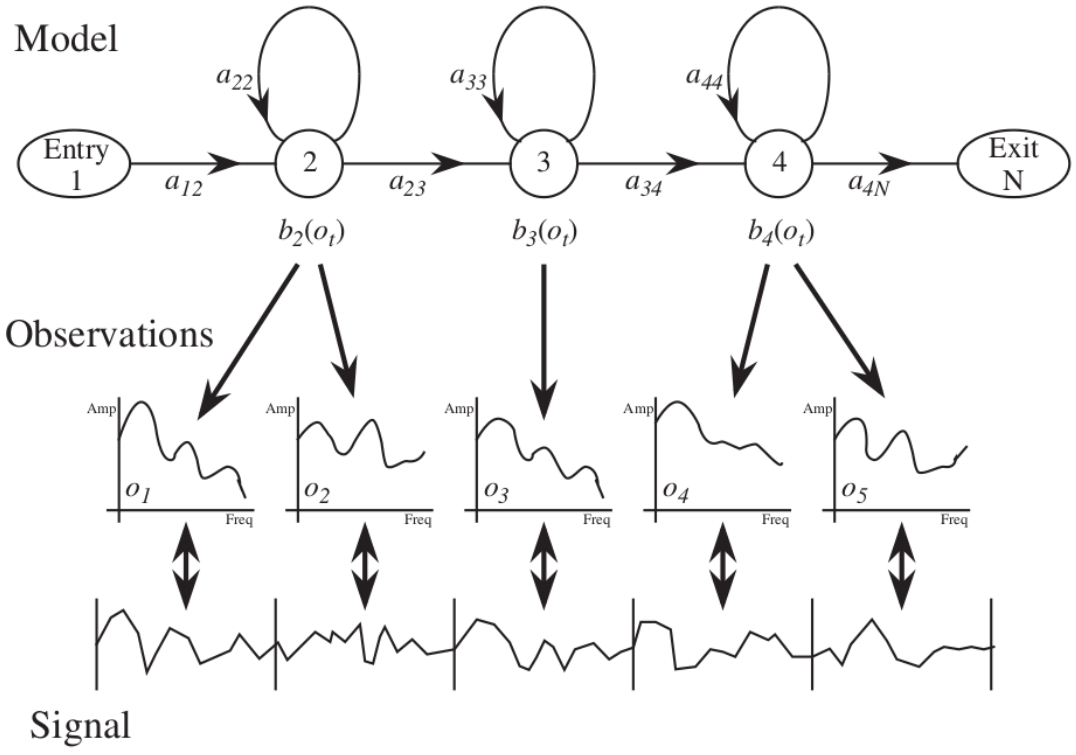
$$P\left(\mathbf{O}|\bar{\lambda}\right) \geq P\left(\mathbf{O}|\lambda\right)$$

δηλαδή, ότι το νέο μοντέλο που προέκυψε είναι πιθανότερο να έχει παράξει το διάνυσμα παρατήρησης.

Επομένως, με βάση την διαδικασία που μόλις αναφέραμε με επαναληπτική χρήση των εξισώσεων θα οδηγηθούμε σε κάποιο σημείο όπου θα πάρουμε την προσέγγιση μεγίστης πιθανοφάνειας του κρυφού μαρκοβιανού μοντέλου. Περισσότερες πληροφορίες όσον αφορά την παραγωγή των επαναληπτικών εξισώσεων, με χρήση της βοηθητικής συνάρτησης του Baum, μπορούμε να βρούμε στο [69].

3.5 Κρυφά Μαρκοβιανά Μοντέλα για Σύνθεση Φωνής και Εικονοσειράς

Τα κρυφά μαρκοβιανά μοντέλα που χρησιμοποιούνται για σύνθεση φωνής ή/και εικονοσειράς, τροποποιούνται σε σχέση με τα γενικά μαρκοβιανά μοντέλα που παρουσιάσαμε προηγουμένως [60]. Έχουν δύο συγκεκριμένες καταστάσεις, την κατάσταση εισόδου, και την κατάσταση εξόδου, οι οποίες δεν παράγουν παρατηρήσεις. Πριν την διαδικασία παραγωγής παρατήρησης το μοντέλο βρίσκεται στην κατάσταση εισόδου, ενώ στο τέλος της διαδικασίας το μοντέλο φτάνει στην κατάσταση εξόδου. Επίσης, εφ' όσον η ομιλία αποτελεί μια ταξινομημένη ακολουθία φωνημάτων, χρησιμοποιείται αριστερή-προς-δεξιά τοπολογία δηλαδή ο πίνακας των πιθανοτήτων μετάβασης είναι άνω τριγωνικός και από μία κατάσταση j μπορούμε να βρεθούμε μόνο σε καταστάσεις k για τις οποίες ισχύει $j \leq k$. Ένα παράδειγμα τέτοιου μαρκοβιανού μοντέλου που χρησιμοποιείται για σύνθεση φωνής φαίνεται στο Σχήμα 3.4.



Σχήμα 3.4: Κρυφό μαρκοβιανό μοντέλο για σύνθεση-αναγνώριση φωνής [60].

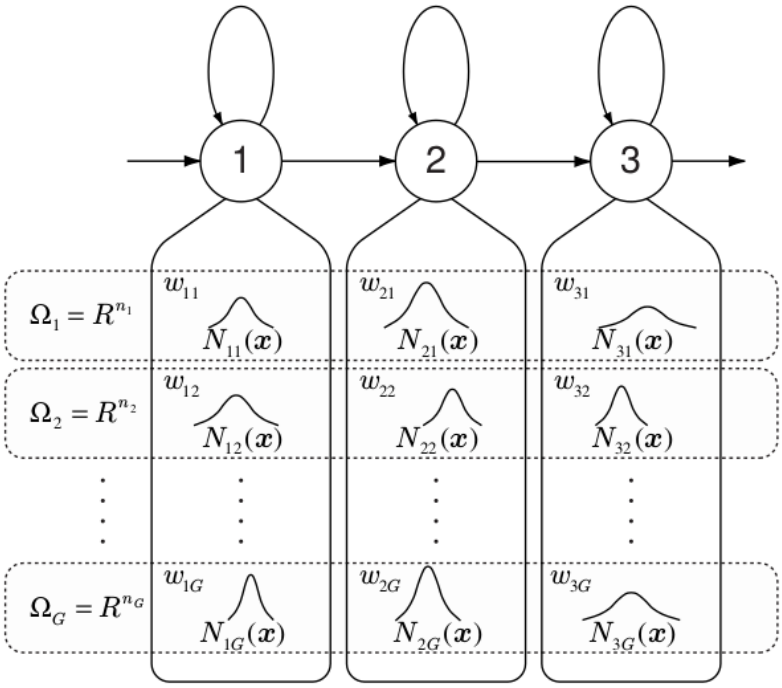
Γενικά, η πιο ευρέως διαδεδομένη συνεχής συνάρτηση για την μοντελοποίηση της κατανομής συνεχών μεταβλητών είναι η περίφημη Γκαουσσιανή κατανομή. Παρ' όλα αυτά, η συνάρτηση αυτή πολλές φορές αδυνατεί να μοντελοποιήσει σύνολα πραγματικών δεδομένων [18]. Για την επίλυση αυτού του προβλήματος, χρησιμοποιούμε *υπέρθεση* πολυμεταβλητών Γκασουσσιανών κατανομών ή όπως καλείται ένα "μείγμα" (mixture) από πολυμεταβλητές Γκαουσσιανές συναρτήσεις:

$$b_{jm}(\boldsymbol{o}_t) = N(\boldsymbol{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_{jm}|}} e^{\frac{-1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})' \boldsymbol{\Sigma}_{\boldsymbol{jm}}^{-1} (\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})} \tag{3.32}$$

$$b_j(\boldsymbol{o}_t) = \sum_{m=1}^{M} c_{jm} b_{jm}(\boldsymbol{o}_t) \tag{3.33}$$

όπου d είναι η διαστατικότητα των δεδομένων και c_{jm}, $\boldsymbol{\mu}_{jm}$, $\boldsymbol{\Sigma}_{jm}$ είναι το βάρος, το διάνυσμα μέσων τιμών, και η μήτρα συνδιακύμανσης της m-οστής συνιστώσας του μείγματος των Γκαουσσιανών συναρτήσεων της κατάστασης j. Είναι προφανές ότι όσο μεγαλύτερος είναι ο αριθμός του πλήθους των πολυμεταβλητών Γκαουσσιανών συναρτήσεων, τόσο μεγαλύτερη είναι και η ακρίβεια με την οποία προσεγίζεται η κατανομή ενός πραγματικού συνόλου δεδομένων.

Μαρκοβιανό Μοντέλο Κατανομής Πολλαπλών Χώρων Αναφέραμε ήδη ότι στο Κεφάλαιο 2, χρησιμοποιούμε επίσης και τη θεμελιώδη συχνότητα της φωνής $f0$ ως τη διέγερση για το φίλτρο σύνθεσης. Η θεμελιώδης συχνότητα αυτή, λαμβάνει ρητές τιμές, στο διάστημα του σήματος όπου έχουμε φωνούμενη ομιλία, και δεν ορίζεται στα διαστήματα αυτά. Το γεγονός αυτό καθιστά αδύνατη την μοντελοποίηση της από ένα κλασικό μαρκοβιανό μοντέλο (συνεχές ή διακριτό). Για την υπέρβαση αυτού του προβλήματος χρησιμοποιούμε έναν διαφορετικό τύπο κρυφού μαρκοβιανού μοντέλου, το μαρκοβιανό μοντέλο κατανομής πολλαπλών χώρων (multi-space probability distribution hidden markov model, MSD-HMM), το οποίο επιτρέπει σε κάθε κατάσταση των μαρκοβιανών μοντέλων την ύπαρξη περισσότερων από μία κατανομών εξόδου - μία για την περίπτωση ρητής θεμελιώδους συχνότητας και μία για την περίπτωση μη ορισμένης θεμελιώδους συχνότητας. Τα MSD-HMM περιγράφονται αναλυτικά στο [55], ενώ ένα παράδειγμα λειτουργίας τους φαίνεται στο Σχήμα 3.5



Σχήμα 3.5: Κρυφό Μαρκοβιανό Μοντέλο με χρήση πολυδιάστατης κατανομής πιθανότητας [55].

3.6 Οπτικοακουστική Σύνθεση Φωνής με Κρυφά Μαρκοβιανά Μοντέλα

Η δομή του συστήματος σύνθεσης φωνής και εικόνας με χρήση κρυφών μαρκοβιανών μοντέλων που υλοποιούμε φαίνεται στο Σχήμα 3.6. Το σύστημα χωρίζεται σε δύο υποκατηγορίες, το *Τμήμα Εκπαίδευσης* και το *Τμήμα Σύνθεσης*.

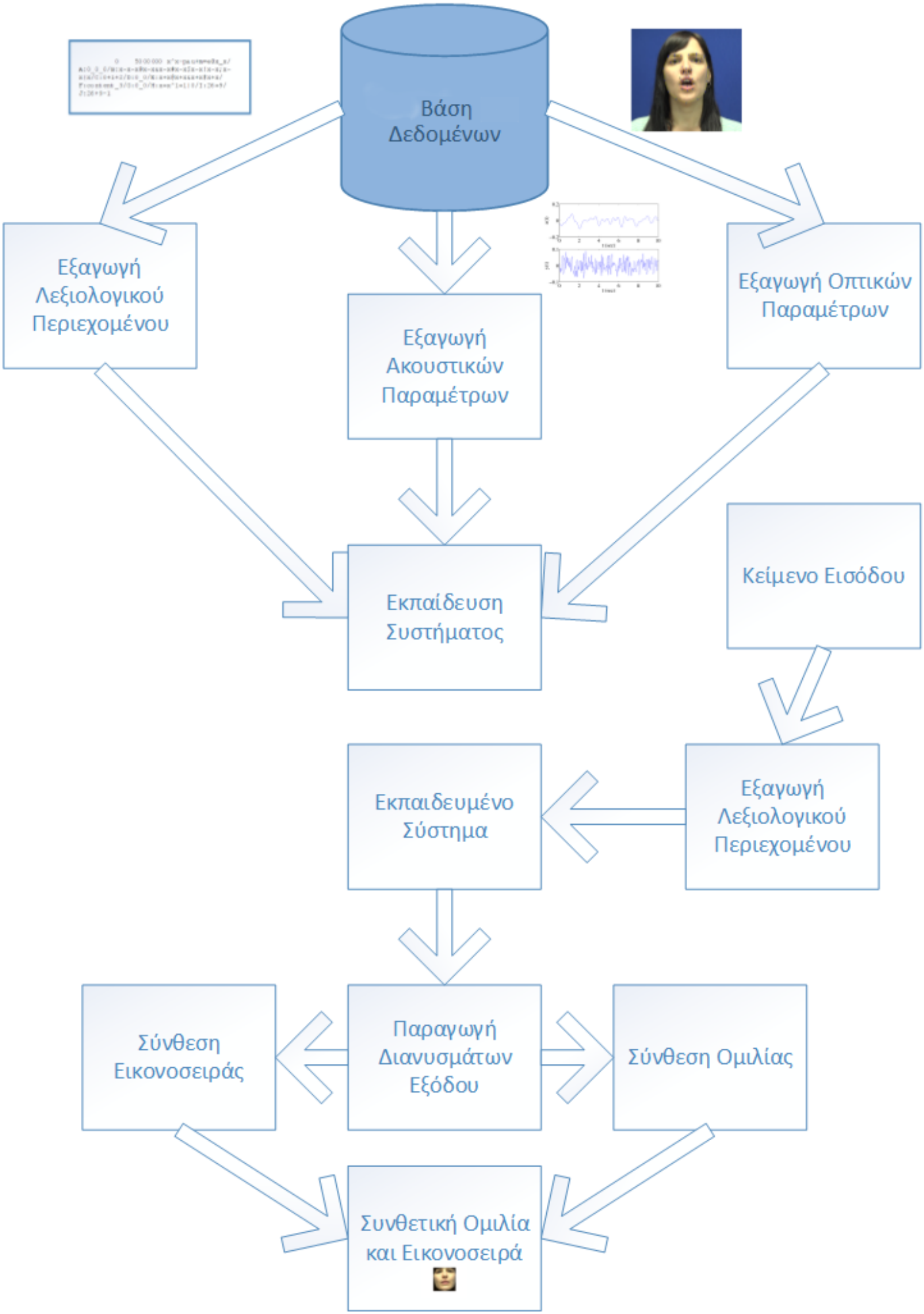
3.6.1 Τμήμα Εκπαίδευσης

Το τμήμα εκπαίδευσης του συστήματος περιλαμβάνει την εξαγωγή χαρακτηριστικών από την βάση δεδομένων CVSP-AV, όπως περιγράψαμε στο Κεφάλαιο 2, ώστε να χρησιμοποιηθούν ως διάνυσμα παρατήρησης για την εκπαίδευση των κρυφών μαρκοβιανών μοντέλων. Στο ίδιο τμήμα γίνεται και η κατάλληλη λεξιλογική ανάλυση της αντίστοιχης προσωδίας των χαρακτηριστικών που θα χρησιμοποιηθούν επίσης για την εκπαίδευση.

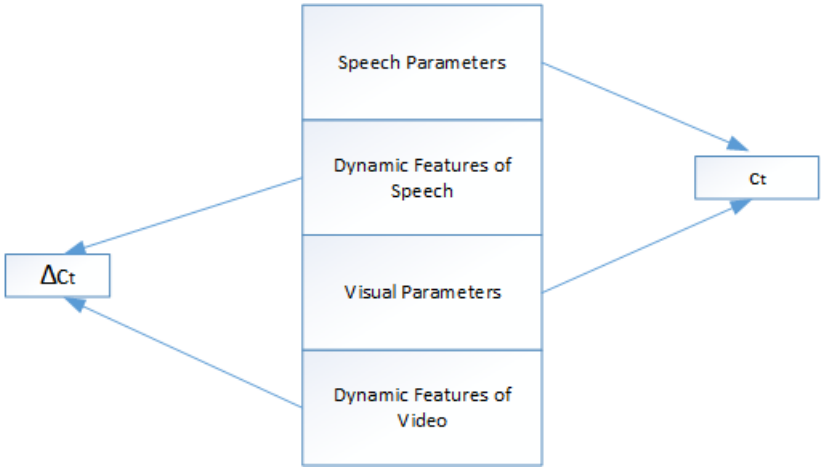
Στην γενική μορφή του, το τμήμα αυτό αναλαμβάνει την εκπαίδευση κρυφών μαρκοβιανών μοντέλων, με κάθε μοντέλο να αντιστοιχεί και σε ένα φώνημα. Τα μοντέλα εκπαιδεύονται με χρήση του αλγορίθμου Baum-Welch χρησιμοποιώντας τα δεδομένα από τη βάση μας.

Observation Vector

Το διάνυσμα παρατήρησης που χρησιμοποιείται στο σύστημα οπτικοακουστικής σύνθεσης φωνής περιέχει τις παραμέτρους εκείνες του ήχου ή της εικόνας, που χαρακτηρίζουν όσο το δυνατόν περισσότερο τα αντίστοιχα μονοδιάστατα και διδιάστατα σήματα, και μπορούν να χρησιμοποιηθούν για την κατασκευή ενός νέου σήματος ομιλίας ή εικόνας. Το διάνυσμα που χρησιμοποιούμε έχει τη



Σχήμα 3.6: Δομή Συστήματος Οπτικοακουστικής Σύνθεσης Φωνής



Σχήμα 3.7: Διάνυσμα Παρατήρησης Συστήματος Οπτικοακουστικής Σύνθεσης Φωνής με Κρυφά Μαρκοβιανά Μοντέλα.

μορφή του Σχήματος 3.7. Παρατηρούμε ότι το διάνυσμα χωρίζεται σε δύο κύρια τμήματα. Το πρώτο τμήμα περιέχει τις εξαχθείσες παραμέτρους από τη βάση δεδομένων. Το δεύτερο τμήμα, περιέχει τα δυναμικά χαρακτηριστικά που υπολογίζονται μεταξύ των ενός αριθμού γειτονικών πλαισίων. Τα δυναμικά αυτά χαρακτηριστικά χρησιμοποιούνται για τον υπολογισμό της χρονικής μεταβολής στο χρόνο. Αν δεν συμπεριλάβουμε στο διάνυσμα παρατήρησης τα δυναμικά αυτά χαρακτηριστικά οδηγούμαστε σε μία step-wise ακολουθία με ασυνέχειες, γεγονός που οδηγεί με τη σειρά του σε μη φυσική ομιλία-ακολουθία εικόνων, καθώς αυτές μεταβάλλονται ομαλά κατά τη ροή του λόγου. Επομένως, αν θεωρήσουμε το διάνυσμα $\boldsymbol{c}_t$ που περιέχει τις παραμέτρους (ακουστικές και οπτικές) που εξάγουμε από ένα πλαίσιο τη στιγμή t, τα δυναμικά χαρακτηριστικά πρώτης τάξης στο πλαίσιο αυτό υπολογίζονται ως συντελεστές παλινδρόμησης από τις γειτονικές στατικές παραμέτρους:

$$\Delta \boldsymbol{c}_t = \sum_{\tau=-L}^{L} w(\tau) \boldsymbol{c}_{t+\tau} \tag{3.34}$$

όπου $w(\tau)_{\tau=-L}^{L}$ οι συντελεστές του παραθύρου που χρησιμοποιείται για τον υπολογισμό των δυναμικών παραμέτρων. Η παραπάνω περίπτωση γενικεύεται και για τη χρήση δυναμικών χαρακτηριστικών μεγαλύτερης τάξης με τον ίδιο τρόπο. Επομένως, το συνολικό διάνυσμα παρατήρησης που χρησιμοποιείται για την εκπαίδευση του συστήματος είναι σε μαθηματική μορφή:

$$\boldsymbol{o}_t = [\boldsymbol{c}_t, \Delta \boldsymbol{c}_t] \tag{3.35}$$

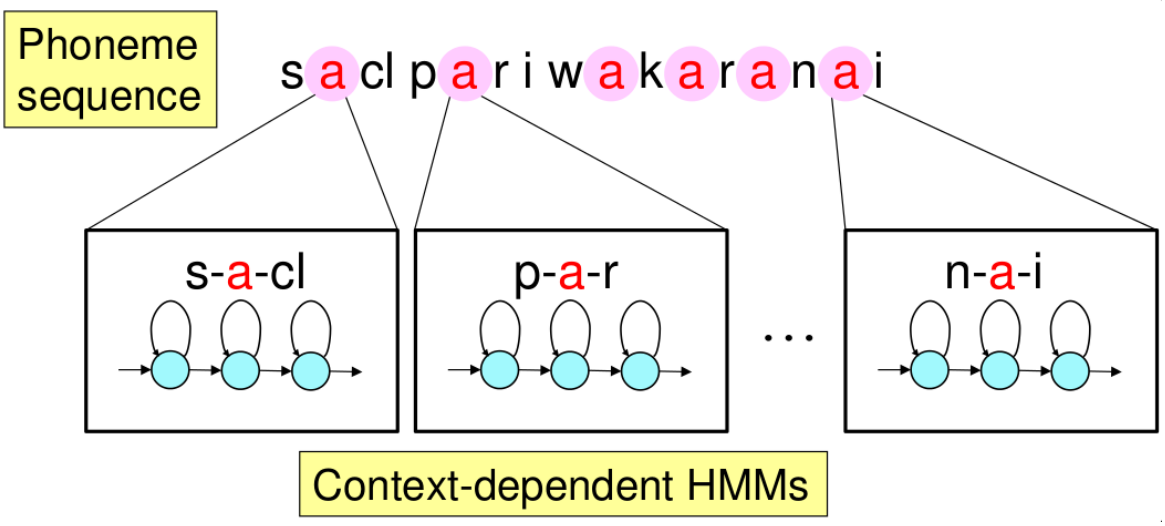
Τώρα, με βάση την ανάλυση των παραμέτρων που έγινε στο Κεφάλαιο 2, το συνολικό διάνυσμα παρατήρησης αναλυτικά δίνεται από:

$$\boldsymbol{o}_t = \begin{bmatrix} \mathbf{c}_{m,t} \\ \Delta \boldsymbol{c}_{m,t} \\ f0_t \\ \Delta f0_t \\ \mathbf{c}_{b,t} \\ \Delta \mathbf{c}_{b,t} \\ \mathbf{c}_{s,t} \\ \Delta \mathbf{c}_{s,t} \\ \mathbf{c}_{tx,t} \\ \Delta \mathbf{c}_{tx,t} \end{bmatrix} \tag{3.36}$$

όπου συμβολίζουμε τους πρώτους M συντελεστές mgc ως $\mathbf{c}_m$, την θεμελιώδη συχνότητα ως $f0$, τους συντελεστές απεριοδικότητας ζώνης ως $\mathbf{c}_{b,t}$, τα βάρη των ιδιοδιανύσμάτων σχήματος που επεξηγούν ένα μεγάλο μέρος της μεταβολής του σχήματος του προσώπου (όπως αναφέραμε στην Υποενότητα 2.2.3) ως $\mathbf{c}_{s,t}$ και βάρη των ιδιοδιανύσματων υφής που επεξηγούν ένα μεγάλο ποσοστό της μεταβολής της υφής ως $\mathbf{c}_{tx,t}$.

Λεξιλογική Πληροφορία και Context Clustering

Εν γένει, η προφορά ενός φωνήματος, τόσο σε οπτικό, όσο και σε ηχητικό επίπεδο εξαρτάται σε μεγάλο βαθμό από το περιβάλλον του φωνήματος. Κάθε μία από τις διαφορετικές προτάσεις που χρησιμοποιούνται για την εκπαίδευση του συστήματος πρέπει επομένως να συνοδεύεται από την αντίστοιχη λεξιλογική πληροφορία για το περιβάλλον του φωνήματος. Λαμβάνοντας υπ' όψιν το περιβάλλον των φωνημάτων, η λεξιλογική πληροφορία που παρέχεται στο σύστημα βρίσκεται στο Παράρτημα 2. Η συμπερίληψη της πληροφορίας σε επίπεδο κρυφών μαρκοβιανών μοντέλων γίνεται με την δημιουργία των context-dependent κρυφών μαρκοβιανών μοντέλων όπως φαίνεται στο Σχήμα 3.8.



Σχήμα 3.8: Context-Dependent Κρυφά Μαρκοβιανά Μοντελα [1].

Ένα μεγάλο πρόβλημα κατά την συμπερίληψη του περιβάλλοντος ενός φωνήματος είναι η έλλειψη δεδομένων στα οποία να συμπεριλαμβάνονται όλα τα διαφορετικά περιβάλλοντα του. Για παράδειγμα, αν θέλουμε να λάβουμε υπ' όψιν το αριστερά (ή το δεξιά) φώνημα ενός φωνήματος, για ένα λεξικό με 45 φωνήματα (+ σιωπή) υπάρχουν συνολικά 2,071 πιθανά μοντέλα. Αν θέλουμε να λάβουμε υπ' όψιν και τα δύο (αριστερά και δεξιά) στην οποία περίπτωση λέμε ότι μοντελοποιούμε ένα τριφώνημα, ο αριθμός των διαφορετικών μοντέλων που χρειάζονται είναι 95,221. Στο παρόν κείμενο χρησιμοποιούμε τα δύο αριστερά και δύο δεξιά φωνήματα ενός φωνήματος (quinphone context),

αλλά και πολλές άλλες παραμέτρους του περιβάλλοντος ενός φωνήματος. Βέβαια, οι ανωτέρω τιμές αποτελούν περισσότερο άνω φράγματα, καθώς πολλά από τα context δεν απαντώνται σχεδόν ποτέ. Παρ' όλα αυτά, η συλλογή μιας βάσης εκπαίδευσης που να περιέχει όλα τα contexts, είναι υπερβολικά επίπονη, και μάλιστα το μέγεθος της είναι απαγορευτικό.

Για την επίλυση του προβλήματος αυτού, επιτρέπουμε στα διαφορετικά μοντέλα που μοιράζονται συγκεκριμένα χαρακτηριστικά να μοιραστούν μεταξύ τους τις πληροφορίες, κάνοντας ομαδοποίηση των καταστάσεων των μοντέλων και "δένοντας τες".

Ο μοιρασμός των πληροφοριών μπορεί να γίνει είτε στα πλαίσια του μοντέλου, δηλαδή μοιρασμός πληροφοριών μεταξύ μοντέλων, είτε στα πλαίσια των καταστάσεων των μοντέλων, δηλαδή διαμοιρασμός πληροφοριών μεταξύ καταστάσεων. Εν γένει, ο διαμοιρασμός στα πλαίσια των καταστάσεων απολαμβάνει καλύτερης απόδοσης, καθώς επιτρέπει το διαμοιρασμό σωστών πληροφοριών (στα πλαίσια μοντέλου είναι δυνατόν να διαμοιραστούν και πληροφορίες τις οποίες δεν μοιράζονται τα δύο μοντέλα), επιτρέποντας τα αριστερά αλλά και δεξιά περιβάλλοντα να μοντελοποιηθούν ξεχωριστά.

Υπάρχουν δύο κύριες μέθοδοι για την ομαδοποίηση των καταστάσεων [60], η μέθοδος οδηγούμενη-από-δεδομένα (data driven clustering) και η μέθοδος δέντρου απόφασης (decision tree based context clustering). Στη συνέχεια περιγράφουμε τις δύο αυτές διαφορετικές μεθόδους.

Data Driven Clustering. Η διαδικασία που ακολουθείται για την ομαδοποίηση σε αυτή τη μέθοδο χωρίζεται στα ακόλουθα βήματα

1. Αρχικά, εκπαιδεύεται ένα σύνολο από κρυφά μαρκοβιανά μοντέλα με Γκαουσσιανές κατανομές εξόδου.
2. Στη συνέχεια, λαμβάνει χώρα μία επαναληπτική διαδικασία κατά την οποία συγχωνεύονται οι κατανομές των καταστάσεων που ομοιάζουν περισσότερο μεταξύ τους. Η συγχώνευση αυτή γίνεται με βάση την ελάχιστη απόσταση $d(i,j)$ μεταξύ των κατανομών:

$$d(i,j) = \{\frac{1}{n}\sum_{k=1}^{n}\frac{(\mu_{ij}-\mu_{jk})^2}{\sigma_{ik}\sigma jk}\}^{1/2} \tag{3.37}$$

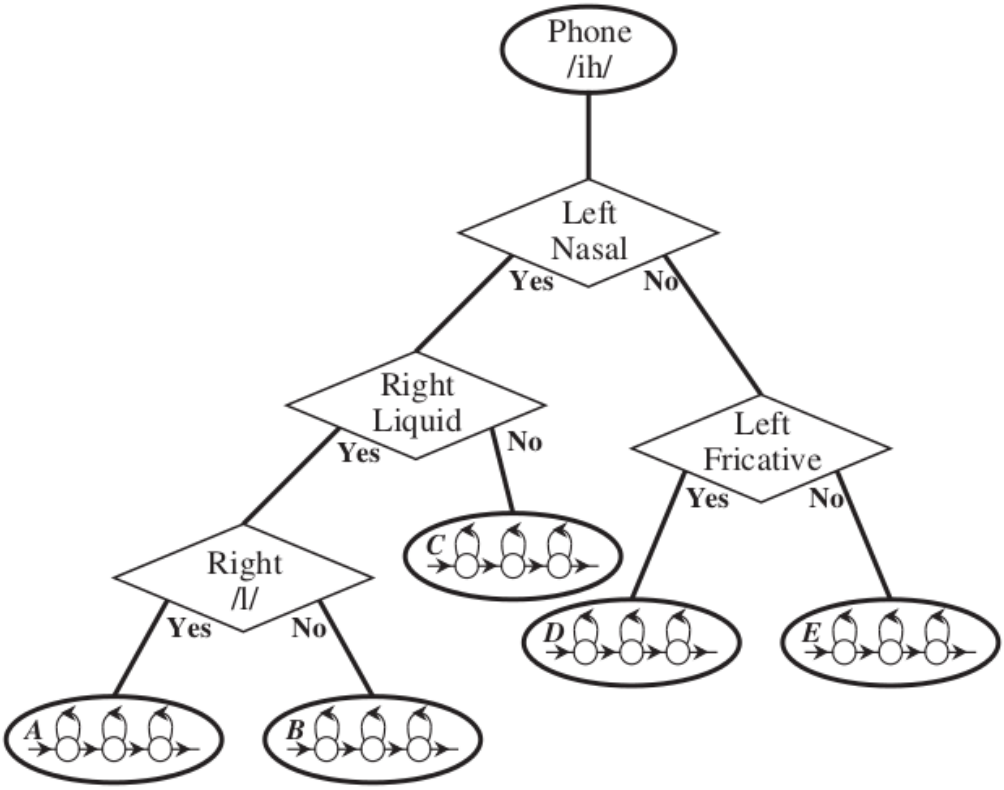
   όπου n είναι η διαστατικότητα των δεδομένων και $\mu_{sk}, \sigma sk$ είναι ο μέσος και η διακύμανση της k-οστής διάστασης της Γκαουσσιανής κατανομής της κατάστασης k.
3. Η διαδικασία αυτή συνεχίζεται έως ότι η ελάχιστη αυτή απόσταση να ξεπεράσει ένα κατώφλι.

Στη συνέχεια, το νέο σύστημα με τις δεμένες καταστάσεις επανεκπαιδεύεται με τον αλγόριθμο Baum-Welch.

Το μεγάλο μειονέκτημα της μεθόδου αυτής έγκειται στο γεγονός ότι απαιτούνται παραδείγματα από κάθε context, για την παραγωγή των αρχικών παραμέτρων του μοντέλου που χρησιμοποιούνται στην διαδικασία του clustering. Η επόμενη μέθοδος που περιγράφουμε, την οποία και χρησιμοποιούμε στο σύστημα μας, επιτρέπει την εύρωστη μοντελοποίηση και των context που δεν απαντώνται στη βάση εκπαίδευσης.

Decision Tree Based Context Clustering. Η μέθοδος αυτή, χρησιμοποιεί μαζί με τα δεδομένα εκπαίδευσης, και τις λεξιλογικές προδιαγραφές, για την συγχώνευση των καταστάσεων. Το κύριο συστατικό της μεθόδου αυτής, είναι ένα δέντρο απόφασης, δηλαδή ένα δυαδικό δέντρο σε κάθε κόμβο του οποίου υπαρχει μία ερώτηση δυαδικής απάντησης που αφορά το context των δεδομένων. Ένα τέτοιο δυαδικό δέντρο φαίνεται στο Σχήμα 3.9. Η διαδικασία που ακολουθείται σε αυτή τη μέθοδο χωρίζεται στα επόμενα βήματα:

1. Οι καταστάσεις στην ίδια θέση των μοντέλων τοποθετούνται στην ρίζα του δέντρου.
2. Σε κάθε κόμβο, επιλέγεται η ερώτηση που χωρίζει τα δεδομένα με βάση τη μεγιστοποίηση ή ελαχιστοποίηση μιας υποκειμενικής συνάρτησης.
3. Ο χωρισμός αυτός συνεχίζεται έως ότου να ικανοποιηθεί κάποιο κριτήριο.
4. Τέλος, συγχωνεύονται οι παράμετροι των δεδομένων σε κάθε φύλλο του δέντρου.



Σχήμα 3.9: Παράδειγμα δυαδικού δέντρου απόφασης για ένα φώνημα [60].

Εν γένει, υπάρχουν δύο κύριες διαφορετικές μέθοδοι που χρησιμοποιούνται για την επιλογή της ερώτησης σε κάθε κόμβο, οι προσεγγίσεις MDL, ML. Στη συνέχεια περιγράφουμε συντόμως και τις δύο.

Προσέγγιση MDL. Ορίζουμε το description length [74] $l_i(x^N)$ για δεδομένα $\{x^N = x_1, ..., x_N\}$ και μοντέλο i από ένα σύνολο μοντέλων $\{1, ..., i, ..., I\}$ ως:

$$l_i(x^N) = -\log P_{\hat{\theta}^{(i)}}(x^N) + \frac{\alpha_i}{2}\log N + \log I \tag{3.38}$$

όπου α_i είναι ο αριθμός των ελεύθερων παραμέτρων (διαστατικότητα) του μοντέλου i και $\hat{\theta}^{(i)}$ είναι οι εκτιμήσεις μεγίστης πιθανοφάνειας των παραμέτρων $\theta^{(i)} = (\theta_1^{(i)}, ..., \theta_{\alpha_i}^{(i)})$ του μοντέλου i.

Παρατηρούμε ότι ο πρώτος όρος είναι πανομοιότυπος με το αρνητικό της λογαριθμισμένης πιθανοφάνειας (log likelihood) των δεδομένων. Ο δεύτερος όρος είναι το μήκος κωδικοποίησης του μοντέλου i και ο τρίτος όρος είναι το κωδικό μήκος που απαιτείται για την επιλογή του μοντέλου i και υποθέτουμε ότι είναι σταθερός.

Τώρα θα δούμε το πως χρησιμοποιείται το ανωτέρο MDL κριτήριο κατά το tree based context clustering. Αρχικά μια κατάσταση S_0 ενός HMM ενός φωνήματος χωρίζεται σε M κόμβους $S_1, ..., S_M$ όπως είδαμε παραπάνω. Στη συνέχεια, υπολογίζεται η λογαριθμισμένη πιθανοφάνεια L του κόμβου S_m που παράγει ένα σύνολο πλαισίων εκπαίδευσης προσεγγιστικά ως εξής:

$$L(S_m) \approx \sum_{t=1}^{T} \log(N((o)_t, \mu_{S_m}, \sum S_m))\gamma_t(S_m) = -\frac{1}{2}(log((2\pi)^K|\sum S_m|) + K), (S_m) \quad (3.39)$$

όπου

$$\gamma_t(S_m) = \frac{\alpha_t(S_m)\beta_t(S_m)}{\sum_s \alpha_t(S_m)\beta_t(S_m)} \quad (3.40)$$

και

$$(S_m) = \sum_{t=1}^{T} \gamma_t(S_m) \quad (3.41)$$

όπου K είναι η διαστατικότητα του διανύσματος δεδομένων $\boldsymbol{o}_t$, $\gamma_t(S_m)$ είναι η a posteriori πιθανότητα ότι το παρατηρούμενο πλαισίο $\boldsymbol{o}_t$ παράγεται από την κατάσταση S_m, και, (S_m) είναι ένας μετρητής της κατάληψης των συνολικών καταστάσεων, που είναι το άθροισμα του $\gamma_t(S_m)$ για όλα τα πλαίσια δεδομένων. Η forward πιθανότητα $\alpha_t(S_m)$, η backward πιθανότητα $\beta_t(S_m)$, το διάνυσμα μέσης τιμής μ_{S_m}, και η μήτρα συνδιακύμανσης $\sum S(m)$ υπολογίζονται από τα δεδομένα εκπαίδευσης. Τότε, το description length l(U) δίνεται ως

$$l(U) \approx -\frac{1}{2}\sum_{m=1}^{M} L(S_m) + KM \log \sum_{m=1}^{M} (S_m) = \frac{1}{2}\sum_{m=1}^{M} (S_m) \log(|\Sigma(S_m)| + KM \log V \quad (3.42)$$

$$V = \sum_{m=1}^{M} S(_m) = (S_0) \quad (3.43)$$

Τώρα, αν ορίσουμε ως $\delta_q(S)$ τη διαφορά μεταξύ του description length l πριν και μετά το διαχωρισμό όταν ο κόμβο S χωρίζεται σε δύο χρησιμοποιώντας την ερώτηση q, τότε έχουμε:

$$\Delta_q(S) = \frac{1}{2}(S_{qy} \log|\Sigma_{S_{qy}}| + (S_{qn}) \log|\Sigma_{S_{qn}}| - (S) \log|\Sigma_S|) + K \log V \quad (3.44)$$

όπου S_qy και S_qn είναι οι κόμβοι που προκύπτουν μετά το διαχωρισμό. Έτσι, ξεκινώντας από έναν αρχικό κόμβο S_0, επιλέγουμε την ερώτηση που θα ελαχιστοποίησει το $\Delta_q(S)$. Αν $\Delta_q(S) > 0$, τότε σταματάμε το διαχωρισμό, αλλιώς αν $\Delta_q(S) < 0$ διαχωρίζουμε τον αρχικό κόμβο σε δύο νέους, για τους οποίους επαναλαμβάνεται εκ νέου η διαδικασία, έως ότου να μην έχουμε άλλους κόμβους.

Maximum Likelihood Approach. Στην περίπτωση του κριτηρίου μεγίστης πιθανοφάνειας, συνεχίζοντας από τις Εξισώσεις 3.39 έως 3.41, επιλέγεται η ερώτηση q που μεγιστοποίει την αύξηση της μεγίστης πιθανοφάνειας μετά τον διαχωρισμό του κόμβου:

$$\delta_q S = L(S_{qy}) + L(S_{qn}) - L(S) \tag{3.45}$$

Ένα μεγάλο μειονέκτημα που παρουσιάζει η προσέγγιση μεγίστης πιθανοφάνειας είναι ότι στις περισσότερες περιπτώσεις, η πιθανοφάνεια αυξάνεται καθώς αυξάνει και ο αριθμός των μονάδων, δηλαδή η αύξηση στην πιθανοφάνεια $\delta_q S$ είναι σχεδόν πάντα θετική. Επομένως, στο τελευταίο κομμάτι του διαχωρισμού, το σύνολο των μοντέλων γίνεται σχεδόν πανομοιότυπο με το σύνολο των συνεχών μοντέλων χωρίς clustering. Έτσι, η προσέγγιση ML απαιτεί μια εξωτερική παράμετρο για τον έλεγχο του μεγέθους του clustering, ο οποίος συνήθως γίνεται χρησιμοποιώντας ένα κατώφλι για την αύξηση της πιθανοφάνειας ή την αύξηση των μονάδων. Το κατώφλι αυτό βελτιστοποιείται μέσω ενός συνόλου δεδομένων εκπαίδεσης, ή μέσω μιας μεθόδου cross-validation. Όπως καταλαβαίνουμε οι διαδικασίες αυτές της βελτιστοποίησης είναι υπολογιστικά ακριβείς, απαιτούν περισσότερα δεδομένα, και δεν έχουν ισχυρή θεωρητική δικαιολόγιση.

Αντιθέτως, το κριτήριο MDL δεν απαιτεί την ύπαρξη εξωτερικού κριτηρίου, και ο όρος $\log V$ στην 3.44 υπολογίζεται αυτόματα από τα δεδομένα εκπαίδευσης.

3.6.2 Τμήμα Σύνθεσης

Παραγωγή Παραμέτρων Εξόδου

Με βάση την ανάλυση που κάναμε για τα κρυφα μαρκοβιανά μοντέλα προηγουμένως, η παραγωγή των παραμέτρων εξόδου από ένα εκπαιδευμένο κρυφό μαρκοβιανό μοντέλο είναι άμεση. Θέλουμε να βρούμε ένα διάνυσμα παρατήρησης εξόδου:

$$\boldsymbol{O} = \left[\boldsymbol{o}_1^\intercal, \boldsymbol{o}_2^\intercal, ..., \boldsymbol{o}_T^\intercal\right] \tag{3.46}$$

έτσι ώστε να μεγιστοποιείται η ακόλουθη πιθανότητα:

$$P\left(\boldsymbol{O}|\lambda\right) = \sum_{\text{all } \boldsymbol{Q}} P\left(\boldsymbol{O}, \boldsymbol{Q}|\lambda\right) \tag{3.47}$$

ως προς το $\boldsymbol{O}$, όπου

$$\boldsymbol{Q} = \left\{\left(q_1, i_1\right), \left(q_2, i_2\right), ..., \left(q_T, i_T\right)\right\} \tag{3.48}$$

είναι η ακολουθία καταστάσεων και μειγμάτων, δηλαδή, το (q, i) δηλώνει το i-οστό μείγμα της κατάστασης q.

Όπως αναφέραμε προηγουμένως, υποθέτουμε ότι το διάνυσμα παραμέτρων αποτελείται από στατικές και συνεχείς παραμέτρους, και είναι της μορφής $\boldsymbol{o}_t = [\boldsymbol{c}_t, \boldsymbol{\Delta c}_t]$.

Επίλυση του προβλήματος μεγιστοποίησης

Ο λογάριθμος της $P\left(\boldsymbol{O}|\boldsymbol{Q},\lambda\right)$ είναι:

$$\log P\left(\boldsymbol{O}|\boldsymbol{Q},\lambda\right)=-\frac{1}{2}\boldsymbol{O}^{\intercal}\boldsymbol{U}^{-1}\boldsymbol{O}+\boldsymbol{O}^{\intercal}\boldsymbol{U}^{-1}\boldsymbol{M}+K \tag{3.49}$$

όπου

$$\boldsymbol{U}^{-1}=diag\left[\boldsymbol{U}_{q_1,i_1}^{-1},\boldsymbol{U}_{q_2,i_2}^{-1},...,\boldsymbol{U}_{q_T,i_T}^{-1}\right] \tag{3.50}$$

και

$$\boldsymbol{M}=\left[\boldsymbol{\mu}_{q_1,i_1}^{\intercal},\boldsymbol{\mu}_{q_2,i_2}^{\intercal},...,\boldsymbol{\mu}_{q_T,i_T}^{\intercal}\right] \tag{3.51}$$

με $\boldsymbol{\mu}_{q_t,i_t}$ το $3M\times 1$ διάνυσμα μέσης τιμής και $\boldsymbol{U}_{q_t,i_t}$ την $3M\times 3M$ μήτρα συνδιακύμανσης που σχετίζονται με το i_t-οστό μείγμα της κατάστασης q_t, και η σταθερά K είναι ανεξάρτητη του O.

Από την εξίσωση 3.49 μπορούμε να δούμε ότι όταν $\boldsymbol{O}=\boldsymbol{M}$, δηλαδή το διάνυσμα παρατήρησης γίνεται ένα διάνυσμα μέσων τιμών, μεγιστοποιείται η $P\left(\boldsymbol{O}|\boldsymbol{Q},\lambda\right)$. Παρ' όλα αυτά, όπως αναφέραμε το διάνυσμα παρατήρησης δεν περιλαμβάνει μόνο στατικές αλλά και δυναμικές παραμέτρους. Αρχικά, θα αναδιατάξουμε τις εξισώσεις 3.34, 3.35 σε μορφή μητρών:

$$\boldsymbol{O}=\boldsymbol{W}\boldsymbol{C} \tag{3.52}$$

με

$$\boldsymbol{C}=[\boldsymbol{c}_1,\boldsymbol{c}_2,...,\boldsymbol{c}_T]^{\intercal} \tag{3.53}$$

$$\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,...,\boldsymbol{w}_T]^{\intercal} \tag{3.54}$$

$$\boldsymbol{w}_t=\left[\boldsymbol{w}_t^{(0)},\boldsymbol{w}_t^{(1)},\boldsymbol{w}_t^{(2)}\right] \tag{3.55}$$

$$\begin{aligned}\boldsymbol{w}_t^{(n)}=[&\mathbf{0}_{M\times M},...,\mathbf{0}_{M\times M},w^{(n)}\left(-L_-^{(n)}\right)\boldsymbol{I}_{M\times M},...,\\&w^{(n)}\left(0\right)\boldsymbol{I}_{M\times M},...,w^{(n)}\left(L_+^{(n)}\right)\boldsymbol{I}_{M\times M},...,\\&\mathbf{0}_{MxM},...,\mathbf{0}_{M\times M}]^{\intercal},\qquad n=0,1,2\end{aligned} \tag{3.56}$$

Επομένως, με βάση τη συνθήκη 3.52 η μεγιστοποίηση της πιθανότητας εξόδου ως προς το $\boldsymbol{O}$ είναι ισοδύναμη με την μεγιστοποίηση της πιθανότητας εξόδου ως προς το $\boldsymbol{C}$. Αντικαθιστώντας επομένως και θέτωντας την παράγωγο του $\log P\left(\boldsymbol{W}\boldsymbol{O}|\boldsymbol{Q},\lambda\right)$ ίση με $\mathbf{0}$ παίρνουμε:

$$\frac{d\log P\left(\boldsymbol{W}\boldsymbol{O}|\boldsymbol{Q},\lambda\right)}{d\boldsymbol{C}}=\boldsymbol{O} \tag{3.57}$$

Η λύση της ανωτέρω εξίσωσης μας οδηγεί στο σύνολο εξισώσεων:

$$\boldsymbol{W}^{\intercal}\boldsymbol{U}^{-1}\boldsymbol{W}\boldsymbol{C} = \boldsymbol{W}^{\intercal}\boldsymbol{U}^{-1}\boldsymbol{M}^{\intercal} \tag{3.58}$$

Η άμεση λύση της 3.58 λόγω του γεγονότος ότι η μήτρα $\boldsymbol{W}^{\intercal}\boldsymbol{U}^{-1}\boldsymbol{W}$ είναι $TM \times TM$ διαστάσεων έχει πολυπλοκότητα $O\left(T^3M^3\right)$. Για να μειωθεί η υπολογιστική απαιτητικότητα καταφεύγουμε σε αποσύμπλεξη $Cholesky$ ή QR λαμβάνοντας υπ' όψιν την ειδική δομή της μήτρας. Ως αποτέλεσμα η πολυπλοκότητα μειώνεται σε $\left(TM^3L^2\right)$) με

$$L = \max_{n\in 1,2, s\in\{-,+\}} L_s^{(n)} \tag{3.59}$$

Global Variance. Ένα πρόβλημα που εμφανίζεται στην σύνθεση φωνής με κρυφά μαρκοβιανά μοντέλα είναι το γεγονός ότι η τροχιά των παραγόμενων παραμέτρων έχει υποστεί μεγάλη ομαλοποίηση (smoothing), λόγω της στατιστικής επεξεργασίας. Η λεία αυτή τροχιά των παραγόμενων παραμέτρων προκαλεί πνιχτή συνθετική φωνή, υποβαθμισμένης φυσικότητας. Για την αντιμετώπιση αυτού του φαινομένου υπάρχουν δύο εναλλακτικές πρακτικές.

Η μία τακτική είναι η αύξηση του αριθμού των μειγμάτων που χρησιμοποιούνται στις Γκαουσσιανές κατανομές γεγονός όμως που οδηγεί σε προβλήματα υπερ-εκπαίδευσης.

Η δεύτερη τακτική [78, 77], είναι η χρήση της συνολικής διακύμανσης (global variance) της τροχιάς.

Η global variance (gv) των διανυσμάτων των στατικών παραμέτρων ορίζεται ως:

$$\boldsymbol{\upsilon}\left(\boldsymbol{C}\right) = \left[\upsilon\left(1\right), \upsilon\left(2\right), ..., \upsilon\left(D\right)\right]^{\intercal} \tag{3.60}$$

όπου

$$\upsilon d = \frac{1}{T}\sum_{t=1}^{T}\left(c_t\left(d\right) - \bar{c}\left(d\right)\right)^2 \tag{3.61}$$

και

$$\bar{c}\left(d\right) = \frac{1}{T}\sum_{\tau=1}^{T} c_\tau\left(d\right) \tag{3.62}$$

Η μέθοδος αυτή υπολογίζει την ακολουθία των στατικών παραμέτρων που παράγονται όχι μόνο με χρήση της πιθανότητας εξόδου των στατικών και δυναμικών παραμέτρων αλλά και με χρήση της GV. Πιο συγκεκριμένα, αντί του κριτηρίου 3.47, μεγιστοποιείται το ακόλουθο κριτήριο:

$$L = \log\left\{p\left(\boldsymbol{O}|\boldsymbol{Q},\boldsymbol{\lambda}\right)^{\omega} p\left(\boldsymbol{\upsilon}\left(\boldsymbol{C}\right)|\boldsymbol{\lambda}_\upsilon\right)\right\} \tag{3.63}$$

Η πιθανότητα $p\left(\boldsymbol{\upsilon}\left(\boldsymbol{C}\right)|\boldsymbol{\lambda}_\upsilon\right)$ μοντελοποιείται με Γκαουσσιανή κατανομή και ουσιαστικά αποτελεί ένας όρος ποινής για την μείωση της GV της τροχιάς.. Οι παράμετροι των μοντέλων $\boldsymbol{\lambda}_\upsilon$ αποτελούνται από το μέσο διάνυσμα $\boldsymbol{\mu}_\upsilon$ και την μήτρα συνδιακύμανσης $\Sigma_\upsilon = \boldsymbol{P}_\upsilon^{-1}$. Η εύρεση του $\boldsymbol{C}$ με βάση το κριτήριο (3.63) χρησιμοποιούνται επαναληπτικές μέθοδοι όπως η Newton-Raphson είτε η Steepest Gradient [78].

3.6.3 Διαδικασία

Τώρα που γνωρίζουμε τη μορφή του διανύσματος εκπαίδευσης και της λεξιλογικής πληροφορίας, καθώς και τη μορφή των κρυφών μαρκοβιανών μοντέλων που χρησιμοποιούμε, μπορούμε να περιγράψουμε γενικώς την διαδικασία που ακολουθείται για την εκπαίδευση του οπτικοακουστικού συστήματος σύνθεσης φωνής με κρυφά μαρκοβιανά μοντέλα:

1) Αρχικά, υπολογίζεται από όλα τα διανύσματα παρατήρησης o_t της βάσης δεδομένων, το διάνυσμα μέσης τιμής και διακύμανσης, το οποίο χρησιμοποιείται για την αρχικοποίηση 29 κρυφών μαρκοβιανών μοντέλων - χωρίς να λαμβάνουμε υπ' όψιν το περιβάλλον των φωνημάτων - ένα για κάθε ένα από τα φωνήματα της ελληνικής γλώσσας (μαζί με την παύση).

2) Τα κρυφά μαρκοβιανά μοντέλα που δημιουργήθηκαν στο πρώτο βήμα εκπαιδεύονται με χρήση του αλγορίθμου Baum-Welch.

3) Στη συνέχεια, μετά από την πρώτη διαδικασία εκπαίδευσης, σχηματίζονται κρυφά μαρκοβιανά μοντέλα για κάθε ένα από τα διαφορετικά περιβάλλοντα για κάθε φώνημα που συναντώνται στη βάση δεδομένων.

4) Τα context-dependent κρυφά μαρκοβιανά μοντέλα που σχηματίστηκαν στο προηγούμενο βήμα εκπαιδεύονται εκ' νέου με τα διανύσματα παρατήρησης που αντιστοιχούν στο περιβάλλον τους.

5) Στη συνέχεια προχωράμε σε tree-based context clustering όπου γίνεται ομαδοποίηση των καταστάσεων των μοντέλων που μοιράζονται συγκεκριμένα χαρακτηριστικά μεταξύ τους που επιλέγονται μέσα από ένα δέντρο ερωτήσεων.

6) Γίνεται πάλι εκπαίδευση των ομαδοποιημένων πια κρυφών μαρκοβιανών μοντέλων.

7) Τα ομαδοποιημένα κρυφά μαρκοβιανά μοντέλα αποσυνδέονται και επαναλαμβάνονται για μία ακόμη φορά οι διαδικασίες των Βημάτων 5 και 6.

8) Γίνεται υπολογισμός της Global Variance.

9) Τώρα που ολοκληρώθηκε η φάση της εκπαίδευσης του συστήματος, σχηματίζονται από τα κρυφά μαρκοβιανά μοντέλα οι προτάσεις προς σύνθεση, και παράγονται τα διανύσματα παρατήρησης τους.

10) Γίνεται σύνθεση της ομιλίας και της εικονοσειράς από τις παραμέτρους που παρήγαν τα κρυφά μαρκοβιανά μοντέλα.

Κεφάλαιο 4

Πρακτική Υλοποίηση και Εφαρμογή του Συστήματος

Στο Κεφάλαιο αυτό, έχοντας πια κάνει την απαραίτητη θεωρητική ανάλυση του οπτικοακουστικού συστήματος σύνθεσης φωνής στο προηγούμενο Κεφάλαιο, θα προχωρήσουμε σε πρακτική υλοποίηση του συστήματος οπτικοακουστικής συνθετικής φωνής. Στη συνέχεια, θα προχωρήσουμε σε αξιολόγηση της ποιότητας της οπτικακουστικής σύνθεσης φωνής από το σύστημα μας, και το πως αυτή επηρεάζεται από διάφορες παραμέτρους και τεχνικές που αναλύσαμε στα προηγούμενα κεφάλαια.

Σαν κύριο εργαλείο για την πρακτική υλοποίηση του συστήματος οπτικοακουστικής σύνθεσης φωνής με χρήση κρυφών μαρκοβιανών μοντέλων, χρησιμοποιούμε το εργαλείο HTS [62] για σύνθεση φωνής που αποτελεί επέκταση του εργαλείο διαχείρισης κρυφών μαρκοβιανών μοντέλων HTK [4]. Με βάση αυτά τα εργαλεία, προχωρούμε σε τροποποιήσεις ώστε να επεκτείνουμε το σύστημα σύνθεσης φωνής σε σύστημα οπτικοακουστικής σύνθεσης φωνής.

Για την αξιολόγηση της ποιότητας της συνθετικής φωνής και εικονοσειράς, προχωρήσαμε σε δημιουργία tests MOS, τα οποία περιγράφουμε αναλυτικά στην Ενότητα 4.3.1. Μελετήσαμε την επίδραση παραμέτρων που αναφέραμε στα προηγούμενα Κεφάλαια, τόσο σε επίπεδο φωνής όσο και σε επίπεδο εικονοσειράς, δείχνοντας ότι ενώ κάποιες τεχνικές (συμπερίληψη δυναμικών χαρακτηριστικών) βελτιώνουν την απόδοση του συστήματος τόσο σε βαθμό συνθετικής φωνής, όσο και σε βαθμό εικονοσειράς, κάποιες άλλες τεχνικές (global variance) έχουν διαφορετική επίπτωση στη συνθετική φωνή (βελτιώση) και διαφορετική επίπτωση στη συνθετική εικονοσειρά (υποβάθμιση).

4.1 Εξαγωγή Ακουστικών & Οπτικών Παραμέτρων και Λεξιλογικού Περιεχομένου

Η εξαγωγή των ακουστικών παραμέτρων για το διάνυσμα παρατήρησης έγινε αρχικά με τη χρήση του εργαλείου STRAIGHT [45] για το MATLAB που αναφέραμε στο Κεφάλαιο 2 και στη συνέχεια με χρήση του πακέτου SPTK [64]. Το α που αντιστοιχεί στη στρέβλωση της συχνότητας επιλέχθηκε με βάση των Πίνακα 2.1 ως 0.42 για να προσεγγίσουμε όσο το δυνατόν καλύτερα την ακουστική κλίμακα Mel, ενώ η τάξη των mgc συντελεστών ήταν 30.

Η εξαγωγή των οπτικών παραμέτρων όπως έχουμε ήδη αναφέρει επίσης στο Κεφάλαιο 2 έγινε με χρήση του aam-tools toolbox [3] για το MATLAB μετά από σημαδοποίηση και εκπαίδευση ενός active appearance μοντέλου με 268 εικόνες με χρήση του εργαλείου [7] και την εφαρμογή του στις εικόνες της βάσης. Πιο συγκεκριμένα, μετά την εκπαίδευση ενός active appearance model από τις 268 εικόνες που σημαδοποιήθηκαν (το οποίο απαρτίζεται από τα 10 πρώτα ιδιοδιανύσματα σχήματος που επεξηγούν το 98% της μεταβολής του σχήματος του προσώπου, και τα 32 πρώτα ιδιοδιανύσματα υφής που επεξηγούν το 95% της μεταβολής της υφής), με χρήση του αλγορίθμου VoTuIc [67], έγινε εφαρμογή του μοντέλου στις περίπου 120,000 εικόνες που απαρτίζουν τη βάση CVSP-AV και αντιστοιχούν στο ουδέτερο συναίσθημα. Στη συνέχεια, οι παράμετροι που εξήχθησαν από την εφαρμογή αυτή α-

ποθηκεύτηκαν σε κατάλληλη μορφή, ώστε να ενσωματωθούν στο διάνυσμα παρατήρησης μαζί με τις ακουστικές παραμέτρους.

Για την εξαγωγή των λεξιλογικών περιεχομένων αναπτύχθηκε με χρήση των εργαλείων Festival [61] και Sail-align [70], ένα πλήρως λειτουργικό front-end λεξιλογικής επεξεργασίας που μετατρέπει το κείμενο εισόδου σε αρχεία λεξιλογικής πληροφορίας κατάλληλα για το σύστημα μας.

4.2 Εκπαίδευση και Σύνθεση

Μετά την εξαγωγή των παραμέτρων προχωρήσαμε σε σύνθεση των οπτικοακουστικών προτάσεων με την ακόλουθη διαδικασία:

1) Υπολογισμός από όλα τα διανύσματα παρατήρησης o_t της βάσης δεδομένων, το διάνυσμα μέσης τιμής και διακύμανσης με χρήση της συνάρτησης HCompV και αρχικοποίηση των μοντέλων με χρήση της συνάρτησης HInit.

2) Εκπαίδευση των hmm που δημιουργήθηκαν στο πρώτο βήμα εκπαιδεύονται με χρήση του αλγορίθμου Baum-Welch με χρήση της συνάρτησης HERest.

3) Σχηματίσμός των hmm για κάθε ένα από τα διαφορετικά περιβάλλοντα για κάθε φώνημα (context-dependent hmms) που συναντώνται στη βάση δεδομένων με χρήση της συνάρτησης HHed χειρισμού κρυφών μαρκοβιανών μοντέλων.

4) Εκ νέου εκπαίδευση των context-dependent hmm με χρήση της συνάρτησης HERest.

5) Εφαρμογή tree-based-context-clustering με χρήση της συνάρτησης HHed.

6) Γίνεται πάλι εκπαίδευση των ομαδοποιημένων πια κρυφών μαρκοβιανών μοντέλων (HERest).

7) Τα ομαδοποιημένα κρυφά μαρκοβιανά μοντέλα αποσυνδέονται (HHed) και επαναλαμβάνονται για μία ακόμη φορά οι διαδικασίες των Βημάτων 5 και 6.

8) Υπολογισμός της Global Variance με χρήση των HERest και HHed.

9) Τώρα που ολοκληρώθηκε η φάση της εκπαίδευσης του συστήματος, σχηματίζονται από τα κρυφά μαρκοβιανά μοντέλα οι προτάσεις προς σύνθεση (HHed), και παράγονται τα διανύσματα παρατήρησης τους με χρήση της συνάρτησης HMGens.

10) Γίνεται σύνθεση της ομιλίας με χρήση του vocoder που περιέχεται στο εργαλείο STRAIGHT και του βίντεο με χρήση του MATLAB και σύνδεση των δύο με χρήση του ffmpeg.

4.3 Αξιολόγηση του Συστήματος Οπτικοακουστικής Σύνθεσης Φωνής

Η αξιολόγηση ενός συστήματος σύνθεσης φωνής και εικονοσειράς αποτελεί εξαιρετικά σημαντική διαδικασία, για την μέτρηση της προόδου των συστημάτων, αλλά και για την επιλογή του κατάλληλου

συστήματος για διαφορετικές εφαρμογές. Έχουμε ήδη αναφέρει στο Κεφάλαιο 1, τα σημαντικότερα χαρακτηριστικά που πρέπει να έχει ένα σύστημα σύνθεσης φωνής - φυσικότητα, κατανόηση του λόγου, και επιτυχή επεξεργασία οποιουδήποτε κειμένου εισόδου. Κατά την αξιολόγηση του συστήματος μας, επεκτείναμε τα χαρακτηριστικά αυτά και στον τομέα της εικονοσειράς, αξιολογώντας επίσης τη φυσικότητα του προσώπου, αλλά και την κατανόηση της κίνησης του στόματος.

4.3.1 Test MOS (Mean Opinion Score)

Μία από τις πιο ευρέως γνωστές μεθόδους που χρησιμοποιούνται κατά την αξιολόγηση ενός συστήματος σύνθεσης είναι το test MOS (Mean Opinion Score), που πρωτοεισήχθηκε από την Διεθνή Ένωση Τηλεπικοινωνιών (International Telecommunication Union), για την αξιολόγηση της ποιότητας των τηλεφωνικών δικτύων [44]. Το test αυτό, αποτελεί ένα ερωτηματολόγιο, στο οποίο οι ερωτηθέντες καλούνται να αξιολογήσουν κάποιο χαρακτηριστικό των συνθετικών προτάσεων σε μία κλίμακα ακεραίων από 1 έως 5 με το 1 να αντιστοιχεί σε *πολύ κακό* και το 5 σε *πολύ καλό* αποτέλεσμα. Η τελική τιμή λαμβάνεται υπολογίζοντας τον αριθμητικό μέσο όλων των διαφορετικών απαντήσεων.

Ανάλογα με τα χαρακτηριστικά που θέλουμε να εξετάσουμε σε ένα σύστημα σύνθεσης φωνής, δημιουργείται και ένα σχετικό test MOS, που περιλαμβάνει ένα σύνολο προτάσεων τις οποίες καλούνται να αξιολογήσουν ως προς τα επιθυμητά χαρακτηριστικά οι ερωτηθέντες.

Το μεγαλύτερο πλεονέκτημα ενός test MOS είναι η αποδοτικότητα του καθώς παρέχει γρήγορη εκτίμηση στην κατανόηση και τη φυσικότητα ενός συστήματος σύνθεσης φωνής [68]. Από το 2005 και μετά τα test MOS χρησιμοποιούνται επίσης στον διεθνή διαγωνισμό Blizzard Challenge που αξιολογεί τα συστήματα συνθετικής φωνής χρησιμοποιώντας κοινές βάσεις δεδομένων [20].

Λαμβάνοντας τα ανωτέρω υπ' όψιν, προχωρήσαμε σε δημιουργία ειδικού ερωτηματολογίου για να αξιολογήσουμε την κατανόηση και την φυσικότητα της συνθετικής φωνής του συστήματος, χαρακτηριστικά εξαιρετικά σημαντικά για ένα σύστημα σύνθεσης φωνής όπως αναφέραμε και στο Κεφάλαιο 1. Το ερωτηματολόγιο απαρτιζόταν από 20 προτάσεις συνθετικής ομιλίας, τις οποίες οι 13 ερωτηθέντες κλήθηκαν να αξιολογήσουν ως προς την φυσικότητα και την κατανόηση τους σε μία κλίμακα από 1 έως 5.

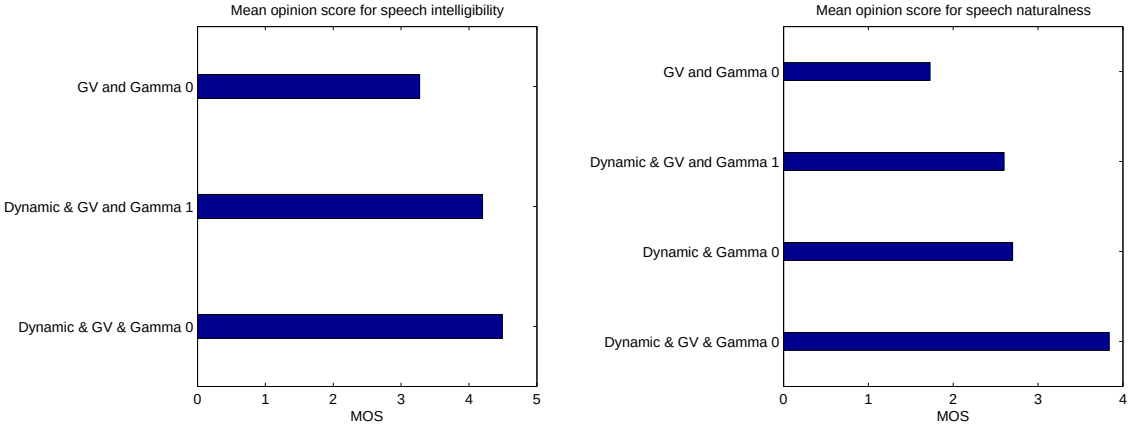
Σκοπός του ερωτηματολογίου ήταν εκτός από την συνολική αξιολόγηση του συστήματος σύνθεσης φωνής, η μελέτη της επίδρασης μερικών παραμέτρων στην φυσικότητα και την κατανόηση της συνθετικής φωνής όπως θα δούμε και εν συνεχεία.

Επιπλέον, επεκτείνοντας την αξιολόγηση στο επίπεδο της οπτικοακουστικής σύνθεσης φωνής, εκτός από τις 20 προτάσεις συνθετικής ομιλίας, οι ερωτηθέντες κλήθηκαν να αξιολογήσουν επίσης την κατανόηση της κίνησης του στόματος της εικονοσειράς, καθώς και την φυσικότητα του προσώπου σε ένα σύνολο από 15 προτάσεις συνθετικής ομιλίας και εικονοσειράς.

Στις επόμενες υποενοτητες θα μελετήσουμε την επίδραση διαφόρων παραμέτρων στην φυσικότητα και την κατανόηση της συνθετικής φωνής και εικονοσειράς, όπως προκύπτει από τα αποτελέσματα του test MOS.

4.3.2 Αξιολόγηση Συνθετικής Φωνής

Κατά την αξιολόγηση της συνθετικής ομιλίας, οι ερωτηθέντες αξιολόγησαν την επίδραση της παραμέτρου γ των mel generalized cepstral συντελεστών που επηρεάζει και την μορφή του φίλτρου σύνθεσης $H(z)$ όπως αναφέραμε στο Κεφάλαιο 2, των δυναμικών χαρακτηριστικών, και της global variance στη φυσικότητα του λόγου και των δυναμικών χαρακτηριστικών και της παραμέτρου γ των mel generalized cepstral συντελεστών στην κατανόηση του λόγου. Τα αποτελέσματα από αυτήν την αξιολόγηση τους φαίνονται στο Σχήμα 4.1 και στον Πίνακα 4.1.



(i) MOS test Κατανόησης της ομιλίας. **(ii)** MOS test Φυσικότητας της ομιλίας.

Σχήμα 4.1: Αξιολόγηση του ήχου (i) Κατανόηση (ii) Φυσικότητα.

MOS	Dynamic Features, GV, $\gamma = 0$	$\gamma = 1$	No GV	No Dynamic Features
Κατανόηση	4.49	4.2	-	3.27
Φυσικότητα	3.84	2.7	2.6	1.7273

Πίνακας 4.1: Αξιολόγηση της ποιότητας του ήχου για διάφορες παραμέτρους.

Όπως είναι εμφανές και από τα διαγράμματα, την μεγαλύτερη επίδραση στην κατανόηση και στη φυσικότητα της ομιλίας έχουν τα δυναμικά χαρακτηριστικά, καθώς βελτιώνουν κατά 1.22 μονάδες την κατανόηση της συνθετικής φωνής και κατά 2.22 μονάδες την φυσικότητα της συνθετικής φωνής, καθώς εξαλείφουν τις ασυνέχειες που εμφανίζονται στη συνθετική φωνή.

Σημαντική είναι επίσης η επίδραση της global variance στη φυσικότητα της φωνής, καθώς η εισαγωγή της στο σύστημα μας βελτιώνει κατά 1.2 μονάδες τη φυσικότητα της φωνής, καθώς αποτρέπει την ομαλοποίηση της τροχιάς των παραγόμενων παραμέτρων, λόγω της στατιστικής επεξεργασίας (Υποενότητα 3.6.2).

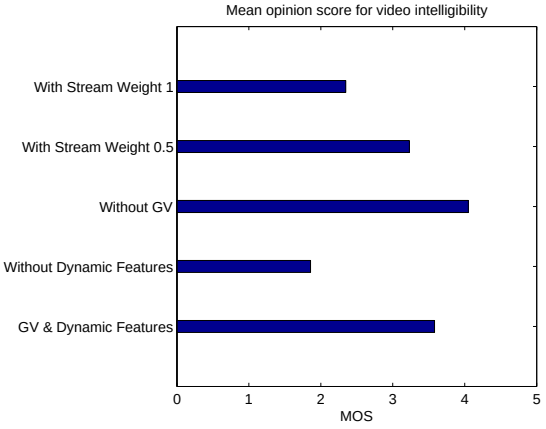
Τέλος, ενώ για $\gamma = 1$ των mel generalized cepstral συντελεστών σε σχέση με $\gamma = 0$ η κατανόηση του ήχου δεν αλλάζει σημαντικά (0.29 μονάδες), η φυσικότητα αλλάζει σημαντικά κατά 1.14 μονάδες.

4.3.3 Αξιολόγηση Συνθετικής Εικόνας

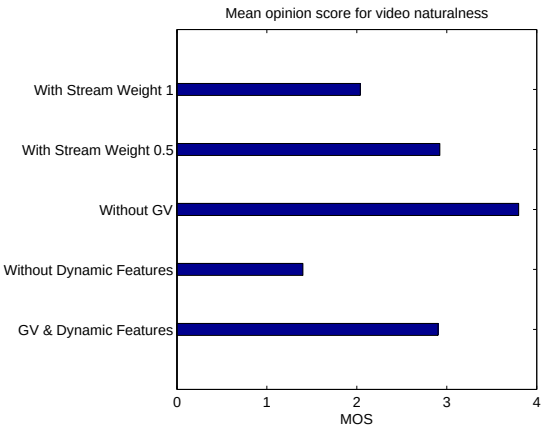
Στη συνέχεια, προχωρήσαμε σε αξιολόγηση της συνθετικής εικονοσειράς. Αυτή τη φορά, μελετήσαμε την επίδραση της global variance, των δυναμικών χαρακτηριστικών, καθώς και του βάρους του διανύσματος (stream weight) των οπτικών παραμέτρων κατά την εκπαίδευση του συστήματος που χρησιμοποιείται για την ευθυγράμμιση των κρυφών μαρκοβιανών μοντέλων σε επίπεδο κατάστασης με χρήση του αλγορίθμου Viterbi. Περισσότερες πληροφορίες μπορούν να βρεθούν στο [42]. Όταν η τιμή του βάρους του διανύσματος είναι ίση με το 0, το διάνυσμα στο οποίο αντιστοιχεί το βάρος, δεν λαμβάνει μέρος στην ευθυγράμμιση αυτή, ενώ όταν η τιμή του βάρους του διανύσματος είναι ίση με 1 λαμβάνει μέρος.

Από την αποτελέσματα και την επεξεργασία τους που πήραμε από το test MOS, προκύπτει το Σχήμα 4.2 και ο Πίνακας 4.2

Πάλι παρατηρούμε ότι τη σημαντικότερη επίδραση έχουν οι δυναμικές παράμετροι, η εισαγωγή των οποίων βελτιώνει κατά 2.2 μονάδες την κατανόηση της εικονοσειράς (κίνηση στόματος) και κατά 2.4 μονάδες την φυσικότητα του, λόγω της εξάλειψης των ασυνεχειών στην εικονοσειρά.



(i) MOS test κατανόησης της εικονοσειράς (κίνηση στόματος).



(ii) MOS test φυσικότητας της εικονοσειράς.

Σχήμα 4.2: Αξιολόγηση της συνθετικής εικονοσειράς: (i) Κατανόηση (ii) Φυσικότητα.

MOS	GV & Dyn. Feat.	No Dyn. Feat.	No GV	$STRW = 0.5$	$STRW = 1$
Κατανόηση	3.5	1.85	4.05	3.28	2.3
Φυσικότητα	3.1	1.4	3.8	2.9	2.03

Πίνακας 4.2: Αξιολόγηση της συνθετικής εικονοσειράς: (i) Κατανόηση (ii) Φυσικότητα.

Σημαντικό είναι να παρατηρήσουμε το γεγονός ότι ενώ η χρήση της global variance στις ακουστικές παραμέτρους βελτιώνει τη φυσικότητα της συνθετικής φωνής, όταν χρησιμοποιείται στις οπτικές παραμέτρους, μειώνει την κατανόηση της εικονοσειράς κατά 0.55 μονάδες, και την φυσικότητα του κατά 0.7, προκαλώντας παραμόρφωση στο συνθετικό πρόσωπο.

Τέλος, παρατηρούμε τη μεγάλη επίδραση που έχει η επιλογή του stream weight των οπτικών παραμέτρων, καθώς κατά τη μεταβολή του από 0.0 σε 1.0 σημειώνεται υποβάθμιση της συνθετικής εικονοσειράς κατά 1.75 μονάδες όσον αφορά την κατανόηση, και κατά 1.67 μονάδες όσον αφορά τη φυσικότητα της εικονοσειράς, με την επιλογή $streamweight = 0.5$ να βρίσκεται σε ενδιάμεσες τιμές. Ως αποτέλεσμα συμπεραίνουμε ότι για την ευθυγράμμιση σε επίπεδο κατάστασης αρκεί να χρησιμοποιούνται μόνο τα ακουστικά χαρακτηριστικά. Σημαντικό γεγονός επίσης είναι το ότι η μεταβολή του stream weight των οπτικών παραμέτρων υποβαθμίζει την φυσικότητα του ήχου κατά 1.7 μονάδες και την κατανόηση του κατά 1.9 μονάδες.

4.4 Συμπεράσματα

Στο Κεφάλαιο αυτό προχωρήσαμε σε πρακτική υλοποίηση ενός συστήματος οπτικοακουστικής σύνθεσης φωνής με χρήση κρυφών μαρκοβιανών μοντέλων. Μετά την υλοποίηση αυτή, προχωρήσαμε σε μελέτη και αξιολόγηση του συνολικού συστήματος τόσο σε επίπεδο σύνθεσης φωνής, όσο και σε επίπεδο οπτικοακουστικής σύνθεσης φωνής. Κατά την αξιολόγηση αυτή, που έγινε μέσω tests MOS, μετρήσαμε την απόδοση του συστήματος όσον αφορά την φυσικότητα και την κατανόηση της συνθετικής ομιλίας, και όσον αφορά την φυσικότητα και την κατανόηση της συνθετικής εικονοσειράς.

Η μέτρηση των αποδόσεων αυτών έγινε και σε βαθμό των παραμέτρων που επηρεάζουν την ποιότητα της παραγόμενης εικονοσειράς, ώστε να μελετήσουμε ποιες παράμετροι και ποιες τεχνικές οδηγούν σε βελτίωση του αποτελέσματος.

Με βάση τα αποτελέσματα των ερωτηματολογίων, παρατηρήσαμε ότι η καλύτερη απόδοση του συ-

στήματος επιτυγχάνεται όταν χρησιμοποιούνται κατά την εκπαίδευση:

- Global Variance για τον ήχο
- Δυναμικές παράμετροι για τον ήχο
- $\gamma = 0$ των mel generalized cepstral coefficients για τον ήχο
- No Global Variance για την εικόνα
- Δυναμικές παράμετροι για την εικόνα
- StreamWeight $= 0$ για τα οπτικά χαρακτηριστικά, και άρα μη συμπερίληψη τους κατά την ευθυγράμμιση των κρυφών μαρκοβιανών μοντέλων σε επίπεδο κατάστασης με χρήση του αλγορίθμου Viterbi, επηρεάζοντας την εικόνα και τον ήχο ταυτόχρονα.

Κεφάλαιο 5

Συναισθηματική Οπτικοακουστική Σύνθεση Φωνής

Η συναισθηματική οπτικοακουστική σύνθεση φωνής αποτελεί ακόμα πεδίο ιδιαίτερα απαιτητικό παρ' όλο που υπάρχουν συστήματα εξαιρετικά ικανοποιητικά για την οπτικοακουστική σύνθεση φωνής χωρίς συναίσθημα, αλλά και επαρκώς ικανοποιητικά συστήματα για την συναισθηματική οπτικοακουστική σύνθεση φωνής [9]. Σε επίπεδο *σύνθεσης φωνής*, έχουν προταθεί αρκετές μέθοδοι για την συναισθηματική σύνθεση φωνής [81], χρησιμοποιώντας είτε μεθοδους προσαρμογής του συναισθήματος του ομιλητή της αρχικής βάσης δεδομένων σε ένα άλλο συναίσθημα, χρησιμοποιώντας ένα μικρό σύνολο από προτάσεις για αυτή την συναισθηματική κατάσταση, είτε με μοντελοποίηση με χρήση ιδιοδιανυσμάτων φωνών.

Κατά την επέκταση ενός συστήματος συναισθηματικής σύνθεσης φωνής σε σύστημα συναισθηματικής οπτικοακουστικής σύνθεσης φωνής, με χρήση active appearance μοντέλων, εμφανίζονται σημαντικά προβλήματα σε επιπεδο μοντελοποίησης του προσώπου, λόγω του γεγονότος ότι τα ιδιοδιανύσματα του μοντέλου εμπεριέχουν πληροφορία τόσο για την κίνηση του στόματος, όσο και για την έκφραση και την στάση του προσώπου, καθιστώντας εξαιρετικά δύσκολο το γεγονός να μοντελοποιηθούν τα χαρακτηριστικά αυτά ξεχωριστά [9].

Στο παρόν Κεφάλαιο, χρησιμοποιώντας την βάση δεδομένων CVSP-AV, επιχειρούμε μια πρώτη επέκταση του συστήματος μας για συναισθηματική οπτικοακουστική σύνθεση φωνής, με βάση τις τρεις άλλες συναισθηματικές καταστάσεις της βάσης δεδομένων (θυμός, χαρά, λύπη). Σκοπός μας είναι να αξιολογήσουμε τόσο τη φυσικότητα και την κατανόηση της συνθετικής φωνής των συστημάτων, όταν χρησιμοποιούνται οι 900 προτάσεις των άλλων συναισθημάτων, αντί του ουδέτερου, όσο και το βαθμό αποτύπωσης των συναισθημάτων, όταν χρησιμοποιείται μόνο συνθετική φωνή, και όταν χρησιμοποιείται συνθετική φωνή και εικονοσειρά.

5.1 Εκπαίδευση και Σύνθεση

Η εξαγωγή των παραμέτρων για τα υπόλοιπα τρία συναισθήματα της βάσης δεδομένων έγινε ακριβώς με τον ίδιο τρόπο που έχουμε περιγράψει για φωνή χωρίς συναίσθημα. Αναφέραμε ήδη ότι ειδικά σε επίπεδο μοντελοποίησης του προσώπου, η δυσκολία της καταγραφής των εκφράσεων του προσώπου, και της ομιλίας, είναι εξαιρετικά δύσκολη. Το γεγονός αυτό μπορούμε να το καταλάβουμε χρησιμοποιώντας και τον Πίνακα 5.1, όπου μπορούμε να δούμε ότι τα ιδιοδιανύσματα σχήματος που μοντελοποιούν το 98% της μεταβολης του προσώπου για τα συναισθήματα της χαράς, της λύπης και του θυμού, αυξάνονται σημαντικά σε αριθμό, σε σχέση με το ουδέτερο συναίσθημα.

Αφού έγινε η εξαγωγή των παραμέτρων με ίδια διαδικασία όπως στο Κεφάλαιο 4, προχωρήσαμε σε εκπαίδευση τριών επιπλέον συστημάτων συναισθηματικής σύνθεσης φωνής, και κατ' επέκταση συναισθηματικής οπτικοακουστικής σύνθεσης φωνής, για τα τρία συναισθήματα της χαράς, της λύπης, και του θυμού.

-	Ουδέτερο	Χαρά	Λύπη	Θυμός
Ιδιοσχήματα	10	18	26	20
Ιδιοδιανύσματα Υφής	32	41	30	20

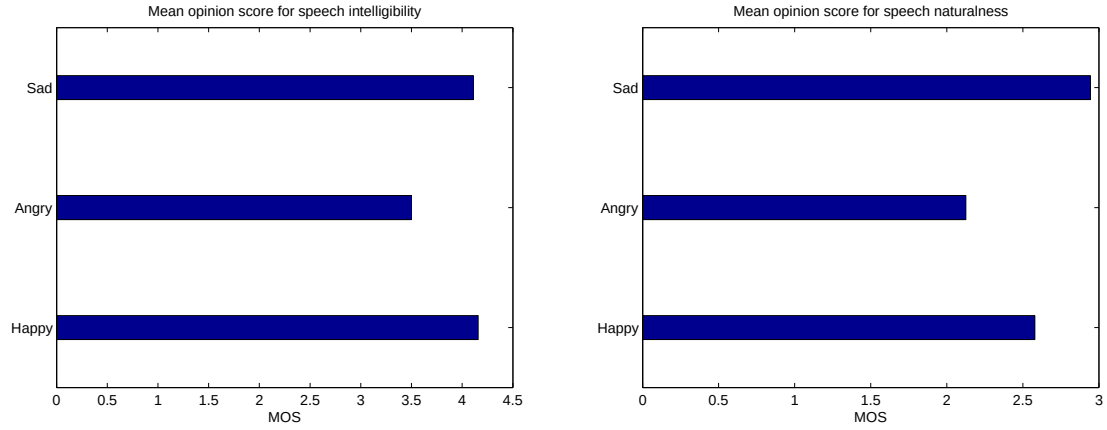
Πίνακας 5.1: Πίνακας αριθμού ιδιοσχημάτων και ιδιοδιανυσμάτων υφής που επεξηγούν το 98 % της μεταβολής του σχήματος και το 95 % της μεταβολής της υφής, αντίστοιχα, για κάθε ένα από τα τέσσερα διαφορετικά συναισθήματα.

5.2 Μελέτη των Συστημάτων Συναισθηματικής Σύνθεσης Φωνής

5.2.1 Μελέτη Φυσικότητας και Κατανόησης

Για την μελέτη και την αξιολόγηση των συστημάτων συναισθηματικής σύνθεσης φωνής, προχωρήσαμε αρχικά σε δημιουργία MOS test το οποίο περιγράψαμε στην Υποενότητα 4.3.1. Με βάση το test αυτό που συμπληρώθηκε επίσης από τους 13 ερωτηθέντες, θελήσαμε σε αρχικό στάδιο, *αγνοώντας* την συναισθηματική κατάσταση της ομιλίας, να μελετήσουμε τη φυσικότητα και την κατανόηση της συνθετικής συναισθηματικής ομιλίας, όταν για την εκπαίδευση του συστήματος χρησιμοποιούνται, αντί για τις 900 προτάσεις που αντιστοιχούν στο ουδέτερο συναίσθημα, οι 900 προτάσεις, από τα άλλα συναισθήματα της βάσης δεδομένων.

Οι ερωτηθέντες αξιολόγησαν ως προς την κατανόηση και την φυσικότητα, συνολικά 5 προτάσεις από κάθε συναίσθημα, αναμειγμένες επιπλέον με το ουδέτερο συναίσθημα (σύνολο 20 προτάσεις). Μετά τη συλλογή των αποτελεσμάτων του test MOS και την επεξεργασία τους λάβαμε τα διαγράμματα φυσικότητας και κατανόησης που φαίνονται στο Σχήμα 5.1 και στον Πίνακα 5.2.



(i) MOS Κατανόησης του ήχου για 3 συναισθήματα **(ii)** MOS Φυσικότητας του ήχου για 3 συναισθήματα.

Σχήμα 5.1: Αξιολόγηση της συναισθηματικής ομιλίας αγνοώντας τη συναισθηματική κατάσταση: (i) Κατανόηση (ii) Φυσικότητα.

Από τα αποτελέσματα, παρατηρούμε ότι η κατανόηση της συνθετικής φωνής, παραμένει υψηλή όταν χρησιμοποιούνται τα τρία επιπλέον συναισθήματα της βάσης δεδομένων, είναι όμως υποβαθμισμένη

MOS	Χαρά	Θυμός	Λύπη
Κατανόηση	4.15	3.5	4.11
Φυσικότητα	2.6	2.2	2.94

Πίνακας 5.2: Πίνακας αποτελεσμάτων φυσικότητας και κατανόησης συνθετικής ομιλίας συναισθημάτων

σε σχέση με τα αποτελέσματα κατανόησης που λάβαμε για το ουδέτερο συναίσθημα στο Σχήμα 4.1.

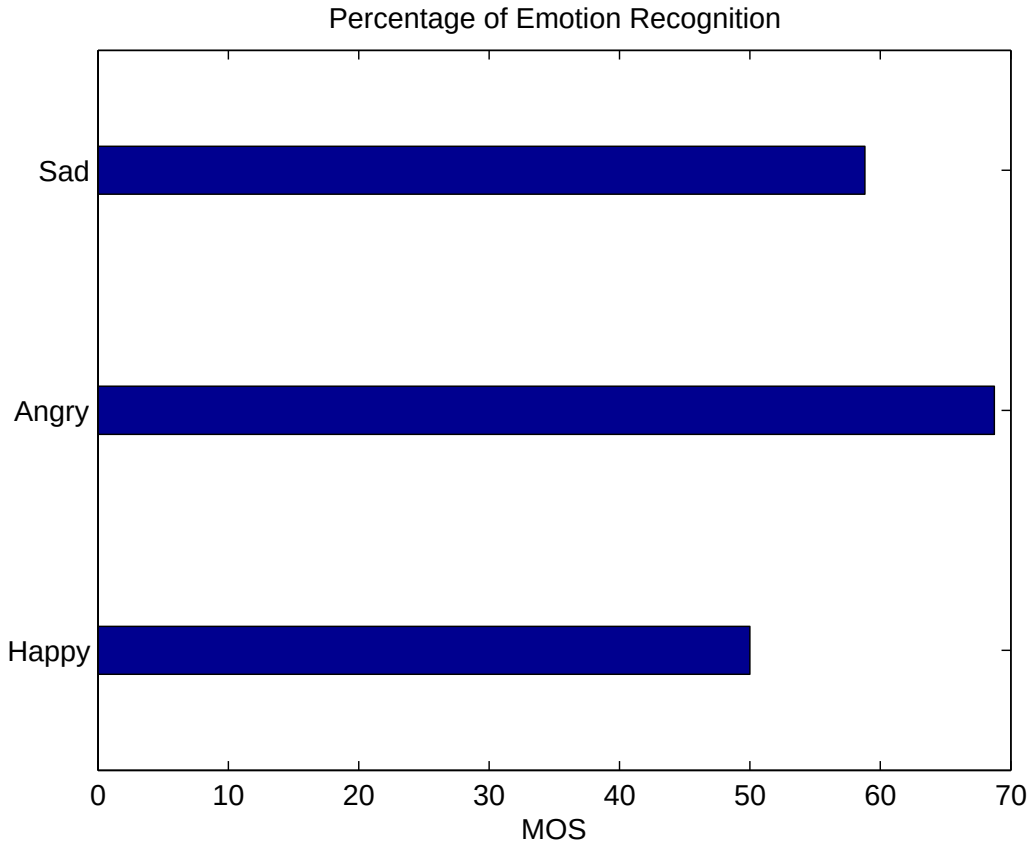
Σε αντίθεση με την κατανόηση της συνθετικής ομιλίας, η φυσικότητα της ομιλίας υποβαθμίζεται σημαντικά, ειδικά όταν για την εκπαίδευση του συστήματος χρησιμοποιούνται οι 900 προτάσεις του συναισθήματος του θυμού.

5.2.2 Μελέτη Αναγνώρισης Συναισθημάτων

Στη συνέχεια, θέλωντας να αξιολογήσουμε το βαθμό στον οποίο αποτυπώνεται κάθε συναίσθημα από τις συνθετικές προτάσεις των συστημάτων συνθετικής φωνής, προχωρήσαμε σε ένα απλό test πολλαπλής επιλογής, κατά το οποίο οι 13 ερωτηθέντες κλήθηκαν να αναγνωρίσουν το συναίσθημα που αποτυπώνεται σε 20 διαφορετικές προτάσεις, όπου ήταν αναμειγμένες συνθετικές προτάσεις από τα συνολικά τέσσερα συστήματα σύνθεσης φωνής (ένα για κάθε μία συναισθηματική κατάσταση).

Κατά τη δημιουργία του test πολλαπλής επιλογής, δόθηκαν σαφείς οδηγίες στους ερωτηθέντες, έτσι ώστε η απόφαση για την αντιστοιχία του συναισθήματος να λαμβάνεται απ' ευθείας μετά το άκουσμα της συνθετικής πρότασης, και όχι αφού ακουστούν όλες οι προτάσεις, έτσι ώστε τα αποτελέσματα του test να είναι όσο το δυνατόν πιο αξιόπιστα. Οι οδηγίες αυτές δόθηκαν λόγω του γεγονότος ότι μετά το άκουσμα όλων των προτάσεων, η αντιστοιχία κάθε πρότασης σε ένα συναίσθημα μπορεί να μην βασιστεί αποκλειστικά στην πρόταση αυτή, αλλά και στη σύγκριση με άλλες προτάσεις, γεγονός που θελήσαμε να αποφύγουμε.

Μετά την επεξεργασία των αποτελεσμάτων λάβαμε τα ποσοστά επιτυχίας για κάθε συναίσθημα που φαίνονται στο Σχήμα 5.2.



Σχήμα 5.2: Επιτυχία αναγνώρισης των τριών συναισθημάτων (από επιλογή μαζί με το ουδέτερο) μόνο με ήχο.

Με την μελέτη του Σχήματος, βλέπουμε ότι ο θυμός αποτελεί τη συναισθηματική κατάσταση που μπορεί και αποτυπώνεται ευκολότερα στη συνθετική ομιλία, παρ' όλο που όπως είδαμε στο Σχήμα 5.1 παρουσιάζει τη μικρότερη φυσικότητα συνθετικής φωνής αν αγνοήσουμε το συναίσθημα, ενώ η χαρά τη δυσκολότερη. Στο ενδιάμεσο βρίσκεται το συναίσθημα της λύπης. Θα πρέπει εδώ να σημειώσουμε το γεγονός, ότι η υποβαθμισμένη φυσικότητα που παρουσιάζουν τα συστήματα συνθετικής φωνής όπως είδαμε στο Σχήμα 5.1, επηρεάζει τα ποσοστά επιτυχίας της αναγνώρισης των συναισθηματικών καταστάσεων.

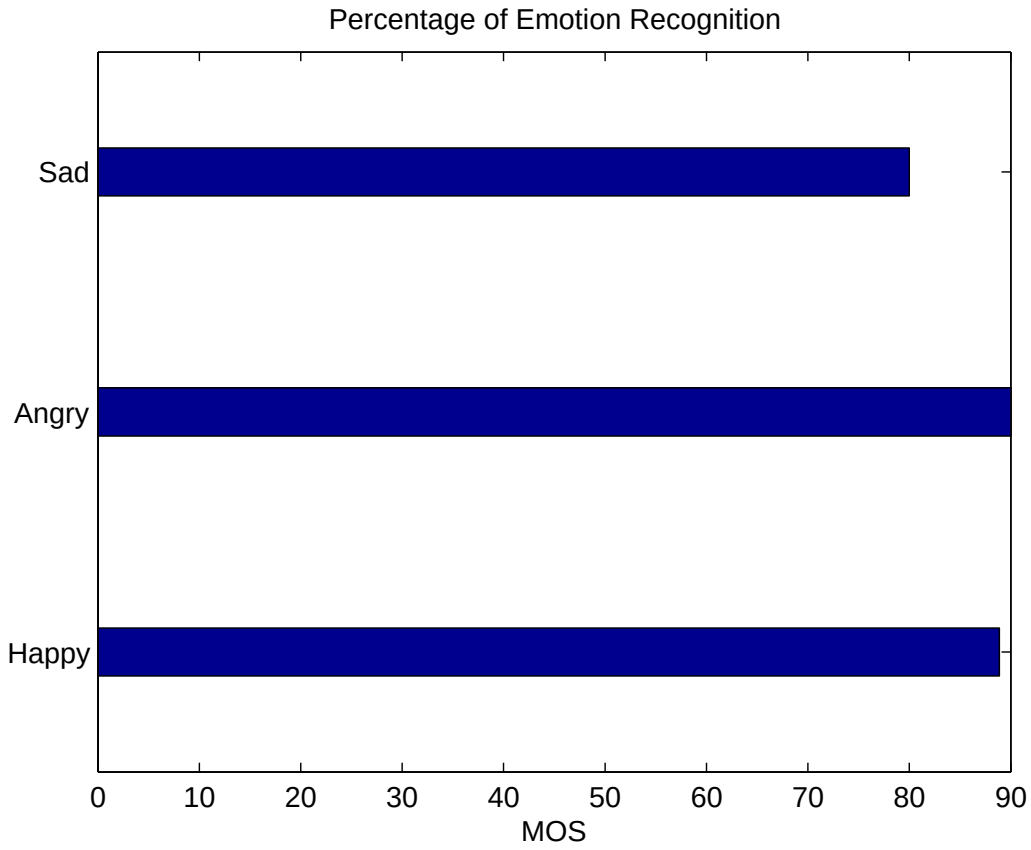
5.3 Μελέτη των Συστημάτων Συναισθηματικής Οπτικοακουστικής Σύνθεσης Φωνής

Στο τέλος, προχωρήσαμε σε επέκταση των τριών επιπλέον συστημάτων σύνθεσης φωνής σε συστήματα οπτικοακουστικής σύνθεσης φωνής, ενσωματώνοντας στην εκπαίδευση και τις παραμέτρους που πήραμε από τα αντίστοιχα παραμετρικά μοντέλα που μοντελοποιούσαν τις κινήσεις του προσώπου για κάθε συναίσθημα. Σκοπός μας σε αυτή την ενότητα είναι να δείξουμε ότι η χρήση οπτικής πληροφορίας αυξάνει σημαντικά την επιτυχία αναγνώρισης του συναισθήματος της συνθετικής πρότασης, σε σύγκριση με την περίπτωση που χρησιμοποιείται μόνο συνθετική φωνή.

Για να αποδείξουμε το γεγονός αυτό, προχωρήσαμε σε εκπαίδευση των συστημάτων με συνολικά 100 προτάσεις, από τις 900 που αντιστοιχούν σε κάθε συναίσθημα, ενσωματώνοντας τις οπτικές παραμέτρους. Μετά την εκπαίδευση και την σύνθεση των προτάσεων, κατασκευάσαμε test πολλαπλής επιλογής, με τις σαφείς οδηγίες που δόθηκαν και στην προηγούμενη ενότητα, χρησιμοποιώντας αντί για σκέτη συνθετική φωνή, συνθετική φωνή και εικονοσειρά αυτή τη φορά.

Μετά την επεξεργασία των αποτελεσμάτων, παίρνουμε τα αποτελέσματα που φαίνονται στο Σχήμα 5.3

Παρατηρούμε ότι η αύξηση της επιτυχίας αναγνώρισης του συναισθήματος σε σχέση μόνο με την συνθετική φωνή είναι εξαιρετική. Παρατηρείται αύξηση της επιτυχίας στο συναίσθημα της χαράς κατά 38%, της λύπης κατά 22%, και του θυμού επίσης κατά 22%, επιβεβαιώνοντας τον αρχικό μας σκοπό που αναφέραμε στην αρχή της Ενότητας, ότι η χρήση της εικονοσειράς αυξάνει δραματικά την αναγνώριση του συναισθήματος.



Σχήμα 5.3: Επιτυχία αναγνώρισης των τριών συναισθημάτων (από επιλογή μαζί με το ουδέτερο) με εικονοσειρά και ήχο.

5.4 Συμπεράσματα

Στο Κεφάλαιο αυτό προχωρήσαμε σε μία πρώτη μελέτη και επέκταση του συστήματος μας σε συναισθηματική σύνθεση φωνής και σε συναισθηματική οπτικοακουστική σύνθεση φωνής. Με την μελέτη αυτή, θελήσαμε να αξιολογήσουμε την απόδοση του συστήματος σύνθεσης φωνής και οπτικοακουστικής σύνθεσης φωνής που υλοποιήσαμε στο Κεφάλαιο 4, όταν για την εκπαίδευση του χρησιμοποιούνται οι προτάσεις για τα διαφορετικά συναισθήματα της χαράς, της λύπης και του θυμού της βάσης δεδομένων CVSP-AV.

Η μελέτη αυτή των συστημάτων που εκπαιδεύτηκαν με τις προτάσεις των συναισθημάτων, έγινε αρχικά σε επιπέδο της κατανόησης και της φυσικότητας της παραγόμενης ομιλίας, αγνοώντας την συναισθηματική κατάσταση που αποτυπώνεται στη συνθετική πρόταση. Τα αποτελέσματα αυτής της μελέτης ήταν ιδιαίτερα ενθαρρυντικά, παρ' όλο που δεν προχωρήσαμε σε καμία βελτιστοποίηση των συστημάτων όσον αφορά τα διαφορετικά συναισθήματα και τις παραμέτρους τους.

Η μελέτη μας επίσης επικεντρώθηκε στην αξιολόγηση της αποτύπωσης των συναισθημάτων στις συνθετικές προτάσεις, τόσο σε επίπεδο σκέτης συνθετικής φωνής, όσο και σε επίπεδο ενσωμάτωσης και εικονοσειράς στις συνθετικές προτάσεις. Τα αποτελέσματα μας ήταν επίσης ιδιαίτερα ενθαρρυντικά, καθώς δείξαμε ότι αν και η αποτύπωση των συναισθημάτων στα μη βελτιστοποιημένα συστήματα συνθετικής φωνής, κυμάνθηκε σε ποσοστά επιτυχίας 50% για την χαρά, 58% για την λύπη, και 69% για τον θυμό, όταν χρησιμοποιούνται για την εκπαίδευση και τα οπτικά χαρακτηριστικά, και κατ' επέκταση έχουμε συνθετική ομιλία και εικονοσειρά, τα ποσοστά επιτυχία αυξάνονται δραματικά, κατά 38% για το συναίσθημα της χαράς, κατά 22% για το συναίσθημα της λύπης και κατά 22% για το συναίσθημα του θυμού.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντική Έρευνα

6.1 Ανακεφαλαίωση και Συνεισφορές

Η παρούσα διπλωματική εργασία παρουσίασε ένα ολοκληρωμένο και άρτιο σύστημα οπτικοακουστικής σύνθεσης φωνής τόσο σε θεωρητικό, όσο και σε πρακτικό επίπεδο. Το σύστημα μας αξιολογήθηκε από 13 ερωτηθέντες με βάση το mean opinion score test, ως προς τη φυσικότητα του και ως προς την κατανόηση του λόγου σε επίπεδο ομιλίας και εικονοσειράς. Όπως είδαμε και στο Κεφάλαιο 4, τα αποτελέσματα ήταν ιδιαίτερα ενθαρρυντικά, ανοίγοντας τον ορίζοντα για περαιτέρω βελτιστοποίηση και εφαρμογές του συστήματος μας, όπως ιδιαίτερα ενθαρρυντικά ήταν και τα αποτελέσματα μιας πρώτης προσέγγισης στην συναισθηματική οπτικοακουστική σύνθεση φωνής που είδαμε στο Κεφάλαιο 5.

Ανακεφαλαιώνοντας, μπορούμε να συνοψίσουμε τις επιστημονικές συνεισφορές της παρούσης διπλωματικής εργασίας στις κάτωθι:

- **Δημιουργία πλήρους βάσης δεδομένων για οπτικοακουστική σύνθεση φωνής.** Πιο συγκεκριμένα, υλοποιήθηκε στα πλαίσια της παρούσης διπλωματικής εργασίας η βάση δεδομένων CVSP-AudioVisual (AV) κατάλληλη για συστήματα σύνθεσης φωνής, και οπτικοακουστικής σύνθεσης φωνής, που βασίζονται σε κρυφά μαρκοβιανά μοντέλα, περιέχοντας 900 προτάσεις για κάθε ένα από τέσσερα διαφορετικά συναισθήματα: ουδέτερο, χαρά, λύπη, θυμός.
- **Θεωρητική ανάλυση των ακουστικών και οπτικών παραμέτρων καθώς και του λεξιλογικού περιεχομένου, που χρησιμοποιούνται για την εκπαίδευση ενός συστήματος οπτικοακουστικής σύνθεσης φωνής με βάση τα κρυφά μαρκοβιανά μοντέλα, καθώς και πρακτική προγραμματιστική υλοποίηση της εξαγωγής των παραμέτρων**. Η πρακτική αυτή προγραμματική υλοποίηση της εξαγωγής των παραμέτρων και του λεξιλογικού περιεχομένου έγινε με χρήση των εργαλείων SPTK [64], STRAIGHT [45], sail-align [70], Festival [61], MATLAB [63], aam-tools [3] και am-tools [7].
- **Θεωρητική ανάλυση και πρακτική υλοποίηση ενός πλήρους συστήματος οπτικοακουστικής σύνθεσης φωνής.** Προχωρήσαμε σε θεωρητική ανάλυση ενός συστήματος οπτικοακουστικής σύνθεσης φωνής επεκτείνοντας την ήδη υπάρχουσα ανάλυση ενός συστήματος σύνθεσης φωνής [81] για σύστημα οπτικοακουστικής σύνθεσης φωνής. Η πρακτική υλοποίηση έγινε επεκτείνοντας το εργαλείο σύνθεσης φωνής με χρήση κρυφών μαρκοβιανών μοντέλων HTS [62] - με χρήση των συστήματων HTK [4], SPTK, STRAIGHT, MATLAB και aam-tools, καθώς και νέων εργαλείων που δημιουργήθηκαν για την υλοποίηση του συστήματος, όπως ένα πλήρες front-end τμήμα επεξεργασίας της λεξιλογικής πληροφορίας κειμένου - σε σύστημα οπτικοακουστικής σύνθεσης φωνής για την Ελληνική γλώσσα.
- **Αξιολόγηση του συστήματος οπτικοακουστικής σύνθεσης φωνής.** Έγινε μελέτη της ποιότητας των παραγόμενων προτάσεων, τόσο όσον αφορά τον ήχο, όσο και την εικονοσειρά, χρησιμοποιώντας tests MOS [44], μελετώντας την επίδραση διαφόρων παραμέτρων και τεχνικών

στην ποιότητα του αποτελέσματος, έτσι ώστε να οδηγηθούμε σε ένα εξαιρετικά ποιοτικό αποτέλεσμα.

- **Μελέτη και αξιολόγηση συναισθηματικής οπτικοακουστικής σύνθεσης φωνής.** Πιο συγκεκριμένα, εφαρμόσαμε το σύστημα οπτικοακουστικής σύνθεσης φωνής που υλοποιήσαμε στα τρία άλλα συναισθήματα της βάσης δεδομένων CVSP-AV, με τα πρώτα αποτελέσματα που είδαμε στο Κεφάλαιο 5 να τυγχάνουν βελτίωσης, αλλά και παράλληλα να είναι ιδιαίτερα ενθαρρυντικά, τόσο σε επίπεδο κατανόησης και φυσικότητας της συνθετικής ομιλίας, όσο και σε επίπεδο αποτύπωσης των συναισθημάτων.

6.2 Προεκτάσεις για μελλοντική έρευνα

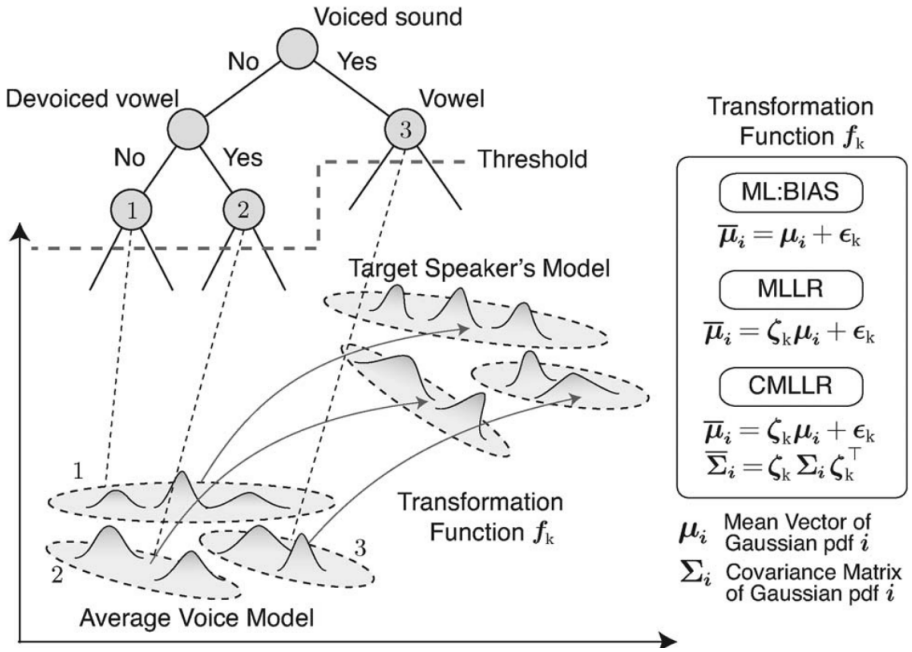
Η οπτικοακουστική σύνθεση φωνής με χρήση κρυφών μαρκοβιανών μοντέλων, χαίρει μεγάλης ευελιξίας και ελαστικότητας. Σημαντικές προεκτάσεις αποτελούν:

- **Προσαρμογή της φωνής ενός ομιλητή στην φωνή άλλου ομιλητή**. Τα συστήματα σύνθεσης φωνής με χρήση κρυφών μαρκοβιανών μοντέλων, με χρήση μεθόδων MLLR (maximum likelihood linear regression) [76, 51, 84], SAT (speaker adaptive training) [8] και CMLLR [40] (constrained maximum likelihood linear regression), παρουσιάζουν σημαντική επιτυχία στην προσαρμογή της φωνής σε μία νέα φωνή για την οποία υπάρχουν λίγα κομμάτια προηχογραφημένης ομιλίας. Ένα διάγραμμα τεχνικών της προσαρμογής αυτής μπορούμε να δούμε στο Σχήμα 6.1.

- **Interpolation**. Τα συστήματα οπτικοακουστικής σύνθεσης φωνής με χρήση κρυφών μαρκοβιανών μοντέλων δύνανται επίσης, να χρησιμοποιήσουν μίξη φωνής και εικονοσειράς από διαφορετικά σύνολα εκπαιδευμένων κρυφών μαρκοβιανών μοντέλων.

- **Συναισθηματική Οπτικοακουστική Σύνθεση Φωνής**. Μέσω της βάσης δεδομένων CVSP-AV που δημιουργήσαμε, αλλά και των δύο προηγούμενων προεκτάσεων που αναφέραμε, μελλοντικός σημαντικός στόχος είναι όχι μόνο η κατασκευή πλήρων και εξαιρετικών συστημάτων συναισθηματικής οπτικοακουστικής σύνθεσης φωνής, αλλά και μίξη των συνόλων των κρυφών μαρκοβιανών μοντέλων για την παραγωγή ενδιάμεσων συναισθημάτων. Απαραίτητη είναι και η επέκταση της εξαγωγής των παραμέτρων της εικονοσειράς, για περαιτέρω βελτίωση του οπτικού αποτελέσματος, χρησιμοποιώντας segmented active appearance models για τις διαφορετικές περιοχές του προσώπου, αλλά και pose invariant active appearance models [9].

- **Ιστοσελίδα οπτικοακουστικής σύνθεσης φωνής**. Στόχος του συνολικού εγχειρήματος, είναι επίσης η δημιουργία μιας πλήρους ιστοσελίδας στο διαδίκτυο, η κατασκευή της οποίας είναι σε εξέλιξη, για οπτικοακουστική σύνθεση φωνής διαφόρων συναισθημάτων, με βάση το κείμενο που εισάγει ο εκάστοτε χρήστης.



Σχήμα 6.1: Τεχνικές για την προσαρμογή ομιλητών σε σύστημα σύνθεσης φωνής [81].

Παράρτημα 1

Προσέγγιση Padé

Στο παράρτημα αυτό θα περιγράψουμε συνοπτικά την προσέγγιση Padé που χρησιμοποιήθηκε στο Κεφάλαιο 2 για την εύρεση ρητής προσέγγισης της συνάρτησης μεταφοράς $D(z)$ και συνήθως αποτελεί καλύτερη προσέγγιση σε σχέση με τη σειρά Taylor.

Η προσέγγιση Padé $[L, M]$ της δυναμοσειράς μιας συνάρτησης $f(x)$ [11]:

$$f(x) = \sum_{k=0}^{\infty} c_k x^k \tag{1.1}$$

ορίζεται ως:

$$R_{L,M}(x) = \frac{\sum_{k=0}^{L} p_k x^k}{1 + \sum_{k=1}^{M} q_k x^k} \tag{1.2}$$

έτσι ώστε οι συντελεστές της δυναμοσειράς της $R_{L,M}(x)$ να είναι ίσοι με τους πρώτους $L + M + 1$ συντελεστές της δυναμοσειράς της $f(x)$:

$$f(x) - R_{L,M}(x) = O\left(x^{L+M+1}\right) \tag{1.3}$$

Για την εύρεση των συντελεστών, μπορούμε να πολλαπλασιάσουμε την εξίσωση 1.3 με τον παρονομαστή της $R_{L,M}$ και να εξισώσουμε τους συντελεστές του x^k για $k = 0, ..., L + M$. Ως αποτέλεσμα παίρνουμε τις εξισώσεις για τους συντελεστές $q_k, k = 1, ..., M$:

$$\sum_{k=1}^{\min(r,M)} q_k c_{r-k} = -c_r \qquad (r = L+1, ..., L+M) \tag{1.4}$$

και τις L εξισώσεις για τους συντελεστές $p_k, k = 0, ..., L$

$$p_k = c_k + \sum_{s=1}^{\min(k,M)} q_s c_{k-s} \qquad (k = 0, ..., L) \tag{1.5}$$

Τις περισσότερες φορές όπου οι εξισώσεις 1.4 έχουν λύση, είναι βολικό να παίρνουμε L = M δηλαδή, διαγώνιες προσεγγίσεις Padé, όπως ακριβώς κάναμε και στο Κεφάλαιο 2 για την προσέγγιση της εκθετικής συνάρτησης. Για $M = 0$ η προσέγγιση Padé αντιστοιχεί στη σειρά McLaurin.

Υπάρχουν αρκετοί αλγόριθμοι για την αποδοτική εύρεση των συντελεστών της προσέγγισης Padé. Ένας από τους αλγορίθμους αυτούς είναι ο αλγόριθμος *epsilon* του Wynn [85], άλλοι αλγόριθμοι

περιλαμβάνουν μετασχηματισμούς ακολουθίας που μπορούν να βρεθούν στο [23], ενώ μπορεί να χρησιμοποιηθεί και ο εκτεταμένος ευκλείδιος αλγόριθμος για την εύρεση του πολυωνυμικού μέγιστου κοινού διαιρέτη.

Παράρτημα 2

Λεξιλογική πληροφορία που παρέχεται στο σύστημα

Το τωρινό φώνημα. Το προηγούμενο φώνημα. Το προ-προηγούμενο φώνημα. Το επόμενο φώνημα. Το μεθεπόμενο φώνημα.
Το γεγονός αν η προηγούμενη συλλαβή τονίζεται. Ο αριθμός των φωνημάτων στην προηγούμενη συλλαβή.
Το γεγονός αν η τωρινή συλλαβή τονίζεται. Ο αριθμός των φωνημάτων στην τωρινή συλλαβή. Η θέση της τωρινής συλλαβής στη λέξη (προς τα εμπρός). Η θέση της τωρινής συλλαβής στη λέξη (προς τα πίσω). Η θέση της τωρινής συλλαβής στη φράση (προς τα εμπρός). Η θέση της τωρνής συλλαβής στη φράση (προς τα πίσω). Ο αριθμός των τονισμένων συλλαβών πριν την τωρινή συλλαβή στη φράση. Ο αριθμός των τονισμένων συλλαβών μετά την τωρινή συλλαβή στη φράση. Ο αριθμός των συλλαβών από την προηγούμενη τονισμένη συλλαβή μέχρι την τωρινή. Ο αριθμός των συλλαβών από την τωρινή τονισμένη συλλαβή έως την επόμενη. Το όνομα του φωνήεντος της τωρινής συλλαβής.
Το γεγονός αν η επόμενη συλλαβή τονίζεται. Ο αριθμός των φωνημάτων στην επόμενη συλλαβή.
Guess part-of-speech της προηγούμενης λέξης. Ο αριθμός των συλλαβών στην προηγούμενη λέξη.
Guess part-of-speech της τωρινής λέξης. Η θέση της τωρινής λέξης στην φράση (προς τα εμπρός). Η θέση της τωρινής λέξης στη φράση (προς τα πίσω). Ο αριθμός των content words (Ουσιαστικά, Ρήματα, Επίθετα και Επιρρήματα) πριν την τωρινή λέξη στην φράση. Ο αριθμός των content words μετά την τωρινή λέξη στη φράση. Ο αριθμός των λέξεων από την προηγούμενη content word έως την τωρινή. Ο αριθμός των λέξεων από την τωρινή λέξη έως την επόμενη content word.
Guess part-of-speech της επόμενης λέξης. Ο αριθμός των συλλαβών στην επόμενη λέξη.
Ο αριθμός των συλλαβών στην προηγούμενη φράση. Ο αριθμός των λέξεων στην προηγούμενη φράση.
Ο αριθμός των συλλαβών στην τωρινή φράση. Ο αριθμός των λέξεων στην τωρινή φράση Η θέση της τωρινής φράσης στη συνολική πρόταση (προς τα εμπρός). Η θέση της τωρινής φράσης στη συνολική πρόταση (προς τα πίσω).
Ο αριθμός των συλλαβών στην επόμενη φράση. Ο αριθμός των λέξεων στην επόμενη φράση.
Ο αριθμός των συλλαβών στη συνολική πρόταση. Ο αριθμός των λέξεων στη συνολική πρόταση. Ο αριθμός των φράσεων στη συνολική πρόταση.

Πίνακας 2.1: Πίνακας Λεξιλογικής Πληροφορίας που παρέχεται στο σύστημα για κάθε φώνημα.

Βιβλιογραφία

[1] HTS Slides. [Online]. Available: http://hts.sp.nitech.ac.jp/.

[2] Intl. Phonetic Alphabet. [Online]. Available: https://www.internationalphoneticassociation.org/.

[3] aam-tools. [Online]. Available: http://cvsp.cs.ntua.gr/software/AAMtools.

[4] HTK. [Online]. Available: http://htk.eng.cam.ac.uk/.

[5] O. Abdel-Hamid, S. Abdou and M. Rashwan, "Improving arabic hmm based speech synthesis quality", in *Proc. of Intl. Conf. on Spoken Language Processing (Interspeech-2006)*, Pittsburgh, PA, 2006.

[6] I. Albrecht et al., "Mixed feelings: expression of non-basic emotions in a musclebased talking head.", *Virtual Reality*, vol. 8, pp. 201–212, 2005.

[7] am-tools. [Online]. Available: http://personalpages.manchester.ac.uk/staff/timothy.f.cootes.

[8] T. Anastasakos et al., "A compact model for speaker-adaptive training", in *Proc. of The Fourth Intl. Conf. on Spoken Language Processing (ICLSP-1996)*, Philadelphia, PA, 1996.

[9] R. Anderson et al., "Expressive visual text-to-speech using active appearance models", in *Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR-2013)*, Portland, OR, 2013.

[10] G. Bailly, "Audiovisual speech synthesis", *Intl. J. of Speech Technology*, vol. 6, pp. 6–331, 2001.

[11] G. A. Jr. Baker and P. Graves-Morris, *Pade Approximants*. Cambridge University Press, 1996.

[12] L. E. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process", *Inequalities*, vol. 3, pp. 1–8, 1972.

[13] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology", *Bulletin of the American Mathematical Society*, vol. 73, pp. 360–363, 1967.

[14] L. E. Baum et al., "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains", *The Annals of Mathematical Statistics*, vol. 41, pp. 164–171, 1970.

[15] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains", *The Annals of Mathematical Statistics*, vol. 37, pp. 1554–1563, 1966.

[16] L. E. Baum and G. R. Sell, "Growth transformations for functions on manifolds", *Pacific J. of Mathematics*, vol. 27, pp. 211–227, 1968.

[17] J. Beskow, "Talking heads – communication, articulation and animation", in *Proc. of Fonetik (Fonetik-1996)*, Nasslingen, Sweden, 1996.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, 2006.

[19] A. W. Black, "Unit selection and emotional speech", in *European Conf. on Speech Communication and Technology (EUROSPEECH-2003)*, Geneva, Switzerland, 2003.

[20] A. W. Black and K. Tokuda, "The blizzard challenge – 2005: Evaluating corpus-based speech synthesis on common datasets", in *Proc. of Intl. Conf. on Spoken Language Processing (Interspeech-2005)*, Lisbon, Portugal, 2005.

[21] F.L. Bookstein, *Morphometric Tools for Landmark Data*. Cambridge Univ. Press, 1991.

[22] M. Brand, "Voice puppetry", in *Proc. of the 26th annual Conf. on Computer graphics and interactive techniques (SIGGRAPH-1999)*, Los Angeles, CA, 1999.

[23] C. Brezenski, "Extrapolation algorithms and pade approximations", *Applied Numerical Mathematics*, vol. 20, pp. 299–318, 1996.

[24] N. Campbell, "Developments in corpus-based speech synthesis: Approaching natural conversational speech", *IEICE trans. on Information and Systems*, pp. 376–383, 2005.

[25] A. Chalamandaris et al., "The ilsp/innoetics text-to-speech system for the blizzard challenge 2013", in *Blizzard Challenge 2013 Workshop*, Reykjavik, Iceland, Barcelona, Spain, 2013.

[26] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech", in *Models and Techniques in Computer Animation*, Lausanne, Switzerland, 1993.

[27] R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen and V. Zue, *Survey of the State of the Art in Human Language Techonology*. Cambridge Univ. Press, 1996.

[28] T. F. Cootes, G. J. Edwards and C.J. Taylor, "Active appearance models", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 681–685, 1998.

[29] T. F. Cootes et al., "Active shape models - their training and applications", *Computer Vision and Image Understanding*, vol. 61, pp. 38–59, 1995.

[30] T.F. Cootes and C.J. Taylor, "Statistical models of appearance for computer vision", 2004.

[31] S. Deena, S. Hou and A. Galata, "Visual speech synthesis by modelling coarticulation dynamics using a non-parametric switching state-space model", in *Proc. of 12th Intl Conf. on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*, Beijing, China, 2010.

[32] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, 1997.

[33] P. Dymarski, *Hidden Markov Models, Theory and Applications*. InTech, 2011.

[34] H. M. El-Bakry, M. Z. Rashad and I. R. Isma'il, "Diphone-based concatenative speech synthesis systems for arabic language", in *Proc. of the 10th WSEAS Intl. Conf. on Circuits, Systems, Electronics, Control and Signal Processing, and Proc. of the 7th WSEAS Intl. Conf. on Applied and Theoretical Mechanics (CSECS-2011/MECHANICS-2011)*, Chicago, IL, 2011.

[35] G. Fant, *Acoustic theory of speech production*. Mouton & Co, 1970.

[36] J. Flanagan, *Speech Analysis, Synthesis, and Perception*. Springer-Verlag, 1972.

[37] J. Flanagan and L. Rabiner, *Speech Synthesis*. Dowden - Hutchinson & Ross Inc., 1973.

[38] G. D. Forney, "The viterbi algorithm: A personal history". April 2005.

[39] T. Fukada et al., "An adaptive algorithm for mel-cepstral analysis of speech", in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP-1992)*, San Francisco, CA, 1992.

[40] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition", *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[41] H. M. Hanson and K.N. Stevens, "A quasiarticulatory approach to controlling acoustic source parameters in a klatt-type formant synthesizer using hlsyn", *J. Acoustical Society of America*, vol. 112, pp. 1158–1182, 2002.

[42] Z. Heiga et al., "Recent development of the hmm-based speech synthesis system (hts)", in *Proc. of Asia-Pacific Signal and Information Processing Association (APSIPA-2009)*, Sapporo, Japan, 2009.

[43] S. Imai and C. Furuichi, "Unbiased estimator of log spectrum and its application to speech signal processing", in *Proc. of European Signal Processing Conf. (EURASIP-1988)*, Grenoble, France, 1988.

[44] International Telecommunication Union, *Methods for subjective determination of transmission quality.*, 1996.

[45] H. Kawahara, J. Estill and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight", in *Intl. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA-2001)*, Firenze, Italy, 2001.

[46] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds", *Speech Communication*, pp. 197–207, 1999.

[47] D. Klatt, "Review of text-to-speech synthesis conversion for english", *J. of the Acoustical Society of America*, vol. 82, pp. 737–793, 1987.

[48] T. Kobayashi and S. Imai, "Spectral analysis using generalized cepstrum", *Acoustics, Speech and Signal Processing*, vol. 32, pp. 1087–1089, 1984.

[49] B. Krogen, "Minimal rules for articulatory speech synthesis", in *Proc. of European Signal Processing Conf. (EUSIPCO-1992)*, Brussels, Belgium, 1992.

[50] S. Krstulovic, A. Hunecke and M. Schroeder, "An hmm-based speech synthesis system applied to german and its adaptation to a limited set of expressive football announcements", in *Proc. of Interspeech*, Antwerp, Belgium, 2007.

[51] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models", *Computer Speech and Language*, vol. 9, pp. 171–185, 2001.

[52] S. Lemmetty, "Review of speech synthesis technology", Master's thesis, Dept. Elect. Eng., Univ. of Technology, Helsinky, 1999.

[53] K. Liu and J. Ostermann, "Realistic facial expression synthesis for an image-based talking head.", in *Intl. Conf. on Multimedia and Expo (ICME-2007)*, Barcelona, Spain, 2011.

[54] A. Lundgren, "An hmm-based text-to-speech system applied to swedish", Master's thesis, Royal Institute of Technology (KTH), Stockholm, 2005.

[55] T. Masuko, *HMM-Based Speech Synthesis and Its Applications*. PhD thesis, Dept. of Institute of Technology, Tokyo, 2002.

[56] Iain Matthews and Simon Baker, "Active appearance models revisited", *Int. J. Computer Vision*, vol. 60, pp. 135–164, 2004.

[57] I. G. Mattingly and T. A. Sebeok, "Speech synthesis for phonetic and phonological models", *Current Trends in Linguistics*, vol. 12, pp. 2451–2487, 1974.

[58] D. C. Mcgurk et al., "Towards perceptually realistic talking heads: Models, methods and mcgurk", in *Symp. on Applied Perception in Graphics and Visualization (APGV-2004)*, Los Angeles, CA, 2004.

[59] H. McGurk and J. Macdonald, "Hearing lips and seeing voices", *Nature*, vol. 264, pp. 746–748, 1976.

[60] Julian James Odell, *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, Univesity of Cambridge, 1995.

[61] Festival. [Online].

[62] HTS. [Online]. Available: http://hts.sp.nitech.ac.jp.

[63] MATLAB. [Online].

[64] SPTK. [Online], "Available: http://sp-tk.sourceforge.net".

[65] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Prentice-Hall, 1989.

[66] J. Ostermann and A. Weissenfeld, "Talking faces - technologies and applications", in *Proc. of Intl. Workshop on Systems, Signal and Image Processing (IWSSIP-2004)*, Poznan, Poland, 2004.

[67] G. Papandreou and P. Maragos, "Adaptive and constrained algorithms for inverse compositional active appearance model fitting", in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR-2008)*, Anchorage, AK, 2008.

[68] M. D. Polkosky and J. R. Lewis, "Expanding the mos: Development and psychometric evaluation of the mos-r and mos-x", *Intl. J. of Speech Technology*, pp. 161–182, 2003.

[69] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall International, Inc., 1993.

[70] sail-align. [Online], "Available: https://github.com/nassosoassos".

[71] D. A. S. Salas, J. L. C. Ruiz and M. G. Mendoza, "Wireless channel model with markov chains using matlab". 2012.

[72] J. Santen et al., *Progress in Speech Synthesis*. Springer-Verlag New York Inc., 1997.

[73] M. Schroeder, "A brief history of synthestic speech", *Speech Communication*, vol. 13, pp. 231–237, 1993.

[74] K. Shinoda and T. Watanabe, "Acoustic modelling based on the mdl principle for speech recognition", in *European Conf. on Speech Communication and Technology (EUROSPEECH-1997)*, Rhodes, Greece, 1997.

[75] S.S. Stevens, J. Volkmann and E.B. Newman, "A scale for the measurement of the psychological magnitude pitch", *J. of the Acoustical Society of America*, vol. 8, pp. 185–190, 1937.

[76] M. Tamura et al., "Adaptation of pitch and spectrum for hmm-based speech synthesis using mllr", in *Proc. of Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2001)*, Salt Lake City, UT, 2001.

[77] T. Toda, "Trajectory training considering global variance for hmm-based speech synthesis", in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2009)*, Taipei, Taiwan, 2009.

[78] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis", *IEICE - Trans. on Informations and Systems*, pp. 816–824, 2007.

[79] K. Tokuda et al., "Spectral estimation of speech by mel-generalized cepstral analysis", *Trans. IEICE*, pp. 1124–1134, 1993.

[80] K. Tokuda et al., "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation", in *Proc. of Intl. Conf. on Spoken Language Processing (ICLSP-1994)*, Yokohama, Japan, 1994.

[81] K. Tokuda, T. Toda and J. Yamagishi, "Speech synthesis based on hidden markov models", *Proc. of the IEEE*, vol. 101, pp. 1234–1252, 2013.

[82] K. Tokuda, H. Zen and A.W. Black, "An hmm-based speech synthesis system applied to english", in *Proc. of IEEE Workshop on Speech Synthesis (SWW-2002)*, Santa Monica, CA, 2002.

[83] L. Wang et al., "Photo-real lips synthesis with trajectory-guided sample selection.", in *Proc. of Speech Synthesis Workshop (SSW-2007)*, Bonn, Germany, 2010.

[84] P. C. Woodland, "Speaker adaptation for continuous density hmms: A review", in *Proc.of the ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition (ISCA-2001)*, Antipolis, France, 2001.

[85] P. Wynn, "On the convergence and stability of the epsilon algorithm", *SIAM J. on Numerical Analysis*, vol. 3, pp. 91–122, 1966.

[86] H. Zen and T. Toda, "An overview of nitech hmm-based speech synthesis system for blizzard challenge 2005", in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH-2005)*, Lisbon, Portugal, 2005.

[87] Χ. Α. Μιναρετζής, "Σύνθεση Φωνής από Κείμενο με Βάση Κρυφά Μαρκοβιανά Μοντέλα", Master's thesis, Dept. Elect. and Comput. Eng., Nat. and Tech. Univ., Athens, 2011.

[88] Α. Χαλαμανδάρης, *Σύγχρονες τεχνικές σχεδίασης και υλοποίησης συστήματος παραγωγής συνθετικής ομιλίας με επεξεργασία στο πεδίο του χρόνου*. PhD thesis, Dept. Elect. and Comput. Eng., Nat. and Tech. Univ., Athens, 2011.