ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Αναγνώριση Φωνής με χρήση Γενικευμένων Φασματικών Ροπών

Robust Speech Recognition using Generalized Spectral Moments

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κουρπά Σταμάτιου

Επιβλέπων καθηγητής: Ποταμιάνος Αλέξανδρος
Εξεταστική επιτροπή: Διγαλάκης Βασίλειος
Ζερβάκης Μιχάλης

Μάρτιος 2013

Περιεχόμενα

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα, πρώτα απ΄ όλους, να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κύριο Ποταμιάνο Αλέξανδρο, για την ανάθεση αυτής της διπλωματικής εργασίας και για την παρακολούθηση και βοήθειά του, καθ΄ όλη τη διάρκειά της. Επίσης, τον κύριο Μπούρο Σωτήρη, για την τεχνική υποστήριξη που μου παρείχε, η οποία ήταν απαραίτητη για την υλοποίηση της παρούσας εργασίας. Τέλος, τους μεταπτυχιακούς φοιτητές Ιωσήφ Ηλία, Κουλουμέντα Βασιλική, Μαλανδράκη Νικόλαο και Μοσχόπουλο Θεοδόση, για την βοήθεια τους σε θεωρητικά αλλά και τεχνικά ζητήματα της διπλωματικής εργασίας.

ΠΕΡΙΛΗΨΗ

Η βελτιστοποίηση της μετάδοσης σήματος φωνής απασχολεί μεγάλο μέρος επιστημόνων και ερευνητών που δραστηριοποιούνται στον τομέα Τηλεπικοινωνιών. Και αυτό διότι μια τέτοια βελτιστοποίηση είναι χρήσιμη για τη βελτίωση πολλών εφαρμογών μετάδοσης σήματος φωνής, όπως η αναγνώριση ομιλίας και η σύνθεση φωνής από κείμενο. Μέχρι στιγμής, έχουν προταθεί διάφορες μέθοδοι γι΄ αυτή τη βελτιστοποίηση. Κάποιες από αυτές αφορούν τη χρήση front-end. Στόχος των front-ends είναι η παραμετροποίηση των ακουστικών χαρακτηριστικών ενός σήματος φωνής, σε διάνυσμα, που το κάθε στοιχείο του αποτελεί ένα χαρακτηριστικό του σήματος. Παραδείγματα τέτοιων χαρακτηριστικών είναι τα MFCC, τα FMP, τα SMAC, αλλά και ο συνδυασμός τους. Στην παρούσα διπλωματική εργασία ερευνούμε το ρόλο που παίζουν διάφορες παράμετροι, που λαμβάνουν χώρα κατά την κατασκευή των front-end χαρακτηριστικών, στην απόδοσή τους, με γνώμονα την ακρίβεια μετάδοσης, όπως αυτή προκύπτει από την πειραματική πλατφόρμα HTK. Στόχος μας είναι η εύρεση κατάλληλων τιμών σε αυτές τις παραμέτρους, προκειμένου να επιτευχθεί η βέλτιστη δυνατή απόδοση, για τις περιπτώσεις που χρησιμοποιούμε τους συνδυασμούς MFCC-FMP και SMAC-FMP. Τα συμπεράσματα που προκύπτουν μπορούν να βοηθήσουν τον κάθε ενδιαφερόμενο στην κατανόηση του ρόλου διαφόρων παραμέτρων στην τελική απόδοση του συστήματος, άρα και στην βαθύτερη κατανόηση της διαδικασίας αναγνώρισης σήματος φωνής. Επίσης, μπορούν να συνεισφέρουν στην περαιτέρω εξέλιξη της έρευνας βελτιστοποίησης της μετάδοσης σήματος φωνής με χρήση κατάλληλων front-end χαρακτηριστικών.

ABSTRACT

The optimization of the speech signal transmission occupies a large number of scientists and researchers who work in the Telecommunication field. This happens due to the fact that such an optimization is useful for many speech signal application's improvement, such as speech recognition and voice synthesis from text. Until now, several methods for this optimization have been suggested. Some of them regard the use of front-end. Front-ends purpose is the parameterization of the speech signal features in a vector, whose each element consists a signal feature. MFCC, FMP, SMAC and there combinations are such feature examples. On this thesis we investigate the role that several parameters, that take place in the front-end features production, play on the front-end features attribution, observing the transmission accuracy taken by the experimental platform HTK. Our purpose is to find the appropriate values for those parameters, so as to achieve the best possible accuracy, for the cases that we use the MFCC-FMP and SMAC-FMP combination. The conclusions that we take can help every interested person to understand the role that some parameters play in the final system accuracy, and as a result, in the deeper undestanding of the speech recognition process. Furthermore, they can contribute to the further optimization of the speech signal transmission research using appropriate front-end features.

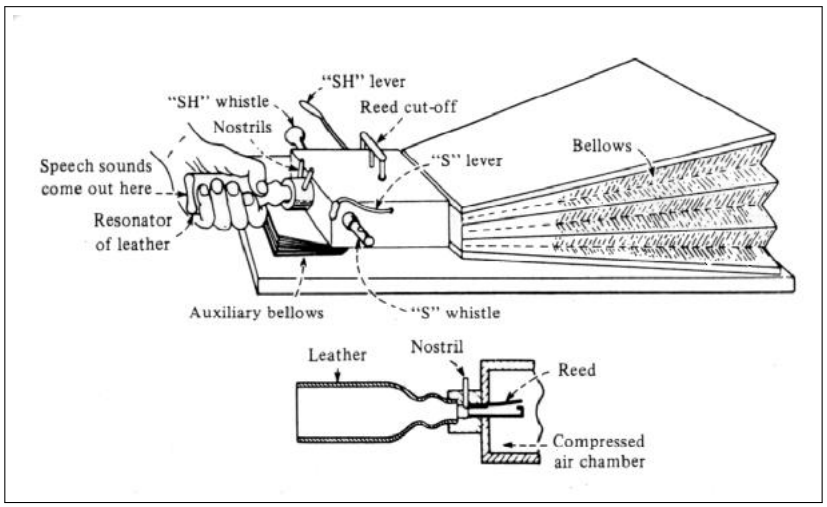
Κεφάλαιο 1

ΕΙΣΑΓΩΓΗ

Λόγω του ότι η ερευνητική περιοχή της συγκεκριμένης διπλωματικής εργασίας έγκειται στη σύνθεση, αναγνώριση και βελτιστοποίηση της μετάδοσης σήματος φωνής, κρίνουμε σκόπιμο να εξηγήσουμε, συνοπτικά, διάφορα εννοιολογικά και ιστορικά ζητήματα που αφορούν στη σύνθεση και αναγνώριση του σήματος φωνής.

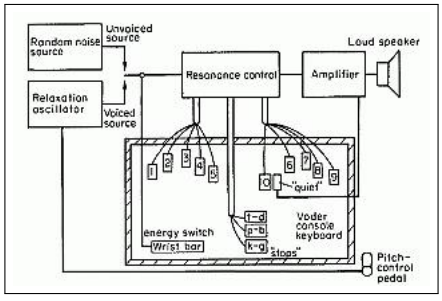
1.1 Σύνθεση φωνής

Με τον όρο "σύνθεση φωνής" εννοούμε τη διαδικασία παραγωγής φωνής με τεχνητό τρόπο, μια διαδικασία που ξεκίνησε από το 2ο μισό του 18ου αιώνα, αρχικά όχι με σκοπό την αναγνώριση ή την κατανόηση του ανθρώπινου λόγου, αλλά την αυτό καθ' αυτό δημιουργία ενός ομιλώντος μηχανήματος, ίσως λόγο του ότι ήδη ήταν γνωστή η χρήση σωλήνων συντονισμού (resonance tubes) για την προσέγγιση της ανθρώπινης φωνητικής οδού. Το 1773, ο Ρώσος επιστήμονας Christian Kratzenstein, καθηγητής φυσιολογίας του Πανεπιστημίου της Κοπεγχάγης, κατάφερε να παράγει ήχους φωνηέντων, με χρήση σωλήνων συντονισμού συνδεδεμένων με οργανικούς σωλήνες (organ pipes). Το 1791 στη Βιέννη, ο Wolfgang von Kempelen κατασκεύασε την περίφημη «Ακουστική - Μηχανική Μηχανή Φωνής". Μια μηχανή που έδινε τη δυνατότητα παραγωγής φωνηέντων και συμφώνων, με τη χρήση μοντέλων της γλώσσας και των χειλιών. Στα μέσα του 1800, ο Charles Wheatstone κατασκεύασε μια έκδοση της μηχανής του Wolfgang von Kempelen,χρησιμοποιώντας αντηχεία από δέρμα, η διαμόρφωση και το σχήμα των οποίων μπορούσαν να τροποποιηθούν ή να ελέγχονται με το χέρι, για να παράγουν διαφορετικούς ήχους που έμοιαζαν με ομιλίες (speech - like sounds), όπως φαίνεται στην παρακάτω εικόνα [18].



Σχήμα 1.1: Η εκδοση του Wheatstone,πάνω στην ομιλούσα μηχανή του Wolfgang von Kempelen, πηγή:[18].

Ακολούθησαν και άλλες μηχανικές κατασκευές τέτοιου τύπου, έως ότου, στο πρώτο μισό του 20ου αιώνα, ο Fletcher και διάφοροι άλλοι επιστήμονες των Εργαστηρίων της Bell (Bell Laboratories) τεκμηρίωσαν τη σχέση που έχει το φάσμα της ομιλίας (speech spectrum) με τα χαρακτηριστικά του ήχου και την ευκρίνεια με την οποία ο ήχος γίνεται αντιληπτός από τον άνθρωπο. Στη δεκαετία του 1930, ο Homer Dudley των Bell Labs , επηρρεασμένος από την έρευνα του Fletcher, κατασκεύασε την πρώτη ηλεκτρονική συσκευή σύνθεσης φωνής, η οποία ονομάστηκε VODER (Voice Operating Demonstrator), και ήταν ένα ηλεκτρονικό ισοδύναμο της μηχανής ομιλίας του Wheatstone, που συνέθετε φωνή με κατάλληλο χειρισμό πλήκτρων. Στις αρχές της δεκαετίας του 1960, πάλι στα Bell Labs, έγινε η πρώτη σύνθεση φωνής με υπολογιστή. Αυτή ήταν και η απαρχή ενός νέου επιστημονικού τομέα, η οποία οδήγησε στην ανάπτυξη των πρώτων ολοκληρωμένων συστημάτων σύνθεσης φωνής από κείμενο [18].



Σχήμα 1.2: Διάγραμμα του VODER, πηγή:[18].

1.2 Αναγνώριση φωνής – Automatic Speech Recognition

Με τον όρο "Αναγνώριση Φωνής (Automatic Speech Recognition - ASR)" εννοούμε την αναγνώριση της ανθρώπινης ομιλίας από τους υπολογιστές. Η διαδικασία της αυτόματης αναγνώρισης φωνής παράγει μία ακολουθία λέξεων από ένα ακουστικό σήμα. Επιπλέον εξάγει το νόημα από την φράση που έχει αναγνωρισθεί, ώστε το σύστημα να μπορεί να απαντήσει στον ομιλητή ή να πραγματοποιήσει μία ενέργεια (π.χ. αναζήτηση σε βάση δεδομένων).

1.3 Εφαρμογές της αναγνώρισης φωνής

Η διαδικασία αναγνώρισης φωνής παρέχει μια σειρά εφαρμογές, για γενικές ή εξειδικευμένες χρήσεις. Παραθέτουμε μερικές από αυτές [32]:

• Προσπέλαση πληροφοριών με ομιλία και όχι με ιεραρχικά menu: Αυτό μπορεί να έχει καθημερινή εφαρμογή στη χρήση ηλεκτρονικών υπολογιστών, όπου από τη μια η προσπέλαση θα γίνεται ταχύτερα και από την άλλη θα εξυπηρετήσει άτομα με αναπηρίες στα μάτια ή τα χέρια ή άτομα που χρειάζεται να έχουν δεσμευμένα τα μάτια ή τα χέρια τους κατά τη διάρκεια της εργασίας τους. Στην αγορά υπάρχουν ήδη προγράμματα αναγνώρισης κειμένου, από την Dragon,την IBM και την Microsoft.

• Τηλεφωνικές εφαρμογές: Η λειτουργία του τηλεφώνου με χρήση της ανθρώπινης φωνής και όχι του πληκτρολογίου, αποτελεί ένα ιδιαίτερα χρηστικό επίτευγμα, τόσο για ταχύτερη κλήση, όσο και για διευκόλυνση ατόμων με αναπηρία ή ηλικιωμένων. Υπάρχουν πολλές τηλεφωνικές εφαρμογές από διάφορες εταιρίες, κυρίως για μικρά λεξιλόγια, για ακολουθίες ψηφίων και για μεμονωμένες λέξεις. Τα επόμενα χρόνια αναμένεται μια έκρηξη εφαρμογών αναγνώρισης ομιλίας μέσω του τηλεφώνου.

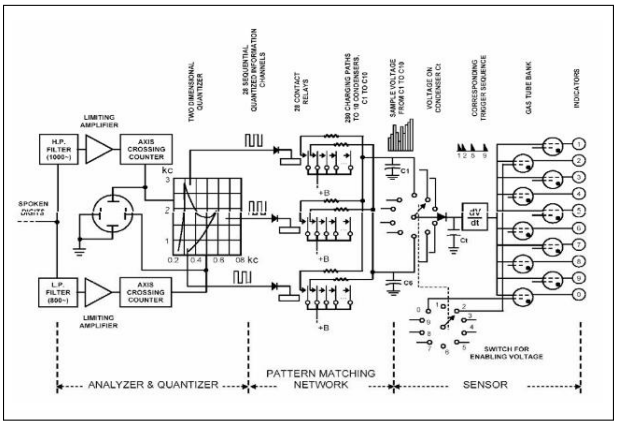
• Υγειονομική περίθαλψη: Η αναγνώριση φωνής μπορεί να εφαρμοστεί στα front-end και back-end της ιατρικής διαδικασία τεκμηρίωσης. Ακόμα, η δημιουργία και η αποτελεσματικότητα των Ηλεκτρονικών Ιατρικών Φακέλων (EMR) βελτιώνεται, αν δημιουργηθούν σε συνδυασμό με μια μηχανή αναγνώρισης φωνής. Π.χ οι αναζητήσεις ή οι συμπληρώσεις της ιατρικής φόρμας μπορούν να εκτελεστούν πιο γρήγορα μέσω της ομιλίας, παρά μέσω του πληκρολογίου.

• Πολεμικά αεροσκάφη: Την τελευταία δεκαετία έχουν λάβει χώρα προσπάθειες για τη δοκιμή και την αξιολόγηση της αναγνώρισης ομιλίας στα μαχητικά αεροσκάφη. Παραδείγματα τέτοιων προγραμμάτων είναι: 1. Το πρόγραμμα των ΗΠΑ για Advanced Fighter Technology Integration (AFTI)/F-16 aircraft (F-16 VISTA). 2. Το πρόγραμμα της Γαλλίας για τα αεροσκάφη Mirage. 3. Το πρόγραμμα της Μεγάλης Βρετανίας, σε μια σειρά τύπων αεροσκαφών.

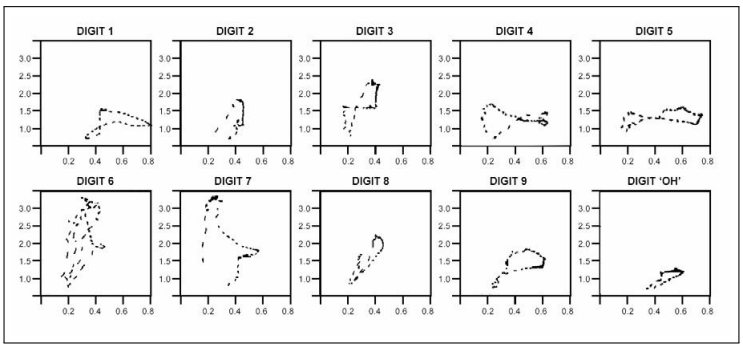
• Ελικόπτερα: Η ανάγκη επίτευξης υψηλής ακρίβειας αναγνώρισης φωνής σε καταστάσεις πίεσης και θορύβου, όπως αυτή του ελικοπτερου, είναι μεγάλη. Στην περίπτωση του ελικοπτέρου είναι ακόμα μεγαλύτερη, επειδή ο πιλότος του ελικοπτέρου συνήθως δεν φοράει μάσκα προσώπου, η οποία θα μειώσει τον ακουστικό θόρυβο στο μικρόφωνο. Μέχρι στιγμής έχουν υλοποιηθεί ή δρομολογηθεί μια σειρά προγράμματα που αφορούν στη βελτίωση της επικονωνίας μέσω ραδιοφώνου, των συστημάτων πλοήγησης, τον έλεγχο ενός αυτοματοποιημένου συστήματος στόχου παράδοσης. Τέτοιου είδους προγράμματα έχουν εφαρμοστεί από μια σειρά χώρες όπως οι ΗΠΑ, η Μεγάλη Βρετανία, η Γαλλία, ο Καναδάς.

1.4 Ιστορική αναδρομή της διαδικασίας αναγνώρισης φωνής

Από τις αρχές του 1950, διάφορες ερευνητικές ομάδες, από διάφορες χώρες, ασχολούνταν με την έρευνα της αυτόματης αναγνώρισης φωνής, εξετάζοντας βασικές αρχές της ακουστικής και της φωνητικής. Το 1952, οι Davis, Biddulph και Balashek, κατασκεύασαν στα εργαστήρια της Bell ένα σύστημα αναγνώρισης μεμονωμένων ψηφίων για έναν ομιλητή, χρησιμοποιώντας τα formant frequencies που υπολογίζονταν (ή εκτιμώνταν) για τα φωνήεντα του καθε ψηφίου [18]. Η εικόνα 1.3 παρουσιάζει το διάγραμμα (block diagram) του παραπάνω συστήματος αναγνώρισης μεμονωμένων ψηφίων, ενώ η εικόνα 1.4 παρουσιάζει τις γραφικές παραστάσεις των τροχιών των formants κατά μήκος των διαστάσεων της πρώτης και της δεύτερης συχνότητας formant (formant frequency) για καθένα από τα δέκα ψηφία, ένα-εννέα και 0, αντίστοιχα. Αυτές οι τροχιές μας βοηθούν να βρούμε την ταυτότητα του άγνωστου ψηφίου, ως το πιο "ταιριαστό" με τη συγκεκριμένη τροχιά.



Σχήμα 1.3: Διάγραμμα αναγνώρισης μεμονωμένων ψηφίων, πηγή:[18].



Σχήμα 1.4: Εικόνες του formant 1 και formant 2 για κάθε ψηφίο, πηγή:[18].

Στη δεκαετία του 1960, αρκετά Ιαπωνικά εργαστήρια ασχολήθηκαν με την κα-

τασκευή hardware ικανού για την εκτέλεση διαδικασίας αναγνώρισης φωνής. Τα πιο αξιοσημείωτα ήταν αυτά της αναγνώρισης φωνηέντων της Suzuki και Nataka στο Radio Research Laboratory του Τόκιο, της αναγνώρισης φωνημάτων των Sakai και Doshita στο Πανεπιστήμιο του Κιότο, και της αναγνώρισης ψηφίων στα εργαστήρια NEC. Παράλληλα, ερευνητικές προσπάθειες έγιναν και στο University College της Αγγλίας, με τη διαδικασία αναγνώρισης φωνημάτων, ώστε να αναγνωριστούν τέσσερα φωνήεντα και εννιά σύμφωνα, όπως επίσης και στο MIT Lincoln Lab για αναγνώριση 19 φωνηέντων.

Η έρευνα στο συγκεκριμένο πεδίο εντάθηκε τη δεκαετία του 1960 και του 1970. Προστέθηκαν και άλλες ερευνητικές ομάδες στις ήδη υπάρχουσες, από τις ΗΠΑ, την Ιαπωνία και την ΕΣΣΔ. Ως σημαντικά επιτεύγματα, ενδεικτικά μπορούμε να αναφέρουμε τη διατύπωση της Γραμμικής Πρόβλεψης Κωδικοποίησης (Linear Predictive Coding – LPC) από τον Itakura και τη διατύπωση της Δυναμικής Χρονικής Στρέβλωσης (Dynamic Time Warping – DTW), στην ουσία ενός γραμμικού προγραμματισμού για τη στοίχιση δύο προτάσεων. Τη δεκαετία του 1980, έγινε μια μεγάλη τομή στη μεθοδολογία της αναγνώρισης φωνής. Πέρασε από έναν κάπως διαισθητικό τρόπο προσέγγισης σε μια αυστηρή στατιστική μοντελοποίηση, μέσω της χρήσης Κρυφών Μαρκοβιανών Μοντέλων (Hidden Marcov Models - HMM). Βέβαια η έννοια του Κρυφού Μαρκοβιανού Μοντέλου ήταν γνωστή από πιο πριν σε μερικά εργαστήρια (πχ στο IBM και στο Ινστιτούτο Αμυντικών Αναλύσεων IDA). Όμως, η μεθοδολογία του μοντέλου αυτού δεν ήταν πλήρης μέχρι τα μέσα της δεκαετίας του 1980, οπότε και έγινε η προτιμότερη μέθοδος για την αναγνώριση φωνής. Υπήρξε και προσπάθεια αξιοποίησης των Νευρωνικών Δικτύων, όμως τα HMM υιοθετήθηκαν σχεδόν από όλα τα εργαστήρια και επικράτησαν σχετικά γρήγορα [18].

Κεφάλαιο 2

ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Εισαγωγή

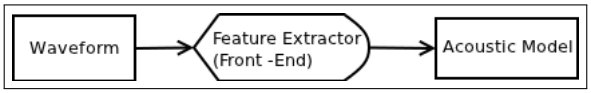
Σε αυτό το κεφάλαιο παραθέτουμε τις απαραίτητες γνώσεις που πρέπει να διαθέτει ο αναγνώστης αυτού του κειμένου, για την παρακολούθηση και την κατανόηση αυτής της διπλωματικής εργασίας. Οι γνώσεις αυτές καθορίζονται από τους άξονες που κινείται η εργασία. Ξεκινάμε, εξηγώντας τη σημασία της εξαγωγής ακουστικών χαρακτηριστικών και το ρόλο του εξαγωγέα τέτοιων χαρακτηριστικών, συνεχίζουμε περιγράφοντας διάφορες μεθόδους παραμετρικής απεικόνισης του σήματος φωνής, και τέλος παρουσιάζουμε τον τρόπο εκτίμησης της ακολουθίας συμβόλων του αρχικού σήματος.

2.2 Εξαγωγή ακουστικών χαρακτηριστικών

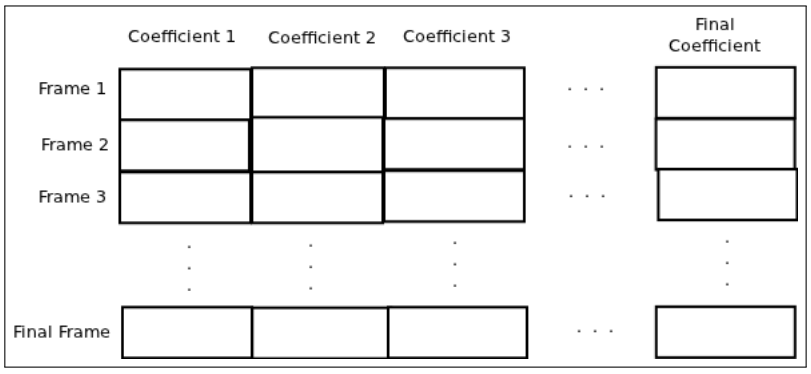
Η εξαγωγή κατάλληλων πληροφοριών – ακουστικών χαρακτηριστικών- από το σήμα φωνής, είναι ένα ερευνητικό αντικείμενο που απασχολεί την επιστημονική κοινότητα, που δραστηριοποιείται στο πεδίο της αναγνώρισης φωνής, εδω και δεκαετίες [8]. Και αυτό διότι ο εξαγωγέας ακουστικών χαρακτηριστικών (front-end) είναι το πρώτο βήμα στο σύστημα αυτόματης αναγνώρισης φωνής (Automatic Speech Recognition) και μετατρέπει ένα ακατέργαστο σήμα σε μια συμπαγή αναπαράσταση, δηλαδή μετατρέπει το σήμα φωνής σε ένα διάνυσμα μικρών διαστάσεων, όπου κάθε στοιχείο του διανύσματος αντιπροσωπεύει ένα χαρακτηριστικό του αρχικού σήματος. Η απόδοση αυτής της φάσης είναι σημαντική για τις επόμενες φάσεις (της αναγνώρισης φωνής), εφόσον επηρρεάζει τη συμπεριφορά και απόδοσή τους [12, 13].
Ιδανικά, τα ακουστικά χαρακτηριστικά πρέπει να ικανοποιούν τις παρακάτω προδιαγραφές:
α) Να μεταφέρουν όλη τη φωνητική πληροφορία του σήματος φωνής.
β) Να είναι ανεξάρτητα από το γένος και την ηλικία του ομιλητή.
γ) Να επηρρεάζονται όσο το δυνατό λιγότερο από το θόρυβο.
δ) Να μοντελοποιούνται εύκολα.

2.3 Front-End

Ο ρόλος του Front -End είναι να μετατρέψει ένα σήμα φωνής (με τη μορφή κυματομορφής) σε ένα διάνυσμα, του οποίου το κάθε στοιχείο αντιπροσωπεύει ένα χαρακτηριστικό του αρχικού σήματος, δηλαδή να το μετατρέψει σε ένα ακουστικό μοντέλο (acoustic model). Το κάθε στοιχείο του διανύσματος περιέχει τα κατάλληλα ακουστικά χαρακτηριστικά για μια σειρά τμήματα του διανύσματος, που ονομάζονται frames, και στα οποία το σήμα φωνής έχει προηγουμένως διαχωριστεί [20].



Σχήμα 2.1: Front - End.



Σχήμα 2.2: Ακουστικό μοντέλο.

Υπάρχουν πολλές μέθοδοι παραμετρικής απεικόνισης ενός σήματος φωνής. Στη συνέχεια θα περιγράψουμε τρεις από αυτές τις μεθόδους, για την κατασκευή MFCC, FMP και SMAC, που είναι και οι μέθοδες που θα μας απασχολήσουν στην παρούσα διπλωματική εργασία.

2.4 MFCC

Οι συντελεστές MFCC είναι από τα πιο διαδεδομένα ακουστικά χαρακτηριστικά για αναπαράσταση φασματικών χαρακτηριστικών σήματος φωνής. Στηρίζονται πάνω στην αποδεδειγμένη διαφοροποίηση που υπάρχει στην αντιληπτική ικανότητα του ανθρώπινου αυτιού στο εύρος ζώνης του φάσματος συχνοτήτων ενός σήματος φωνής. Έχει αποδειχθεί ότι οι μεγαλύτερες συχνότητες του φάσματος συχνοτήτων ενός σήματος φωνής (speech spectrum) περιέχουν λιγότερο διακριτή πληροφορία φωνημάτων από ότι τα σημεία με χαμηλή ή μέτρια συχνότητα (χαμηλότερη από 3kHz). Η κλίμακα Mel ικανοποιεί την παραπάνω προδιαγραφή και γι΄ αυτό χρησιμοποιείται για την εξαγωγή των MFCC χαρακτηριστικών [19]. Ορίζει γραμμικές

αποστάσεις κάτω από το 1 kHz και λογαριθμικές πάνω από τα 1kHz. Έτσι, εφαρμόζουμε τριγωνικά φίλτρα, των οποίων οι κεντρικές συχνότητες ισαπέχουν στις χαμηλές συχνότητες και είναι λογαριθμικά τοποθετημένες στις υψηλές. Τα φίλτρα αυτά, στην ουσία κάνουν δειγματοληψία της ενέργειας του σήματος σε διαφορετικές ζώνες. Στη συνέχεια θα εξηγήσουμε τα στάδια κατασκευής των MFCCs.

2.4.1 Προέμφαση – Pre-emphasis

Στο στάδιο αυτό περνάμε το σήμα από ένα high-pass φίλτρο, το οποίο δίνει έμφαση στις υψηλές συχνότητες. Με αυτό τον τρόπο αυξάνουμε την ενέργεια του σήματος στις υψηλότερες συχνότητες και αντισταθμίζουμε το κομμάτι με τις υψηλές συχνότητες, που ήταν υποβαθμισμένο στο αρχικό σήμα. Επίσης, ενισχύουμε τη σημασία των formants υψηλών συχνοτήτων [12, 21].

$$x^{'}[n] = x[n] - ax[n-1] \tag{2.1}$$

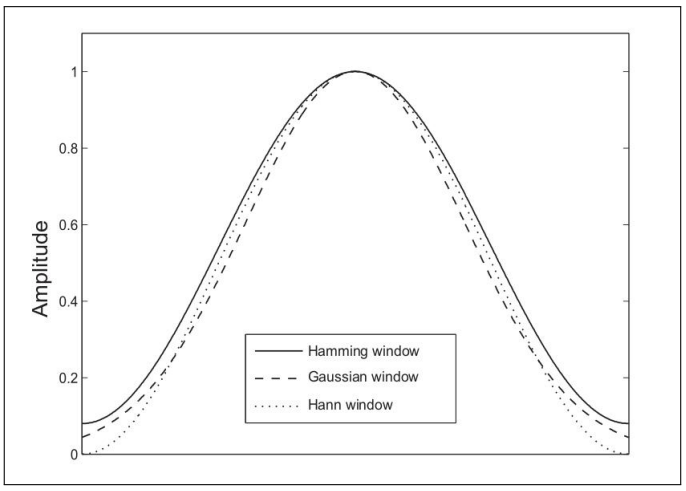
Όπου α είναι συνήθως μεταξύ 0.9 και 1 [12, 21]. Π.χ. αν το $\alpha = 0.95$, τότε το 95% του κάθε δείγματος θα προέρχεται από το προηγούμενο δείγμα [12].

2.4.2 Πλαισίωση – Framing

Γενικά το σήμα φωνής δεν είναι ένα στάσιμο σήμα, αλλά αν το χωρίσουμε σε μικρά κομμάτια, μπορεί να θεωρηθεί ως τέτοιο. Αυτό συμβαίνει σε αυτό το στάδιο. Με λίγα λόγια, η συνεχής ροή δειγμάτων ηχητικού σήματος διαιρείται σε πλαίσια (frames) των N δειγμάτων, όπου το N κυμαίνεται από 20-40 msec. Τα διαδοχικά πλαίσια απέχουν μεταξύ τους M δείγματα, όπου $M < N$. Επομένως, τα πλαίσια είναι επικαλυπτόμενα και η συνήθης επικάλυψη είναι 50% (M=N/2). Είναι σημαντικό να υπάρχει επικάλυψη επειδή με αυτό τον τρόπο δεν θα έχουμε χάσιμο πληροφορίας στα άκρα των πλαισίων [12, 23]. Για παράδειγμα, αν χρησιμοποιήσουμε N=200 και M=100 (επικάλυψη 50%) και το ηχητικό σήμα δειγματοληπτηθεί με συχνότητα δειγματοληψίας Fs = 8kHz, δηλαδή με περίοδο δειγματοληψίας Ts = 1/Fs = 0.125 msec, τότε το σήμα διαιρείται σε πλαίσια διάρκειας N × Ts = 25 msec τα οποία απέχουν κατά M × Ts = 12.5 msec και εξάγεται ένα πλαίσιο κάθε 12.5 msec.

2.4.3 Παραθύρωση - Windowing

Σε αυτό το βήμα, θέλουμε να ελαχιστοποιήσουμε τις ασυνέχειες, δηλαδή την πιθανή φασματική παραμόρφωση στην αρχή και στο τέλος του πλαισίου [25]. Για να γίνει αυτό θα επιβάλλουμε μια "μείωση" του σήματος στην αρχή και στο τέλος του πλαισίου. Αυτή η "μείωση" επιτυγχάνεται με την εφαρμογή ενός παραθύρου, της κατηγορίας των raised cosine παραθύρων, σε κάθε πλαίσιο. Τριών ειδών τέτοια παράθυρα φαίνονται στο παρακάτω σχήμα.



Σχήμα 2.3: Σύγκριση παραθύρων, πηγή: [23]

Από τα διάφορα τέτοιου τύπου παράθυρα, εκείνο που χρησιμοποιείται στην αναγνώριση φωνής είναι το παράθυρο Hamming, που δίνεται από την παρακάτω εξίσωση:

$$w[n] = 0.54 - 0.46cos\frac{2\pi n}{N-1} \tag{2.2}$$

όπου 0⩽ N ⩽ N-1, και N είναι ο αριθμός των δειγμάτων σε κάθε πλαίσιο.
Το αποτέλεσμα της παραθυροποίησης είναι

$$Y[n] = X[n]W[n] \tag{2.3}$$

όπου X[n] είναι το σήμα εισόδου, δηλαδή το κατάλληλο πλαίσιο, W[n] είναι το παράθυρο Hamming και Y[n] είναι το πλαίσιο μετά την παραθυροποίηση.

2.4.4 Μετασχηματισμός Fourier - FFT

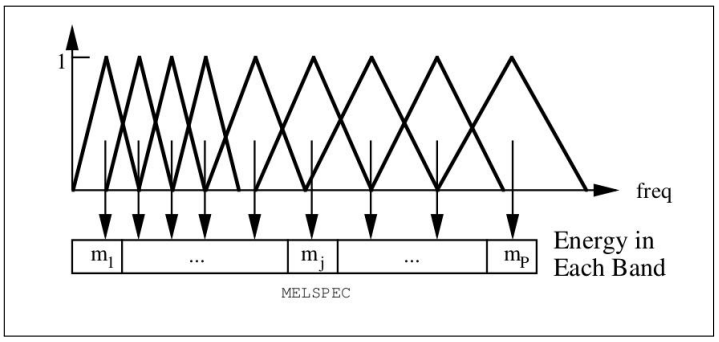
Η ανάλυση των οργάνων ανθρώπινης ακοής έχει αποδείξει ότι το κυματικό σήμα φωνής χωρίζεται σύμφωνα με τις συχνότητες. Επομένως, πρέπει να μετατρέψουμε κάθε ένα από τα παραπάνω πλαίσια από το πεδίο του χρόνου στο πεδίο της συχνότητας και να πάρουμε το φάσμα του κάθε πλαισίου. Αυτό επιτυγχάνεται παίρνοντας το μέτρο του μετασχηματισμού Fourier [23], ο οποίος δίνεται από την ακόλουθη σχέση

$$Y'[k] = \sum_{n=0}^{N-1} Y[n] \exp\frac{-j2\pi kn}{N} \tag{2.4}$$

Όπου n και N είναι ο δείκτης και ο συνολικός αριθμός των στοιχείων του πλαισίου που θα μετασχηματιστεί, αντίστοιχα, k ο δείκτης για κάθε μετασχηματισμένο στοιχείο, $x[n]$ είναι το κάθε στοιχείο του πλαισίου που θα μετασχηματιστεί και $Y[k]$ το κάθε μετασχηματισμένο στοιχείο.

2.4.5 Συστοιχία mel φίλτρων

Όπως αναφέρθηκε στο προηγούμενο υποκεφάλαιο, το ανθρώπινο αυτί χωρίζει τους ήχους ανάλογα με τις συχνότητες. Όμως η ανάλυση της κλίμακας των συχνοτήτων δεν είναι γραμμική. Αυτή η συμπεριφορά προσομοιώνεται με τη χρήση συστοιχιών φίλτρων (filter banks). Υπάρχουν αρκετά είδη συστοιχιών φίλτρων. Από αυτές εκείνη που χρησιμοποιείται στην εύρεση των MFCC είναι η συστοιχία φίλτρων mel. Στην ουσία κατασκευάζουμε μια διάταξη τριγωνικών ζωνοπερατών φίλτρων, που η κεντρική συχνότητα του κάθε φίλτρου αντιστοιχεί στην αντίστοιχη συχνότητα της κλίμακας mel, η μέγιστη τιμή του κάθε φίλτρου είναι η μονάδα και τα φίλτρα έχουν επικάλυψη 50%. Για κάθε πλαίσιο, φιλτράρουμε το φάσμα που προέκυψε από το παραπάνω υποκεφάλαιο, με καθένα από αυτά τα φίλτρα και αρθροίζουμε. Ως αποτέλεσμα, για κάθε φίλτρο προκύπτει ένα νούμερο [23]. Επομένως, μετά το φιλτράρισμα, το κάθε πλαίσιο θα αποτελείται από τόσους συντελεστές, όσα και τα φίλτρα. Το παρακάτω σχήμα δείχνει αυτή την συστοιχία τριγωνικών φίλτρων.



Σχήμα 2.4: Συστοιχία φίλτρων με κλίμακα mel, πηγή: [25], σελ.60

Στο κεφάλαιο με τίτλο ¨Κλίμακα mel¨, που βρίκεται στο Παράρτημα, εξηγούμε τους τύπους εύρεσης των συχνοτήτων με κλίμακα mel και παρέχουμε πληροφορίες για τη χρήση της συγκεκριμένης κλίμακας.

2.4.6 DCT

Σε αυτό το βήμα, μετατρέπουμε το λογάριθμο του φάσματος mel για το κάθε πλαίσιο, από το πεδίο της συχνότητας στο πεδίο του χρόνου. Επειδή, τόσο οι συντελεστές του mel φάσματος, όσο και ο λογάριθμός τους είναι πραγματικοί αριθμοί, μπορούμε να τους μεταφέρουμε στο πεδίο του χρόνου χρησιμοποιώντας το Διακριτό Μετασχηματισμό Συνημιτόνου (Discrete Cosine Transform - DCT).

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j cos(\frac{\pi i}{N}(j - 0.5)) \tag{2.5}$$

Όπου N είναι ο αριθμός των φίλτρων mel, m είναι το λογαριθμισμένο διάνυσμα του κάθε πλαισίου, i είναι ο δείκτης του κάθε MFCC και c είναι το διάνυσμα των MFCC.

Το αποτέλεσμα της διαδικασίας αυτής είναι οι συντελεστές MFCC .

Σε αυτό το σημείο θα θέλαμε να αναφέρουμε τους λόγους για τους οποίους χρησιμοποιούμε λογάριθμο και μέθοδο DCT.
• Με τη χρήση του λογαρίθμου πετυχαίνουμε δύο θετικά αποτελέσματα. Πρώτον, ενισχύουμε την προσωμοίωση εκεί όπου ο ήχος δεν είναι γραμμικός. Δεύτερον, ακριβώς επειδή το αποτέλεσμα του λογαρίθμου είναι άθροισμα, αν υπάρχει θόρυβος στο αρχικό σήμα, αυτός μπορεί πιο εύκολα να αφαιρεθεί μετά από λογαρίθμιση [23].
• Με τη χρήση DCT για τη μετατροπή από το πεδίο της συχνότητας στο πεδίο του χρόνου, μειώνουμε τη συσχέτιση των MFCC χαρακτηριστικών. Αυτό μας δίνει τη δυνατότητα, σε μετέπειτα επεξεργασία των χαρακτηριστικών αυτών, να μπορούμε να χρησιμοποιήσουμε διαγώνιους πίνακες συνδιασποράς (diagonal covariance matrices), όπως πχ αν χρησιμοποιήσουμε Κρυφά Μαρκοβιανά Μοντέλα. Ενώ αν δεν είχαμε χρησιμοποιήσει τη μέθοδο DCT, αλλά άλλη μέθοδο, πιθανόν να είχαμε χαρακτηριστικά με μεγαλύτερη συσχέτιση και τότε θα έπρεπε να χρησιμοποιήσουμε πλήρη πίνακα συνδιασποράς (full covariance matrix), κάτι που θα αύξαναι το πληροφοριακό φορτίο (computational load) [23].

2.4.7 Συντελεστές Δέλτα και Συντελεστές Επιτάχυνσης

Το τελευταίο βήμα της κατασκευής των MFCC χαρακτηριστικών περιλαμβάνει την κατασκευή των συντελεστών Δέλτα και Επιτάχυνσης (Delta and Acceleration Coefficients). Αυτό συμβαίνει διότι το σήμα φωνής δεν μένει σταθερό στο χρόνο, αλλά αλλάζει. Με την προσθήκη των συντελεστών Δέλτα και Επιτάχυνσης, πετυχαίνουμε τη σημαντική βελτίωση της απόδοσης του συτήματος, διότι οι συντελεστές αυτοί σχετίζονται με τις αλλαγές το φάσμα των MFCC χαρακτηριστικών στην πάροδο του χρόνου [25, 27]. Με την προσθήκη των συντελεστών αυτών, η απόδοση του συστήματος αναγνώρισης φωνής ενισχύεται σημαντικά [25, 27]. Ύστερα από την κατασκευή των συντελεστών αυτών, ο αριθμός των MFCC χαρακτηριστικών έχει τριπλασιαστεί, λόγω του ότι αν έχουμε N συντελεστές MFCC, κατασκευάζουμε N συντελεστές Δέλτα και N συντελεστές Επιτάχυνσης. Στο παράρτημα περιγράφουμε τον τρόπο κατασκευής των συντελεστών Δέλτα και Επιτάχυνσης.

2.5 FMP

Έχει αποδειχθεί ότι σε ένα σήμα φωνής εμπεριέχεται η διαμόρφωση πλάτους (Amplitude Modulation-AM) και συχνότητας (Frequency Modulation-FM). Βάσει αυτής της απόδειξης, το σήμα φωνής μπορεί να παρασταθεί ως ένα AM–FM μοντέλο. Όμως, σε ένα τέτοιο μοντέλο έχουμε συχνότητες μη σταθερές κατά τη διάρκεια μιας περιόδου, αλλά κυμαινόμενες γύρω από μια κεντρική συχνότητα. Επίσης, ούτε τα πλάτη είναι σταθερά κατά τη διάρκεια μιας περιόδου. Αυτό το γεγονός δημιουργεί προβλήματα στην εξαγωγή χαρακτηριστικών [7]. Τα προβλήματα αυτά ξεπερνιώνται με την κατασκευή των FMP χαρακτηριστικών. Στη συνέχεια εξηγούμε τη μορφή και τη σύνθεση ενός AM–FM μοντέλου, όπως και τη διαδικασία κατασκευής των FMP χαρακτηριστικών και την αξία τους.

2.5.1 Το μοντέλο φωνής AM-FM

Το AM–FM μοντέλο φωνής είναι ένα μη-γραμμικό μοντέλο, που περιγράφει το σήμα φωνής ως ένα συνδυασμό διαμόρφωσης πλάτους (AM) και διαμόρφωσης συχνότητας (FM). Περιγράφεται από την ακόλουθη εξίσωση [5]:

$$r(t) = \alpha(t)cos[2\pi(f_c(t) + \int q(\tau)d\tau) + \theta] \tag{2.6}$$

όπου f_c είναι η κεντρική συχνότητα συντονισμού (center value of the formant frequency), $q(t)$ είναι το σήμα διαμόρφωσης συχνότητας (frequency modulating signal) και $a(t)$ είναι το σήμα διαμόρφωσης πλάτους (time-varying amplitude). Η στιγμιαία συχνότητα του σήματος (instantaneous frequency) ορίζεται ως $f(t) = f_c + q(t)$. Το σήμα φωνής $s(t)$ μοντελοποιείται ως το άθροισμα K τέτοιων $AM - FM$ σημάτων, ένα για κάθε συντονισμό (formant) [5].

$$s(t) = \sum_{k=1}^{K} r_k(t) \tag{2.7}$$

2.5.2 Πολυζωνική αποδιαμόρφωση

Όπως έχουμε ήδη αναφέρει, στο $AM - FM$ μοντέλο, οι συχνότητες συντονισμού (formant frequencies) δεν είναι σταθερές κατά τη διάρκεια μιας περιόδου, αλλά κυμαίνονται γύρω από μια κεντρική συχνότητα (center frequency). Για να εκτιμήσουμε, λοιπόν, τις συχνότητες συντονισμού και το εύρος ζώνης, θα βασιστούμε στην εκτίμηση της στιγμιαίας συχνότητας και του στιγμιαίου εύρους ζώνης του σήματος φωνής. Εφόσον το σήμα φωνής $s(t)$ μοντελοποιείται ως άθροισμα K $AM - FM$ σημάτων, ένα για κάθε συντονισμό (formant), μας ενδιαφέρει να βρούμε το στιγμιαίο πλάτος ($\alpha(t)$) και τη στιγμιαία συχνότητα ($f(t)$) του παραπάνω σήματος, για κάθε formant. Συνήθως, ο αριθμος των formants δεν ξεπερνά τα έξι, δηλαδή κ=1,...,6. Υπάρχουν διάφορες μέθοδοι τέτοιας αποδιαμόρφωσης. Η μέθοδους που θα χρησιμοποιήσουμε είναι ο αλγόριθμος διαχωρισμού ενέργειας (ESA-Energy Seperation Algorithm) πάνω στο σήμα $s(t)$. Ο αλγόριθμος αυτός βασίζεται στον διαχωρισμό ενέργειας και χρησιμοποιεί το μη-γραμμικό τελεστή ενέργειας Teager-Kaiser (TEO) [5]. Για να βρούμε, λοιπόν, τις τιμές των ($\alpha(t)$) και $f(t)$ για κάθε formant, φιλτράρουμε το σήμα $s(t)$ σε διαφορετικές φασματικές ζώνες (όσες και τα formants) και έπειτα κάνουμε τη διαδικασία του ESA για κάθε φασματική ζώνη. Επίσης, αποδιαμόρφωση μπορεί να γίνει και με χρήση μετασχηματισμού Hilbert και σήματος Gabor. Σε αυτή την περίπτωση, το στιγμιαίο πλάτος είναι το μέτρο του αναλυτικού σήματος και η στιγμαία συχνότητα είναι η παράγωγος της φάσης του. Η αποδιαμόρφωση με χρήση μετασχηματισμού Hilbert έχει παρόμοια αποτελέσματα με αυτήν με χρήση ESA, αλλά ο αλγόριθμος διαχωρισμού ενέργειας έχει μικρότερο υπολογιστικό κόστος. Αυτός είναι ένας παράγοντας που συντελεί στο να προτιμάται ο ESA έναντι του Hilbert [5]. Ο αλγόριθμος ESA και ο τελεστής ενέργειας TEO περιγράφονται στο Παράρτημα.

Έχοντας υπολογίσει τις τιμές της στιγμιαίας συχνότητας και του στιγμιαίου εύρους ζώνης για κάθε formant, θα προχωρήσουμε στον υπολογισμό της συχνότητας συντονισμού και του εύρους ζώνης για κάθε formant. Υπάρχουν δύο μέθοδοι εκτίμησης της συχνότητας και του εύρους ζώνης του κάθε formant. Ο πρώτος χρησιμοποιεί στάθμιση (weighting) και ο δεύτερος δεν τη χρησιμοποιεί. Στο επόμενο υποκεφάλαιο εξηγούμε τις δύο αυτές μεθόδους.

2.5.3 Εκτίμηση των συχνοτήτων συντονισμού και του εύρους ζώνης του κάθε formant

Η εκτίμηση της συχνότητας, αν δεν χρησιμοποιούμε στάθμιση (unweighted mean frequency), ορίζεται ως [5]:

$$F_u = \frac{1}{T}\int_{t_0}^{t_0+T} f(t)dt \tag{2.8}$$

όπου t_0 και T είναι η αρχή και διάρκεια του πλαίσιου ανάλυσης, αντίστοιχα. Η εκτίμηση του εύρους ζώνης, αν δεν χρησιμοποιούμε στάθμιση (standard deviation), ορίζεται ως [5]:

$$[B_u]^2 = \frac{1}{T}\int_{t_0}^{T} (f(t) - F_u)^2 dt \tag{2.9}$$

όπου t_0 και T είναι η αρχή και διάρκεια του πλαίσιου ανάλυσης, αντίστοιχα. Η εκτίμηση της συχνότητας, αν χρησιμοποιήσουμε στάθμιση (weighting), ορίζεται ως [5]:

$$F_w = \frac{\int_{t_0}^{t_0+T} f(t)[\alpha(t)]^2 dt}{\int_{t_0}^{t_0+T} [\alpha(t)]^2 dt} \tag{2.10}$$

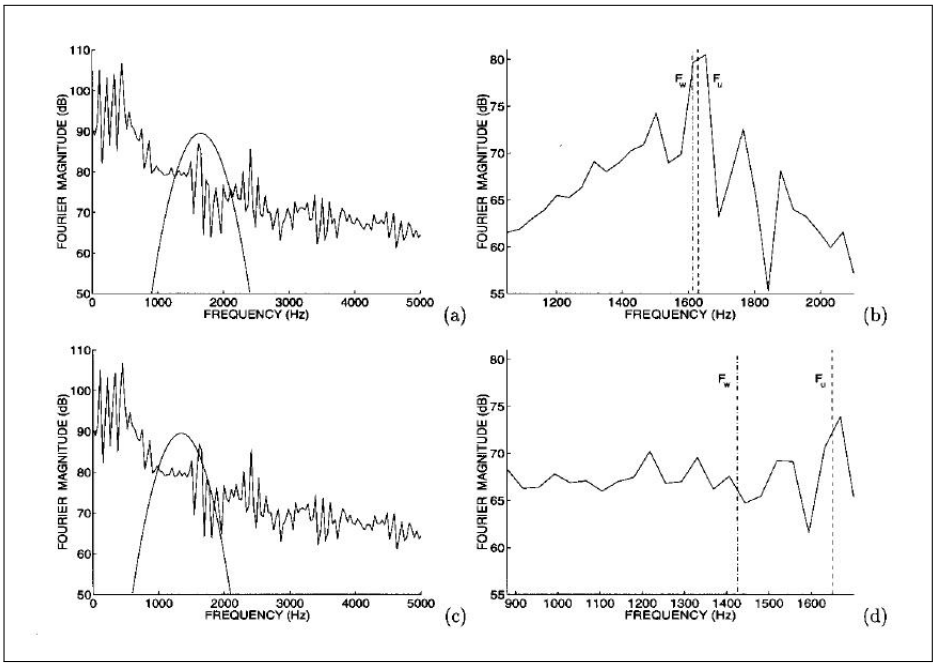
Η εκτίμηση του εύρους ζώνης, αν χρησιμοποιούμε στάθμιση, ορίζεται ως [5]:

$$[B_w]^2 = \frac{\int_{t_0}^{t_0+T} [\dot{\alpha}(t)/2\pi]^2 + (f(t) - F_u)^2[\alpha(t)]^2 dt}{\int_{t_0}^{t_0+T} [\alpha(t)]^2 dt} \tag{2.11}$$

όπου ο όρος $[\alpha(t)]^2$ χρησιμοποιείται ως βάρος και ο όρος $[\dot{\alpha}(t)/2\pi]^2$ εκφράζει τη συνεισφορά του στιγμιαίου πλάτους $a(t)$ στο εύρος ζώνης. Η βασική διαφορά στη συμπεριφορά της F_u σε σχέση με την F_w είναι η εξής: Η F_u επικεντρώνεται στην μεγαλύτερη κορυφή του φάσματος, ενώ η F_w λαμβάνει υπόψη όλη τη φασματική ζώνη [5].

Ως αποτέλεσμα, αν έχουμε έμφωνη περιοχή, η F_u ταυτίζεται με την ισχυρότερη αρμονική του εκάστοτε εύρους ζώνης. Αν όμως δεν υπάρχει μια κορυφή ξεκάθαρα μεγαλύτερη από τις υπόλοιπες, στην περιοχή που μας ενδιαφέρει, η F_u μπορεί να έχει κακή συμπεριφορά. Αντίθετα, η F_w, ακριβώς διότι λαμβάνει υπόψη όλη τη φασματική ζώνη, προσεγγίζει καλύτερα τη συχνότητα συντονισμού. Επομένως, η F_u είναι καταλληλότερη για την εκτίμηση της θεμελιώδους συχνότητας, ενώ η F_w είναι καταλληλότερη για την εκτίμηση των formants. Αυτό αποδεικνύεται και από το επόμενο σχήμα. Στο σχήμα αυτό έχουμε δύο περιπτώσεις. Στην πρώτη περίπτωση (εικόνα a, εικόνα b) έχουμε την παρουσίαση του φάσματος Fourier ενός πλαισίου φωνής των $25ms$ και της απόκρισης συχνότητας ενός Gabor φίλτρου με κεντρική συχνότητα $1600Hz$ (εικόνα a) καθώς και την απόκριση συχνότητας του φιλτραρισμένου σήματος (εικόνα b). Στην δεύτερη περίπτωση (εικόνα c, εικόνα d) έχουμε την παρουσίαση του φάσματος Fourier ενός πλαισίου φωνής των $25ms$ και της απόκρισης συχνότητας ενός Gabor φίλτρου με κεντρική συχνότητα $1300Hz$ (εικόνα c) καθώς και την απόκριση συχνότητας του φιλτραρισμένου σήματος(εικόνα d). Για κάθε μια από τις παραπάνω δύο περιπτώσεις, απεικονίζονται οι εκτιμήσεις των F_w, F_u. Παρατηρούμε ότι όταν το φίλτρο περικλύει ένα ισχυρό formant, όπως στην πρώτη περίπτωση, οι τιμές των F_w, F_u σχεδόν συμπίπτουν και είναι αρκετά

ακριβείς. Αντίθετα, όταν δεν έχουμε κάποιο ισχυρό formant, η F_u κατευθύνεται προς την μεγαλύτερη κορυφή, ενώ η F_w κατευθύνεται προς τη μέση τιμή, που είναι που κοντά στη συχνότητα συντονισμού [5]. Εφόσον, λοιπόν, η F_w είναι καταλληλότερη για την εκτίμηση των formants, θα χρησιμοποιήσουμε, στον υπολογισμό του FMP, τη σταθμισμένη συχνότητα και το σταθμισμένο εύρος ζώνης για κάθε formant. Ο υπολογισμός αυτός παρουσιάζεται στο επόμενο υποκεφάλαιο.



Σχήμα 2.5: Εικόνα a: Φάσμα Fourier ενός πλαισίου φωνής των $25ms$ και απόκριση συχνότητας ενός Gabor φίλτρου με κεντρική συχνότητα $1600Hz$, εικόνα b: απόκριση συχνότητας του φιλτραρισμένου σήματος, εικόνα c: φάσμα Fourier ενός πλαισίου φωνής των $25ms$ και απόκριση συχνότητας ενός Gabor φίλτρου με κεντρική συχνότητα $1300Hz$, εικόνα d: απόκριση συχνότητας του φιλτραρισμένου σήματος, πηγή: [5]

2.5.4 Εκτίμηση του FMP

Έχοντας υπολογίσει τη σταθμισμένη συχνότητα και το σταθμισμένο εύρος ζώνης, μπορούμε πλέον να υπολογίσουμε την τιμή του FMP (Frequency Modulation Percentage) για το formant που μας ενδιαφέρει. Η τιμή του FMP, για το formant που μας ενδιαφέρει, υπολογίζεται από τον παρακάτω τύπο [7]:

$$FMP_i = B_i/F_i \tag{2.12}$$

όπου B είναι το σταθμισμένο εύρος ζώνης του formant, F είναι η σταθμισμένη συχνότητα του formant, και $i = 1, ..., K$ ο αριθμός της εκάστοτε φασματικής ζώνης, δηλαδή του εκάστοτε formant. Για την ανάλυση σε φασματικές ζώνες χρησιμοποιούνται συνήθως Gabor φίλτρα, σε κατάλληλη συστοιχία, με τις συ-

χνότητες τους κατανεμημένες στην κλίμακα Mel [7]. Η περιγραφή των φίλτρων Gabor βρίσκεται στο Παράρτημα.

2.5.5 Σημασία εύρεσης του FMP

Η εύρεση του FMP είναι χρήσιμη όταν έχουμε $AM - FM$ μοντέλο με συχνότητες συντονισμού (formant frequencies) μη σταθερές κατά τη διάρκεια μιας περιόδου, αλλά κυμαινόμενες γύρω από μια κεντρική συχνότητα (center frequency). Αυτό συμβαίνει επειδή παρέχει πληροφορίες σχετικά με τη διαρκώς χρονικά διακυμαινόμενη δομή του formant, επωφελούμενη και από την πολύ καλή ανάλυση χρόνου του ESA. Στην ουσία, με την εύρεση του FMP χαρτογραφούμε τις διακυμάνσεις του formant [7].

2.6 SMAC

Το SMAC front-end αποτελείται από την πρώτη κανονικοποιημένη κεντρική φασματική ροπή (first normalized central spectral moment) και από λίγους, χαμηλής τάξης, cepstral συντελεστές (low order cepstral coefficients). Στο Παράρτημα αναλύεται ο τρόπος εύρεσης της φασματικής ροπής και οι ιδιότητες της, ενώ στα επόμενα υποκεφάλαια εξηγούμε τη δομή, τον τρόπο εύρεσης και τη σημασία του SMAC front-end.

2.6.1 Δομή και τρόπος εύρεσης του SMAC front-end

Όπως είπαμε προηγουμένως, το SMAC front-end αποτελείται από την κανονικοποιημένη κεντρική φασματική ροπή πρώτης τάξης ($N_c^1, \gamma = 2$) και από κάποιους συντελεστές cepstrum χαμηλής τάξης. Η επιλογή της N_c^1 έναντι της N^1, δηλαδή της κανονικοποιημένης φασματικής ροπής πρώτης τάξης, έγινε για αριθμητικούς κυρίως λόγους. Θεωρητικά είναι ισοδύναμες, αφού διαφέρουν κατά μία σταθερά, αλλά η N_c^1 είναι προτιμότερη, καθώς οι τιμές της έχουν μηδενική μέση τιμή [28]. Οι συντελεστές της κανονικοποιημένης κεντρικής φασματικής ροπής πρώτης τάξης προσεγγίζουν τις συχνότητες συντονισμού (formant frequencies) όταν το εύρος ζώνης των φίλτρων που χρησιμοποιούνται είναι μεγάλο, και μοντελοποιούν τις φασματικές κορυφές του σήματος φωνής στην απεικόνισή του σε πυκνόγραμμα, όπως δείξαμε στο Παράρτημα. Όμως με αυτό τον τρόπο χάνεται η πληροφορία των φασματικών υψών μεταξύ των φασματικών κορυφών και συλλέγεται μόνο η πληροφορία που αφορά τις φασματικές κορυφές. Γι΄ αυτό το λόγο προσθέτουμε μερικούς φασματικούς συντελεστές χαμηλής τάξης, οι οποίοι συλλέγουν τις πληροφορίες που αφορούν τη χονδρική φασματική περιβάλλουσα. Δεν χρειάζεται να κάνουμε κάποιον περαιτέρω μετασχηματισμό, όπως DCT, διότι οι συντελεστές των φασματικών ροπών είναι σε μεγάλο βαθμό ασυσχέτιστοι [1].

Όπως δείξαμε στο Παράρτημα, η φασματική ροπή πρώτης τάξης παρουσιάζει ευαισθησία όσον αφορά τις αρμονικές του pitch, όταν οι συστοιχίες φίλτρων που χρησιμοποιούνται αποτελούνται από φίλτρα με στενό εύρος ζώνης (δηλαδή το εύρος ζώνης είναι συγκρίσιμο με την απόσταση των διαδοχικών αρμονικών) άρα και μικρή επικάλυψη. Στην περίπτωση, μάλιστα, που έχουμε συστοιχία φίλτρων σε κλίμακα Mel, όπως συμβαίνει στην κατασκευή των SMAC front-ends, η ευαισθησία αυτή είναι εντονότερη στα χαμηλότερα φίλτρα, που έχουν στενότερο εύρος ζώνης από ότι τα υψηλότερα. Αυτό το φαινόμενο δεν είναι επιθυμητό στη διαδικασία

αναγνώρισης φωνής, διότι το pitch συνήθως δεν μεταφέρει φωνητική πληροφορία [28]. Για να αντιμετωπίσουμε το φαινόμενο αυτό, μεγαλώνουμε την επικάλυψη, στο πεδίο της συχνότητας, ανάμεσα στα γειτονικά φίλτρα της συστοιχίας. Με αυτό τον τρόπο αυξάνεται το εύρος ζώνης των φίλτρων της συστοιχίας, με αποτέλεσμα κάθε φίλτρο να περιλαμβάνει περισσότερες αρμονικές και να μπορούμε να κάνουμε καλύτερη εκτίμηση. Παλαιότερες μελέτες προσπαθούσαν να επιλύσουν το παραπάνω πρόβλημα είτε μειώνοντας τον αριθμό των φίλτρων που χρησιμοποιούνταν, είτε χρησιμοποιώντας γραμμική κλίμακα για τη συστοιχία φίλτρων. Συστοιχίες φίλτρων με μεγάλη επικάλυψη στο πεδίο της συχνότητας έχουν, επίσης, χρησιμοποιηθεί στην εκτίμηση των formants και στην αναγνώριση ομιλητή [1].

Για τον υπολογισμό των συντελεστών SMAC χρησιμοποιούμε μια συστοιχία φίλτρων Gabor, που περιλαμβάνουν 12 φίλτρα στενού εύρους ζώνης (8kHz) στην περιοχή συχνοτήτων μέχρι τα 4kHz και 16 φίλτρα ευρύτερου εύρους ζώνης (16kHz) στην περιοχή συχνοτήτων από τα 4kHz έως τα 8kHz. Παρόλο που μπορούμε να πετύχουμε παρόμοια ή και καλύτερη απόδοση χρησιμοποιώντας περισσότερα φίλτρα, θέλουμε η διάσταση του διανύσματος SMAC να είναι μικρή. Γι΄ αυτό το λόγο δεν θα προσθέσουμε περισσότερα φίλτρα. Η επικάλυψη μεταξύ των γειτονικών φίλτρων πετυχαίνεται με τη ρύθμιση του εύρους ζώνης τους. Το εύρος ζώνης των 236 Mels έχει πειραματικά αποδειχθεί ότι είναι κοντά στη βέλτιστη τιμή. Για ευκολία, χρησιμοποιούμε την ίδια συστοιχία φίλτρων για τον υπολογισμό των φασματικών συντελεστών χαμηλής τάξης, που προσθέτουμε στο διάνυσμα του SMAC front-end. Έχει πειραματικά αποδειχθεί ότι η πρόσθεση των φασματικών συντελεστών μηδενικής και πρώτης τάξης, δηλαδή C0 και C1 αντίστοιχα, βελτιώνει την απόδοση του συστήματος. Ο συντελεστής C0 περιέχει πληροφορία για την ενέργεια του σήματος και ο συντελεστής C1 για την φασματική κλίση. Η προσθήκη περισσότερων συντελεστών, στην καλύτερη περίπτωση προσδίδει μικρή βελτίωση στην απόδοση, στη χειρότερη οδηγεί σε μείωση της απόδοσης. Τέλος, το διάνυσμα του SMAC συμπληρώνεται με τους συντελεστές Δέλτα και Επιτάχυνσης (Delta and Acceleration coefficients). Πλέον, μέχρι να αποδειχθεί κάτι άλλο, χρησιμοποιούμε τον όρο SMAC όταν αναφερόμαστε στο front-end που χρησιμοποιεί 12 φίλτρα στενού εύρους ζώνης (8kHz) στην περιοχή συχνοτήτων μέχρι τα 4kHz και 16 φίλτρα ευρύτερου εύρους ζώνης (16kHz) στην περιοχή συχνοτήτων από τα 4kHz έως τα 8kHz, και περιέχει μόνο τους συντελεστές C0 και C1.

Όπως θα δείξουμε στο επόμενο υποκεφάλαιο, η εύρεση των συντελεστών SMAC βελτιώνει σημαντικά τα αποτελέσματα του τομέα αναγνώρισης φωνής.

2.6.2 Απόδοση και σημασία εύρεσης του SMAC front-end

Η απόδοση του SMAC front-end αξιολογήθηκε σε μια σειρά πειραματικές μετρήσεις αναγνώρισης φωνής. Οι μετρήσεις έλαβαν χώρα για σήματα φωνής από τις βάσεις δεδομένων TIMIT, AURORA 2 και AURORA 3. Στην περίπτωση της TIMIT μετρήθηκε η απόδοση αναγνώρισης φωνημάτων, σε συνθήκες καθαρής ηχογράφησης, για διαφορετικό αριθμό φίλτρων και πρόσθετων φασματικών συντελεστών στο διάνυσμα του SMAC (είτε μόνο C0, είτε C0 και C1, είτε C0 και C2, είτε C0 και C3), ενώ έγινε σύγκριση της απόδοσης του SMAC front-end με το MFCC front-end για τις προηγούμενες περιπτώσεις. Στην περίπτωση της AURORA 2 και του AURORA 3 (τόσο για την περίπτωση της Ισπανικής ηχογράφησης όσο και για την περίπτωση της Ιταλικής) μετρήθηκε η απόδοση αναγνώρισης λέξεων για διαφορετικά επίπεδα προσθετικού θορύβου και συγκρίθηκε

με την αντίστοιχη απόδοση των MFCC, PLP, RASTA-PLP, καθώς και με την περίπτωση καταπίεσης του θορύβου με χρήση Wiener Filtering (WF). Από τα πειράματα αυτά προέκυψε ότι το SMAC front-end συμπεριφέρεται ελαφρώς καλύτερα από το MFCC στην περίπτωση καθαρής ηχογράφησης (με μικρές διακυμάνσεις που δεν έχουν ιδιαίτερη στατιστική σημασία) και είναι εύρωστο σε συνθήκες προσθετικού θορύβου, εφόσον συμπεριφέρεται αισθητά καλύτερα από το MFCC και λίγο καλύτερα από ότι το RASTA-PLP. Επιπλέον, εξακολουθεί να συμπεριφέρεται καλύτερα από το MFCC και όταν εφαρμόζουμε Wiener Filtering [1].

Εν κατακλείδι, το SMAC front-end παρουσιάζει καλύτερη απόδοση, συγκριτικά με άλλα front-ends, και αυτό επιτεύχθηκε με τις ακόλουθες καινοτομίες, κατά την κατασκευή του. Πρώτον, την εισαγωγή της πληροφορίας της φασματικής περιβάλλουσας στο διάνυσμα των χαρακτηριστικών, με την προσθήκη των C0 και C1 συντελεστών. Δεύτερον, τη χρήση συστοιχίας φίλτρων Gabor με μεγαλύτερα εύρη ζώνης (ή αντίστοιχα μεγαλύτερη επικάλυψη συχνότητας) άρα μειωμένη ευαισθησία της φασματικής ροπής στις αρμονικές του pitch [1, 28].

2.7 Άλλες μέθοδοι παραμετροποίησης

Παραμετροποίηση ενός σήματος φωνής μπορεί να γίνει όχι μόνο ακολουθώντας μια από τις προηγούμενες μεθόδους, αλλά και εφαρμόζοντας τον συνδυασμό τους. Για την ακρίβεια, τον συνδυασμό των MFCC και FMP χαρακτηριστικών ή των SMAC και FMP χαρακτηριστικών. Με αυτό τον τρόπο προσπαθούμε να επιτύχουμε βελτίωση της απόδοσης. Στη συνέχεια θα περιγράψουμε δύο μεθόδους συνδυασμού των παραπάνω χαρακτηριστικών.

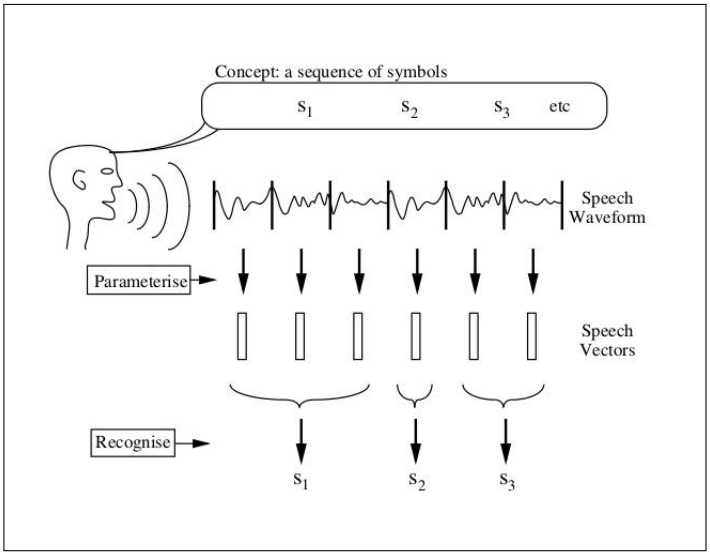
Η πρώτη μέθοδος, η οποία λαμβάνει χώρα μετά τη διαδικασία κατασκευής των διανυσμάτων των front-ends και πριν τη διαδικασία εκπαίδευσης και ελέγχου -τις οποίες θα περιγράψουμε στο επόμενο κεφάλαιο-, έγκειται στην ένωση των χαρακτηριστικών σε ένα διάνυσμα. Συγκεκριμένα, δημιουργούμε ένα νέο διάνυσμα, του οποίου τα πρώτα στοιχεία είναι τα στοιχεία του πρώτου front-end και τα επόμενα στοιχεία είναι εκείνα του δεύτερου front-end. Δηλαδή, στην περίπτωση που συνδέουμε τα χαρακτηριστικά MFCC και FMP, το νέο διάνυσμα θα περιέχει 57 χαρακτηριστικά, εκ των οποίων τα 39 πρώτα θα είναι τα στοιχεία του MFCC διανύσματος και τα επόμενα 18 θα είναι τα στοιχεία του FMP διανύσματος. Ενώ, στην περίπτωση που συνδέουμε τα χαρακτηριστικά MFCC και SMAC, το νέο διάνυσμα θα περιέχει 60 χαρακτηριστικά, εκ των οποίων τα 42 πρώτα θα είναι τα στοιχεία του SMAC διανύσματος και τα επόμενα 18 θα είναι τα στοιχεία του FMP διανύσματος. Έπειτα συνεχίζουμε με τις διαδικασίες της εκπαίδευσης και του ελέγχου, αλλά με το νέο πλέον διάνυσμα χαρακτηριστικών.

Η δεύτερη μέθοδος λαμβάνει χώρα μετά τη διαδικασία εκπαίδευσης και πριν τη διαδικασία ελέγχου, όταν δηλαδή έχουμε δημιουργήσει τα Κρυφά Μαρκοβιανά Μοντέλα (HMM) για το κάθε front-end, αλλά δεν έχουμε βρει την ακρίβεια μετάδοσης. Σε αυτήν τη μέθοδο διαιρούμε τα χαρακτηριστικά του κάθε front-end σε streams και τα συνδυάζουμε. Συγκεκριμένα, δημιουργούμε ένα νέο HMM που περιέχει τον ίδιο αριθμό καταστάσεων με τα άλλα δύο, αλλά η κάθε κατάστασή του απαρτίζεται από δύο streams, το πρώτο εκ των οποίων περιλαμβάνει τα διανύσματα της αντίστοιχης κατάστασης του πρώτου front-end και το άλλο τα διανύσματα της αντίστοιχης κατάστασης του δεύτερου front-end. Αν προσθέσουμε τιμή βάρους (weight) σε κάθε ένα από τα streams της κάθε κατάστασης, αλλάζουμε τη συνεισφορά του κάθε front-end στο HMM, με αποτέλεσμα τη βελτίωση ή χειροτέρευση

της απόδοσης. Επίσης, ενώ ο αριθμός των συντελεστών έχει αυξηθεί, ο αριθμός των παραμέτρων του Κρυφού Μαρκοβιανού Μοντέλου δεν έχει αυξηθεί αισθητά, κάτι που είναι ιδιαίτερα επιθυμητό στη χρήση του μοντέλου κατά τη διαδικασία Εκπαίδευσης και Ελέγχου. Με αυτό τον τρόπο έχουμε δημιουργήσει ένα multi-stream HMM, που θα απαρτίζεται είτε από τα front-ends MFCC και FMP, είτε από τα front-ends SMAC και FMP, και το οποίο μετά θα περάσει από τη διαδικασία Ελέγχου.

2.8 Εκπαίδευση του συστήματος, εύρεση της ακρίβειας

Τα συστήματα αναγνώρισης φωνής, σε γενικές γραμμές, θεωρούν ότι το σήμα φωνής είναι ένα μήνυμα κωδικοποιημένο ως μια ακολουθία από ένα ή περισσότερα σύμβολα. Για να επιτευχθεί η διαδικασία αναγνώρισης της ακολουθίας αυτής, το αρχικό σήμα, που είναι μια συνεχής κυματομορφή, χωρίζεται σε πλαίσια, το καθένα από τα οποία μετατρέπεται σε ένα διάνυσμα συντελεστών και ο κάθε συντελεστής αποτυπώνει ένα χαρακτηριστικό του πλαισίου. Από αυτά τα χαρακτηριστικά μπορούμε να εκτιμήσουμε ποιά ήταν η ακολουθία συμβόλων του αρχικού σήματος και να εξετάσουμε το βαθμό κατά τον οποίο η ακολουθία που εκτιμήσαμε συμπίπτει με την αρχική. Κάτι τέτοιο αποτυπώνεται στην επόμενη εικόνα [25].

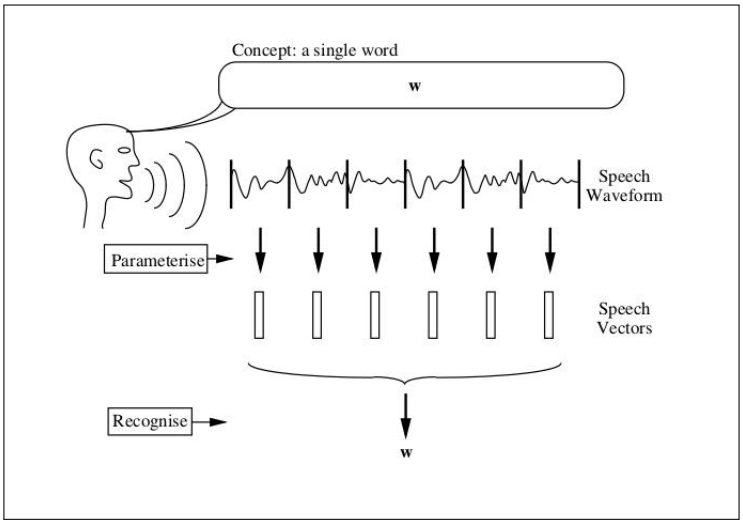


Σχήμα 2.6: Κωδικοποίηση και αποκρυπτογράφηση μηνύματος, πηγή: [25]

Τέτοιοι συντελεστές είναι τα MFCC, τα FMP και τα SMAC, τα οποία περιγράψαμε στα παραπάνω κεφάλαια. Ο τρόπος εύρεσης των παραπάνω εκτιμήσεων είναι ζήτημα που θα μας απασχολήσει σε αυτό το κεφάλαιο.

Η εκτίμηση της αρχικής ακολουθίας συμβόλων έχει δύο σημαντικά προβλήματα. Πρώτον, η αντιστοίχιση συμβόλου-φωνής δεν είναι μοναδική, διότι αρκετές φορές, διαφορετικά σύμβολα δίνουν τον ίδιο ήχο. Επιπλέον, υπάρχουν διαφοροποιήσεις από κυματομορφή σε κυματομορφή, εξαιτίας της διάθεσης του ομιλητή, του

περιβάλλοντος ηχογράφησης κ.α. Δεύτερον, τα όρια μεταξύ των συμβόλων δεν μπορούν να προσδιοριστούν ρητά από την κυματομορφή του σήματος φωνής. Ως εκ τούτου, δεν μπορούμε να αντιμετωπίσουμε την κυματομορφή ως μια ακολουθία από συνεχόμενα στατικά μοντέλα. Το δεύτερο πρόβλημα, δηλαδή η μη γνώση του πού τοποθετούνται τα όρια κάθε συμβόλου, μπορεί να αντιμετωπιστεί με το να περιοριστούμε στην αναγνώριση μιας μεμονομένης λέξης. Αυτό συνεπάγεται με το ότι η κυματομορφή αντιστοιχεί σε ένα σύμβολο (λέξη) που επιλέγεται από ένα λεξιλόγιο, όπως φαίνεται και στο ακόλουθο σχήμα. Για την αναγνώριση αυτή θα χρησιμοποιήσουμε Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models) [25], τη λογική των οποίων θα εξηγήσουμε στη συνέχεια.



Σχήμα 2.7: Πρόβλημα αναγνώρισης μεμονομένης λέξης, πηγή: [25]

2.8.1 Κρυφά Μαρκοβιανά Μοντέλα HMM)

Θεωρούμε ότι η κάθε λέξη μπορεί να παρουσιαστεί ως μια ακολουθία διανυσμάτων ή παρατηρήσεων O, καθορισμένων ως $O = o_1, o_2, ..., o_T$, όπου o_T είναι η παρατήρηση τη χρονική στιγμή t. Το πρόβλημα της αναγνώρισης μιας μεμονομένης λέξης, που αναφέραμε στο προηγούμενο κεφάλαιο, μπορεί θεωρηθεί ως το πρόβλημα υπολογισμού της πιθανότητας

$$\arg\max_i P(w_i|O) \tag{2.13}$$

όπου w_i είναι η i-οστή λέξη. Για τον υπολογισμό της παραπάνω πιθανότητας χρησιμοποιούμε τον κανόνα του Bayes, από όπου παίρνουμε:

$$P(w_i|O) = \frac{P(O|w_i)P(w_i)}{P(O)} \tag{2.14}$$

Καταλαβαίνουμε, λοιπόν, ότι για μια δεδομένη πιθανότητα $P(w_i)$, η πιο πιθανή λέξη εξαρτάται από την πιθανότητα $P(O|w_i)$. Όμως, η ακολουθία παρατηρήσεων O έχει δύο διαστάσεις, άρα ο απευθείας υπολογισμός της δεσμευμένης από κοινού

πιθανότητας $P(o_1, o_2, ..., o_T|w_i)$ για κάθε λέξη, είναι περίπλοκος και χρονοβόρος. Αυτό το ζήτημα λύνεται με τη χρήση Κρυφών Μαρκοβιανών Μοντέλων, όπου αντικαθιστάται ο υπολογισμός της πιθανότητας $P(O|w_i)$ από τον ευκολότερο υπολογισμό των παραμέτρων του Κρυφού Μαρκοβιανού Μοντέλου. Θεωρούμε, λοιπόν, ότι η ακολουθία των διανυσμάτων που αντιπροσωπεύει κάθε λέξη, κατασκευάζεται από ένα Κρυφό Μαρκοβιανό Μοντέλό. Περαιτέρω στοιχεία για τον ορισμό και τις παραμέτρους του Κρυφού Μαρκοβιανού Μοντέλου παρατίθενται στο Παράρτημα.

Στην αναγνώριση φωνής, αφού κατασκευάσουμε το Κρυφό Μαρκοβιανό Μοντέλο, κάνουμε εκπαίδευση (training) των παραμέτρων του και, στη συνέχεια, χρησιμοποιούμε τα αποτελέσματα της εκπαίδευσης για την αναγνώριση φωνής (testing). Στη συνέχεια, θα εξηγήσουμε, συνοπτικά, τις διαδικασίες εκπαίδευσης και ελέγχου.

2.8.2 Διαδικασία της εκπαίδευσης

Στη διαδικασία της εκπαίδευσης (training), εκτιμάμαι τις παραμέτρους του Κρυφού Μαρκοβιανού Μοντέλου (HMM), που έχει προηγουμένως κατασκευαστεί. Για την επίτευξη αυτού του σκοπού, χρησιμοποιούμε τον αλγόριθμο Baum-Welch. Ο αλγόριθμος αυτός δίνει απάντηση στην ακόλουθη ερώτηση: "Αν γνωρίζουμε την ακολουθία των παρατηρήσεων O, ποιές είναι οι πιο πιθανές παράμετροι του HMM"; Με λίγα λόγια, βρίσκει το HMM (λ) που μεγιστοποιεί την πιθανότητα $P(O|\lambda)$. Για την εκτέλεση του αλγορίθμου Baum-Welch χρειάζονται άλλοι δύο αλγόριθμοι: ο αλγόριθμος Forward και ο αλγόριθμος Backward. Ο αλγόριθμος Forward δίνει την πιθανότητα να βρισκόμαστε τη χρονική στιγμή t στην κατάσταση j και η ακολουθία των παρατηρήσεων μέχρι εκείνη τη χρονική στιγμή να είναι $o_1, ..., o_t$. Ο αλγόριθμος Backward δίνει την πιθανότητα να βρισκόμαστε τη χρονική στιγμή t στην κατάσταση j και η ακολουθία των παρατηρήσεων που ακολουθεί να είναι $o_{t+1}, ..., o_T$ [25]. Έχοντας υπολογίσει τις πιθανότητες που προκύπτουν από τους αλγόριθμους Forward και Backward, μπορούμε να υλοποιήσουμε τον αλγόριθμο Baum-Welch. Στο Παράρτημα περγράφουμε τον τρόπο υλοποίησης καθενός από τους παραπάνω αλγορίθμους.

Σε αυτό το σημείο θα θέλαμε να σημειώσουμε ότι για τον υπολογισμό των πιθανοτήτων Forward και Backward χρησιμοποιείται ένας αρκετά μεγάλος αριθμός πιθανοτήτων. Αυτό σημαίνει ότι οι πιθανότητες αυτές γίνονται πολύ μικρά νούμερα. Για να αποφευχθεί αυτό, ο υπολογισμός αυτών των πιθανοτήτων στο HTK, γίνεται με χρήση λογαρίθμου. Πραγματοποιείται με την εντολή $HRest$ αν χρησιμοποιούμε εκπαίδευση απομονωμένης οντότητας (isolated-unit training) ή με την εντολή $HERest$ αν χρησιμοποιούμε εκπαίδευση ενσωματωμένης οντότητας (embedded-unit training) [25]. Δηλαδή, η εντολή $HRest$ χρησιμοποιείται όταν εκτιμούμε τις παραμέτρους ενός μόνο HMM, ενώ η εντολή $HERest$ λαμβάνει ως είσοδο ένα σύνολο από HMMs και εκτιμά τις παραμέτρους για κάθε HMM αυτού του συνόλου.

Έφόσον ολοκληρωθεί η διαδικασία εκπαίδευσης και έχουν εκτιμηθεί οι πιο πιθανές παράμετροι του HMM, μπορούμε να βρούμε την πιο πιθανή αλληλουχία καταστάσεων για το HMM, χρησιμοποιώντας τον αλγόριθμο Viterbi, όπως θα εξηγήσουμε στη συνέχεια.

2.8.3 Διαδικασία της αναγνώρισης

Στη διαδικασία της αναγνώρισης βρίσκουμε την πιο πιθανή αλληλουχία καταστάσεων για ένα HMM, αν γνωρίζουμε την αλληλουχία παρατηρήσεων και αφού προηγουμένος έχουμε βρει τις πιο πιθανές παραμέτρους του HMM, μέσω της διαδικασίας εκπαίδευσης. Για να βρούμε την πιο πιθανή αλληλουχία καταστάσεων κάνουμε χρήση του αλγόριθμου Viterbi, ο οποίος υπολογίζεται στο HTK με την εντολή $HVite$. Αν π.χ. μια λέξη αντιστοιχεί σε ένα HMM, τότε με τον αλγόριθμο Viterbi βρίσκουμε την πιο πιθανή ακολουθία καταστάσεων του HMM γι΄ αυτή τη λέξη [25]. Στο Παράρτημα εξηγούμε τον τρόπο υλοποίησης του αλγόριθμου αυτού.

Κεφάλαιο 3

ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΔΙΚΟΥ ΜΑΣ ΣΥΣΤΗΜΑΤΟΣ

3.1 Στόχος

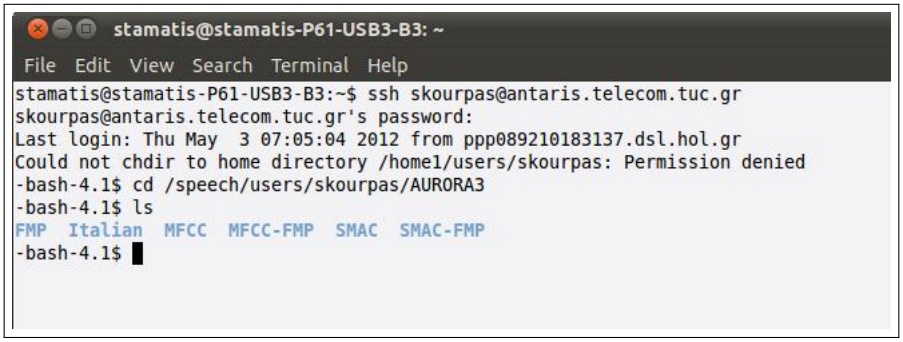
Από τα παραπάνω, δημιουργήθηκε η ιδέα της έρευνας του κατά πόσο επηρρεάζουν την ακρίβεια μετάδοσης ενός multistream συστήματος μετάδοσης σήματος φωνής, οι αλλαγές σε κάποιους παράγοντες. Επικετρωθήκαμε σε δύο περιπτώσεις. Πρώτον, στην περίπτωση που έχουμε front-end τύπου MFCC-FMP. Δεύτερον, στην περίπτωση που έχουμε front-end τύπου SMAC-FMP. Ακόμη, η ανίχνευση του πόσο και αν βελτιώνεται η ακρίβεια μετάδοσης του συστήματος αν κάνουμε αλλαγές σε αυτούς τους παράγοντες.

3.2 Υλική υποδομή

Για τις παραπάνω περιπτώσεις, εφαρμόσαμε τα χαρακτηριστικά των front-ends στη βάση δεδομένων AURORA 3 για την περίπτωση της Ισπανικής γλώσσας (AURORA 3 Speech Database - Spanish Task). Αυτή η βάση δεδομένων μας δίνει τη δυνατότητα αναγνώρισης λέξεων (word level recognition task). Περιέχει ηχογραφημένα αρχεία, που περιέχουν τους αριθμούς της Ισπανικής γλώσσας, από το ένα έως το εννέα, συμπεριλαμβανομένου και του μηδέν. Τα αρχεία είναι δειγματοληπτημένα στα $8kHz$ και ηχογραφημένα από δύο ειδών μικρόφωνα, ένα κοντινής και ένα μακρινής απόστασης. Είναι φυσικό ότι το μικρόφωνο μακρινής απόστασης επηρρεάζεται, από τον θόρυβο, περισσότερο από ότι το μικρόφωνο κοντινής απόστασης. Η ηχογράφηση γίνεται για τρεις καταστάσεις οδήγησης σε αυτοκίνητο, με αντίστοιχα SNR 12, 9 και 5 dB. Στην πρώτη κατάσταση, που ονομάζεται ΄Κατάσταση Καλής Ταύτισης' (Well-Matched - WM), τόσο τα δεδομένα για τη διαδικασία εκπαίδευσης (training), όσο και τα δεδομένα για τη διαδικασία ελέγχου (testing) συλλέγονται με μικρόφωνο κοντινής απόστασης, τα επηρρεάζει ελάχιστα ο θόρυβος και ταυτίζονται αρκετά μεταξύ τους. Στην δεύτερη κατάσταση, που ονομάζεται ΄Κατάσταση Μέτριας Μη Ταύτισης' (Medium Mismatch - MM) τα δεδομένα για τη διαδικασία εκπαίδευσης (training), συλλέγονται με μικρόφωνο

κοντινής απόστασης, ενώ τα δεδομένα για τη διαδικασία ελέγχου (testing) συλλέγονται με μικρόφωνο μακρινής απόστασης, και διαφέρουν στα επίπεδα θορύβου -χαμηλός και μέτριος θόρυβος για τη διαδικασία εκπαίδευσης, υψηλός θόρυβος για τη διαδικασία ελέγχου-. Τέλος, στην τρίτη κατάσταση, που ονομάζεται Ϋκατάσταση Υψηλής Μη Ταύτισης' (High Mismatch - HM) τα δεδομένα για τη διαδικασία εκπαίδευσης συλλέγονται με μικρόφωνο κοντινής απόστασης, σε αντίθεση με τα δεδομένα για τη διαδικασία ελέγχου που συλλέγονται με μικρόφωνο μακρινής απόστασης, ενώ υπάρχει μεγάλη απόκλιση στα επίπεδα θορύβου, μεταξύ των δύο διαδικασιών [26].

Για την κατασκευή των χαρακτηριστικών front-end χρησιμοποιήσαμε κώδικα Perl και κώδικα Matlab. Για τη διαδικασία εύρεσης της ακρίβειας μετάδοσης χρησιμοποιήσαμε την πειραματική πλατφόρμα HTK [25]. Τόσο οι κώδικες Perl και Matlab, όσο και οι κώδικες του HTK αποθηκεύτηκαν στον υποφάκελο skourpas του φακέλου speech του Εργαστηρίου Τηλεπικοινωνιών και Δικτύων του Πολυτεχνείου Κρήτης. Τα προγράμματα Perl, Matlab και η πειραματική πλατφόρμα HTK βρίσκονται στον server του συγκεκριμένου εργαστηρίου. Ως εκ τούτου, όποτε θέλαμε να τα χρησιμοποιήσουμε, τα καλούσαμε από το κατάλληλο μονοπάτι (path). Τρέχαμε, δηλαδή το πρόγραμμα μέσω του server του εργαστηρίου. Η προσβαση στον φάκελο speech γινόταν με χρήση τη εντολής ssh μέσω οποιουδήποτε από τους τρεις εξυπηρετές, τον antaris, τον orion, την fryne. Πχ πληκτρολογούσαμε στο terminal την εντολή ssh skourpas@antaris.telecom.tuc.gr, έπειτα δίναμε τον κατάλληλο κωδικό πρόσβασης και εισερχόμασταν στον φάκελο speech. Στο παρακάτω σχήμα φαίνεται αυτή η μέθοδος.



Σχήμα 3.1: Τρόπος εισόδου στον φάκελο αποθήκευσης της εργασίας

Στη συνέχεια περιγράφουμε τον τρόπο υλοποίησης καθενός από τα front-end χαρακτηριστικά που θα χρησιμοποιήσουμε στο σύστημά μας.

3.3 Υλοποίηση του front-end MFCC-FMP

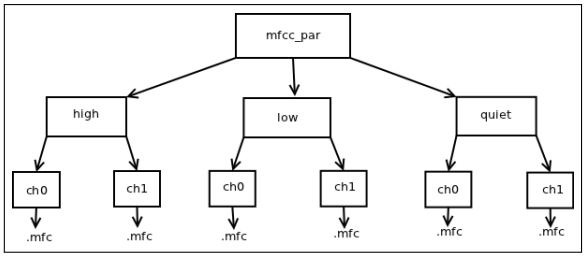
3.3.1 Εισαγωγή

Για την υλοποίηση του front-end MFCC-FMP χρησιμοποιήσαμε τα θεωρητικά δεδομένα που παρατίθενται στο κεφάλαιο 3 και αφορούν την κατασκευή των $MFCC$, των FMP, καθώς και την κατασκευή multistream. Με αυτό το θεωρητικό υπόβαθρο, κατασκευάσαμε αρχικά τους 39 συντελεστές $MFCC$, έπειτα τους

18 συντελεστές FMP βασιζόμενοι στα (word segments) των $MFCC$, και μετά τους 57 συντελεστές $MFCC-FMP$. Τέλος, θέσαμε κατάλληλες τιμές στα βάρη του κάθε ενός συντελεστή και βρήκαμε την ακρίβεια μετάδοσης. Στη συνέχεια, παρατίθενται το κάθε βήμα της υλοποίησης.

3.3.2 Υλοποίηση των συντελεστών MFCC και εύρεση της ακρίβειας

Για την κατασκευή των 39 συντελεστών MFCC χρησιμοποιήσαμε ως αρχεία εισόδου τα wav αρχεία της βάσης δεδομένων AURORA 3, τα οποία είναι αποθηκευμένα σε αρχεία, ανάλογα με το αν η ηχογράφηση έγινε από μικρόφωνο κοντινής ή μακρινής απόστασης και από το επίπεδο του θορύβου που επικρατούσε εκείνη τη χρονική περίοδο. Το παρακάτω σχήμα δείχνει τον συγκεκριμένο τρόπο αποθήκευσης.



Σχήμα 3.2: Μονοπάτι αποθήκευσης των .wav αρχείων

Όπου το αρχείο mfcc_par περιέχει όλα τα .wav αρχεία, τα high, low, quiet είναι τα αρχεία για υψηλό, μέτριο και χαμηλό επίπεδο θορύβου, αντίστοιχα, και ch0, ch1 είναι τα αρχεία για μικρόφωνο κοντινής και μακρινής απόστασης, αντίστοιχα.

Στόχος μας είναι η μετατροπή καθενός από αυτά τα αρχεία σε έναν πίνακα που έχει ως γραμμές τα πλαίσια και ως στήλες τους 39 συντελεστές MFCC. Δηλαδή περιέχει όλους τους συντελεστές MFCC για κάθε πλαίσιο. Για να επιτευχθεί αυτό, φτιάξαμε κώδικα perl που δημιουργεί ένα προσωρινό matlab αρχείο τύπου $tmpmat...m$ που περιέχει τα αρχεία εισόδου και τα αντίστοιχα αρχεία εξόδου για κάθε αρχείο .wav. Στη συνέχεια, ο κώδικας perl καλεί τη matlab, ώστε αυτή να πραγματοποιήσει την κατασκευή των MFCC συντελεστών για κάθε αρχείο .wav. Στο παρακάτω σχήμα παρουσιάζεται ένα τμήμα του $tmpmat...m$ αρχείου.

```
path(path,'/teras/speech/users/skourpas/AURORA3/MFCC/14_16');
spanish_mfcc('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030b1.0.wav','/speech/users/skourpas/AURORA3/MFCC/14_16/features/mfcc_par/high/ch0/v
10030b1.0.mfc');
spanish_mfcc('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030c1.0.wav','/speech/users/skourpas/AURORA3/MFCC/14_16/features/mfcc_par/high/ch0/v
10030c1.0.mfc');
spanish_mfcc('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030c3.0.wav','/speech/users/skourpas/AURORA3/MFCC/14_16/features/mfcc_par/high/ch0/v
10030c3.0.mfc');
spanish_mfcc('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030c4.0.wav','/speech/users/skourpas/AURORA3/MFCC/14_16/features/mfcc_par/high/ch0/v
10030c4.0.mfc');
spanish_mfcc('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i1.0.wav','/speech/users/skourpas/AURORA3/MFCC/14_16/features/mfcc_par/high/ch0/v
10030i1.0.mfc');
spanish_mfcc('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i2.0.wav','/speech/users/skourpas/AURORA3/MFCC/14_16/features/mfcc_par/high/ch0/v
10030i2.0.mfc');
spanish_mfcc('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i3.0.wav','/speech/users/skourpas/AURORA3/MFCC/14_16/features/mfcc_par/high/ch0/v
10030i3.0.mfc');
spanish_mfcc('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i4.0.wav','/speech/users/skourpas/AURORA3/MFCC/14_16/features/mfcc_par/high/ch0/v
10030i4.0.mfc');
```

Σχήμα 3.3: Προσωρινό matlab αρχείο για την κατασκευή των συντελεστών MFCC

Στη συνέχεια, υλοποιήσαμε σε κώδικα matlab το κάθε βήμα της θεωρίας εύρεσης MFCC. Αρχικά, ο κώδικας μετατρέπει το σήμα εισόδου σε ένα διάνυσμα με ακέραια ψηφία. Έπειτα πραγματοποιείται η διαδικασία της προέμφασης, με συντελεστή 0.97. Ακολουθεί η δημιουργία των πλαισίων, στην οποία ορίζουμε ως μέγεθος πλαισίου τα 200 δείγματα και ως επικάλυψη τα 100 δείγματα, έχουμε, δηλαδή, επικάλυψη 50%. Το κάθε πλαίσιο πολλαπλασιάζεται με ένα παράθυρο Hamming και το αποτέλεσμα τους υπόκειται σε μετασχηματισμό Fourier. Στη συνέχεια φιλτράρουμε το φάσμα που προέκυψε με mel φίλτρο και αθροίζουμε. Ως αποτέλεσμα του αθροίσματος προκύπτει ένας συντελεστής για κάθε πλαίσιο. Στην υλοποίησή μας έχουμε θέσει 40 mel φίλτρα. Επαναλαμβάνουμε τη διαδικασία του φιλτραρίσματος για καθένα από τα 40 mel φίλτρα. Επομένως, μετά το φιλτράρισμα προκύπτουν 40 συντελεστές για κάθε πλαίσιο. Αφού υπολογίσουμε το λογάριθμο του κάθε πλαισίου (που τώρα πια απαρτίζεται από 40 νούμερα), μεταφέρουμε τους συντελεστές από το πεδίο της συχνότητας στο πεδίο του χρόνου, με χρήση του Διακριτού Μετασχηματισμού Συνημιτόνου (DCT). Ως αποτέλεσμα του DCT προκύπτει ένα διάνυσμα με 13 ψηφία, τα οποία είναι οι συντελεστές MFCC. Ακολουθεί η διαδικασία CMS. Τέλος, υπολογίζοντας τους συντελεστές Δέλτα και Επιτάχυνσης, προστίθενται στους αρχικούς 13 συντελεστές άλλοι 13 από τη διαδικασία εύρεσης των συντελεστών Δέλτα και άλλοι 13 από τη διαδικασία εύρεσης των συντελεστών Επιτάχυνσης. Προκύπτει, λοιπόν, το τελικό διάνυσμα των MFCC, που αποτελείται από 39 συντελεστές. Στη συνέχεια παρουσιάζεται πίνακας που περιέχει τις συναρτήσεις που περιλαμβάνει ο κώδικας.

Όνομα συνάρτησης	**Λειτουργία συνάρτησης**
spanish_mfcc	Υλοποιεί όλη τη διαδικασία δημιουργίας των $MFCC$ και καλεί και τις υπόλοιπες συναρτήσεις.
loadbin2	Διαβάζει τα δεδομένα από το αρχείο .wav και δημιουργεί ένα διάνυσμα με ακέραια ψηφία των δεκαέξι bits ($int16$).
mfb2cep	Πραγματοποιεί τη διαδικασία Διακριτού Μετασχηματισμού Συνημιτόνου (Discrete Cosine Transform - DCT).
append_deltas	Βρίσκει τους συντελεστές Δέλτα (Delta coefficients) και τους συντελεστές Επιτάχυνσης (Acceleration coefficients).
htk_write	Αποθηκεύει το διάνυσμα των συντελεστών MFCC στο δυαδικό αρχείο εξόδου.

Για την εύρεση της ακρίβειας χρησιμοποιούμε την πειραματική πλατφόρμα HTK. Η εύρεση της ακρίβειας χωρίζεται σε δύο μέρη. Στο πρώτο μέρος πραγματοποιούμε τη διαδικασία εκπαίδευσης (training) και στο δεύτερο μέρος τη διαδικασία ελέγχου (testing). Υπολογίζουμε την ακρίβεια μετάδοσης για κάθε μια από τις καταστάσεις Well-Matched - WM, Medium Mismatch - MM και High Mismatch. Στη διαδικασία της εκπαίδευσης, στόχος μας είναι να καταλήξουμε σε ένα κατάλληλο Κρυφό Μαρκοβιανό Μοντέλο για κάθε λέξη. Αρχικά, ορίζουμε ένα πρωτότυπο HMM μοντέλο (prototype model). Οι παράμετροι αυτού του μοντέλου δεν είναι σημαντικοί και ο ρόλος του έγκειται στον ορισμό της τοπολογίας του μοντέλου. Το πρωτότυπο μοντέλο αποτελείται από 14 καταστάσεις από αριστερά προς τα δεξιά (14 state left-right). Η κάθε κατάσταση έχει δύο διανύσματα, το διάνυσμα

της μέσης τιμής, με όλους τους συντελεστές του ίσους με μηδέν και το διάνυσμα της διακύμανσης, με όλους τους συντελεστές του ίσους με ένα. Καθένα από αυτά τα διανύσματα έχει 39 συντελεστές. Έχοντας ως αφετηρία το πρωτότυπο μοντέλο, υπολογίζουμε το νέο διανυσμα μέσης τιμής και διακύμανσης, και θέτουμε όλες τις καταστάσεις να έχουν το ίδιο διάνυσμα μέσης τιμής και διακύμανσης. Κατασκευάζουμε το ίδιο μοντέλο για κάθε λέξη. Προσθέτουμε το σιωπηλό μοντέλο (silence model), που έχει 3 καταστάσεις. Στη συνέχεια, επανεκτιμούμε το μοντέλο χρησιμοποιώντας τα δεδομένα που βρίσκονται σε κατάλληλα script files και κάνοντας χρήση του ενσωματωμένου - embedded αλγόριθμου Baum-Welch. Πραγματοποιούμε μια σειρά επανεκτιμήσεις. Προσθέτουμε το μοντέλο μικρής παύσης (short pause model-sp), που έχει μια κατάσταση και η κατασκευή του βασίζεται στο silence model. Για την ακρίβεια, η κατάστασή του είναι ίδια με τη μεσαία καταστση του silence model. Ταυτίζουμε τη μοναδική κατάσταση του μοντέλου sp με την κεντρική κατάσταση του silence μοντέλου. Συνεχίζουμε τις επανεκτιμήσεις και προσθέτουμε τους κατάλληλους mixture components μετά τη δωδέκατη και την εικοστή επανεκτίμηση. Σταματάμε στο εικοστό όγδοο Κρυφό Μαρκοβιανό Μοντέλο, όταν και θεωρούμε ότι έχουμε μια ικανοποιητική σύγκλιση αποτελεσμάτων, επομένως έχουμε βρει αρκετά πιθανές παραμέτρους του HMM. Τότε το κάθε HMM αποτελείται από 14 καταστάσεις, με κάθε κατάσταση να αποτελείται από 16 Κανονικές Κατανομές (Gaussian Mixtures). Έπειτα προχωράμε στη διαδικασία ελέγχου, όπου με χρήση του αλγόριθμου Viterbi βρίσκουμε την πιο πιθανή αλληλουχία λέξεων-αριθμών για κάθε ηχογραφημένο αρχείο και, τέλος, υπολογίζουμε την ακρίβεια μετάδοσης για κάθε πρόταση.

Οι εντολές HTK που χρησιμοποιούμε στις διαδικασίες εκπαίδευσης και ελέγχου παρατείθενται, συνοπτικά, στον ακόλουθο πίνακα:

Όνομα εντολής	**Λειτουργία εντολής**
HCompV	Σκανάρει τα αρχεία δεδομένων, υπολογίζει το νέο διάνυσμα μέσης τιμής και διακύμανσης και θέτει όλες τις καταστάσεις του Κρυφού Μαρκοβιανού Μοντέλου (HMM) να έχουν το ίδιο διάνυσμα μέσης τιμής και διακύμανσης. Επίσης, κατασκευάζει το αρχείο vFloors, το οποίο περιέχει ένα διάνυσμα διακύμανσης ίσο με 0.01 φορές το διάνυσμα της διακύμανσης που υπολογίστηκε πριν. Αυτό το διάνυσμα χρησιμοποιείται για να θέσουμε όριο στις τιμές της διακύμανσης στα επόμενα βήματα.
macro_gen	Κατασκευάζει το αρχείο macros, που περιέχει τα στοιχεία περιγραφής του HMM (μέγεθος διανύσματος κ.τ.λ.) και το διάνυσμα διακύμανσης του vFloors.
models_lmixsil	Κατασκευάζει το αρχείο models, που περιέχει ένα αντίγραφο του HMM για κάθε λέξη-αριθμό. Ακόμα, περιέχει το HMM για το σιωπηλό μοντέλο sil. Με αυτό τον τρόπο αποφεύγουμε να έχουμε ξεχωριστό αρχείο για κάθε λέξη-αριθμό.
HERest	Επανεκτιμά τις παραμέτρους των HMMs που βρίσκονται στο αρχείο models και το διάνυσμα διακύμανσης του αρχείου macros, κάνοντας χρήση του embedded Baum-Welch αλγόριθμου.
spmodel_gen	Κατασκευάζει το μοντέλο μικρής παύσης-sp.
HHEd	Την πρώτη φορά που την καλούμε, συνδέει το μοντέλο sp με το μοντέλο sil, ταυτίζοντας τη μοναδική κατάσταση του sp με τη μεσαία κατασταση του sil. Ακόμα αλλάζει κάποιες τιμές στους πίνακες διακύμανσης του sil και του sp. Τις άλλες φορές πραγματοποιεί mixtures σε διάφορα HMMs.
HVite	Βρίσκει, με χρήση του αλγόριθμου Viterbi, την πιο πιθανή αλληλουχία λέξεων-αριθμών για κάθε ηχογραφημένο αρχείο.
HResult	Υπολογίζει την ακρίβεια μετάδοσης για κάθε πρόταση, δηλαδή για το σύνολο των λέξεων του αρχείου και για κάθε λέξη.

Πίνακας 3.1: Εντολές των διαδικασιών εκπαίδευσης και ελέγχου

3.3.3 Υλοποίηση των συντελεστών FMP και εύρεση της ακρίβειας

Για την κατασκευή των 18 συντελεστών FMP χρησιμοποιήσαμε ως αρχεία εισόδου τα wav αρχεία της βάσης δεδομένων AURORA 3, που χρησιμοποιήσαμε και στην κατασκευή των συντελεστών MFCC. Ο στόχος μας είναι ο ίδιος με εκείνο της κατασκευής των MFCC συντελεστών, δηλαδή η μετατροπή καθενός από αυτά τα αρχεία σε έναν πίνακα που έχει ως γραμμές του τα πλαίσια και ως στήλες του τους 18 συντελεστές FMP. Δηλαδή περιέχει όλους τους συντελεστές FMP για κάθε πλαίσιο. Για να επιτευχθεί αυτό ακολουθήσαμε την ίδια διαδικασία με εκείνη της κατασκευής MFCC συντελεστών, δηλαδή φτιάξαμε κώδικα perl που δημιουργεί ένα προσωρινό matlab αρχείο τύπου *tmpmat...m* που περιέχει τα αρχεία εισόδου και τα αντίστοιχα αρχεία εξόδου για κάθε αρχείο .wav . Στη συνέχεια, ο κώδικας perl καλεί τη matlab , ώστε αυτή να πραγματοποιήσει την κατασκευή των FMP συντελεστών, για κάθε αρχείο .wav. Στο παρακάτω σχήμα παρουσιάζεται ένα τμήμα του *tmpmat...m* αρχείου.

```
path(path,'/teras/speech/users/skourpas/AURORA3/FMP/14_16');
spanish_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030b1.0.wav','/speech/users/skourpas/AURORA3/FMP/14_16/features_fmp/mfcc_par/high/ch0
/v10030b1.0.fmp');
spanish_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030c1.0.wav','/speech/users/skourpas/AURORA3/FMP/14_16/features_fmp/mfcc_par/high/ch0
/v10030c1.0.fmp');
spanish_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030c3.0.wav','/speech/users/skourpas/AURORA3/FMP/14_16/features_fmp/mfcc_par/high/ch0
/v10030c3.0.fmp');
spanish_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030c4.0.wav','/speech/users/skourpas/AURORA3/FMP/14_16/features_fmp/mfcc_par/high/ch0
/v10030c4.0.fmp');
spanish_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i1.0.wav','/speech/users/skourpas/AURORA3/FMP/14_16/features_fmp/mfcc_par/high/ch0
/v10030i1.0.fmp');
spanish_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i2.0.wav','/speech/users/skourpas/AURORA3/FMP/14_16/features_fmp/mfcc_par/high/ch0
/v10030i2.0.fmp');
spanish_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i3.0.wav','/speech/users/skourpas/AURORA3/FMP/14_16/features_fmp/mfcc_par/high/ch0
/v10030i3.0.fmp');
spanish_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i4.0.wav','/speech/users/skourpas/AURORA3/FMP/14_16/features_fmp/mfcc_par/high/ch0
/v10030i4.0.fmp');
```

Σχήμα 3.4: Προσωρινό matlab αρχείο για την κατασκευή των συντελεστών FMP

Στη συνέχεια υλοποιήσαμε σε κώδικα matlab το κάθε βήμα της θεωρίας εύρεσης των FMP. Αρχικά ο κώδικας μετατρέπει το σήμα εισόδου σε ένα διάνυσμα με ακέραια ψηφία. Έπειτα, αφού κανονικοποιήσουμε τα ψηφία του διανύσματος, πραγματοποιούμε τη διαδικασία της προέμφασης, με συντελεστή 0.97. Στη συνέχεια, υπολογίζουμε τα πλαίσια, για τα οποία ορίζουμε ως μέγεθος πλαισίου τα 200 δείγματα και ως επικάλυψη τα 100 δείγματα, εχουμε, δηλαδή, επικάλυψη 50%. Ακολουθεί η εκτίμηση του στιγμιαίου πλάτους και της στιγμιαίας συχνότητας του σήματος, για κάθε πλαισίο, με χρήση του αλγόριθμου διαχωρισμού ενέργειας ESA. Ως αποτέλεσμα θα έχουμε για κάθε πλαίσιο έξι τιμές στιγμιαίου πλάτους και έξι τιμές στιγμιαίας συχνότητας, διότι ορίσαμε έξι εύρη ζώνης για κάθε πλαίσιο. Μετά κάνουμε median filtering σε κάθε στιγμιαίο πλάτος και στιγμιαία συχνότητα του σήματος. Έπειτα, αφού υπολογίσουμε τη σταθμισμένη συχνότητα και το σταθμισμένο εύρος ζώνης, υπολογίζουμε το FMP. Συνολικά, δηλαδή, έχουμε υπολογίσει έξι τιμές του FMP για κάθε πλαίσιο. Τέλος, υπολογίζοντας τους συντελεστές Δέλτα και Επιτάχυνσης, προστίθενται στους αρχικούς έξι συντελεστές, άλλοι έξι από τη διαδικασία εύρεσης συντελεστών Δέλτα και άλλοι έξι από τη διαδικασία εύρεσης των συντελεστών Επιτάχυνσης. Προκύπτει, λοιπόν το τελικό διάνυσμα των FMP, που αποτελείται από 18 συντελεστές. Στη συνέχεια παρουσιάζεται πίνακας που περιέχει τις συναρτήσεις που περιλαμβάνει ο κώδικας.

Όνομα συνάρτησης	**Λειτουργία συνάρτησης**
spanish_fmp	Υλοποιεί όλη τη διαδικασία δημιουργίας των FMP και καλεί και τις υπόλοιπες συναρτήσεις.
loadbin2	Διαβάζει τα δεδομένα από το αρχείο .wav και δημιουργεί ένα διάνυσμα με ακέραια ψηφία των δεκαέξι bits ($int16$).
mbdemod	Πραγματοποιεί πολυζωνική αποδιαμόρφωση. Στην ουσία, αφού φιλτράρει το σήμα με χρήση φίλτρου Gabor, υπολογίζει το στιγμιαίο πλάτος και τη στιγμιαία συχνότητα, με χρήση του αλγόριθμου διαχωρισμού ενέργειας ESA.
append_deltas	Βρίσκει τους συντελεστές Δέλτα (Delta coefficients) και τους συντελεστές Επιτάχυνσης (Acceleration coefficients).
htk_write	Αποθηκεύει το διάνυσμα των συντελεστών FMP στο δυαδικό αρχείο εξόδου.

Πίνακας 3.2: Συναρτήσεις της διαδικασίας κατασκευής FMP συντελεστών.

Για την εύρεση της ακρίβειας χρησιμοποιούμε, όπως και στην περίπτωση των MFCC συντελεστών, την πειραματική πλατφόρμα HTK, χωρίζουμε τη διαδικασία σε διαδικασία εκπαίδευσης και ελέγχου και επαναλαμβάνουμε τη διαδικασία για τις καταστάσεις Well-Matched - WM, Medium Mismatch - MM και High Mismatch. Όμως, για να βρούμε το κατάλληλο Κρυφό Μαρκοβιανό Μοντέλο (HMM) για κάθε λέξη, θα στηριχτούμε στα label files που προέκυψαν από τη διαδικασία εκπαίδευσης των MFCC συντελεστών. Για την ακρίβεια, θα κατασκευάσουμε το κάθε HMM ξεχωριστά, εντάσσωντας την εκάστοτε λέξη σε συγκεκριμένα χρονικά περιθώρια, τα ίδια με αυτά των MFCC συντελεστών για την αντίστοιχη λέξη. Τα περιθώρια αυτά θα προκύψουν από τα παραπάνω label files.

Αρχικά, ορίζουμε ένα πρωτότυπο HMM μοντέλο, με 14 καταστάσεις από αριστερά προς τα δεξιά και με κάθε κατάσταση αποτελούμενη από δύο διανύσματα, το διάνυσμα της μέσης τιμής, με όλους τους συντελεστές του ίσους με μηδέν και το διάνυσμα της διακύμανσης, με όλους τους συντελεστές του ίσους με ένα, όπως και στην περίπωση των MFCC. Όμως, τώρα τα διανύσματα του μοντέλου έχουν 18 συντελεστές. Έχοντας και πάλι ως αφετηρία το πρωτότυπο μοντέλο, υπολογίζουμε το νέο διανυσμα μέσης τιμής και διακύμανσης, και θέτουμε όλες τις καταστάσεις να έχουν το ίδιο διάνυσμα μέσης τιμής και διακύμανσης. Υπολογίζουμε τις αρχικές παραμέτρους του κάθε μοντέλου, δηλαδή των αριθμών ένα έως εννέα και της σιωπηλής κατάστασης-sil ξεχωριστά, με χρήση του αλγόριθμου Viterbi. Με αυτό τον τρόπο, το κάθε μοντέλο φτιάχνεται ανεξάρτητα από όλα τα άλλα μοντέλα. Για τους παραπάνω υπολογισμούς χρησιμοποιούμε το Master Label File - MLF της διαδικασίας εκπαίδευσης όταν έχουμε MFCC συντελεστές, που έχει προκύψει αμέσως πριν την εισαγωγή του μοντέλου μικρής παύσης - sp. Στη συνέχεια πραγματοποιούμε μια σειρά επανεκτιμήσεων για κάθε μοντέλο χωριστά, έχοντας ως είσοδο το παραπάνω MLF και κάνοντας χρήση του αλγόριθμου Baum-Welch, όχι όμως του ενσωματωμένου - embedded, όπως στην περίπτωση των MFCC. Υπολογίζουμε το μοντέλο sp και πραγματοποιούμε τις απαραίτητες αλλαγές στους πίνακες διακύμανσης του sil και του sp. Έπειτα κάνουμε επανεκτιμήσεις, με τον ίδιο τρόπο όπως πριν, αλλά με το MLF που προέκυψε στο τέλος της διαδικασίας

εκπαίδευσης όταν έχουμε MFCC συντελεστές. Συνεχίζουμε τις επανεκτιμήσεις και τις προσθήκες κατάλληλου αριθμού mixtures σε κάθε κάθε HMM την κατάλληλη στιγμή. Όταν φτάσουμε σε σύγκλιση αποτελεσμάτων στις επανεκτιμήσεις, ενώνουμε όλα τα HMMs σε ένα αρχείο και συνδέουμε το sp με το sil μοντέλο, ταυτίζοντας τη μοναδική κατάσταση του μοντέλου sp με την κεντρική κατάσταση του sil μοντέλου. Τότε έχουμε κατασκευάσει HMMs με 14 καταστάσεις και με 16 Κανονικές Κατανομές (Gaussian Mixtures) για κάθε κατάσταση. Έπειτα πραγματοποιούμε τη διαδικασία ελέγχου, η οποία είναι ίδια με αυτήν της περίπτωσης των MFCC.

Οι εντολές HTK που χρησιμοποιούμε στις διαδικασίες εκπαίδευσης και ελέγχου παρατείθενται, συνοπτικά, στον ακόλουθο πίνακα:

Όνομα εντολής	**Λειτουργία εντολής**
HCompV	Σκανάρει τα αρχεία δεδομένων, υπολογίζει το νέο διάνυσμα μέσης τιμής και διακύμανσης και θέτει όλες τις καταστάσεις του Κρυφού Μαρκοβιανού Μοντέλου (HMM) να έχουν το ίδιο διάνυσμα μέσης τιμής και διακύμανσης. Επίσης, κατασκευάζει το αρχείο vFloors, το οποίο περιέχει ένα διάνυσμα διακύμανσης ίσο με 0.01 φορές το διάνυσμα της διακύμανσης που υπολογίστηκε πριν. Αυτό το διάνυσμα χρησιμοποιείται για να θέσουμε όριο στις τιμές της διακύμανσης στα επόμενα βήματα.
macro_gen	Κατασκευάζει το αρχείο macros, που περιέχει τα στοιχεία περιγραφής του HMM (μέγεθος διανύσματος κ.τ.λ.) και το διάνυσμα διακύμανσης του vFloors.
HInit	Υπολογίζει τις αρχικές παραμέτρους του κάθε HMM, κάνοντας χρήση του αλγόριθμού Viterbi.
HRest	Επανεκτιμά τις παραμέτρους του HMM, κάνοντας χρήση του αλγόριθμου Baum-Welch.
spmodel_gen	Κατασκευάζει το μοντέλο μικρής παύσης-sp.
HHEd	Συνδέει το μοντέλο sp με το μοντέλο sil, ταυτίζοντας τη μοναδική κατάσταση του sp με τη μεσαία καταστση του sil. Ακόμα αλλάζει κάποιες τιμές στους πίνακες διακύμανσης του sil και του sp. Επίσης, πραγματοποιεί mixtures σε διάφορα HMMs.
HVite	Βρίσκει, με χρήση του αλγόριθμου Viterbi, την πιο πιθανή αλληλουχία λέξεων-αριθμών για κάθε ηχογραφημένο αρχείο.
HResult	Υπολογίζει την ακρίβεια μετάδοσης για κάθε πρόταση, δηλαδή το σύνολο λέξεων του αρχείου, και για κάθε λέξη.

Πίνακας 3.3: Εντολές των διαδικασιών εκπαίδευσης και ελέγχου

3.3.4 Υλοποίηση των baseline συντελεστών MFCC-FMP και εύρεση της ακρίβειας

Για την κατασκευή των 57 συντελεστών MFCC-FMP χρησιμοποιήσαμε, και πάλι, ως αρχεία εισόδου τα wav αρχεία της βάσης δεδομένων AURORA 3, που χρησιμοποιήσαμε και στην κατασκευή των συντελεστών MFCC και FMP. Ο στόχος μας είναι ο ίδιος με εκείνο της κατασκευής των MFCC και FMP συντελεστών, δηλαδή η μετατροπή καθενός από αυτά τα αρχεία σε έναν πίνακα που έχει ως γραμμές του τα πλαίσια και ως στήλες του τους 57 συντελεστές MFCC-FMP. Δηλαδή περιέχει όλους τους συντελεστές MFCC-FMP για κάθε πλαίσιο. Για να επιτευχθεί αυτό ακολουθήσαμε την ίδια διαδικασία με εκείνη της κατασκευής MFCC και FMP συντελεστών, προσαρμοσμένη, όμως, στην κατασκευή των MFCC-FMP συντελεστών. Στο παρακάτω σχήμα παρουσιάζεται ένα τμήμα του προσωρινού matlab αρχείου *tmpmat...m*, που περιέχει τα αρχεία εισόδου και τα αντίστοιχα αρχεία εξόδου για κάθε αρχείο .wav .

```
path(path,'/teras/speech/users/skourpas/AURORA3/MFCC-FMP/14_16');
spanish_mfcc_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030b1.0.wav','/speech/users/skourpas/AURORA3/MFCC-FMP/14_16/features/mfcc_par/high/ch0/v10030b1.0.mfc');
spanish_mfcc_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030c1.0.wav','/speech/users/skourpas/AURORA3/MFCC-FMP/14_16/features/mfcc_par/high/ch0/v10030c1.0.mfc');
spanish_mfcc_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030c3.0.wav','/speech/users/skourpas/AURORA3/MFCC-FMP/14_16/features/mfcc_par/high/ch0/v10030c3.0.mfc');
spanish_mfcc_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030c4.0.wav','/speech/users/skourpas/AURORA3/MFCC-FMP/14_16/features/mfcc_par/high/ch0/v10030c4.0.mfc');
spanish_mfcc_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i1.0.wav','/speech/users/skourpas/AURORA3/MFCC-FMP/14_16/features/mfcc_par/high/ch0/v10030i1.0.mfc');
spanish_mfcc_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i2.0.wav','/speech/users/skourpas/AURORA3/MFCC-FMP/14_16/features/mfcc_par/high/ch0/v10030i2.0.mfc');
spanish_mfcc_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i3.0.wav','/speech/users/skourpas/AURORA3/MFCC-FMP/14_16/features/mfcc_par/high/ch0/v10030i3.0.mfc');
spanish_mfcc_fmp('/speech/data/AURORA3/Spanish/disk1_3/speechdata/high/ch0/v10030i4.0.wav','/speech/users/skourpas/AURORA3/MFCC-FMP/14_16/features/mfcc_par/high/ch0/v10030i4.0.mfc');
```

Σχήμα 3.5: Προσωρινό matlab αρχείο για την κατασκευή των συντελεστών MFCC-FMP.

Στη συνέχεια υλοποιήσαμε σε κώδικα matlab το κάθε βήμα της θεωρίας εύρεσης των MFCC-FMP. Στην ουσία, πρώτα υπολογίσαμε τους συντελεστές MFCC και τους συντελεστές FMP, και έπειτα συνενώσαμε τα διανύσματα των συντελεστών αυτών, κατασκευάζοντας με αυτό τον τρόπο ένα νέο διάνυσμα με 57 συντελεστές, εκ των οποίων οι 39 πρώτοι είναι οι συντελεστές MFCC και οι επόμενοι 18 είναι οι συντελεστές FMP. Το νέο αυτό διάνυσμα είναι το διάνυσμα των συντελεστών MFCC-FMP, που όπως και στους συντελεστές MFCC και FMP, αποτελείται από 14 καταστάσεις με 16 Κανονικές Κατανομές (Gaussian Mixtures) για κάθε κατάσταση. Έπειτα, πραγματοποιούμε τις διαδικασίες εκπαίδευσης και ελέγχου για κάθε μια από τις καταστάσεις WM, MM και HM. Οι διαδικασίες αυτές είναι ίδιες με αυτές της περίπτωσης των MFCC συντελεστών, προσαρμοσμένες, βέβαια, σε διανύσματα 57 συντελεστών.

3.3.5 Υλοποίηση των multistream συντελεστών MFCC-FMP και εύρεση της ακρίβειας

Για την κατασκευή των multistream συντελεστών MFCC-FMP και την εύρεση της ακρίβειας, πραγματοποιούμε την ακόλουθη διαδικασία: Χρησιμοποιώντας τα ίδια αρχεία εισόδου που χρησιμοποιήσαμε και στην κατασκευή των συντελεστών MFCC, FMP και MFCC-FMP, κατασκευάζουμε, ξεχωριστά, τους 39 συντελεστές MFCC, τους 18 συντελεστές FMP και τους 57 συντελεστες MFCC-FMP. Στη συνέχεια, διεξάγουμε τη διαδικασία εκπαίδευσης για το καθένα από τα παρα-

πάνω front-ends και καταλήγουμε σε ένα HMM για το καθένα από αυτά. Έπειτα, κατασκευάζουμε ένα νέο HMM, που περιλαμβάνει τον ίδιο αριθμό καταστάσεων με τα HMMs των MFCC, FMP και MFCC-FMP, αλλά η κάθε του κατάσταση αποτελείται από δύο streams. Το πρώτο stream περιέχει τα διανύσματα της αντίστοιχης κατάστασης του MFCC HMM και το δεύτερο περιέχει τα διανύσματα της αντίστοιχης κατάστασης του FMP HMM. Δηλαδή, το πρώτο stream περιέχει τα διανύσματα μέσης τιμής και διακύμανσης για όλες τις Γκαουσιανές Κατανομές της αντίστοιχης κατάστασης του MFCC HMM και το δεύτερο stream περιέχει τα αντίστοιχα διανύσματα της αντίστοιχης κατάστασης του FMP HMM. Ο πίνακας συνδιακύμανσης της κάθε κατάστασης είναι αυτός του MFCC-FMP HMM για την αντίστοιχη κατάσταση. Επίσης, προσθέτουμε τιμές βάρους (weight) σε κάθε ένα από τα δύο streams, ώστε να μειώσουμε ή να αυξήσουμε τη συνεισφορά του στο HMM. Όταν έχουμε δημιουργήσει το νέο HMM, πραγματοποιούμε τη διαδικασία του ελέγχου, προκειμένου να βρούμε την ακρίβεια μετάδοσης και επαναλαμβάνουμε τη διαδικασία για κάθε μία από τις καταστάσεις WM, MM και HM.

3.4 Υλοποίηση του front-end SMAC-FMP

3.4.1 Εισαγωγή

Ακολουθήσαμε την ίδια λογική με την υλοποίηση του front-end MFCC-FMP , προσαρμοσμένη όμως στα χαρακτηριστικά $SMAC$ και FMP και στο κατάλληλο θεωρητικό υπόβαθρο γι΄ αυτά. Χρησιμοποιήσαμε ως αρχεία εισόδου τα δια wav αρχεία που χρησιμοποιήσαμε και για την κατασκευή του front-end MFCC-FMP. Επίσης, η διαδικασία μέχρι την κατασκευή του matlab αρχείου, που δημιουργεί τους συντελεστές που κάθε φορά επιθυμούμε, παραμένει η ίδια. Στη συνέχεια, παρατίθεται η υλοποίηση των συντελεστών SMAC, FMP και SMAC-FMP .

3.4.2 Υλοποίηση των συντελεστών SMAC και εύρεση της ακρίβειας

Για την κατασκευή των 42 συντελεστών SMAC ακολουθήσαμε τα παρακάτω βήματα: Αφού μετατρέψαμε το σήμα εισόδου σε ένα διάνυσμα με ακέραια ψηφία, πραγματοποιήσαμε προέμφαση, με συντελεστή 0.97. Έπειτα δημιουργήσαμε πλαίσια 200 δειγμάτων με επικάλυψη 100 δείγματα, δηλαδή 50% επικάλυψη. Για κάθε πλαίσιο, υπολογίσαμε το μέτρο του μετασχηματισμού Fourier και το φιλτράραμε με μια συστοιχία από 12 φίλτρα Gabor, με κεντρικές συχνότητες κατανεμημένες στην κλίμακα Mel. Στη συνέχεια, υπολογίζουμε την κανονικοποιημένη κεντρική φασματική ροπή πρώτης τάξης για κάθε ένα από τα 12 στοιχεία του πλαισίου. Έπειτα υπολογίζουμε τους φασματικούς συντελεστές μηδενικής και πρώτης τάξης, εκτελώντας τα ακόλουθα βήματα: υπολογίζουμε το λογάριθμο του κάθε πλαισίου και τον μετατρέπουμε στο πεδίο του χρόνου, με χρήση του Διακριτού Μετασχηματισμού Συνημιτόνου (DCT), από τον οποίο και θα προκύψουν δύο συντελεστές, που στη συνέχεια υπόκειται σε διαδικασία CMS. Οι συντελεστές αυτοί θα είναι οι C0 και C1, που είναι ο φασματικός συντελεστής μηδενικής και πρώτης τάξης, αντίστοιχα. Τοποθετούμε την κανονικοποιημένη κεντρική φασματική ροπή πρώτης τάξης και τους παραπάνω φασματικούς συντελεστές σε ένα διάνυσμα, το οποίο έχει 14 στοιχεία (οι 12 συντελεστές της κεντρικής φασματικής ροπής πρώτης τάξης και οι δύο φασματικοί συντελεστές). Ακολουθεί η εύρεση των συντελεστών Δέλτα

και Επιτάχυνσης, οι οποίοι προστίθενται στο υπάρχον διάνυσμα, με αποτέλεσμα το τελικό διάνυσμα να έχει 42 συντελεστές. Για την ακρίβεια, έχει 14 αρχικούς συντελεστές, 14 συντελεστές Δέλτα και 14 συντελεστές Επιτάχυνσης. Οι συναρτήσεις που περιλάμβάνει ο κώδικας, περιέχονται στον ακόλουθο πίνακα.

Όνομα συνάρτησης	**Λειτουργία συνάρτησης**
smac8_spanish	Υλοποιεί όλη τη διαδικασία δημιουργίας των SMAC και καλεί και τις υπόλοιπες συναρτήσεις.
gabor	Κατασκευάζει φίλτρο Gabor, με κεντρική συχνότητα κατανεμημένη στην κλίμακα Mel.
loadbin2	Διαβάζει τα δεδομένα από το αρχείο .wav και δημιουργεί ένα διάνυσμα με ακέραια ψηφία των δεκαέξι bits ($int16$).
mfb2cep	Πραγματοποιεί τη διαδικασία Διακριτού Μετασχηματισμού Συνημιτόνου (Discrete Cosine Transform - DCT).
append_deltas	Βρίσκει τους συντελεστές Δέλτα (Delta coefficients) και τους συντελεστές Επιτάχυνσης (Acceleration coefficients).
htk_write	Αποθηκεύει το διάνυσμα των συντελεστών SMAC στο δυαδικό αρχείο εξόδου.

Πίνακας 3.4: Συναρτήσεις της διαδικασίας κατασκευής SMAC συντελεστών

Στη συνέχεια, λαμβάνουν χώρα οι διαδικασίες εκπαίδευσης και ελέγχου για τις καταστάσεις WM, MM και HM. Η διαδικασίες αυτές είναι ίδιες με αυτές της περίπτωσης των MFCC συντελεστών του front-end MFCC-FMP, με τη διαφορά ότι τώρα έχουμε Κρυφά Μαρκοβιανά Μοντέλα με 16 καταστάσεις και τρία Gaussian mixtures για κάθε κατάσταση.

3.4.3 Υλοποίηση των συντελεστών FMP και εύρεση της ακρίβειας

Για την κατασκευή των 18 συντελεστών FMP ακολουθήσαμε την ίδια διαδικασία με αυτήν της κατασκευής των FMP συντελεστών στην περίπτωση του front-end MFCC-FMP, τόσο για το ποιά είναι τα αρχεία εισόδου, όσο και για τις συναρτήσεις και τις παραμέτρους στη διαδικασία κατασκευής των συντελεστών FMP. Επίσης, την ίδια διαδικασία με αυτήν των FMP συντελεστών της περίπτωσης του front-end MFCC-FMP ακολουθήσαμε και κατά τις διαδικασίες εκπαίδευσης και ελέγχου για τις καταστάσεις WM, MM και HM, με τη διαφορά, βέβαια, ότι στηριζόμαστε στα MLF αρχεία των SMAC συντελεστών και όχι των MFCC και έχουμε HMMs με 16 καταστάσεις και τρία Gaussian mixtures για κάθε κατάσταση.

3.4.4 Υλοποίηση των baseline συντελεστών SMAC-FMP και εύρεση της ακρίβειας

Την ίδια διαδικασία με αυτήν του front-end MFCC-FMP ακολουθήσαμε και στην περίπτωση της κατασκευής των 60 συντελεστών SMAC-FMP. Μόνο που

τώρα κατασκευάζουμε 42 SMAC συντελεστές και τους συνενώνουμε με τους 18 FMP συντελεστές, με αποτέλεσμα να φτιαχτεί ένα διάνυσμα 60 στοιχείων, κάθε ένα από τα οποία είναι ένας συντελεστής SMAC-FMP. Όσον αφορά τις διαδικασίες εκπαίδευσης και ελέγχου για τις καταστάσεις WM, MM και HM, είναι ίδιες με αυτές της περίπτωσης των SMAC συντελεστών, προσαρμοσμένες, όμως, σε διανύσματα 60 συντελεστών και όχι 42.

3.4.5 Υλοποίηση των multistream συντελεστών SMAC-FMP και εύρεση της ακρίβειας

Τόσο η κατασκευή των multistream συντελεστών SMAC-FMP, όσο και η εύρεση της ακρίβειας, για κάθε μία από τις καταστάσεις WM, MM και HM, περιλαμβάνει την ίδια διαδικασία με αυτήν της κατασκευής και της εύρεσης της ακρίβειας για τους multistream συντελεστές MFCC-FMP, προσαρμοσμένη όμως στα δεδομένα ότι έχουμε SMAC συντελεστές στη θέση των MFCC συντελεστών και πλέον κατασκευάζουμε HMMs με 16 καταστάσεις και τρία Gaussian mixtures για κάθε κατάσταση.

Κεφάλαιο 4

ΜΕΤΡΗΣΕΙΣ – ΠΕΙΡΑΜΑΤΑ

4.1 Εισαγωγή

Όπως έχουμε ήδη αναφέρει, το σήμα φωνής $s(t)$ μοντελοποιείται ως το άθροισμα Κ $AM-FM$ σημάτων $r(t)$, ένα για κάθε συντονισμό (formant), όπως δείχνει και ο παρακάτω μαθηματικός τύπος:

$$s(t) = \sum_{k=1}^{K} r_k(t) \tag{4.1}$$

Η ανάλυση του παραπάνω σήματος και η εκτίμηση των παραμέτρων του στιγμιαίου πλάτους ($a(t)$) και της στιγμιαίας συχνότητας ($f(t)$) του για κάθε formant, γίνεται με τη διαδικασία της αποδιαμορφωσης. Εφόσον, όμως, το σήμα έχει παραπάνω από μία $AM-FM$ συνιστώσες, πρέπει πρώτα να φιλτραριστεί σε διαφορετικές φασματικές ζώνες, και στη συνέχεια να γίνει αποδιαμόρφωση της κάθε ζώνης ξεχωριστά. Χρειάζεται, δηλαδή να γίνει πολυζωνική ανάλυση (multiband analysis) σε φασματικές ζώνες με φίλτρα Gabor σε κατάλληλη συστοιχία. Μπορούμε να πραγματοποιήσουμε πολλών ειδών μεθόδους αποδιαμόρφωσης. Στα πλαίσια της συγκεκριμένης εργασίας θα επικεντρωθούμε στις ακόλουθες δύο μεθόδους: Στην πρώτη μέθοδο, οι συστοιχίες των φίλτρων Gabor είναι σταθερών συχνοτήτων μέσα σε ένα επιθυμητό εύρος ζώνης, κατανεμημένες στην κλίμακα mel. Στη δεύτερη μέθοδο είναι μεταβλητών συχνοτήτων και ακολουθούν τις συχνότητες συντονισμού του φωνητικού σωλήνα [30].

Με βάση τις παραπάνω μεθόδους, πραγματοποιούμε μια σειρά πειραματικές μετρήσεις της ακρίβειας μετάδοσης, για κάθε κατηγορία front-end που περιγράψαμε στο προηγούμενο κεφάλαιο. Σκοπός των μετρήσεων είναι η εύρεση και βελτιστοποίηση της ακρίβειας μετάδοσης, ανάλογα, πρώτον με την τιμή που έχει το Median φίλτρο του FMP και, δεύτερον ανάλογα με τον τρόπο που διεξάγεται η διαδικασία της αποδιαμόρφωσης (demodulation) στο FMP. Ως εκ τούτου, τα πειράματα μπορούν να χωριστούν σε δύο βασικές κατηγορίες. Η πρώτη κατηγορία εξετάζει τη διακύμανση της ακρίβειας μετάδοσης, για μια σειρά τιμές του Median φίλτρου του FMP, όταν στο FMP λαμβάνει χώρα αποδιαμόρφωση με χρήση φίλτρων Gabor με συστοιχίες σταθερών συχνοτήτων. Η δεύτερη κατηγορία εξετάζει τη διακύμαν-

ση της ακρίβειας μετάδοσης, για μια σειρά τιμές του Median φίλτρου του FMP, όταν στο FMP λαμβάνει χώρα, για κάθε πλαίσιο, αποδιαμόρφωση με φίλτρα Gabor μεταβλητών συχνοτητων. Επίσης, εφόσον κατασκευάσαμε multistream συντελεστές, πειραματιζόμαστε πάνω στις τιμές των βαρών του κάθε stream, ώστε να εξετάσουμε πώς τα βάρη επηρρεάζουν την απόδοση, αλλά και ποιός συνδυασμός βαρών είναι ο καταλληλότερος. Διεξάγαμε πειράματα πάνω στα ακόλουθα ζεύγη βαρών:

Βάρος πρώτου stream	**Βάρος δεύτερου stream**
1.0	0.1
0.8	0.2
1.0	1.0
1.0	0.2
0.5	0.5

Πίνακας 4.1: Ζεύγη βαρών για το κάθε multistream

Όπου στην περίπτωση των MFCC-FMP front-ends το πρώτο νούμερο του κάθε ζεύγους δίνεται στο MFCC stream και το δεύτερο στο FMP stream, ενώ στην περίπτωση των SMAC-FMP front-ends το πρώτο δίνεται στο SMAC stream και το δεύτερο στο FMP stream. Επειδή όμως, και στις δύο περιπτώσεις, τα καλύτερα αποτελέσματα προέκυπταν για τα ζεύγη 1.0-0.1 και 0.8-0.2, σε όσα αποτελέσματα παραθέτουμε στο κεφάλαιο αυτό και περιλαμβάνουν ζεύγη βαρών, καταγράφουμε τα αποτελέσματα μόνο γι΄ αυτά τα ζεύγη. Σε αυτό το κεφάλαιο θα παρουσιάσουμε τα πειραματικά μας αποτελέσματα για κάθε μια από τις παραπάνω κατηγορίες πειραμάτων, για τους συνδυασμούς βαρών και για κάθε μία από τις υλοποιήσεις των front-ends, ξεκινώντας από την περίπτωση των MFCC-FMP front-ends και συνεχίζοντας με την περίπτωση των SMAC-FMP front-ends. Τα αποτελέσματα αφορούν και τις τρεις καταστάσεις WM, MM και HM.

4.2 Αποτελέσματα του MFCC-FMP front-end

Για την κατασκευή των MFCC-FMP front-ends, κατασκευάσαμε, αρχικά, τα MFCC front-ends, έπειτα τα FMP front-ends, μετά τα baseline MFCC-FMP front-ends και, τέλος, τα multistream MFCC-FMP front-ends. Τα αποτελέσματα της ακρίβειας μετάδοσης (%) του MFCC front-end για τις τρεις καταστάσεις WM, MM, HM, παρουσιάζονται στον παρακάτω πίνακα:

WM Accuracy (%)	**MM Accuracy (%)**	**HM Accuracy (%)**
92.44	84.26	56.81

Πίνακας 4.2: Αποτελέσματα για το MFCC

Στη συνέχεια, παρουσιάζουμε τα αποτελέσματα της ακρίβειας μετάδοσης για τα FMP front-ends, τα baseline MFCC-FMP front-ends και τα multistream MFCC-FMP front-ends, για κάθε μία από τις δύο μεθόδους αποδιαμόρφωσης.

4.2.1 Αποτελέσματα για την περίπτωση που έχουμε αποδιαμόρφωση με σταθερές συχνότητες

Για την περίπτωση που πραγματοποιούμε αποδιαμόρφωση με χρήση φίλτρων Gabor με συστοιχίες σταθερών συχνοτήτων κατανεμημένες στην κλίμακα mel, οι δύο παρακάτω πίνακες δείχνουν τα πειραματικά αποτελέσματα για τα front-ends FMP και baseline MFCC-FMP, αντίστοιχα.

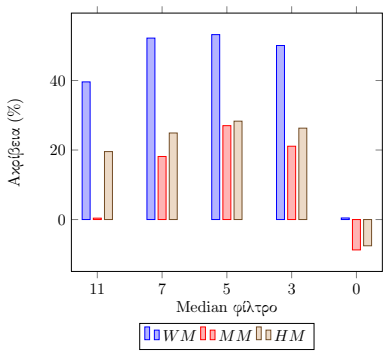
Median φίλτρο	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	39.60	0.40	19.55
7	52.20	18.14	24.90
5	53.20	27.01	28.30
3	50.07	21.09	26.29
0	0.45	-8.74	-7.55

Πίνακας 4.3: Αποτελέσματα για το FMP

Median φίλτρο	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	92.27	84.13	56.06
7	91.43	80.30	56.27
5	91.98	81.00	54.62
3	91.00	77.79	57.44
0	87.92	73.61	58.86

Πίνακας 4.4: Αποτελέσματα για το baseline MFCC-FMP

Ενώ τα αποτελεσματα του πίνακα για τα FMP front-ends παρουσιάζονται παρακάτω με τη μορφή σχεδιαγράμματος:



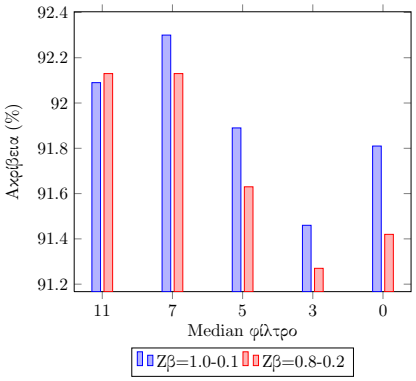
Σχήμα 4.1: FMP, για τις καταστάσεις WM, MM, HM

Ο παρακάτω πίνακας αποτυπώνει τα πειραματικά αποτελέσματα για το multi-stream MFCC-FMP, όπου, όμως, παίρνουμε αποτελέσματα για δύο διαφορετικά ζεύγη τιμών βαρών στα streams. Στην πρώτη περίπτωση έχουμε δώσει την τιμή βάρους 1.0 στο MFCC stream και 0.1 στο FMP stream. Στην δεύτερη περίπτωση έχουμε δώσει την τιμή 0.8 στο MFCC stream και 0.2 στο FMP stream.

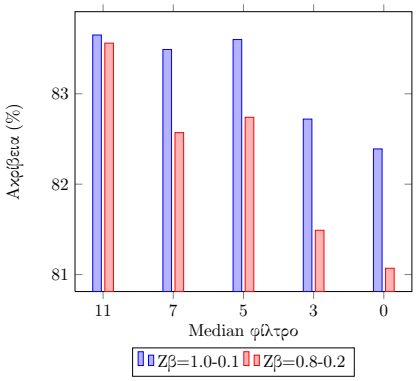
Median φίλτρο	**Weights**	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	1.0 - 0.1	92.09	83.65	55.22
11	0.8 - 0.2	92.13	83.56	51.25
7	1.0 - 0.1	92.30	83.49	56.12
7	0.8 - 0.2	92.13	82.57	54.86
5	1.0 - 0.1	91.89	83.60	56.06
5	0.8 - 0.2	91.63	82.74	54.86
3	1.0 - 0.1	91.46	82.72	55.73
3	0.8 - 0.2	91.27	81.49	54.11
0	1.0 - 0.1	91.81	82.39	56.36
0	0.8 - 0.2	91.42	81.07	55.31

Πίνακας 4.5: Αποτελέσματα για το multistream MFCC-FMP

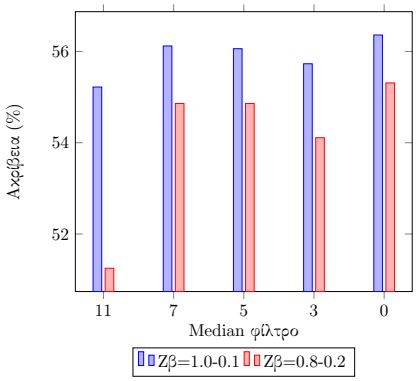
Τα παραπάνω αποτελέσματα, για κάθε μια από τις καταστάσεις WM, MM, HM, αποτυπώνονται παρακάτω με τη μορφή σχεδιαγραμμάτων:



(α′) Κατάσταση WM



(β′) Κατάσταση MM



(γ′) Κατάσταση HM

Σχήμα 4.2: Multistream MFCC-FMP, για τα ζεύγη βαρών (Zβ)

Από τους παραπάνω πίνακες και σχεδιαγράμματα, προκύπτει ότι: Για την περίπτωση του FMP front-end η καλύτερη απόδοση συμβαίνει στην περίπτωση που έχουμε δώσει στο Median φίλτρο την τιμή 5. Στο multistream MFCC-FMP τα αποτελέσματα για το ζεύγος βαρών 1.0-0.1 είναι καλύτερα από αυτά του ζεύγους 0.8-0.2, εκτός από τα αποτελέσματα της κατάστασης WM στην περίπτωση που έχουμε τιμή Median φίλτρου το 11. Αλλά και τότε η διαφορά τους είναι μόλις 0.04, και εφόσον αυτό το φαινόμενο εμφανίζεται μόνο μια φορά θεωρούμε ότι δεν χρειαζεται να ληφθεί υπόψη. Οπότε θα επικεντρωθούμε στη σύγκριση των αποτελεσμάτων μόνο για το 1.0-0.1. Εκεί παρατηρούμε ότι δεν έχουμε μια καλύτερη περίπτωση για όλες τις καταστάσεις ταυτόχρονα. Για την ακρίβεια, η καλύτερη απόδοση για την κατάσταση WM συμβαίνει όταν το Median φίλτρο έχει την τιμή 7, για την κατάσταση MM όταν έχει την τιμή 11 και για την κατάσταση HM την τιμή 0. Όμως, σε γενικές γραμμές, τα καλύτερα αποτελέσματα προκύπτουν όταν το φίλτρο παίρνει την τιμή 7, και αυτό για τους ακόλουθους λόγους: Πρώτον, τότε έχουμε τη βέλτιστη ακρίβεια μετάδοσης για την κατάσταση WM. Δεύτερον, τα αποτελέσματα για τις καταστάσεις MM και HM είναι, αντιστοίχως, τα τρίτα και δεύτερα καλύτερα, αλλά όχι με μεγάλη διαφορά από τα πρώτα. Τρίτον, είναι η μόνη περίπτωση μεταξύ των τριών περιπτώσεων που συγκρίνουμε, που έχουμε την μεγαλύτερη απόδοση σε δύο από τις τρεις καταστάσεις ταυτόχρονα, με όποια

από τις άλλες περιπτώσεις τη συγκρίνουμε. Συγκεκριμένα, αν τη συγκρίνουμε με την περίπτωση που το φίλτρο έχει την τιμή 11, έχει καλύτερη απόδοση στις καταστάσεις WM και HM όταν το μέγεθος του Median φίλτρου είναι ίσο με 7, από ότι όταν είναι ίσο με 11, και μόνο στην κατάσταση MM ισχύει το αντίθετο. Αν τη συγκρίνουμε με την περίπτωση που το φίλτρο έχει την τιμή 0, έχει καλύτερη απόδοση για τις καταστάσεις WM και MM όταν το μέγεθος του Median φίλτρου είναι ίσο με 7, από ότι όταν είναι ίσο με 0, και μόνο στην κατάσταση HM ισχύει το αντίθετο.

4.2.2 Αποτελέσματα για την περίπτωση που έχουμε αποδιαμόρφωση με χρήση φάσματος μεταβλητών συχνοτήτων

Στον παρόν υποκεφάλαιο, μελετάμε τις τιμές της ακρίβειας μετάδοσης, όταν η αποδιαμόρφωση γίνεται με χρήση φίλτρων Gabor μεταβλητών συχνοτητων. Για την ακρίβεια, πειραματιζόμαστε πάνω στις αλλαγές της ακρίβειας που προκαλούνται από τις αλλαγές στο φάσμα μεταβλητών συχνοτήτων των συστοιχιών των φίλτρων Gabor. Επειδή, παράλληλα, πειραματιζόμαστε και βάση του μεγέθους του Median φίλτρου, αλλά και του ζεύγους βαρών για τα streams, μας ενδιαφέρει να βρούμε τη βέλτιστη τριάδα φάσματος συχνοτήτων, μεγέθους φίλτρου και ζεύγους βαρών. Για να το πετύχουμε αυτό, πειραματιζόμαστε αρχικά πάνω στις τιμές της ακρίβειας που προκύπτουν για διάφορες τιμές μεγέθους φίλτρου, όταν έχουμε πάρει ένα συγκεκριμένο φάσμα μεταβλητών συχνοτήτων. Στην περίπτωσή μας, το φάσμα αυτό παίρνει τιμές στο διάστημα από 68.8 Hz έως 368.8 Hz, διότι θεωρούμε ότι οι συχνότητες του φάσματος μεταβάλλονται στο διάστημα -150 έως 150 Hz. Βρίσκουμε, λοιπόν, το βέλτιστο μέγεθος φίλτρου και μετά, κρατώντας αυτό το μέγεθος φίλτρου, βρίσκουμε τις τιμές ακρίβειας για διάφορα φάσματα μεταβλητών συχνοτήτων και ζευγών βαρών. Το φάσμα που θα μας δώσει τη βέλτιστη τιμή ακρίβειας, μαζί με το βέλτιστο μέγεθος φίλτρου και το κατάλληλο ζεύγος βαρών, είναι η επιθυμητή τριάδα παραμέτρων.

Στα παρακάτω δύο σχήματα αποτυπώνονται οι τιμές της ακρίβειας μετάδοσης για τις καταστάσεις WM, MM και HM για τα FMP front-ends και τα baseline MFCC-FMP front-ends, ανάλογα με την τιμή του Median φίλτρου.

Median φίλτρο	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	17.19	-14.22	18.95
7	49.71	20.05	31.1
5	57.96	29.08	33.77
3	65.16	34.23	37.23
0	60.64	29.94	31.49

Πίνακας 4.6: Αποτελέσματα για το FMP

Median φίλτρο	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	91.31	83.80	57.77
7	91.17	84.15	56.81
5	91.20	84.50	58.80
3	90.55	84.22	58.20
0	90.07	82.98	59.07

Πίνακας 4.7: Αποτελέσματα για το baseline MFCC-FMP

Ο παρακάτω πίνακας δείχνει τα πειραματικά αποτελέσματα για το multistream MFCC-FMP, όπου, όμως, παίρνουμε αποτελέσματα για δύο διαφορετικά ζεύγη τιμών βαρών στα streams, τα ίδια με εκείνα του αντίστοιχου πίνακα της περίπτωσης που έχουμε αποδιαμόρφωση με σταθερές συχνότητες.

Median φίλτρο	**Weights**	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	1.0 - 0.1	91.94	84.77	54.32
11	0.8 - 0.2	89.44	80.15	48.81
7	1.0 - 0.1	92.61	85.01	55.49
7	0.8 - 0.2	92.78	85.47	53.02
5	1.0 - 0.1	92.60	84.90	56.15
5	0.8 - 0.2	92.87	85.45	53.83
3	1.0 - 0.1	92.55	84.39	55.91
3	0.8 - 0.2	92.84	84.42	54.44
0	1.0 - 0.1	92.56	84.13	56.15
0	0.8 - 0.2	92.61	83.93	54.29

Πίνακας 4.8: Αποτελέσματα για το multistream MFCC-FMP

Παρατηρούμε ότι, με βάση τον πίνακα των multistream MFCC-FMP front-ends, η, σε γενικές γραμμές, καλύτερη απόδοση προκύπτει όταν έχουμε μέγεθος του Median φίλτρου ίσο με 5 και ζεύγος βαρών 0.8-0.2. Και αυτό διότι: Στην περίπτωση αυτή έχουμε τη βέλτιστη απόδοση για την κατάσταση WM και τη δεύτερη καλύτερη για την MM, με διαφορά από την πρώτη, που είναι για μέγεθος φίλτρου 7, μόνο 0.02. Βέβαια, δεν έχουμε την καλύτερη απόδοση για την κατάσταση HM, αλλά και πάλι ο καλύτερος συνδυασμός των αποτελεσμάτων των τριών καταστάσεων παραμένει αυτός, αφού για οποιοδήποτε άλλο συνδυασμό περιέχει μέγιστη τιμή, δηλαδή για μέγεθος φίλτρου 7 ή 0, δεν πετυχαίνουμε παρά μόνο μία από τις τρεις τιμές μέγιστη και τις άλλες δύο να διαφέρουν αρκετά από τη μέγιστη τιμή τους. Βέβαια, η καλύτερη απόδοση για την HM συμβαίνει και πάλι όταν έχουμε μέγεθος φίλτρου το 5, αλλά με άλλο ζεύγος βαρών, το ζεύγος 1.0 -0.1. Επιπλέον, η καλύτερη απόδοση, με βάση τον πίνακα των FMP front-ends, προκύπτει όταν το μέγεθος του Median φίλτρου είναι ίσο με 3. Θα πειραματιστούμε, λοιπόν, πάνω στα φάσματα μεταβλητών συχνοτήτων των συστοιχιών των Gabor φίλτρων, για τα ίδια front-ends, παίρνοντας πρώτα το μέγεθος του Median φίλτρου ίσο με 5 και έπειτα ίσο με 3. Δηλαδή, πειραματιζόμαστε πάνω στη μετακίνηση του φάσματος μεταβλητών συχνοτήτων κατά συγκεκριμένα Hz. Στους επόμενους δύο πίνακες φαίνονται τα αποτελέσματα για τις τιμές της ακρίβειας, όσον αφορά τα FMP front-ends και τα baseline MFCC-FMP front-ends, για την περίπτωση που το μέγεθος του Median φίλτρου είναι ίσο με 5. Η πρώτη στήλη των πινάκων δείχνει το κατά

πόσα Hz μετακινήθηκε το φάσμα, τόσο προς την αριστερή όσο και προς τη δεξιά πλευρά του άξονα συχνοτήτων.

Μετακίνηση (Hz)	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	64.40	38.01	34.86
-90:90	64.45	31.32	35.19
-150:150	57.96	29.06	33.77
-210:210	47.14	25.80	31.76
-270:270	43.21	14.86	24.54

Πίνακας 4.9: Αποτελέσματα για το FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 5

Μετακίνηση (Hz)	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	91.19	82.72	59.28
-90:90	90.96	84.20	59.07
-150:150	91.20	84.50	58.80
-210:210	91.58	84.35	58.68
-270:270	90.34	85.14	56.99

Πίνακας 4.10: Αποτελέσματα για το baseline MFCC-FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 5

Έχοντας βρει τα αποτελέσματα για τα FMP και baseline MFCC-FMP front-ends, εξετάζουμε, τώρα, την ακρίβεια, για την περίπτωση των multistream MFCC-FMP front-ends, δίνοντας ως τιμές βαρών των streams τις ίδιες που δώσαμε και στην παραπάνω περιπτώση multistream MFCC-FMP front-end.

Μετακίνηση (Hz)	**Weights**	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	1.0 - 0.1	92.58	84.39	55.70
-30:30	0.8 - 0.2	92.78	84.39	54.26
-90:90	1.0 - 0.1	92.66	84.39	55.97
-90:90	0.8 - 0.2	92.99	84.37	53.92
-150:150	1.0 - 0.1	92.60	84.90	56.15
-150:150	0.8 - 0.2	92.87	85.45	53.83
-210:210	1.0 - 0.1	92.50	84.81	55.91
-210:210	0.8 - 0.2	92.78	84.75	54.02
-270:270	1.0 - 0.1	92.70	84.81	56.48
-270:270	0.8 - 0.2	92.64	84.61	54.32

Πίνακας 4.11: Αποτελέσματα για το multistream MFCC-FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 5

Στους επόμενους τρεις πίνακες αποτυπώνονται τα αντίστοιχα, με τους παραπάνω τρεις πίνακες αποτελέσματα, όσον αφορά τα FMP front-ends, τα baseline MFCC-FMP front-ends και τα multistream MFCC-FMP front-ends, για την περίπτωση που το μέγεθος του Median φίλτρου είναι ίσο με 3.

Μετακίνηση (Hz)	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	62.67	34.18	33.71
-90:90	64.65	38.01	38.65
-150:150	65.16	34.23	37.23
-210:210	54.73	26.96	35.85
-270:270	46.52	19.83	26.98

Πίνακας 4.12: Αποτελέσματα για το FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 3

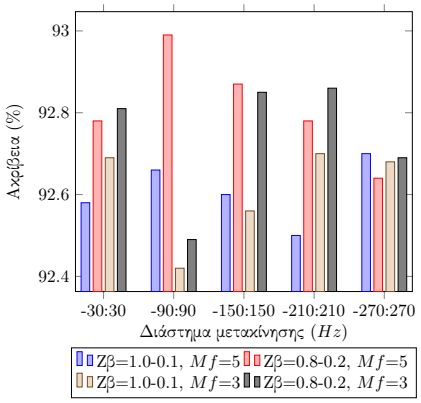
Μετακίνηση (Hz)	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	91.42	82.50	60.66
-90:90	90.95	83.60	59.70
-150:150	90.50	84.22	58.20
-210:210	90.18	83.78	57.71
-270:270	90.33	84.57	59.46

Πίνακας 4.13: Αποτελέσματα για το baseline MFCC-FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 3

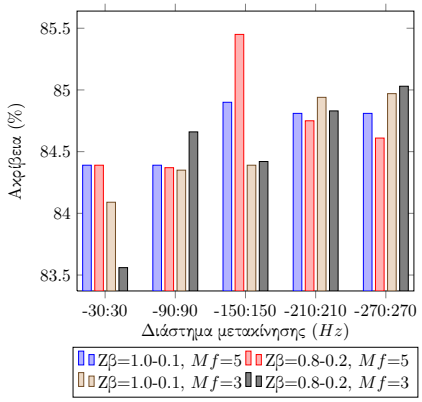
Μετακίνηση (Hz)	**Weights**	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	1.0 - 0.1	92.69	84.09	55.79
-30:30	0.8 - 0.2	92.81	83.56	53.74
-90:90	1.0 - 0.1	92.42	84.35	56.09
-90:90	0.8 - 0.2	92.49	84.66	54.62
-150:150	1.0 - 0.1	92.56	84.39	55.91
-150:150	0.8 - 0.2	92.85	84.42	54.44
-210:210	1.0 - 0.1	92.70	84.94	55.91
-210:210	0.8 - 0.2	92.86	84.83	53.89
-270:270	1.0 - 0.1	92.68	84.97	56.00
-270:270	0.8 - 0.2	92.69	85.03	53.92

Πίνακας 4.14: Αποτελέσματα για το multistream MFCC-FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 3

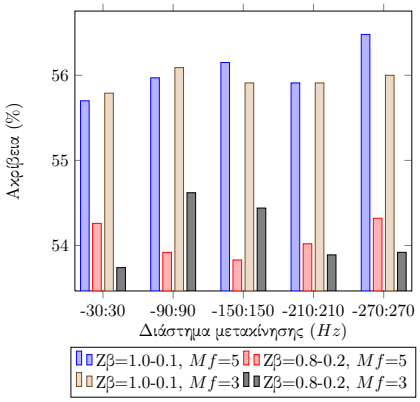
Στη συνέχεια παρουσιάζουμε σε διαγράμματα τα αποτελέσματα των multistream MFCC-FMP front-ends, τόσο για μέγεθος Median φίλτρου το 5, όσο και το 3, ώστε να συγκρίνουμε γραφικά τα αποτελέσματα. Το κάθε διάγραμμα παρουσιάζει τα αποτελέσματα για μια απο τις καταστάσεις WM, MM, HM.



(α′) Κατάσταση WM



(β′) Κατάσταση MM



(γ′) Κατάσταση HM

Σχήμα 4.3: Multistream MFCC-FMP, ανάλογα με το ζεύγος βαρών (Zβ) και το Median φίλτρο (Mf)

Από τα αποτελέσματα των δύο πινάκων για τα multistream MFCC-FMP front-ends, αλλά και με τη βοήθεια των παραπάνω γραφημάτων, παρατηρούμε ότι τα αποτελεματα του πίνακα για τιμή Median φίλτρου το 5 υπερτερούν σε σχέση με αυτά για τιμή φίλτρου το 3, διότι: Παρόλο που σε κάποιες περιπτώσεις παρουσιάζει καλύτερα αποτελέσματα η περίπτωση για μέγεθος φίλτρου το 5 και σε άλλες για μέγεθος το 3, όταν έχουμε την τιμή 5 παρουσιάζονται πιο πολλές φορές καλύτερες τιμές για τις ίδιες συνθήκες, για την ακρίβεια 17 έναντι 13 όταν έχουμε την τιμή 3. Επιπλέον, τότε εμφανίζονται περισσότερες φορές καλύτερες τιμές σε δύο καταστάσεις ταυτόχρονα, και μια φορά και στις τρεις καταστάσεις ταυτόχρονα. Ακόμη, η μέγιστη τιμή για κάθε μια από τις τρεις καταστάσεις WM, MM και HM εμφανίζεται τότε. Επιλέγουμε ως σύνολο παραμέτρων, το μέγεθος του Median φίλτρου να είναι ίσο με 5, το διάστημα μετακίνησης να είναι -150:150 Hz και το ζεύγος βαρών 0.8-0.2. Σε αυτό το συμπέρασμα καταλήγουμε διότι: Πρώτον, τότε έχουμε την μέγιστη τιμή για την κατάσταση MM. Δεύτερον, έχουμε τη δεύτερη μεγαλύτερη τιμή της κατάστασης WM, τόσο όταν το φίλτρο έχει τιμή 5, όσο και για το σύνολο των δύο πινάκων. Δηλαδή, καμία τιμή αυτής της κατάστασης, στον πίνακα για τιμή φίλτρου το 3, δεν την ξεπερνά. Τρίτον, παρόλο που δεν έχουμε

την καλύτερη τιμή για την κατάσταση HM, δεν υπάρχει συνδυασμός των τριών καταστάσεων που να είναι προτιμότερος, σε κανέναν από τους δύο πίνακες, διότι οποιοσδήποτε άλλος συνδυασμός δεν δίνει τις τιμές για δύο από τις τρεις καταστάσεις να είναι ταυτόχρονα τόσο κοντά στις μέγιστες τιμές τους ή η τιμή της μιας κατάστασης να ταυτίζεται με τη μέγιστη τιμή της και της άλλης να είναι τόσο κοντά σε αυτήν. Τέλος, συγκρίνοντας την τριάδα τιμών ακρίβειας που επιλέχτηκε ως βέλτιστη παραπάνω, με την βέλτιστη τριάδα όταν έχουμε multistream MFCC-FMP front-ends με φάσμα σταθερών συχνοτήτων, που βρήκαμε στο προηγούμενο υποκεφάλαιο, παρατηρούμε ότι, για τις καταστάσεις WM και MM προκύπτουν καλύτερα αποτελέσματα όταν χρησιμοποιούμε το φάσμα μεταβλητών συχνοτήτων και μόνο στην κατάσταση HM ισχύει το αντίθετο. Οπότε, στη διαδικασία αναγνώρισης σήματος φωνής με χρήση multistream MFCC-FMP front-ends είναι προτιμότερη η χρήση φάσματος μεταβλητών συχνοτήτων στα φίλτρα Gabor με ταυτόχρονη χρήση μέγεθους του Median φίλτρου το 5, διάστημα μετακίνησης το -150:150 Hz και ζεύγος βαρών το 0.8-0.2.

4.3 Αποτελέσματα του SMAC-FMP front-end

Αφού εξετάσαμε τα αποτελέσματα της ακρίβειας για την περίπτωση των MFCC-FMP front-ends, συνεχίζουμε με την περίπτωση των SMAC-FMP front-ends. Από την κατασκευή, τη διαδικασία εκπαίδευσης και τη διαδικασία ελέγχου των SMAC front-ends, με τον τρόπο που περιγράψαμε σε παραπάνω κεφάλαιο, προέκυψαν τα ακόλουθα αποτελέσματα για την ακρίβεια μετάδοσης (%), στις τρεις καταστάσεις WM, MM και HM.

WM Accuracy (%)	**MM Accuracy (%)**	**HM Accuracy (%)**
94.25	89.21	77.68

Πίνακας 4.15: Αποτελέσματα για το SMAC

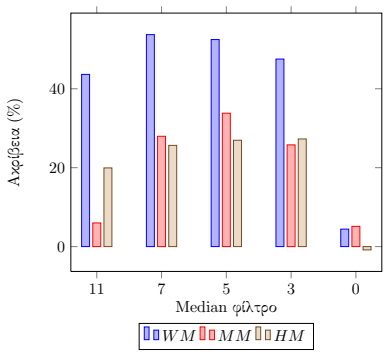
Έχοντας βρει τα αποτελέσματα της ακρίβειας για την περίπτωση των SMAC front-ends, συνεχίζουμε, κατ΄ αναλογία με την περίπτωση των MFCC-FMP front-ends, στην εύρεση της ακρίβειας μετάδοσης για τα FMP front-ends, τα baseline SMAC-FMP front-ends και τα multistream SMAC-FMP front-ends, για τις ίδιες περιπτώσεις αποδιαμόρφωσης, μεγέθους του Median φίλτρου και βαρών των streams, που χρησιμοποιούμε και στα MFCC-FMP front-ends.

4.3.1 Αποτελέσματα για την περίπτωση που έχουμε αποδιαμόρφωση με σταθερές συχνότητες

Ξεκινάμε, όπως και στην περίπτωση των MFCC-FMP front-ends, από την περίπτωση που έχουμε αποδιαμόρφωση όπου το φάσμα συχνοτήτων των συστοιχιών των Gabor φίλτρων είναι σταθερό. Τα αποτελέσματα της ακρίβειας για την περίπτωση των FMP front-ends, για διάφορες τιμές μεγέθους του Median φίλτρου, αποτυπωμένα σε πίνακα και σε διάγραμμα, είναι τα ακόλουθα:

Median φίλτρο	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	43.64	5.99	19.94
7	53.72	27.96	25.65
5	52.47	33.81	26.95
3	47.53	25.78	27.28
0	4.43	5.13	-0.84

Πίνακας 4.16: Αποτελέσματα για το FMP



Σχήμα 4.4: FMP, για τις καταστάσεις WM, MM, HM

Ενώ, τα αποτελέσματα για την περίπτωση των baseline SMAC-FMP front-ends είναι τα ακόλουθα:

Median φίλτρο	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	92.42	84.77	74.02
7	91.06	83.14	73.41
5	91.19	81.80	77.53
3	90.38	79.07	71.52
0	87.92	76.07	73.50

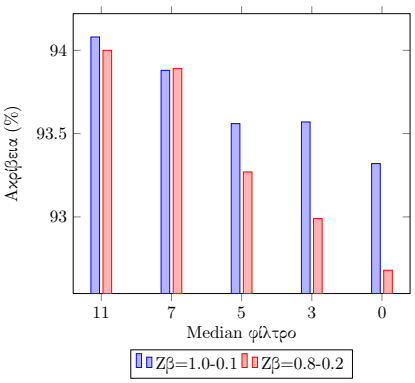
Πίνακας 4.17: Αποτελέσματα για το baseline SMAC-FMP

Συνεχίζουμε με την εύρεση της ακρίβειας για την περίπτωση των multistream SMAC-FMP front-ends, ανάλογα με μέγεθος του Median φίλτρου, σε συνδυασμό με το βάρος του κάθε stream.

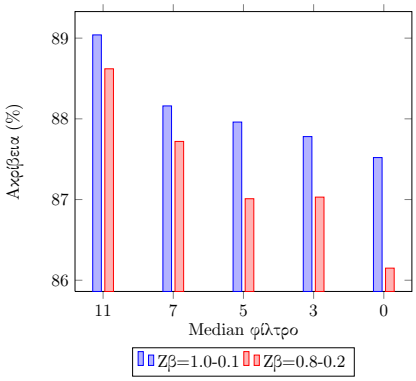
Median φίλτρο	**Weights**	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	1.0 - 0.1	94.08	89.04	76.90
11	0.8 - 0.2	94.00	88.62	75.94
7	1.0 - 0.1	93.88	88.16	76.99
7	0.8 - 0.2	93.89	87.72	76.51
5	1.0 - 0.1	93.56	87.96	76.78
5	0.8 - 0.2	93.27	87.01	76.27
3	1.0 - 0.1	93.57	87.78	77.02
3	0.8 - 0.2	92.99	87.03	76.45
0	1.0 - 0.1	93.32	87.52	76.36
0	0.8 - 0.2	92.68	86.15	75.31

Πίνακας 4.18: Αποτελέσματα για το multistream SMAC-FMP

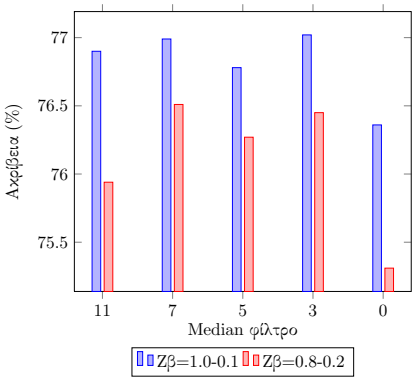
Στη συνέχεια παραθέτουμε τα αποτελέσματα του παραπάνω πίνακα σε σχεδιαγράμματα, ανάλογα με την κατάσταση WM, MM, HM.



(α′) Κατάσταση WM



(β′) Κατάσταση MM



(γ′) Κατάσταση HM

Σχήμα 4.5: Multistream SMAC-FMP, για τα ζεύγη βαρών (Zβ)

Από τον προηγούμενο πίνακα και τα σχεδιαγράμματα, προκύπτουν τα ακόλουθα συμπεράσματα: Τα αποτελέσματα για το ζεύγος βαρών 1.0-0.1 είναι καλύτερα από αυτά του ζεύγους 0.8-0.2. Εκτός από τη περίπτωση όπου η τιμή του Median φίλτρου είναι 7 και η κατάσταση είναι WM. Και εκεί όμως διαφέρουν μόνο κατά 0.01, και σε συνδυασμό με το ότι είναι μόνο μία περίπτωση, μας οδηγεί να μην τη λάβουμε υπόψη. Επικεντρωνόμαστε, λοιπόν, στα ζεύγη 1.0-0.1. Από αυτά, τα καλύτερα αποτελέσματα, για τις καταστάσεις WM και MM, προκύπτουν όταν το Median φίλτρο παίρνει την τιμή 11, ενώ για την κατάσταση HM όταν παίρνει την τιμή 3. Όμως, τα τρίτα καλύτερα αποτελέσματα για την HM προέρχονται όταν η τιμή του Median φίλτρου είναι 11, και μάλιστα δεν διαφέρουν πολύ από τα πρώτα. Οπότε μπορούμε να πούμε ότι τα καλύτερα αποτελέσματα ακρίβειας προκύπτουν όταν το φίλτρο παίρνει την τιμή 11 και το ζεύγος βαρών είναι 1.0-0.1.

4.3.2 Αποτελέσματα για την περίπτωση που έχουμε αποδιαμόρφωση με χρήση μεταβλητών συχνοτήτων

Σε αυτό το υποκεφάλαιο πειραματιστήκαμε πάνω στις αλλαγές που προκαλούν στην ακρίβεια, το φάσμα μεταβλητών συχνοτήτων των συστοιχιών των φίλτρων

Gabor, σε συνδυασμό με το μεγέθος του Median φίλτρου, και την εύρεση του βέλτιστου συνδυασμού. Για το σκοπό αυτό ακολουθήσαμε την ίδια διαδικασία με αυτήν του αντίστοιχου υποκεφαλαίου των MFCC-FMP front-ends. Τα παρακάτω δύο σχήματα περιέχουν τις τιμές της ακρίβειας μετάδοσης για τις καταστάσεις WM, MM και HM, για τα FMP front-ends και τα baseline SMAC-FMP front-ends, ανάλογα με την τιμή του Median φίλτρου, κρατώντας το ίδιο φάσμα μεταβλητών συχνοτήτων με αυτό που είχαμε κρατήσει στην αντίστοιχη περίπτωση των MFCC-FMP front-ends, δηλαδή το διάστημα από από 68.8 Hz έως 368.8 Hz, διότι θεωρούμε ότι οι συχνότητες του φάσματος μεταβάλλονται στο διάστημα -150 έως 150 Hz.

Median φίλτρο	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	23.62	0.90	23.91
7	55.78	30.46	31.58
5	61.95	36.30	35.67
3	65.84	38.83	37.65
0	58.68	31.41	25.23

Πίνακας 4.19: Αποτελέσματα για το FMP

Median φίλτρο	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	93.63	87.01	74.20
7	93.67	87.17	76.36
5	93.47	86.79	76.66
3	93.19	85.89	76.81
0	92.20	84.35	77.02

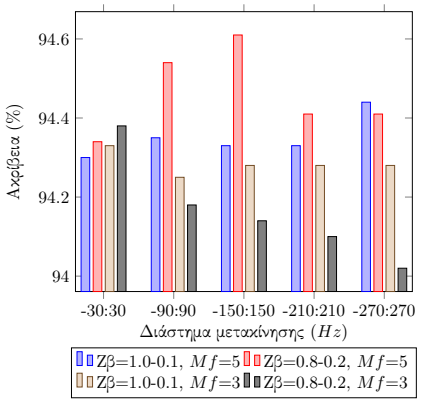
Πίνακας 4.20: Αποτελέσματα για το baseline SMAC-FMP

Κατ΄ αναλογία με την περίπτωση των MFCC-FMP front-ends, προκύπτουν τα ακόλουθα αποτελέσματα για το multistream SMAC-FMP, όπου συνδυάζουμε το μέγεθος του Median φίλτρου με τα βάρη των streams, για τις ίδιες τιμές που δίνουμε και στην περίπτωση των MFCC-FMP front-ends.

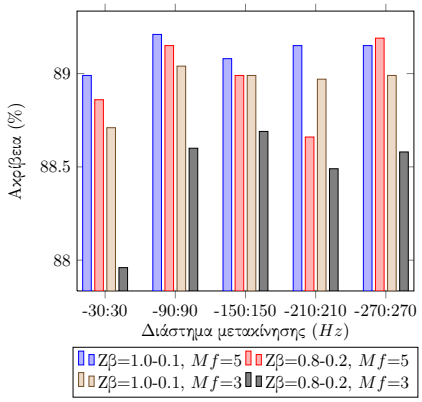
Median φίλτρο	**Weights**	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
11	1.0 - 0.1	94.33	89.08	77.05
11	0.8 - 0.2	94.61	88.99	75.85
7	1.0 - 0.1	94.33	88.80	77.29
7	0.8 - 0.2	94.38	88.71	76.36
5	1.0 - 0.1	94.18	89.04	77.41
5	0.8 - 0.2	93.97	88.66	76.84
3	1.0 - 0.1	94.28	88.99	77.32
3	0.8 - 0.2	94.14	88.69	76.87
0	1.0 - 0.1	94.29	88.84	77.20
0	0.8 - 0.2	94.43	88.64	76.66

Πίνακας 4.21: Αποτελέσματα για το multistream SMAC-FMP

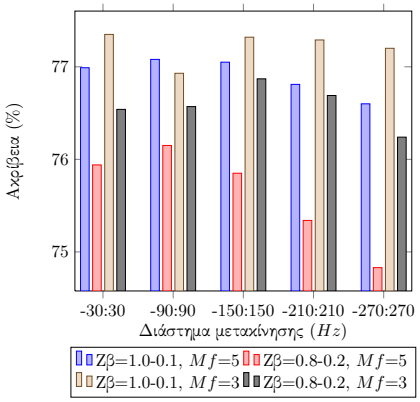
Ενώ τα αντίστοιχα σχεδιαγράμματα για κάθε κατάσταση WM, MM, HM, είναι τα ακόλουθα:



(α′) Κατάσταση WM



(β′) Κατάσταση MM



(γ′) Κατάσταση HM

Σχήμα 4.6: Multistream SMAC-FMP, ανάλογα με το ζεύγος βαρών (Zβ) και το Median φίλτρο (Mf)

Λαμβάνοντας υπόψη τον πίνακα για το FMP, η καλύτερη απόδοση συμβαίνει όταν έχουμε μέγεθος του Median φίλτρου ίσο με 3. Ενώ, λαμβάνοντας υπόψη τον πίνακα και τα σχεδιαγράμματα για το multistream SMAC-FMP, η, σε γενικές γραμμές, καλύτερη απόδοση είναι όταν έχουμε μέγεθος του Median φίλτρου ίσο με 11 και ζεύγος βαρών 1.0-0.1, και αυτό διότι τότε η κατάσταση MM παίρνει τη μεγαλύτερη τιμή της και οι τιμές των άλλων δύο καταστάσεων δεν απέχουν πολύ από τις μέγιστες τιμές τους. Οπότε, ως σύνολο τιμών μπορούμε να πούμε ότι είναι το καλύτερο από όλο τον πίνακα. Εφόσον, λοιπόν, έχουμε δύο διαφορετικές τιμές του Median φίλτρου, θα πειραματιστούμε, όπως και στην περίπτωση των MFCC-FMP front-ends, δύο φορές. Την πρώτη φορά το μέγεθος του Median φίλτρου είναι ίσο με 11 και τη δεύτερη ίσο με 3. Και τις δύο φορές πειραματιζόμαστε πάνω στο φάσμα μεταβλητών συχνοτήτων των συστοιχιών των Gabor φίλτρων, που λαμβάνουν χώρα στη διαδικασία της αποδιαμόρφωσης. Στους επόμενους δύο πίνακες φαίνονται τα αποτελέσματα για τις τιμές της ακρίβειας, όσον αφορά τα SMAC front-ends και τα baseline SMAC-FMP front-ends, όταν το μέγεθος του Median φίλτρου είναι ίσο με 11.

Μετακίνηση (Hz)	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	49.34	18.49	23.31
-90:90	43.14	13.89	25.11
-150:150	23.62	0.90	23.91
-210:210	2.87	-12.81	17.41
-270:270	-6.96	-17.48	15.52

Πίνακας 4.22: Αποτελέσματα για το FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 11

Μετακίνηση (Hz)	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	92.79	87.54	74.92
-90:90	93.30	86.86	75.19
-150:150	93.63	87.01	74.20
-210:210	93.35	88.22	74.71
-270:270	93.28	87.70	74.62

Πίνακας 4.23: Αποτελέσματα για το baseline SMAC-FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 11

Έπειτα, βρίσκουμε τις τιμές της ακρίβειας για την περίπτωση των multistream SMAC-FMP front-ends, για τις ίδιες τιμές βαρών των streams και μετακίνησης του φάσματος μεταβλητών συχνοτήτων, με αυτές της αντίστοιχης περίπτωσης των multistream SMAC-FMP front-ends, όπου λαμβάνει χώρα αποδιαμόρφωση σταθερού φάσματος συχνοτήτων. Οι τιμές αυτές περιέχονται στον παρακάτω πίνακα:

Μετακίνηση (Hz)	**Weights**	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	1.0 - 0.1	94.30	88.99	76.99
-30:30	0.8 - 0.2	94.34	88.86	75.94
-90:90	1.0 - 0.1	94.35	89.21	77.08
-90:90	0.8 - 0.2	94.54	89.15	76.15
-150:150	1.0 - 0.1	94.33	89.08	77.05
-150:150	0.8 - 0.2	94.61	88.99	75.85
-210:210	1.0 - 0.1	94.33	89.15	76.81
-210:210	0.8 - 0.2	94.41	88.66	75.34
-270:270	1.0 - 0.1	94.44	89.15	76.60
-270:270	0.8 - 0.2	94.41	89.19	74.83

Πίνακας 4.24: Αποτελέσματα για το multistream SMAC-FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 11

Στη συνέχεια παραθέτουμε τους αντίστοιχους με τους τρεις παραπάνω, πίνακες με τα αποτελέσματα της ακρίβειας μετάδοσης, για τα FMP, τα baseline SMAC-FMP και τα multistream SMAC-FMP front-ends, όταν το μέγεθος του Median φίλτρου είναι ίσο με 3.

Μετακίνηση (Hz)	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	60.30	34.54	29.02
-90:90	65.31	40.72	31.49
-150:150	65.84	38.83	37.65
-210:210	59.77	35.29	34.56
-270:270	54.13	31.26	29.83

Πίνακας 4.25: Αποτελέσματα για το FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 3

Μετακίνηση (Hz)	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	92.56	84.11	76.09
-90:90	93.40	85.54	77.14
-150:150	93.19	85.89	76.81
-210:210	93.22	86.02	76.39
-270:270	93.11	86.37	75.91

Πίνακας 4.26: Αποτελέσματα για το baseline SMAC-FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 3

Μετακίνηση (Hz)	**Weights**	**WM Accuracy (%)**	**MM Accuracy (%)**	**HM Accuracy (%)**
-30:30	1.0 - 0.1	94.33	88.71	77.35
-30:30	0.8 - 0.2	94.38	87.96	76.54
-90:90	1.0 - 0.1	94.25	89.04	76.93
-90:90	0.8 - 0.2	94.18	88.60	76.57
-150:150	1.0 - 0.1	94.28	88.99	77.32
-150:150	0.8 - 0.2	94.14	88.69	76.87
-210:210	1.0 - 0.1	94.28	88.97	77.29
-210:210	0.8 - 0.2	94.10	88.49	76.69
-270:270	1.0 - 0.1	94.28	88.99	77.20
-270:270	0.8 - 0.2	94.02	88.58	76.24

Πίνακας 4.27: Αποτελέσματα για το multistream SMAC-FMP για διάφορα φάσματα μεταβλητών συχνοτήτων, όταν το μέγεθος του Median φίλτρου είναι ίσο με 3

Συγκρίνοντας τα αποτελέσματα των πινάκων για τα multistream SMAC-FMP front-ends, παρατηρούμε ότι: Για την κατάσταση WM τα αποτελέσματα για την περίπτωση που δίνουμε στο Median φίλτρο την τιμή 11 είναι καλύτερα από όταν δίνουμε την τιμή 3, εκτός από την περίπτωση που έχουμε μετακίνηση στο διάστημα -30:30 Hz. Και τότε ακόμα, όμως, η διαφορά είναι μόλις 0.03 και 0.04 για τις περιπτώσεις ζεύγους βαρών 1.0 -0.1 και 0.8-0.2, αντίστοιχα, και εφόσον συμβαίνει μόνο μια φορά αυτό, θεωρούμε ότι δεν χρειάζεται να το λάβουμε υπόψη. Επίσης, τα αποτελέσματα για την κατάσταση MM είναι καλύτερα όταν η τιμή του Median φίλτρου είναι 11. Αντίθετα, τα περισσότερα αποτελέσματα για την κατάσταση HM είναι καλύτερα όταν η τιμή του Median φίλτρου είναι 3. Εφόσον, λοιπόν, υπάρχει αυτή η διαφοροποίηση, θα καταλήξουμε στο ποιά τιμή Median φίλτρου δίνει τα

καλύτερα αποτελέσματα, βάσει της συνολικής εικόνας των αποτελεσμάτων. Καταλήγουμε, λοιπόν, στο συμπέρασμα ότι τα καλύτερα αποτελέσματα προκύπτουν όταν το Median φίλτρο έχει την τιμή 11. Στη συνέχεια, επικεντρώνοντας την προσοχή μας στον πίνακα που το Median φίλτρο είναι ίσο με 11, παρατηρούμε ότι: Για τις καταστάσεις MM και HM, τα αποτελέσματα για το ζεύγος βαρών 1.0-0.1 είναι καλύτερα από τα αντίστοιχα του ζεύγους 0.8-0.2. Το αντίθετο ισχύει για την κατάσταση WM, εκτός από την περίπτωση που έχουμε μετακίνηση στο διάστημα από -270 έως 270 Hz. Επειδή, όμως, όπως προαναφέραμε, όταν υπάρχει διαφοροποίηση βγάζουμε συμπεράσματα βάσει της συνολικής εικόνας των αποτελεσμάτων, συμπεραίνουμε ότι το ζεύγος βαρών με τα βέλτιστα, σε γενικές γραμμές, αποτελέσματα είναι το 1.0-0.1. Από αυτό το ζεύγος, τα καλύτερα αποτελέσματα προκύπτουν όταν έχουμε μετακίνηση στο διάστημα συχνοτήτων -90 έως 90 Hz, διότι τότε έχουμε τα καλύτερα αποτελέσματα για τις καταστάσεις MM και HM, και το δεύτερο καλύτερο αποτέλεσμα για την κατάσταση WM. Άρα, οι επιθυμητές παράμετροι για την περίπτωση του multistream SMAC-FMP, όταν έχουμε φάσματα μεταβλητών συχνοτήτων, είναι η τιμή 11 για το μέγεθος του Median φίλτρου, -90 έως 90 Hz για τη μετακίνηση συχνοτήτων, και 1.0-0.1 για τις τιμές των βαρών των streams. Ενώ, συγκρίνοντας το με το multistream SMAC-FMP όταν έχουμε φάσμα σταθερών συχνοτήτων, καταλήγουμε σε δύο συμπεράσματα: Πρώτον, ότι και οι δυο κατηγορίες multistream SMAC-FMP παρουσιάζουν τα βέλτιστα αποτελέσματά τους για τις ίδιες παραμέτρους μέγεθος του Median φίλτρου και βάρη των streams. Δεύτερον, ότι το multistream SMAC-FMP όταν έχουμε φάσμα μεταβλητών συχνοτήτων οδηγεί σε καλύτερα αποτελέσματα από όταν έχουμε φάσμα σταθερών συχνοτήτων. Τέλος, αν συγκρίνουμε τα βέλτιστα αποτελέσματα του multistream SMAC-FMP με τα βέλτιστα αποτελέσματα του multistream MFCC-FMP που βρήκαμε στο προηγούμενο κεφάλαιο, παρατηρούμε ότι τα αποτελέσματα για το multistream SMAC-FMP είναι καλύτερα από αυτά του multistream MFCC-FMP, και για τις τρεις καταστασεις WM, MM και HM. Επομένως, το front-end που οδηγεί στα βέλτιστα αποτελέσματα ακρίβειας μετάδοσης είναι το Multistream SMAC-FMP, όταν έχουμε δώσει τις ακόλουθες τιμές στις παραμέτρους: Μέγεθος του Median φίλτρου = 11, διάστημα μετακίνησης του φάσματος μεταβλητών συχνοτήτων = -90:90 Hz και ζεύγος βαρός των streams = 1.0-0.1. Οι αντίστοιχες τιμές ακρίβειας μετάδοσης είναι οι εξής:

Κατάσταση	**Ακρίβεια (%)**
WM	94.35
MM	89.21
HM	77.08

Πίνακας 4.28: Αποτελέσματα ακρίβειας μετάδοσης για τις βέλτιστες παραμέτρους του Multistream SMAC-FMP front-end.

Κεφάλαιο 5

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΚΑΤΕΥΘΥΝΣΕΙΣ ΜΕΛΛΟΝΤΙΚΗΣ ΕΡΕΥΝΑΣ

Αντικείμενο της εργασίας μας ήταν η βελτιστοποίηση της μετάδοσης σήματος φωνής, με χρήση front-end χαρακτηριστικών. Επικεντρωθήκαμε στα χαρακτηριστικά multistream MFCC-FMP και multistream SMAC-FMP, που τα εφαρμόσαμε στη βάση δεδομένων AURORA 3 για την Ισπανική γλώσσα (AURORA 3 Speech Database - Spanish Task), η οποία περιέχει τους αριθμούς της Ισπανικής γλώσσας από το ένα έως και το εννιά, συμπεριλαμβανομένου και του μηδέν, σε δειγματοληπτημένα στα 8 kHz αρχεία, ηχογραφημένα από δύο ειδών μικρόφωνα, ένα κοντινής και ένα μακρινής απόστασης, για τρεις καταστάσεις οδήγησης σε αυτοκίνητο. Την Κατάσταση Καλής Ταύτισης-WM την Κατάσταση Μέτριας Μη Ταύτισης-MM και την Κατάσταση Υψηλής Μη Ταύτισης-HM. Αυτή η βάση δεδομένων μας δίνει τη δυνατότητα αναγνώρισης λέξης. Διεξάγαμε πειράματα πάνω σε μια σειρά παραμέτρους της κατασκευής των παραπάνω front-end χαρακτηριστικών και παρατηρήσαμε κάθε φορά το αν και κατά πόσο βελτιώνεται η ακρίβεια μετάδοσης, σε κάθε μια από τις καταστάσεις WM, MM και HM, αλλά και σαν συνολική εικόνα. Στόχος μας ήταν να βρούμε τον βέλτιστο συνδυασμό παραμέτρων, δηλαδή αυτόν που δίνει τα βέλτιστα αποτελέσματα για την ακρίβεια μετάδοσης του σήματος φωνής.

Με βάση τα πειράματα που διεξάγαμε καταλήξαμε στα ακόλουθα συμπεράσματα. Πρώτον, ότι η χρήση φίλτρων Gabor με φάσματα μεταβλητών συχνοτήτων, κατά τη διαδικασία κατασκευής των χαρακτηριστικών FMP, βελτιώνει την απόδοση του συστήματος, τόσο όσον αφορά την ακρίβεια που προκύπτει από τα FMP χαρακτηριστικά, όσο και από το συνδυασμό τους με τα MFCC ή τα SMAC. Δεύτερον, τα πειράματα που διεξάγαμε πάνω στο ζήτημα του αν αυξάνεται η απόδοση με τη χρήση multistreams, τόσο στα MFCC-FMP χαρακτηριστικά, όσο και στα SMAC-FMP χαρακτηριστικά, έδειξαν ότι με την απόθεση κατάλληλου βάρους σε κάθε stream, η απόδοση του συστήματος βελτιώνεται. Άρα είναι προτιμότερη η χρήση multistreams με κατάλληλα βάρη. Τρίτον, ασχοληθήκαμε με την παρα-

πέρα βελτίωση της ακρίβειας μετάδοσης, ερευνώντας τις αλλαγές που προκαλούν στο σύστημα οι αλλαγές σε κάποιες παραμέτρους. Επικεντρωθήκαμε στην έρευνα για δύο παραμέτρους, που και οι δύο λαμβάνουν χώρα κατά την κατασκευή των χαρακτηριστικών FMP. Η πρώτη παράμετρος είναι το μέγεθος του Median φίλτρου και η δεύτερη το διάστημα στο οποίο μετακινούνται οι συχνότητες του φάσματος μεταβλητών συχνοτήτων. Αυτές οι παράμετροι συνδυάστηκαν και με τα βάρη που δίνουμε στα streams, ούτως ώστε να προκύψει η τριάδα παραμέτρων που θα μας δώσει τα καλύτερα αποτελέσματα, τόσο για τα MFCC-FMP, όσο και για τα SMAC-FMP. Έχοντας βρει την παραπάνω καταλληλότερη τριάδα παραμέτρων, μπορέσαμε πλέον να αποφασίσουμε ποιός συνδυασμός front-end χαρακτηριστικών, MFCC-FMP ή SMAC-FMP, οδηγεί σε καλύτερα αποτελέσματα. Συγκρίνοντας, λοιπόν, τα αποτελέσματά τους, βρήκαμε ότι τα multistream SMAC-FMP χαρακτηριστικά οδηγούν σε καλύτερα αποτελέσματα από ότι τα multistream MFCC-FMP, και για τις τρεις καταστάσεις WM, MM και HM. Επομένως, είναι προτιμότερη η χρήση multistream SMAC-FMP χαρακτηριστικών, με τις συγκεκριμένες τιμές στις παραμέτρους που βρήκαμε, παρά η αντίστοιχη με multistream MFCC-FMP χαρακτηριστικά. Οι τιμές αυτές είναι: Μέγεθος του Median φίλτρου = 11, διάστημα μετακίνησης του φάσματος μεταβλητών συχνοτήτων = -90:90 Hz και ζεύγος βαρός των streams = 1.0-0.1. Οι αντίστοιχες τιμές ακρίβειας μετάδοσης είναι οι ακόλουθες:

Κατάσταση	**Ακρίβεια (%)**
WM	94.35
MM	89.21
HM	77.08

Πίνακας 5.1: Αποτελέσματα ακρίβειας μετάδοσης για τις βέλτιστες παραμέτρους του Multistream SMAC-FMP front-end.

Δυστυχώς, λόγω του περιορισμού στο χρόνο, αλλά και στα θέματα έρευνας που περικλύει κάθε διπλωματική εργασία, δεν μπορέσαμε να επεκτείνουμε την έρευνά μας σε μια σειρά θέματα, που μπορεί να οδηγούσαν σε μεγαλύτερη βελτίωση της απόδοσης του συστήματός μας ή σε ποιό ολοκληρωμένα συμπεράσματα για το ρόλο που διαδραματίζουν διάφορες παράμετροι στο σύστημά μας. Κάποια από τα θέματα αυτά αφορούν, κατ΄ αρχήν, την έρευνα πάνω σε άλλες τιμές που μπορούν να πάρουν οι παράμετροι της μετακίνησης του φάσματος και του μεγέθους του Median φίλτρου και στο αν βελτιώνεται τότε η απόδοση. Τιμές ενδιάμεσες αυτών που έχουμε δώσει, αλλά και μεγαλύτερες ή μικρότερες. Επιπλέον, την έρευνα πάνω στις τιμές άλλων παραμέτρων του συστήματός μας και στο συνδυασμό τους με τις παραπάνω, παράμετροι δηλαδή με τις οποίες δεν ασχοληθήκαμε σε βάθος. Τέτοιες είναι ο συντελεστής προέμφασης, ο αριθμός των Mel φίλτρων και ο πολλαπλασιαστής εύρους ζώνης στα φίλτρα Gabor. Επίσης, τη χρήση της μεθόδου HTD έναντι της ESA κατά τη διαδικασία κατασκευής των FMP. Ακόμη, το ζήτημα του χρόνου υλοποίησης, το πώς δηλαδή θα κάνουμε το πρόγραμμά μας πιο γρήγορο και με λιγότερο υπολογιστικό κόστος, διότι ο χρόνος που χρειαζόμαστε για να προκύψουν τα αποτελέσματα είναι ιδιαίτερα μεγάλος. Τέλος, την επέκταση των πειραμάτων στη βάση δεδομένων AURORA 3, για την Ιταλική γλώσσα (AURORA 3 Italian task), προκειμένου να ληφθούν πιο ολοκληρωμένα συμπεράσματα, μιας και θα έχουμε αποτελέσματα από δύο tasks ταυτόχρονα, αλλά και να μπορέσουμε να βρούμε ποιά

front-end χαρακτηριστικά και με ποιές παραμέτρους βελτιώνουν την απόδοση του συστήματος σε αυτό το task.

Πιστεύουμε ότι μελλοντικές επεκτάσεις της παρούσας εργασίας θα δώσουν απαντήσεις σε πολλά από τα παραπάνω θέματα, θα οδηγήσουν σε περαιτέρω βελτίωση των αποτελεσμάτων, και θα συνεισφέρουν σημαντικά στην επέκταση των γνώσεων πάνω στο επιστημονικό πεδίο της αναγνώρισης σήματος φωνής.

Παράρτημα Α′

ΠΑΡΑΡΤΗΜΑ

Α′.1 Κρυφά Μαρκοβιανά Μοντέλα (HMM)

Ένα Κρυφό Μαρκοβιανό Μοντέλο είναι μια μηχανή πεπερασμένων καταστάσεων, στην οποία το σύστημα που μοντελοποιείται θεωρείται μια διαδικασία Markov. Αποτελείται από τις ακόλουθες παραμέτρους:

1. Ένα σύνολο καταστάσεων N. Σε μια τυχαία χρονική στιγμή t το μοντέλο βρίσκεται στην κατάσταση q_t, όπου $q_t = 1, 2, ..., N$. Το μοντέλο αλλάζει μία κατάσταση κάθε χρονική στιγμή.

2. Ένα σύνολο παρατηρήσεων $O = o_1, o_2, ..., o_T$, όπου T είναι η τελική χρονική στιγμή. Αν τη χρονική στιγμή t, βρισκόμαστε στην κατάσταση j, τότε εξάγεται μια παρατήρηση o_t, βάσει της συνάρτησης πυκνότητας πιθανότητας

$$B = b_j(o_t) = P[o_t|q_t = j] \tag{Α′.1}$$

3. Η μετάβαση από την κατάσταση i στην κατάσταση j γίνεται βάσει μιας πιθανότητας μετάβασης a_{ij}, όπου

$$A = a_{ij} = P[q_{t+1} = j|q_t = i] \tag{Α′.2}$$

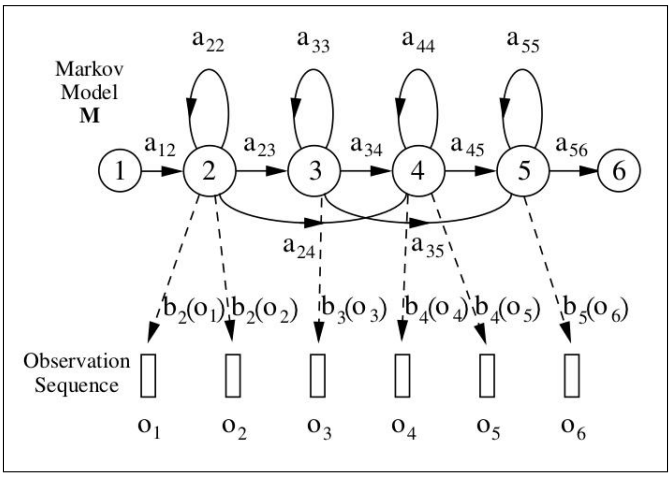
   όπου $1 \leq i, j \leq N$.

4. Αν το Κρυφό Μαρκοβιανό Μοντέλο ξεκινά από μιά κατάσταση i, αυτό αντιστοιχίζεται στην κατανομή πιθανότητας

$$\pi = \{\pi_i\} = P[q_1 = i], 1 \leqslant i \leqslant N. \tag{Α′.3}$$

Συνοπτικά, ένα Κρυφό Μαρκοβιανό Μοντέλο συμβολίζεται ως $\lambda = (A, B, \pi)$. Σε ένα τέτοιο μοντέλο, μόνο το σύνολο των παρατηρήσεων O είναι γνωστό, ενώ η ακολουθία των καταστάσεων που οδηγούν στην ακολουθία των παρατηρήσεων δεν είναι γνωστή, είναι κρυφή. Γι′ αυτό ονομάζεται Κρυφό Μαρκοβιανό Μοντέλο. Η αρχιτεκτονή τους διαφέρει, ανάλογα με την εφαρμογή. Στη διαδικασία αναγνώρισης φωνής χρησιμοποιείται αρχιτεκτονική μεταβάσεων από αριστερά προς τα δεξιά (left-to-right) [29]. Αν χρησιμοποιήσουμε την πειραματική πλατφόρμα HTK , η αρχική

και η τελική κατάσταση ενος Κρυφού Μαρκοβιανού Μοντέλου δεν συμπληρώνεται, ώστε να διευκολυνθεί η κατασκευή των σύνθετων μοντέλων [25]. Στο παρακάτω σχήμα, απεικονίζεται ένα τέτοιο μοντέλο.



Σχήμα Α΄.1: Κρυφό Μαρκοβιανό Μοντέλο, με τέσσερις καταστάσεις, εκτός της αρχικής και της τελικής, πηγή: [25]

Οι κατανομές πιθανότητας που χρησιμοποιούνται ποικίλουν, ανάλογα με την περίπτωση του HMM που χρησιμοποιούμε. Αν χρησιμοποιούμε την πειραματική πλατφόρμα HTK, τότε η πιθανότητα της κάθε κατάστασης να εξάγει μια παρατήρηση ($b_j(o_t)$) τη χρονική στιγμή t, δίνεται από το συνδυασμό Γκαουσιανών κατανομών (Gaussian Mixtures). Επιπλέον, αν κάθε διάνυσμα παρατηρήσεων που εξάγεται τη χρονική στιγμή t διασπάται σε ανεξάρτητα διανύσματα (streams), δηλαδή αν έχουμε την περίπτωση multistream HMM, η $b_j(o_t)$ δίνεται από τον παρακάτω μαθηματικό τύπο [25]:

$$b_j(o_t) = \prod_{s=1}^{S} [\sum_{m=1}^{M_s} c_{jsm} N(o_{st}; \mu_{jsm} \Sigma_{jsm})]^{\gamma_s} \qquad (\text{Α΄.4})$$

Όπου M_s είναι ο αριθμός των κανονικών κατανομών (mixture components), s είναι ο αριθμός των ανεξάρτητων διανυσμάτων, c_{jsm} είναι το βάρος της m-ιοστής κατανομής και $N(o; \mu, \Sigma)$ είναι η πολυδιάστατη κανονική κατανομή, με μέση τιμή μ και πίνακα αυτοσυσχέτισης Σ και γ_s είναι το βάρος του κάθε stream.

Α΄.2 Κλίμακα mel

Έρευνες της ψυχοακουστικής έχουν αποδείξει ότι η ανθρώπινη αντίληψη για τις συχνότητες ενός σήματος φωνής δεν ακολουθεί γραμμική κλίμακα, διότι η αντιληπτική ικανότητα του ανθρώπινου αυτιού δεν είναι ομοιόμορφα κατανεμημένη στο ακουστικό φάσμα των συχνοτήτων (δηλαδή στο εύρος ζώνης που αντιλαμβάνεται το αυτί). Είναι, λοιπόν, χρήσιμο να αναλύουμε το σήμα φωνής στο πεδίο της συχνότητας, σε κλίμακες που προσεγγίζουν τον τρόπο αντίληψης των ήχων από το ανθρώπινο αυτί. Για το λόγο αυτό έχουν προταθεί διάφορες κλίμακες συ-

χνότητας/εύρους ζώνης, με κυριότερες τις εξής:
- Μουσική κλίμακα.
- Κλίμακα Mel.
- Κλίμακα Bark.
- Κλίμακα ERB (Equivalent Rectangular Bandwidth) [31].

Από τις παραπάνω μεθόδους, η πιο δημοφιλής είναι η κλίμακα mel (από τη λέξη melody) και η πρώτη που προέκυψε από πειραματικές μετρήσεις. Όπως αποδείχτηκε, για κάθε τόνο με πραγματική συχνότητα f μετρημένη σε Hz, υπάρχει ένας υποκειμενικός τόνος στην κλίμακα mel. Δηλαδή δεν υπάρχει μόνο το πραγματικό φάσμα συχνοτήτων, αλλά και το υποκειμενικό, αυτό δηλαδή που αντιλαμβάνεται το ανθρώπινο αυτί, το οποίο προσεγγίζεται με την κλίμακα mel. Η κλίμακα αυτη αποτελει αναδίπλωση του άξονα συχνοτήτων σε λογαριθμική κλίμακα. Είναι γραμμική μέχρι το 1kHz και λογαριθμική στις υψηλότερες συχνότητες [31]. Μετριέται σε mels και δίνεται από την παρακάτω σχέση [19]:

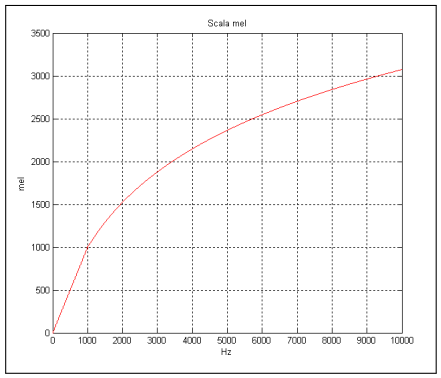
$$m = 2595log(1 + \frac{f}{700}) \tag{Α΄.5}$$

Ενώ η αντίστροφη σχέση (από mel σε συχνότητες) είναι η εξής:

$$f = 700(10^{m/2595} - 1) \tag{Α΄.6}$$

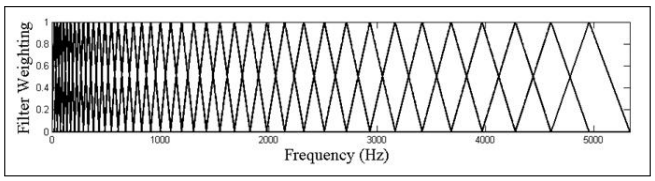
Όπου f είναι η συχνότητα σε Hz και m είναι τα αντίστοιχα mels.

Στο παρακάτω σχήμα έχουμε την απεικόνιση συχνοτήτων Hz σε mels, με χρήση των παραπάνω σχέσεων.



Σχήμα Α΄.2: Κλίμακα Mel

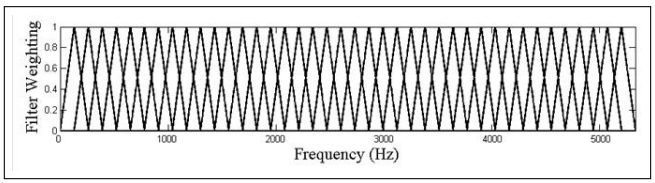
Μια προσέγγιση για να προσομοιώσουμε το υποκειμενικό φάσμα, που αναφέραμε παραπάνω, είναι η χρησιμοποίηση μιας διάταξης (μιας συστοιχίας φίλτρων-filterbank) τριγωνικών ζωνοπερατών φίλτρων, με ένα φίλτρο για κάθε mel συχνότητα. Στο παρακάτω σχήμα απεικονίζεται μια τέτοια προσομοίωση, όταν έχουμε θέσει τους mel συντελεστές ίσους με 40.



Σχήμα Α΄.3: Κλίμακα Mel, πηγή: [19]

Παρατηρούμε ότι έχουμε περισσότερα φίλτρα στις χαμηλές συχνότητες και λιγότερα στις υψηλές. Προσεγγίζουμε καλύτερα το φάσμα συχνοτήτων που αντιλαμβάνεται το αυτί.

Αν δεν χρησιμοποιήσουμε τη mel κλίμακα και τα φίλτρα είναι ομοιόμορφα κατανεμημένα στο πεδίο των συχνοτήτων, το αποτέλεσμα απεικονίζεται στο παρακάτω σχήμα.



Σχήμα Α΄.4: Ομοιόμορφα κατανεμημένα φίλτρα, πηγή: [19]

Όμως, τότε δεν προσεγγίζουμε τόσο καλά το φάσμα συχνοτήτων που αντιλαμβάνεται το αυτί.

Α΄.3 Συντελεστές Δέλτα και Συντελεστές Επιτάχυνσης

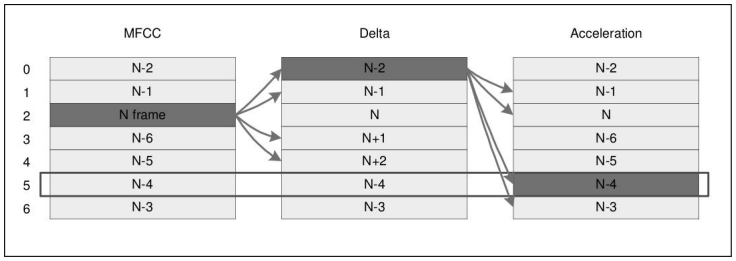
Τους συντελεστές Δέλτα και Επιτάχυνσης τους χρησιμοποιούμε όταν θέλουμε να εξάγουμε front-end χαρακτηριστικά, πχ MFCC, FMP ή SMAC από ένα σήμα φωνής που δεν μένει σταθερό στο χρόνο, αλλά αλλάζει. Τότε προκύπτει η ανάγκη να προσθέσουμε στα front-end χαρακτηριστικά και χαρακτηριστικά που σχετίζονται με τις αλλαγές στο φάσμα των front-end χαρακτηριστικών στην πάροδο του χρόνου. Και αυτά τα χαρακτηριστικά είναι οι συντελεστές Δέλτα και Επιτάχυνσης (Delta and Acceleration Coefficients), με την προσθήκη των οποίων η απόδοση ενός συστήματος αναγνώρισης φωνής μπορεί να ενισχυθεί σημαντικά.

Οι συντελεστές Δέλτα βρίσκονται με την παρακάτω εξίσωση:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{Α΄.7}$$

όπου d_t είναι ο συντελεστής Δέλτα τη χρονική στιγμή t και $c_{t-\theta}$, $c_{t+\theta}$ είναι ο αντίστοιχος προηγούμενος και ο αντίστοιχος επόμενος συντελεστής front-end. Η παράμετρος θ ορίζεται από εμάς κάθε φορά. Ο αντίστοιχος τύπος χρησιμοποιείται για την εύρεση των συντελεστών Επιτάχυνσης, μόνο που τότε τα $c_{t-\theta}$, $c_{t+\theta}$ είναι οι αντίστοιχοι συντελεστές Δέλτα. Αν πχ είχαμε ορίσει το συντελεστή θ ίσο με

δύο, τότε, για την κατασκευή του συντελεστή Δέλτα, θα χρησιμοποιούνταν οι front-end συντελεστές των προηγούμενων δύο πλαισίων και των επόμενων δύο πλαισίων. Αυτό σημαίνει ότι μετά την εύρεση των front-end συντελεστών για το N-οστό πλαίσιο, μπορούμε να βρούμε τους Δέλτα συντελεστές για το $N-1$, $N+1$, $N-2$ και $N+2$ πλαίσιο. Και το $N-2$ πλαίσιο δεν θα αλλάξει, διότι τα επόμενα front-ends δεν πρόκειται να το επηρρεάσουν. Με τον ίδιο τρόπο βρίσκουμε και τους συντελεστές Επιτάχυνσης. Αυτό απεικονίζεται στο επόμενο σχήμα, όπου έχουμε επτά πλαίσια [25, 27].



Σχήμα Α΄.5: Σχεδιάγραμμα εύρεσης συντελεστών Δέλτα και Επιτάχυνσης, πηγή: [27]

Α΄.4 Τελεστής ενέργειας Teager-Kaiser (TEO)

Όταν έχουμε ένα AM-FM σήμα φωνής r(t), ο τελεστής ενέργειας Teager-Kaiser βρίσκει την ενέργεια του σήματος και υπολογίζεται με την ακόλουθη εξίσωση, όταν πρόκειται για σήματα συνεχούς χρόνου [7]:

$$\Psi[r(t)] = [\dot{r(t)}]^2 - r(t)\ddot{r(t)} \tag{Α΄.8}$$

ενώ, όταν πρόκειται για σήματα διακριτού χρόνου, υπολογίζεται από την εξίσωση:

$$\Psi[r(n)] = r^2(n) - r(n-1)r(n+1) \tag{Α΄.9}$$

Ο συγκεκριμένος τελεστής έχει την ιδιότητα να ανιχνεύει την ενέργεια ενός γραμμικού ταλαντωτή, γι΄ αυτό και ονομάζεται τελεστής ενέργειας. Έχει πολύ καλή ανάλυση χρόνου και χαμηλή πολυπλοκότητα, γι΄ αυτό και είναι ιδιαίτερα διαδεδομένος [7].

Α΄.5 ESA

Για ένα AM-FM σήμα φωνής r(t), η στιγμιαία συχνότητα και το στιγμιαίο πλάτος, με χρήση του ESA, υπολογίζονται, αντίστοιχα, ως εξής [7]:

$$f(t) \approx (\frac{1}{2\pi})\sqrt{\frac{\Psi[\dot{r}(t)]}{\Psi[r(t)]}} \tag{Α΄.10}$$

$$\alpha(t) \approx \frac{\Psi[r(t)]}{\sqrt{\Psi[\dot{r}(t)]}} \tag{Α΄.11}$$

Παρόμοιες εξισώσεις και αλγόριθμοι υπάρχουν και για την περίπτωση που κάνουμε ESA σε σήμα διακριτού χρόνου (DESA) [7].

Α΄.6 Φίλτρα Gabor

Οι συστοιχίες φίλτρων που χρησιμοποιούνται στην πολυζωνική ανάλυση και αποδιαμόρφωση, αποτελούνται από φίλτρα Gabor. Η κρουστική απόκριση και η απόκριση συχνότητας ενός φίλτρου Gabor δίνονται, αντίστοιχα, από τους παρακάτω τύπους [5]:

$$h(t) = \exp(-\alpha^2 t^2)\cos(\omega_c(t)) \tag{Α΄.12}$$

$$H(\omega) = \frac{\sqrt{\pi}}{2\alpha}(\exp[\frac{-(\omega-\omega_c)^2}{4\alpha^2}] + \exp[\frac{-(\omega+\omega_c)^2}{4\alpha^2}]) \tag{Α΄.13}$$

όπου ω_c είναι η κεντρική συχνότητα και a είναι η παράμετρος που καθορίζει το εύρος ζώνης του φίλτρου. Το αποδοτικό RMS του εύρους ζώνης του φίλτρου Gabor είναι ίσο με $a/(2\pi)$. Τα φίλτρα Gabor χρησιμοποιούνται για διάφορους λόγους, εκ των οποίων οι σηματικότεροι είναι, πρώτον, η ιδιότητά τους να έχουν βέλτιστη διακριτική ικανότητα, τόσο στο πεδίο του χρόνου, όσο και στο πεδίο της συχνότητας, και δεύτερον, το γεγονός ότι η απόκριση συχνότητά τους δεν παρουσιάζει ισχυρούς δευτερεύοντες λοβούς [5].

Α΄.7 Υπολογισμός της φασματικής ροπής

Αν υποθέσουμε ότι έχουμε ένα σήμα φωνής διακριτού χρόνου $x(n)$, το οποίο φιλτράρεται με μία συστοιχία από K ζωνοπερατά φίλτρα, με κεντρικές συχνότητες ω_k, τότε θα προκύψουν τα διακριτά ζωνοπερατά σήματα $x_k(n)$, που θα δίνονται, στο πεδίο του χρόνου και της συχνότητας, από την ακόλουθη σχέση:

$$x_k(n) = x(n) * h_k(n) \longleftrightarrow X_k(\omega) = X(\omega)H_k(\omega) \tag{Α΄.14}$$

όπου $h_k(n)$ είναι η κρουστική απόκριση και $H_k(\omega)$ είναι η απόκριση συχνότητας για το k-οστό φίλτρο. Η m-οστή φασματική ροπή και κεντρική φασματική ροπή για κάθε σήμα $x_k(n)$, για μια αυθαίρετη σταθερά γ, ορίζονται αντίστοιχα ως [1]:

$$S^m(k) = \int_0^\pi |X_k(\omega)|^\gamma \omega^m d\omega \tag{Α΄.15}$$

$$S_c^m(k) = \int_0^\pi |X_k(\omega)|^\gamma (\omega-\omega_k)^m d\omega \tag{Α΄.16}$$

Ο όρος "κεντρική" χρησιμοποιείται λόγω της χρήσης της κεντρικής συχνότητας του φίλτρου. Οι αντίστοιχες κανονικοποιημένες φασματικές ροπές ορίζονται ως [1]:

$$N^m(k) = S^m(k)/S^0(k) \tag{Α΄.17}$$

$$N_c^m(k) = S_c^m(k)/S_c^0(k) \tag{Α΄.18}$$

Από τις τέσσερις παραπάνω σχέσεις προκύπτει ότι $S_c^0(k) \triangleq S^0(k)$ και $N_c^1(k) \triangleq N^1(k) - \omega_k$. Η φασματική ροπή μηδενικής τάξης, $S^0(k)$, για γ=2, είναι ισοδύναμη με την ενέργεια του ζωνοπερατού σήματος που προήλθε από το k-οστό φίλτρο. Άρα το διάνυσμα S^0 ισοδυναμεί με το διάνυσμα που χρησιμοποιείται στον υπολογισμό των $MFCC$. Η φασματική ροπή πρώτης τάξης, N^1 προσεγγίζει τη σταθμισμένη συχνότητα συντονισμού (weighted average formant frequency) για κάθε φασματική ζώνη, η οποία επίσης χρησιμοποιείται στη διαδικασία αναγνώρισης φωνής. Επιπλέον, οι εκτιμήσεις για τη φασματική ροπή μηδενικής τάξης και τη φασματική ροπή πρώτης τάξης, στο πεδίο του χρόνου, χρησιμοποιούνται και αυτές με τη σειρά τους στη διαδικασία αναγνώρισης φωνής [1]. Στο επόμενο υποκεφάλαιο θα δείξουμε τη σχέση που έχει η φασματική ροπή με το φάσμα ισχύος και τις φασματικές κορυφές.

Α΄.8 Σχέση της φασματικής ροπής με το φάσμα ισχύος και τις φασματικές κορυφές

Αν $h_k(n)$ είναι η κρουστική απόκριση ενός πραγματικού Gabor φίλτρου, τότε η απόκριση συχνότητας του φίλτρου αυτού, όπως δείξαμε και σε παραπάνω κεφάλαιο, εκφράζεται από τη σχέση

$$H_k(\omega) = (\sqrt{\pi}/2\alpha)(e^{(\omega-\omega_k)^2/4\alpha^2} + e^{(\omega+\omega_k)^2/4\alpha^2}) \tag{Α΄.19}$$

όπου a είναι μια παράμετρος που καθορίζει το εύρος ζώνης του φίλτρου. Για τον υπολογισμό της φασματικής ροπής χρησιμοποιείται, συνήθως, μόνο ο θετικός συντελεστής της παραπάνω σχέσης, εφόσον ο μαθηματικός τύπος της φασματικής ροπής περιέχει ολοκλήρωμα με θετικούς όρους, άρα για θετικές συχνότητες. Επομένως, η απόκριση συχνότητας θα είναι

$$H^+_{k(\omega)} = (\sqrt{\pi}/2\alpha)(e^{(\omega-\omega_k)^2/4\alpha^2}) \tag{Α΄.20}$$

Αν πάρουμε το ολοκλήρωμα, ως προς ω_k, της φασματικής ροπής μηδενικής τάξης και της παραπάνω απόκρισης συχνότητας, προκύπτουν, αντίστοιχα:

$$\begin{aligned}\frac{dS^0(k)}{d\omega_k} = \frac{d}{d\omega_k}\int_0^\pi |X_k(\omega)|^\gamma d\omega &= \int_0^\pi \frac{d|X_k(\omega)|^\gamma}{d\omega_k} d\omega \\ &\simeq \int_0^\pi |X(\omega)|^\gamma \frac{d|H^+_{k(\omega)}|^\gamma}{d\omega_k} d\omega\end{aligned} \tag{Α΄.21}$$

$$\begin{aligned}\frac{d|H_k^+(\omega)|^\gamma}{d\omega_k} &= (\sqrt{\pi}/2\alpha)^\gamma \frac{de^{-\gamma(\omega-\omega_k)^2/4\alpha^2}}{d\omega_k} \\ = (\sqrt{\pi}/2\alpha)^\gamma 2(\gamma/4\alpha^2)(\omega-\omega_k)e^{-\gamma(\omega-\omega_k)^2/4\alpha^2} &= (\gamma/2\alpha^2)(\omega-\omega_k)|H_k^+(\omega)|^\gamma\end{aligned} \tag{Α΄.22}$$

Αντικαθιστώντας το ολοκλήρωμα της απόκρισης συχνότητας στο ολοκλήρωμα

της φασματικής ροπής μηδενικής τάξης προκύπτει:

$$\begin{aligned}\frac{dS^0(k)}{d\omega_k} &\simeq \frac{\gamma}{2\alpha^2}\int_0^\pi |X(\omega)|^\gamma |H_k^+(\omega)|^\gamma(\omega-\omega_k)d\omega \\ &\simeq \frac{\gamma}{2\alpha^2}\int_0^\pi |X_k(\omega)|^\gamma(\omega-\omega_k)d\omega = \frac{\gamma}{2\alpha^2}S_c^1(k)\end{aligned} \tag{Α΄.23}$$

Δηλαδή:

$$S_c^1(k) \simeq \frac{2\alpha^2}{\gamma}\frac{dS^0(k)}{d\omega_k} \tag{Α΄.24}$$

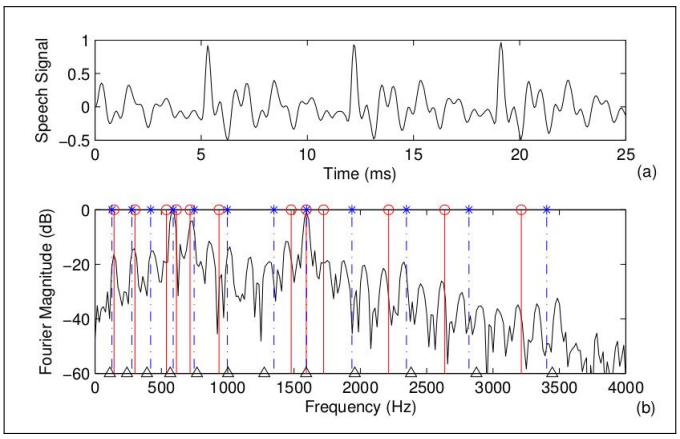
Εφόσον $N_c^m(k) = S_c^m(k)/S_c^0(k)$, τότε $N_c^1(k) = S_c^1(k)/S_c^0(k)$, άρα $N_c^1(k)S_c^0(k) \simeq \frac{2\alpha^2}{\gamma}\frac{dS^0(k)}{d\omega_k}$. Επομένως:

$$N_c^1(k) \simeq \frac{2\alpha^2}{\gamma S^0(k)}\frac{dS^0(k)}{d\omega_k} = \frac{2\alpha^2}{\gamma}\frac{dlog(S^0(k))}{d\omega_k} \tag{Α΄.25}$$

Από την παραπάνω σχέση προκύπτουν δύο συμπεράσματα. Πρώτον, ότι η κανονικοποιημένη κεντρική φασματική ροπή πρώτης τάξης είναι ανάλογη της παραγώγου του λογαρίθμου, ως προς το ω_k, της φασματικής ροπής μηδενικής τάξης, δηλαδή είναι ανάλογη με την λογαριθμική φασματική ισχύ (log power specrtum), η οποία χρησιμοποιείται στο $MFCC front-end$. Αν η κεντρική συχνότητα του φίλτρου (ω_k) βρίσκεται πριν τη φασματική κορυφή, τότε το ολοκλήρωμα στην παραπάνω σχέση είναι θετικό. Αν βρίσκεται μετά τη φασματική κορυφή, το ολοκλήρωμα είναι αρνητικό. Άρα η εκτίμηση κυμαίνεται γύρω από τη φασματική κορυφή. Δεύτερον, ότι όσο μικρότερη είναι η παράμετρος a, τόσο πιο κοντά θα είναι η τιμή της $N^1(k)$ με την κεντρική συχνότητα ω_k. Επομένως, το εύρος ζώνης του φίλτρου παίζει σημαντικό ρόλο. Και αυτό αποδεικνύεται από την εξής αλληλουχία:

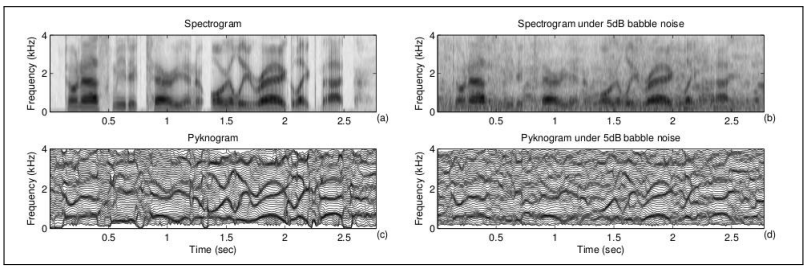
$$\alpha \to 0 \Longrightarrow N_c^1(k) \to 0 \Rightarrow N^1(k) \to \omega_k \tag{Α΄.26}$$

Όπως είπαμε προηγουμένως η εκτίμηση του $N_c^1(k)$ κυμαίνεται γύρω από τη φασματική κορυφή. Αν, λοιπόν, το φίλτρο είναι στενό (δηλαδή η παράμετρος α είναι πολύ μικρή ή τείνει προς το 0) η εκτίμηση παρουσιάζει ευαισθησία, διότι τείνει να επιλέξει την ισχυρότερη αρμονική του pitch γύρω από την κεντρική συχνότητα του φίλτρου. Αυτό φαίνεται στο παρακάτω σχήμα, που απεικονίζει τις εκτιμήσεις για τη φασματική ροπή πρώτης τάξης, για δύο συστοιχίες φίλτρων Gabor, με κεντρικές συχνότητες κατανεμημενες στην κλίμακα Mel. Το εύρος ζώνης των φίλτρων της πρώτης συστοιχίας είναι στα 118 Mels, δηλαδή η συστοιχία αποτελείται από στενά φίλτρα και είναι περίπου ισοδύναμη με την κλασική συστοιχία τριγωνικών φίλτρων με επικάλυψη 50%, που χρησιμοποιείται στον υπολογισμό των $MFCC$. Το εύρος ζώνης των φίλτρων της δεύτερης συστοιχίας είναι στα 236 Mels, που είναι ισοδύναμο με 70% επικάλυψη, άρα τα φίλτρα είναι πιο πλατιά από ότι της πρώτης συστοιχίας. Όπως φαίνεται από το παρακάτω σχήμα, η $N^1(k)$ για τα στενά φίλτρα συμπίπτει σε μεγάλο βαθμό με την πλησιέστερη, στην κεντρική συχνότητα του φίλτρου, αρμονική. Ενώ για την περίπτωση των πλατύτερων φίλτρων, οι τιμές της $N^1(k)$ είναι πλησιέστερα της συχνότητας συντονισμού (formant frequency) [1].



Σχήμα Α΄.6: (a): Πλαίσιο 25-ms (το φώνημα $/ae/$, άντρας ομιλητής, (b): Το αντίστοιχο φάσμα DFT (μέχρι τα $4kHz$, και σε υπέρθεση οι φασματικές ροπές πρώτης τάξης για δύο συστοιχίες φίλτρων με σταθερό εύρος ζώνης 118 και 236 $Mels$, αντίστοιχα. Οι τιμές της φασματικής ροπής που αντιστοιχούν σε συστοιχία με φίλτρα στενού εύρους ζώνης, απεικονίζονται με αστερίσκους και διακεκομμένες γραμμές. Οι τιμές της φασματικής ροπής που αντιστοιχούν σε συστοιχία με φίλτρα ευρύτερου εύρους ζώνης, απεικονίζονται με κύκλους και συνεχείς γραμμές. Οι κεντρικές συχνότητες των φίλτρων, που είναι ίδιες και για τα δύο φίλτρα, απεικονίζονται με τρίγωνα πάνω στον άξονα x. Πηγή: [1]

Η ιδιότητα της φασματικής ροπής πρώτης τάξης να προσεγγίζει την τοπική φασματική κορυφή, την καθιστά ιδιαίτερα χρήσιμη για την φασματική εκτίμηση σε ενθόρυβη κατάσταση, και αυτό διότι τότε οι φασματικές κορυφές δεν επηρρεάζονται σημαντικά. Αυτό φαίνεται στο παρακάτω σχήμα, που παρουσιάζει τα πυκνογράμματα και τα φασματογράμματα μιας πρότασης από το σήμα φωνής TI-MIT σε αθόρυβη και ενθόρυβη κατάσταση. Τα πυκνογράμματα κατασκευάστηκαν με χρήση συστοιχιών φίλτρων Gabor με 64 γραμμικά κατανεμημένα φίλτρα μέχρι τα 4kHz και σταθερό εύρος ζώνης 400Hz. Συγκεκριμένα, στα σχήματα a, c παρουσιάζεται το φασματόγραμμα και το πυκνόγραμμα για το σήμα φωνής, όταν δεν υπάρχει θόρυβος. Στα σχήματα b, d παρουσιάζεται το φασματόγραμμα και το πυκνόγραμμα για το σήμα φωνής, όταν έχει προστεθεί θόρυβος babble στα 5dB, που προήλθε από τη βάση δεδομένων NoiseX92. Όπως βλέπουμε από το πυκνόγραμμα, οι εκτιμήσεις της φασματικής ροπής δεν επηρρεάζονται σημαντικά από τον θόρυβο, όσο βέβαια η φασματική κορυφή παραμένει πάνω από το επίπεδο του θορύβου [1]. Η φασματική ροπή πρώτης τάξης είναι ακόμα πιο χρήσιμη όταν τα φίλτρα έχουν πλατύ εύρος ζώνης, διότι τότε προσεγγίζει την συχνότητα συντονισμού.



Σχήμα Α΄.7: (a): φασματόγραμμα μιας πρότασης από το σήμα φωνής TIMIT χωρίς θόρυβο, (c): πυκνόγραμμα μιας πρότασης από το σήμα φωνής TIMIT χωρίς θόρυβο, (b): φασματόγραμμα μιας πρότασης από το σήμα φωνής TIMIT με θόρυβο (d): πυκνόγραμμα μιας πρότασης από το σήμα φωνής TIMIT με θόρυβο, πηγή: [1]

Α΄.9 Οι αλγόριθμοι Forward, Backward και Baum-Welch

Ο αλγόριθμος Forward δίνει την πιθανότητα να βρισκόμαστε τη χρονική στιγμή t στην κατάσταση j και η ακολουθία των παρατηρήσεων μέχρι εκείνη τη χρονική στιγμή είναι $o_1, ..., o_t$ [25]. Η πιθανότητα Forward υπολογίζεται από τον τύπο:

$$\alpha_j(t) = P(s_t = j, o_1, ..., o_t|\lambda) \tag{Α΄.27}$$

Όπου λ είναι το HMM και s_t είναι η κατάσταση που βρισκόμαστε τη χρονική στιγμή t. Η πιθανότητα να βρισκόμαστε στην κατάσταση j τη χρονική στιγμή t και να εμφανίζεται η παρατήρηση o_t, μπορεί να εξαχθεί και από το άθροισμα των Forward πιθανοτήτων για όλες τις δυνατές προηγούμενες καταστάσεις i, λαμβάνοντας υπόψη την πιθανότητα μετάβασης α_{ij} [25]. Επομένως, η πιθανότητα Forward , για ένα αριθμό καταστάσεων N, μπορεί να υπολογιστεί αποτελεσματικά και από τον τύπο:

$$\alpha_j(t) = [\sum_{i=2}^{N-1} \alpha_i(t-1)\alpha_{ij}]b_j(o_t) \tag{Α΄.28}$$

Όπου α_{ij} είναι η πιθανότητα μετάβασης από την κατάσταση i στην κατάσταση j, N είναι το σύνολο των καταστάσεων, $b_j(o_t)$ είναι η συνάρτηση πυκνότητας πιθανότητας να εξάγεται μια παρατήρηση o από την κατάσταση j τη χρονική στιγμή t και $\alpha_i(t-1)$ είναι η πιθανότητα Forward για την κατάσταση i και τη χρονική στιγμή $t-1$.

Ο αλγόριθμος Backward δίνει την πιθανότητα να βρισκόμαστε τη χρονική στιγμή t στην κατάσταση j και η ακολουθία των παρατηρήσεων που ακολουθεί να είναι $o_{t+1}, ..., o_T$. Η πιθανότητα Backward υπολογίζεται από τον τύπο:

$$\beta_j(t) = P(o_{t+1}, ..., o_T|s_t = j, \lambda) \tag{Α΄.29}$$

Όπου λ είναι το HMM και s_t είναι η κατάσταση που βρισκόμαστε τη χρονική στιγμή t. Με την ίδια λογική όπως στην πιθανότητα Forward, η πιθανότητα Ba-

ckward , για ένα αριθμό καταστάσεων N, μπορεί να υπολογιστεί αποτελεσματικά από τον τύπο:

$$\beta_j(t) = \sum_{j=2}^{N-1} \alpha_{ij} b_j(o_{t+1}) \beta_j(t+1) \tag{Α΄.30}$$

Όπου N είναι ο αριθμός των καταστάσεων, α_{ij} είναι η πιθανότητα μετάβασης από την κατάσταση i στην κατάσταση j, $b_j(o_{t+1})$ είναι η συνάρτηση πυκνότητας πιθανότητας να εξάγεται μια παρατήρηση o από την κατάσταση j τη χρονική στιγμή $t+1$ και $b_j(o_{t+1})$ είναι η πιθανότητα Backward τη χρονική στιγμή $t+1$ για την κατάσταση j.

Από ότι φαίνεται από τις παραπάνω σχέσεις, η πιθανότητα Forward είναι από κοινού πιθανότητα, ενώ η πιθανότητα Backward είναι δεσμευμένη πιθανότητα. Αυτή η ασυμμετρία είναι σκόπιμη, ώστε να μπορούμε να υπολογίζουμε την πιθανότητα κάθε κατάστασης παίρνοντας το αποτέλεσμα των δύο αυτών πιθανοτήτων [25]. Έτσι βρίσκουμε την από κοινού πιθανότητα να έχουμε μια ακολουθία παρατηρήσεων O όταν βρισκόμαστε στην κατάσταση j για το συγκεκριμένο HMM που μας ενδιαφέρει, όπως φαίνεται από την παρακάτω σχέση:

$$\alpha_j(t)\beta_j(t) = P(O, s(t) = j|\lambda) \tag{Α΄.31}$$

Όπου λ είναι το HMM, O είναι η ακολουθία των παρατηρήσεων, $s(t)$ είναι η κατάσταση τη χρονική στιγμή t και $\alpha_j(t)$, $\beta_j(t)$ είναι οι πιθανότητες Forward και Backward, αντίστοιχα, για εκείνη τη χρονική στιγμή.

Ως εκ τούτου, προκύπτει η πιθανότητα $L_j(t)$ να βρισκόμαστε στην κατάσταση j τη χρονική στιγμή t, όταν η ακολουθία των παρατηρήσεων είναι O και το HMM είναι το λ:

$$L_j(t) = P(s(t) = j|O, \lambda) = \frac{P(O, s(t) = j|\lambda)}{P(O|\lambda)} = \frac{1}{P}\alpha_j(t)\beta_j(t) \tag{Α΄.32}$$

Όπου $P = P(O|\lambda)$, $s(t)$ είναι η κατάσταση που βρισκόμαστε τη χρονική στιγμή t και $\alpha_j(t)$, $\beta_j(t)$ είναι η πιθανότητες Forward και Backward, αντίστοιχα, για τη χρονική στιγμή t και την κατάσταση j.

Με γνώση των παραπάνω πιθανοτήτων, μπορούμε να υλοποιήσουμε τον αλγόριθμο Baum-Welch. Ο αλγόριθμος αυτός αποτελείται από τα παρακάτω βήματα [25]:

1. Αρχικοποίηση της μέσης τιμής και της διασποράς. Συνήθως χρησιμοποιούμε ένα πρωτότυπο HMM, του οποίου η κάθε κατάσταση έχει ένα διάνυσμα μέσης τιμής που ο κάθε συντελεστής του είναι 0 και ένα διάνυσμα διακύμανσης που ο κάθε συντελεστής του είναι 1.

2. Υπολογισμός των πιθανοτήτων Forward και Backward, για όλες τις καταστάσεις j και τις χρονικές στιγμές t.

3. Για κάθε κατάσταση j και χρονική στιγμή t, χρησιμοποιούμε την πιθανότητα $L_j(t)$ και το διάνυσμα παρατηρήσεων o_t, ώστε να υπολογίσουμε τη νέα μέση τιμή και διασπορά για την κατάσταση αυτή. Η νέα μέση τιμή και διασπορά υπολογίζονται, αντίστοιχα, από τις παρακάτω σχέσεις:

$$\mu_j = \frac{\sum_{t=1}^{T} L_j(t) o_t}{\sum_{t=1}^{T} L_j(t)} \tag{Α΄.33}$$

$$\Sigma_j = \frac{\sum_{t=1}^{T} L_j(t)(o_t - \mu_j)(o_t - \mu_j)'}{\sum_{t=1}^{T} L_j(t)} \tag{Α΄.34}$$

4. Υπολογισμός των παραμέτρων του HMM βάσει των αποτελεσμάτων των παραπάνω σχέσεων.

5. Αν η τιμή του $P = P(O|\lambda)$ είναι μεγαλύτερη από την προηγούμενη που είχε υπολογιστεί, σταματάμε και δίνουμε ως παραμέτρους αυτές που υπολογίστηκαν τελευταίες. Αλλιώς συνεχίζουμε τα βήματα 2-5.

Α΄.10 Αλγόριθμος Viterbi

Για ένα HMM λ, έστω $\phi_j(t)$ η εκτίμηση μέγιστης πιθανοφάνειας, να έχουμε την αλληλουχία παρατηρήσεων $o_1, ..., o_t$ όταν βρισκόμαστε στην κατάσταση j τη χρονική στιγμή t. Αυτή η πιθανότητα μπορεί να υπολογιστεί με την παρακάτω σχέση:

$$\phi_j(t) = \max_i \phi_i(t-1)\alpha_{ij} b_j(o_t) \tag{Α΄.35}$$

όπου

$$\phi_1(1) = 1 \tag{Α΄.36}$$

$$\phi_j(1) = \alpha_{1j} b_j(o_t) \tag{Α΄.37}$$

Όπου $1 < j < N$, α_{ij} είναι η πιθανότητα μετάβασης από την κατάσταση i στην κατάσταση j και $b_j(o_t)$ είναι η συνάρτηση πυκνότητας πιθανότητας να εξαχθεί η παρατήρηση o από την κατάσταση j τη χρονική στιγμή t. Η εκτίμηση μέγιστης πιθανοφάνειας $P(O|\lambda)$ δίνεται, τότε, από τον παρακάτω τύπο:

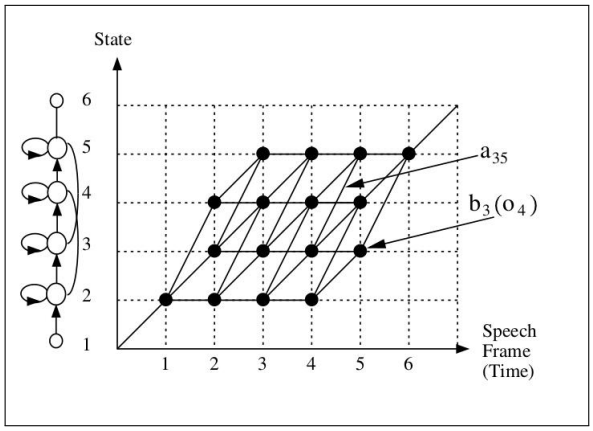
$$\phi_N(T) = \max_i \{\phi_i(T)\alpha_{iN}\} \tag{Α΄.38}$$

Επειδή χρειάζεται να γίνουν επανεκτιμήσεις (re-estimations), ο απευθείας υπολογισμός των παραπάνω σχέσεων θα οδηγήσει σε υποχείλιση (underflow). Γι΄ αυτό χρησιμοποιούμε λογαριθμικές πιθανοφάνειες (log likelihoods). Οπότε, παίρνουμε:

$$\psi_j(t) = \max_i \{\psi_i(t-1) + log(\alpha_{ij})\} + log(b_j(o_t)) \tag{Α΄.39}$$

Η παραπάνω σχέση είναι η βάση του αλγόριθμου Viterbi. Όπως φαίνεται στο παρακάτω σχήμα, ο αλγόριθμος αυτός βρίσκει το καλύτερο μονοπάτι σε ένα πίνακα, όπου η κάθετη διάσταση αντιπροσωπεύει τις καταστάσεις του HMM και η οριζόντια διάσταση αντιπροσωπεύει τα πλαίσια της ομιλίας (χρόνος). Κάθε μεγάλος κύκλος, στο σχήμα, είναι η λογαριθμική πιθανότητα να έχουμε το συγκεκριμένο πλαίσιο τη συγκεκριμένη χρονική στιγμή. Κάθε γραμμή μεταξύ των κύκλων είναι η λογαριθμική πιθανότητα μετάβασης. Η λογαριθμική πιθανότητα κάθε μονοπατιού υπολογίζεται αθροίζοντας τις λογαριθμικές πιθανότητες μετάβασης και τις

λογαριθμικές τελικές πιθανότητες του μονοπατιού. Η διαδρομή που διαγράφουν τα μονοπάτια είναι από τα αριστερά προς τα δεξιά. Κάθε χρονική στιγμή t το μερικό μονοπάτι (partial path) $\psi_i(t-1)$ είναι γνωστό σε όλες τις καταστάσεις i. Έτσι η παραπάνω σχέση υπολογίζει το τελικό μονοπάτι $\psi_j(t)$ επεκτεινόμενη, κάθε χρονική στιγμή, από το ένα μερικό μονοπάτι στο άλλο [25].



Σχήμα Α΄.8: Αλγόριθμος Viterbi, πηγή: [25]

Βιβλιογραφία

[1] Pirros Tsiakoulis, Alexandros Potamianos, Dimitrios Dimitriadis, *Spectral Moment Features Augmented by Low Order Cepstral Coefficients for Robust ASR.*

[2] Pirros Tsiakoulis, Alexandros Potamianos, Dimitrios Dimitriadis, *SHORT-TIME INSTANTANEOUS FREQUENCY AND BANDWIDTH FEATURES FOR SPEECH RECOGNITION.*

[3] Dimitris Dimitriadis, Petros Maragos, Alexandros Potamianos, *Robust AM-FM Features for Speech Recognition.*

[4] Alexandros Potamianos, Petros Maragos, *Time-frequency distributions for automatic speech recognition.*

[5] Alexandros Potamianos, Petros Maragos, *Speech formant frequency and bandwidth tracking using multiband energy demodulation.*

[6] Dimitrios Dimitriadis, Petros Maragos, *Robust Energy Demodulation Based on Continuous Models with Application to Speech Recognition.*

[7] D. Dimitriadis, N. Katsamanis, P. Maragos, G. Papandreou, V. Pitsikalis, *TOWARDS AUTOMATIC SPEECH RECOGNITION IN ADVERSE ENVIRONMENTS.*

[8] Bojan Kotnik, Damjan Vlaj, Zdravko Kačič, Bogomir Horvat, *ROBUST MFCC FEATURE EXTRACTION ALGORITHM USING EFFICIENT ADDITIVE AND CONVOLUTIONAL NOISE REDUCTION PROCEDURES.*

[9] Benjamin J. Shannon, Kuldip K. Paliwal, *MFCC Computation from Magnitude Spectrum of Higher Lag Autocorrelation Coefficients for Robust Speech Recognition.*

[10] Pirros Tsiakoulis, Alexandros Potamianos, *STATISTICAL ANALYSIS OF AMPLITUDE MODULATION IN SPEECH SIGNALS USING AN AM–FM MODEL.*

[11] Dimitrios Dimitriadis, Petros Maragos, *Continuous energy demodulation methods and application to speech analysis.*

[12] Lindasalwa Muda, Mumtaj Begam, I. Elamvazuthi, *Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques.*

[13] Md Jahangir Alam, Patrick Kenny, Douglas O'Shaughnessy, *A Study of Low-variance Multi-taper Features for Distributed Speech Recognition.*

[14] Beth Logan, *Mel-Frequency Cepstral Coefficients for Music Modeling.*

[15] Matthew Nicholas Stuttle, *A Gaussian Mixture Model Spectral Representation for Speech Recognition.*

[16] Qifeng Zhu, Markus Iseli, Xiaodong Cui, Abeer Alwan, *Noise Robust Feature Extraction for ASR using the Aurora 2 Database.*

[17] L. R. Rabiner, B. Juang, *Fundamentals Of Speech Recognition. Prentice Hall, Englewood Cliffs.*

[18] B.H. Juang, Lawrence R. Rabiner, *Automatic Speech Recognition – A Brief History of the Technology Development.*

[19] Seyed Hamidreza Mohammadi, Hossein Sameti, Amirhossein Tavanaei, Ali Soltani-Farani, *FILTER-BANK DESIGN BASED ON DEPENDENCIES BETWEEN FREQUENCY COMPONENTS AND PHONEME CHARACTERISTICS.*

[20] Zheng – Hua Tan, Borge Lindberg, *Automatic Speech Recognition on Mobile Devices and over Communication Networks.*

[21] http://mirlab.org/jang/books/audiosignalprocessing/speechFeatureMfcc.asp?title=12-2%20MFCC, *12-2 MFCC.*

[22] Dimitrios Dimitriadis, Petros Maragos, Alexandros Potamianos, *MODULATION FEATURES FOR SPEECH RECOGNITION.*

[23] Rok Gajsek, France Miheli˘c, *Comparison of speech parameterization techniques for Slovenian language.*

[24] 5th International Conference, ICISTM 2011, Gurgaon, India, March 2011, Proceedings, *Information Intelligence, Systems, Technology and Management.*

[25] Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, Phil Woodland, *The HTK Book.*

[26] Ji Ming, Baochun Hou, Queen's University Belfast, University of Hertfordshire, United Kingdom, *Speech Recognition in Unknown Noisy Conditions.*

[27] Kisun You, Hoyoun Kim, Wonyong Sung, *Implementation of an International Quality Assessment System for a Handheld Device.*

[28] Πύρρος Τσιάκουλης, *Σύνθεση φωνής με υπολογιστική αεροδυναμική ανάλυση του ανθρωπινού ηχητικού σωλήνα και σύγκριση με κλασσικές μεθόδους. Τελική τεχνική αναφορά για το ερευνητικό πρόγραμμα ΠΕΝΕΔ 2003.*

[29] Nafiz ARICA, Fatos. T. YARMAN-VURAL Department of Computer Engineering, METU-TURKEY, *A NEW HMM TOPOLOGY FOR SHAPE RECOGNITION.*

[30] Πύρρος Τσιάκουλης, *Σύνθεση φωνής με υπολογιστική αεροδυναμική ανάλυση του ανθρωπινού ηχητικού σωλήνα και σύγκριση με κλασσικές μεθόδους. Διδακτορική Διατριβή του Πύρρου Τσιάκουλη. Αθήνα, Απρίλιος 2010.*

[31] Debalina Ghosh, Depanwita Sarkar Debnath, Saikat Bose, *A Comparative Study of Performance of FPGA based Mel Filter Bank & Bark Filter Bank, Department of Microelectronics & VLSI Design, Techno India, SaltLake, Kolkata, India.*

[32] http://en.wikipedia.org/wiki/Speech_recognition, *Applications.*