Εθνικο Μετσοβιο Πολυτεχνειο
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Κατάτμηση και Κατηγοριοποίηση Φωνής, Θορύβου και Σιωπής με χρήση Μετρικών Κριτηρίων και Στατιστικών HMM/GMM Μοντέλων, με Εφαρμογές σε Ηχητικά Τμήματα από Δελτία Ειδήσεων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Αγγελικής Ν. Μεταλληνού

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

Αθήνα, Ιούλιος 2007

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Κατάτμηση και Κατηγοριοποίηση Φωνής, Θορύβου και Σιωπής με χρήση Μετρικών Κριτηρίων και Στατιστικών HMM/GMM Μοντέλων, με Εφαρμογές σε Ηχητικά Τμήματα από Δελτία Ειδήσεων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Αγγελικής Ν. Μεταλληνού

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 19 Ιουλίου 2007.

.....	.....	.....
Πέτρος Μαραγκός	Μιλτιάδης Αναγνώστου	Νεκτάριος Κοζύρης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2007.

.....

Αγγελική Ν. Μεταλληνού
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Στη Μαργαρίτα, την αδερφή μου, και στους γονείς μου

Ever Tried. Ever Failed. No Matter. Try Again. Fail Again. Fail Better.
Samuel Beckett

Περίληψη

Στην παρούσα διπλωματική εργασία μελετάται το πρόβλημα επεξεργασίας ηχητικών αρχείων και συγκεκριμένα το πεδίο του εντοπισμού αλλαγών και συνακόλουθης κατάτμησης των αρχείων αλλά και το πεδίο της στατιστικής μοντελοποίησης ηχητικών κλάσεων με σκοπό την κατηγοριοποίηση ηχητικών τμημάτων. Αρχικά γίνεται μία αναλυτική παρουσίαση των μεθόδων και των συστημάτων που παρουσιάζονται στην βιβλιογραφία για την κατάτμηση και την κατηγοριοποίηση ηχητικών αρχείων. Ακολούθως, αναφορικά με το πρόβλημα της κατάτμησης, παρουσιάζονται ποικίλα χαρακτηριστικά που μπορούν να εξαχθούν από το ηχητικό σήμα, κριτήρια κατάτμησης, με έμφαση στα μετρικά κριτήρια, και αλγόριθμοι κατάτμησης. Προτείνεται επίσης ένας νέος αλγόριθμος με σκοπό τη βελτίωση των αποτελεσμάτων της κατάτμησης. Αναφορικά με το πρόβλημα της στατιστικής μοντελοποίησης με στόχο την κατηγοριοποίηση, παρουσιάζονται τα βήματα και οι σχεδιαστικές αποφάσεις για τη δημιουργία ενός συστήματος κατάτμησης και κατηγοριοποίησης βασισμένου σε στατιστικά Κρυφά Μαρκοβιανά Μοντέλα και Γκαουσιανά Μοντέλα Μιγμάτων. Στη συνέχεια αναλύονται, υλοποιούνται και συγκρίνονται θεωρητικά και πειραματικά διάφορες προσεγγίσεις για την κατάτμηση και κατηγοριοποίηση τμημάτων και εισάγεται μία καινούρια ιδέα, η έννοια των καμπύλων ποσοστών. Εντέλει, παρουσιάζεται ένα συνολικό σύστημα που συνδυάζει τις επιμέρους υπομονάδες που υλοποιήθηκαν και μπορεί να εφαρμοστεί για την κατάτμηση και κατηγοριοποίηση πραγματικών ηχητικών αρχείων από δελτία ειδήσεων.

Λέξεις κλειδιά

Σηματοδότηση Ηχητικών Αρχείων, Κατάτμηση Ηχητικών Αρχείων, Μετρικά Κριτήρια Κατάτμησης, Bayesian Information Criterion, Mel Frequency Cepstral Coeffiecients, Teager Energy Cepstral Coefficients, Φράκταλς, Φράκταλ Διάσταση Ηχητικού Σήματος, Κατηγοριοποίηση Ηχητικών Αρχείων, Στατιστική Μοντελοποίηση, Κρυφά Μαρκοβιανά Μοντέλα, Γκαουσιανά Μοντέλα Μιγμάτων, Δελτία Ειδήσεων.

Abstract

This diploma thesis deals with the problem of audio stream processing and specifically with the issue of event detection and subsequent audio stream segmentation as well as the issue of statistical modeling of audio classes for classification. Initially, we present the most prominent methods and systems that appear in the literature for audio stream segmentation and classification. Afterwards, concerning the issue of segmentation, we present various features that can be extracted from the audio signal, various segmentation criteria, with emphasis on metric-based criteria, and segmentation algorithms. We also propose a new algorithm in order to improve audio segmentation results. Concerning the issue of statistical modeling for audio classification, we present the basic steps and design decisions for the creation of an audio segmentation and classification system that is based on the use of Hidden Markov Models and Gaussian Mixture Models. Afterwards, we study, implement and compare, both theoretically and experimentally, various methods for audio stream segmentation and classification and we propose a new idea, the percentage curves. Finally, we present a complete system that combines the subsystems that have been implemented and can be used for segmentation and classification of real audio streams from broadcast news.

Keywords

Audio Diarization, Audio Segmentation, Metric-Based Segmentation Criteria, Bayesian Information Criterion, Mel Frequency Cepstral Coefficients, Teager Energy Cepstral Coefficients, Fractals, Fractal Dimension of an Audio Signal, Audio Classification, Statistical Modeling, Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), Broadcast News.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή Πέτρο Μαραγκό, επιβλέποντα της διπλωματικής μου εργασίας, για την εμπιστοσύνη που μου έδειξε ήδη από πολύ νωρίς στη διάρκεια της φοιτητικής μου σταδιοδρομίας, για τις ανεκτίμητες συμβουλές του, που ήταν απαραίτητες για την επιτυχή ολοκλήρωση της παρούσας εργασίας, αλλά και για την καθοδήγησή του τόσο σε θέματα έρευνας όσο και σε θέματα σχετικά με τη μελλοντική μου σταδιοδρομία. Επίσης, θα ήθελα να ευχαριστήσω τον μεταδιδακτορικό ερευνητή Δημήτρη Δημητριάδη για την αδιάκοπη καθοδήγηση και βοήθειά του στην εκπόνηση της εργασίας αυτής. Επιπλέον θα ήθελα να ευχαριστήσω τους υποψήφιους διδάκτορες Νάσσο Κατσαμάνη και Βασίλη Πιτσικάλη (post doc πλέον) αλλά και όλους του υποψήφιους διδάκτορες του Εργαστηρίου Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων για τις ανεκτίμητες συμβουλές και την υποστήριξη που μου προσέφεραν. Αισθάνομαι ότι η εμπειρία της εργασίας μου στο Εργαστηρίου Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων θα με καθοδηγεί και θα με βοηθάει στην μελλοντική μου πορεία.

Θα ήθελα θερμά να ευχαριστήσω τους γονείς μου που με την αγάπη και την καθοδήγηση που μου προσέφεραν με διαμόρφωσαν ως άνθρωπο και έκαναν δυνατά όσα έχω μέχρι τώρα καταφέρει ή προσπαθήσει στη ζωή μου. Τους ευχαριστώ από την καρδιά μου για την υποστήριξη και την υπομονή τους. Ιδιαίτερα θέλω να ευχαριστήσω τη Μαργαρίτα, τη μικρή μου αδερφή, που πάντα με αγαπάει και πάντα με προστατεύει. Στη Μαργαρίτα και στους γονείς μου θα ήθελα να αφιερώσω την παρούσα διπλωματική εργασία

Θα ήθελα να ευχαριστήσω τους φίλους μου για την αγάπη, τη φιλία, την υποστήριξη και την κατανόησή τους, χωρίς τους οποίους η ζωή μου θα ήταν άδεια. Τους ευχαριστώ για τις ωραίες συζητήσεις, τα γέλια, τα ξενύχτια, τα ταξίδια και τις εμπειρίες και νοιώθω ότι αν και μπορεί στο μέλλον να βρισκόμαστε μακριά στην πραγματικότητα θα είμαστε πάντα κοντά.

Τέλος, ευχαριστώ τον αναγνώστη της εργασίας αυτής για το ενδιαφέρον και την προσοχή του.

Περιεχόμενα

Λίστα Πινάκων

Λίστα Εικόνων

Κεφάλαιο 1

Εισαγωγή

1.1 Ορισμός του Προβλήματος και Εφαρμογές

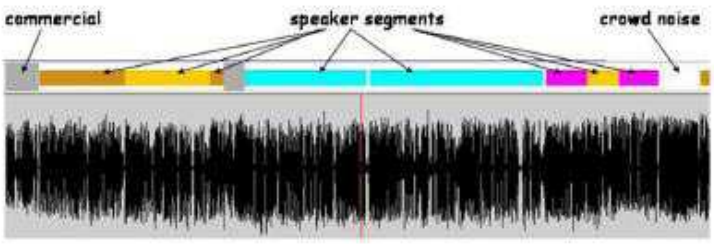
Η σύγχρονη εποχή χαρακτηρίζεται από ραγδαία τεχνολογική πρόοδο και από ολοένα αυξανόμενες δυαντότητες παραγωγής και διακίνησης τεράστιου όγκου πληροφορίας. Σε αυτή τη νέα εποχή της πληροφορίας που αναπτύσσεται μπροστά στα μάτια μας, γίνεται ολοένα και πιο σαφές το γεγονός ότι η γνώση είναι δύναμη και η πρόσβαση στην πληροφορία είναι ενα ανεκτίμητο πλεονέκτημα. Την ίδια στιγμή η ύπαρξη τεράστιου διαθέσιμου όγκου μη επεξεργασμένης πληροφορίας δημιουργεί νέες προκλήσεις. Ίσως το μεγάλο στοίχημα για την σύγχρονη τεχνολογία δεν είναι τόσο η ελεύθερη διακίνηση και πρόσβαση στην πληροφορία όσο η αποτελεσματική διαχείρισή της, ώστε να είναι δυνατή η εξαγωγή χρήσιμων συμπερασμάτων με αποδοτικό και έξυπνο τρόπο.

Όσον αφορά το πεδίο της επεξεργασίας σημάτων και ήχου, είναι δυνατή τόσο η διακίνηση όσο και η αποθήκευση τεράστιου όγκου πληροφορίας που βρίσκεται σε μορφή αρχείων ήχου, όπως δελτία ειδήσεων, ηχητικά μηνύματα ή συνομιλία από audio conferencing. Η αποτελεσματική διαχείριση των αρχείων αυτών και η εξαγωγή κατάλληλης σημασιολογικής πληροφορίας αλλά και μετά-πληροφορίας σχετικά με την ύπαρξη θορύβου/μουσικής στο αρχείο ή σχετικά με το πλήθος και την ταυτότητα των ομιλητών αποτελεί μία πρόκληση για την έρευνα στο πεδίο επεξεργασίας ηχητικών αρχείων. Επιπλέον, η έρευνα τρόπων εξόρυξης πληροφορίας δεν περιορίζεται σε ηχητικά αρχεία αλλά μπορεί να εφαρμοστεί ευρέως και στο γενικότερο πρόβλημα εξόρυξης δεδομένων από πολυμεσικές εφαρμογές, όπου για παράδειγμα η πληροφορία από το ηχητικό σήμα θα συνδυάζεται κατάλληλα με την οπτική πληροφορία ενός video αρχείου.

Παρακάτω θα περιγραφεί το γενικότερο πρόβλημα κατηγοριοποιήσης ηχητικού αρχείου

(audio diarization) καθώς και τα πιο ειδικά προβλήματα εντοπισμού γεγονότων σε ηχητικό αρχείο (event detection) και στατιστικής μοντελοποίησης ηχητικών κλάσεων με χρήση Κρυφών Μαρκοβιανών Μοντέλων με τελικό στόχο την κατηγοριοποίηση ηχητικών υποτμημάτων.

Γενικά ένα ηχητικό αρχείο (audio stream) αποτελείται από καταγεγραμμένους ήχους διάφορων ηχητικών πηγών, όπως διαφορετικοί ομιλητές, τμήματα μουσικής, διάφοροι τύποι θορύβου κλπ. Ο όρος audio diarization αναφέρεται στη σηματοδότηση και κατηγοριοποίηση αυτών των audio-streams. Στην απλούστερη περίπτωση θα κατηγοριοποιούσαμε το αρχείο σε περιοχές ομιλίας και μη ομιλίας ενώ σε μία πιο λεπτομερή περίπτωση diarization θα κατηγοριοποιούσαμε τα τμήματα μη ομιλίας σε μουσική, θόρυβο κλπ, ενώ στα τμήματα ομιλίας θα σημειώναμε επίσης τις αλλαγές ομιλητών και θα συσχετίζαμε κάθε τμήμα ομιλίας με ένα συγκεκριμένο ομιλητή. Η τελευταία εργασία ονομάζεται speaker diarization. Η δυσκολία του προβλήματος καθορίζεται επιπλέον και από την εκ των προτέρων γνωστή πληροφορία που πιθανόν έχουμε για τα audio streams που εξετάζουμε. Για παράδειγμα, γνώση της δομής του audio stream (πχ μουσική ακολουθούμενη από ομιλία) ή γνώση του πλήθους των ομιλητών διευκολύνει την αντιμετώπιση του προβλήματος. Στην εικόνα 1.1 φαίνεται ένα παράδειγμα diarization ενός audio stream.



Εικόνα 1.1: Παράδειγμα diarization για ένα audio stream ([45])

Το πρόβλημα εντοπισμού γεγονότων (event detection) σε ένα audio stream είναι ένα υποπρόβλημα του audio diarization το οποίο επικεντρώνεται στον εντοπισμό σημείων που αντιστοιχούν στην αρχή κάποιας αλλαγής, για παράδειγμα μετάβαση από ομιλία σε μουσική ή αλλαγή ομιλητών ή αρχή χειροκροτημάτων κλπ. Ο ορισμός μίας σημαντικής αλλαγής διαφέρει ανάλογα με τη εφαρμογή και εξαρτάται από το πόσο ευαίσθητο θέλουμε να είναι το σύστημά μας. Σκοπός του εντοπισμού γεγονότων είναι να αντιστοιχίσει ένα αρχείο φωνής σε μία ακολουθία ομογενών ακουστικά τμημάτων, όπου ο όρος ομογενής εξαρτάται από τις ηχητικές κλάσεις που έχουμε ορίσει στο πρόβλημά μας.

Το επόμενο βήμα είναι η κατηγοριοποίηση των τμημάτων αυτών σε κατάλληλες κατηγορίες όπως ομιλία, θόρυβος, μουσική. Για το πρόβλημα της κατηγοριοποίησης μία ευρέως διαδεδομένη και αποτελεσματική μεθοδολογία είναι η στατιστική μοντελοποίηση των κατηγοριών του προβλήματος με χρήση μοντέλων όπως τα Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models) και τα Μοντέλα Γκαουσιανών Μιγμάτων (Gaussian Mixture Models). Τα στατιστικά μοντέλα εκπαιδεύονται με χρήση δειγμάτων της κατηγορίας που αντιπροσωπεύουν και χρησιμοποιούνται για την κατηγοριοποίηση νέων δειγμάτων που εισέρχονται στο σύστημα.

Τόσο το πρόβλημα της κατάτμησης ηχητικών αρχείων όσο και το πρόβλημα της κατηγοριοποίησης των υποτμημάτων με χρήση στατιστικών μοντέλων θα εξεταστούν στην παρούσα εργασία. Τελικός στόχος είναι η υλοποίηση ενός αποδοτικού συστήματος που θα μπορεί να εφαρμοστεί για το πρόβλημα diarization με πραγματικά δεδομένα από δελτία ειδήσεων.

Η ύπαρξη συστημάτων ικανών να εκτελέσουν αποτελεσματικό diarization και speaker diarization σε μη επεξεργασμένα ηχητικά αρχεία, βρίσκει πληθώρα εφαρμογών τόσο σε θέματα αποθήκευσης και μεταφοράς πληροφορίας, όσο και σε θέματα αναγνώρισης ομιλίας. Επίσης τέτοια συστήματα σχετίζονται με εφαρμογές αποδοτικής εξόρυξης πληροφορίας και αποτελεσματικού χειρισμού της.

Για παράδειγμα, πολλές σύγχρονες εφαρμογές Internet βασίζονται στη μεταφορά ηχητικής πληροφορίας μέσω διαδικτύου χρησιμοποιώντας Voice Over IP πρωτόκολλα (VoIP). Η αναγνώριση των ηχητικών τμημάτων που δεν περιέχουν χρήσιμη πληροφορία, όπως τα τμήματα θορύβου, και η απόρριψη τους από το σύστημα θα μπορούσε να βοηθήσει ώστε να μεταφέρεται μόνο η πραγματικά χρήσιμη πληροφορία και, κατά συνέπεια, να αυξηθεί η ταχύτητα και να μειωθεί το απαιτούμενο εύρος ζώνης τέτοιων εφαρμογών.

Ομοίως, ένα σύστημα ικανό να απορρίπτει την άχρηστη πληροφορία, όπως ο θόρυβος, θα έβρισκε εφαρμογή στην προεπεξεργασία ηχητικών δεδομένων που συλλέγονται για αποθήκευση. Κατά συνέπεια, θα μπορούσαν να αποθηκεύονται μόνο τα χρήσιμα ηχητικά δεδομένα, μειώνοντας τον απαιτούμενο διαθέσιμο αποθηκευτικό χώρο αλλά και την ταχύτητα εύρεσης κάποιου ηχητικού τμήματος.

Επιπλέον ένα σύστημα διαχωρισμού ηχητικής πληροφορίας σε ομιλία, θόρυβο και μουσική θα μπορούσε να είναι το πρώτο στάδιο ενός συστήματος αναγνώρισης ομιλίας ή/και μουσικής. Η αρχική κατηγοριοποίηση των ηχητικών υποτμημάτων είναι απαραίτητη ώστε να οδηγούνται στο σύστημα αναγνώρισης ομιλίας μόνο τα τμήματα ομιλίας και στο σύστημα αναγνώρισης μουσικής μόνο τα τμήματα μουσικής.

Το ευρύτερο πρόβλημα audio-diarization έχει πληθώρα χρήσιμων εφαρμογών σε πεδία

εξόρυξης πληροφορίας. Για παράδειγμα, για την εφαρμογή diarization σε δελτία ειδήσεων, ένα αποτελεσματικό σύστημα θα έδινε τη δυνατότητα να εντοπίσουμε γρήγορα συγκεκριμένο ομιλητή μέσα από μία βάση δεδομένων δελτίων. Επίσης θα προσέφερε τη δυνατότητα αυτόματης εύρεσης ορισμένων λέξεων κλειδιών και τμημάτων ομιλίας που σχετίζονται με ένα επιλεγμένο θέμα. Επιπλέον καθίσταται δυνατή η εξαγωγή στατιστικών συμπερασμάτων και μετα-πληροφορίας από τη βάση δεδομένων των δελτίων ειδήσεων. Για παράδειγμα θα μπορούσαν να τεθούν ερωτήσεις όπως ποια είναι τα θέματα που απασχόλησαν για περισσότερο από x λεπτά τα δελτία ειδήσεων ορισμένων σταθμών σε μία ορισμένη χρονική περίοδο, ή ποιοι ομιλητές ακούστηκαν πιο συχνά σε δελτία ειδήσεων σε κάποια συγκεκριμένη χρονική περίοδο. Τα παραπάνω διευκολύνουν το γρήγορο, έξυπνο και αποτελεσματικό χειρισμό μεγάλου όγκου ηχητικής πληροφορίας.

Συμπερασματικά, η αποδοτική επεξεργασία μεγάλου όγκου πληροφορίας είναι μία από τις σημαντικότερες τεχνολογικές προκλήσεις της σύχρονης εποχής. Η μελέτη του προβλήματος επεξεργασίας ηχητικών αρχείων, ώστε να είναι δυνατή η αυτόματη εξόρυξη ποικίλων μορφών πληροφορίας από αυτά, είναι μία περιοχή με μεγάλο ερευνητικό ενδιαφέρον και πληθώρα σημαντικών πρακτικών εφαρμογών.

1.2 Συνοπτική Περιγραφή των Τεχνολογικών Προβλημάτων που εξετάστηκαν στην παρούσα εργασία

Στην παρούσα εργασία εξετάζουμε τα προβλήματα εντοπισμού γεγονότων σε ηχητικό αρχείο (event detection) και στατιστικής μοντελοποίησης ηχητικών κλάσεων με χρήση Κρυφών Μαρκοβιανών Μοντέλων με τελικό στόχο την κατηγοριοποίηση ηχητικών υποτμημάτων.

Σχετικά με το πρόβλημα εντοπισμού γεγονότων και συνακόλουθης κατάτμησης ενός ηχητικού αρχείου μελετάμε χαρακτηριστικά που μπορούν να εξαχθούν από το σήμα αλλά και κριτήρια εντοπισμού αλλαγών με έμφαση στα μετρικά κριτήρια. Επιπλέον, παρουσιάζουμε και υλοποιούμε αλγορίθμους κατάτμησης που χρησιμοποιούν διάφορα διανύσματα χαρακτηριστικών και διάφορα μετρικά κριτήρια. Επίσης προτείνεται ένας νέος αλγόριθμος, ο οποίος ονομάζεται αλγόριθμος δεύτερου περάσματος και σκοπός του είναι η επικύρωση των αλλαγών που βρίσκονται από κάποιον προηγούμενο αλγόριθμο εντοπισμού αλλαγών. Στο πειραματικό μέρος, συγκρίνεται η απόδοση ποικίλων διανυσμάτων χαρακτηριστικών στο πρόβλημα της κατάτμησης. Επίσης, μελετάται κατά πόσο ο αλγόριθμος δεύτερου περάσματος επιτυγχάνει να βελτιώσει τα αποτελέσματα των υπάρχντων αλγορίθμων που

εξετάζονται.

Σχετικά με το πρόβλημα στατιστικής μοντελοποίησης, υλοποιείται ένα πλήρες σύστημα για την κατάμηση και κατηγοριοποίηση audio-streams με χρήση HMM και GMM μοντέλων. Αρχικά παρουσιάζονται αναλυτικά, τα βήματα και οι σχεδιαστικές αποφάσεις που πρέπει να ληφθούν για τη σχεδίαση και την υλοποίηση ενός τέτοιου συστήματος, καθώς και τα λογισμικά εργαλεία που χρησιμοποιούνται. Στη συνέχεια παρουσιάζονται διάφορες παραλλαγές του συστήματος αυτού, που χρησιμοποιούν διαφορετικούς, περισσότερο ή λιγότερο σύνθετους αλγορίθμους κατάτμησης και κατηγοριοποίησης. Μία από τις παραλλαγές που παρουσιάζονται βασίζεται σε μία καινούρια ιδέα που εισάγεται σε αυτή την εργασία, τις καμπύλες ποσοστών. Στο πειραματικό μέρος, συγκρίνεται η απόδοση διάφρων παραλλαγών μέσω πειραμάτων τόσο σε συνθετικά δεδομένα όσο και σε πραγματικά δεδομενα από δελτία ειδήσεων.

Τέλος, παρουσιάζεται ένα συνολικό σύστημα που συνδυάζει τα υποσυστήματα κατάτμησης audio-stream και κατάτμησης/κατηγοριοποίησης με χρήση στατιστικών μοντέλων. Εξηγείται πως τα σύστημα θα μπορούσε να χρησιμοποιηθεί σε πραγματικά δελτία, ποιες παραλλαγές του θα μπορούσαν να δοκιμαστούν και δίνονται ιδέες σχετικά με τις πιθανές μελλοντικές επεκτάσεις του συστήματος.

1.3 Συνοπτική Περιγραφή των Αποτελεσμάτων της παρούσας εργασίας

Η παρούσα εργασία επιχειρεί μία μελέτη της περιοχής κατάτμησης και κατηγοριοποίησης ηχητικών αρχείων, δίνοντας βάρος στον εντοπισμό αλλαγών με χρήση μετρικών κριτήριων αλλά και στην κατηγοριοποίηση ηχητικών τμημάτων με χρήση στατιστικών μοντέλων. Επίσης, στα πλαίσια της εργασίας υλοποιήθηκε ένα πλήρες σύστημα κατάτμησης και κατηγοριοποίησης ηχητικών αρχείων που μπορεί να εφαρμοστεί σε αρχεία από δελτία ειδήσεων.

Σχετικά με το πρόβλημα εντοπισμού αλλαγών (event detection) υλοποιήθηκε ένα σύστημα κατάτμησης ηχητικών αρχείων που λειτουργεί σε δύο στάδια κατάτμησης. Στο πρώτο στάδιο χρησιμοποιείται ένας αλγόριθμος κατάτμησης που υπάρχει στη βιβλιογραφία και χρησιμοποιεί το μετρικό κριτήριο BIC και άλλα μετρικά κριτήρια, για τον εντοπισμό αλλαγών. Στο δεύτερο στάδιο χρησιμοποιείται ένας νέος αλγόριθμος που εισάγεται στην παρούσα εργασία και σκοπός του είναι να επικυρώσει της αλλαγές του πρώτου σταδίου δηλαδή να αποφασίσει αν είναι πραγματικές αλλαγές. Δοκιμάζονται ποικίλα διανύσματα χαρακτηριστικών και για τα δύο στάδια. Το σύστημα στην τελική μορφή του, σύμφωνα με

τα πειραματικά αποτελέσματα σε πραγματικά δεδομένα από δελτία ειδήσεων, επιτυγχάνει να εντοπίσει με επιτυχία τις αλλαγές μεταξύ ομιλίας και μη ομιλίας ενώ εντόπίζει ικανοποιητικά τις αλλαγές μεταξύ ομιλητών. Επίσης τα αποτελέσματα υποδεικνύουν ότι το δεύτερο στάδιο επιτυγχάνει να βετλιώσει τα αποτελέσματα του πρώτου σταδίου, απορρίπτοντας αλλαγές που δεν αντιστοιχούν σε πραγματικές αλλαγές.

Σχετικά με το πρόβλημα κατάτμησης και κατηγοριοποίησης με χρήση στατιστικών μοντέλων, υλοποιείται ένα σύστημα βασισμένο σε HMM και GMM μοντέλα το οποίο μπορεί να συνδυαστεί με το προηγούμενο σύστημα κατάτμησης και να εφαρμοστεί σε πραγματικά ηχητικά αρχεία από δελτία ειδήσεων. Το σύστημα αυτό λειτουργεί με διάφορους εναλλακτικούς αλγόριθμους που είτε κατατάσσουν στην κατάλληλη κλάση τα υποτμήματα που δέχονται από το σύστημα κατάτμησης είτε επιχειρούν μία πιο λεπτομερή κατάτμηση και κατηγοριοποίηση του αρχείου. Επίσης εισάγεται μία νέα ιδέα, η έννοια των καμπύλων ποσοστων, και παρουσιάζονται αλγόριθμοι για την επεξεργασία ηχητικών αρχείων που βασίζονται στην ιδέα των καμπύλων ποσοστών. Τα πειραματικα αποτελέσματα υποδεικνύουν ότι η μεθοδος κατάτμησης/κατηγοριοποίησης με χρήση καμπύλων ποσοστών δίνει πιο ομαλό και καλύτερης ποιότητας τελικό αποτέλεσμα από το αποτέλεσμα που δίνουν απλούστερες μέθοδοι επεξεργασίας ενώ σε κάποιες περιπτώσεις οδηγεί και σε μεγαλύτερα ποσοστά επιτυχίας. Επίσης τα πειράματα μας δείχνουν ότι το σύστημα λειτουργεί πολύ καλά στο πρόβλημα της κατηγοριοποίησης ηχητικών αρχειων σε ομιλία και μη ομιλία και λειτουργεί λιγότερο ικανοποιητικά στο πιο δύσκολο πρόβλημα κατάταξης σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία. Εντούτοις, το σύστημα έχει πιθανόν περιθώρια βελτίωσης όταν θα είναι διαθέσιμα περισσότερα δεδομένα για την εκπαίδευση των στατιστικών μοντέλων.

Συμπερασματικά, στην παρούσα εργασία εισάγονται δύο νέες ιδέες, ο αλγόριθμος του δεύτερου σταδίου για τον εντοπισμό των αλλαγών και η ιδέα των καμπύλων ποσοστών για χρήση στο σύστημα απόφασης με τα στατιστικά μοντέλα. Επίσης γίνεται προσπάθεια να συνδυαστούν αποτελεσματικά οι νέες ιδέες με τους υπάρχοντες αλγορίθμους για τη δημιουργία ενός πλήρους συστήματος κατάτμησης και κατηγοριοποίηση audio-streams απο δελτία ειδήσεων. Το τελικό σύστημα λειτουργεί καλά για το πρόβλημα εντοπισμού αλλαγών μεταξύ ομιλίας και μη ομιλίας και κατάταξης των τμημάτων σε ομιλία και μη ομιλία. Μελλοντικά πρέπει να γίνουν βελτιώσεις ώστε το σύστημα να χειριστεί περισσότερο ικανοποιητικά δυσκολοτερα προβλήματα όπως ο εντοπισμός αλλαγών ομιλητή και η κατάταξη των ηχητικών τμημάτων σε πιο λεπτομερείς ηχητικές κλάσεις.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

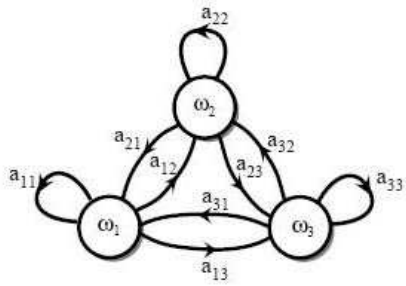
2.1 HMM και GMM μοντέλα

Τα μαρκοβιανά μοντέλα είναι στοχαστικές διαδικασίες που έχουν βρει μεγάλη εφαρμογή σε προβλήματα αναγνώρισης ήχου και φωνής.

2.1.1 Μαρκοβιανά μοντέλα πρώτης τάξης

Θεωρούμε μία ακολουθία καταστάσεων σε διαδοχικούς χρόνους, όπου η κατάσταση σε χρόνο t συμβολίζεται με ω(t). Μία συγκεκριμένη ακολουθία μήκους T συμβολίζεται με $\omega^T = \{\omega(1), \omega(2), \ldots, \omega(T)\}$. Σημειώνουμε ότι το σύστημα μπορεί να επισκεπτεί ξανά μία κατάσταση και ότι δεν είναι απαραίτητο να επισκεπτεί όλες τις καταστάσεις.

Το μοντέλο μας για την παραγωγή μίας ακολουθίας περιγράφεται από πιθανότητες μετάβασης $P(\omega_j(t+1)|\omega_i(t)) = a_{ij}$, δηλαδή τη χρονικά εξαρτημένη πιθανότητα του να έχουμε την κατάσταση ω_j σε χρόνο $t+1$ δεδομένου ότι έχουμε την κατάσταση ω_i σε χρόνο t. Δεν είναι απαραίτητο να υπάρχει συμμετρία στις πιθανότητες μετάβασης (γενικά $a_{ij} \neq a_{ji}$) και το σύστημα μπορεί να επισκεπτεί μία συγκεκριμένη κατάσταση πολλές φορές διαδοχικά (γενικά $a_{ij} \neq 0$). Το μοντέλο ονομάζεται πρώτης ταξης καθώς η πιθανότητα μετάβασης μεταξύ καταστάσεων σε χρόνο $t+1$ εξαρτάται μόνο από την πιθανότητα μετάβασης σε χρόνο t. Ένα τέτοιο μοντέλο φαίνεται στην εικόνα 2.1.



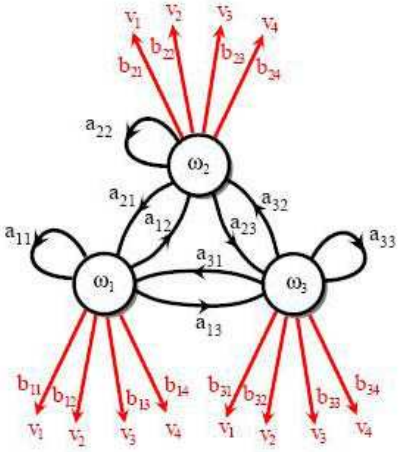
Εικόνα 2.1: Ένα μαρκοβιανό μοντέλο πρώτης τάξης ([12])

2.1.2 Κρυφά Μαρκοβιανά μοντέλα πρώτης τάξης-Hidden Markov Models (HMMs)

Θεωρούμε και σε αυτή την περίπτωση ότι σε κάθε χρονική στιγμή t το μοντέλο βρίσκεται σε μία κατάσταση ω(t) αλλά θεωρούμε επιπλέον ότι το μοντέλο εκπέμπει κάποιο ορατό σύμβολο υ(t). Όπως και με τις καταστάσεις, έχουμε και μία ακολουθία ορατών συμβόλων $V^T = \{\upsilon(1), \upsilon(2), \ldots, \upsilon(T)\}$. Σε κάθε κατάσταση έχουμε και μία πιθανότητα εκπομπής ενός συγκεκριμένου ορατού συμβόλου $\upsilon_k(t)$, δηλαδή $P(\upsilon_k(t)|\omega_j(t)) = b_{jk}$. Επειδή έχουμε πρόσβαση μόνο στην ακολουθία ορατών συμβόλων V^T ενώ η ακολουθία καταστάσεων ω^T δεν είναι παρατηρήσιμη, ονομάζουμε το μοντέλο Κρυφό Μαρκοβιανό Μοντέλο (Hidden Markov Model-HMM). Ένα τέτοιο μοντέλο φαίνεται στην εικόνα 2.2.

Απαιτούμε να υπάρχει πάντα κάποια μετάβαση από το χρονική στιγμή t στην $t+1$ (ακόμα και προς την ίδια κατάσταση), συνεπώς έχουμε την συνθήκη κανονικοποιησης $\sum_j a_{ij} = 1$ για όλα τα j.

Αν η διαδικασία εκπομπής συμβόλων εξόδου είναι διακριτή τότε και η κατανομή των πιθανοτήτων εξόδου είναι διακριτή και ισχύει: $\sum_k b_{jk} = 1$. Στην περίπτωση αυτή έχουμε ένα διακριτό HMM. Διαφορετικά, έχουμε ένα συνεχές HMM και οι κατανομές εξόδου είναι οι από κοινού συναρτήσεις πυκνότητας πιθανότητας ενός τυχαίου διανύσματος $\upsilon(t)$, με τιμή $b_j(\upsilon(t))$,όπου $\upsilon(t) = [\upsilon_1(t), \upsilon_2(t), \ldots, \upsilon_d(t)]^T$, όπου d είναι η διάσταση του $\upsilon(t)$. Κάποιες από τις κατηγορίες των συνεχών HMMs είναι τα Gaussian HMMs και Gaussian Mixture HMMs.



Εικόνα 2.2: Ένα κρυφό μαρκοβιανό μοντέλο. ([12])

2.1.3 Gaussian Hidden Markov Models

Τα Gaussian HMMs είναι συνεχή HMMs με γκαουσιανές κατανομές εξόδου. Δηλαδή αν συμβολίσουμε τη μεταβλητή $\upsilon(t)$ με x_t, έχουμε:

$$b_j(x_t) = N(x_t; \mu_j; \Sigma_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\{(1/2)\sum_{k=1}^{d} \frac{(x_{kt} - \mu_{jk})^2}{\sigma_{jk}^2}\}$$

με παραμέτρους τις μέσες τιμές μ_j, τον πίνακα συμμεταβλητότητας Σ_j και d είναι η διάσταση του $\upsilon(t)$. Ο αριθμός των παραμέτρων είναι:

$$d + \frac{d \times (d+1)}{2}$$

Συχνά χρησιμοποιείται διαγώνιος πίνακας συμμεταβλητότητας άρα οι κατανομές εξόδου γράφονται στη μορφή:

$$b_j(x_t) = \frac{1}{(2\pi)^{d/2} \prod_{k=1}^{d} \sigma_{jk}} \exp\{(1/2)\sum_{k=1}^{d} \frac{(x_{kt} - \mu_{jk})^2}{\sigma_{jk}^2}\}$$

όπου σ_{jk} για $k = 1, \ldots, d$ είναι τα διαγώνια στοιχεία του $d \times d$ διαγώνιου πίνακα συμμεταβλητότητας Σ.

2.1.4 Gaussian Mixture Hidden Markov Models

Επειδή μία απλή γκαουσιανή κατανομή μπορεί να μην επαρκεί για να μοντελοποιήσει την κατανομή του x σε μία κατάσταση, χρησιμοποιούνται μοντέλα στα οποία η κατανομές εξόδου είναι γραμμικοί συνδυασμοί γκαουσιανών κατανομών που ονομάζονται μίγματα. Τέτοια μοντέλα ονομάζονται Gaussian Mixture HMMs και οι κατανομές εξόδου τους περιγράφονται από τη σχέση:

$$b_j(x_t) = \sum_{m=1}^{M} c_{jm} N(x_t; \mu_{jm}; \Sigma_{jm})$$

όπου M είναι το πλήθος των γκαουσιανών μιγμάτων και $\sum_{m=1}^{M} c_{jm} = 1$.

2.1.5 Gaussian Mixture Models - GMMs

Σε προβλήματα αναγνώρισης ή κατηγοριοποίησης όπου δεν είναι ιδιαίτερα σημαντική η χρονική εξέλιξη του αρχείου ήχου (για παράδειγμα θέλουμε να μάθουμε αν ένα audio stream περιέχει φωνή ή μουσική αλλά δεν μας ενδιαφέρει να κάνουμε αναγνώριση φωνής) μπορούμε να χρησιμοποιήσουμε Gaussian Mixture HMMs με μία μόνο κατάσταση. Τα μοντέλα αυτά ονομάζονται Gaussian Mixture Models ή GMMs και σκοπός τους είναι η κατηγοριοποίηση της μοναδικής τους κατάστασης και όχι η εύρεση μίας πιθανής ακολουθίας καταστάσεων.

2.1.6 Βασικά Προβλήματα των HMMs

Τα 3 βασικά προβλήματα που σχετίζονται με τα HMMs παρουσιάζονται συνοπτικά παρακάτω:

Το πρόβλημα της Αξιολόγησης (Scoring) Θεωρούμε ότι έχουμε ένα HMM με πλήρως ορισμένες πιθανότητες μετάβασης a_{ij} και πιθανότητες εξόδου b_{jk}. Υπολογίστε την πιθανότητα μία συγκεκριμένη ακολουθία ορατών καταστάσεων V^T να προέρχεται από αυτό το μοντέλο.

Το πρόβλημα της Αποκωδικοποίησης (Decoding - State Estimation) Θεωρούμε ότι έχουμε ένα HMM και ένα σύνολο παρατηρήσεων V^T. Προσδιορίστε την πιο πιθανή ακολουθία των κρυφών καταστάσεων ω_T που οδήγησε σε αυτές τις παρατηρήσεις.

Το πρόβλημα της Μάθησης (Training) Θεωρούμε ότι έχουμε μία αρχική υποτυπώδη δομή ενός μοντέλου, δηλαδή τον αριθμό των κρυφών καταστάσεων και τον αριθμό των ορατών συμβόλων εξόδου, αλλά όχι τις πιθανότητες a_{ij} και b_{jk}. Με δεδομένο ένα σύνολο παρατηρήσεων ορατών συμβόλων τα οποία χρησιμοποιούμε για εκπαίδευση, να προσδιοριστούν οι παραπάνω πιθανότητες, δηλαδή να εκπαιδευτεί το μοντέλο με βάση τα διαθέσιμα δεδομένα.

Για την επίλυση του προβλήματος Αξιολόγησης χρησιμοποιούνται οι αλγόριθμοι HMM Forward και HMM Backward, για την επίλυση του προβλήματος Αποκωδικοποίησης χρησιμοποιείται ο αλγόριθμος Viterbi και για την επίλυση του προβλήματος Μάθησης χρησιμοποιείται ο αλγόριθμος Expectation-Maximization (EM). Περισσότερες πληροφορίες για τους ορισμούς και την επίλυση των παραπάνω προβλημάτων υπάρχουν στο [12].

Τα GMM και HMM μοντέλα είναι μοντέλα που έχουν μελετηθεί εκτενώς και είναι ιδιαίτερα διαδεδομένα σε εφαρμογές επεξεργασίας audio stream και αναγνώρισης ομιλητών. Γενικά μπορούμε να πούμε ότι σε εφαρμογές με σημαντική εκ των προτέρων γνώση για την χρονική συμπεριφορά του audio stream χρησιμοποιούνται συχνά HMMs ενώ σε πιο γενικές εφαρμογές κατηγοριοποίησης audio stream και αναγνώρισης ομιλητών, όπου δεν υπάρχει πληροφορία για τη χρονική εξέλιξη των audio stream, χρησιμοποιούνται συνήθως GMMs. Τα πλεονεκτήματα των GMMs είναι ότι βασίζονται σε ένα στατιστικό μοντέλο που έχει μελετηθεί σε βάθος, είναι υπολογιστικά αποδοτικά, και δεν εξαρτώνται από τα χρονικά χαρακτηριστικά της ομιλίας αλλά μόνο από την κατανομή των ακουστικών χαρακτηριστικών.

Αν και το γενικό μοντέλο του GMM, όπως περιγράφηκε στα στοιχεία θεωρίας, υποστηρίζει πλήρεις πίνακες συμμεταβλητότητας, πολύ συχνά χρησιμοποιούνται μόνο διαγώνιοι πίνακες συμμεταβλητότητας. Αυτό συμβαίνει επειδή πρώτον η μοντελοποίηση πυκνότητας πιθανότητας με έναν πλήρη πίνακα συμμεταβλητότητας GMM τάξης M, μπορεί να γίνει εξίσου καλά και με χρήση ενός διαγώνιου πίνακα συμμεταβλητότητας μεγαλύτερης τάξης. Δεύτερον, τα GMMs με διαγώνιους πίνακες είναι υπολογιστικά πιο αποδοτικά καθώς δεν απαιτούνται στη φάση εκπαίδευσης επαναλαμβανόμενες αντιστροφές του $D \times D$ πλήρους πίνακα συμμεταβλητότητας και τρίτον έχει παρατηρηθεί πειραματικά ότι τα GMMs με διαγώνιους πίνακες αποδίδουν καλύτερα από τα GMMs με πλήρεις πίνακες.

Κεφάλαιο 3

State of the Art

Στην ενότητα αυτή περιγράφεται το State of the Art για το πρόβλημα Event Detection σε audio streams και γενικότερα για το πρόβλημα audio diarization.

3.1 Εξαγωγή Χαρακτηριστικών

Έχουν αναπτυχθεί διάφορα χαρακτηριστικά τα οποία περιέχουν ποικίλες μορφές πληροφορίας για το audio stream υπό εξέταση, όπως πληροφορίες για το φάσμα, το πλάτος του σήματος, το συχνοτικό του περιεχόμενο κλπ. Επίσης υπάρχουν εξειδικευμένα χαρακτηριστικά τα οποία μετρούν ιδιότητες που διαφέρουν από μουσική σε φωνή ή σε θόρυβο και χρησιμοποιούνται για την κατάταξη των audio streams σε διάφορες κατηγορίες.

Τα πιο ευρέως διαδεδομένα χαρακτηριστικά είναι οι συντελεστές Mel Frequency Cepstral Coefficients (MFCC). Οι συντελεστές αυτοί χρησιμοποιούνται είτε στην απλή τους μορφή είτε μαζί με τις πρώτες ή και δεύτερες παραγώγους τους. Έχουν αναπτυχθεί αρκετές παραλλαγές των συντελεστών αυτών όπως για παράδειγμα οι TECC συντελεστές(Teager Energy Cepstral Coefficients) που βασίζονται στη Teager ενέργεια και παρουσιάζουν αυξημένη ανθεκτικότητα σε θορυβώδη σήματα [11]. Χρησιμοποιούνται επίσης ως χαρακτηριστικά οι λογάριθμοι των ενεργειών του φιλτραρισμένου σήματος με φίλτρα σε κλίμακα mel [19], καθώς και τα PMVDR χαρακτηριστικά (Perceptual Minimum Variance Distortionless Response) [19] και [51]. Όπως εξηγείται στο [51] η μοντελοποίηση του spectrum του σήματος με PMVDR δημιουργεί ένα περισσότερο ομαλό spectral envelope.

Επίσης, χρησιμοποιούνται ευρέως χαρακτηριστικά που βασίζονται στην ανάλυση Linear Prediction(LP) και Perceprual Linear Prediction(PLP), όπως τα LP cepstrum, Line Spectrum Pair(LSP). Περισσότερες πληροφορίες για τις μεθόδους αυτές και τα χαρακτηριστικά

που βασίζονται σε αυτές υπάρχουν στα [6] και [30].

Υπάρχουν κάποια πολύ διαδεδομένα μονοδιάστατα χαρακτηριστικά όπως τα mean square amplitude ή root mean square amplitude (RMS), maximum amplitude (ή envelope), short time energy (STE) και zero-crossing rate (ZCR) του σήματος. Πολλές φορές, αντί για τα χαρακτηριστικά, αυτά χρησιμοποιούνται είτε μέσες τιμές, είτε αποκλίσεις τους, είτε πιο σύνθετες παραλλαγές τους πχ high zero crossing rate που ορίζεται ως το ποσοστό των πλαισίων των οποίων το ZCR υπερβαίνει το $1.5 \times average(ZCR)$ ή low short time energy ratio που είναι ανάλογο του πλήθους των πλαισίων για τα οποία η STE είναι μικρότερη από $0.5 \times average(STE)$ ([29]). Αρκετά μονοδιάστατα χαρακτηριστικά, κάποια από τα οποία είναι εξειδικευμένα στο διαχωρισμό φωνής από μουσική παρουσιάζονται στο [42]. Ενδεικτικά αναφέρουμε τα spectral flux, spectral centroid, spectral rolloff point και pulse metric. Επίσης στο [41] εισάγονται χαρακτηριστικά όπως previous local difference και next local difference. Άλλα χαρακτηριστικά που αναφέρονται στη βιβλιογραφία είναι τα mean per frame entropy, average probability dynamism, phone distribution match, background label energy ratio ([48]). Ακόμα, έχει γίνει χρήση νευρωνικών δικτύων για τον υπολογισμό χαρακτηριστικών entropy και dynamism για την διαφοροποίηση φωνής από μουσική, χαρακτηριστικά τα οποία εμφανίζουν υψηλά ποσοστά επιτυχίας [3].

Επίσης έχουν αναπτυχθεί χαρακτηριστικά που βασίζονται στη θεωρία για τον Energy Separation αλγόριθμο (ESA) [31], τα οποία χρησιμοποιούν τον αλγόριθμο ESA σε συνδυασμό με την Teager ενέργεια του σήματος. Τέτοια χαρακτηριστικά είναι τα maximum average Teager Energy (MTE), mean instantaneous amplitude (MIA) και mean instantaneous frequency (MIF), τα οποία προτείνονται ως εναλλακτικά των πιο συνηθισμένων mean square amplitude, mean absolute amplitude και average zero-crossings rate χαρακτηριστικών και παρουσιάζουν αυξημένη ανθεκτικότητα στο θόρυβο [13].

Η μελέτη των χαρακτηριστικών που περιγράφουν το συχνοτικό περιεχόμενο του σήματος έχει ιδιαίτερο ενδιαφέρον. Συνήθως οι ήχοι που παράγονται από μουσικά όργανα έχουν το χαρακτηριστικό ότι είναι αρμονικοί δηλαδή περιέχουν μία κύρια συχνότητα και αρκετά ακέραια πολλαπλάσιά της. Αντίθετα η φωνή είναι ένας ήχος που περιέχει μικτά αρμονικά, δηλαδή έμφωνα, τμήματα αλλά και μη αρμονικά, δηλαδή άφωνα, τμήματα. Έτσι είναι σημαντικά χαρακτηριστικά όπως η θεμελιώδης συχνότητα (pitch) αλλά και χαρακτηριστικά που σχετίζονται με τις αρμονικές ιδιότητες ενός σήματος. ([24] και [46]).

Τέλος έχουν μελετηθεί και μη γραμμικά χαρακτηριστικά όπως αυτά που βασίζονται στις fractal διαστάσεις του υπό εξέταση σήματος, η θεωρία των οποίων αναπτύσσεται στο [33] και [32].

3.2 Rule-Based Προσεγγίσεις

Με τον όρο αλλαγές αναφερόμαστε σε αξιοσημείωτα γεγονότα σε ένα audio stream, όπως μετάβαση από ομιλία σε μουσική ή θόρυβο και αντίστροφα ή εναλλαγή ομιλητών, ή μετάβαση σε διαφορετικό επίπεδο θορύβου, ή άλλες αλλαγές λιγότερο ή περισσότερο αισθητές ανάλογα με την ευαισθησία εντοπισμού που θέλουμε να πετύχουμε. Οι προσεγγίσεις για τον εντοπισμό αλλαγών (event detection) σε audio streams κατατάσσονται σε τρεις βασικές κατηγορίες: τις rule-based προσεγγίσεις που περιγράφονται σε αυτή την ενότητα, τις metric- based και τις model-based προσεγγίσεις που θα περιγραφούν αργότερα.

Στις rule-based προσεγγίσεις ο εντοπισμός αλλαγών και η κατηγοριοποίηση του audio stream γίνονται με βάση διάφορους κανόνες που εφαρμόζονται στο σύνολο των χαρακτηριστικών που εξάγονται από το σήμα. Ένα παράδειγμα τέτοιας προσέγγισης υπάρχει στο [29] για την κατηγοριοποίηση της μη ομιλίας σε θόρυβο, μουσική και σιωπή. Επίσης στο [41] χρησιμοποιείται ένα δέντρο απόφασης για την κατάταξη των ηχητικών πλαισίων.

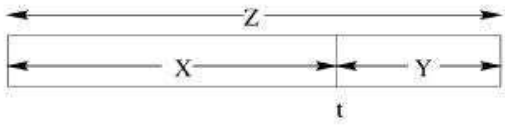
Το πιο ευρέως χρησιμοποιούμενο χαρακτηριστικό για τη χρήση σε τέτοιους κανόνες είναι η ενέργεια και οι προσεγγίσεις που χρησιμοποιούν την ενέργεια ονομάζονται energy-based προσεγγίσεις. Τέτοιες μέθοδοι βασίζονται στη μέτρηση της ενέργειας του σήματος σε διάφορα εύρη συχνοτήτων και στη χρήση κατωφλιών για τον εντοπισμό της σιωπής, και θεωρούν τα σημεία σιωπής ως σημεία αλλαγής ([47] και [25]).

Γενικά, οι rule-based μέθοδοι είναι σχετικά απλές υπολογιστικά μέθοδοι καθώς βασίζονται σε απλούς κανόνες ενός δέντρου απόφασης και σε κατώφλια, δεν χρειάζονται φάση εκπαίδευσης και μπορούν να λειτουργήσουν σε πραγματικό χρόνο. Εντούτοις, εξαιτίας τις χρήσης κατωφλιών υστερούν σε ανθεκτικότητα στο θόρυβο και τα κατώφλια πρέπει να αναπροσαρμόζονται ανάλογα με τα δεδομένα, ενώ δεν είναι πάντα δυνατό να βρεθεί μία βέλτιστη τιμή.

3.3 Metric-based Προσεγγίσεις

Οι metric-based προσεγγίσεις μετρούν ουσιαστικά τη διαφορά μεταξύ δύο διαδοχικών παραθύρων που μετατοπίζονται στο audio stream που εξετάζουμε, όπως φαίνεται στο σχήμα 3.1 και εντοπίζουν αλλαγή στο σημείο ανάμεσα στα δύο παράθυρα αν η διαφορά ξεπερνάει κάποιο κατώφλι (δηλαδή αν τα σήματα των δύο παραθύρων προέρχονται από διαφορετικές πηγές). Οι διαφορές των μεθόδων αυτών έγκεινται τόσο στα μέτρα διαφοράς που χρησιμοποιούνται όσο και στις αποφάσεις σχετικά με τα κατώφλια.

Μία ευρέως διαδεδομένη μέθοδος αυτής της κατηγορίας που δίνει πολύ καλά αποτελέ-



Εικόνα 3.1: Δύο γειτονικά παράθυρα με ακολουθίες διανυσμάτων χαρακτηριστικών X και Y, γύρω από τη χρονική στιγμή t, για την οποία θέλουμε να αποφασίσουμε αν αντιστοιχεί σε σημείο αλλαγής.

σματα είναι η μέθοδος BIC (Bayesian Information Criterion) που περιγράφεται στο [8]. Αυτή η μέθοδος ψάχνει για σημεία αλλαγής σε ένα παράθυρο ελέγχοντας αν τα δεδομένα του παραθύρου περιγράφονται καλύτερα από μία κατανομή (δεν υπάρχει αλλαγή) ή δύο κατανομές (υπάρχει αλλαγή). Αν βρεθεί αλλαγή τα παράθυρο αρχικοποιείται από το σημείο αλλαγής και μετά αλλιώς το παράθυρο αυξάνεται. Η αναζήτηση συνεχίζεται μέχρι να εξεταστεί ολόκληρο το audio-stream. Μειονέκτημα της μεθόδου είναι ότι χρειάζεται αρκετά δεδομένα (μεγάλο παράθυρο αναζήτησης) για το σωστό υπολογισμό του κριτηρίου κι έτσι έχει υψηλά ποσοστά αποτυχίας σε μικρά παράθυρα, συνεπώς δεν εντοπίζει γρήγορες εναλλαγές. Επίσης, ο υπολογισμός του BIC είναι υπολογιστικά ακριβός, της τάξης $O(n^2)$. Μία βελτίωση του παραπάνω αλγορίθμου παρουσιάζεται στο [19], όπου παρουσιάζονται δύο προσεγγίσεις του κριτηρίου BIC, οι οποίες λειτουργούν καλά για μικρά παράθυρα. Οι προσεγγίσεις αυτές είναι τα κριτήρια T^2 και WMD (Weighted Mean Distance).

Κάποια άλλα μέτρα που χρησιμοποιούνται στη βιβλιογραφία είναι η συμμετρική Kullback-Leibler απόσταση (K-L) καθώς και μία παραλλαγή της η divergence shape distance (DSD).([28], [20]).

Ένα άλλο συνηθισμένο κριτήριο είναι το LLR κριτήριο (LogLikelihood Ratio), που χρησιμοποιείται για παράδειγμα στο [23]. Στο [2] παρουσιάζεται μία νέα παραλλαγή του LLR, το κριτήριο LLRC το οποίο οδηγεί σε βελτιωμένη απόδοση, συγκρίσιμη με την απόδοση του BIC.

Επίσης, ένα κριτήριο που βασίζεται στην τεχνική Vector Quantization περιγράφεται στο [20]. Η προσέγγιση VQ βασίζεται στη γενικευμένη απόσταση μεταξύ δύο σειρών από διανύσματα χαρακτηριστικών, τις οποίες συμβολίζουμε με S^A και S^B. Το μέτρο διαφοροποίησης VQ (VQ distortion measure, VQD), μεταξύ του S^B και του κωδικού-βιβλίου (codebook) C^A, που δημιουργήθηκε μετά την ομαδοποίηση των χαρακτηριστικών στο S^A, ορίζεται ως:

$$VQD(C^A, S^B) = \frac{1}{T}\sum_{t=1}^{T} \arg\min_{1\leq k\leq K}\{d(C_k^A, S_t^B)\}$$

όπου το C_k^A αντιπροσωπεύει το k-οστό κωδικό-διάνυσμα στο C^A, το S_t^B αντιπροσωπεύει το t-οστό διάνυσμα χαρακτηριστικών στην ακολουθία S^B και d είναι η ευκλείδεια απόσταση.

Τέλος ένα νέο κριτήριο μέτρησης απόστασης μεταξύ δύο παραθύρων είναι το Weighted squared Euclidean Distance (WED) το οποία προτείνεται στο [25]. Το κριτήριο αυτό βασίζεται στην ευκλείδεια απόσταση μεταξύ των διανυσμάτων χαρακτηριστικών των δύο πλαισίων υπό σύγκριση, ενώ χρησιμοποιούνται επίσης και βάρη που εξαρτώνται από τις διακυμάνσεις των διανυσμάτων χαρακτηριστικών. Τέλος εφαρμόζεται σιγμοιδής συνάρτηση στα βάρη ώστε να αυξηθεί η ικανότητά τους να διακρίνουν διαφορετικά παράθυρα.

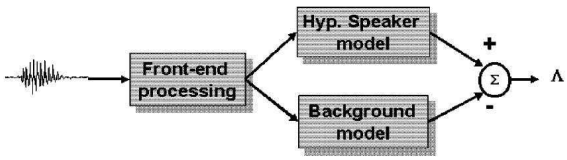
3.4 Model-Based Προσεγγίσεις (HMM/GMM μοντέλα)

Για να κάνουμε μία εισαγωγή στη θεωρία των model-based προσεγγίσεων θεωρούμε αρχικά το πρόβλημα της αναγνώρισης ενός ομιλητή μέσα σε ένα audio-stream. Ο εντοπισμός του ομιλητή μπορεί να θεωρηθεί ως ο έλεγχος δύο υποθέσεων, της H_0, δηλαδή ότι το τμήμα του audio stream προέρχεται από τον ομιλητή, και της H_1, δηλαδή ότι το τμήμα δεν προέρχεται από τον ομιλητή. Ο βέλτιστος τρόπος για να αποφασίσουμε μεταξύ των δύο υποθέσεων είναι ο έλεγχος των λόγων πιθανοφάνειας που παρουσιάζεται παρακάτω:

$$\frac{p(Y|H_0)}{p(Y|H_1)} \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{reject } H_0 \end{cases}$$

όπου $p(Y|H_i), i = 0, 1$ είναι η συνάρτηση πυκνότητας πιθανότητας (ή αλλιώς πιθανοφανεια) για την υπόθεση H_i, υπολογισμένη σε ένα τμήμα Y του audio stream. Σκοπός είναι να υπολογιστούν οι δύο πιθανοφάνειες, $p(Y|H_i), i = 0, 1$.

Στην εικόνα 3.2 φαίνεται ένα τυπικό σύστημα αναγνώρισης με χρήση μοντέλων.



Εικόνα 3.2: Model-Based σύστημα για την αναγνώριση ομιλητή([40]).

Αρχικά γίνεται η εξαγωγή των διανυσμάτων χαρακτηριστικών X του audio stream (front-end processing), τα οποία χρησιμοποιούνται στη συνέχεια για τον υπολογισμό της

πιθανοφάνειας. Μαθηματικά, η υπόθεση H_0 αναπαρίσταται από ένα μοντέλο, που συμβολίζεται με λ_{hyp} και χαρακτηρίζει τον ομιλητή που θέλουμε να αναγνωρίσουμε στο χώρο των χαρακτηριστικών που εξάγαμε. Η εναλλακτική υπόθεση αναπαρίσταται από το μοντέλο $\lambda_{\overline{hyp}}$. Συχνά, για τη σύγκριση πιθανοφανειών, αντί για το λόγο πιθανοφάνειας που παρουσιάστηκε παραπάνω, χρησιμοποιούνται λογάριθμοι, όπως στην παρακάτω συνάρτηση:

$$\Lambda(X) = \log\ p(X|\lambda_{hyp}) - \log\ p(X|\lambda_{\overline{hyp}}).$$

Το επόμενο σημαντικό βήμα είναι η επιλογή της συνάρτησης πυκνότητας πιθανότητας $p(X|\lambda)$, δηλαδή ο προσδιορισμός του μοντέλου που θα χρησιμοποιήσουμε. Τα GMM και HMM μοντέλα είναι μοντέλα που έχουν μελετηθεί εκτενώς και είναι ιδιαίτερα διαδεδομένα σε εφαρμογές επεξεργασίας audio streams και αναγνώρισης ομιλητών.

Για την εκπαίδευση των μοντέλων HMM ή GMM, με δεδομένο ένα σύνολο διανυσμάτων χαρακτηριστικών τα οποία χρησιμοποιούνται ως δεδομένα εκπαίδευσης, χρησιμοποιείται συνήθως ο αλγόριθμος Expectation-Maximization (EM) ([12]). Παρόλα αυτά έχουν προταθεί στη βιβλιογραφία και πιο αποδοτικοί αλγόριθμοι από τον EM για χρήση σε real-time συστήματα, των οποίων η απόδοση προσεγγίζει αυτη του αλγορίθμου EM ([28]).

Ένα πολύ απλό σύστημα που χρησιμοποιεί GMMs συνήθως περιέχει μόνο δύο μοντέλα, για ομιλία και μη ομιλία. Αν θέλουμε να κάνουμε πιο λεπτομερή κατηγοριοποίηση των τμημάτων των audio streams μπορούμε να προσθέσουμε μοντέλα για το φύλο του ομιλητή ή σχετικά με το εύρος ζώνης (bandwidth) του καναλιού ([16]). Επίση μπορούμε να έχουμε ξεχωριστά μοντέλα για μουσική, θόρυβο, ομιλία με μουσική και ομιλία με θόρυβο. Η μοντελοποίηση της ομιλίας υπό διάφορες συνθήκες αποτρέπει τη λανθασμένη κατηγοριοποίηση ηχογραφημένης ομιλίας ως θόρυβο ή μουσική όταν η ηχογράφηση έχει γίνει υπό συνθήκες θορύβου ή μουσικής αντίστοιχα ([18], [50] και [36]). Γενικά τα μοντέλα που εκπαιδεύουμε είναι τόσο λεπτομερή όσο λεπτομερής θέλουμε να είναι η κατηγοριοποίηση του audio stream.

Στη βιβλιογραφία αναφέρονται διάφορες τιμές σχετικά με το πλήθος των γκαουσιανών σε ένα GMM, και συνήθως στο στάδιο της εκπαίδευσης επιλέγεται αρχικά ένα μικρό πλήθος γκαουσιανών (πχ 2 ή 4) το οποίο αυξάνεται σταδιακά μέχρι την τελική τιμή του (πχ 64, 128 ή ακόμα περισσότερες). Όταν έχουμε ένα μη κατατμημένο audio stream, χρησιμοποιούμε για την αποκωδικοποίηση τον αλγόριθμο Viterbi, είτε σε ένα πέρασμα είτε σε διαδοχικά περάσματα με προαιρετική προσαρμογή (adaptation), ώστε να αναγνωρίσουμε τις διάφορες περιοχές. Αλλιώς αν το audio stream έχει ήδη κατατμηθεί σε ομογενή κομμάτια, το κάθε

κομμάτι μπορεί να κατηγοριοποιηθεί ξεχωριστά.

Ένα model-based σύστημα που χρησιμοποιεί GMMs με μεγάλη επιτυχία είναι το σύστημα της ομάδας LIMSI ([16]). Το σύστημα αυτό θα περιγραφεί με συντομία παρακάτω και ένα διάγραμμα της λειτουργίας του φαίνεται στην εικόνα 3.3.

Αρχικά εξάγονται από το σήμα τα τμήματα ομιλίας χρησιμοποιώντας Viterbi αποκωδικοποίηση και GMMs για ομιλία, ομιλία με θόρυβο, ομιλία με μουσική, μουσική, θόρυβο ή σιωπή.(Speech Activity Detection-SAD). Στη συνέχεια, μέσα στα τμήματα ομιλίας, βρίσκονται τα σημεία αλλαγής ομιλητών χρησιμοποιώντας κάποιο metric-based κριτήριο (Choping in small segments). Για κάθε ένα από τα τμήματα εκπαιδεύεται ένα GMM μοντέλο.

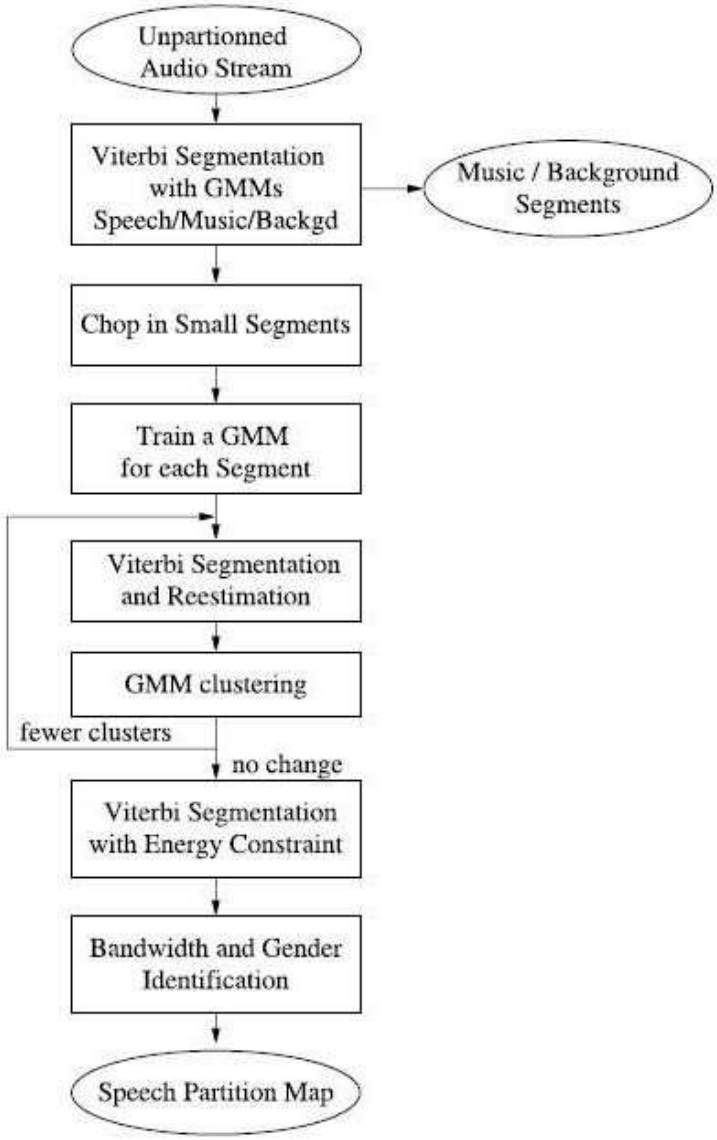
Ακολουθεί μια διαδοχική διαδικασία GMM Κατάτμησης και ομαδοποίησης(Iterative GMM Segmentation \ Clustering Procedure) η οποία έχει ως σκοπό τη δημιουργία ομογενών κλάσεων εναλλάσοντας διαδικασίες Viterbi αποκωδικοποίησης, επανυπολογισμού GMM και συγχώνευσης κοντινών κλάσεων. Στόχος είναι η εύρεση ενός διαχωρισμού των τμημάτων ομιλίας ώστε να μεγιστοποιηθεί η συνάρτηση:

$$\sum_{i=1}^{N} log\ f(s_i|M_{c_i}) - aN - \beta K$$

όπου $S = (s_1, \ldots, s_N)$ είναι ο διαχωρισμός των τμημάτων ομιλίας σε μία ακολουθία N τμημάτων, c_i είναι η ετικέτα κλάσης για το κάθε τμήμα s_i (ανάμεσα στις K διαφορετικές κλάσεις), $f(s_i|M_{c_i})$ είναι η πιθανότητα εμφάνισης του τμήματος s_i με δεδομένο το μοντέλο M_{c_i}, και α και β είναι οι ποινές για τα τμήματα και τις κλάσεις αντίστοιχα.

Στη συνέχεια, ακολουθεί μια νέα Viterbi αποκωδικοποίηση για τον λεπτομερή καθορισμό των ορίων του κάθε τμήματος με χρήση περιορισμού βασισμένου στην ενέργεια (Viterbi Segmentation with Energy Constraint). Τέλος εκπαιδεύονται κατάλληλα GMM για κατηγοριοποίηση των τμημάτων ανάλογα με το φύλο του ομιλητή και το bandwith (Bandwidth and Gender Identification).

Αν θέλουμε να αναγνωρίσουμε ένα συγκεκριμένο τύπο audio-stream, για παράδειγμα έναν ομιλητή έναντι όλων των άλλων πιθανών περιπτώσεων ήχου (όπως στην περίπτωση που παρουσιάστηκε αρχικά), μπορούμε αντί να συγκρίνουμε τα δεδομένα με όλα τα πιθανά μοντέλα να εκπαιδεύσουμε ένα γενικό μοντέλο για την περίπτωση το τμήμα που εξετάζουμε να μην ανήκει στην επιθυμητή υπόθεση (περίπτωση $\lambda_{\overline{hyp}}$). Η προσέγγιση αυτή προτείνεται στο [40], όπου χρησιμοποιούνται τα γενικά μοντέλα UBM (Universal Background Model). Στο σύστημα GMM-UBM που παρουσιάζεται, αντί να χρησιμοποιηθεί ο αλγόριθμος EM για την εκπαίδευση όλων των μοντέλων, εκπαιδεύεται αρχικά ένα UBM και στη συνέχεια



Εικόνα 3.3: Το σύστημα LIMSI ([16]).

παράγονται τα μοντέλα για τον κάθε ομιλητή προσαρμόζοντας το UBM στα δεδομένα του συγκεκριμένου ομιλητή (adaptation).

Αν και συνήθως σε συστήματα εύρεσης εναλλαγών ομιλητών οι ομιλητές κατηγοριοποιούνται ως "ομιλητής1", "ομιλητής2", κλπ είναι δυνατό να έχουμε αναγνώριση της ταυτότητας του ομιλητή εκπαιδεύοντας για παράδειγμα ξεχωριστά μοντέλα για ομιλητές που είναι πιθανό να μιλούν σε δελτία ειδήσεων(πχ κεντρικούς εκφωνητές ή πολιτικούς) και συμπεριλαμβάνοντας τα μοντέλα αυτά στο σύστημα. Άλλα συστήματα που χρησιμοποιούν γλωσσολογική πληροφορία περιγράφονται στα ([7], [43])

Μία πρόσφατη προσέγγιση η οποία εφαρμόζει ταυτόχρονα τις διαδικασίες κατάτμησης και ομαδοποίησης βασίζεται στη θεωρία για τα evolutive-HMM (E-HMM) μοντέλα που αναπτύχθηκε στο ([34]). Σε αυτή την τεχνική, ο εντοπισμός ενός ομιλητή επηρρεάζει τόσο τον εντοπισμό των άλλων ομιλητών όσο και τα όρια των τμημάτων. Το συνολικό audio stream αναπαρίσταται από ένα εργοδικό HMM στο οποίο κάθε κατάσταση αναπαριστά έναν ομιλητή και οι μεταβάσεις μοντελοποιούν τις αλλαγές ομιλητών. Το αρχικό HMM έχει μία μόνο κατάσταση. Σε κάθε επανάληψη, ένα μικρό τμήμα, το οποίο θεωρούμε ότι προέρχεται από έναν μη εντοπισμένο ομιλητή, επιλέγεται και χρησιμοποιείται για να δημιουργηθεί ένα μοντέλο του ομιλητή αυτού μεσω Bayesian adaptation του UBM. Μία νέα κατάσταση προστίθεται τότε στο HMM ώστε να αναπαραστήσει τον ομιλητή και οι πιθανότητες μετάβασης μεταβάλλονται κατάλληλα. Μία νέα κατάτμηση δημιουργείται από την Viterbi αποκωδικοποιήση του audio stream με βάση το νέο HMM και κάθε μοντέλο προσαρμόζεται χρησιμοποιώντας την καινούρια κατάτμηση. Η φάση κατάτμησης συνεχίζεται μέχρι να μην μεταβάλλονται πλέον οι ετικέτες του κάθε τμήματος.Η διαδικασία προσθήκης νέων ομιλητών συνεχίζεται μέχρι να τελειώσουν τα δεδομένα, ή να σταματήσει να υπάρχει κέρδος σχετικά με την πιθανοφάνεια από την προσθήκη.

3.5 Υβριδικές Προσεγγίσεις

Στην ενότητα αυτή παρουσιάζονται κάποια συστήματα που χρησιμοποιούν συνδυασμό των προσεγγίσεων που παρουσιάστηκαν παραπάνω ή και επιπλέον ιδέες, όπως για παράδειγμα χρήση νευρωνικών δικτύων.

Στο [4] παρουσιάζεται ένα σύστημα που βασίζεται στο LIMSI ([16]) αλλά περιέχει κάποιες αλλαγές που οδηγούν σε βελτίωση της απόδοσης. Η κυριότερη αλλαγή είναι η αντικατάσταση του σταδίου Iterative GMM segmentation \ Clustering procedure με ένα στάδιο

BIC clustering, δηλαδή ομαδοποίηση των κλάσεων με χρήση του κριτηρίου BIC. Πιο συγκεκριμένα στο σύστημα αυτό έχουμε αρχικά, όπως και στο LIMSI, εξαγωγή των τμημάτων φωνής από το συνολικό σήμα, εντοπισμό των αλλαγών ομιλητή και αντίστοιχη κατάτμηση (με χρήση κατάλληλων GMMs). Ακολουθεί BIC clustering όπου κάθε κλάση με αρχικοποιείται με ένα τμήμα και τα τμήματα ομαδοποιούνται με βάση το ΔBIC κριτήριο:

$$\Delta BIC = (n_i + n_j)log|\Sigma| - n_i log|\Sigma_i| - n_j log|\Sigma_j| - \lambda P$$

όπου Σ είναι ο πίνακας συμμεταβλητότητας του τμήματος που προέκυψε από τη συγχώνευση των δύο κλάσεων, Σ_i και Σ_j είναι οι πίνακες συμμεταβλητότητας των κλάσεων c_i και c_j αντίστοιχα και n_i και n_j είναι ο αριθμός των πλαισίων στις κλάσεις c_i και c_j αντίστοιχα. Η ποινή P είναι

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1))logn$$

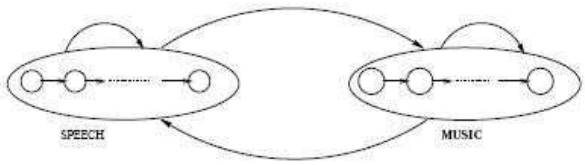
όπου d είναι η διάσταση του χώρου του διανύσματος χαρακτηριστικών.

Δύο κλάσεις συγχωνεύονται όταν $\Delta BIC < 0$ και σε κάθε επανάληψη επιλέγονται για συγχώνευση οι κλάσεις με το περισσότερο αρνητικό ΔBIC. Η διαδικασία τερματίζεται όταν το ΔBIC μεταξύ όλων των κλάσεων είναι μεγαλύτερο ή ίσο του μηδενός.

Στη συνέχεια ακολουθεί ένα δεύτερο προαιρετικό στάδιο clustering για ομαδοποίηση σε πιο λεπτομερείς κλάσεις (πχ ομιλία με μουσική, ομιλία με θόρυβο) και ένα τελευταίο στάδιο με σκοπό να αφαιρεθούν σύντομα τμήματα σιωπής που πιθανόν να έχουν απομείνει.

Επιπλέον, ένα σύστημα που συνδυάζει HMMs με νευρωνικά δίκτυα και δίνει πολύ καλά αποτελέσματα παρουσιάζεται στο [3]. Συγκεκριμένα χρησιμοποιείται ένα Multilayer Perceptron (MLP) για τον υπολογισμό των εκ των υστέρων πιθανοτήτων (posterior probabilities) των φωνημάτων ομιλίας με δεδομένα διανύσματα χαρακτηριστικών που αντιστοιχούν σε ένα παράθυρο συγκεκριμένης διάρκειας. Οι πιθανότητες που εξάγονται από το MLP χρησιμοποιούνται για τον υπολογισμό δύο χαρακτηριστικών, Entropy και Dynamism. Τα χαρακτηριστικά αυτά δίνονται ως είσοδο σε ένα HMM σύστημα που κάνει κατάταξη σε μουσική και ομιλία. Συγκεκριμένα, χρησιμοποιείται ένα πλήρως συνδεδεμένο HMM με δύο καταστάσεις (μουσική και ομιλία), και κάθε μία από τις κατάστάσεις αυτές περιέχει διαδοχικές υποκαταστάσεις που είναι παρόμοιες και μοντελοποιούνται με μίγματα γκαουσιανών κατανομών. Η τοπολογία φαίνεται στην εικόνα 3.4.

Επίσης, μία εναλλακτική προσέγγιση που συνδυάζει διάφορες ιδέες προτείνεται στο [29]. Αρχικά γίνεται κατηγοριοποίηση του σήματος σε ομιλία και μη ομιλία χρησιμοποιώντας τεχνικές που βασίζονται σε κλασσικές ιδέες επεξεργασίας προτύπων, όπως ο KNN



Εικόνα 3.4: Η τοπολογία του HMM classifier του συστήματος [3].

αλγόριθμος (K- nearest neighbour algorithm) και linear spectral pairs vector quantization (LSP-VQ). Στο δεύτερο στάδιο γίνεται κατηγοριοποίηση των τμημάτων μη ομιλίας σε μουσική, θόρυβο και σιωπή, χρησιμοποιώντας rule-based προσέγγιση βασισμένη σε ένα set καινούριων, εξειδικευμένων χαρακτηριστικών. Τέλος, η επεξεργασία των τμημάτων ομιλίας και η δημιουργία κλάσεων ομιλητών γίνεται με χρήση GMM μοντέλων. Το σύστημα έχει καλή απόδοση και ταχύτητα που επιτρέπει τη χρήση του σε real time εφαρμογές.

Τέλος, αναφέρουμε το σύστημα για ακολουθιακή ομαδοποίηση ομιλητών (sequential speaker clustering) που προτάθηκε στο ([27]). Το σύστημα αυτό είναι μία κομψή λύση στην ομαδοποίηση ομιλητών, όταν η δημιουργία των κλάσεων χρειάζεται να γίνεται online, χωρίς να είναι εκ των προτέρων γνωστά όλα τα δεδομένα. Το σύστημα αυτό εξετάζει σειριακά τα τμήματα ήχου και αποφασίζει αν ταιριάζουν σε κάποια από τις διαθέσιμες κλάσεις ομιλητών με βάση κατάλληλα κριτήρια απόστασης λόγων πιθανοφάνειας. Αν βρεθεί ταίριασμα, ανανεώνονται τα δεδομένα της αντίστοιχης κλάσης, αλλιώς το τμήμα αρχικοποιεί μία νέα κλάση ομιλητή. Η τεχνική αυτή είναι πολύ γρήγορη και μπορεί να χρησιμοποιηθεί σε real-time συστήματα.

Κεφάλαιο 4

Κατάτμηση Ηχητικών Τμημάτων με Χρήση Μετρικών Κριτηρίων

4.1 Σκοπός

Στο κεφάλαιο αυτό μελετάται το πρόβλημα της εύρεσης αλλαγών σε ηχητικά τμήματα (audio - streams) και της κατάτμησής τους με βάση τα εντοπισμένα σημεία αλλαγής.

Ο όρος αλλαγή είναι αρκετά γενικός και εξαρτάται από την εκάστοτε εφαρμογή και από τη λεπτομέρεια κατάτμησης που επιθυμούμε να επιτύχουμε. Για τις περιπτώσεις audio-streams σημεία αλλαγής μπορούν να οριστούν τα σημεία όπου υπάρχει μετάβαση από ομιλία σε μη ομιλία ή αντίστροφα. Λέγοντας μη ομιλία θα μπορούσαμε να αναφερόμαστε σε σιωπή, μουσική, διάφορες μορφές θορύβου ανάλογα με την εφαρμογή που εξετάζουμε. Σε περίπτωση που επιθυμούμε μία πιο λεπτομερή κατάτμηση θα μπορούσαμε να ορίσουμε ως σημεία αλλαγής και τα σημεία αλλαγής ομιλητών ή τα σημεία όπου παρατηρείται αλλαγή στις συνθήκες περιβάλλοντος της ηχογράφησης, για παράδειγμα σημεία έναρξης κάποιου θορύβου στο background κατά τη διάρκεια της ομιλίας ενός ομιλητή. Παρατηρούμε λοιπόν ότι ο ορισμός των αλλαγών είναι σε αρκετές περιπτώσεις υποκειμενικός και εξαρτώμενος της εφαρμογής και συνακόλουθα η ακρίβεια των αλγορίθμων κατάτμησης θα μπορούσε να προσαρμοστεί στην επιθυμητή ακρίβεια εντοπισμού αλλαγών.

Σχετικά με τη διαδικασία εύρεσης αλλαγών, υπάρχουν στη βιβλιογραφία αρκετές διαφορετικές προσεγγίσεις του προβλήματος, πολλές από τις οποίες αναφέρθηκαν συνοπτικά στην ενότητα State of the Art. Στην παρούσα ενότητα θα παρουσιαστούν πιο αναλυτικά κάποιες από αυτές και τα αποτελέσματα που δίνουν με εφαρμογή σε πραγματικά δελτία ειδήσεων. Πιο συγκεκριμένα, θα μελετηθούν συγκριτικά διάφορα χαρακτηριστικά που μπορούν να εξα-

χθούν από το ηχητικό σήμα ώστε να χρησιμοποιηθούν στη διαδικασία εντοπισμού αλλαγών. Στη συνέχεια, θα μελετηθεί πληθώρα κριτηρίων για την εύρεση αλλαγών. Θα επικεντρωθούμε σε μετρικά κριτήρια (Metric-based) καθώς οι μέθοδοι στατιστικής μοντελοποίησης του σήματος θα εξεταστούν στο επόμενο κεφάλαιο. Τελικά, θα παρουσιαστούν αλγόριθμοι που εξάγουν χαρακτηριστικά από το σήμα και χρησιμοποιούν metric-based κριτήρια για να βρουν αλλαγές.

Εξετάζεται πληθώρα μονοδιάστατων και πολυδιάστατων χαρακτηριστικών που έχουν προταθεί στη βιβλιογραφία. Επίσης, εξετάζεται το πρόβλημα του αποτελεσματικού συνδυασμού των χαρακτηριστικών αυτών ώστε να επιτύχουμε μεγαλύτερα ποσοστά επιτυχίας εντοπισμού αλλαγών, από ότι με τη χρήση του κάθε χαρακτηριστικού ξεχωριστά. Η ανάγκη για αποτελεσματική συνδυαστική χρήση πολλών ανομοιόμορφων χαρακτηριστικών του σήματος οδηγεί σε αλγορίθμους πολλών περασμάτων που θα παρουσιαστούν αναλυτικά στην ενότητα της μελέτης αλγορίθμων. Ένας τέτοιος αλγόριθμος είναι ο αλγόριθμος δεύτερου περάσματος που αποτελεί μία καινούρια ιδέα που αναπτύχθηκε κατά τη διάρκεια αυτής της εργασίας ώστε να συνδυαστούν πολυδιάστατα χαρακτηριστικά όπως οι συντελεστές Mel Frequency Cepstral Coefficients (MFCC) με μονοδιάστατα χαρακτηριστικά του σήματος όπως η fractal διάσταση (fractal dimension) και το μέγιστο πλάτος του σήματος (max amplitude).

Τα πειράματα έχουν γίνει σε τμήματα από πραγματικά δελτία ειδήσεων. Ενδιαφερόμαστε για την εύρεση αλλαγών μεταξύ ομιλίας και μη ομιλίας αλλά και αλλαγών μεταξύ ομιλητή, ίδιου ή διαφορετικού φύλου. Η περίπτωση της μη ομιλίας για δελτία ειδήσεων περιλαμβάνει τη σιωπή και το θόρυβο, όπου στο θόρυβο κατατάσσουμε και τη μουσική. Τα τμήματα δελτίων που χρησιμοποιούνται στα πειράματα έχουν επιλεχθεί χειροκίνητα ώστε να καλύπτουν όλο το δυνατό εύρος περιπτώσεων αλλαγών και οι αλλαγές σε αυτά έχουν σημειωθεί επίσης χειροκίνητα. Επίσης τα τμήματα αυτά έχουν χωριστεί σε κατηγορίες ανάλογα με τη δυσκολία εύρεσης των σημείων αλλαγής που περιέχουν. Λεπτομέρειες σχετικές με την κατηγοριοποίηση των ηχητικών τμημάτων αλλά και με τις συνθήκες διεξαγωγής των πειραμάτων υπάρχουν στη ενότητα των πειραματικών αποτελεσμάτων. Στην ίδια ενότητα παρουσιάζονται αναλυτικά αποτελέσματα των πειραμάτων που έγιναν και γίνεται σχολιασμός τους.

Συμπερασματικά, σκοπός του κεφαλαίου αυτού είναι τα εξετάσει με ικανοποιητική πληρότητα την περιοχή της κατάτμησης audio-streams σε σχέση με τα χρησιμοποιούμενα χαρακτηριστικά του σήματος αλλά και σε σχέση με τους metric-based αλγορίθμους εντοπισμού αλλαγών που υπάρχουν στη βιβλιογραφία. Επίσης προτείνεται ένας νέος αλγόριθμος δεύτερου περάσματος ο οποίος φαίνεται ότι έχει δυνατότητες να βελτιώσει τα αποτελέσματα

των ευρέως χρησιμοποιούμενων αλγορίθμων αυτής της περιοχής.

4.2 Εξαγωγή Χαρακτηριστικών του Ηχητικού Σήματος

Στην ενότητα αυτή παρουσιάζονται διάφορα χαρακτηριστικά που υπάρχουν στη βιβλιογραφία. Τέτοια χαρακτηριστικά εξάγονται από το σήμα με σκοπό να χρησιμοποιηθούν από αλγορίθμους εντοπισμού αλλαγών. Τυπικά, σε μεθόδους εξαγωγής χαρακτηριστικών χωρίζουμε το σήμα σε διαδοχικά παράθυρα με συγκεκριμένο μήκος και συγκεκριμένη επικάλυψη μεταξύ διαδοχικών παραθύρων. Από κάθε τέτοιο παράθυρο εξάγουμε ένα ή περισσότερα χαρακτηριστικά τα οποία θεωρούμε ότι είναι ικανά να μοντελοποιήσουν το σήμα και να χρησιμοποιηθούν αντί αυτού στη διαδικασία εντοπισμού αλλαγών. Ονομάζουμε πολυδιάστατα τα χαρακτηριστικά που απαιτούν την εξαγωγή περισσότερων του ενός συντελεστών για κάθε πλαίσια (frame) σήματος ενώ ονομάζουμε μονοδιάστατα τα χαρακτηριστικά που αποτελούνται από έναν συντελεστή ανά frame σήματος. Σημειώνουμε ότι πριν την εξαγωγή των χαρακτηριστικών γίνεται παραθύρωση των frames του σήματος με κάποιο κατάλληλο φίλτρο, συνήθως με παράθυρο Hamming.

Τα πολυδιάστατα χαρακτηριστικά που υλοποιήθηκαν και μελετήθηκαν παρουσιάζονται παρακάτω.

4.2.1 Mel Frequency Cepstral Coefficients - MFCC

Οι συντελεστές Mel Frequency Cepstral Coefficients είναι συντελεστές που χρησιμοποιούνται εκτενώς σε προβλήματα εντοπισμού αλλαγών αλλά και αναγνώρισης ηχητικών σημάτων. Η συμπεριφορά και η απόδοση των συντελεστών αυτών έχει μελετηθεί στη βιβλιογραφία και έχει αποδειχθεί ότι δίνουν πολύ καλά αποτελέσματα.

Οι συντελεστές MFCC εξάγονται από μία αναπαράσταση του cepstrum του σήματος, με τη διαφορά ότι η συστοιχία των φίλτρων που χρησιμοποιείται για το φιλτράρισμα του αρχικού frame είναι κατανεμημένη στην κλίμακα Mel. Ο τύπος μετατροπής από τη γραμμική κλίμακα στη κλίμακα Mel φαίνεται παρακάτω:

$$f_{mel} = 2595 \log(1 + \frac{f_{lin}}{700})$$

όπου f_{mel} είναι οι συχνότητες στην κλίμακα mel και f_{lin} είναι οι συχνότητες στην αρχική γραμμική κλίμακα.

Η χρήση της κλίμακας mel προτιμάται καθώς η κατανομή των φίλτρων στην κλίμακα αυτή μοντελοποιεί καλύτερα την απόκριση του ανθρώπινου συστήματος παραγωγής φωνής από ότι η χρήση γραμμικά κατανεμημένων φίλτρων.

Η μέθοδος παραγωγής των MFCC χαρακτηριστικών είναι συνοπτικά η εξής:

1. Παίρνουμε τον μετασχηματισμό Fourier ενός παραθυροποιημένου Frame του σήματος
2. Θεωρούμε συστοιχία φίλτρων ισοκατανεμημένων στην κλίμακα mel. Τα φίλτρα που χρησιμοποιούνται είναι συνήθως τριγωνικά.
3. Αναλύουμε το Frame με βάση τη συστοιχία των φίλτρων και υπολογίζουμε την ενέργεια της απόκρισης για κάθε ένα από τα φίλτρα με είσοδο το Frame υπό εξέταση.
4. Παίρνουμε το λογάριθμο των ενεργειών
5. Παίρνουμε το διακριτό μετασχηματισμό συνημιτόνου για κάθε έναν από τους συντελεστές και κρατάμε τους N πρώτους συντελεστές. Εξάγουμε τελικά N συντελεστές για κάθε Frame.

Σχετικά με τη υλοποίηση και τα πειράματα, τα χαρακτηριστικά που χρησιμοποιούνται για την κατάτμηση είναι 13 MFCC συντελεστές. Το φίλτρα που χρησιμοποιούνται για την εξαγωγή των συντελεστών είναι 35 τρίγωνα κεντραρισμένα σε κάθε μία από τις 35 συχνότητες που βρίσκονται ισοκατανεμημένες στην κλίμακα mel.

Παρατηρείται πειραματικά ότι η χρήση επικαλυπτόμενων πλαισίων για την εξαγωγή των MFCC οδηγεί σε υπερκατάτμηση και σε εύρεση περισσότερων του ενός κοντινών σημείων αλλαγής που αντιστοιχούν στην ίδια πραγματική αλλαγή. Κατά συνέπεια επιλέγουμε τα πλαίσια να μην είναι είναι επικαλυπτόμενα. Το μήκος του frame επιλέγεται ίσο με 40msec.

Επίσης, για να συνδυάσουμε τα πλεονεκτήματα της χρήσης επικαλυπτόμενων frames με την εύρεση πιο γενικών και ομαλών χαρακτηριστικών χρησιμοποιούμε μία τεχνική εύρεσης μέσων όρων. Συγκεκριμένα, χωρίζουμε το κάθε μη επικαλυπτόμενο frame των 40msec σε 3 επικαλυπτόμενα frames των 20msec με 10 msec επικάλυψη. Βρίσκουμε τους MFCC συντελεστές σε αυτά τα επικαλυπτόμενα frames και παίρνουμε μέσο όρο για να βρούμε τους τελικούς MFCC συντελεστές που αντιστοιχούν στο αρχικό frame των 40msec. Από τα πειραματικά αποτελέσματα φαίνεται ότι αυτή η τεχνική δίνει καλύτερη απόδοση από τη χρήση μη επικαλυπτόμενων frames των 40msec, στην περίπτωση εντοπισμού αλλαγών μεταξύ ομιλίας και μη ομιλίας. Η τεχνική αυτή θα αναφέρεται στο εξής ως τεχνική averaging.

4.2.2 Teager Energy Cepstral Coefficients - TECC

Οι συντελεστές Teager Energy Cepstrum Coefficients (TECC) αποτελούν μία παραλλαγή των MFCC χαρακτηριστικών, η εξαγωγή των οποίων βασίζεται στον υπολογισμό της

ενέργειας Teager. Παρακάτω περιγράφεται συνοπτικά το θεωρητικό υπόβαθρο για τον υπολογισμό της ενέργειας Teager καθώς και η διαδικασία εξαγωγής των χαρακτηριστικών TECC. Περισσότερες λεπτομέρειες περιέχονται στο [11] ενώ το θεωρητικό υπόβαθρο περιέχεται στο [31].

Σύμφωνα με το νόμο του Νεύτωνα για την κίνηση ενός ταλαντωτή με μάζα m και σταθερά ελατηρίου k, ισχύει:

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0$$

Η λύση της παραπάνω εξίσωσης είναι:

$$x(t) = a\cos(\varphi(t))$$

Η συνολική ενέργεια του συστήματος είναι το άθροισμα της κινητικής και δυναμικής του ενέργειας και δίνεται από τον παρακάτω τύπο:

$$E = \frac{1}{2}kx^2 + \frac{1}{2}m\dot{x}^2 \quad \Rightarrow \quad E = \frac{1}{2}m\omega^2 a^2$$

όπου ω = dφ(t)/dt.

Με βάση τα παραπάνω οι Teager και Kaiser πρότειναν τον τελεστή Teager-Kaiser Ψ :

$$\Psi[x(t)] = \dot{x}^2(t) - x(t)\ddot{x}(t)$$

Όταν ο τελεστής Ψ εφαρμοστεί στο AM - FM σήμα $x(t) = a(t)\cos(\varphi(t))$, δίνει:

$$\Psi[x(t)] \cong a^2(t)\dot{\varphi}^2(t)$$

Συνεπώς αντί για τη χρήση της κλασσικής προσέγγισης x^2, της ενέργειας ενός σήματος, μπορεί να χρησιμοποιηθεί εναλλακτικά η Teager ενέργεια. Έτσι, η εξαγωγή των TECC συντελεστών γίνεται με τον ίδιο τρόπο με την εξαγωγή των MFCC συντελεστών, μόνο που για τον υπολογισμό της ενέργειας του κάθε φίλτρου χρησιμοποιείται η Teager ενέργεια. Ο συντελεστής Teager Ψ για ένα διακριτό σήμα ορίζεται όπως φαίνεται παρακάτω, όπου η παράγωγος του σήματος προσεγγίζεται από 1-sample differences:

$$\Psi[x(n)] = x^2(n) - x(n-1)x(n+1)$$

Η χρήση Teager ενέργειας απαιτεί και την τροποποίηση του αλγορίθμου για την εξαγωγή των χαρακτηριστικών. Καταρχήν δεν ισχύει πλέον για την Teager ενέργεια το θεώρημα του Parceval, το οποίο αναφέρει ότι:

$$E = \int_{-\infty}^{+\infty} x^2(t)dt = \int_{-\infty}^{+\infty} |X(\omega)|^2 d\omega$$

Άρα δεν ισχύει ότι η μέση ενέργεια του σήματος μπορεί να υπολογιστεί είτε στο πεδίο του χρόνου είτε στο πεδίο της συχνότητας. Η μέση Teager ενέργεια πρέπει να υπολογιστεί στο πεδίο του χρόνου. Αυτό σημαίνει ότι αν μεταφέρουμε το σήμα μας στο πεδίο της συχνότητας για να το φιλτράρουμε αποφεύγοντας έτσι την υπολογιστικά απαιτητική συνέλιξη που χρειάζεται στο πεδίο του χρόνου, στη συνέχεια θα πρέπει να μεταφέρουμε το αποτέλεσμα του φιλτραρίσματος πίσω στο πεδίο του χρόνου με ανάστροφο FFT. Κάτι τέτοιο δεν χρειάζεται να γίνει κατά τον υπολογισμό των MFCC, όπου η 'κλασσική' ενέργεια υπολογίζεται απευθείας στο πεδίο της συχνότητας, γι' αυτό και ο υπολογισμός των MFCC γίνεται πολύ γρήγορα.

Ανακεφαλαιώνοντας λοιπόν, αν επιλέξουμε να μην μεταφερθούμε στο χώρο της συχνότητας για τον υπολογισμό του φιλτραρισμένου σήματος, έχουμε λόγω της συνέλιξης επιπλέον πολυπλοκότητα σε σχέση με τα MFCC, $O(Q \cdot n^2)$, όπου Q είναι ο αριθμός των φίλτρων. Αν αντίθετα επιλέξουμε να μεταφερθούμε στο χώρο της συχνότητας για τον υπολογισμό του φιλτραρισμένου σήματος, έχουμε λόγω του ανάστροφου FFT που πρέπει να γίνει στη συνέχεια επιπλέον πολυπλοκότητα, $O(Q \cdot n \cdot \log n)$. Η υλοποίηση είναι καλύτερο να γίνει ακολουθώντας τον δεύτερο τρόπο που είναι πιο γρήγορος, και πάλι όμως ο υπολογισμός των TECC αργεί αισθητά και ο χρόνος υπολογισμού κυριαρχεί στο συνολικό χρόνο επεξεργασίας του audio-stream.

Ένας επιπλέον παράγοντας αργοπορίας είναι η χρήση gabor φίλτρων, ο υπολογισμός των οποίων είναι πιο χρονοβόρος από τον υπολογισμό των τριγωνικών φίλτρων.

Στην υλοποίηση που έγινε εξάγονται 13 TECC συντελεστές. Το φίλτρα που χρησιμοποιούνται για την εξαγωγή των συντελεστών είναι 35 gabor φίλτρα κεντραρισμένα σε κάθε μία από τις 35 συχνότητες που βρίσκονται ισοκατανεμημένες στην κλίμακα mel. Επίσης, για να βελτιώσουμε την απόδοση, υπολογίζουμε τους συντελεστές σε μη επικαλυπτόμενα frames των 40msec και χρησιμοποιούμε την τεχνική averaging, που περιγράφηκε στην ενότητα των MFCC συντελεστών.

4.2.3 Perceptual Minimum Variance Distortionless Response - PMVDR

Επιπρόσθετα από τους συντελεστές LFE χρησιμοποιούμε τα χαρακτηριστικά Perceptual Minimum Variance Distortionless Response (PMVDR)([51],[35],[44]).

Όπως εξηγείται στο [35], το MVDR spectrum χρησιμοποιείται επειδή μοντελοποιεί καλύτερα έμφωνους ήχους φωνής, στους οποίους τα spectra που βασίζονται σε LPC μοντέλα

τείνουν να υπερεκτιμούν τα μέσα και υψηλά pitch της φωνής και περιέχουν ανεπιθύμητες, απότομες κορυφές. Η μοντελοποίηση με MVDR δημιουργεί ένα περισσότερο ομαλό spectral envelope.

Θα εξηγηθούν με συντομία κάποια βασικά σημεία στον υπολογισμό του MVDR spectrum που αναφέρονται στο [51]. Στη μέθοδο εκτίμησης του MVDR spectrum, η ενέργεια του σήματος σε μία συχνότητα ω_l καθορίζεται φιλτράροντας το σήμα με ένα ειδικά σχεδιασμένο FIR φίλτρο h(n) και μετρώντας την ενέργεια της εξόδου. Το φίλτρο h(n) είναι σχεδιασμένο ώστε να ελαχιστοποιεί την ενέργεια εξόδου με τον περιορισμό ότι η απόκριση στη συχνότητα ω_l έχει κέρδος μονάδα. Το MVDR spectrum τάξης Q γράφεται ως εξής:

$$P_{MV}(\omega) = \frac{1}{\sum_{k=-Q}^{Q} \mu(k) e^{-j\omega k}} = \frac{1}{|B(e^{j\omega})|^2}$$

Οι παράμετροι μ(k) υπολογίζονται χρησιμοποιώντας τους LPC συντελεστές α_k και διακύμανση του λάθους πρόβλεψης P_e.

$\mu(k) = \frac{1}{P_e} \sum_{i=0}^{Q-k} (Q+1-k-2i)\alpha_i \alpha_{i+k}^*, \qquad k: 0, \ldots, Q$

$\mu(k) = \mu(-k), \qquad k: -Q, \ldots, -1$

Τα βήματα που χρησιμοποιούμε για τον υπολογισμό των συντελεστών PMVDR φαίνονται παρακάτω. Ο χαρακτηρισμός perceptual στους συντελεστές οφείλεται στο ότι πραγματοποιούμε warping του φάσματος του σήματος στην κλίμακα mel.

1. Χωρίζουμε το σήμα σε frames.

2. Εφαρμόζουμε hamming window σε κάθε frame.

3. Μεταφέρουμε το σήμα στο χώρο της συχνότητας με FFT παίρνουμε το τετράγωνο του μέτρου του φάσματος.

4. Κάνουμε warping του τετραγώνου του μέτρου του φάσματος στην κλίμακα mel. Το warping γίνεται από συχνότητα 0 μέχρι τη συχνότητα δειγματοληψίας.

5. Με IFFT επιστρέφουμε στο πεδίο του χρόνου. Επειδή το warped φάσμα δεν είναι συμμετρικό, το σήμα στο πεδίο του χρόνου είναι μιγαδικό.

6. Εκτελούμε τον αλγόριθμο Levinson-Durbin για να βρούμε τους LPC συντελεστές α_i, οι οποίοι είναι μιγαδικοί αριθμοί.

7. Μετατρέπουμε τους LPC συντελεστές α_i στους MVDR συντελεστές μ(k) σύμφωνα με τον τύπο που παρουσιάστηκε παραπάνω. Οι συντελεστές μ(k) έχουν ερμιτιανή συμμετρία.

8. Υπολογίζουμε το MVDR spectrum τάξης Q.

9. Από το spectrum N_{cep} συντελεστές cepstrum, οι οποίοι είναι οι ζητούμενοι συντελεστές PMVDR.

Στην υλοποίηση επιλέγουμε να χρησιμοποιήσουμε μη επικαλυπτόμενα frames των 40msec. Επιλέγουμε επίσης να δημιουργήσουμε το MVDR cepstrum τάξης Q=8 και παίρνουμε 20 PMVDR συντελεστές.

4.2.4 Παράγωγοι MFCC και TECC συντελεστών

Εκτός από τους ίδιους τους συντελεστές που παρουσιάστηκαν παραπάνω, είναι ενδιαφέρον να μελετήσουμε πως αυτοί μεταβάλλονται με το χρόνο και αν η πληροφορία για τη μεταβολή τους θα μπορούσε να μας βοηθήσει στο πρόβλημα του εντοπισμού αλλαγών από ομιλία σε μη ομιλία και αντίστροφα. Για το λόγο αυτό χρησιμοποιούμε τις παραγώγους τόσο των MFCC και TECC συντελεστών.

Έτσι όπως θα παρουσιαστεί στη συνέχεια, έγιναν πειράματα για να διαπιστωθεί πως συμπεριφέρεται ο αλγόριθμος πρώτου περάσματος όταν οι συντελεστές που χρησιμοποιούνται είναι οι πρώτες παράγωγοι των 13 MFCC ή TECC συντελεστών.

Σε σχέση με τον υπολογισμό της παραγώγου υλοποιήθηκε μία εκδοχή της παραγώγου, σύμφωνα με τον τύπο:

$$d_t = \frac{\sum_{\theta=1}^{\Theta}(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta}\theta^2}$$

ο οποίος δίνεται στο [52]. Στον παραπάνω τύπο d_t είναι ο delta συντελεστής, δηλαδή η πρώτη παράγωγος, σε χρόνο t, c_t είναι ο στατικός συντελεστής σε χρόνο t ενώ θ είναι το παράθυρο στο οποίο υπολογίζουμε την παράγωγο. Για τα πειράματά μας τυπική τιμή παραθύρου είναι θ=3.

Εκτός από τα παραπάνω πολυδιάστατα χαρακτηριστικά υλοποιήθηκαν και δοκιμάστηκαν πολλά μονοδιάστατα χαρακτηριστικά, τα οποία παρουσιάζονται παρακάτω.

4.2.5 Επιλογή της Teager ενέργειας του πιο ενεργού καναλιού

Στη συνέχεια περιγράφεται μία μέθοδος για τον εντοπισμό της Teager ενέργειας του πιο ενεργού καναλιού και της χρήσης της ως μοναδικού συντελεστή, η οποία περιγράφεται αναλυτικά στα [13] και [5].

Σύμφωνα με τη μέθοδο αυτή, υπολογίζουμε τον μέσο όρο της Teager ενέργειας του φιλτραρισμένου σήματος για κάθε φίλτρο και τελικά επιλέγουμε τη μέγιστη από αυτές τις ενέργειες ως μοναδικό συντελεστή. Έτσι, για κάθε frame έχουμε έναν συντελεστή, που είναι η μέση Teager ενέργεια του περισσότερου ενεργού καναλιού:

$$i = arg \max_{1 \leq k \leq K} (\overline{\Psi[(s * h_k)(n)]})$$

$$Max\ Average\ Teager\ Energy = (\overline{\Psi[(s * h_i)(n)]})$$

όπου Ψ είναι ο τελεστής Teager για διακριτό σήμα και h_i είναι είναι το (gabor) φίλτρο με τη μεγαλύτερη ενέργεια.

4.2.6 Συντελεστής Fractal Dimension

Στο σημείο αυτό θα ακολουθήσει μία συνοπτική περιγραφή της διάστασης ενός fractal αλλά και της θεωρίας σχετικά με τον αποδοτικό υπολογισμό της fractal διάστασης ενός σήματος φωνής. Αναλυτικοί ορισμοί για τη διάσταση ενός fractal αλλά και αποδείξεις των ιδιοτήτων που θα αναφερθούν στη συνέχεια, περιέχονται στα [32] και [33].

Έστω ότι η συνεχής, πραγματική συνάρτηση S(t), $0 \leq t \leq T$ αναπαριστά ένα σήμα φωνής μικρής διάρκειας και ότι το συμπαγές, επίπεδο σύνολο

$$F = (t, S(t)) \in R^2 : 0 \leq t \leq T,$$

αναπαριστά το γράφο του σήματος. Ο Mandelbrot (1982) όρισε τη fractal διάσταση του F ίση με την Hausdorff διάστασή του, D_H. Γενικά ισχύει $1 \leq D_H \leq 2$. Το σήμα S ονομάζεται fractal αν ο γράφος του είναι ενα fractal σύνολο, δηλαδή αν η διάσταση του D_H είναι αυστηρά μεγαλύτερη του 1, όπου 1 είναι η τοπολογική διάσταση του F.

Θα εξετάσουμε μία άλλη διάσταση η οποία σχετίζεται στενά με την D_H, την Minkowski-Bouligand διάσταση D_M.

Minkowski-Bouligand διάσταση Η διάσταση αυτή βασίζεται στην ιδέα του Minkowski για την εύρεση του μήκους καμπύλων F. Εφαρμόζουμε dilation στο F, με structuring element δίσκο ακτίνας ε, δημιουργούμε δηλαδή την ένωση των δίσκων αυτών, που είναι κεντραρισμένοι σε όλα τα σημεία του F, και συνεπώς δημιουργούμε ένα Minkowski cover. Βρίσκουμε το εμβαδό A(ε) του dilated συνόλου και θέτουμε το μήκος του ίσο με $\lim_{\varepsilon \to 0} L(\varepsilon)$, όπου L(ε) = A(ε)/2ε. Τότε, η Minkowski-Bouligand διάσταση D_M είναι η σταθερά D, στο νόμο $L(\varepsilon) \propto \varepsilon^{1-D}$, καθώς $\varepsilon \to 0$, τον οποίο το μήκος L(ε) υπακούει, αν το F είναι fractal.

Ισχύει γενικά ότι $1 \leq D_H \leq D_M \leq 2$. Στο εξής θα ασχοληθούμε μόνο με τη διάσταση D_M, την οποία θα αποκαλούμε fractal διάσταση D, λόγω των παρακάτω χρήσιμων ιδιοτήτων της:

1. Σχετίζεται στενά με τη διάσταση D_H και έτσι μπορεί να εκφράσει τα fractal χαρακτηριστικά ενός σήματος
2. Συμπίπτει με την διάσταση D_H σε πολλές περιπτώσεις πρακτικής σημασίας.
3. Είναι πολύ πιο εύκολα υπολογίσιμη από την D_H.
4. Θα εφαρμοστεί σε δειγματοληπτημένα σήματα, όπου ούτως ή άλλως οι περισσότερες μέθοδοι δίνουν προσεγγιστικά αποτελέσματα.

Αποδεικνύεται ότι η διάσταση D δεν αλλάζει αν αντικαταστήσουμε τους δίσκους στο Minkowski cover του F, με κάποιο άλλο συμπαγές, επίπεδο σχήμα B. Συνεπώς, αν $\varepsilon B = \{\varepsilon b : b \in B\}$, είναι ένα σχήμα B σε κλίμακα ε, παίρνουμε κατανομές εμβαδού για διάφορα σχήματα σε διάφορες κλίμακες:

$$A_B(\varepsilon) = area(F \oplus \varepsilon B)$$

όπου $F \oplus \varepsilon B$ είναι:

$$F \oplus \varepsilon B = \{z + \varepsilon b \in R^2 : z \in F, b \in B\}$$

Η απειροστή τάξη της πολυκλιμακωτής συνάρτησης εμβαδού δίνει τη fractal διάσταση του F, δηλαδή:

$$D = 2 - \lim_{\varepsilon \to 0} \frac{\log[A_B(\varepsilon)]}{\log(\varepsilon)}$$

Θεωρώντας τώρα ότι $A_B(\varepsilon) \propto \varepsilon^{2-D}$, καθώς $\varepsilon \to 0$, έχουμε:

$$\log[A_B(\varepsilon)] = (2 - D)\log(\varepsilon) + constant, \quad \varepsilon \to 0$$

Συνεπώς, η διάσταση D Μπορεί να υπολογιστεί στην πράξη χρησιμοποιώντας την μέθοδο ελαχίστων τετραγώνων και μετρώντας την κλίση της γραμμής, που προσεγγίζει μία ευθεία σε ορισμένη περιοχή κλιμάκων ε, στη γραφική παράσταση του $\log[A_B(\varepsilon)]$ συναρτήσει του $\log \varepsilon$.

Παρατηρούμε όμως ότι για την υλοποίηση του dilation $F \oplus \varepsilon B$ εφαρμόσαμε δισδιάστατη επεξεργασία του μονοδιάστατου σήματος S(t), αυξάνοντας έτσι την υπολογιστική

πολυπλοκότητα. Για να αποφύγουμε την αύξηση της πολυπλοκότητας, θα υπολογίσουμε το εμβαδό $A_B(\varepsilon)$ χρησιμοποιώντας μονοδιάστατη επεξεργασία του σήματος S(t). Για το σκοπό αυτό ορίζουμε τις μορφολογικές πράξεις dilation και erosion του σήματος S(t) από μία πραγματική συνάρτηση G(t), ως εξής:

$$(S \oplus G)(t) = \sup_x \{S(x) + G(t-x)\}$$

$$(S \ominus G)(t) = \inf_x \{S(x) - G(x-t)\}$$

Οι παραπάνω πράξεις dilation και erosion του σήματος είναι μονοδιάστατες και παρόμοιες με τη συνέλιξη και τη συσχέτιση αντίστοιχα. Αν τώρα επιλέξουμε οποιοδήποτε συμπαγές, απλά συνδεδεμένο και συμμετρικό επίπεδο σύνολο και ορίσουμε:

$$G_\varepsilon(t) = sup\{y \in R : (t, y) \in \varepsilon B\}$$

ως τη συνάρτηση (structuring element), της οποίας ο γράφος είναι το άνω όριο του εB, παίρνουμε το επιθυμητό εμβαδό $A_B(\varepsilon)$ ως εξής:

$$A_B(\varepsilon) = \int_0^T [(S \oplus G_\varepsilon)(t) - (S \ominus G_\varepsilon)(t)]dt + O(\varepsilon^2)$$

όπου $S \oplus G_\varepsilon$ και $S \ominus G_\varepsilon$ είναι οι μονοδιάστατες πράξεις dilation και erosion που ορίστηκαν παραπάνω.

Για την περίπτωση διακριτού σήματος S[n], n=0,1,...,N, οι τύποι που παρουσιάστηκαν παίρνουν τη μορφή που φαίνεται παρακάτω και απτελούν το διακριτό αλγόριθμο για την υλοποίηση των προηγούμενων ιδεών:

$$S \oplus G[n] = \max_{-1 \le k \le 1} \{S[n+k] + G[k]\}, \quad \varepsilon = 1$$

$$S \ominus G[n] = \min_{-1 \le k \le 1} \{S[n+k] - G[k]\}, \quad \varepsilon = 1$$

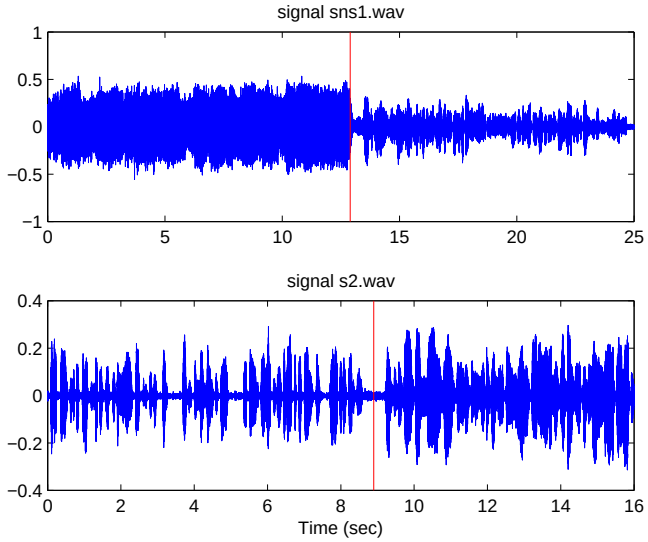
$$S \oplus G_{\varepsilon+1} = (S \oplus G_\varepsilon) \oplus G, \quad \varepsilon \ge 2$$

$$S \ominus G_{\varepsilon+1} = (S \ominus G_\varepsilon) \ominus G, \quad \varepsilon \ge 2$$

όπου ε=1,2,...,ε_{max}. Για τον υπολογισμό των εμβαδών $A_B[\varepsilon]$ αντικαθιστούμε το ολοκλήρωμα $\int_0^T$ με την άθροιση $\sum_{n=0}^{N}$.

Στην παρούσα υλοποίηση, υπολογίζουμε τη fractal διάσταση του διακριτού σήματος φωνής που εξετάζουμε σε κάθε frame μήκους 40msec. Άρα για κάθε frame έχουμε έναν συντελεστή fractal dimension. Ο συντελεστής αυτός θα μπορούσε να χρησιμοποιηθεί μαζί με τους 13 MFCC ή τους 13 TECC συντελεστές, για να δημιουργήσει ένα νέο σύνολο χαρακτηριστικών που θα μπορούσε να χρησιμοποιηθεί από τον αλγόριθμο πρώτου περάσματος. Εντούτοις, επειδή παρατηρήθηκε πειραματικά ότι η προσθήκη του συντελεστή fractal dimension δεν βελτιώνει την απόδοση των MFCC ή TECC χαρακτηριστικών, η διάσταση fractal θα αξιοποιηθεί καλύτερα ξεχωριστά σε ένα δεύτερο πέρασμα.

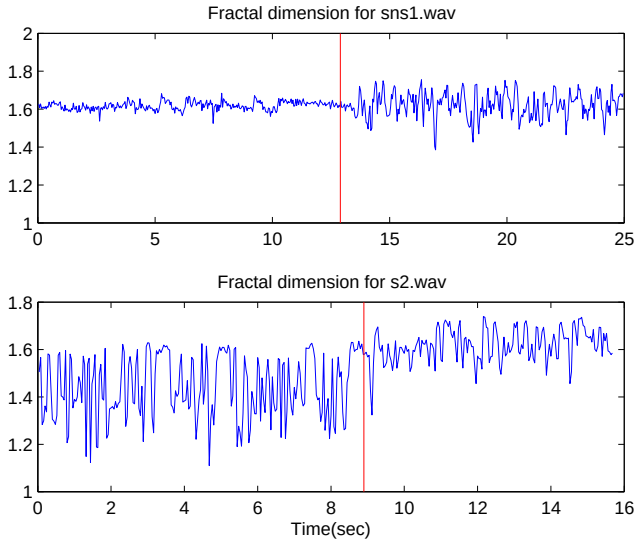
Στην εικόνα 4.1 φαίνονται τα ηχητικά σήματα σε δύο περιπτώσεις. Στην πρώτη περίπτωση έχουμε μετάβαση από μη ομιλία (μουσική) σε ομιλία ενώ στη δεύτερη περίπτωση έχουμε αλλαγή ομιλητών.



Εικόνα 4.1: Στην πάνω εικόνα φαίνεται το ηχητικό σήμα για μετάβαση από μουσική σε φωνή(αλλαγή στα 12.9 sec) ενώ στην κάτω εικόνα φαίνεται το ηχητικό σήμα για αλλαγή ομιλητών(αλλαγή στα 8.9 sec). Η κάθετη γραμμή δείχνει το σημείο αλλαγής.

Στη εικόνα 4.2 φαίνεται η Fractal διάσταση ενός σήματος σε αυτές τις δύο περιπτώσεις. Παρατηρούμε ότι η fractal διάσταση διαφέρει αισθητά μεταξύ ομιλίας και μη ομιλίας γεγονός που δείχνει ότι το χαρακτηριστικό αυτό θα μπορούσε να βοηθήσει στον εντοπισμό αλλαγών

κατηγορίας 1. Η fractal διάσταση διαφέρει λιγότερο αισθητά μεταξύ ομιλητών όπως φαίνεται στη δεύτερη περίπτωση.



Εικόνα 4.2: Στην πάνω εικόνα φαίνεται η fractal διάσταση για μετάβαση από μουσική σε φωνή(αλλαγή στα 12.9 sec) ενώ στην κάτω εικόνα φαίνεται η fractal διάσταση για αλλαγή ομιλητών(αλλαγή στα 8.9 sec). Η κάθετη γραμμή δείχνει το σημείο αλλαγής.

4.2.7 RMS Τιμή

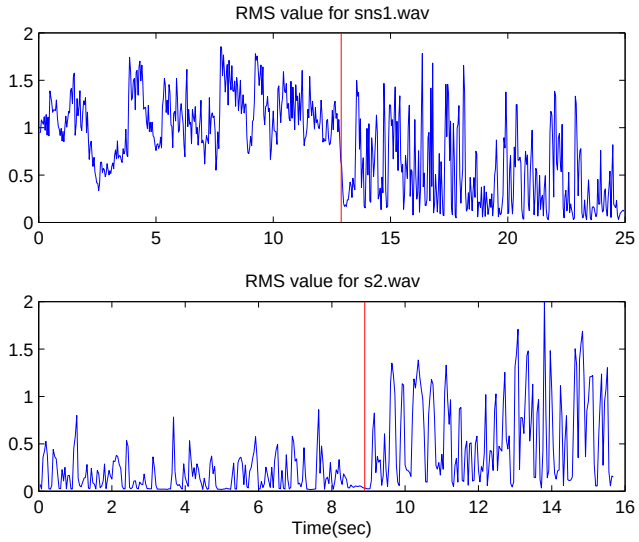
Η RMS (root mean square) τιμή είναι είναι ένα αρκετά δημοφιλές χαρακτηριστικό για τον εντοπισμό αλλαγών σε ένα σήμα, το οποίο χρησιμοποιείται για παράδειγμα στα [41] και [38].

Στην υλοποίηση μας υπολογίζουμε μία RMS τιμή σήματος για κάθε 40msec σήματος σύμφωνα με τον τύπο:

$$RMS = \sqrt{\sum_{i=1}^{N} x^2(i)}$$

όπου $x(i), i = 1, \ldots, N$ είναι οι τιμές του σήματος σε ένα πλαίσιο 40msec.

Έχουμε δηλαδή ένα μονοδιάστατο χαρακτηριστικό για κάθε 40msec σήματος. Ένα παράδειγμα φαίνεται στην εικόνα 4.3 για τις 2 περιπτώσεις που παρουσιάστηκαν και στην εικόνα 4.2, δηλαδή για μετάβαση από μη ομιλία σε ομιλία και για αλλαγή ομιλητών. Και για τις 2 περιπτώσεις παρατηρούμε διαφορά στο σήμα πριν και μετά το σημείο αλλαγής.



Εικόνα 4.3: Στην πάνω εικόνα φαίνονται οι RMS τιμές για μετάβαση από μουσική σε φωνή (αλλαγή στα 12.9 sec) ενώ στην κάτω εικόνα φαίνονται οι RMS τιμές για αλλαγή ομιλητών (αλλαγή στα 8.9 sec). Η κάθετη γραμμή δείχνει το σημείο αλλαγής.

4.2.8 Μέγιστο Πλάτος

Χρησιμοποιούμε επιπλέον ένα χαρακτηριστικό που προτείνεται στο [41],που είναι το μέγιστο πλάτος του σήματος (max amplitude) ή envelope του σήματος. Το μέγιστο πλάτος υπολογίζεται για κάθε 40msec όπως φαίνεται παρακάτω:

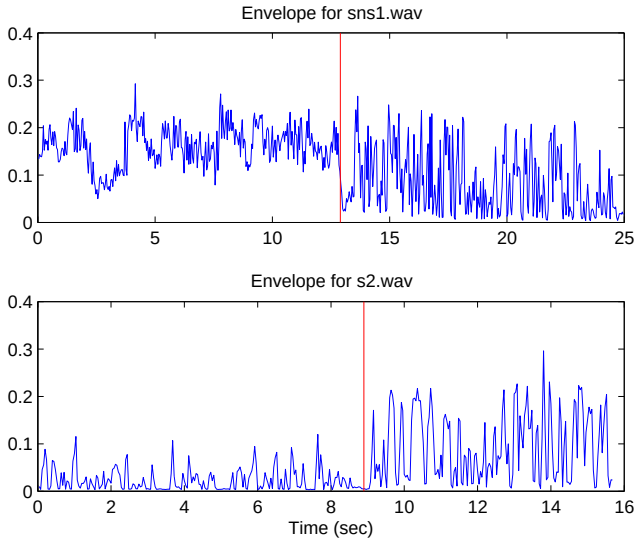
$$Envelope = \max_{i=1,\ldots,N} x(i)$$

όπου $x(i), i = 1, \ldots, N$ είναι οι τιμές του σήματος σε ένα πλαίσιο 40msec.

Ένα παράδειγμα φαίνεται στην εικόνα 4.4 για τις 2 περιπτώσεις που παρουσιάστηκαν και προηγουμένως, δηλαδή για μετάβαση από μη ομιλία σε ομιλία και για αλλαγή ομιλητών. Και για τις 2 περιπτώσεις παρατηρούμε διαφορά στο σήμα πριν και μετά το σημείο αλλαγής.

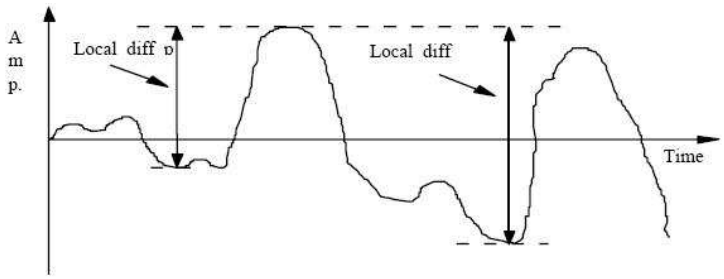
4.2.9 Τοπική Προηγούμενη Διαφορά Πλάτους

Η Τοπική Προηγούμενη Διαφορά Πλάτους (Previous Local Difference) είναι ένα χαρακτηριστικό που προτείνεται στο [41] και ορίζεται ως η διαφορά πλάτους μεταξύ της μέγιστης κορυφής και της προηγούμενης ελάχιστης κοιλάδας μέσα σε ένα πλαίσιο. Ο υπολογισμός



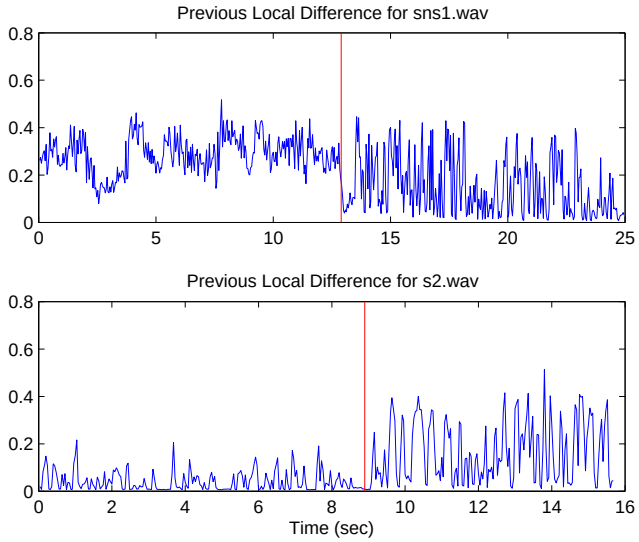
Εικόνα 4.4: Στην πάνω εικόνα φαίνονται οι envelope τιμές για μετάβαση από μουσική σε φωνή (αλλαγή στα 12.9 sec) ενώ στην κάτω εικόνα φαίνονται οι envelope τιμές για αλλαγή ομιλητών (αλλαγή στα 8.9 sec). Η κάθετη γραμμή δείχνει το σημείο αλλαγής.

του χαρακτηριστικού αυτού στην υλοποίηση μας γίνεται για κάθε πλαίσιο 40msec και ο τρόπος υπολογισμού φαίνεται στην εικόνα 4.5.



Εικόνα 4.5: Ο τρόπος υπολογισμού των χαρακτηριστικών Previous Local Difference και Next Local Difference για ένα frame ([41])

Ένα παράδειγμα φαίνεται στην εικόνα 4.6 για τις 2 περιπτώσεις που παρουσιάστηκαν και προηγουμένως, δηλαδή για μετάβαση από μη ομιλία σε ομιλία και για αλλαγή ομιλητών. Και για τις 2 περιπτώσεις παρατηρούμε διαφορά στο σήμα πριν και μετά το σημείο αλλαγής.



Εικόνα 4.6: Στην πάνω εικόνα φαίνονται οι Previous Local Difference τιμές για μετάβαση από μουσική σε φωνή (αλλαγή στα 12.9 sec) ενώ στην κάτω εικόνα φαίνονται οι Previous Local Difference τιμές για αλλαγή ομιλητών (αλλαγή στα 8.9 sec).Η κάθετη γραμμή δείχνει το σημείο αλλαγής.

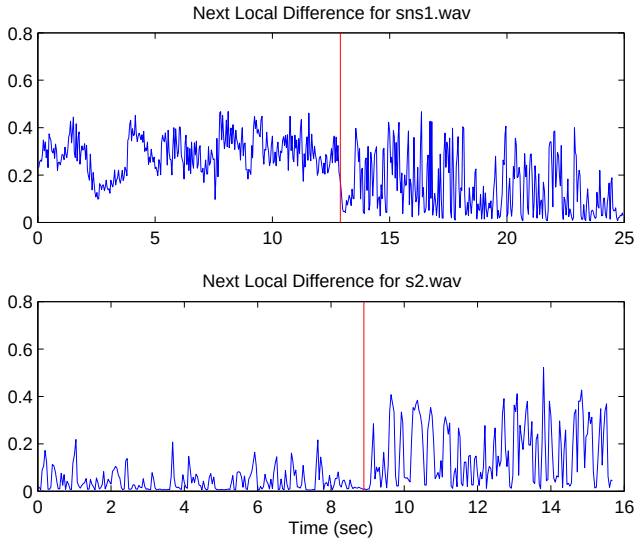
4.2.10 Τοπική Επόμενη Διαφορά Πλάτους

Η Τοπική Επόμενη Διαφορά Πλάτους (Next Local Difference) είναι ένα χαρακτηριστικό που προτείνεται στο [41] και είναι αντίστοιχη με την Τοπική Προηγούμενη Διαφορά Πλάτους. Ορίζεται ως η διαφορά πλάτους μεταξύ της μέγιστης κορυφής και της επόμενης ελάχιστης κοιλάδας μέσα σε ένα πλαίσιο. Ο υπολογισμός του χαρακτηριστικού αυτού στην υλοποίηση μας γίνεται για κάθε πλαίσιο 40msec και ο τρόπος υπολογισμού φαίνεται στην εικόνα 4.5.

Ένα παράδειγμα φαίνεται στην εικόνα 4.7 για τις 2 περιπτώσεις που παρουσιάστηκαν και προηγουμένως, δηλαδή για μετάβαση από μη ομιλία σε ομιλία και για αλλαγή ομιλητών. Και για τις 2 περιπτώσεις παρατηρούμε διαφορά στο σήμα πριν και μετά το σημείο αλλαγής.

4.2.11 Χρήση Παραγώγων των μονοδιάστατων χαρακτηριστικών

Εκτός από τα ίδια τα χαρακτηριστικά που παρουσιάστηκαν παραπάνω, είναι ενδιαφέρον να μελετήσουμε πως αυτά μεταβάλλονται με το χρόνο και αν η πληροφορία για τη μεταβολή τους θα μπορούσε να μας βοηθήσει στο πρόβλημα του εντοπισμού αλλαγών από ομιλία σε μη ομιλία και αντίστροφα καθώς και μεταξύ ομιλητών. Στη πραγματικότητα, τα πειράματα



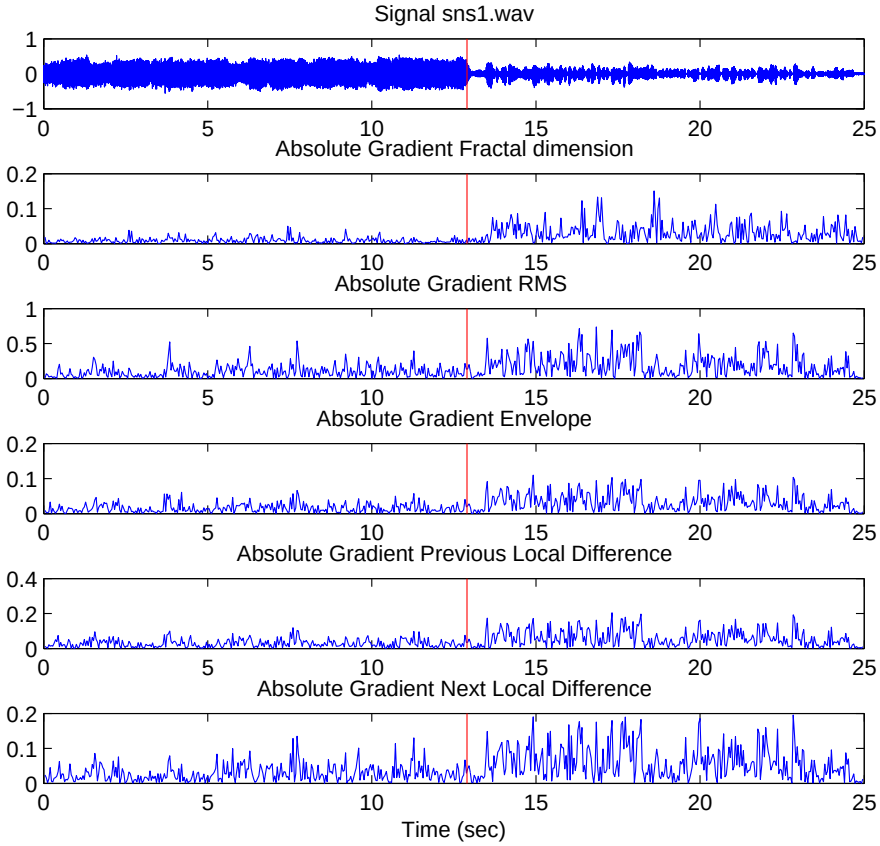
Εικόνα 4.7: Στην πάνω εικόνα φαίνονται οι Next Local Difference τιμές για μετάβαση από μουσική σε φωνή (αλλαγή στα 12.9 sec) ενώ στην κάτω εικόνα φαίνονται οι Next Local Difference τιμές για αλλαγή ομιλητών (αλλαγή στα 8.9 sec).Η κάθετη γραμμή δείχνει το σημείο αλλαγής.

που έγιναν υποδεικνύουν ότι είναι καλύτερο να χρησιμοποιήσουμε την απόλυτη τιμή της παραγώγου των παραπάνω χαρακτηριστικών, αντί των ίδιων των χαρακτηριστικών, καθώς έτσι παίρνουμε καλύτερα αποτελέσματα τόσο ως προς το ποσοστό επιτυχίας όσο και ως προς το false alarm.

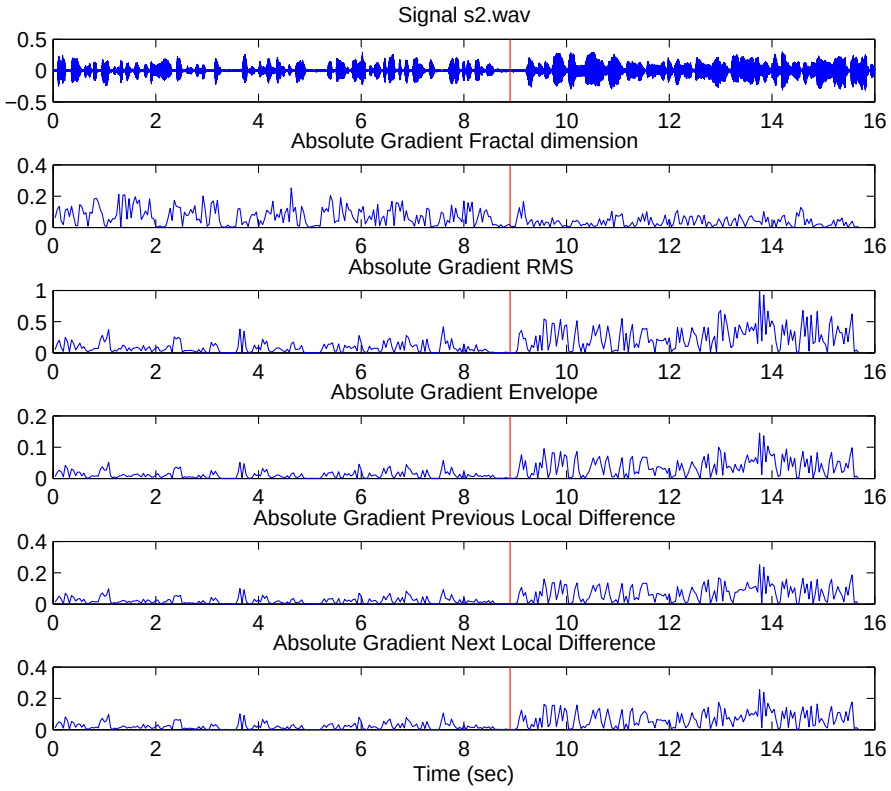
Στην εικόνα 4.8 φαίνονται το ηχητικό σήμα και οι απόλυτες τιμές των παραγώγων για τα 5 χαρακτηριστικά που παρουσιάστηκαν παραπάνω για το παράδειγμα της μετάβασης από μη ομιλία(μουσική) σε ομιλία (αρχείο sns1.wav).

Αντίστοιχα στην εικόνα 4.9 φαίνονται το ηχητικό σήμα και οι απόλυτες τιμές των παραγώγων για τα 5 χαρακτηριστικά που παρουσιάστηκαν παραπάνω για το παράδειγμα της αλλαγής ομιλητών (αρχείο s2.wav).

Και στα 2 παραπάνω παραδείγματα παρατηρούμε μεταβολή των χαρακτηριστικών πριν και μετά τα σημεία αλλαγής, γεγονός που μας επιτρέπει να χρησιμοποιήσουμε τα χαρακτηριστικά αυτά για την εύρεση σημείων αλλαγής.



Εικόνα 4.8: Το ηχητικό σήμα και οι απόλυτες τιμές των παραγώγων για 5 χαρακτηριστικά για την περίπτωση μετάβασης από μουσική σε ομιλία στο αρχείο sns1.wav (αλλαγή στα 12.9 sec).Η κάθετη γραμμή δείχνει το σημείο αλλαγής.



Εικόνα 4.9: Το ηχητικό σήμα και οι απόλυτες τιμές των παραγώγων για 5 χαρακτηριστικά για την περίπτωση αλλαγής ομιλητών στο αρχείο s2.wav (αλλαγή στα 8.9 sec).Η κάθετη γραμμή δείχνει το σημείο αλλαγής.

4.3 Μελέτη Μετρικών Κριτηρίων Κατάτμησης

Στην ενότητα αυτή θα παρουσιαστούν διάφορα metric-based κριτήρια κατάτμησης. Οι metric-based προσεγγίσεις μετρούν ουσιαστικά τη διαφορά μεταξύ δύο διαδοχικών παραθύρων που μετατοπίζονται στο audio stream που εξετάζουμε και εντοπίζουν αλλαγή στο σημείο ανάμεσα στα δύο παράθυρα αν η διαφορά ξεπερνάει κάποιο κατώφλι (δηλαδή αν τα σήματα των δύο παραθύρων προέρχονται από διαφορετικές πηγές). Οι διαφορές των μεθόδων αυτών έγκεινται τόσο στα μέτρα διαφοράς που χρησιμοποιούνται όσο και στις αποφάσεις σχετικά με τα κατώφλια.

4.3.1 Κριτήριο Bayesian Information Criterion - BIC

Μία ευρέως διαδεδομένη μέθοδος που δίνει καλά αποτελέσματα είναι η μέθοδος BIC (Bayesian Information Criterion) που περιγράφεται στο [8].

Έστω ότι έχουμε τα δεδομένα.

$$H_0 = x_1, x_2, \ldots, x_n \sim N(\mu_0, \Sigma_0)$$

και εξετάζουμε την αλλαγή στο πλαίσιο i, δηλαδή:

$$H_1 = x_1, x_2, \ldots, x_i \sim N(\mu_1, \Sigma_1) \quad \textit{και} \quad H_2 = x_{i+1}, x_{i+2}, \ldots, x_n \sim N(\mu_2, \Sigma_2)$$

Ορίζουμε την πιθανότητα αλλαγής στο πλαίσιο i :

$$R(i) = N \cdot log|\Sigma| - N_1 \cdot log|\Sigma_1| - N_2 \cdot log|\Sigma_2|$$

το penalty P, όπου d είναι η διάσταση του χώρου χαρακτηριστικών

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1)log|N|)$$

και το βάρος λ=1.

Είναι:

$$BIC(i) = R(i) - \lambda P$$

Αποφασίζουμε ότι υπάρχει αλλαγή στο σημείο i αν

$$\max_i\{BIC(i)\} > threshold$$

4.3.2 Κριτήριο Weighted Mean Distance - WMD

Το κριτήριο αυτό χρησιμοποιείται στο [51] όταν δεν υπάρχουν αρκετά δεδομένα ώστε να εφαρμόσουμε BIC ή T2 .

Υπολογίζουμε την απόσταση μεταξύ των διαστημάτων H_1 και H_2 από τον τύπο:

$$WMD = \frac{a \cdot b}{a+b} \cdot (\mu_1 - \mu_2)^T I^{-1} (\mu_1 - \mu_2)$$

Όπου a,b είναι το πλήθος των πλαισίων στα H_1 και H_2 αντίστοιχα και I ειναι ο μοναδιαίος πίνακας.

Αποφασίζουμε ότι υπάρχει αλλαγή στο σημείο i αν

$$\max_i \{WMD(i)\} > threshold$$

4.3.3 Κριτήριο T^2

Το κριτήριο αυτό χρησιμοποιείται στο [51] όταν δεν υπάρχουν αρκετά δεδομένα ώστε να εφαρμόσουμε BIC .

Υπολογίζουμε την απόσταση μεταξύ των διαστημάτων H_1 και H_2 από τον τύπο:

$$T^2 = \frac{a \cdot b}{a+b} \cdot (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$$

Όπου a,b είναι το πλήθος των πλαισίων στα H_1 και H_2 και Σ είναι ο πίνακας αυτοσυσχέτισης στο διάστημα H_0.

Αποφασίζουμε ότι υπάρχει αλλαγή στο σημείο i αν

$$\max_i \{T^2(i)\} > threshold$$

4.3.4 Κριτηριο Kullback-Leibler - KL2

Το κριτήριο αυτό αναφέρεται για παράδειγμα στο [20].

Υπολογίζουμε την απόσταση μεταξύ των διαστημάτων H_1 και H_2 από τον τύπο:

$$KL2 = \frac{1}{2} Tr[(\Sigma_1 - \Sigma_2)(\Sigma_2^{-1} - \Sigma_1^{-1})] + \frac{1}{2} Tr[(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T]$$

όπου Σ_1 και Σ_2 είναι οι πίνακες αυτοσυσχέτισης στα διαστήματα H_1 και H_2 αντίστοιχα.

Αποφασίζουμε ότι υπάρχει αλλαγή στο σημείο i αν

$$max\{KL2(i)\} > threshold$$

4.3.5 Κριτηριο Divergence Shape Distance - DSD

Το κριτήριο αυτό αναφέρεται για παράδειγμα στο [20] και αποτελεί μία απλοποίηση του KL2.

Υπολογίζουμε την απόσταση μεταξύ των διαστημάτων H_1 και H_2 από τον τύπο:

$$DSD = \frac{1}{2}Tr[(\Sigma_1 - \Sigma_2)(\Sigma_2^{-1} - \Sigma_1^{-1})]$$

όπου Σ_1 και Σ_2 είναι οι πίνακες αυτοσυσχέτισης στα διαστήματα H_1 και H_2 αντίστοιχα.

4.3.6 Κριτήριο Weighted squared Euclidean Distance - WED

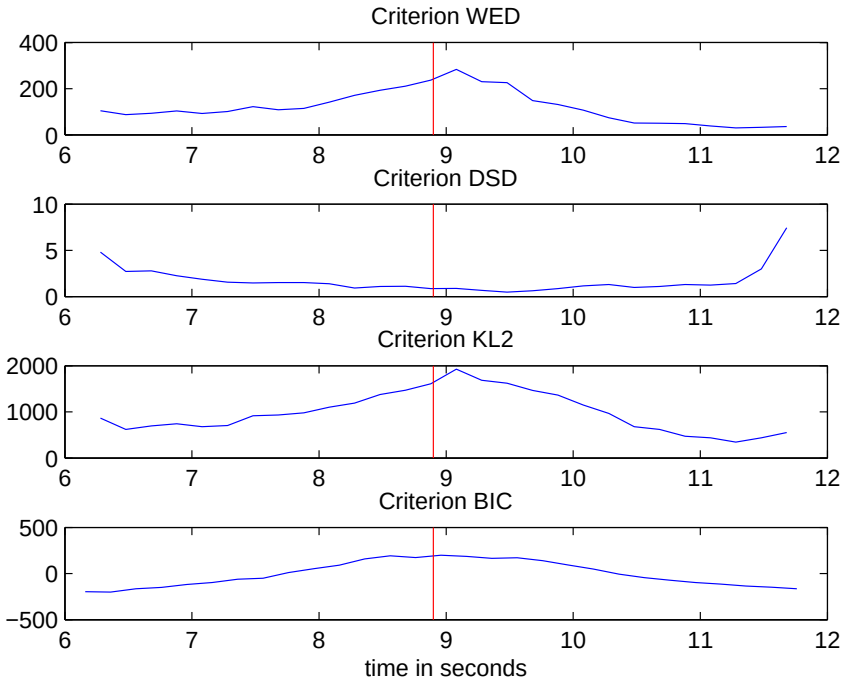
Ένα εναλλακτικό κριτήριο μέτρησης απόστασης μεταξύ δύο παραθύρων είναι το Weighted squared Euclidean Distance (WED) το οποία προτείνεται στο [25]. Τα κριτήριο αυτό βασίζεται στην ευκλείδεια απόσταση μεταξύ των διανυσμάτων χαρακτηριστικών των δύο πλαισίων υπό σύγκριση, ενώ χρησιμοποιούνται επίσης και βάρη που εξαρτώνται από τις διακυμάνσεις των διανυσμάτων χαρακτηριστικών. Τέλος εφαρμόζεται σιγμοιδής συνάρτηση στα βάρη ώστε να αυξηθεί η ικανότητά τους να διακρίνουν διαφορετικά παράθυρα.

4.3.7 Γραφική Σύγκριση των Μετρικών Κριτηρίων

Από τα πειράματα που έγιναν για τη σύγκριση των παραπάνω κριτηρίων παρουσιάζονται ενδεικτικά κάποια γραφικά αποτελέσματα για δύο χαρακτηριστικές περιπτώσεις αλλαγών σε ηχητικό σήμα. Με τον τρόπο αυτό δίνουμε μια ιδέα των αποτελεσμάτων των παραπάνω κριτηρίων για την εύρεση αλλαγών. Τα πειράματα γίνονται στα ηχητικά σήματα που έχουν ήδη παρουσιαστεί στην εικόνα 4.1.

Στην εικόνα 4.10 παρατηρούμε τις τιμές των 4 κριτηρίων WED, DSD, KL2 και BIC για μία περίπτωση αλλαγής ομιλητών. Από το υπό εξέταση σήμα έχουν εξαχθεί 13 MFCC συντελεστές (χωρίς πρώτες και δεύτερες παραγώγους). Για την εξαγωγή των χαρακτηριστικών χρησιμοποιήθηκαν 35 τριγωνικά φίλτρα. Θεωρούμε επίσης frames των 40msec και δεν υπάρχει επικάλυψη διαδοχικών frames. Η εικόνα δείχνει τις τιμές που παίρνει το εκάστοτε κριτήριο για διαδοχικά χρονικά σημεία, κάνοντας κάθε φορά την υπόθεση ότι το υπό εξέταση σημείο είναι ένα σημείο αλλαγής. Παρατηρούμε ότι οι καμπύλες τείνουν να έχουν κορυφή γύρω από το πραγματικό σημείο αλλαγής που βρίσκεται στα 8.9sec, αν και υπάρχουν αποκλίσεις από την πραγματική τιμή. Κατά συνέπεια, οι τιμές των κριτηρίων μεγιστοποιούνται κοντά στα πραγματικά σημεία αλλαγής, γεγονός που οδηγεί στον εντοπισμό των αλλαγών. Πιο συγκεκριμενα, παρατηρούμε ότι τα κριτήρια WED, KL2 και

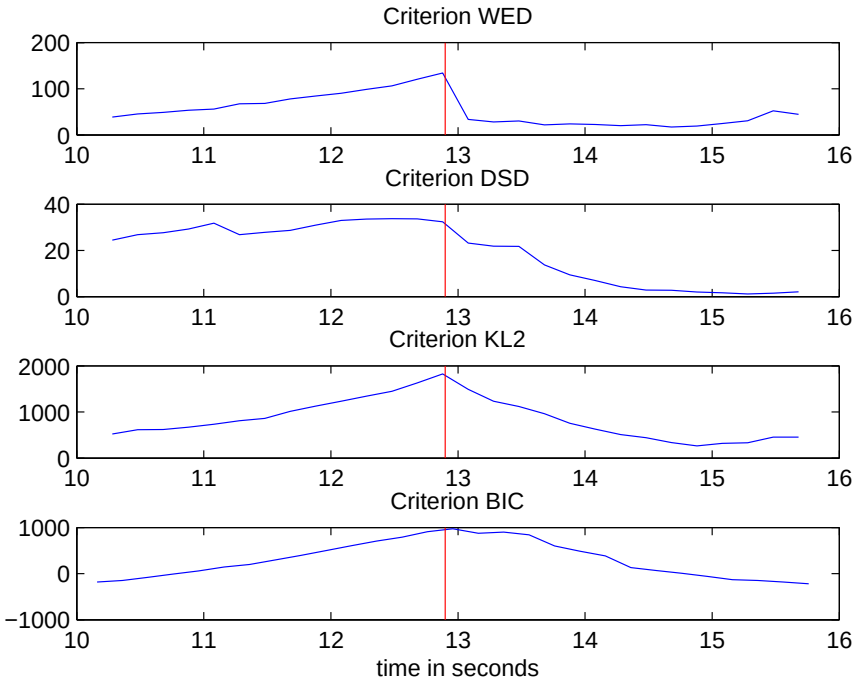
BIC λειτουργούν καλά, ενώ το DSD αποτυγχάνει να εντοπίσει την πραγματική αλλαγή και μεγιστοποιείται στα άκρα του υπό εξέταση διαστήματος. Τέλος παρατηρούμε ότι, αν και οι μορφές των καμπύλων έχουν κάποιες ομοιότητες, οι τιμές των κριτηρίων διαφέρουν πολύ μεταξύ τους με το KL2 να δίνει γενικά τις υψηλότερες τιμές ενώ το BIC να δίνει τιμές κοντά στο 0. Κατά συνέπεια, τα κατώφλια που θα χρησιμοποιήσουμε για τον εντοπισμό των αλλαγών εξαρτώνται από το κριτήριο που χρησιμοποιείται κάθε φορά.



Εικόνα 4.10: Καμπύλες τιμών 4 κριτηρίων εντοπισμού αλλαγών για την περίπτωση αλλαγής ομιλητών στο αρχείο s2.wav (αλλαγή στα 8.9 sec)

Ένα δεύτερο παράδειγμα δίνεται στην εικόνα 4.11 παρατηρούμε τις τιμές των 4 κριτηρίων WED, DSD, KL2 και BIC για μία περίπτωση αλλαγής από θόρυβο(μουσική) σε ομιλία. Χρησιμοποιήθηκαν και εδώ 13 MFCC χαρακτηριστικά. Το πραγματικό σημείο αλλαγής βρίσκεται στα 12.9sec. Παρατηρούμε ότι και σε αυτήν την περίπτωση οι τιμές των κριτηρίων BIC, KL2 και WED μεγιστοποιούνται γύρω από το πραγματικό σημείο αλλαγής ενώ το DSD αποτυγχάνει να εντοπίσει την αλλαγή. Τέλος, είναι ενδιαφέρον να σημειώσουμε ότι σε αυτήν την περίπτωση το κριτήριο BIC παίρνει υψηλές τιμές γύρω από το σημεία αλλαγής, δηλαδή αισθητά πάνω από το 0. Αυτό ίσως δηλώνει ότι το BIC Μπορεί να εντοπίσει με

μεγαλύτερη ευκολία αλλαγές μεταξύ ομιλίας και μη ομιλίας από ότι αλλαγές ομιλητών. Το γεγονός αυτό θα αναλυθεί λεπτομερέστερα σε επόμενες ενότητες όπου θα εξεταστεί η απόδοση του BIC για διάφορες κατηγορίες αλλαγών.



Εικόνα 4.11: Καμπύλες τιμών 4 κριτηρίων εντοπισμού αλλαγών για την περίπτωση αλλαγής από μουσική σε ομιλία στο αρχείο sns1.wav (αλλαγή στα 12.9 sec)

4.4 Ανάπτυξη Αλγορίθμων Κατάτμησης

Στην ενότητα αυτή παρουσιάζονται οι αλγόριθμοι που υλοποιήθηκαν για τον εντοπισμό των αλλαγών σε audio-streams. Η βασική διαδικασία εντοπισμού αλλαγών γίνεται με χρήση κριτηρίου BIC. Το κριτήριο αυτό μπορεί να συνδυαστεί με πολυδιάστατα χαρακτηριστικά του σήματος όπως τα MFCC ή τα TECC χαρακτηριστικά. Στη συνέχεια, για να βελτιώσουμε τα αποτελέσματα του πρώτου περάσματος παρουσιάζουμε ένα δεύτερο πέρασμα που έχει ως στόχο να αποφασίσει αν ένα υποψήφιο σημείο αλλαγής που βρέθηκε από το πρώτο πέρασμα είναι πραγματικό σημείο αλλαγής. Για να πάρει αυτήν την απόφαση το δεύτερο πέρασμα χρησιμοποιεί κριτήρια πιθανότητας. Το δεύτερο πέρασμα χρησιμοποιεί μονοδιάστατα χαρακτηριστικά όπως η Fractal διάσταση ή η RMS τιμή.

4.4.1 Πρώτο Πέρασμα με χρήση BIC ή προσεγγίσεών του

Ο αλγόριθμος πρώτου περάσματος για τον εντοπισμό αλλαγών βασίζεται σε συνδυασμό των συστημάτων που περιγράφονται στα [19] και [8]. Συγκεκριμένα, γίνεται χρήση του κριτηρίου BIC (Bayesian Information Criterion) ενώ σε περιπτώσεις που τα διαθέσιμα δεδομένα δεν είναι αρκετά για τον αξιόπιστο υπολογισμό του BIC, χρησιμοποιούνται τα κριτήρια T^2 και WMD (Weighted Mean Distance).

Ο αλγόριθμος που υλοποιήθηκε χρησιμοποιεί τα παραπάνω 3 κριτήρια για τον εντοπισμό αλλαγών. Συγκεκριμένα αρχίζει τα εξετάζει το συνολικό audio-stream παίρνοντας ένα ελάχιστο παράθυρο [start, end] στην αρχή και εξετάζοντας την ύπαρξη αλλαγής στο παράθυρο αυτό. Αν δεν βρεθεί αλλαγή το τελικό άκρο του παραθύρου αυξάνεται ενώ αν βρεθεί αλλαγή παίρνουμε πάλι ένα ελάχιστο παράθυρο μετατοπισμένο δεξιά του σημείου αλλαγής. Με αυτήν την τεχνική εξασφαλίζουμε ότι σε κάθε παράθυρο που γίνεται αναζήτηση σημείου αλλαγής βρίσκεται το πολύ ένα τέτοιο σημείο.

Όσον αφορά την επιλογή κριτηρίου αυτή εξαρτάται από το πλήθος των δεδομένων που έχουμε στη διάθεσή μας, δηλαδή από το μήκος του παραθύρου που εξετάζουμε. Το κριτήριο WMD χρειάζεται σχετικά λίγα δεδομένα για να δώσει αξιόπιστη εκτίμηση, το κριτήριο T^2 χρειάζεται περισσότερα δεδομένα από το WMD αλλά λιγότερα από το BIC και το BIC είναι το κριτήριο που χρειάζεται τα περισσότερα δεδομένα για να δώσει αξιόπιστη εκτίμηση. Ουσιαστικά τα WMD και T^2 αποτελούν προσεγγίσεις του BIC όταν έχουμε λίγα δεδομένα.

Άρα η λογική που ακολουθούμε είναι να ξεκινάμε από ένα ελάχιστο παράθυρο [start,end] στο οποίο χρησιμοποιούμε WMD και το οποίο αυξάνουμε σταδιακά κατά ένα βήμα έστω step δηλαδή $end = end + step$. Όταν ξεπεραστεί ένα πρώτο κατώφλι μήκους χρησιμοποιούμε T^2 και όταν ξεπεραστεί ένα δεύτερο κατώφλι μήκους χρησιμοποιούμε BIC. Κάθε

φορά που εντοπίζεται αλλαγή στην επόμενη επανάληψη παίρνουμε ένα ελάχιστο παράθυρο μετατοπισμένο δεξιά του σημείου αλλαγής, άρα χρησιμοποιούμε και πάλι WMD.

Όσον αφορά την εύρεση αλλαγής σε ένα συγκεκριμένο διάστημα [start,end], τα 3 κριτήρια χωρίζουν το προς εξέταση διάστημα H_0 σε 2 υποδιαστήματα H_1 και H_2:

$$H_0 = x_1, x_2, \ldots, x_n \sim N(\mu_0, \Sigma_0)$$

$$H_1 = x_1, x_2, \ldots, x_i \sim N(\mu_1, \Sigma_1) \quad \textit{και} \quad H_2 = x_{i+1}, x_{i+2}, \ldots, x_n \sim N(\mu_2, \Sigma_2)$$

και εξετάζουν την πιθανότητα αλλαγής στο σημείο i. Το σημείο i προς εξέταση μεταβάλλεται από την αρχή προς το τέλος του [start,end] αν και βέβαια η αρχική τιμή του βρίσκεται κάποιο ελάχιστο αριθμό σημείων δεξιά του start ώστε να εξασφαλίζεται η ύπαρξη αρκετών δεδομένων για τον υπολογισμό του H_1. Ομοίως η τελική τιμή του i βρίσκεται κάποιο ελάχιστο αριθμό σημείων αριστερά του end ώστε να εξασφαλίζεται η ύπαρξη αρκετών δεδομένων για τον υπολογισμό του H_2. Κατά συνέπεια, ένα σημείο αλλαγής που βρίσκεται πολύ κοντά στα άκρα του [start,end] θα βρεθεί με χειρότερη ακρίβεια από ένα σημείο που βρίσκεται πιο κεντρικά στο [start, end].

Σχετικά με τα χαρακτηριστικά που εξάγουμε από το σήμα για να χρησιμοποιηθούν από το πρώτο πέρασμα, δοκιμάζονται διάφορα πολυδιάστατα χαρακτηριστικά όπως τα MFCC, τα TECC, οι παράγωγοί τους αλλά και συνδυασμοί των πολυδιάστατων χαρακτηριστικών με μονοδιάστατα όπως η fractal διάσταση. Τα αποτελέσματα τέτοιων πειραμάτων θα παρουσιαστούν στην ενότητα των πειραματικών αποτελεσμάτων.

Σχετικά με τον αλγόριθμο έχουμε να παρατηρήσουμε τα παρακάτω:

Κατώφλια Ο αλγόριθμος χρησιμοποιεί πολλά κατώφλια τα οποία πρέπει να ρυθμιστούν κατάλληλα. Πράγματι, κάθε ένα από τα κριτήρια BIC, T^2 και WMD χρησιμοποιεί κατώφλι για τον εντοπισμό αλλαγής. Επιπλέον ο αλγόριθμος, χρησιμοποιεί κατώφλια για να ορίσει το ελάχιστο μήκος διαστήματος [start, end] αλλά και τα μήκη διαστήματος στα οποία αλλάζουμε το κριτήριο από WMD σε T^2 και από T^2 σε BIC. Οι τιμές που χρησιμοποιήθηκαν για να κατώφλια είτε προτείνονται στη βιβλιογραφία, είτε βρίσκεται πειραματικά ότι δίνουν καλά αποτελέσματα ως προς τα ποσοστά επιτυχίας εντοπισμού αλλαγών.

Επικαλυπτόμενα frames Ο αλγόριθμος εξετάζει μη επικαλυπτόμενα παράθυρα για την εύρεση αλλαγών. Όμως αν επιλέξουμε να έχουμε επικαλυπτόμενα frames για την εξαγωγή των χαρακτηριστικών τότε είναι πιθανό αλγόριθμος να εντοπίσει δύο φορές

την ίδια αλλαγή. Πράγματι, παρατηρείται και πειραματικά ότι όταν χρησιμοποιούμε επικαλυπτόμενα frames κατά το στάδιο στη εξαγωγή χαρακτηριστικών, έχουμε υπερκατάτμηση και βρίσκονται περισσότερα του ενός κοντινά σημεία αλλαγής που αντιστοιχούν στην ίδια πραγματική αλλαγή. Για το λόγο αυτό χρησιμοποιούμε μη επικαλυπτόμενα frames και παίρνουμε μέσους όρους των χαρακτηριστικών μέσα στο ίδιο frame, όπως εξηγήθηκε σε προηγούμενη ενότητα.

Πολυπλοκότητα αλγορίθμου Ο αλγόριθμος εύρεσης αλλαγών περιέχει έναν εξωτερικό βρόχο (loop) στο συνολικό audio-stream που εξετάζουμε και καθένα από τα κριτήρια περιέχει επίσης έναν βρόχο στο σύνολο του διαστήματος προς εξέταση. Κατά συνέπεια ο αλγόριθμος έχει πολυπλοκότητα $O(n^2)$ όπου n είναι ο αριθμός των πλαισίων (frames) του συνολικού audio-stream που εξετάζουμε.

Ταχύτητα αλγορίθμου Εδώ πρέπει να σημειωθεί ότι η ταχύτητα δεν εξαρτάται μόνο από το πλήθος των χαρακτηριστικών που χρησιμοποιούνται αλλά και από το πλήθος των αλλαγών που τελικά βρίσκονται σε ένα stream προς εξέταση. Πράγματι, αν έχουμε εντοπισμό πολλών αλλαγών, χρησιμοποιούνται συχνότερα τα κριτήρια WMD και T^2 που είναι σχετικά γρήγορα, ενώ αν δεν βρίσκονται αλλαγές καταλήγουμε να έχουμε συνεχή χρήση του BIC σε μεγάλα παράθυρα, διαδικασία που είναι περισσότερο χρονοβόρα. Η παρατήρηση αυτή δείχνει ότι όσον αφορά τον τομέα της ταχύτητας είναι καλύτερο να έχουμε υπερκατάτμηση από υποκατάτμηση στο πρώτο πέρασμα.

4.4.2 Αλγόριθμος Δεύτερου Περάσματος

Ο αλγόριθμος αυτός είναι μία καινούρια ιδέα που χρησιμοποιεί μονοδιάστατα χαρακτηριστικά που βρίσκονται για κάθε frame (των 40msec) του audio-stream με σκοπό να μειώσει τα ποσοστά false alarm, δηλαδή να αποφασίσει ποιες από τις αλλαγές που βρέθηκαν στο πρώτο πέρασμα δεν αντιστοιχούν σε πραγματικές αλλαγές. Συγκεκριμένα, ο αλγόριθμος αυτός δέχεται ως είσοδο τα πιθανά σημεία αλλαγής, που βρίσκονται από τον αλγόριθμο πρώτου περάσματος, και αποφασίζει ότι πραγματικά σημεία αλλαγής είναι αυτά στα οποία η πιθανότητα αλλαγής ξεπερνάει κάποιο κατώφλι. Για τον υπολογισμό της πιθανότητας αλλαγής σε ένα σημείο χρησιμοποιείται η μέθοδος που παρουσιάζεται στο [38].

Ο αλγόριθμος αυτός αναπτύχθηκε με σκοπό την μείωση των ποσοστών false alarm, δηλαδή των σημείων αλλαγής που βρίσκονται από το πρώτο πέρασμα και δεν αντιστοιχούν σε πραγματική αλλαγή. Επιπλέον, ο αλγόριθμος αυτός αξιοποιεί μονοδιάστατα χαρακτηριστικά όπως η fractal διάσταση και η RMS τιμή. Τα χαρακτηριστικά αυτά, αν και περιέχουν χρήσιμη πληροφορία για το πρόβλημα εντοπισμού αλλαγών κυρίως μεταξύ ομιλίας και

μη ομιλίας, είναι δύσκολο να χρησιμοποιηθούν στο πρώτο πέρασμα μαζί με πολυδιάστατα χαρακτηριστικά όπως οι MFCC συντελεστές. Πειραματικά αποτελέσματα που θα παρουσιαστούν στην αντίστοιχη ενότητα δείχνουν ότι ο συνδυασμός διαφορετικών διανυσμάτων χαρακτηριστικών σε ένα κοινό συνολικό διάνυσμα χαρακτηριστικών δεν οδηγεί απαραίτητα σε βελτίωση των αποτελεσμάτων, ακόμα και αν τα επιμέρους χαρακτηριστικά περιέχουν χρήσιμη πληροφορία. Κατά συνέπεια ο συνδυασμός τους θα πρέπει να γίνει με πιο σύνθετους τρόπους, πιθανόν σε πολλά περάσματα. Ο αλγόριθμος δεύτερου περάσματος είναι μία ιδέα προς αυτή την κατεύθυνση.

Αλγόριθμος Δεύτερου Περάσματος με χρήση ενός Μονοδιάστατου Χαρακτηριστικού

Για τον υπολογισμό της πιθανότητας επιλέγουμε ένα από τα μονοδιάστατα χαρακτηριστικά που έχουν παρουσιαστεί στην ενότητα της εξαγωγής χαρακτηριστικών, για παράδειγμα απόλυτη τιμή της παραγώγου της Fractal διάστασης, απόλυτη τιμή της παραγώγου της RMS τιμής κλπ. και εξετάζουμε την γραφική παράσταση του χαρακτηριστικού αυτού ανά 40msec σήματος. Η μέθοδος είναι να μοντελοποιήσουμε τη γραφική παράσταση με τη γενικευμένη κατανομή χ^2, όπως γίνεται στο [38]. Στη συνέχεια προσδιορίζουμε την πιθανότητα να υπάρχει αλλαγή γύρω από ένα υποψήφιο σημείο αλλαγής που έχει ήδη βρεθεί από τον αλγόριθμο πρώτου περάσματος.

Συγκεκριμένα, η γενικευμένη κατανομή χ^2 ορίζεται από τη συνάρτηση πυκνότητας πιθανότητας:

$$p(x) = \frac{x^a e^{-bx}}{b^{a+1}\Gamma(a+1)}, \quad x \geq 0.$$

όπου $\Gamma(.)$ είναι η κατανομή Γάμμα και οι παράμετροι a και b σχετίζονται με τη μέση τιμή και τη διακύμανση του σήματος που εξετάζουμε:

$$a = \frac{\mu^2}{\sigma^2} - 1$$

$$b = \frac{\sigma^2}{\mu^2}$$

Έστω λοιπόν ότι ένα γνωστό πιθανό σημείο αλλαγής είναι το check. Θεωρούμε τότε δύο παράθυρα πριν και μετά το σημείο αλλαγής υπό εξέταση, έστω [start,check] και [check,finish]. Μετράμε την ομοιότητα των δύο διαστημάτων χρησιμοποιώντας το παρακάτω μέτρο ομοιότητας που βασίζεται στις συναρτήσεις πυκνότητας πιθανότητας που μοντελοποιούν τα δύο υπό εξέταση διαστήματα:

$$\rho(p_1, p_2) = \int \sqrt{p_1(x)p_2(x)}dx$$

Το παραπάνω μέτρο παίρνει τιμές στο διάστημα [0,1], όπου η τιμή 1 σημαίνει όμοιες κατανομές και η τιμή 0 σημαίνει εντελώς μη επικαλυπτόμενες κατανομές. Για το λόγο αυτό η τιμή $1 - \rho(p_1, p_2)$, που είναι γνωστή ως απόσταση Matusita, μπορεί να ερμηνευτεί ως η απόσταση μεταξύ του περιεχομένου των δύο διαστημάτων.

Για τη γενικευμένη χ^2 κατανομή το μέτρο ομοιότητας εξαρτάται από τις παραμέτρους a_i, b_i των δύο υπό εξέταση διαστημάτων:

$$\rho(p_1, p_2) = \frac{\Gamma(\frac{a_1+a_2}{2}+1)}{\sqrt{\Gamma(a_1+1)\Gamma(a_2+1)}} \frac{2^{\frac{a_1+a_2}{2}+1} b_1^{\frac{a_2+1}{2}} b_2^{\frac{a_1+1}{2}}}{(b_1+b_2)^{\frac{a_1+a_2}{2}+1}}$$

Συνεπώς για το σημείο check υπολογίζουμε την τιμή D που μας δίνει την πιθανότητα το σημείο αυτό να είναι ένα σημείο αλλαγής:

$$D = 1 - \rho(p_1, p_2)$$

Στην πραγματικότητα μετράμε τις τιμές D για μία περιοχή γύρω από το υποψήφιο σημείο αλλαγής, δίνουμε κατάλληλα βάρη στις τιμές D, δηλαδή μεγαλύτερα βάρη στις τιμές που αντιστοιχούν σε σημείο πολύ κοντά στο υποψήφιο σημείο και μικρότερα βάρη στις τιμές που αντιστοιχούν σε σημεία πιο απομακρυσμένα από το υποψήφιο σημείο, και τελικά επιλέγουμε το σημείο που αντιστοιχεί στο μεγαλύτερο σταθμισμένο D. Τέλος αν το D αυτό ξεπερνά κάποιο κατώφλι δεχόμαστε το σημείο ως σημείο πραγματικής αλλαγής αλλιώς το απορρίπτουμε ως false alarm. Καθώς το μέγεθος D εκφράζει πιθανότητα, λαμβάνει τιμές στο διάστημα [0,1]. Πειραματικά βρέθηκε ότι καλές τιμές του D κυμαίνονται γύρω στο 0.6.

Τελος, σημειώνουμε ότι κατά την υλοποίηση του αλγορίθμου για να έχουμε καλύτερες τιμές παραμέτρων, τροποποιήσαμε ελαφρώς τους τύπους που δίνουν τις παραμέτρους a και b, χρησιμοποιώντας λογαρίθμους, όπως φαίνεται παρακάτω:

$$a = \log(\frac{\mu^2}{\sigma^2} - 1)$$

$$b = \log(\frac{\sigma^2}{\mu^2})$$

Για τα σημεία που μας δίνει το δεύτερο πέρασμα είναι χρήσιμο να κάνουμε και ένα τελικό πέρασμα που θα εντοπίζει ζευγάρια σημείων αλλαγής που απέχουν μεταξύ τους

κάποια απόσταση μικρότερη από μία ελάχιστη απόσταση. Τα σημεία αυτά τότε θεωρείται ότι αντιστοιχούν στην ίδια πραγματική αλλαγή άρα κρατάμε μόνο το σημείο που έχει την μεγαλύτερη πιθανότητα αλλαγής και απορρίπτουμε το γειτονικό του. Με τον τρόπο αυτό μειώνουμε περισσότερο το false alarm. Στα πειράματά μας θεωρούμε ως ελάχιστη απόσταση δύο γειτονικών σημείων τα 0.5sec.

Στα πειράματα που έγιναν χρησιμοποιήσαμε ως κατώφλι πιθανότητας για το δεύτερο πέρασμα την τιμή 0.6 για όλα τα χαρακτηριστικά που παρουσιάστηκαν στην ενότητα εξαγωγής χαρακτηριστικών, εκτός από την απόλυτη τιμή της παραγώγου της fractal διάστασης, για την οποία χρησιμοποιήθηκε το κατώφλι 0.55.

Αλγόριθμος Δεύτερου Περάσματος με χρήση πολλών Μονοδιάστατων Χαρακτηριστικών

Μέχρι τώρα παρουσιάστηκε μέθοδος υπολογισμού της πιθανότητας αλλαγής με βάση ένα μόνο χαρακτηριστικό. Είναι χρήσιμο να μπορέσουμε να συνδυάσουμε περισσότερα του ενός χαρακτηριστικά κατά την απόφαση για το αν ένα σημείο αλλαγής είναι πραγματικό ή όχι ώστε να βελτιώσουμε την τελική απόδοση. Για το λόγο αυτό θα επεκτείνουμε την ιδέα που παρουσιάστηκε παραπάνω.

Η ιδέα είναι να επιλέξουμε ένα σύνολο χαρακτηριστικών και να υπολογίσουμε τα πιθανά σημεία αλλαγής ξεχωριστά για κάθε μονοδιάστατο χαρακτηριστικό. Έτσι για κάθε χαρακτηριστικό παίρνουμε ένα σύνολο σημείων αλλαγής. Είναι λογικό ότι αν πολλά από τα χαρακτηριστικά που χρησιμοποιήσαμε έχουν βρει σημείο αλλαγής μέσα σε μία γειτονιά τότε υπάρχει αυξημένη πιθανότητα να υπάρχει σε αυτή τη γειτονιά κάποιο σημείο αλλαγής ενώ αν λίγα ή κανένα χαρακτηριστικό βρήκαν σημείο αλλαγής σε μία γειτονιά τότε το πιθανότερο είναι ότι δεν θα υπάρχει εκεί κάποια αλλαγή. Η απόφαση αυτή είναι πιο σίγουρη καθώς βασίζεται σε περισσότερα του ενός χαρακτηριστικά αντί μόνο σε ένα.

Στην υλοποίηση μας, προβάλλουμε όλα τα σημεία αλλαγής που βρέθηκαν από όλα τα χαρακτηριστικά σε έναν άξονα και κάνουμε ομαδοποίηση (clustering) των σημείων αυτών. Όταν βρεθεί αυτή η ομαδοποίηση κρατάμε για κάθε κλάση το πλήθος των σημείων που την αποτελούν καθώς και το κέντρο της κλάσης, δηλαδή τη μέση τιμή των σημείων που την αποτελούν. Αν το πλήθος των σημείων ξεπερνά ή ισούται με κάποιο κατώφλι τότε επιλέγουμε το κέντρο της κλάσης ως πραγματικό σημείο αλλαγής αλλιώς το απορρίπτουμε. Το κατώφλι που θα επιλέξουμε εξαρτάται από το πόσο αυστηροί θα είμαστε στην απόφαση ότι ένα σημείο αλλαγής είναι πραγματικό. Για παράδειμα αν έχουμε χρησιμοποιήσει 5 χαρακτηριστικά τότε το πλήθος των σημείων ενός cluster θα κυμαίνεται λογικά από 1

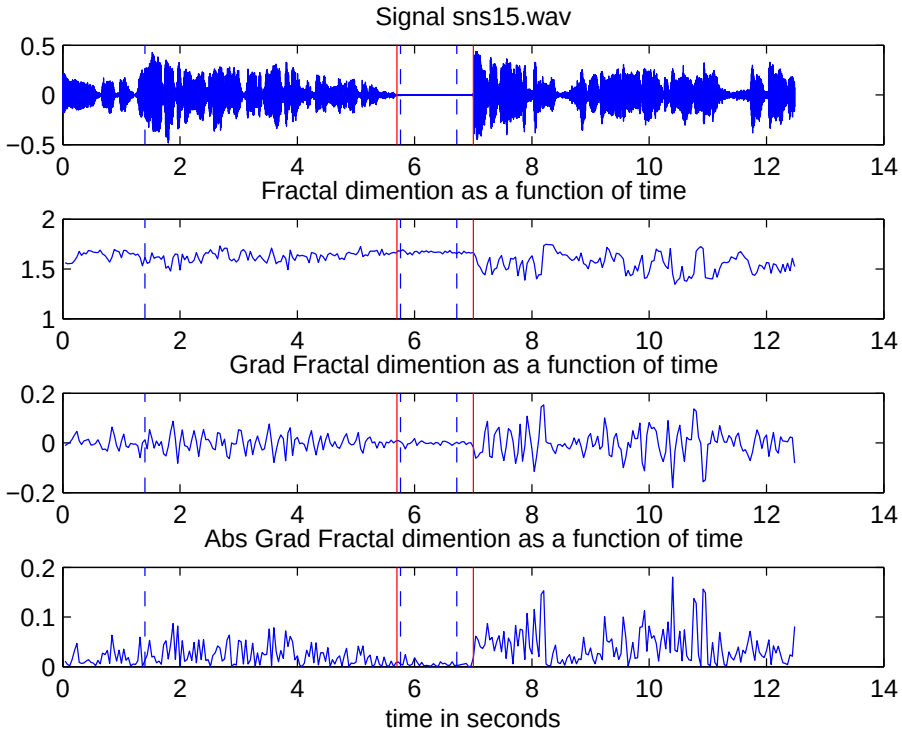
εώς 5 (σημειώνουμε εδώ ότι από τα σημεία που βρέθηκαν για κάθε χαρακτηριστικό έχουν συγχωνευθεί τα πολύ κοντινά ζεύγη άρα κάθε χαρακτηριστικό μπορεί να συνεισφέρει το πολύ ένα σημείο σε μία γειτονιά). Κατά συνέπεια πιθανές τιμές κατωφλιού είναι από 1 εώς 5, όπου το 5 οδηγεί σε μία πολύ αυστηρή απόφαση που απαιτεί όλα τα χαρακτηριστικά να έχουν βρει αλλαγή στη συγκεκριμένη γειτονιά ενώ το 1 απαιτεί τουλάχιστον 1 από τα χαρακτηριστικά να έχει βρει αλλαγή στη συγκεκριμένη γειτονιά.

Με τον τρόπο αυτό βρίσκουμε σημεία αλλαγής βασιζόμενοι στην πληροφορία περισσότερων του ενός χαρακτηριστικών. Στην υλοποίηση μας ως μέγιστη επιτρεπτή απόσταση για τα σημεία ενός cluster επιλέγουμε τα 0.7sec. Επίσης όπως θα φανεί στην ενότητα των πειραματικών αποτελεσμάτων, έχουν γίνει πειράματα για διάφορες τιμές κατωφλιού από 1 εως 3, ενώ χρησιμοποιούνται τα 5 χαρακτηριστικά που παρουσιάστηκαν στην ενότητα της εξαγωγής χαρακτηριστικών.

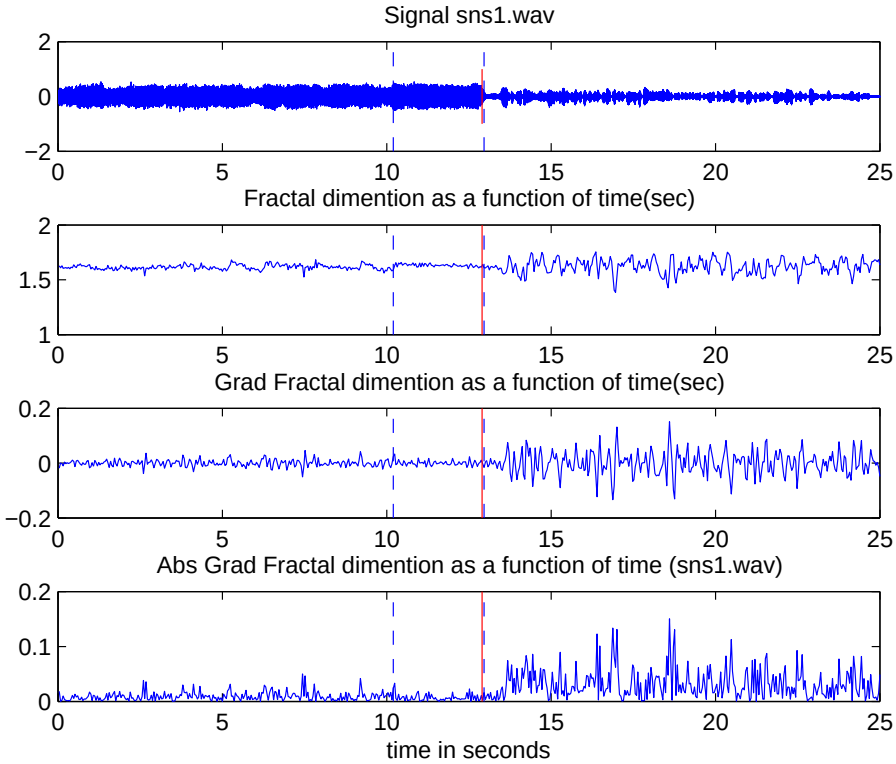
Ένα γραφικό παράδειγμα χρήσης του Δεύτερου Περάσματος

Ένα παράδειγμα φαίνεται στο σχήμα 4.12 όπου παρουσιάζονται το ηχητικό σήμα, η γραφική παράσταση της fractal dimension συναρτήσει του χρόνου, η παράγωγος της γραφικής παράστασης αυτής καθώς και η απόλυτη τιμή της παραγώγου. Στο παράδειγμα αυτό υπάρχουν 2 πραγματικές αλλαγές, από ομιλία σε σιωπή στο σημείο 5.7 sec και από σιωπή σε ομιλία στο σημείο 7.0 sec (δείγμα sns15.wav). Επίσης στα σχήματα σημειώνονται και οι τρεις πιθανές αλλαγές που βρέθηκαν στα σημεία 1.4 sec, 5.76 sec και 6.72 sec από τον αλγόριθμο πρώτου περάσματος χρησιμοποιώντας 13 MFCC averaged χαρακτηριστικά. Οι ιδιότητες της γραφικής παράστασης μπορούν να χρησιμοποιηθούν ώστε να επιβεβαιωθούν οι αλλαγές στα 5.76 sec και 6.72 sec και να απορριφθεί η αλλαγή στα 1.4 sec.

Ένα παράδειγμα φαίνεται στο σχήμα 4.13 όπου ομοίως παρουσιάζονται το ηχητικό σήμα, η γραφική παράσταση της fractal dimension συναρτήσει του χρόνου, η παράγωγος της γραφικής παράστασης αυτής καθώς και η απόλυτη τιμή της παραγώγου. Στο παράδειγμα αυτό υπάρχει 1 πραγματική αλλαγή, από θόρυβο σε ομιλία στο σημείο 12.9 sec. Επίσης στα σχήματα σημειώνονται και οι δύο πιθανές αλλαγές που βρέθηκαν στα σημεία 10.2 sec και 12.96 sec από τον αλγόριθμο πρώτου περάσματος χρησιμοποιώντας 13 MFCC averaged χαρακτηριστικά. Οι ιδιότητες της γραφικής παράστασης μπορούν να χρησιμοποιηθούν ώστε να επιβεβαιωθεί η αλλαγή στα 12.96 sec και να απορριφθεί η αλλαγή στα 10.2 sec.



Εικόνα 4.12: Το ηχητικό σήμα, Fractal dimension, Παράγωγος της Fractal Dimension και Απόλυτη τιμή της Παραγώγου για το δείγμα sns15.wav. Οι πραγματικές αλλαγές φαίνονται με κόκκινη γραμμή ενώ οι πιθανές αλλαγές (13 MFCC averaged)με διακεκομμένη μπλε γραμμή.



Εικόνα 4.13: Το ηχητικό σήμα, Fractal dimension, Παράγωγος της Fractal Dimension και Απόλυτη τιμή της Παραγώγου για το δείγμα sns1.wav. Η πραγματική αλλαγή φαίνεται με κόκκινη γραμμή ενώ οι πιθανές αλλαγές (13 MFCC averaged)με διακεκομμένη μπλε γραμμή.

4.5 Πειραματικά Αποτελέσματα

Στην ενότητα αυτή θα παρουσιαστεί μία επιλογή των πειραμάτων που έγιναν χρησιμοποιώντας τους αλγορίθμους πρώτου και δεύτερου περάσματος και κάποια από τα χαρακτηριστικά που αναλύθηκαν νωρίτερα

4.5.1 Διαχωρισμός των Αλλαγών σε Κατηγορίες

Τα audio-streams που εξετάζουμε προέρχονται από δελτία ειδήσεων στα οποία κυρίαρχο ρόλο παίζει η ομιλία ενώ , ανάλογα το δελτίο, υπάρχουν αρκετά τμήματα με θόρυβο ή σιωπή. Για το λόγο αυτό, η κατηγοριοποίηση των αλλαγών που θα επιχειρηθεί εδώ βασίζεται στο κατά πόσο και με ποιον τρόπο διαχωρίζονται δύο τμήματα ομιλίας σε ένα audio-stream. Θα θεωρηθούν δύο ομάδες δεδομένων, αυτά που περιέχουν ομιλία και είναι χρήσιμα για το σύστημα για το σύστημα που αναπτύσσουμε καθώς σε αυτά τα δεδομένα θα επιχειρηθεί αργότερα αναγνώριση φωνής, και αυτά που δεν περιέχουν ομιλία και πρέπει να αφαιρεθούν από τα δεδομένα καθώς δεν περιέχουν χρήσιμη πληροφορία. Λέγοντας μη ομιλία εννοούμε θόρυβο, σιωπή ή μουσική.

Η τρεις κατηγορίες αλλαγών που θεωρούμε φαίνονται παρακάτω:

Κατηγορία 1 Περιλαμβάνει αλλαγές μεταξύ ομιλίας και μη ομιλίας. Σε αυτή την κατηγορία το τμήμα μη ομιλίας, αν βρίσκεται ανάμεσα σε δύο τμήματα ομιλίας, διαρκεί πάνω από 2 sec (και συνήθως είναι θόρυβος), έτσι ώστε να μην μπορεί να θεωρηθεί παύση ή σύντομος θόρυβος μεταξύ εναλλαγής ομιλητών.

Κατηγορία 2 Περιλαμβάνει αλλαγές που είναι παύσεις μεταξύ ομιλητών,περιπτώσεις δηλαδή όπου ανάμεσα σε δύο τμήματα ομιλίας έχουμε τμήμα μη ομιλίας (συνήθως σιωπή ή λίγος θόρυβος) μήκους 0.7-2 sec.

Κατηγορία 3 Περιλαμβάνει εναλλαγές ομιλητών, όπου ανάμεσα σε 2 τμήματα ομιλίας έχουμε μία πολύ μικρή παύση, μικρότερη από 0.7 sec ή και καθόλου παύση. Επίσης η κατηγορία αυτή περιλαμβάνει και περιπτώσεις όπου κατά το χρονικό διάστημα μετάβασης οι δύο ομιλητές μιλούν ταυτόχρονα.

Τα πειραματικά αποτελέσματα μας δείχνουν ότι οι αλλαγές της κατηγορίας 1 είναι και οι πιο εύκολες στον εντοπισμό τους, οι αλλαγές της κατηγορίας 2 είναι δυσκολότερες να εντοπιστούν, ενώ οι πιο δύσκολες στον εντοπισμό τους είναι οι αλλαγές της κατηγορίας 3. Παρατηρούμε δηλαδή ότι το πρόβλημα του εντοπισμού αλλαγών ομιλητή είναι δυσκολότερο από το πρόβλημα εντοπισμού αλλαγών μεταξύ ομιλίας και μη ομιλίας και ότι όσο

μικρότερη είναι η παύση που παρεμβάλλεται (αν παρεμβάλλεται) μεταξύ δύο ομιλητών, τόσο δυσκολότερος γίνεται ο εντοπισμός της αλλαγής των ομιλητών αυτών.

Σημειώνουμε εδώ ότι τα δεδομένα έχουν είναι τμήματα από πραγματικά δελτία ειδήσεων που έχουν επιλεγεί επειδή περιέχουν κάποια αλλαγή και έχουν κατηγοριοποιηθεί ανάλογα με τον τύπο αλλαγής που περιέχουν. Το σημείο αλλαγής έχει σημειωθεί χειροκίνητα για κάθε ένα από αυτά τα τμήματα. Στα πειράματα εξετάζουμε την ικανότητα των αλγορίθμων πρώτου και δεύτερου περάσματος στον εντοπισμό των αλλαγών.

4.5.2 Μέτρα αξιολόγησης της απόδοσης

Τα μέτρα που χρησιμοποιούμε ορίζονται ως εξής:

$$Success = \frac{Number\ of\ change\ points\ that\ were\ correctly\ detected}{Number\ of\ real\ change\ points}$$

$$False\ Alarm = \frac{Number\ of\ change\ points\ that\ were\ wrongly\ detected}{Number\ of\ change\ points\ that\ were\ detected}$$

Η ακρίβεια εντοπισμού αλλαγής επιλέγεται ίση με 0.5sec, εκτός και αν αναφέρεται διαφορετικά. Δηλαδή θεωρούμε ότι μία αλλαγή έχει βρεθεί αν το σημείο που εντοπίστηκε βρίσκεται σε μία περιοχή μικρότερη ή ίση των 0.5sec γύρω από το πραγματικό σημείο αλλαγής.

4.5.3 Πειράματα σε αλλαγές κατηγορίας 1, για διάφορα περάσματα και για διάφορα χαρακτηριστικά

Σε αυτή τη σειρά πειραμάτων επικεντρωθήκαμε σε αλλαγές κατηγορίας 1.Επιλέχθηκαν 20 audio-streams μήκους από 10sec εώς 25sec τα οποία περιέχουν συνολικά 39 πραγματικές αλλαγές από ομιλία σε μη ομιλία ή αντίστροφα.Λέγοντας μη ομιλία εννοούμε θόρυβο, μουσική, σιωπή ή κάποιος συνδυασμός των προηγουμένων. Πάντως στις περισσότερες περιπτώσεις μη ομιλία είναι θόρυβος.

Δοκιμάστηκαν διάφορα σύνολα χαρακτηριστικών από τα οποία παρουσιάζουμε τα εξής:

set1 13 MFCC χαρακτηριστικά με χρήση μη επικαλυπτόμενων frames των 40msec στα οποία έχει εφαρμοστεί διαδικασία averaging όπως περιγράφεται στην αντίστοιχη ενότητα

set2 13 TECC χαρακτηριστικά με χρήση μη επικαλυπτόμενων frames των 40msec στα οποία έχει εφαρμοστεί διαδικασία averaging όπως περιγράφεται στην αντίστοιχη ενότητα

set3 13 delta MFCC χαρακτηριστικά, δηλαδή πρώτες παράγωγοι των χαρακτηριστικών του set1.

set4 13 delta TECC χαρακτηριστικά, δηλαδή πρώτες παράγωγοι των χαρακτηριστικών του set2.

set5 14 MFCC+ fractal dimension χαρακτηριστικά, από τα οποία έχουμε 13 MFCC στα οποία έχει εφαρμοστεί διαδικασία averaging και ένα χαρακτηριστικό fractal dimension. Όλα τα χαρακτηριστικά υπολογίστηκαν με χρήση μη επικαλυπτόμενων frames των 40msec

set6 14 delta MFCC+ delta fractal dimension χαρακτηριστικά, από τα οποία έχουμε 13 delta MFCC στα οποία έχει εφαρμοστεί διαδικασία averaging και ένα χαρακτηριστικό delta fractal dimension. Όλα τα χαρακτηριστικά υπολογίστηκαν με χρήση μη επικαλυπτόμενων frames των 40msec

Επίσης για τα σύνολα χαρακτηριστικών set1 εως set4 βρέθηκαν ποσοστά επιτυχίας για εφαρμογή ενός ή δύο περασμάτων εντοπισμού αλλαγών και με αυτόν τον τρόπο εξετάζεται η απόδοση του αλγορίθμου δεύτερου περάσματος. Για το δεύτερο πέρασμα χρησιμοποιούμε ένα μονοδιάστατο χαρακτηριστικό και συγκεκριμένα την απόλυτη τιμή της παραγώγου της Fractal διάστασης. Για τα set5 και set6 δεν έχει νόημα να χρησιμοποιήσουμε δεύτερο πέρασμα καθώς έχουμε ενσωματώσει τον συντελεστή fractal dimension στο σύνολο χαρακτηριστικών του πρώτου περάσματος.

Για το πρώτο πέρασμα επιλέγουμε εμπειρικά τα κατώφλια που μας δίνουν καλά πειραματικά αποτελέσματα ως προς τα ποσοστά Success και False Alarm. Τα κατώφλια για το πρώτο πέρασμα είναι:

BIC Κατώφλι=50

Tsquare Κατώφλι=1600

WMD Κατώφλι=1600

Απόδοση συστήματος με χρήση 13 MFCC averaged συντελεστών

Στον πίνακα 4.1 φαίνονται στατιστικά στοιχεία για την απόδοση του συστήματος για το set χαρακτηριστικών set1, για ένα ή δύο περάσματα. Για το δεύτερο πέρασμα γίνονται δύο πειράματα για διαφορετικές τιμές κατωφλιού. Το κατώφλι αυτό είναι το κατώφλι το οποίο πρέπει να ξεπερνά η πιθανότητα αλλαγής για να θεωρείται το υπο εξέταση σημείο ως σημείο πραγματικής αλλαγής (αναλυτικές λεπτομέρειες αναφέρονται στην ενότητα περιγραφής του αλγορίθμου δεύτερου περάσματος). Η ακρίβεια εύρεσης αλλαγής είναι range=0.5.

Πίνακας 4.1: Αποτελέσματα εύρεσης αλλαγών για το σύνολο χαρακτηριστικών set1: 13 MFCC averaged και range = 0.5

set1	1 pass	2 passes,thres=0.55	2 passes, thres=0.6
Success	89.74%	82.05%	79.48%
False Alarm	28.57%	15.79%	13.89%
Correct Changes Detected	35	32	31
False Changes Detected	14	6	5
Missed Real Changes	4	7	8
Total Changes Detected	49	38	36
Total Real Changes	39	39	39

Παρατηρούμε ότι ο αλγόριθμος πρώτου περάσματος οδηγεί σε υπερκατάτμηση του αρχικού audio-stream αλλά και σε αρκετά υψηλά ποσοστά επιτυχίας. Την ίδια στιγμή τα ποσοστά False alarm είναι επίσης αρκετά υψηλά. Για να αντιμετωπίσουμε την υπερκατάτμηση εφαρμόζουμε τον αλγόριθμο δεύτερου περάσματος.

Ο αλγόριθμος κατορθώνει να μειώσει σε σημαντικό βαθμό τα ποσοστά false alarm αν και την ίδια στιγμή ευθύνεται για μία μικρή αλλά αισθητή μείωση στα ποσοστά επιτυχίας. Όπως είναι αναμενόμενο, όσο πιο αυστηρό (υψηλό) είναι το κατώφλι το οποίο πρέπει να ξεπερνά η πιθανότητα αλλαγής για να θεωρείται το αντίστοιχο σημείο ως σημείο πραγματικής αλλαγής, τόσο πιο μεγάλη είναι η μείωση του false alarm αλλά και η μείωση του ποσοστού επιτυχίας.

Απόδοση συστήματος με χρήση 13 TECC averaged συντελεστών

Στον πίνακα 4.2 φαίνονται στατιστικά στοιχεία για την απόδοση του συστήματος για το set χαρακτηριστικών set2, για ένα ή δύο περάσματα. Η ακρίβεια εύρεσης αλλαγής είναι range=0.5.

Πίνακας 4.2: Αποτελέσματα εύρεσης αλλαγών για το σύνολο χαρακτηριστικών set2: 13 TECC averaged και range = 0.5

set2	1 pass	2 passes,thres=0.55	2 passes, thres=0.6
Success	89.74%	84.62%	82.05%
False Alarm	30.0%	15.38%	13.51%
Correct Changes Detected	35	33	32
False Changes Detected	15	6	5
Missed Real Changes	4	6	7
Total Changes Detected	50	39	37
Total Real Changes	39	39	39

Τα πειραματικά αποτελέσματα δείχνουν ότι τα 13 TECC averaged χαρακτηριστικά έχουν ελαφρώς καλύτερη απόδοση από τα 13 MFCC averaged χαρακτηριστικά, για τις διαφορετικές τιμές του range και κυρίως όταν χρησιμοποιείται και ο αλγόριθμος δεύτερου περάσματος. Γενικά, τα σχόλια που μπορούν να γίνουν είναι λίγο πολύ ίδια με αυτά της προηγούμενης ενότητας. Πρέπει όμως να τονιστεί ότι η χρήση TECC χαρακτηριστικών αυξάνει κατά πολύ το χρόνο επεξεργασίας του audio-stream, επειδή η εξαγωγή των χαρακτηριστικών αυτών είναι χρονοβόρα.

Απόδοση συστήματος με χρήση 13 delta MFCC averaged συντελεστών

Στον πίνακα 4.3 φαίνονται στατιστικά στοιχεία για την απόδοση του συστήματος για το set χαρακτηριστικών set3, για ένα ή δύο περάσματα.

Πίνακας 4.3: Αποτελέσματα εύρεσης αλλαγών για το σύνολο χαρακτηριστικών set3: 13 delta MFCC averaged και range = 0.5

set3	1 pass	2 passes,thres=0.55	2 passes, thres=0.6
Success	66.67%	64.10%	61.54%
False Alarm	10.34%	3.84%	0.0%
Correct Changes Detected	26	25	24
False Changes Detected	3	1	0
Missed Real Changes	13	14	15
Total Changes Detected	29	26	24
Total Real Changes	39	39	39

Είναι εμφανές ότι η χρήση πρώτων παραγώγων οδηγεί σε υποκατάτμηση του audio-stream, δημιουργώντας ιδιαίτερα χαμηλά ποσοστά false alarm αλλά και πολύ χαμηλότερα ποσοστά επιτυχίας από αυτά των set1 και set2. Κατά συνέπεια, τα χαρακτηριστικά αυτά είναι μάλλον ακατάλληλα για την εφαρμογή του αλγορίθμου δεύτερου περάσματος, ο οποίος θεωρεί ότι δέχεται ένα υπερκατατμημένο audio-stream και στοχεύει στο να μειώσει το false alarm. Όπως είναι αναμενόμενο, η εφαρμογή του αλγορίθμου δεύτερου περάσματος επιδεινώνει τα αποτελέσματα, όσον αφορά το ποσοστό επιτυχίας, αν και το false alarm πρακτικά μηδενίζεται.

Απόδοση συστήματος με χρήση 13 delta TECC averaged συντελεστών

Στον πίνακα 4.4 φαίνονται στατιστικά στοιχεία για την απόδοση του συστήματος για το set χαρακτηριστικών set4, για ένα ή δύο περάσματα.

Πίνακας 4.4: Αποτελέσματα εύρεσης αλλαγών για το σύνολο χαρακτηριστικών set4: 13 delta TECC averaged και range = 0.5

set4	1 pass	2 passes,thres=0.55	2 passes, thres=0.6
Success	69.23%	61.54%	61.54%
False Alarm	10.0%	4.0%	4.0%
Correct Changes Detected	27	24	24
False Changes Detected	3	1	1
Missed Real Changes	12	15	15
Total Changes Detected	30	25	25
Total Real Changes	39	39	39

Και εδώ η χρήση πρώτων παραγώγων οδηγεί σε υποκατάτμηση του audio-stream, δημιουργώντας ιδιαίτερα χαμηλά ποσοστά false alarm αλλά και πολύ χαμηλότερα ποσοστά επιτυχίας από αυτά των set1 και set2. Άρα τα σχόλια που μπορούν να γίνουν είναι παρόμοια με αυτά της προηγούμενης ενότητας.

Απόδοση συστήματος με χρήση 13 MFCC averaged και 1 fractal dimension συντελεστών

Στην ενότητα αυτή δοκιμάζεται η απόδοση του συστήματος αν δοκιμάσουμε να ενσωματώσουμε τον συντελεστή fractal dimension στο πρώτο πέρασμα μαζί με τους 13 MFCC averaged συντελεστές, και δεν χρησιμοποιήσουμε τον αλγόριθμο δεύτερου περάσματος. Έτσι στον πίνακα 4.5 φαίνονται στατιστικά στοιχεία για την απόδοση του συστήματος για το set χαρακτηριστικών set5.

Τα αποτελέσματα που παίρνουμε είναι συγκρίσιμα με τα αποτελέσματα των set1(13 MFCC averaged) και set2(13 TECC averaged). Παρατηρούμε ότι τα αποτελέσματα είναι γενικά χειρότερα τόσο σχετικά με το ποσοστό επιτυχίας όσο και με το ποσοστό false alarm, σε σχέση με τα αποτελέσματα που παίρνουμε για ένα πέρασμα και για τις αντίστοιχες τιμές του range για τα set χαρακτηριστικών set1 και set2. Αυτό μας υποδεικνύει ότι θα ήταν καλύτερο να χρησιμοποιήσουμε την πληροφορία που μας προσφέρουν τα χαρακτη-

Πίνακας 4.5: Αποτελέσματα εύρεσης αλλαγών για το σύνολο χαρακτηριστικών set5: 13 MFCC averaged + 1 fractal dimension και για ένα πέρασμα

set5	range = 0.2 sec	range = 0.3 sec	range = 0.5 sec
Success	71.79%	84.62%	87.18%
False Alarm	44.0%	34.0%	32.0%
Correct Changes Detected	28	33	34
False Changes Detected	22	17	16
Missed Real Changes	11	6	5
Total Changes Detected	50	50	50
Total Real Changes	39	39	39

ριστικά fractal dimension με κάποιον άλλο τρόπο, όπως για παράδειγμα χρησιμοποιώντας τον αλγόριθμο δεύτερου περάσματος που περιγράφηκε παραπάνω.

Απόδοση συστήματος με χρήση 13 delta MFCC averaged και 1 delta fractal dimension συντελεστών

Στην ενότητα αυτή δοκιμάζουμε να χρησιμοποιήσουμε τις παραγώγους των συντελεστών MFCC και fractal dimension(set6). Έτσι στον πίνακα 4.6 φαίνονται στατιστικά στοιχεία για την απόδοση του συστήματος για το set χαρακτηριστικών set6 για ένα πέρασμα.

Πίνακας 4.6: Αποτελέσματα εύρεσης αλλαγών για το σύνολο χαρακτηριστικών set6: 13 delta MFCC averaged + 1 delta fractal dimension και για ένα πέρασμα

set6	range = 0.2 sec	range = 0.3 sec	range = 0.5 sec
Success	33.34%	51.28%	61.54%
False Alarm	50.0%	23.07%	7.69%
Correct Changes Detected	13	20	24
False Changes Detected	13	6	2
Missed Real Changes	26	19	15
Total Changes Detected	26	26	26
Total Real Changes	39	39	39

Τα αποτελέσματα που παίρνουμε είναι συγκρίσιμα με τα αποτελέσματα των set3(13 delta MFCC averaged) και set4(13 delta TECC averaged). Παρατηρούμε όμως ότι τα αποτελέσματα είναι γενικά χειρότερα τόσο σχετικά με το ποσοστό επιτυχίας όσο και με το ποσοστό false alarm, σε σχέση με τα αποτελέσματα που παίρνουμε για ένα πέρασμα και για τις αντίστοιχες τιμές του range για τα set χαρακτηριστικών set3 και set4.

Συμπεράσματα

Στην αναφορά αυτή εξετάστηκαν λεπτομερώς διάφορα set χαρακτηριστικών και η λειτουργία του συστήματος για ένα ή δύο περάσματα εντοπισμού αλλαγών. Τα βασικότερα συμπεράσματα συνοψίζονται παρακάτω:

1. Επιτυγχάνουμε την καλύτερη απόδοση του συστήματος αν χρησιμοποιήσουμε 13 TECC averaged συντελεστές και αλγόριθμο δεύτερου περάσματος. Πράγματι οι παραπάνω συνθήκες έδωσαν αποτελέσματα $Success = 84.62\%$ και $FalseAlarm = 15.38\%$, με $range = 0.5$ και $threshold = 0.55$. Η χρήση 13 MFCC averaged συντελεστών και αλγορίθμου δεύτερου περάσματος δίνει ελαφρώς χειρότερα αποτελέσματα ($Success = 82.05\%$ και $FalseAlarm = 15.79\%$, με $range = 0.5$ και $threshold = 0.55$) αλλά είναι υπολογιστικά πολύ πιο αποδοτική, καθώς μειώνει το συνολικό χρόνο επεξεργασίας στο 1/3 σε σχέση με τα TECC χαρακτηριστικά.

2. Ο αλγόριθμος δεύτερου περάσματος που είναι μία καινούρια ιδέα λειτουργεί ικανοποιητικά για αλλαγές κατηγορίας 1 και επιτυγχάνει μείωση του false alarm. Πράγματι, σε όλα τα πειράματα που διεξήχθησαν ο αλγόριθμος δεύτερου περάσματος κατάφερε να μειώσει σημαντικά το false alarm που είχε προκύψει από το πρώτο πέρασμα, αν και ευθύνεται για μία (περισσότερο ή λιγότερο σημαντική ανάλογα με το set χαρακτηριστικών) μείωση του ποσοστού επιτυχίας. Επιπλέον, ο αλγόριθμος εντοπισμού αλλαγών δεύτερου περάσματος είναι υπολογιστικά πολύ γρήγορος και δεν δημιουργεί αισθητές καθυστερήσεις.

4.5.4 Πειράματα σε αλλαγές κάθε κατηγορίας, για διάφορα περάσματα

Σε αυτή τη σειρά πειραμάτων εξετάζεται η απόδοση του δεύτερου περάσματος για όλες τις κατηγορίες αλλαγών και για κάθε ένα από τα χαρακτηριστικά που μελετήθηκαν καθώς και για τον συνδυασμό τους. Τα πειράματα έδειξαν ότι οι απόλυτες τιμές των παραγώγων των χαρακτηριστικών δίνουν τις περισσότερες φορές καλύτερα αποτελέσματα από τα ίδια τα χαρακτηριστικά έτσι θα παρουσιαστούν μόνο τα αποτελέσματα για τις απόλυτες τιμές των

παραγώγων των χαρακτηριστικών.

Το πρώτο πέρασμα χρησιμοποιεί 13 MFCC averaged χαρακτηριστικά.

Τα κατώφλια για το πρώτο πέρασμα φαίνονται παρακάτω. Παρατηρούμε ότι έχουμε μειώσει το κατώφλι του BIC σε σχέση με την προηγούμενη σειρά πειραμάτων, καθώς με αυτές τις τιμές παρατηρούμε καλύτερα αποτελέσματα.

BIC Κατώφλι=50

Tsquare Κατώφλι=1600

WMD Κατώφλι=1600

Η ακρίβεια εύρεσης μίας αλλαγής είναι 0.5 sec πριν και μετά το πραγματικό σημείο αλλαγής.

Στους πίνακες που θα παρουσιαστούν στη συνέχεια χρησιμοποιούνται οι παρακάτω συντομογραφίες. Με τον όρο γειτονικά ζεύγη σημείων εννοούμε σημεία που απέχουν μεταξύ τους λιγότερο από 0.5sec

pass1 Εφαρμογή μόνο του πρώτου περάσματος (BIC) και τελικά συγχώνευση των γειτονικών ζευγών σημείων.

fractal Εφαρμογή μετά το πρώτο πέρασμα και του δεύτερου περάσματος με χαρακτηριστικό την απόλυτη τιμή τη παραγώγου της fractal διάστασης και με τιμή κατωφλιού πιθανότητας 0.55. Τέλος γίνεται συγχώνευση των γειτονικών ζευγών σημείων.

rms Εφαρμογή μετά το πρώτο πέρασμα και του δεύτερου περάσματος με χαρακτηριστικό την απόλυτη τιμή τη παραγώγου της rms τιμής και με τιμή κατωφλιού πιθανότητας 0.6. Τέλος γίνεται συγχώνευση των γειτονικών ζευγών σημείων.

env Εφαρμογή μετά το πρώτο πέρασμα και του δεύτερου περάσματος με χαρακτηριστικό την απόλυτη τιμή τη παραγώγου της envelope τιμής και με τιμή κατωφλιού πιθανότητας 0.6. Τέλος γίνεται συγχώνευση των γειτονικών ζευγών σημείων.

p.l.d. Εφαρμογή μετά το πρώτο πέρασμα και του δεύτερου περάσματος με χαρακτηριστικό την απόλυτη τιμή τη παραγώγου της previous local difference τιμής και με τιμή κατωφλιού πιθανότητας 0.6. Τέλος γίνεται συγχώνευση των γειτονικών ζευγών σημείων.

n.l.d. Εφαρμογή μετά το πρώτο πέρασμα και του δεύτερου περάσματος με χαρακτηριστικό την απόλυτη τιμή τη παραγώγου της next local difference τιμής και με τιμή κατωφλιού πιθανότητας 0.6. Τέλος γίνεται συγχώνευση των γειτονικών ζευγών σημείων.

Κατηγορία 1

Για τα αποτελέσματα που θα παρουσιαστούν εδώ χρησιμοποιήθηκαν δεδομένα που ανήκουν στην κατηγορία 1 και προέρχονται από δελτίο ειδήσεων της NET (181006.wav)

Στον πίνακα 4.7 φαίνονται τα αποτελέσματα του δεύτερου περάσματος για κάθε ένα απο τα πιθανά χαρακτηριστικά σε σύγκριση με τη χρήση ενός περάσματος μόνο (πρώτη στηλη). Παρατηρούμε ότι όλα τα χαρακτηριστικά αποδίδουν πολύ καλά και φαίνεται ότι τα envelope, previous local difference και next local difference πιθανόν υπερτερούν των fractal και rms. Τέλος, παρατηρούμε ένα αρκετά υψηλό ποσοστό false alarm. Το ποσοστό αυτό θα μπορούσε να μειωθεί αν αυξάναμε τα κατώφλια πιθανότητας (οι τιμές των οποίων έχουν ήδη αναφερθεί) αλλά έτσι θα μειωνόταν και το ποσοστό επιτυχίας.

Πίνακας 4.7: Αποτελέσματα εύρεσης αλλαγών κατηγορίας 1 για τα διάφορα χαρακτηριστικά που μπορούν να χρησιμοποιηθούν στον αλγόριθμο δεύτερου περάσματος

	pass1	fractal	rms	env	p.l.d.	n.l.d.
Success	97.44%	89.74%	89.74%	94.87%	94.87%	94.87%
False Alarm	32.14%	20.45%	20.45%	19.56%	19.56%	19.56%
Correct Changes Detected	38	35	35	37	37	37
False Changes Detected	18	9	9	9	9	9
Missed Real Changes	1	4	4	2	2	2
Total Changes Detected	56	44	44	46	46	46
Total Real Changes	39	39	39	39	39	39

Στον πίνακα 4.8 φαίνονται τα αποτελέσματα του δεύτερου περάσματος για την περίπτωση που χρησιμοποιούμε ταυτόχρονα πληροφορία και από τα 5 χαρακτηριστικά, όπως εξηγήθηκε στην επέκταση του δεύτερου περάσματος νωρίτερα, σε σύγκριση με τη χρήση ενός περάσματος μόνο (πρώτη στήλη). Δοκιμάζονται διάφορα κατώφλια για το δεύτερο πέρασμα.

Παρατηρούμε από τον πίνακα 4.8 ότι το αποτέλεσμα της επέκτασης του αλγορίθμου για κατώφλια 2 ή 3 είναι πολύ καλό, και ίσο με το καλύτερο δυνατό αποτέλεσμα του πίνακα 4.7. Δηλαδή η επέκταση αποδίδει πολύ καλά αν απαιτήσουμε τουλάχιστον 2 ή 3 από τα 5

Πίνακας 4.8: Αποτελέσματα εύρεσης αλλαγών κατηγορίας 1 για την επέκταση του αλγορίθμου δεύτερου περάσματος και για διάφορα πιθανά κατώφλια

	pass1	thres=3	thres=2	thres=1
Success	97.43%	94.87%	94.87%	97.43%
False Alarm	32.14%	19.56%	19.56%	25.49%
Correct Changes Detected	38	37	37	38
False Changes Detected	18	9	9	13
Missed Real Changes	1	2	2	1
Total Changes Detected	56	46	46	51
Total Real Changes	39	39	39	39

χαρακτηριστικά να βρίσκουν μία αλλαγή για να τη θεωρήσουμε πραγματική. Αν μειώσουμε το κατώφλι σε ένα αρκούμαστε σε τουλάχιστον 1 από τα 5 χαρακτηριστικά, γεγονός που αυξάνει το ποσοστό false alarm αλλά την ίδια στιγμή αυξάνει και το ποσοστό επιτυχίας στην τιμή που είχε πριν το δεύτερο πέρασμα.

Κατηγορία 2

Για τα αποτελέσματα που θα παρουσιαστούν εδώ χρησιμοποιήθηκαν δεδομένα που ανήκουν στην κατηγορία 2 και προέρχονται από δελτίο ειδήσεων του MEGA (171006.wav)

Στον πίνακα 4.9 φαίνονται τα αποτελέσματα του δεύτερου περάσματος για κάθε ένα από τα πιθανά χαρακτηριστικά σε σύγκριση με τη χρήση ενός περάσματος μόνο (πρώτη στήλη). Παρατηρούμε ότι τα περισσότερα χαρακτηριστικά αποδίδουν αρκετά καλά και φαίνεται ότι τα rms και previous local difference πιθανόν υπερτερούν αν και το πλήθος των σημείων αλλαγής δεν είναι τέτοιο ώστε να μπορούμε να βγάλουμε ασφαλή συμπεράσματα σχετικά με τη διαφοροποίηση της απόδοσης που προκαλεί το κάθε χαρακτηριστικό.

Στον πίνακα 4.10 φαίνονται τα αποτελέσματα του δεύτερου περάσματος για την περίπτωση που χρησιμοποιούμε ταυτόχρονα πληροφορία και από τα 5 χαρακτηριστικά, όπως εξηγήθηκε στην επέκταση του δεύτερου περάσματος νωρίτερα, σε σύγκριση με τη χρήση ενός περάσματος μόνο (πρώτη στήλη). Δοκιμάζονται διάφορα κατώφλια για το δεύτερο πέρασμα. Παρατηρούμε ότι για αυτή την κατηγορία αλλαγών, που είναι δυσκολότερες στον εντοπισμό τους από τις αλλαγές της κατηγορίας 1, η επέκταση του δεύτερου περάσματος λειτουργεί καλά για μικρότερα κατώφλια. Αυτό συμβαίνει επειδή πολλές αλλαγές γίνονται αντιληπτές από μόνο ένα ή δύο από τα χρησιμοποιούμενα χαρακτηριστικά και επίσης επειδή

Πίνακας 4.9: Αποτελέσματα εύρεσης αλλαγών κατηγορίας 2 για τα διάφορα χαρακτηριστικά που μπορούν να χρησιμοποιηθούν στον αλγόριθμο δεύτερου περάσματος

	pass1	fractal	rms	env	p.l.d.	n.l.d.
Success	91.17%	55.88%	76.47%	64.70%	73.52%	67.64%
False Alarm	18.42%	13.63%	3.70%	8.33%	0	8%
Correct Changes Detected	31	19	26	22	25	23
False Changes Detected	7	3	1	2	0	2
Missed Real Changes	3	15	8	12	9	11
Total Changes Detected	38	22	27	24	25	25
Total Real Changes	34	34	34	34	34	34

διαφορετικά χαρακτηριστικά αντιλαμβάνονται πιθανώς και διαφορετικές αλλαγές.

Πίνακας 4.10: Αποτελέσματα εύρεσης αλλαγών κατηγορίας 2 για την επέκταση του αλγορίθμου δεύτερου περάσματος και για διάφορα πιθανά κατώφλια

	pass1	thres=3	thres=2	thres=1
Success	91.17%	70.58%	73.52%	76.47%
False Alarm	18.42%	7.69%	7.4%	7.14%
Correct Changes Detected	31	24	25	26
False Changes Detected	7	2	2	2
Missed Real Changes	3	10	9	8
Total Changes Detected	38	26	27	28
Total Real Changes	34	34	34	34

Κατηγορία 3

Για τα αποτελέσματα που θα παρουσιαστούν εδώ χρησιμοποιήθηκαν δεδομένα που ανήκουν στην κατηγορία 3 και προέρχονται από δελτίο ειδήσεων της NET (181006.wav)

Στον πίνακα 4.11 φαίνονται τα αποτελέσματα του δεύτερου περάσματος για κάθε ένα από τα πιθανά χαρακτηριστικά σε σύγκριση με τη χρήση ενός περάσματος μόνο (πρώτη στήλη). Αυτή η κατηγορία αλλαγών είναι η πιο δύσκολη στον εντοπισμό της και παρατηρούμε ότι

τα χαρακτηριστικά δεν αποδίδουν ικανοποιητικά. Το μόνο χαρακτηριστικό που φαίνεται να αποδίδει ελαφρώς καλύτερα είναι η rms τιμή.

Στον πίνακα 4.12 φαίνονται τα αποτελέσματα του δεύτερου περάσματος για την περίπτωση που χρησιμοποιούμε ταυτόχρονα πληροφορία και από τα 5 χαρακτηριστικά, όπως εξηγήθηκε στην επέκταση του δεύτερου περάσματος νωρίτερα, σε σύγκριση με τη χρήση ενός περάσματος μόνο (πρώτη στήλη). Δοκιμάζονται διάφορα κατώφλια για το δεύτερο πέρασμα. Παρατηρούμε ότι για αυτή την κατηγορία αλλαγών, που είναι οι πιο δύσκολες στον εντοπισμό τους, η επέκταση του δεύτερου περάσματος οδηγεί σε σημαντική βελτίωση της απόδοσης ειδικά για κατώφλι ίσο με 1, σε σχέση με τα αποτελέσματα του πίνακα 4.11. Αυτό συμβαίνει επειδή πολλές αλλαγές γίνονται αντιληπτές από μόνο ένα από τα χρησιμοποιούμενα χαρακτηριστικά και επίσης επειδή διαφορετικά χαρακτηριστικά αντιλαμβάνονται πιθανώς και διαφορετικές αλλαγές. Έτσι ο συνδυασμός της πληροφορίας πολλών χαρακτηριστικών οδηγεί σε αποτελέσματα αισθητά καλύτερα από αυτά που δίνει το κάθε χαρακτηριστικό ξεχωριστά.

Πίνακας 4.11: Αποτελέσματα εύρεσης αλλαγών κατηγορίας 3 για τα διάφορα χαρακτηριστικά που μπορούν να χρησιμοποιηθούν στον αλγόριθμο δεύτερου περάσματος

	pass1	fractal	rms	env	p.l.d.	n.l.d.
Success	72.34%	42.55%	55.31%	31.91%	34.04%	38.29%
False Alarm	37%	28.57%	25.71%	11.76%	27.27%	28%
Correct Changes Detected	34	20	26	15	16	18
False Changes Detected	20	8	9	2	6	7
Missed Real Changes	13	27	21	32	31	29
Total Changes Detected	54	28	35	17	22	25
Total Real Changes	47	47	47	47	47	47

Συμπεράσματα

Εξετάσαμε αναλυτικά την απόδοση του αλγορίθμου δεύτερου περάσματος για τις διάφορες κατηγορίες αλλαγών και για τα διάφορα χαρακτηριστικά που θα μπορούσαν να χρησιμοποιηθούν. Δείξαμε ότι χαρακτηριστικά όπως η απόλυτη τιμή της παραγώγου της envelope τιμής ενός σήματος ή η απόλυτη τιμή της παραγωγου της rms τιμής ενός σήματος καθώς και άλλα χαρακτηριστικά μπορούν να αποδόσουν το ίδιο καλά ή και καλύτερα από την απόλυτη τιμή της παραγώγου της fractal διάστασης ενός σήματος. Επίσης δείξαμε ότι ο

Πίνακας 4.12: Αποτελέσματα εύρεσης αλλαγών κατηγορίας 3 για την επέκταση του αλγορίθμου δεύτερου περάσματος και για διάφορα πιθανά κατώφλια

	pass1	thres=3	thres=2	thres=1
Success	72.34%	38.29%	55.31%	65.95%
False Alarm	37.03%	25%	21.21%	27.9%
Correct Changes Detected	34	18	26	31
False Changes Detected	20	6	7	12
Missed Real Changes	13	29	21	16
Total Changes Detected	54	24	33	43
Total Real Changes	47	47	47	47

συνδυασμός των χαρακτηριστικών αυτών στο δεύτερο πέρασμα λειτουργεί γενικά καλύτερα από τη μεμονωμένη χρήση τους για την εύρεση αλλαγών.

Η κυριότερη βελτίωση που προσφέρει η επέκταση του δεύτερου περάσματος για πολλά χαρακτηριστικά είναι στην εύρεση δύσκολων αλλαγών, δηλαδή αλλαγών ομιλητή όπου δεν υπάρχει παύση μεταξύ των δύο ομιλητών (αλλαγές κατηγορίας 3). Στην περίπτωση αυτή για κατώφλι=1 παρατηρούμε αποτελέσματα απόδοσης αισθητά καλύτερα από τα αποτελέσματα απόδοσης του κάθε χαρακτηριστικού ξεχωριστά. Πάντως παρατηρείται μια γενική βελτίωση της απόδοσης για όλες τις κατηγορίες αλλαγών.

Σχετικά με το κατώφλι που χρησιμοποιεί η επέκταση του δεύτερου περάσματος, παρατηρούμε ότι για αλλαγές κατηγορίας 1 ένα κατώφλι 2 ή 3 λειτουργεί καλά, ενώ μικρότερο κατώφλι οδηγεί σε υψηλότερο false alarm. Αντίθετα, για τις αλλαγές κατηγορίας 2 και 3 είναι καλύτερα να χρησιμοποιήσουμε χαμηλότερο κατώφλι, δηλαδή 1.

Κατά συνέπεια, για τη χρήση της επέκτασης του αλγορίθμου σε ένα audio-stream που γενικά περιέχει αλλαγές και των 3 κατηγοριών χωρίς να ξέρουμε εκ των προτέρων τι είδους αλλαγή είναι η αλλαγή που εντοπίζεται στο πρώτο πέρασμα, πρέπει να κάνουμε έναν συμβιβασμό σχετικά με την τιμή κατωφλιού που θα επιλέξουμε για το δεύτερο πέρασμα. Ο συμβιβασμός αυτός εξαρτάται από το πόσο υψηλό false alarm μπορούμε να ανεχθούμε και από το τι ποσοστό επιτυχίας απαιτούμε στην εύρεση δύσκολων αλλαγών.

Τέλος σημειώνουμε ότι επειδή το δεύτερο πέρασμα βασίζεται σε χαρακτηριστικά, όπως η Fractal διάσταση, η RMS τιμή, η τιμή μέγιστου πλάτους, των οποίων οι τιμές διαφοροποιούνται αισθητά μεταξύ ομιλίας και μη ομιλίας αλλά δεν διαφέρουν τόσο πολύ μεταξύ ομιλίας διαφορετικών ομιλητών το δεύτερο πέρασμα τείνει να απορρίπτει είτε μη πραγματικές αλ-

λαγές είτε αλλαγές μεταξύ ομιλητών, ενώ κρατάει τη μεγάλη πλειοψηφία των αλλαγών μεταξύ φωνής και μη φωνής. Κατά συνέπεια, εξετάζοντας αν μία αλλαγή έχει περάσει το κατώφλι πιθανότητας του δεύτερου περάσματος μπορούμε να βγάλουμε κάποια πιθανοτικά συμπεράσματα σχετικά με τον τύπο της αλλαγής αυτής.

4.6 Συμπεράσματα από τη Μελέτη του Προβλήματος Κατάτμησης

Στην ενότητα αυτή επιχειρήθηκε μία ανάλυση του προβλήματος της κατάτμησης audio-streams με έμφαση στη μελέτη χαρακτηριστικών, στη μελέτη μετρικών κριτηρίων κατάτμησης (metric-based) και στην υλοποίηση αλγορίθμων κατάτμησης.

Όσον αφορά τη μελέτη χαρακτηριστικών, παρουσιάστηκαν πλήθος μονοδιάστατων και πολυδιάστατων χαρακτηριστικών που χρησιμοποιούνται στη βιβλιογραφία. Στη συνέχεια παρουσιάστηκαν διάφορα μετρικά κριτήρια για τον εντοπισμό αλλαγών σε audio-streams και παρουσιάστηκαν γραφικά τα αποτελέσματα των κριτηρίων αυτών για κάποιες περιπτώσεις αλλαγών.

Ακολούθως παρουσιάστηκαν με λεπτομέρεια οι 2 αλγόριθμοι εντοπισμού αλλαγών που υλοποιήθηκαν και χρησιμοποιήθηκαν στα πειράματα. Ο αλγόριθμος πρώτου περάσματος χρησιμοποιεί το κριτήριο BIC ή προσεγγίσεις του και προτάθηκε στα [19] και [8]. Ο αλγόριθμος αυτός εντοπίζει πιθανά σημεία αλλαγής.

Ο αλγόριθμος δεύτερου περάσματος είναι μία καινούρια ιδέα που βασίζεται στον υπολογισμό πιθανότητας αλλαγής σε ένα σημείο, που προτείνεται στο [38]. Ο αλγόριθμος αυτός αποφασίζει αν ένα σημείο αλλαγής που βρέθηκε από το πρώτο πέρασμα αντιστοιχεί σε πραγματικό σημείο αλλαγής. Σκοπός του αλγορίθμου είναι να μειωθεί το υψηλό ποσοστό false alarm που συνήθως παίρνουμε από το πρώτο πέρασμα αλλά και να χρησιμοποιηθούν στη διαδικασία εντοπισμού αλλαγών μονοδιάστατα χαρακτηριστικά όπως η Fractal διάσταση κα η RMS τιμή, που είναι δύσκολο να συνδυαστούν με τα πολυδιάστατα χαρακτηριστικά του πρώτου περάσματος.

Έγιναν πειράματα που μελετούν τη συμπεριφορά των διάφορων χαρακτηριστικών που μπορούν να χρησιμοποιηθούν τόσο στο πρώτο όσο και στο δεύτερο πέρασμα. Επίσης, τα πειράματα εξετάζουν κατά πόσο το δεύτερο πέρασμα βελτιώνει την απόδοση του πρώτου περάσματος. Τα αποτελέσματα που εκτελέστηκαν υποδεικνύουν ότι το δεύτερο πέρασμα είτε με ένα είτε με πολλά χαρακτηριστικά μπορεί να οδηγήσει σε βελτίωση των ποσοστών Success και False Alarm

Συγκεκριμένα, το πρώτο πέρασμα, με χρήση 13 MFCC, λειτουργεί πολύ καλά για τον εντοπισμό αλλαγών στις κατηγορίες 1 και 2 (ποσοστά success περίπου 95% και 90% αντίστοιχα) ενώ λειτουργεί ικανοποιητικά για την κατηγορία 3 (ποσοστό success περίπου 70%). Εντούτοις, δημιουργεί υψηλά ποσοστά false alarm της τάξης του 30%-40% σε κάθε κατηγορία.

Το δεύτερο πέρασμα με χρήση ενος χαρακτηριστικού (ενός από αυτά που παρουσιάστηκαν στην ενότητα των χαρακτηριστικών για το δεύτερο πέρασμα) λειτουργεί πολύ καλά για την κατηγορία 1, επιτυγχάνοντας μείωση του false alarm περίπου σε 15%-20%, με μικρή μείωση του ποσοστού success (90%). Αντίθετα για τις υπόλοιπες κατηγορίες και ειδικά για την κατηγορία 3, που είναι και η δυσκολότερη, το δεύτερο πέρασμα δεν λειτουργεί καλά, μειώνοντας πάρα πολύ το ποσοστά success.

Το δεύτερο πέρασμα με χρήση πολλών χαρακτηριστικών επιτυγχάνει βελτίωση των αποτελεσμάτων για τις κατηγορίες 2 και 3, μειώνοντας το false alarm και κρατώντας το ποσοστό success σε ικανοποιητικό επίπεδο. Συγκεκριμένα, για χαμηλό κατώφλι, το δεύτερο πέρασμα επιτυγχάνει για την κατηγορία 2 success 75% και false alarm 8% περίπου και για την κατηγορία 3 success 60% και false alarm 25% περίπου.

Σχετικά με την ταχύτητα, η επεξεργασία ενός δελτίου μεγέθους περίπου 1 ώρας και 15' απαιτεί λίγο λιγότερο από 2 ώρες και 30', με χρήση του συνολικού συστήματος δηλαδή πρώτου και δεύτερου περάσματος(με πολλά χαρακτηριστικά). Το μεγαλύτερο ποσοστό του χρόνου επεξεργασίας φαίνεται να καταναλώνεται από το πρώτο πέρασμα, λόγω της χρήσης του κριτηρίου BIC.

Συμπερασματικά, θα μπορούσαμε να πούμε ότι το σύστημα είναι ικανό να εντοπίσει με επιτυχία αλλαγές μεταξύ ομιλίας και μη ομιλίας και να εντοπίσει σχετικά ικανοποιητικά αλλαγές ομιλητή. Ο χρήση των διαφόρων εναλλακτικών επιλογών για τα κατώφλια και για τα διάφορα περάσματα εξαρτάται από τις απαιτήσεις που έχουμε ως προς το ποσοστά success και false alarm αλλά και από το πώς οι αλλαγές που βρέθηκαν θα συνδυαστούν με τα αποτελέσματα του συστήματος με GMM μοντέλα που θα παρουσιαστεί στο επόμενο κεφάλαιο.

Κεφάλαιο 5

Κατάτμηση και Κατηγοριοποίηση Ηχητικών Τμημάτων με Χρήση Στατιστικών Μοντέλων

5.1 Σκοπός

Στην ενότητα αυτή μελετάται το πρόβλημα της στατιστικής μοντελοποίησης του σήματος με χρήση Κρυφών Μαρκοβιανών Μοντέλων (Hidden Markov Models - HMMs). Σκοπός είναι να γίνει κατάτμηση του ηχητικού σήματος σε τμήματα που ανήκουν σε ομογενείς κλάσεις αλλά και κατηγοριοποίηση των τμημάτων στην αντίστοιχη κλάση.

Τα δεδομένα του συνόλου εκπαίδευσης(train set) και του συνόλου δοκιμής (test set) προέρχονται από πραγματικά δελτία ειδήσεων. Τα δελτία αυτά πρέπει να έχουν κατάλληλες ετικέτες ανάλογα με τις διαφορετικές κατηγορίες ήχου που περιέχουν ώστε να μπορούν να χρησιμοποιηθούν από το σύστημα (κατάλληλο transcription). Τέτοιες κατηγορίες ήχου μπορεί να είναι φωνή, θόρυβος, σιωπή, μουσική ενώ η φωνή μπορεί να κατηγοριοποιηθεί επιπλέον σε αντρική και γυναικεία αλλά και σε φωνή με παρουσία ή απουσία θορύβου στο background.

Για την εκπαίδευση των HMMs αλλά και για την αποκωδικοποίηση των δεδομένων του test set χρησιμοποιούμε τα εργαλεία που προσφέρει το πρόγραμμα HTK [52]. Το πρόγραμμα αυτό αναπτύχθηκε από την ομάδα επεξεργασίας φωνής του πανεπιστημίου Cambridge. Χρησιμοποιείται ευρέως από την επιστημονική κοινότητα για την εκπαίδευση και την αποκωδικοποίηση μοντέλων καθώς προσφέρει πληθώρα χρήσιμων εργαλείων και τα αποτελέσματα του θεωρούνται αξιόπιστα. Αρχικά, γίνεται μία σύντομη περιγραφή του προγράμματος HTK

και των εργαλείων που προσφέρει για στατιστική μοντελοποίηση.

Στη συνέχεια δίνεται μία γενική περιγραφή των βημάτων που ακολουθήθηκαν για το σχεδιασμό ενός συστήματος που χρησιμοποιεί HMM και GMM μοντέλα. Το σύστημα αυτό πρέπει να κατατάσει διαδοχικά Frames του ηχητικού σήματος σε μία από τις διαθέσιμες κλάσεις του προβλήματος και με βάση τα αποτελέσματα να κάνει κατάτμηση και κατηγοριοποίηση του σήματος. Αναφέρονται τα γενικά βήματα που ακολουθούνται στο σχεδιασμό τέτοιων συστημάτων, οι σχεδιαστικές αποφάσεις που λήφθησαν για τη συγκεκριμένη εφαρμογή αλλά και τα προβλήματα που αντιμετωπίστηκαν.

Ακολουθεί μία περιγραφή ενός απλού συστήματος κατηγοριοποίησης ηχητικών τμημάτων. Το σύστημα αυτό δέχεται έτοιμα τα τμήματα και έχοντας εκ των προτέρων εκπαιδεύσει κατάλληλα μοντέλα για κάθε κλάση (ομιλία, θόρυβος, σιωπή, κλπ) κατατάσσει τα τμήματα στην πιο κατάλληλη κλάση. Το σύστημα αυτό δεν επιχειρεί να κάνει κατάτμηση του τμήματος που δέχεται και υποθέτει ότι το τμήμα ανήκει εξολοκλήρου σε μία από τις διαθέσιμες κλάσεις.

Στη συνέχεια περιγράφουμε ένα πιο σύνθετο σύστημα που έχει σκοπό την κατάτμηση και κατηγοριοποίηση ηχητικών τμημάτων. Δεν γίνεται πλέον η παραδοχή ότι τα τμήματα που δέχεται το σύστημα ανήκουν εξολοκλήρου σε μία κλάση, είναι δηλαδή ομογενή. Κατά συνέπεια το σύστημα προσπαθεί να χωρίσει το ηχητικό τμήμα εισόδου σε ομογενή υποτμήματα που ανήκουν σε κάποια από τις διαθέσιμες κλάσεις. Προς αυτήν την κατεύθυνση χρησιμοποιούνται διάφορες τεχνικές που αναλύονται στην αντίστοιχη ενότητα. Επίσης εισάγεται μία νέα ιδέα, συγκεκριμένα η έννοια των καμπύλων ποσοστών και παρουσιάζονται αλγόριθμοι για το διαχωρισμό του ηχητικού τμήματος σε ομογενή υποτμήματα με βάση τα αποτελέσματα των GMM μοντέλων και την ανάλυση των καμπύλων ποσοστών. Η απόδοση των διάφορων μεθόδων παρουσιάζεται και συγκρίνεται στην ενότητα των πειραματικών αποτελεσμάτων.

Εντέλει παρουσιάζεται η συμπεριφορά και η απόδοση όλων των αλγορίθμων που αναπτύχθηκαν και γίνεται σύγκριση των αποτελεσμάτων. Μελετώνται θέματα όπως η σύγκριση των διαφορετικών τεχνικών, η επίδραση της αλλαγής κάποιων παραμέτρων στα τελικά αποτελέσματα, η εξέλιση των ποσοστών επιτυχίας με βάση το χρόνο του test set που έχουμε διαθέσιμο (time sufficient statistics). Τα πειράματα έγιναν κυρίως σε δεδομένα από πραγματικά δελτία ειδήσεων. Εντούτοις, για την καλύτερη κατανόηση της λειτουργίας των συστημάτων που αναπτύχθηκαν έγιναν και πειράματα σε συνθετικά δεδομένα, κάποια από τα οποία παρουσιάζονται επιλεκτικά

Σκοπός είναι να μελετηθεί με ικανοποιητική πληρότητα το πρόβλημα της στατιστικής μοντελοποίησης ηχητικού σήματος και να αναπτυχθούν διάφορες παραλλαγές ενός συστή-

ματος που θα μπορούν να χρησιμοποιηθούν σε πραγματικές εφαρμογές για την κατάτμηση και κατηγοριοποίηση audio-streams. Τελική επιδίωξή μας ήταν ο αποτελεσματικός συνδυασμός του συστήματος κατάτμησης που περιγράφηκε στο προηγούμενο κεφάλαιο με το σύστημα στατιστικής κατάτμησης και κατηγοριοποίησης που περιγράφεται στο παρόν κεφάλαιο.

5.2 Περιγραφή των Δυνατοτήτων του Προγράμματος HTK

Το HTK είναι ένα πρόγραμμα που έχει αναπτυχθεί από την ομάδα επεξεργασίας φωνής του πανεπιστημίου Cambridge. Χρησιμοποιείται ευρέως από την επιστημονική κοινότητα για την εκπαίδευση και την αποκωδικοποίηση HMM και GMM μοντέλων καθώς προσφέρει πληθώρα χρήσιμων εργαλείων. Το πρόγραμμα αυτό είναι γραμμένο σε γλώσσα C και είναι ουσιαστικά μία συλλογή συναρτήσεων για την εκπαίδευση, αποκωδικοποίηση, προσαρμογή μαρκοβιανών μοντέλων και άλλων συγγενών λειτουργιών που χρειάζονται σε προβλήματα στατιστικής μοντελοποίησης. Μία αναλυτική περιγραφή του HTK και των δυνατοτήτων του υπάρχει στο [52]. Στη παρούσα ενότητα θα γίνει μία σύντομη περιγραφή των συναρτήσεων που χρησιμοποιήσαμε για τη δημιουργία του συστήματος κατάτμησης/κατηγοριοποίησης.

Εξαγωγή χαρακτηριστικών Για την εξαγωγή διανύσματος χαρακτηριστικών από το κάθε Frame σήματος χρησιμοποιείται η συνάρτηση **HCopy**. Η HCopy είναι ένα γενικό εργαλείο που μετατρέπει την κυματομορφή του σήματος υπό εξέταση σε διανύσματα χαρακτηριστικών, όπως τα MFCC ή LPC χαρακτηριστικά, αλλά μπορεί να κάνει μετατροπές από ένα χαρακτηριστικό σε ένα άλλο. Για τον καθορισμό λεπτομερειών όπως τα χαρακτηριστικά που θέλουμε να εξάγουμε αλλά και το μήκος του κάθε Frame και της επικάλυψης μεταξύ των Frames, χρησιμοποιούνται configuration files συγκεκριμένης δομής και πληθώρα παραμέτρων εισόδου.

Αρχικοποίηση Μοντέλων Η **HInit** χρησιμοποιείται για να παράγει αρχικές εκτιμήσεις των παραμέτρων ενός HMM Μοντέλου χρησιμοποιώντας τις παρατηρήσεις του συνόλου εκπαίδευσης. Δέχεται ώς είσοδο ένα αρχικό μοντέλο με μηδενικές ή τυχαίες τιμές παραμέτρων και εκτελεί επαναληπτικά τον αλγόριθμο Viterbi ώστε να δώσει μία εκτίμηση των παραμέτρων. Δέχεται επίσης configuration files και πληθώρα παραμέτρων που καθορίζουν λεπτομέρειες της εκτέλεσης.

Εκπαίδευση των μοντέλων Για την εκπαίδευση των μοντέλων χρησιμοποιείται η **HRest**. Η συνάρτηση αυτή χρησιμοποιεί τον βασικό αλγόριθμο Baum-Welch για τον υπολογισμό των παραμέτρων ενός HMM μοντέλου με βάση τα δεδομένα που του training

set. Η HRest έχει υλοποιηθεί ώστε να λειτουργεί με HMMs που έχουν προηγουμένως αρχικοποιηθεί με την HInit. Επίσης, η HREst υποστηρίζει πληθώρα λειτουργιών όπως η εκπαίδευση μοντέλων με πολλαπλά μίγματα σε κάθε κατάσταση, η χρήση πλήρων ή διαγώνιων πινάκων αυτοσυσχέτισης, η χρήση συσχετισμένων παραμέτρων ή μιγμάτων (parameter or mixture tying) κλπ. Κατά τη διάρκεια της εκπαίδευσης μπορούμε να αυξάνουμε σταδιακά τον αριθμό των μιγμάτων κάθε κατάστασης χρησιμοποιώντας τη συνάρτηση **HHed**, η οποία τροποποιεί τον ορισμό της δομής του HMM που εκπαιδεύουμε ώστε να προστίθενται μίγματα. Μετά από μία κλήση της HHed πρέπει να γίνει επανεκπαίδευση του μοντέλου με χρήση της HRest.

Διαδικασία Δοκιμής-Αποκωδικοποίησης Η συνάρτηση **HVite** χρησιμοποιείται για την αποκωδικοποίηση ενός αρχείου ήχου εισόδου και την κατηγοριοποίση/ αντιστοίχιση των Frames του αρχείου σε κάθε ένα από τα διαθέσιμα μοντέλα. Για τη σωστή λειτουργία της η HVite χρειάζεται μία περιγραφή του δικτύου των HMMs που είναι διαθέσιμα και ένα κατάλληλο λεξικό για τα HMMs. Η HVite υποστηρίζει πλήθος λειτουργιών ανάμεσα στις οποιές είναι και η αντιστοίχιση των Frames εισόδου στις διάφορες καταστάσεις ενός HMM μοντέλου (state alignment).

Υπολογισμός Στατιστικών Αποτελεσμάτων Για την ανάλυση της απόδοσης χρησιμοποιείται η συνάρτηση **HResults** η οποία παράγει στατιστικά αποτελέσματα. Η HResults δέχεται το transcription το οποίο θέλουμε να αξιολογήσουμε και ένα transcription αναφοράς και τα συγκρίνει παράγοντας χρήσιμα στατιστικά μέτρα απόδοσης όπως %Correct και %Accurate.

Προσαρμογή των Μοντέλων Σε πολλά συστήματα αναγνώρισης είναι χρήσιμο τα μοντέλα που έχουν εκπαιδευτεί κατά τη διάρκεια της διαδικασίας εκπαίδευσης να μεταβάλλονται κατά τη διάρκεια των διαδικασιών δοκιμής και χρήσης του συστήματος, έτσι ώστε να προσαρμόζονται στα νέα δεδομένα που δέχεται το σύστημα. Για τη λειτουργία της προσαρμογής παρέχεται η συνάρτηση **HEAdapt**. Η συνάρτηση αυτή εκτελεί προσαρμογή ενός συνόλου HMM μοντέλων χρησιμοποιώντας είτε Maximum Likelihood Linear Regression (MLLR) είτε Maximum A-Posteriori (MAP) είτε και τα δύο.

Σημειώνουμε επιπλέον ότι επειδή οι παραπάνω συναρτήσεις δέχονται ένα μεγάλο πλήθος παραμέτρων αλλά διάφορα configuration αρχεία που καθορίζουν τη λειτουργία τους, δεν συνηθίζεται να καλούνται απευθείας από τη γραμμή εντολών. Η κλήση τους γίνεται

συνήθως μέσω κατάλληλων perl scripts τα οποία καθορίζουν κατάλληλα τις παραμέτρους και διαχειρίζονται τα αρχεία εισόδου και εξόδου που χρειάζονται οι συναρτήσεις.

5.3 Γενικά Στοιχεία Σχεδίασης Συστήματος Κατάτμησης και Κατηγοριοποίησης Βασισμένο σε HMM/GMM Μοντέλα

Ο σχεδιασμός του συστήματος κατάτμησης/κατηγοριοποίησης audio-streams που βασίζεται σε HMM/GMM μοντέλα ακολουθεί τα παρακάτω βήματα:

Χωρισμός των Διαθέσιμων Δεδομένων σε σύνολα Εκπαίδευσης και Δοκιμής Το πρώτο βήμα για τη σχεδίαση ενός συστήματος βασισμένου σε HMM και GMM μοντέλα είναι η εύρεση ικανοποιητικού πλήθους δεδομένων που θα χρησιμοποιηθούν στις φάσεις εκπαίδευσης και δοκιμής του συστήματος. Η μορφή των διαθέσιμων δεδομένων και η ευκολία εύρεσης τους εξαρτάται από την εκάστοτε εφαρμογή. Τα δεδομένα πρέπει να χωριστούν σε σύνολο εκπαίδευσης (training set) με βάση το οποίο θα γίνει η εκπαίδευση των παραμέτρων και σε σύνολο δοκιμής (test set) με βάση το οποίο θα γίνει η δοκιμή της απόδοσης του συστήματος. Τα δύο σύνολα αυτά πρέπει να είναι ξένα μεταξύ τους και το training set είναι συνήθως πολύ μεγαλύτερο από το test set. Γενικά είναι καλό να έχουμε μεγάλο πλήθος δεδομένων εκπαίδευσης ώστε να μπορούμε να υπολογίσουμε με αξιόπιστο και γενικό τρόπο τις παραμέτρους των μοντέλων.

Στην εφαρμογή κατάτμησης και κατηγοριοποίησης από audio-streams τα διαθέσιμα δεδομένα είναι μεταγραφές (Transcriptions) από πραγματικά δελτία ειδήσεων. Τα transcriptions είναι σε μορφή xml αρχείων που παρέχουν πληροφορίες για την αρχή και το τέλος ηχητικών τμημάτων του δελτίου. Τέτοια τμήματα κατηγοριοποιούνται ως ομιλία, σιωπή, θόρυβος, μουσική ενώ για τα τμήματα της ομιλίας μπορούμε να εξάγουμε από το transcription επιπλέον πληροφορία σχετικά με το φύλο του ομιλητή και με την ύπαρξη ή όχι θορύβου στο background.

Σημειώνουμε εδώ ότι αν και η ηχογράφηση ενός δελτίου ειδήσεων είναι εύκολο να γίνει και από την άποψη αυτή η συλλογή των ηχητικών δεδομένων δεν παρουσιάζει δυσκολία, η δημιουργία ενός σωστού transcription για ένα δελτίο ειδήσεων είναι χρονοβόρα και κοπιαστική εργασία. Επίσης, η καλή εκπαίδευση των μοντέλων εξαρτάται από την ορθότητα των transcriptions δηλαδή από τη σωστή σήμανση των ηχητικών περιοχών. Κατά συνέπεια, η

εύρεση μεγάλου πλήθους δεδομένων εκπαίδευσης είναι ένα πρόβλημα που θα μας απασχολήσει κατά τη σχεδίαση και την υλοποίηση του συστήματος. Θεωρούμε ότι μία ικανοποιητική αναλογία συνόλων εκπαίδευσης και δοκιμής είναι περίπου 10:1 δηλαδή χρήση 10 δελτίων (της 1 ώρας περίπου) για εκπαίδευση και δοκιμή με χρήση ενός δελτίου.

Καθορισμός των κλάσεων του συστήματος Επίσης, πρέπει να οριστούν οι κλάσεις του προβλήματος δηλαδή οι κλάσεις στις οποίες θέλουμε να κατατάξουμε κάθε frame του εξεταζόμενου audio-stream. Κάθε μία κλάση θα αντιπροσωπεύεται από ένα GMM μοντέλο με κάποιο πλήθος γκαουσιανών μιγμάτων. Το πλήθος των μιγμάτων εξαρτάται από την πολυπλοκότητα της κλάσης που θέλουμε να μοντελοποιήσουμε αλλά και από την υπαρξη ικανοποιητικού πλήθους διαθέσιμων δεδομένων εκπαίδευσης του μοντέλου. Ο ορισμός των κλάσεων καθορίζεται βασικά από τις προδιαγραφές του συστήματος που θέλουμε να υλοποιήσουμε αλλά επηρεάζεται και από τα διαθέσιμα δεδομένα.

Πιθανές κλάσεις για την εφαρμογή σε audio-streams είναι ομιλία, ο θόρυβος, η μουσική, η σιωπή, η αντρική/γυναικεία ομιλία, η ομιλία με παρουσία ή απουσία θορύβου κλπ. Εντούτοις τα δελτία ειδήσεων δεν περιέχουν ίση ποσότητα ηχητικών δεδομένων από κάθε μία από τις παραπάνω κλάσεις.

Πιο συγκεκριμένα, από μετρήσεις που έγιναν σε πραγματικά δελτία του training set βρέθηκε ότι συνήθως το 95% του συνολικού δελτίου αποτελείται από ομιλία ενώ το 4% είναι θόρυβος ή μουσική και το 1% σιωπή. Επίσης το 50% της ομιλίας είναι από αντρική και το υπόλοιπο 50% είναι γυναικεία ομιλία. Τέλος περίπου το 75% της ομιλίας γίνεται χωρίς την παρουσία θορύβου στο background ενώ στο υπόλοιπο 25% ακούγεται κάποιο είδος θορύβου κατά τη διάρκεια της ομιλίας.

Σχετικά με τα transcriptions σημειώνουμε ότι η σήμανση των σιωπών των ομιλητή μπορεί να είναι υποκειμενική και συχνά αμελείται. Επίσης η παρουσία ή απουσία θορύβου είναι αρκετά υποκειμενική σε κάποιες περιπτώσεις και εξαρτάται από την ακουστική ευαισθησία του ατόμου που γράφει το transcription.

Σύμφωνα με τα παραπάνω καταλήγουμε στην επιλογή κλάσεων όπως ομιλία, μη ομιλίας, αντρικής ομιλία και γυναικείας ομιλία. Στην περίπτωση της μη ομιλίας κατατάσσουμε το θόρυβο, τη σιωπή και τη μουσική. Αποφασίσαμε να συγχωνεύσουμε τις 3 προηγούμενες κλάσεις στην περίπτωση της μη ομιλίας έτσι ώστε να εξασφαλίσουμε την ύπαρξη περισσότερων δεδομένων για την εκπαίδευση του μοντέλου. Επίσης, προτιμήσαμε να κατατάξουμε περαιτέρω την ομιλία σε αντρική και γυναικεία επειδή μία τέτοια κατάταξη είναι πιο αντικειμενική από την κατάταξη σε καθαρή και θορυβώδη ομιλία και επιπλέον επειδή έχουμε περίπου ίση ποσότητα αντρικής και γυναικείας ομιλίας στα διαθέσιμα δελτία.

Εξαγωγή Χαρακτηριστικών από το audio-stream Πρέπει να επιλεγούν τα χαρακτηριστικά που θα εξάγουμε από τα δεδομένα ώστε να εκπαιδεύσουμε τα μοντέλα του συστήματος. Στη βιβλιογραφία για τέτοια συστήματα συναντάται συχνά η χρήση συντελεστών MFCC (Mel Frequency Cepstral Coefficients) μαζί με τις πρώτες και δεύτερες παραγώγους τους.

Στην υλοποίησή μας χρησιμοποιούνται 39 συνολικά συντελεστές, δηλαδή 12 Mel Frequency Cepstral Coefficients (MFCC) και το συντελεστή της ενέργειας μαζί με τις πρώτες και τις δεύτερες παραγώγους τους. Σημειώνουμε ότι υπάρχει διαφοροποίηση σε σχέση με τα χαρακτηριστικά που χρησιμοποιήθηκαν στο σύστημα κατάτμησης. Πράγματι, στο σύστημα κατηγοριοποίησης χρησιμοποιούμε επιπλέον τις πρώτες και δεύτερες παραγώγους των MFCC συντελεστών καθώς αυτές περιέχουν χρήσιμη πληροφορία για τις ηχητικές κλάσεις που θέλουμε να μοντελοποιήσουμε. Για την παραγωγή των χαρακτηριστικών χρησιμοποιούνται 35 φίλτρα, και παράθυρο των 40msec με 10msec overlap. Η εξαγωγή των χαρακτηριστικών γίνεται με χρήση της συνάρτησης HCopy του HTK.

Εκπαίδευση των κατάλληλων μοντέλων Στη συνέχεια γίνεται εκπαίδευση των κατάλληλων μοντέλων, συγκεκριμένα δημιουργείται ένα GMM μοντέλο για κάθε μία από τις κλάσεις που υπάρχουν στα διαθέσιμα transcriptions, δηλαδή $class_i, i = 1, \ldots, n$. Η διαδικασία της εκπαίδευσης περιλαμβάνει την αρχικοποίηση των GMM μοντέλων με τα δεδομένα που ανήκουν στις αντίστοιχες κλάσεις και με χρήση της HInit. Στη συνέχεια γίνεται και υπολογισμός των παραμέτρων των μοντέλων με χρήση των δεδομένων εκπαίδευσης της αντίστοιχης κλάσης και με χρήση της HRest.

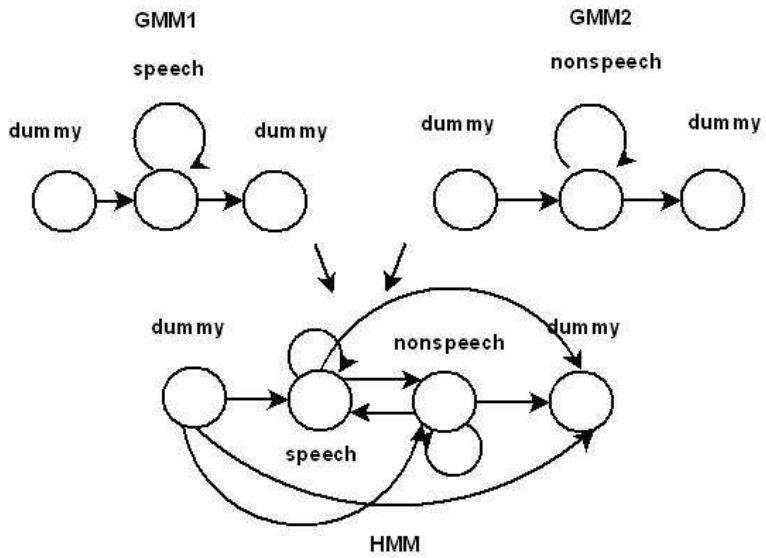
Αρχίζουμε την εκπαίδευση χρησιμοποιώντας μία γκαουσιανή κατανομή για την μοντελοποίηση της κάθε κλάσης. Σταδιακά μπορούμε να αυξήσουμε τον αριθμό των γκαουσιανών κατανομών που αντιστοιχούν στο κάθε GMM χρησιμοποιώντας την HHed για να αυξήσουμε τα γκαουσιανά μίγματα του GMM και την HRest για να επανεκτιμήσουμε τις παραμέτρους του GMM. Αυτό επαναλαμβάνεται μέχρι να έχουμε τον απαιτούμενο αριθμό μιγμάτων σε κάθε GMM.

Ο αριθμός των μιγμάτων που θα χρησιμοποιηθεί εξαρτάται από την πολυπλοκότητα της αντίστοιχης κλάσης αλλά και από το πλήθος των διαθέσιμων δεδομένων για την εκπαίδευση των αντίστοιχων μοντέλων. Όσο περισσότερα μίγματα χρησιμοποιούμε τόσο περισσότερα δεδομένα πρέπει να διαθέτουμε για τον αξιόπιστο υπολογισμό των παραμέτρων των μοντέλων. Στην υλοποίησή μας τυπικές τιμές είναι 8 μίγματα για τα μοντέλα ομιλίας, αντρικής και γυναικείας ομιλίας και 4 μίγματα για το μοντέλο μη ομιλίας.

Με αυτό τον τρόπο παίρνουμε n GMM μοντέλα, όπου n ο αριθμός των κλάσεων του

προβλήματος, τα οποία σύμφωνα με τη σύμβαση του HTK αποτελούνται από 3 καταστάσεις, από τις οποίες η μεσαία μοντελοποιεί την αντίστοιχη κλάση και οι άλλες δύο αποτελούν dummy states. Θα μπορούσαμε να χρησιμοποιήσουμε αυτά τα GMMs στο στάδιο αποκωδικοποίησης (decoding), ώστε να κατατάξουμε κάθε frame του test set στην κατάλληλη κλάση. Όμως για λόγους που έχουν να κάνουν περισσότερο με την καλύτερη χρήση της μνήμης του υπολογιστικού συστήματος και την καλύτερη απόδοση, χρησιμοποιούμε μία ισοδύναμη μοντελοποίηση του προβλήματος με HMMs.

Συγκεκριμένα, δημιουργούμε ένα HMM που περιέχει τις μεσαίες καταστάσεις από τα GMMs που εκπαιδεύσαμε νωρίτερα και δύο επιπλέον καταστάσεις, την αρχική και την τελική dummy κατάσταση. Ουσιαστικά στο HMM αυτό συγχωνεύουμε τα GMMs που εκπαιδεύσαμε, όπως φαίνεται σχηματικά στην εικόνα 5.1. Η κάθε κατάσταση (εκτός από τις dummy states) μοντελοποιεί μία από τις κλάσεις του προβλήματος και οι μεταβάσεις από τη μία κατάσταση στην άλλη μοντελοποιούν τις αλλαγές μεταξύ κλάσεων σε ένα audio-stream. Ο πίνακας πιθανοτήτων μετάβασης δημιουργείται έτσι ώστε να δίνει για κάθε κατάσταση ίδια πιθανότητα μετάβασης στην ίδια κατάσταση και στις υπόλοιπες καταστάσεις και μικρή πιθανότητα μετάβασης στην τελική dummy κατάσταση. Ομοίως από την αρχική dummy κατάσταση μπορούμε να μεταβούμε με ίση πιθανότητα σε μία από τις καταστάσεις και με μικρή πιθανότητα στην τελική dummy κατάσταση. Τα παραπάνω υλοποιήθηκαν με κατάλληλα perl scripts.



Εικόνα 5.1: Συγχώνευση GMM μοντέλων σε ένα HMM μοντέλο.

Με αυτή τη μοντελοποίηση μετατρέπουμε το πρόβλημα decoding σε πρόβλημα εύρεσης alignment σε ένα HMM μοντέλο, όπου κάθε κατάσταση αντιπροσωπεύει μια κλάση.

Δοκιμή του συστήματος Τελικά έχουμε το στάδιο της δοκιμής του συστήματος στα δεδομένα του test set (decoding). Για το σκοπό αυτό χρησιμοποιείται η συνάρτηση HVite του HTK η οποία κάνει alignment των test δεδομένων στις καταστάσεις του HMM μοντέλου και επιστρέφει ως αποτέλεσμα τις χρονικές στιγμές που το audio-stream βρίσκεται σε κάθε κατάσταση. Δεδομένου ότι κάθε κατάσταση μοντελοποιεί μία κλάση, βρίσκουμε έτσι τις κλάσεις στις οποίες ανήκουν τα διαδοχικά frames του audio-stream.

Το σύστημα αυτό μας δίνει τη δυνατότητα τόσο να κατατάξουμε ένα τμήμα σε μία από τις διαθέσιμες κλάσεις αν τα Frames του ανήκουν στην κλάση αυτή, όσο και να εντοπίσουμε αλλαγές μεταξύ των κλάσεων και να κάνουμε κατάτμηση του συνολικού τμήματος σε ομογενή υποτμήματα.

Λεπτομέρειες Υλοποίησης Σχετικά με την υλοποίηση του συστήματος, αναφέρουμε ότι για τη δημιουργία, εκπαίδευση και δοκιμή των μοντέλων, χρησιμοποιήθηκαν οι κατάλληλες συναρτήσεις του HTK (γραμμένες σε C). Για λειτουργίες όπως η κλήση των συναρτήσεων αυτών, η ανάγνωση αρχείων εισόδου και η εγγραφή σε αρχεία εξόδου χρησιμοποιήθηκαν perl script. Επίσης, perl scripts χρησιμοποιήθηκαν και για διάφορες άλλες λειτουργίες όπως για ανάγνωση των trs αρχείων, που ουσιαστικά είναι XML αρχεία και περιέχουν τα transcriptions των δελτίων ειδήσεων, για εγγραφή των trascriptions σε μορφή που μπορεί να χρησιμοποιηθεί από το HTK, για εγγραφή των αποτελεσμάτων του HTK σε trs αρχεία για την καλύτερη παρουσίαση τους, για εξαγωγή στατιστικών στοιχείων σχετικά με την απόδοση του συστήματος (confusion matrices), κλπ. Τέλος, ο συντονισμός και ο έλεγχος του συνολικού συστήματος γίνεται μέσω του Matlab.

5.4 Περιγραφή Συστήματος Κατηγοριοποίησης

Σε αυτή την ενότητα εξετάζεται ένα υποσύνολο του προβλήματος κατάτμησης και κατηγοριοποίησης, συγκεκριμένα μόνο το πρόβλημα της κατηγοριοποίησης ομογενών ακουστικών τμημάτων με χρήση HMM. Συνεπώς το σύστημα που θα περιγραφεί παρακάτω δέχεται ως είσοδο ένα τμήμα audio-stream το οποίο θεωρείται ομογενές και το κατατάσσει σε κάποια από τις διαθέσιμες κλάσεις. Τα ομογενή τμήματα προέρχονται από το στάδιο εύρεσης αλλαγών σε audio-stream και κατάτμησής του. Αλγόριθμοι εύρεσης αλλαγών σε audio -sreams έχουν περιγραφεί αναλυτικά στο προηγούμενο κεφάλαιο.

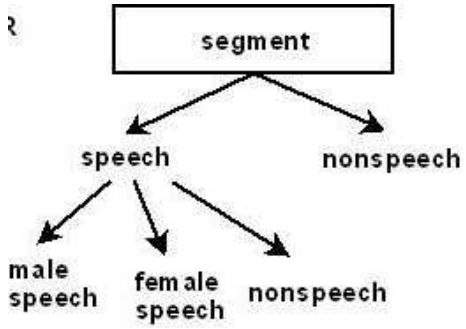
Άρα το αρχικό σύστημα που σχεδιάζουμε περιέχει πλήρως διαχωρισμένα τα στάδια της κατάτμησης και της κατηγοριοποίησης. Η χρήση HMM/GMM μοντέλων περιορίζεται μόνο στην κατηγοριοποίηση.

Το επόμενο βήμα είναι να καθοριστεί ποιές θα είναι οι διαθέσιμες κλάσεις στις οποίες θα μπορεί να κατηγοριοποιηθεί ένα τμήμα. Για κάθε μία από αυτές τις κλάσεις θα εκπαιδευτεί ένα GMM και τελικά τα GMMs θα συγχωνευτούν σε ένα HMM που θα χρησιμοποιηθεί για την κατάταξη. Προκαταρκτικά πειράματα που έγιναν για την εύρεση των κατάλληλων κλάσεων υποδεικνύουν ότι το σύστημα αποδίδει καλύτερα όταν το πλήθος των κλάσεων είναι μικρό. Κατά συνέπεια, είναι καλύτερα η κατάταξη να γίνεται σε στάδια, όπου το κάθε στάδιο να περιέχει περισσότερη λεπτομέρεια. Το πλήθος των κλάσεων του κάθε σταδίου είναι συνήθως 2 ή 3.

Στο πρώτο στάδιο έχουμε 2 μοντέλα, ένα για ομιλία και ένα για μή ομιλία. Η περίπτωση της μη ομιλίας περιλαμβάνει θόρυβο, σιωπή και μουσική. Σε δεύτερο επίπεδο γίνεται περαιτέρω κατηγοριοποίηση της ομιλίας σε αντρική και γυναικεία. Ως αποτέλεσμα παίρνουμε ένα labeling των τμημάτων του δελτίου δοκιμής σε ομιλία και μη ομιλία από το πρώτο στάδιο, ενώ από το δεύτερο στάδιο παίρνουμε ένα labeling σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία.

Για να αποφασίσει το σύστημα σε ποια κλάση θα κατατάξει το τμήμα υπό εξέταση υπολογίζει για όλα τα frames που ανήκουν σε μία συγκεκριμένη κλάση το ποσοστό του συνολικού μήκους τους ως προς το συνολικό μήκος του τμήματος. Στη συνέχεια, ορίζονται κατάλληλα κατώφλια για την κατάταξη, τέτοια ώστε να ευνοούν την κατάταξη σε φωνή. Αυτό γίνεται επειδή θεωρείται γενικά προτιμότερο να κατατάξουμε λανθασμένα ένα τμήμα μη ομιλίας σε ομιλία παρά το αντίστροφο. Τα τμήματα που έχουν αναγνωριστεί ως μη ομιλία δεν θα απορριφθούν από το συνολικό σύστημα, κατά συνέπεια λανθασμένη κατάταξη της ομιλίας θα μπορούσε να οδηγήσει σε απώλεια χρήσιμης ομιλίας, γεγονός που προσπαθούμε να αποφύγουμε.

Ένα διάγραμμα του συστήματος φαίνεται στην εικόνα 5.2. Εκεί παρατηρούμε ότι στο δεύτερο επίπεδο που γίνεται λεπτομερής κατηγοριοποίηση της φωνής έχει προστεθεί και η κλάση της μη φωνής. Μέσα από πειράματα αυτό έχει φανεί χρήσιμο. Σε πρώτο επίπεδο προτιμούμε να κατηγοριοποιήσουμε φωνή ως μη φωνή παρά το αντίστροφο άρα ευνοούμε την κατάταξη των τμημάτων σε φωνή. Σε δεύτερο επίπεδο, που πιο λεπτομερής πληροφορία είναι διαθέσιμη, κάποια από τα τμήματα μη φωνής που λανθασμένα κατηγοριοποιήθηκαν ως φωνή μπορούν να ξεδιαλεχθούν.



Εικόνα 5.2: Σύστημα κατάταξης τμηματων audio-stream.

5.5 Περιγραφή Συστήματος Κατάτμησης και Κατηγοριοποίησης

Στην ενότητα αυτή ασχολούμαστε με το γενικότερο πρόβλημα της κατάτμησης και κατηγοριοποίησης με χρήση HMM και GMM μοντέλων. Δεν κάνουμε πλέον την παραδοχή ότι τα τμήματα που δέχεται το σύστημα ανήκουν εξολοκλήρου σε μία κλάση, είναι δηλαδή ομογενή. Αντίθετα, με βάση την κατηγοριοποίηση των frames του τμήματος υπό εξέταση προσπαθούμε να χωρίσουμε το τμήμα σε υποτμήματα όσο το δυνατόν πιο ομογενή και να κατατάξουμε το κάθε υποτμήμα στην περισσότερο ταιριαστή κλάση. Το σύστημα που σχεδιάζουμε πρέπει να λειτουργεί σε κάθε περίπτωση, όχι μόνο όταν οι κλάσεις είναι σαφώς διαχωρισμένες μεταξύ τους και υπάρχουν σχετικά ομογενή υποτμήματα σε υπό εξέταση τμήμα. Πρέπει επίσης να λειτουργεί και σε δύσκολες περιπτώσεις όπου οι κλάσεις είναι δύσκολα διαχωρίσιμες και το σύνολο των δεδομένων εκπαίδευσης είναι μικρό. Σε τέτοιες περιπτώσεις το σύστημα ξεχωρίζει εκείνα τα υποτμήματα με μεγάλη ανομοιογένεια και παίρνει κάποια λογική απόφαση για αυτά, όπως το να τα κατατάξει στην κλάση όπου ανήκει η πλειοψηφία των frames τους. Δηλαδή στη χειρότερη περίπτωση η συμπεριφορά του συστήματός μας γίνεται ίδια με την συμπεριφορά του απλού συστήματος που παρουσιάσαμε στην προηγούμενη αναφορά. Τέλος, το σύστημα πρέπει να είναι υπολογιστικά αποδοτικό, δηλαδή να λειτουργεί με γραμμικό τρόπο ως προς το συνολικό αριθμό των (ομογενών) τμημάτων που βρίσκουμε αλλά και να βρίσκει τα ομογενή τμήματα με τον καλύτερο δυνατό τρόπο.

5.5.1 Περιγραφή του Προβλήματος και Σχεδιασμός του Συστήματος

Θεωρούμε ότι έχουμε κάποιες γνωστές κλάσεις και ένα υπό εξέταση audio-stream το οποίο αποτελείται από υποτμήματα καθένα από τα οποία ανήκει σε μία από τις υπάρχουσες κλάσεις. Σκοπός μας είναι να βρούμε τα σημεία του audio-stream στα οποία γίνεται αλλαγή από τη μία κλάση στην άλλη και στη συνέχεια να κατατάξουμε το κάθε υποτμήμα που ορίζουν δύο σημαία αλλαγής στην κατάλληλη κλάση.

Για το σκοπό αυτό εκπαιδεύουμε ένα GMM μοντέλο για κάθε μία από τις κλάσεις. Στη συνέχεια συγχωνεύουμε τα GMM μοντέλα σε ένα HMM μοντέλο όπου η μετάβαση από μία κατάσταση σε άλλη συμβολίζει μετάβαση από μία κλάση σε άλλη.

Το σύστημα δέχεται το υπό εξέταση τμήμα και κατατάσσει διαδοχικά τα frames του στην πιο ταιριαστή κλάση. Επειδή το κάθε frame έχει μικρή διάρκεια συχνά καταλήγουμε να παίρνουμε ως αποτέλεσμα ένα υπερκατατμημένο τμήμα με συχνές εναλλαγές από τη μία κλάση στην άλλη από frame σε frame. Το πρόβλημα είναι πώς θα εξάγουμε από το αποτέλεσμα αυτό κάποιες σχετικά ομογενείς περιοχές ώστε να τις κατατάξουμε στην κατάλληλη κατηγορία και πώς θα βρούμε με ακρίβεια τα όρια μεταξύ των διαφορετικών περιοχών.

Ένα παράδειγμα θα παρουσιαστεί παρακάτω, το οποίο επιλέγεται λόγω της δυσκολίας που παρουσιάζει. Για το παράδειγμα χρησιμοποιήθηκαν συνθετικά δεδομένα. Πειραματιζόμαστε με τρεις κλάσεις που μοντελοποιούνται με 2 γκαουσιανές κατανομές και είναι σχετικά δύσκολα διαχωρίσιμες:

class1 Μίγμα 1:Μέση τιμή 0 και διακύμανση 1, Μίγμα 2: Μέση τιμή 1 και διακύμανση 1

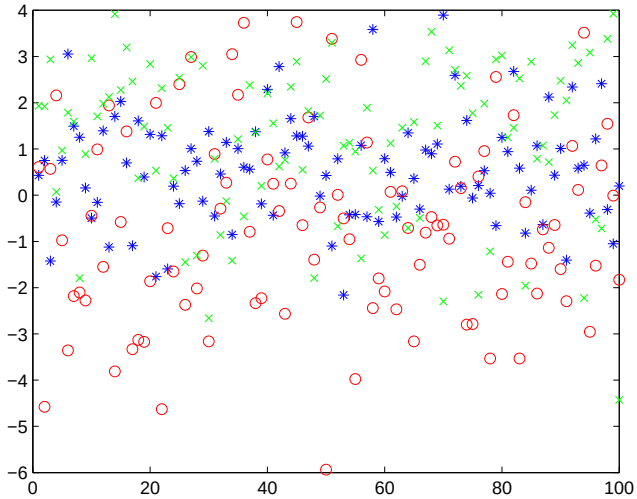
class2 Μίγμα 1:Μέση τιμή 0 και διακύμανση 2, Μίγμα 2: Μέση τιμή -1 και διακύμανση 2

class3 Μίγμα 1:Μέση τιμή 2 και διακύμανση 1, Μίγμα 2: Μέση τιμή 1 και διακύμανση 2

Στο σχήμα 5.3 φαίνονται 100 παρατηρήσεις από κάθε μία από τις παραπάνω κλάσεις.

Στον πίνακα 5.5 φαίνονται τα αποτελέσματα όταν χρησιμοποιήθηκαν 100 δεδομένα ανά κλάση στο training set και η μοντελοποίηση της κάθε κλάσης έγινε με 1 γκαουσιανή κατανομή. Τα αποτελέσματα αυτά υπολογίστηκαν με χρήση της συνάρτησης HResults του HTK. Η απόδοση είναι μέτρια δεδομένης της δυσκολίας διαχωρισμού αλλά και του ότι είχαμε σχετικά λίγα δεδομένα εκπαίδευσης και χρησιμοποιήθηκε μόνο μία γκαουσιανή κατανομή.

Η εικόνα 5.4 δείχνει πως φαίνεται το αντίστοιχο transcription όπου στην πάνω γραμμή φαίνεται το αποτέλεσμα του συστήματος και στην κάτω γραμμή φαίνεται το σωστό transcription(δηλαδή class1 για 0-1sec, class2 για 1-2sec και class3 για 2-3sec). Σημειώνουμε

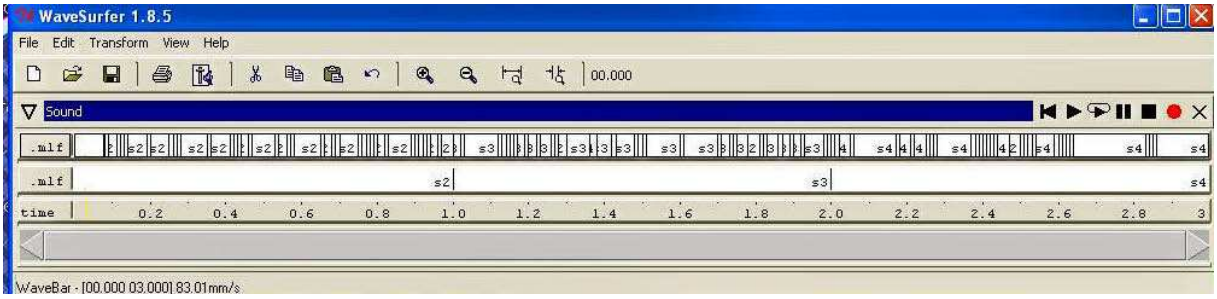


Εικόνα 5.3: 100 παρατηρήσεις από κάθε μία από τις 3 κλάσεις του πειράματος: class1 (mean=0, std=1,mean=1, std=1, μπλε χρώμα και συμβολισμός με *), class2 (mean=0, std=2, mean=-1, std=2, κόκκινο χρώμα και συμβολισμός με o), class3 (mean=2, std=1, mean=1, std=2, πράσινο χρώμα και συμβολισμός με x).

Πίνακας 5.1: Αποτελέσματα του πειράματος όπου το training set είχε 100 παρατηρήσεις ανά κλάση και η μοντελοποίηση των κλάσεων έγινε με 1 γκαουσιανή.

className	Hits	FAs	Actual
class1	76	29	100
class2	74	15	100
class3	79	27	100
Overall	229	71	300
%Correct	76.33		
%Accuracy	76.33		

ότι η κλάση 1 συμβολίζεται με s2, η κλάση 2 με s3 και η κλάση 3 με s4. Κάθε γραμμή δείχνει ένα transcription, όπου τα σημεία αλλαγής σημειώνονται με κατακόρυφες γραμμές και μέσα σε κάθε τμήμα που ορίζεται από 2 σημεία αλλαγής γράφεται η αντίστοιχη κλάση που αναγνώρισε το σύστημα.



Εικόνα 5.4: Αποτελέσματα, σε μορφή transcription, του συστήματος για τις 3 κλάσεις. Στην πάνω γραμμή φαίνεται το αποτέλεσμα ενώ στην κάτω γραμμή φαίνεται το transcription αναφοράς

Παρατηρούμε ότι αν και υπάρχει σε μεγάλο βαθμό υπερκατάτμηση και πολλά μικρά τμήματα σε μία περιοχή κατατάσσονται λανθασμένα σε άλλες κλάσεις, η πλειοψηφία των τμημάτων μίας περιοχής κατατάσσεται στη σωστή κλάση και επίσης παρατηρείται αλλαγή κλάσεων κοντά στα όρια μεταξύ δύο περιοχών. Αυτές είναι οι πληροφορίες που θέλουμε να εξάγουμε από το αποτέλεσμα ώστε έχοντας ως δεδομένο ένα transcription όπως αυτό της πάνω γραμμής να παράγουμε ένα transcription όπως αυτό της κάτω γραμμής.

5.5.2 Χρήση Καμπύλων Ποσοστών

Για να εξάγουμε από το υπερκατατμημένο αποτέλεσμα που μας δίνει το σύστημα με τα GMMs τα όρια των ομογενών περιοχών, εισάγουμε μία νέα ιδέα, τις καμπύλες ποσοστών. Αρχικά ορίζουμε τις καμπύλες ποσοστών οι οποίες μας δίνουν πληροφορία για το που αρχίζουν και τελειώνουν ομογενή τμήματα, δηλαδή περιοχές που ανήκουν σε μία συγκεκριμένη κλάση. Στη συνέχεια, από την καμπύλη ποσοστών για την κάθε κλάση εξάγουμε τα όρια των υποδιαστημάτων που ανήκουν στην κλάση αυτή. Τέλος, έχοντας ως δεδομένα κάποια πιθανά όρια υποδιαστημάτων για την κάθε κλάση στο συνολικό τμήμα υπό εξέταση, επιχειρούμε να τα συνδυάσουμε με ένα ικανοποιητικό και αποδοτικό τρόπο ώστε να παράγουμε ένα transcription χωρίς κενά διαστήματα και επικαλύψεις και με σωστή κατανομή των διαστημάτων στις αντίστοιχες κλάσεις.

Καμπύλες Ποσοστών

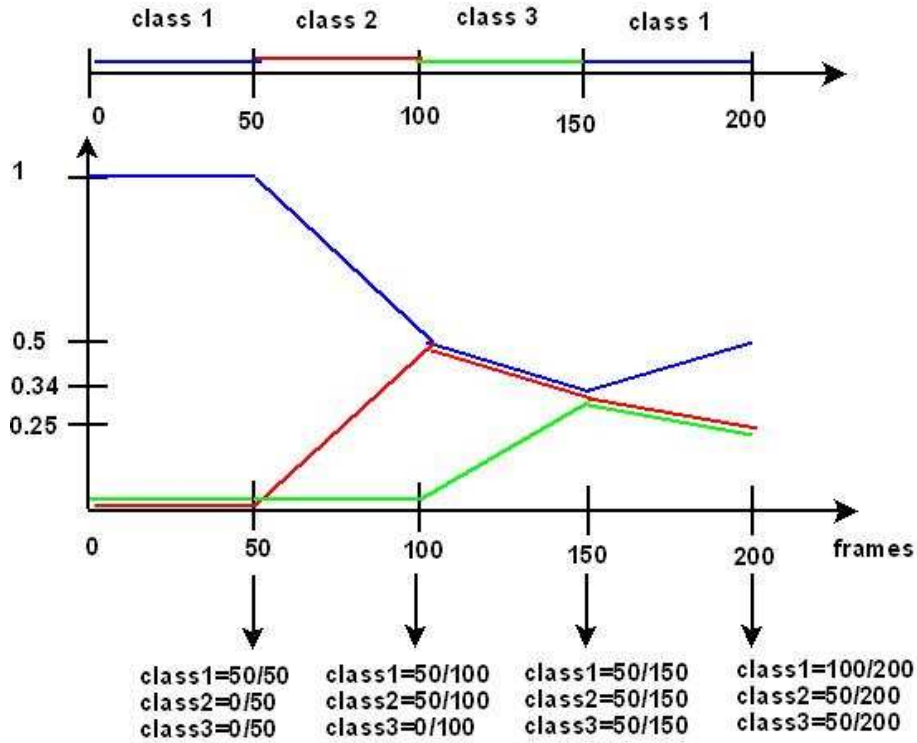
Ως αποτέλεσμα από το σύστημα με τα GMM μοντέλα παίρνουμε μία αλληλουχία υποτμημάτων του τμήματος εξέτασης με τα σημεία αρχής και τέλους του κάθε υποτμήματος και την κατάταξη του υποτμήματος αυτού (δηλαδή παίρνουμε ένα transcription). Τα υποτμήματα έχουν διαφορετικά μήκη, με τα ελάχιστο δυνατό μήκος να είναι όσο και ένα Frame και το μέγιστο δυνατό μήκος να είναι όσο το μήκος του συνολικού τμήματος υπό εξέταση. Για να δημιουργήσουμε τις καμπύλες ποσοστών διαβάζουμε διαδοχικά τα υποτμήματα του transcription και σε κάθε σημείο αλλαγής υπολογίζουμε το μήκος της κάθε κλάσης μέχρι το σημείο αυτό ώς προς το συνολικό μήκος μέχρι το σημείο αυτό. Δηλαδή σε κάθε σημείο αλλαγής που εντοπίζουμε υπολογίζουμε το ποσοστό μήκους (δηλαδή frames) που έχει καταλάβει η κάθε κλάση μέχρι το σημείο αυτό. Στη συνέχεια δημιουργούμε τη γραφική παράσταση των ποσοστών συναρτήσει των σημείων αλλαγής. Σημειώνουμε ότι τα σημεία αυτά δεν ισαπέχουν.

Ένα απλό παράδειγμα υπολογισμού καμπύλων ποσοστών φαίνεται στο σχήμα 5.5. Έχουμε ένα transcription που περιέχει 200 frames, από τα οποία τα 50 πρώτα ανήκουν στην κλάση 1, τα επόμενα 50 στην κλάση 2, τα επόμενα 50 στην κλάση 3 και τα τελευταία 50 στην κλάση 1.

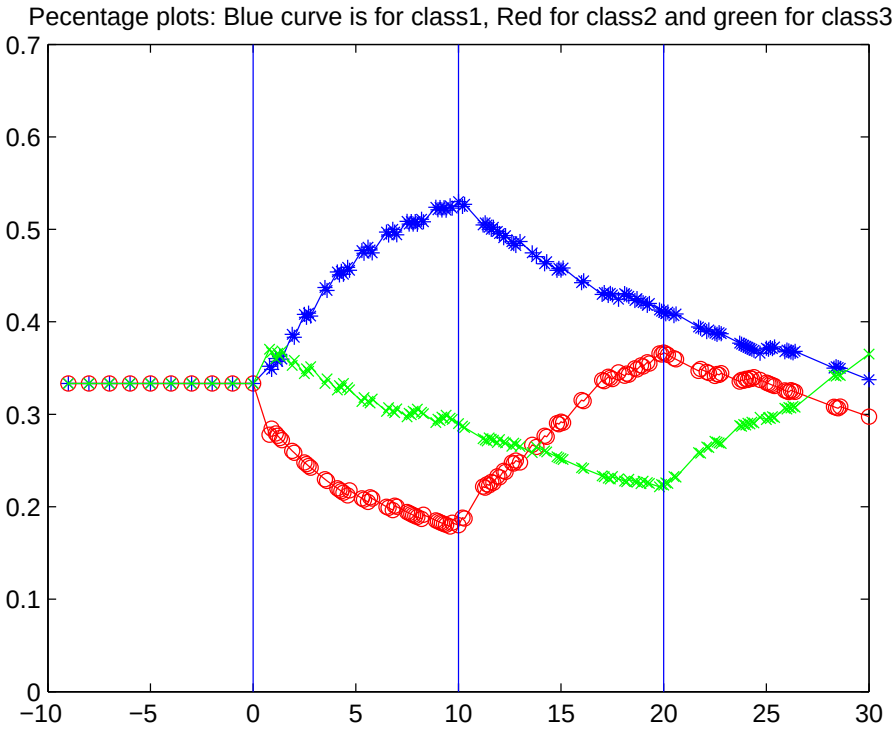
Παρατηρούμε ότι γενικά μπορούμε να εντοπίσουμε τις περιοχές της κάθε κλάσης από τα τμήματα που η αντίστοιχη καμπύλη ποσοστών είναι αύξουσα, δηλαδή έχει θετική κλίση. Αυτό συμβαίνει επειδή όταν προστίθεται ένα νέο τμήμα που ανήκει σε κάποια κλάση τα ποσοστά της κλάσης αυτής αυξάνονται ενώ τα ποσοστά των άλλων κλάσεων μειώνονται μέχρι το επόμενο σημείο αλλαγής. Εξαίρεση αποτελεί το τμήμα των καμπύλων από την αρχή μέχρι το πρώτο σημείο αλλαγής, όπου η κλάση του τμήματος αλλαγής έχει καμπύλη ποσοστών στο 1 με κλίση 0, και όλες οι υπόλοιπες κλάσεις έχουν καμπύλες ποσοστών στο 0 με κλίση 0. Αυτή η ειδική περίπτωση αντιμετωπίζεται εύκολα αν, πριν δημιουργήσουμε τις καμπύλες, προσθέσουμε κάποια λίγα dummy ποσοστά στην αρχή όλων των κλάσεων, τα οποία να είναι ίδια για όλες τις κλάσεις.

Στην εικόνα 5.6 βλέπουμε τις καμπύλες ποσοστών για το πείραμα που παρουσιάστηκε στην προηγούμενη ενότητα (περιγραφή του προβλήματος). Ο οριζόντιος άξονας μετράται σε seconds*0.1 (αντί για seconds) για να βλέπουμε τις καμπύλες με μεγαλύτερη ακρίβεια. Υπενθυμίζουμε ότι η κλάση 1 είναι στην περιοχή [0, 10] (sec*0.1), η κλάση 2 είναι στην περιοχή [10,20] (sec*0.1) και η κλάση 3 είναι στην περιοχή [20,30] (sec*0.1). Έχουμε προσθέσει τα ίδια dummy δεδομένα στην αρχή κάθε κλάσης και τα σημεία αλλαγής σημειώνονται πάνω στην κάθε καμπύλη ποσοστών (και δεν ισαπέχουν).

Παρατηρούμε ότι όντως τα τμήματα όπου η κάθε καμπύλη είναι (γενικά) αύξουσα ορί-



Εικόνα 5.5: Υπολογισμός των καμπύλων ποσοστών σε ένα απλό παράδειγμα.



Εικόνα 5.6: Υπολογισμός των καμπύλων ποσοστών για το πείραμα των 3 κλάσεων.Η κλάση 1(συμβολισμός με μπλε γραμμή και *) είναι στην περιοχή [0,10], η κλάση 2(συμβολισμός με κόκκινη γραμμή και o) στην περιοχή [10,20] και η κλάση 3 (συμβολισμός με πράσινη γραμμή και x) στην περιοχή [20,30].

ζουν τις υποπεριοχές του τμήματος που ανήκουν στην αντίστοιχη κλάση. Κατά συνέπεια, οι καμπύλες ποσοστών μπορούν όντως να μας δώσουν χρήσιμη πληροφορία για τα όρια και την κατηγοριοποίηση των υποτμημάτων, ακόμα και όταν το αρχικό transcription είναι υπερκατατμημένο.

Εξαγωγή Ορίων των Κλάσεων από τις Καμπύλες Ποσοστών

Παρουσιάζουμε εδώ μία μέθοδο και τον αντίστοιχο αλγόριθμο για την εξαγωγή χρήσιμης πληροφορίας σχετικά με τα όρια της κάθε κλάσης από την καμπύλη ποσοστών της.

Η ιδέα είναι να ομαλοποιήσουμε αρχικά τις καμπύλες για την κάθε κλάση και στη συνέχεια να πάρουμε την παράγωγό τους. Η ομαλοποίηση αυτή μπορεί να γίνει με ένα απλό median φιλτράρισμα. Στη συνέχεια, για την κάθε καμπύλη υπολογίζουμε την παράγωγό της και επιλέγουμε τα τμήματα που η παράγωγος αυτή είναι θετική. Με αυτό τον τρόπο, εξάγουμε τις περιοχές που η καμπύλη ποσοστών είναι αύξουσα, δηλαδή της περιοχές του τμήματος που ανήκουν στην αντίστοιχη κλάση.

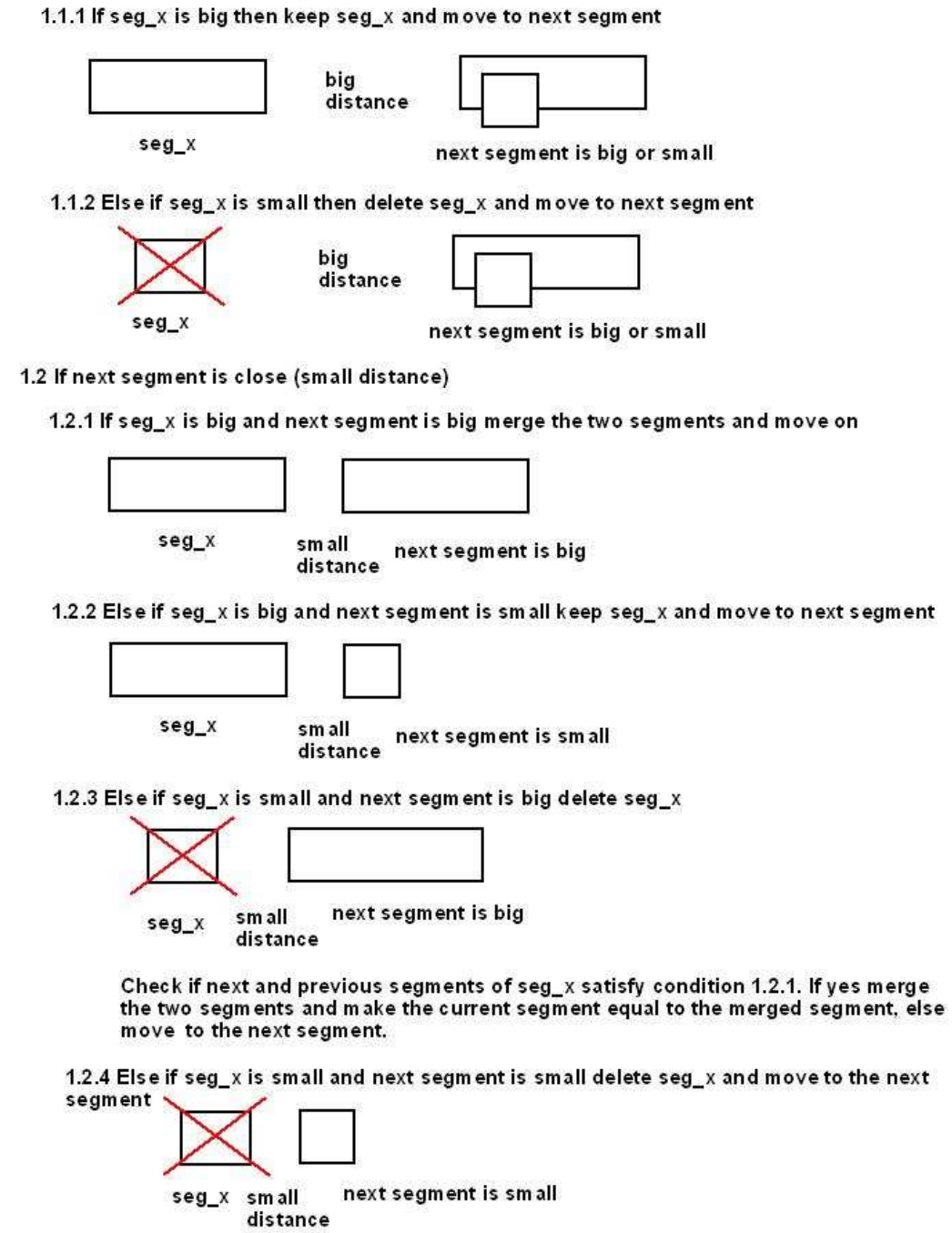
Εντούτοις, ενδέχεται να υπάρχουν μικρά αύξοντα τμήματα στις καμπύλες ποσοστών που να μην αντιστοιχούν σε πραγματικά τμήματα αλλά σε κάποια λανθασμένη κατηγοριοποίηση μικρού αριθμού frames. Αυτό έχει ως αποτέλεσμα να πάρουμε για μία κλάση, εκτός από τις πραγματικές περιοχές της και καποιες μικρές περιοχές που αντιστοιχούν σε σφάλματα κατάταξης. Για να αντιμετωπίσουμε αυτό το πρόβλημα χρησιμοποιούμε τον αλγόριθμο merge-delete ο οποίος παρουσιάζεται παρακάτω. Σκοπός του είναι να ομαλοποιήσει το αποτέλεσμα συγχωνεύοντας μεγάλα τμήματα που βρίσκονται πολύ κοντά μεταξύ τους και διαγράφοντας πολύ μικρά τμήματα.

Η λειτουργία του αλγορίθμου φαίνεται στο σχήμα 5.7.

Ο αλγόριθμος merge-delete είναι γραμμικός ως προς το πλήθος των τμημάτων αφού επεξεργάζεται σειριακά τα τμήματα που έχουμε εξάγει από την καμπύλη ποσοστών και αποφασίζει αν θα τα κρατήσει, αν θα τα διαγράψει ή αν θα τα συγχωνεύσει. Επίσης ο αλγόριθμος merge-delete απαιτεί τον ορισμό δύο κατωφλιών, ένα που καθορίζει το όριο μεταξύ μεγάλου και μικρού τμήματος και ένα που καθορίζει το όριο μεταξύ μεγάλης και μικρής απόστασης.

Παρατηρούμε ότι σε κάθε βήμα του αλγορίθμου τα τμήματα που βρίσκονται πίσω από το τρέχον τμήμα είναι μεγάλα και απέχουν μεγάλη απόσταση μεταξύ τους. Αυτό συμβαίνει επειδή για κάθε τρέχον τμήμα, το προηγούμενο τμήμα του, αν υπάρχει, είναι μεγάλο και βρίσκεται σε μεγάλη απόσταση από το τρέχον. Συνεπώς, επαγωγικά προς τα πίσω αποδεικνύεται η προηγούμενη πρόταση.

Χωρίζουμε με αυτόν τον τρόπο το audio-stream σε τμήματα. Τμήματα που ανήκουν



Εικόνα 5.7: Η λειτουργία του αλγορίθμου merge-delete για κάθε τμήμα που παίρνουμε από την επεξεργασία μίας καμπύλης ποσοστών.

στην ίδια κλάση είναι μεγάλα και απέχουν μεγάλη απόσταση μεταξύ τους. Τμήματα όμως που ανήκουν σε διαφορετικές κλάσεις μπορούν να είναι διαδοχικά, ή επικαλυπτόμενα ή σε κάποια απόσταση μεταξύ τους. Το πρόβλημα του συνδυασμού των τμημάτων που παράγει ο αλγόριθμος merge-delete, για όλες τις διαθέσιμες κλάσεις ώστε να παραχθεί ένα σωστό transcription εξετάζεται παρακάτω.

Συνδυασμός των Ορίων των Διαφορετικών Κλάσεων

Ο συνδυασμός των ο τμημάτων που παίρνουμε από την κάθε καμπύλη ποσοστών ώστε να παραχθεί ένα όσο το δυνατόν καλύτερο transcription είναι το πιο απαιτητικό κομμάτι του συστήματος κατάτμησης/κατηγοριοποίησης. Θέλουμε να παράγουμε transcription που να περιέχει τις υποπεριοχές που αναγνωρίστηκαν αλλά να μην περιέχει κενά ή επικαλύψεις. Πρέπει να αποφασιστεί η συμπεριφορά του συστήματος σε όλες τις δυνατές περιπτώσεις.

Η δυσκολία του προβλήματος συνδυασμού των επιμέρους περιοχών έγκειται στο ότι πρέπει να θεωρήσουμε την πιο γενική περίπτωση για να εξασφαλίσουμε ότι το σύστημα θα λειτουργεί ακόμα και σε κακές/παθολογικές περιπτώσεις. Στις συνήθεις περιπτώσεις τα όρια που βρίσκουμε για κάθε κλάση έχουν κάποιες μικρές επικαλύψεις ή κενά μεταξύ τους. Όμως σε περιπτώσεις που το σύστημα αδυνατεί να αναγνωρίσει σωστά τα δεδομένα εισόδου, οι καμπύλες ποσοστών μπορεί να περιέχουν λανθασμένη ή αντιφατική πληροφορία και να οδηγούν στην εξαγωγή περιοχών διαφορετικών κλάσεων που, αν συνδυαστούν, δημιουργούν μεγάλα κενά ή μεγάλες επικαλύψεις. Το σύστημα πρέπει να είναι αρκετά γενικό ώστε να χειρίζεται λογικά και τέτοιες δύσκολες περιπτώσεις αναγνώρισης. Περιοχές στις οποίες υπάρχουν επικαλυπτόμενα τμήματα ή κενά θα μπορούσαν να αναγνωριστούν ως κλάση no decision, δηλαδή αδυναμία αναγνώρισης. Εναλλακτικά, τα τμήματα αυτά θα μπορούσαν να καταταχθούν στην κλάση στην οποία ανήκει η πλειοψηφία των frames τους.

Θα παρουσιάσουμε τον αλγόριθμο συνδυασμού τον τμημάτων διαφορετικών κλάσεων ξεκινώντας από το χαμηλότερο επίπεδο αφαίρεσης και συνεχίζοντας σταδιακά προς τα πάνω.

Έστω ότι έχουμε ως δεδομένο ένα transcription με τις κλάσεις από 1 εως i-1, κάποια κενά τμήματα και πιθανόν κάποια από τα τμήματα της κλάσης i και θέλουμε να τοποθετήσουμε το επόμενο στη σειρά τμήμα της κλάσης i. Για το σκοπό αυτό χρησιμοποιούνται δύο συναρτήσεις, η putPosition_usual και η putPosition_sameStart. Η μόνη παραδοχή που μπορούμε να κάνουμε είναι ότι τα ήδη υπάρχοντα τμήματα που ανήκουν σε κάποια κλάση είναι μεγάλα καθώς προέρχονται από τη merge-delete που κρατάει μόνο μεγάλα τμήματα και επίσης οι αλγόριθμοι(putPosition_usual και putPosition_sameStart) δημιουργούν μόνο μεγάλα τμήματα(όπως θα φανεί στην περιγραφή τους).

Η putPosition_usual χρησιμοποιείται για την γενικότερη περίπτωση όπου δεν υπάρχει

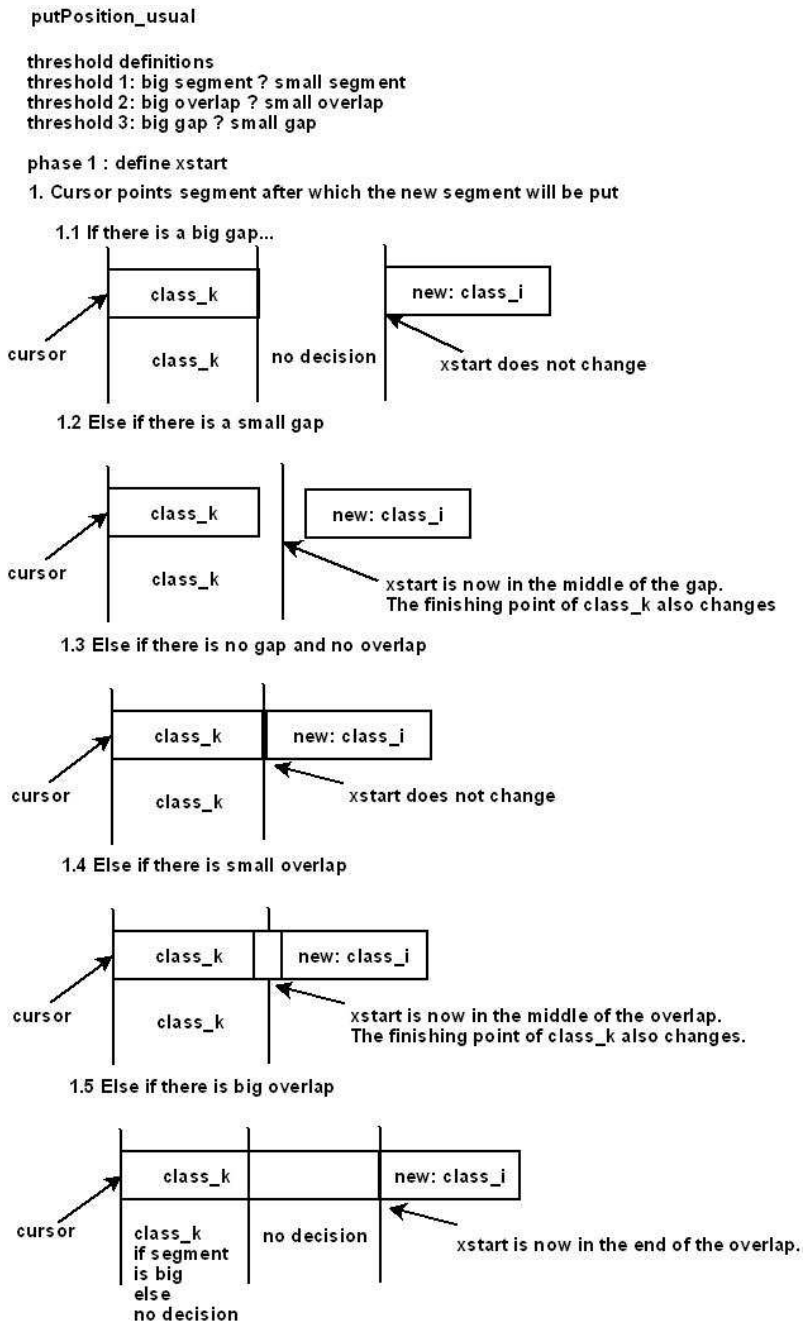
κάποια άλλη κλάση στο ήδη υπάρχον Transcription που να αρχίζει ακριβώς στο σημείο που αρχίζει και το καινούριο τμήμα που θέλουμε να εισάγουμε. Η putPosition_usual αποτελείται από 3 διαδοχικές φάσεις που περιγράφονται στα σχήματα 5.8, 5.9 και 5.10. Η πρώτη φάση έχει να κάνει με τον καθορισμό του σημείου αρχής της καινούριας περιοχής, η δεύτερη φάση έχει να κάνει με την ύπαρξη άλλων κλάσεων στο εσωτερικό της καινούριας περιοχής και η τρίτη φάση καθορίζει το τελικό σημείο της καινούριας περιοχής.

Η putPosition_sameStart χρησιμοποιείται για την ειδική περίπτωση όπου υπάρχει κάποια άλλη κλάση στο ήδη υπάρχον Transcription που να αρχίζει ακριβώς στο σημείο που αρχίζει και το καινούριο τμήμα που θέλουμε να εισάγουμε. Η putPosition_sameStart περιγράφεται στα σχήματα 5.11 και 5.12.

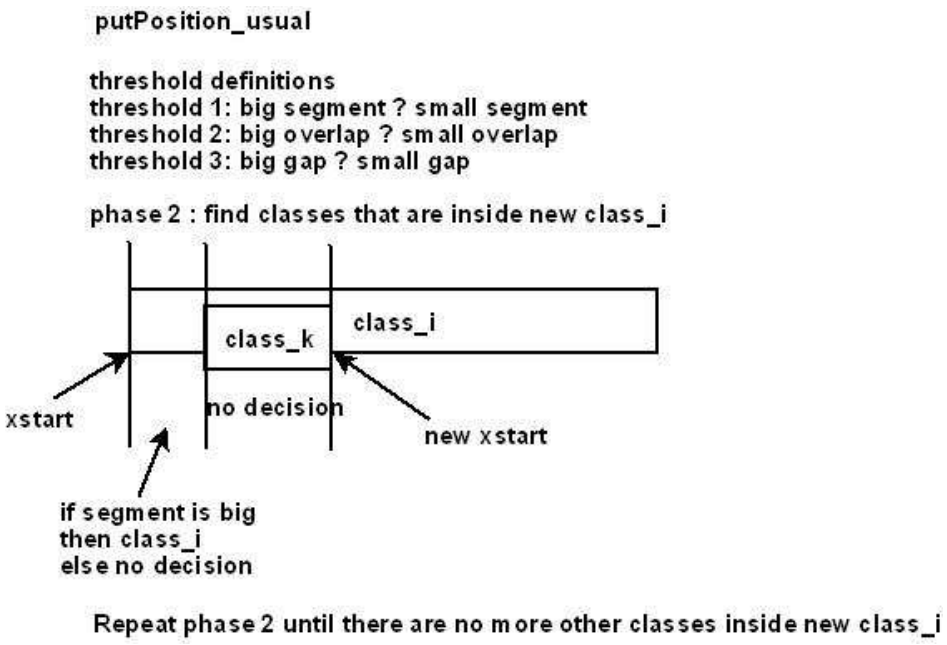
Έχοντας υλοποιήσει τις παραπάνω συναρτήσεις,μπορούμε να δημιουργήσουμε τη γενικότερη συνάρτηση putPosition, οι οποία συνδυάζει τις 2 παραπάνω. Ακολουθεί ο ψευδοκώδικας, όπου με space συμβολίζουμε το ήδη υπάρχον transcription. Ο δείκτης cursor δείχνει κάθε φορά το τμήμα μετά το οποία θα τοποθετηθεί το καινούριο τμήμα.

```
[space,cursor]=putPosition(xstart,xfinish,xclass,space,cursorStart)

Αρχίζοντας από το cursorStart αύξησε τον cursor και διέσχισε το
Transcription space για να βρεις το κατάλληλο σημείο να τοποθετήσεις
το νέο τμήμα.

Αν ο cursor δείξει τμήμα που να αρχίζει στο ίδιο σημείο με το νέο
τμήμα τότε
space=putPosition_sameStart(xstart,xfinish,xclass,cursor,space)

Αλλιώς βρες το πρώτο υπάρχον τμήμα με μεγαλύτερο xstart, κάνε
cursor=cursor-1 και
space=putPosition_usual(xstart,xfinish,xclass,cursor,space) (Ομοίως
και στην περίπτωση που δεν υπάρχει τμήμα με μεγαλύτερο xstart)

Επέστρεψε τις μεταβλητές space και cursor
```

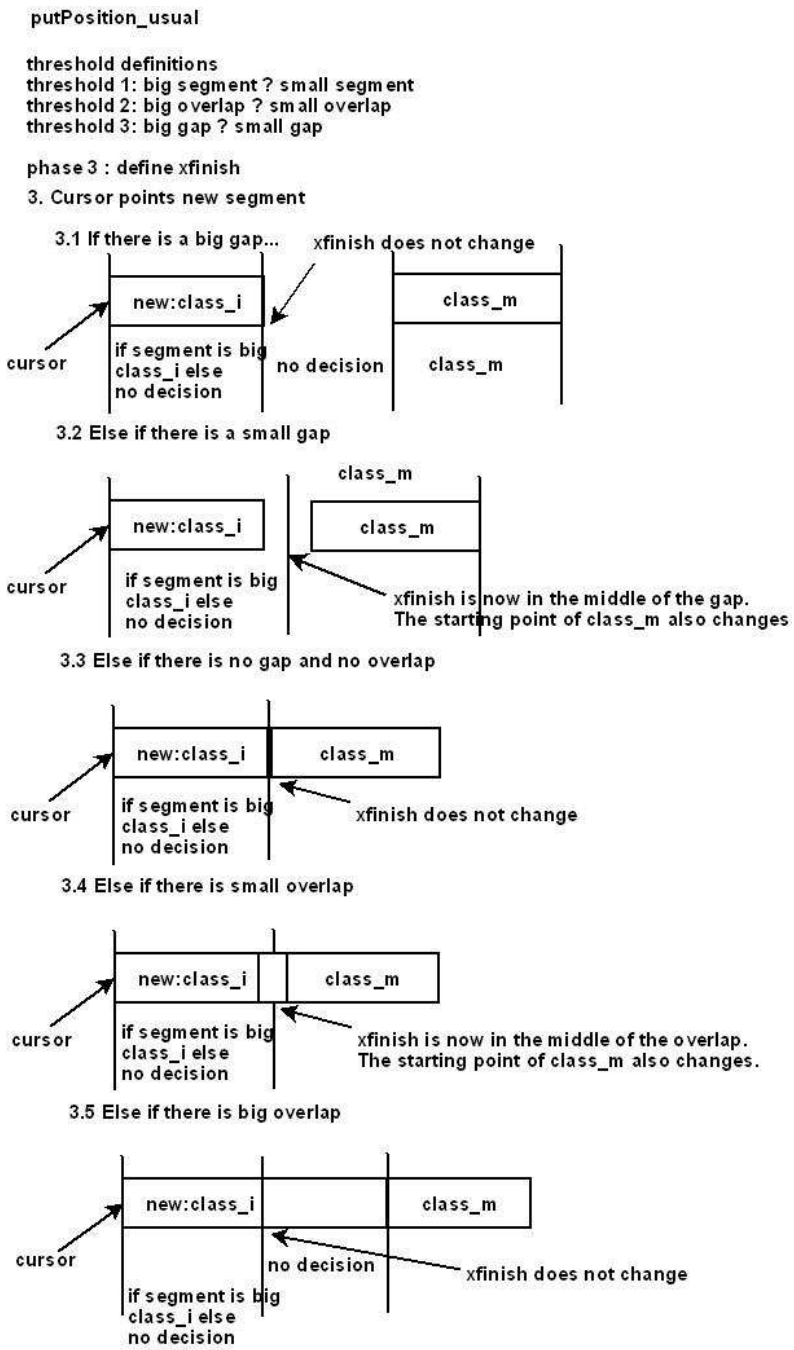
Στη συνέχεια μπορούμε να ορίσουμε μία συνάρτηση, για παράδειγμα την putClass που θα τοποθετεί σειριακά τα τμήματα μίας κλάσης i, σε ένα ήδη υπάρχον Transcription που περιέχει τις κλάσεις 1 εώς i-1. Η συνάρτηση putClass θα χρησιμοποιεί εσωτερικά την putPosition. Αξίζει να αναφερθεί πιο αναλυτικά η λειτουργία της μεταβλητής cursor. Κάθε



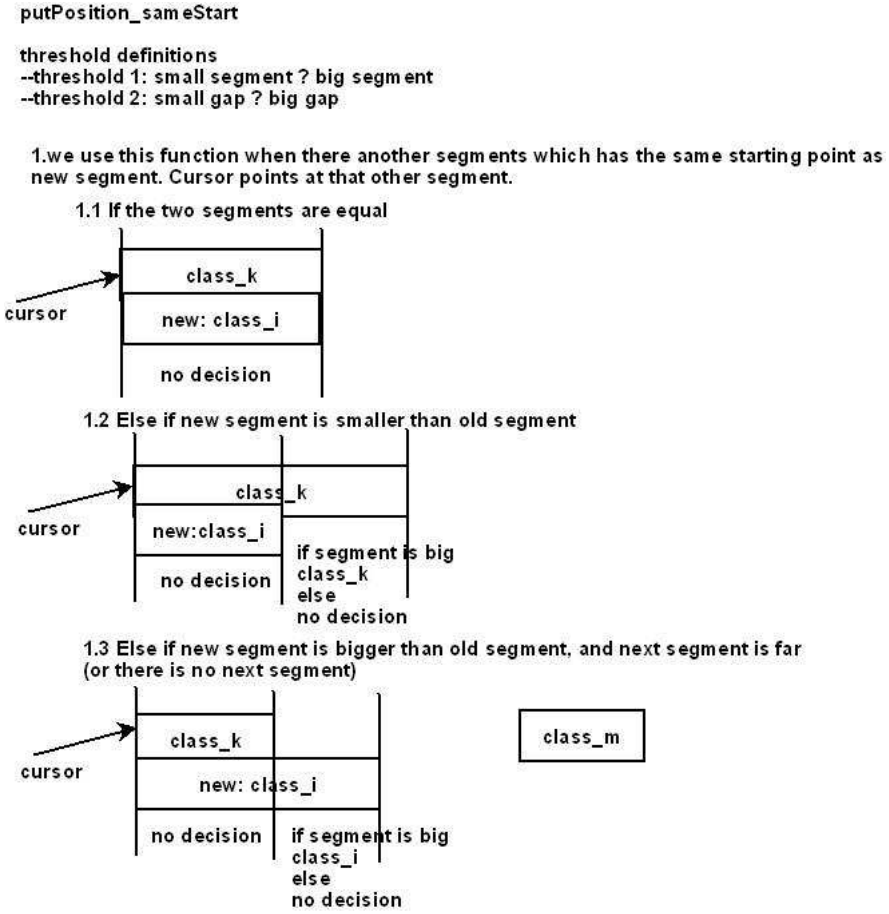
Εικόνα 5.8: Η πρώτη φάση του αλγορίθμου putPosition_usual.



Εικόνα 5.9: Η δεύτερη φάση του αλγορίθμου putPosition_usual.



Εικόνα 5.10: Η τρίτη φάση του αλγορίθμου putPosition_usual.



Εικόνα 5.11: Η λειτουργία του αλγορίθμου putPosition_sameStart.

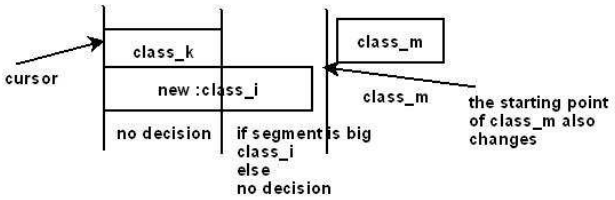
putPosition_sameStart

threshold definitions
--threshold 1: small segment ? big segment
--threshold 2: small gap ? big gap

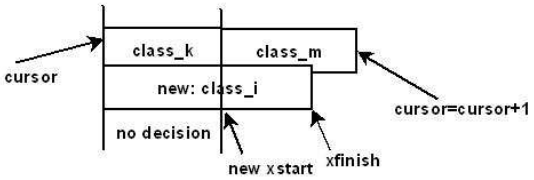
1.we use this function when there another segments which has the same starting point as new segment. Cursor points at that other segment.

1.1-1.3... (previous diagram)

1.4 Else if new segment is bigger than old segment and next segment has a small gap with new segment

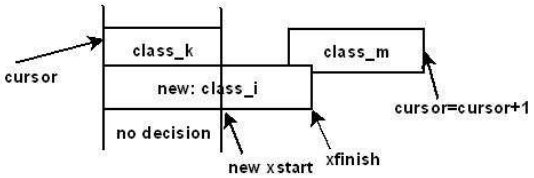


1.5 Else if new segment is bigger than old segment and next segment is exactly next to old segment (special case)



we call again putPosition_sameStart with arguments the new xstart, xfinish and the new cursor...

1.6 Else if new segment is bigger than old segment and next segment has an overlap with new segment



we call putPosition_usual with arguments the new xstart, xfinish and the new cursor...

Εικόνα 5.12: Η λειτουργία του αλγορίθμου putPosition_sameStart (συνέχεια).

φορά που η putPosition επιστρέφει, η cursor δείχνει το τμήμα του transcription μετά το οποίο τοποθετήσαμε το νέο τμήμα. Έτσι, αν θέλουμε να τοποθετήσουμε το επόμενο διαθέσιμο τμήμα της κλάσης i και ψάχνουμε για την κατάλληλη θέση, αρκεί να ψάξουμε από το σημείο cursor και μετά (καθώς τα τμήματα της κάθε κλάσης είναι με αύξουσα σειρά, όπως δίνονται από τη merge-delete). Δηλαδή, δεν είναι απαραίτητο να αρχίσουμε να ψάχνουμε το transcription από την αρχή και αυτό μειώνει την πολυπλοκότητα της putClass από τετραγωνική σε γραμμική. Κατά συνέπεια, η PutClass ειναι $O(n)$, όπου n ο αριθμός των συνολικών τμημάτων του transcription στο συγκεκριμένο πέρασμα (σημειώνουμε ότι ο αριθμός των συνολικών τμημάτων αυξάνει καθώς προσθέτουμε νέες κλάσεις).

Αν θεωρήσουμε τώρα τη συνάρτηση Merger που χρησιμοποιεί εσωτερικά την putClass για να τοποθετήσει σειριακά τις κλάσεις στο transcription, καταλαβαίνουμε ότι η πολυπλοκότητα της merger είναι στη χειρότερη περίπτωση $O(n * m)$,όπου m το πλήθος των κλάσεων και n το μέγιστο πλήθος τμημάτων του transcription (δηλαδή το πλήθος των τμημάτων μετά την προσθήκη της τελευταίας κλάσης). Όμως δεδομένου ότι το σύστημά μας έχει σχεδιαστεί ώστε κάθε φορά να επιλέγει μεταξύ μικρού αριθμού κλάσεων, δηλαδή m=2,3 ή 4, μπορούμε να θεωρήσουμε προσεγγιστικά ότι η Merger είναι $O(n)$. Συνεπώς, η συνάρτηση συνδυασμού των κλάσεων που περιγράψαμε είναι υπολογιστικά αποδοτική.

Τέλος, σημειώνουμε ότι δεδομένου ότι η Merger δέχεται τμήματα κλάσεων από την merge-delete, δηλαδή τα τμήματα της κάθε κλάσης είναι μεγάλα και απέχουν μεγάλη απόσταση μεταξύ τους, παράγει transcription που πληρεί κάποιες προυποθέσεις. Οι προυποθέσεις αυτές εξασφαλίζονται από το σχεδιασμό των putPosition_usual και putPosition_sameStart. Συγκεκριμένα στο παραγόμενο transcription υπάρχουν μόνο μεγάλα τμήματα καθώς τα μικρά τμήματα έχουν συγχωνευθεί με τα γειτονικά τους. Επίσης τυχόν διαδοχικά τμήματα που ανήκουν στην ίδια τάξη έχουν συγχωνευτεί στο ίδιο τμήμα. Τέλος, ακόμα και τα no decision τμήματα είναι μεγάλα. Εντούτοις, το transcription Μπορεί να έχει κάποιες ατέλειες όπως το να περιέχει μεγάλα κενά, τμήματα δηλαδή που δεν έχουν καταταγεί σε κάποια κλάση ούτε στη no decision. Η αντιμετώπιση τέτοιων περιπτώσεων παρουσιάζεται στην επόμενη ενότητα.

Τελική Επεξεργασία των Ορίων

Η συνάρτηση Merger που δημιουργήσαμε, μας δίνει transcriptions τα οποία μπορούν να περιέχουν κάποιες ατέλειες. Για παράδειγμα το transcription θα μπορούσε να περιέχει κάποιο μεγάλο τμήμα που να είναι κενό, δηλαδή να μην ανήκει σε κάποια κλάση αλλά και ούτε να είναι no decision. Ας δούμε μία περίπτωση που κάτι τέτοιο μπορεί να συμβεί.

Έστω ότι υπάρχει μία περιοχή του σήματος στην οποία το σύστημα μας δεν μπορεί

να αποφασίσει την κλάση και κατατάσσει τα διαδοχικά frames σε διαφορετικές κλάσεις. Αυτό θα δημιουργήσει καμπύλες ποσοστών που σε αυτή την περιοχή θα έχουν μικρά σκαμπανεβάσματα και η επεξεργασία των καμπύλων θα δημιουργήσει μικρά τμήματα για κάθε κλάση σε αυτή την περιοχή. Στη συνέχεια, ο αλγόριθμος merge-delete θα διαγράψει αυτά τα μικρά τμήματα των κλάσεων σε αυτή την περιοχή και τελικά δεν θα υπάρχει τμήμα των διαθέσιμων κλάσεων που να αντιστοιχεί σε αυτή την περιοχή. Κατά συνέπεια ο merger θα παράγει transcription με μεγάλο κενό (αν το κενό ήταν μικρό, ο merger θα το συγχώνευε με τα γειτονικά του).

Έτσι είναι απαραίτητο να κάνουμε ένα επιπλέον πέρασμα ώστε να εντοπίσουμε κενά στο transcription και να τα κατατάξουμε no decision.

Στη συνέχεια, πρέπει να αποφασίσουμε πως θα χειριστούμε τα τμήματα no decision του παραγόμενου transcription. Η απλούστερη λύση είναι να τα αφήσουμε ως έχουν ώστε το τελικό αποτέλεσμα να περιέχει τμήματα που δεν κατατάχθηκαν πουθενά. Μία άλλη λύση είναι για κάθε no decision τμήμα να μετρήσουμε πόσα frames του κατατάχθηκαν σε κάθε κλάση και να κατατάξουμε το τμήμα στην κλάση που πλειοψηφεί.

Συμπεριφορά του Αλγορίθμου σε ένα παράδειγμα

Το παράδειγμα με τις 3 κλάσεις, με βάση το οποίο παρουσιάσαμε τις καμπύλες ποσοστών, μπορεί να θεωρηθεί, ως προς το βαθμό δυσκολίας του, ως μία μέση περίπτωση για το σύστημα. Θα εξετάσουμε πως συμπεριφέρεται το σύστημα για την περίπτωση αυτή. Υπενθυμίζουμε ότι το σωστό transcription είναι class1 για [0,1sec], class2 για [1,2sec] και class3 για [2,3sec].

Η επεξεργασία των καμπύλων ποσοστών μας δίνει τα διαστήματα:

class 1 Διαστήματα [0, 0.96 sec] και [2.5, 2.53 sec]

class 2 Διάστημα [0.94, 2 sec] και το σημείο 2.41sec

class 3 Διαστήματα [0, 0.12 sec], [0.76, 0.77 sec] και [1.99, 3sec]

Ο αλγόριθμος merge-delete παίρνει τα παραπάνω διαστήματα και μας δίνει τα διαστήματα:

class 1 Διάστημα [0, 0.96 sec]

class 2 Διάστημα [0.94, 2 sec]

class 3 Διάστημα [1.99, 3sec]

Ο αλγόριθμος merger μας δίνει τελικά το transcription:

class 1 Διάστημα [0 0.95sec]

class 2 Διάστημα [0.95, 1.995sec]

class 3 Διάστημα [1.995, 3sec]

Παρατηρούμε ότι το αποτέλεσμα του Merger δεν έχει καν ανάγκη του τελικού βήματος προσαρμογής των ορίων. Έχουμε επιτύχει το στόχο μας σε μεγάλο βαθμό καθώς από το αρχικό υπερκατατμημένο Transcription εισόδου παράγαμε ένα transcription με ελάχιστη απόσταση από το σωστό και χωρίς περιττές κατατμήσεις.

Σχολιασμός

Η παραπάνω ανάλυση αλλά και τα πειραματικά αποτελέσματα που θα παρουσιαστούν στην αντίστοιχη ενότητα δείχνουν ότι το σύστημα έχει πολλές δυνατότητες, καθώς μπορεί δεχόμενο ένα υπερκατατμημένο transcription να παράγει ένα σωστό transcription με υπολογιστικά αποδοτικό τρόπο, δηλαδή γραμμικό ως προς το συνολικό πλήθος των τμημάτων του transcription.

Εντούτοις, υπάρχουν πολλά σημεία του συστήματος τα οποία πρέπει να μελετηθούν και να βελτιωθούν. Για παράδειγμα οι αλγόριθμοι των σταδίων εξαγωγής περιοχών από τις καμπύλες ποσοστών και συνδυασμού των περιοχών χρησιμοποιούν πολλά κατώφλια για να καθορίσουν αν ένα τμήμα είναι μεγάλο ή μικρό, αν μία απόσταση είναι μεγάλη ή μικρή, αν μία επικάλυψη είναι μεγάλη η μικρή, κλπ. Στα πειράματα μας ορίσαμε τα κατώφλια αυτά στα 0.4sec, όμως θα έπρεπε να γίνει προσπάθεια να βρεθεί εμπειρικά κάποια καλή τιμή ή θα έπρεπε να σχεδιαστεί ένας τρόπος αυτόματου υπολογισμού των κατωφλιών.

Επιπλέον, σε περιπτώσεις που έχουμε μικρή επικάλυψη ή μικρό κενό μεταξύ δύο διαδοχικών κλάσεων, αναπροσαρμόζουμε τα όριά τους, ώστε το σημείο που διαχωρίζει τις κλάσεις να βρίσκεται στη μέση της επικάλυψης ή του κενού. Αυτή ή επιλογή δεν είναι απαραίτητα η καλύτερη δυνατή και να μπορούσαμε να δοκιμάσουμε να τοποθετήσουμε το σημείο αλλαγής με κάποιο άλλο κριτήριο. Για παράδειγμα θα μπορούσαμε να το τοποθετήσουμε σε ένα σημείο της περιοχής αυτής που το σήμα έχει ελάχιστη ενέργεια, ώστε να μην κόψουμε στη μέση φράσεις ή άλλη ακουστική πληροφορία.

Τέλος, η απόφαση που παίρνουμε για τα τμήματα no decision θα μπορούσε να είναι πιο σύνθετη από μία απλή πλειοψηφία των frames. Ανάλογα με το πρόβλημα, η κάθε μία από τις διαθέσιμες κλάσεις μπορεί να έχει συχνότητα εμφάνισης και βαρύτητα. Για παράδειγμα

στο πρόβλημα των δελτίων ειδήσεων η κλάση της φωνής έχει περισσότερη βαρύτητα από τη κλάση του θορύβου. Θα προτιμούσαμε να κατατάξουμε λανθασμένα θόρυβο ως φωνή παρά να κατατάξουμε λανθασμένα φωνή ως θόρυβο, επειδή στην τελευταία περίπτωση θα καταλήγαμε να χάσουμε τμήματα ομιλίας. Έτσι θα μπορούσαμε με χρήση βαρών να ευνοήσουμε την επιλογή κάποιας σημαντικής ή συχνής κλάσης, κατά την εξέταση ενός no decision τμήματος, εις βάρος άλλων λιγότερο σημαντικών ή πιο σπάνιων κλάσεων.

5.5.3 Χρήση Median Filetring για ομαλοποίηση του αποτελέσματος του HMM classifier

Στην ενότητα αυτή αναφέρεται μία πολύ απλή ιδέα για την ομαλοποίηση του υπερκατατμημένου αποτελέσματος που παράγεται από το σύστημα κατάταξης με GMM μοντέλα. Συγκεκριμένα, μπορούμε να εφαρμόσουμε median φιλτράρισμα στο υπερκατατμημένο αποτέλεσμα με σκοπό να συγχωνεύσουμε μικρά τμήματα με γειτονικά μεγάλα τμήματα. Η μέθοδος αυτή παρέχει έναν εύκολο και απλό τρόπο ομαλοποίησης του αποτελέσματος, όμως δεν μπορεί να χειριστεί αποτελεσματικά δύσκολες περιπτώσεις στις οποίες έχουμε διαδοχικά μικρά τμήματα τα οποία ανήκουν σε διαφορετικές κλάσεις. Τα τμήματα αυτά δεν μπορούν να συγχωνευθούν και έτσι παραμένουν ως έχουν. Κατά συνέπεια το τελικό transcription μπορεί να είναι υπερκατατμημένο.

5.5.4 Συνδυασμός Median Filtering και Καμπύλων Ποσοστών

Οι δύο μέθοδοι που παρουσιάστηκαν παραπάνω, δηλαδή η τεχνική median filtering και η χρήση καμπύλων ποσοστών, μπορούν να συνδυαστούν ώστε να επιτύχουμε το καλύτερο δυνατό αποτέλεσμα. Συγκεκριμένα μπορούμε να εφαρμόσουμε Median filtering στο αποτέλεσμα των GMM Μοντέλων και στη συνέχεια να υπολογίσουμε τις καμπύλες ποσοστών από το ομαλοποιημένο αποτέλεσμα. Στη συνέχεια μπορούν να εφαρμοστούν χωρίς καμία αλλαγή οι αλγόριθμοι που παρουσιάστηκαν στην ενότητα της χρήσης καμπύλων ποσοστών, ώστε να παράγουμε το Transcription εξόδου.

Το ερώτημα που τίθεται είναι κατά πόσο η χρήση των καμπύλων ποσοστών και των σχετικών αλγορίθμων MergeDelete και Merger βελτιώνουν την απόδοση που θα είχαμε με ένα απλό median filtering. Τα πειραματικά αποτελέσματα σε πραγματικά δεδομένα που παρουσιάζονται στην αντίστοιχη ενότητα δείχνουν ότι πράγματι η χρήση της μεθόδου με median filtering και με τις καμπύλες ποσοστών βελτιώνει τα ποσοστά επιτυχίας σε σχέση με το απλό median filtering και επίσης παράγει transcriptions με λιγότερες κατατμήσεις και χωρίς υπερβολικά μικρά υποτμήματα.

5.6 Πειραματικά Αποτελέσματα

Στην ενότητα αυτή παρουσιάζονται αποτελέσματα των πειραμάτων που έγιναν τόσο σε συνθετικά δεδομένα όσο και σε πραγματικά δεδομένα από δελτία ειδήσεων.

5.6.1 Πειράματα σε Συνθετικά Δεδομένα

Στην ενότητα αυτή περιγράφονται πειράματα που έγιναν σε συνθετικά δεδομένα. Συγκεκριμένα, αντί να χρησιμοποιήσουμε την HCopy για να εξάγουμε χαρακτηριστικά από πραγματικά audio-streams, δημιουργούμε μέσω του Matlab ένα αρχείο παρόμοιο με αυτό που θα δημιουργούσε η HCopy που περιέχει συνθετικά χαρακτηριστικά. Δηλαδή περιέχει χαρακτηριστικά που έχουν παραχθεί με τυχαίο τρόπο έτσι ώστε να ανήκουν σε γκαουσιανές κατανομές με συγκεκριμένη μέση τιμή και διακύμανση. Σε πρώτη φάση εκπαιδεύουμε κατάλληλα μοντέλα με αυτά τα συνθετικά δεδομένα και στη συνέχεια χρησιμοποιούμε δεδομένα που έχουν παραχθεί με τον ίδιο τρόπο για να εξετάσουμε την απόδοση του συστήματος. Όπως περιγράφηκε σε προηγούμενη ενότητα, για κάθε κλάση που δημιουργούμε εκπαιδεύουμε ένα GMM και στη συνέχεια συγχωνεύουμε τα GMMs σε ένα HMM και κάνουμε decoding με alignment.

Τα πειράματα σε συνθετικά δεδομένα και σε ελεγχόμενο πειραματικό περιβάλλον έχουν ως σκοπό να βοηθήσουν στην καλύτερη κατανόηση της λειτουργίας του συστήματος και των αλγορίθμων κατάτμησης και κατηγοριοποίησης που παρουσιάστηκαν.

Θεωρούμε ότι στο πρόβλημά μας έχουμε 3 κλάσεις στις οποίες θέλουμε να κατατάξουμε τα υποτμήματα του audio-stream υπό εξέταση. Οι κλάσεις αυτές παράγονται συνθετικά μέσω του Matlab και τα χαρακτηριστικά τους φαίνονται αναλυτικά παρακάτω.

class1 Η κλάση 1 περιέχει 4 γκαουσιανές κατανομές

1. Μέση τιμή 0 και διακύμανση 1
2. Μέση τιμή 1 και διακύμανση 1
3. Μέση τιμή -1 και διακύμανση 0.5
4. Μέση τιμή 1.5 και διακύμανση 0.5

class2 Η κλάση 2 περιέχει 4 γκαουσιανές κατανομές

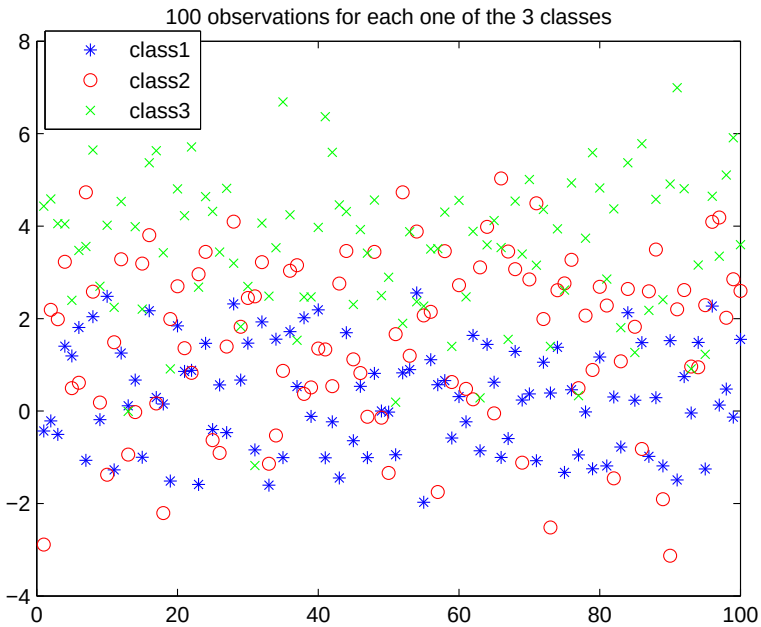
1. Μέση τιμή 0 και διακύμανση 2
2. Μέση τιμή 1 και διακύμανση 2

3. Μέση τιμή 2 και διακύμανση 1
4. Μέση τιμή 3 και διακύμανση 1

class3 Η κλάση 3 περιέχει 2 γκαουσιανές κατανομές

1. Μέση τιμή 3 και διακύμανση 2
2. Μέση τιμή 4 και διακύμανση 1

Το πρόβλημα διαχωρισμού των κλάσεων αυτών είναι αρκετά δύσκολο καθώς υπάρχει επικάλυψη μεταξύ τους. Στο σχήμα 5.13 φαίνονται 100 παρατηρήσεις από την κάθε κλάση. Παρατηρούμε ότι υπάρχει επικάλυψη μεταξύ των κλάσεων 1 και 2 και μεταξύ των κλάσεων 2 και 3, συνεπώς οι κλάσεις αυτές έχουν αρκετά μεγάλη δυσκολία διαχωρισμού.



Εικόνα 5.13: Στην εικόνα φαίνονται 100 παρατηρήσεις από κάθε μία από τις 3 κλάσεις του προβλήματος. Με μπλέ * σημειώνεται η κλάση 1, με κόκκινο ο η κλάση 2 και με πράσινο x η κλάση 3

Επιλέγουμε ένα audio-stream δοκιμής το οποίο θα χρησιμοποιηθεί σε όλα τα πειράματα που θα γίνουν. Το audio-stream δοκιμής έχει μήκος 5sec και το transcription του φαίνεται παρακάτω:

class 1 [0, 1sec]

class 2 [1, 2sec]

class 3 [2, 3.5sec]

class 2 [3.5, 4sec]

class 1 [4, 5sec]

Το transcription που επιλέξαμε είναι αρκετά δύσκολο καθώς οι μεταβάσεις είναι μεταξύ κλάσεων με μεγάλη επικάλυψη.

Αποτελέσματα του HMM/GMM Συστήματος Αναγνώρισης

Η απόδοση μετράται βρίσκοντας τα ποσοστά %Correct και %Accurate. Τα ποσοστά αυτά βρίσκονται με τη βοήθεια της συνάρτησης του HTK HResults, όπου γίνεται σύγκριση του transcription εξόδου του συστήματος με ένα transcription αναφοράς. Το κάθε ένα από τα transcriptions περιέχει για κάθε διαδοχικό υποτμήμα των 10msec του δελτίου την αντίστοιχη κατάταξη (πχ class1 ή class2).Ο ορισμός των ποσοστών %Correct και %Accurate είναι:

$$\%Correct = \frac{H}{N} \times 100\%$$

$$\%Accurate = \frac{H - I}{N} \times 100\%$$

όπου N είναι το συνολικό πλήθος των υποτμημάτων που εξετάζονται στο transcription, H είναι το πλήθος των υποτμημάτων που κατατάχθηκαν σωστά και I είναι το συνολικό πλήθος των τμημάτων που κατατάχθηκαν σε λάθος κατηγορία (Insertions).

Πείραμα 1 Στο πείραμα 1 χρησιμοποιούμε 50frames, που αντιστοιχούν σε χρόνο 0.5sec για την εκπαίδευση της κάθε κλασης. Στη συνέχεια βλέπουμε πως μεταβάλλεται η απόδοση του συστήματος αναγνώρισης, μεταβάλλοντας τον αριθμό των γκαουσιανών κατανομών που θεωρούμε για το κάθε μοντέλο που εκπαιδεύουμε. Συγκεκριμένα στο πείραμα (1α) έχουμε εκπαιδεύσει όλα τα μοντέλα με μία γκαουσιανή κατανομή, στο πείραμα (1β) έχουμε εκπαιδεύσει όλα τα μοντέλα με 2 γκαουσιανές κατανομές, στο πείραμα (1γ) με 4 γκαουσιανές και στο πείραμα (1δ) με 8 γκαουσιανές. Τα αποτελέσματα φαίνονται στον πίνακα 5.2.

Πίνακας 5.2: Αποτελέσματα των 4 πειραμάτων (1α),(1β),(1γ) και (1δ). Για την εκπαίδευση του κάθε μοντέλου χρησιμοποιούμε 50 παρατηρήσεις από την αντίστοιχη κλάση. Στο πείραμα α η μοντελοποίηση έγινε με 1 γκαουσιανή, στο β με 2 γκαουσιανές, στο γ με 4 γκαουσιανές και στο δ με 8 γκαουσιανές

experiment 1a			
Class Name	Hits	FAs	Actual
class1	153	47	192
class2	69	81	137
class3	124	26	167
Overall	346	154	496
%Correct	70.77		
%Accurate	68.75		
experiment 1b			
Class Name	Hits	FAs	Actual
class1	161	39	191
class2	98	52	164
class3	119	31	144
Overall	378	122	499
%Correct	75.95		
%Accurate	75.75		
experiment 1c			
Class Name	Hits	FAs	Actual
class1	156	44	183
class2	98	52	174
class3	117	33	141
Overall	371	129	498
%Correct	74.90		
%Accurate	74.30		
experiment 1d			
Class Name	Hits	FAs	Actual
class1	139	61	161
class2	80	70	157
class3	131	19	181
Overall	350	150	499
%Correct	70.34		
%Accurate	69.94		

Παρατηρούμε, ότι τα ποσοστά επιτυχίας και ακρίβειας είναι αρκετά χαμηλά, γεγονός αναμενόμενο λόγω του μικρού πλήθους δεδομένων εκπαίδευσης. Η αύξηση των γκαουσιανών από 1 σε 2 οδηγεί σε μία αισθητή βελτίωση, καθώς η μοντελοποίηση των κλάσεων του προβλήματος γίνεται με καλύτερο τρόπο. Εντούτοις, η περαιτέρω αύξηση των γκαουσιανών κατανομών δεν οφελεί αλλά βλάπτει την απόδοση. Ιδιαίτερα στην περίπτωση των 8 γκαουσιανών τα ποσοστά είναι χαμηλά. Αυτό συμβαίνει όχι μόνο επειδή οι γκαουσιανές είναι περισσότερες από τις πραγματικές γκαουσιανές των κλάσεων αλλά κυρίως επειδή δεν έχουμε αρκετά δεδομένα για τον αξιόπιστο υπολογισμό των παραμέτρων τους. Κατά συνέπεια, γίνεται overtraining των μοντέλων στα δεδομένα εκπαίδευσης και το σύστημα δεν είναι αρκετά γενικό ώστε να κατατάξει σωστά το audio-stream δοκιμής. Συνολικά, τα καλύτερα αποτελέσματα τα έχουμε για 2 γκαουσιανές ενώ τα χειρότερα για 1 και 8 γκαουσιανές.

Πείραμα 2 Στο πείραμα 2 χρησιμοποιούμε 100frames, που αντιστοιχούν σε χρόνο 1sec για την εκπαίδευση της κάθε κλασης. Στη συνέχεια βλέπουμε πως μεταβάλλεται η απόδοση του συστήματος αναγνώρισης, μεταβάλλοντας τον αριθμό των γκαουσιανών κατανομών που θεωρούμε για το κάθε μοντέλο που εκπαιδεύουμε. Συγκεκριμένα στο πείραμα (2α) έχουμε εκπαιδεύσει όλα τα μοντέλα με μία γκαουσιανή κατανομή, στο πείραμα (2β) έχουμε εκπαιδεύσει όλα τα μοντέλα με 2 γκαουσιανές κατανομές, στο πείραμα (2γ) με 4 γκαουσιανές και στο πείραμα (2δ) με 8 γκαουσιανές. Τα αποτελέσματα φαίνονται στον πίνακα 5.3.

Παρατηρούμε ότι για την περίπτωση της 1 γκαουσιανής κατανομής τα ποσοστά είναι ιδαίτερα χαμηλά, καθώς μία κατανομή δεν μπορεί να μοντελοποιήσει ικανοποιητικά κλάσεις που έχουν 2 ή 4 κατανομές. Εντούτοις, αυξάνοντας τις γκαουσιανές παρατηρούμε μεγάλη βελτίωση των ποσοστών επιτυχίας και ακρίβειας. Τα ποσοστά αυτά είναι καλύτερα από αυτά του πειράματος 1, καθώς στο πείραμα 2 έχουμε περισσότερα δεδομένα για τον αξιόπιστο υπολογισμό των παραμέτρων. Το καλύτερο αποτέλεσμα το έχουμε για 4 γκαουσιανές ενώ έχουμε καλά αποτελέσματα και στην περίπτωση των 2 γκαουσιανών. Για 8 γκαουσιανές, τα αποτελέσματα είναι πιο χαμηλά κυρίως λόγω της έλλειψης επαρκούς πλήθους δεδομένων για την εκπαίδευση των παραμέτρων 8 κατανομών.

Καμπύλες ποσοστών για ένα παράδειγμα

Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα των μεθοδολογιών και αλγορίθμων κατάτμησης και κατηγοριοποίησης που περιγράφηκαν στην ενότητα της χρήσης καμπύλων ποσοστών. Συγκεκριμένα, για ένα επιλεγμένο πειράματα από αυτά που παρουσιάστηκαν

Πίνακας 5.3: Αποτελέσματα των 4 πειραμάτων (2α),(2β),(2γ) και (2δ). Για την εκπαίδευση του κάθε μοντέλου χρησιμοποιούμε 100 παρατηρήσεις από την αντίστοιχη κλάση. Στο πείραμα α η μοντελοποίηση έγινε με 1 γκαουσιανή, στο β με 2 γκαουσιανές, στο γ με 4 γκαουσιανές και στο δ με 8 γκαουσιανές

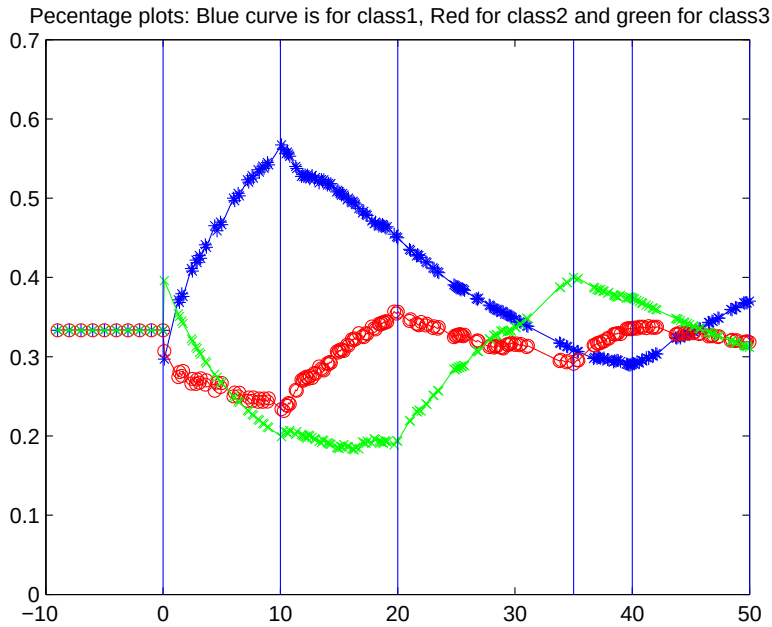
experiment 2a			
Class Name	Hits	FAs	Actual
class1	134	66	173
class2	69	81	160
class3	124	27	166
Overall	326	174	499
%Correct	65.73		
%Accurate	65.13		
experiment 2b			
Class Name	Hits	FAs	Actual
class1	160	40	192
class2	97	53	159
class3	122	28	147
Overall	379	121	498
%Correct	76.31		
%Accurate	75.70		
experiment 2c			
Class Name	Hits	FAs	Actual
class1	167	33	189
class2	105	45	162
class3	124	26	148
Overall	396	104	499
%Correct	79.56		
%Accurate	79.16		
experiment 2d			
Class Name	Hits	FAs	Actual
class1	150	50	173
class2	101	49	188
class3	112	38	138
Overall	363	137	499
%Correct	72.95		
%Accurate	72.55		

στην προηγούμενη ενότητα,το (2β), παρουσιάζονται οι καμπύλες ποσοστών και όλα τα ενδιάμεσα αποτελέσματα των αλγορίθμων μέχρι την παραγωγή του τελικού transcription.

Υπενθυμίζουμε ότι στο πείραμα 2α χρησιμοποιούμε 100 frames, που αντιστοιχούν σε χρόνο 1sec για την εκπαίδευση της κάθε κλάσης και μοντελοποιούμε την κάθε κλάση με 2 γκαουσιανές κατανομές. Η απόδοση του συστήματος αναγνώρισης frames δίνει Correct=76.31% και Accuracy=75.7%.

Σημειώνουμε ότι όλοι οι χρόνοι στους πίνακες και στις καμπύλες ποσοστών σημειώνονται σε sec*0.1.

Οι καμπύλες ποσοστών φαίνονται στο σχήμα 5.14. Παρατηρούμε ότι οι καμπύλες δείχνουν σε γενικές γραμμές που βρίσκεται η κάθε κλάση, δηλαδή η καμπύλη κάθε κλάσης είναι αύξουσα στην περιοχή της κλάσης αυτής. Τα τελικά αποτελέσματα είναι αρκετά κοντά στο σωστό transcription, όπως φαίνεται από τον πίνακα 5.3.



Εικόνα 5.14: Οι καμπύλες ποσοστών, πριν την ομαλοποίηση για το πείραμα (2β). Η κλάση 1 συμβολίζεται με μπλε γραμμή και *, η κλάση 2 με κόκκινη γραμμή και ο και η κλάση 3 με πράσινη γραμμή και x. Στο πείραμα έγινε χρήση 100 παρατηρήσεων για κάθε κλάση στο train set και μοντελοποίηση με 2 γκαουσιανές.

Πίνακας 5.4: Αποτελέσματα για το πείραμα (2β), δηλαδή χρήση 100 παρατηρήσεων για κάθε κλάση στο train set και μοντελοποίηση με 2 γκαουσιανές. Φαίνονται διαδοχικά τα διαστήματα που εξάγονται από τις καμπύλες ποσοστών,η επεξεργασία τους από τους αλγορίθμους merge-delete, merger και majority-process και η παραγωγή του τελικού Transcription

percPlots		
class1	class2	class3
0 10.5	10.5 19.9	0 0.1
39.9 50	25.2 25.3	15.1 15.3
	28.5 29.7	16.2 18.1
	34.5 41.2	19 34.5
	44.4 44.7	
merge-delete		
class1	class2	class3
0 10.5	10.5 19.9	19 34.5
39.9 50	34.5 41.2	
merger		
transcription		
0	10.5	class1
10.5	19.45	class2
19.45	34.5	class3
34.5	40.55	class2
40.55	50	class1
majority-process		
transcription		
0	10.5	class1
10.5	19.45	class2
19.45	34.5	class3
34.5	40.55	class2
40.55	50	class1

5.6.2 Πειράματα σε Πραγματικά Δεδομένα

Σε αυτήν την ενότητα μελετούμε την απόδοση του συστήματός μας σε πραγματικά δεδομένα. Χωρίζουμε το σύνολο των διαθέσιμων δελτίων ειδήσεων και των αντίστοιχων transcriptions σε train και test set. Εκπαιδεύουμε τα μοντέλα του συστήματός μας με τα δελτία του train set και βρίσκουμε την απόδοση για τα δελτία του test set.

Το σύστημα που έχουμε υλοποιήσει κατατάσσει αρχικά τα frames σε ομιλία και μη ομιλία (χρησιμοποιώντας δηλαδή ένα HMM με 2 αντίστοιχες καταστάσεις) και στο δεύτερο στάδιο κατατάσσει τα τμήματα μη ομιλίας σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία. Ως αποτέλεσμα παράγονται 2 κατατμήσεις (transcriptions) του δελτίου. Η πρώτη το χωρίζει σε τμήματα ομιλίας και μη ομιλίας και η δεύτερη σε τμήματα αντρικής ομιλίας, γυναικείας ομιλίας και μη ομιλίας.

Η απόδοση μετράται βρίσκοντας τα ποσοστά %Correct και %Accurate. Τα ποσοστά αυτά βρίσκονται με τη βοήθεια της συνάρτησης του HTK HResults, όπου γίνεται σύγκριση του transcription εξόδου του συστήματος με ένα transcription αναφοράς. Το κάθε ένα από τα transcriptions περιέχει για κάθε διαδοχικό υποτμήμα των 30msec του δελτίου την αντίστοιχη κατάταξη (πχ ομιλία ή μη ομιλία).Ο ορισμός των ποσοστών %Correct και %Accurate είναι:

$$\%Correct = \frac{H}{N} \times 100\%$$

$$\%Accurate = \frac{H - I}{N} \times 100\%$$

όπου N είναι το συνολικό πλήθος των υποτμημάτων που εξετάζονται στο transcription, H είναι το πλήθος των υποτμημάτων που κατατάχθηκαν σωστά και I είναι το συνολικό πλήθος των τμημάτων που κατατάχθηκαν σε λάθος κατηγορία (Insertions).

Σκοπός είναι να μελετηθεί και να συγκριθεί η απόδοση 3 διαφορετικών αλγορίθμων. Ο πρώτος είναι αυτός που δέχεται έτοιμα τα τμήματα του σταδίου κατάτμησης και κατηγοριοποιεί το κάθε τμήμα στην κλάση όπου ανήκει η πλειοψηφία των frames του. Ο αλγόριθμος αυτός περιγράφεται στην ενότητα 5.4 και θα αναφέρεται στο εξής ως αλγόριθμος Majority. Ο δεύτερος εκτελεί επιπλέον κατάτμηση στα τμήματα που δέχεται από το στάδιο κατάτμησης με βάση τα αποτελέσματα του HMM μοντέλου. Στη συνέχεια ομαλοποιεί τα αποτελέσματα των μοντέλων εκτελώντας απλό median filtering. Ο αλγόριθμος αυτός περιγράφεται στην ενότητα 5.5.3 και θα αναφέρεται στο εξής ως αλγόριθμος Smoothing. Ο τρίτος εξεταζόμενος αλγόριθμος εκτελεί τα βήματα του αλγορίθμου Smoothing αλλά στη συνέχεια με βάση τα αποτελέσματα που παίρνει κατασκευάζει τις καμπύλες ποσοστών και

χρησιμοποιεί τους αλγορίθμους merge-delete και merger για να παράγει το τελικό transcription. Ο αλγόριθμος αυτός περιγράφεται στην ενότητα 5.5.4 και θα αναφέρεται στο εξής ως αλγόριθμος Two-Phase.

Τέλος σημειώνουμε ότι η απόδοση του κάθε αλγορίθμου κρίνεται όχι μόνο από τα ποσοστά %Correct και %Accurate αλλά και από την δυνατότητα παραγωγής ενός Transcription που δεν θα είναι υπερκατατμημένο και θα περιέχει υποτμήματα ικανοποιητικά μεγάλου μήκους, για παράδειγμα μήκους πάνω από 0.3sec.

Στα πειράματά μας το train set αποτελείται από 8 ολόκληρα δελτία ειδήσεων, μήκους περίπου 1 ώρας το καθένα. Τα 5 από αυτά προέρχονται από τη NET και τα υπόλοιπα 3 από το MEGA. Το test set αποτελείται από 1 δελτίο ειδήσεων μήκους 1 ώρας και 5 λεπτών περίπου που προέρχεται από τη NET και δεν έχει χρησιμοποιηθεί στο train set.

Χρησιμοποιήθηκαν 2 HMM μοντέλα. Το πρώτο περιείχε 2 καταστάσεις για ομιλία και μη ομιλία. Η κατάσταση της μη ομιλίας εκπαιδεύτηκε με 4 γκαουσιανά μίγματα ενώ η κατάσταση της ομιλίας εκπαιδεύτηκε με 8 γκαουσιανά μίγματα λόγω του μεγαλύτερου πλήθους διαθέσιμων δεδομένων. Το δεύτερο HMM περιείχε 3 καταστάσεις για αντρική ομιλία, γυναικεία ομιλία και μη ομιλία. Οι καταστάσεις αντρικής και γυναικείας ομιλίας εκπαιδεύτηκαν με 8 γκαουσιανά μίγματα και η κατάσταση της μη ομιλίας με 4 γκαουσιανά μίγματα. Ως χαρακτηριστικά χρησιμοποιήθηκαν 13 MFCC με τις πρώτες και δεύτερες παραγώγους τους. Επιλέξαμε frame=40msec και overlap=10msec.

Ως είσοδο και για τους 3 εξεταζόμενους αλγορίθμους θεωρούμε τις κατατμήσεις που βρέθηκαν από το στάδιο κατάτμησης, εκτελώντας τον αλγόριθμο BIC χωρίς δεύτερο πέρασμα. Το τελικό αποτέλεσμα συγκρίνεται με ένα transcription αναφοράς.

Αναφέρουμε ότι το Transcription αναφοράς περιέχει 396 αλλαγές, είτε μεταξύ ομιλίας και μη ομιλίας είτε μεταξύ αλλαγών ομιλητή. Η σήμανση των αλλαγών έχει γίνει με το χέρι και χωρίς μεγάλη ευαισθησία. Για παράδειγμα δεν σημειώνονται ως αλλαγές οι αλλαγές έντασης του θορύβου ή οι αλλαγές έντασης της φωνής κάποιου ομιλητή ή οι αλλαγές στο θόρυβο που ακούγεται στο background κατά τη διάρκεια μίας ομιλίας. Τέλος σημειώνουμε ότι τα τμήματα στα οποία έχει χωριστεί το δελτίο έχουν μέσο μήκος 10sec και το μικρότερο τμήμα έχει μήκος 0.52sec

Το Transcription που προέρχεται από τον αλγόριθμο BIC περιέχει 464 αλλαγές. Τα τμήματα στα οποία έχει χωριστεί το δελτίο έχουν μέσο μήκος 8.5sec και το μικρότερο τμήμα έχει μήκος 0.49sec.

Συγκρίνοντας τις αλλαγές του BIC με τις αλλαγές του transcription αναφοράς παίρ-

νουμε τα αποτελέσματα που φαίνονται στον πίνακα 5.5

Πίνακας 5.5: Αποτελέσματα από τη σύγκριση αλλαγών του transcription αναφοράς και αλλαγών που βρέθηκαν με τον αλγόριθμο BIC σε δελτίο μήκους 1 ώρας περίπου

	BIC
Success	75%
False Alarm	35.9%
Correct Changes Detected	297
False Changes Detected	167
Missed Real Changes	99
Total Changes Detected	464
Total Real Changes	396

Με δεδομένες αυτές τις αλλαγές εισόδου και τα τμήματα που ορίζουν εξετάζουμε τη συμπεριφορά των αλγορίθμων Majority, Smoothing και Two-Phase ως προς την κατηγοριοποίηση και περαιτέρω κατάτμηση. Σημειώνουμε ότι τα αποτελέσματα των αλγορίθμων αξιολογούνται χωρίς να κάνουμε σε αυτά κάποια επιπλέον επεξεργασία, όπως συγχώνευση γειτονικών όμοιων τμημάτων ή συγχώνευση πολύ μικρών τμημάτων με τα γειτονικά τους, εκτός κι αν αναφέρεται διαφορετικά.

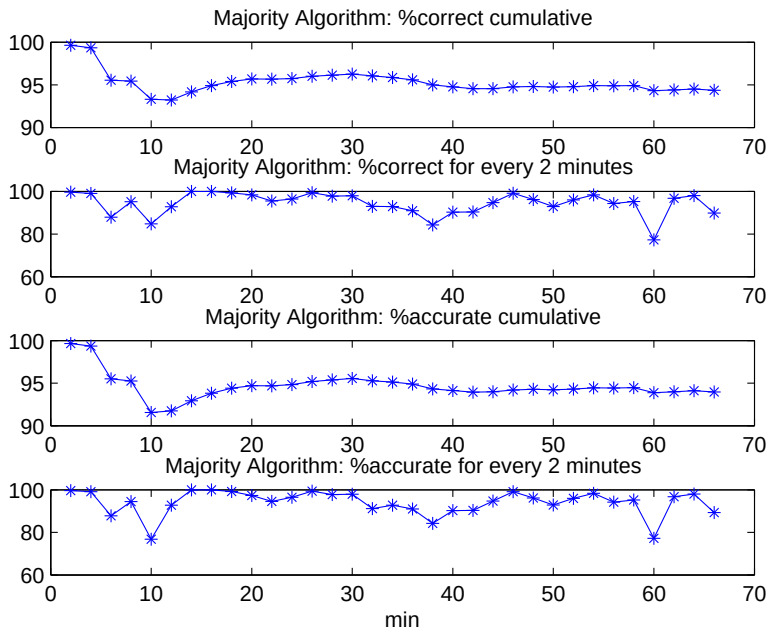
Κατηγοριοποίηση σε Ομιλία και μη Ομιλία Αρχικά εξετάζουμε την περίπτωση κατηγοριοποίησης σε ομιλία και μη ομιλία. Στον πίνακα 5.6 φαίνονται τα ποσοστά Correct και Accurate που βρέθηκαν από την HResults για τα πρώτα 20λεπτά του δελτίου, τα πρώτα 40 λεπτά του δελτίου και για το συνολικό δελτίο.

Επίσης στην εικόνα 5.15 φαίνονται τα ποσοστά Correct και Accurate για τον αλγόριθμο Majority. Στην πρώτη γραμμή της εικόνας φαίνονται τα συγκεντρωτικά ποσοστά Success μέχρι την τρέχουσα χρονική στιγμή. Τα ποσοστά μετρώνται κάθε 2 λεπτά. Στην δεύτερη γραμμή φαίνονται τα ποσοστά Success για κάθε δίλεπτο που μετράμε. Ομοίως στην τρίτη γραμμή της εικόνας φαίνονται τα συγκεντρωτικά ποσοστά Accurate μέχρι την τρέχουσα χρονική στιγμή και στην τέταρτη γραμμή φαίνονται τα ποσοστά Accurate για κάθε δίλεπτο που μετράμε. Παρόμοιες γραφικές παραστάσεις παρουσιάζονται στις εικόνες 5.16 και 5.17 για τους αλγορίθμους Smoothing και Two-Phase αντίστοιχα

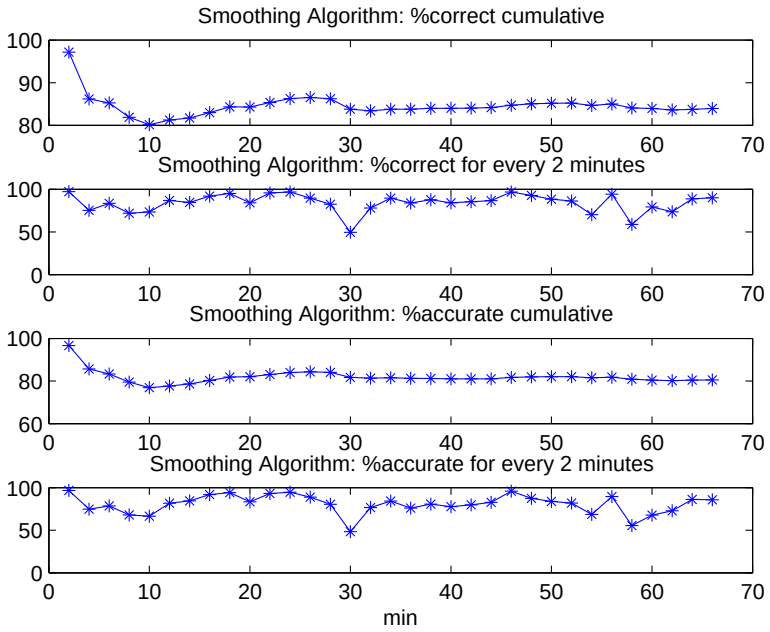
Παρατηρούμε ότι η απόδοση του αλγορίθμου Majority που κατατάσσει τα τμήματα εισόδου στην κλάση όπου ανήκει η πλειοψηφία των frames τους, είναι αισθητά ανώτερη

Πίνακας 5.6: Ποσοστά Correct και Accurate για τους 3 αλγόριθμους και για διάφορα μήκη του εξεταζόμενου δελτίου για το πρόβλημα κατάταξης σε ομιλία και μη ομιλία

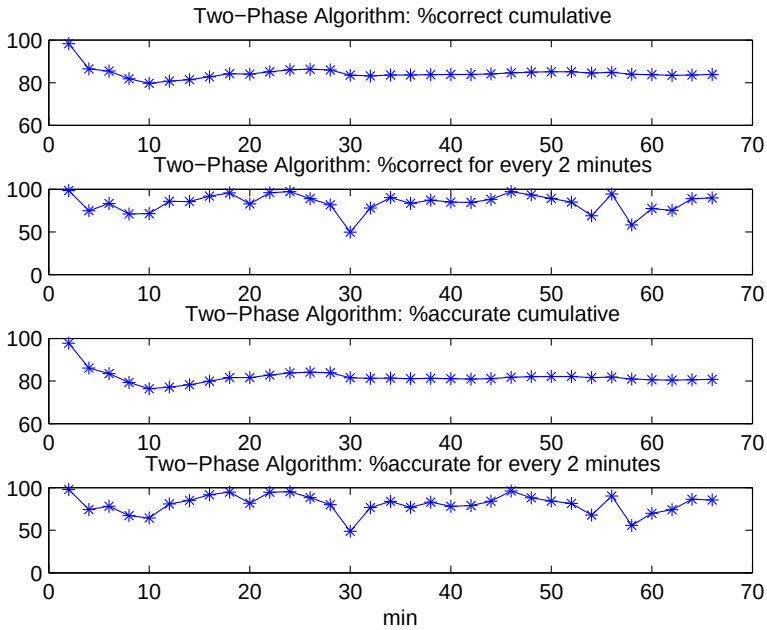
speech-nonspeech		Majority	Smoothing	Two-Phase
first 20min	%Correct	95.69	84.3	84.1
	%Accurate	94.7	81.99	81.7
first 40min	%Correct	94.76	83.9	83.9
	%Accurate	94.13	81.03	81.1
total	%Correct	94.51	83.94	83.86
	%Accurate	94.11	80.5	80.7



Εικόνα 5.15: Συγκεντρωτικά και στιγμιαία ποσοστά success και accurate για τον αλγόριθμο Majority και για την κατηγοριοποίηση σε ομιλία και μη ομιλία.



Εικόνα 5.16: Συγκεντρωτικά και στιγμιαία ποσοστά success και accurate για τον αλγόριθμο Smoothing και για την κατηγοριοποίηση σε ομιλία και μη ομιλία.



Εικόνα 5.17: Συγκεντρωτικά και στιγμιαία ποσοστά success και accurate για τον αλγόριθμο Two-Phase και για την κατηγοριοποίηση σε ομιλία και μη ομιλία.

από την απόδοση των αλγορίθμων Smoothing και Two-Phase που επιχειρούν να κάνουν περαιτέρω κατάτμηση του audio-stream εισόδου με βάση τα αποτελέσματα της κατάταξης. Αυτό καταρχήν οφείλεται στο γεγονός ότι ο αλγόριθμος BIC λειτουργεί αρκετά καλά με την έννοια ότι τα περισσότερα τμήματα που βρίσκει είναι ομογενή δηλαδή ανήκουν είτε εξολοκλήρου στην κλάση της ομιλίας είτε στην κλάση της μη ομιλίας. Άλλωστε, όπως αναλύθηκε στο κεφάλαιο 4, ο BIC βρίσκει με μεγάλα ποσοστά επιτυχίας τις αλλαγές μεταξύ ομιλίας και μη ομιλίας. Κατά συνέπεια, η παραδοχή των ομογενών τμημάτων που κάνει ο BIC ισχύει για το συγκεκριμένο πρόβλημα και έτσι τα αποτελέσματα είναι πολύ καλά.

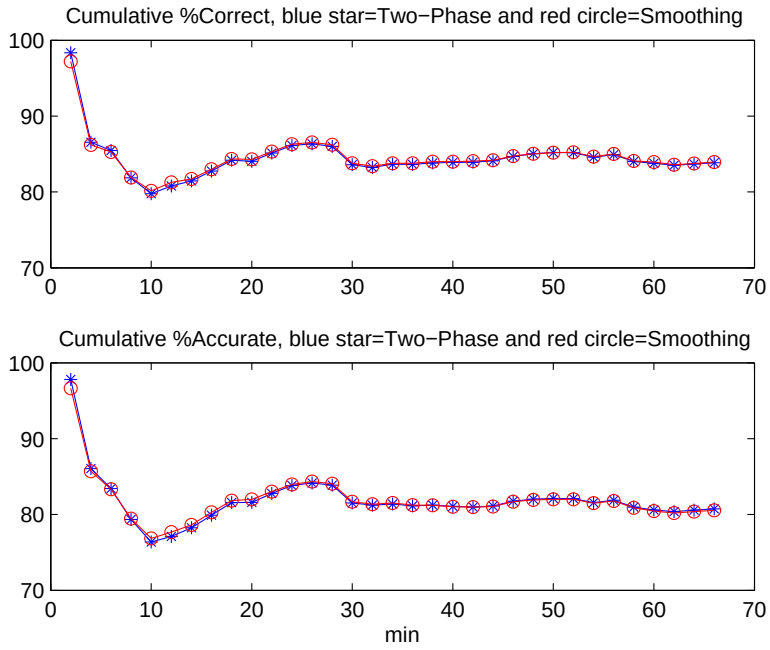
Επίσης, πρέπει να προσθέσουμε ότι η ακρίβεια/ευαισθησία της κατάτμησης που επιτυγχάνει ο BIC είναι αντίστοιχη με την ευαισθησία του transcription αναφοράς. Αντίθετα, οι αλγόριθμοι Smoothing και Two-Phase επιχειρούν μία πιο λεπτομερή κατάτμηση από αυτή του Transcription αναφοράς. Αυτή η λεπτομερής κατάτμηση πιθανόν δεν είναι απαραίτητη για την απλή περίπτωση διαχωρισμού μεταξύ ομιλίας και μη ομιλίας κατά συνέπεια παρατηρούμε μείωση της απόδοσης για τους αλγόριθμους αυτού. Εντούτοις, και οι δύο αλγόριθμοι λειτουργούν ικανοποιητικά και μάλιστα επιτυγχάνουν παρόμοια ποσοστά Correct και Accurate.

Σχετικά με τα λάθη κατάταξης των αλγορίθμων, μελετώντας τα τμήματα και τις αντίστοιχες κατατάξεις που μας δίνουν οι αλγόριθμοι, παρατηρούμε ότι το πιο κοινό λάθος είναι η λανθασμένη κατηγοριοποίηση της θορυβώδους ομιλίας σε θόρυβο. Αυτό είναι ένα αναμενόμενο πρόβλημα δεδομένου ότι πολλές φορές και το ανθρώπινο αυτί δυσκολεύεται να διακρίνει την πολύ θορυβώδη ομιλία από τον θόρυβο. Ένα άλλο συχνό λάθος είναι η λανθασμένη κατάταξη των τμημάτων σιωπής σε ομιλία. Το λάθος αυτό οφείλεται σε δύο λόγους. Πρώτον στο γεγονός ότι τα διαθέσιμα δελτία του train set περιείχαν πολύ μικρό ποσοστό σιωπής και έτσι είναι δύσκολη η εκπαίδευση του μοντέλου της μη ομιλίας (που περιέχει και τη σιωπή), ώστε να αναγνωρίζει ικανοποιητικά τις σιωπές. Δεύτερον, η ανθρώπινη ομιλία περιέχει συχνά μικρές ή μεγαλύτερες παύσεις και συχνά είναι υποκειμενικό για ακόμα και για τον άνθρωπο να αποφασίσει αν μία μικρή παύση θα έπρεπε να κατηγοριοποιηθεί ως σιωπή ή θα έπρεπε να ονομαστεί ομιλία καθώς περιέχεται μέσα σε ένα τμήμα ομιλίας.

Σχετικά με τις στιγμιαίες συγκεντρωτικές γραφικές παραστάσεις των ποσοστών Correct και Accurate για κάθε 2 λεπτά του δελτίου, έχουμε να παρατηρήσουμε ότι αν και τα στιγμιαία ποσοστά αλλάξουν αισθητά ανά 2 λεπτά και εξαρτώνται από τα ποσοστά θορύβου και τη δυσκολία κατάταξης των τμημάτων του κάθε δίλεπτου, τα συγκεντρωτικά ποσοστά συγκλίνουν σταδιακά σε κάποια τιμή. Πράγματι, μετά τα 15 πρώτα λεπτά φαίνεται ότι τα ποσοστά έχουν συγκλίνει σε μία τιμή. Κατά συνέπεια, φαίνεται ότι είναι ασφαλές να

μετρήσουμε τα ποσοστά επιτυχίας στα 15 πρώτα λεπτά ενός δελτίου, ώστε να έχουμε μία ικανοποιητική προσέγγιση των ποσοστών για το συνολικό δελτίο.

Ακολουθεί λεπτομερής σύγκριση των αποτελεσμάτων των αλγορίθμων Smoothing και Two-Phase. Όπως φαίνεται στην εικόνα 5.18 τα συγκεντρωτικά ποσοστά απόδοσης, δηλαδή Correct και Accurate, για τους δύο αλγορίθμους είναι περίπου τα ίδια. Εντούτοις, τα Transcriptions που παράγει ο κάθε αλγόριθμος είναι αρκετά διαφορετικά μεταξύ τους ως προς το πλήθος των τμημάτων και ως προς το μέσο μήκος του κάθε τμήματος. Τέτοιες πληροφορίες φαίνονται στον πίνακα 5.7.



Εικόνα 5.18: Συγκεντρωτικά ποσοστά success και accurate για τους αλγόριθμους Smoothing και Two-Phase και για την κατηγοριοποίηση σε ομιλία και μη ομιλία.Με κόκκινους κύκλους φαίνονται τα ποσοστά του Smoothing και με μπλε αστέρια τα ποσοστά του Two-Phase.

Από τον πίνακα 5.7 φαίνεται ότι η ποιότητα του transcription που παράγει ο αλγόριθμος Two-Phase είναι αισθητά καλύτερη από την ποιότητα του Transcription του Smoothing. Πράγματι πάνω από τα μισά τμήματα από αυτά που παράγει ο Smoothing έχουν μήκος μικρότερο από 0.8 sec ενώ το 47% είναι μικρότερα από 0.4sec, το 33% είναι μικρότερα από

Πίνακας 5.7: Πληροφορίες σχετικά με τα μήκη (σε sec) των τμημάτων των Transcriptions που παράγονται από τους αλγορίθμους Two-Phase και Smoothing για δελτίο μήκους 1 ώρας περίπου και για κατάταξη σε ομιλία και μη ομιλία

speech-nonspeech		
	Two-Phase	Smoothing
total segments	1396	2916
min length	0.01	0.01
max lenght	93.52	79.51
mean length	2.84	1.3
length <0.1	3	624
length <0.2	13	957
length <0.4	40	1372
length <0.8	367	1807

0.2 sec και το 21.4% είναι μικρότερα από 0.1sec. Κατά συνέπεια, η μεγάλη πλειοψηφία των τμημάτων που παράγονται από τον Smoothing είναι υπερβολικά μικρά και δεν μπορούν να χρησιμοποιηθούν σε κάποιο επόμενο στάδιο επεξεργασίας (πχ αναγνώριση των τμημάτων φωνής) Ένα επιπλέον σοβαρό μειονέκτημα που προκύπτει από τη μελέτη του Transcription είναι ότι τα υπερβολικά μικρά τμήματα (μικρότερα από 0.2 sec) εμφανίζονται διαδοχικά και σε ομάδες, κατά συνέπεια δεν μπορούν να συγχωνευθούν με τα γειτονικά τους. Είναι εμφανές ότι το transcription του Smoothing απαιτεί σημαντική μεταεπεξεργασία ώστε να μπορέσει να χρησιμοποιηθεί από κάποιο επόμενο στάδιο αναγνώρισης.

Η επεξεργασία αυτή φαίνεται να εκτελείται από τον αλγόριθμο Two-Phase ο οποίος αν και δεν επιτυγχάνει αύξηση των ποσοστών Correct και Accurate, επιτυγχάνει την παραγωγή ενός πιο ομαλού transcription με αρκούντως μεγάλα υποτμήματα. Τα υπερβολικά μικρά τμήματα είναι λίγα και βρίσκονται συνήθως ανάμεσα σε μεγάλα υποτμήματα, γι αυτό μπορούν να συγχωνευθούν εύκολα με εφαρμογή κάποιων απλών κανόνων.

Για να δείξουμε κατά πόσο τα 2 transcriptions προσφέρονται για μία επεξεργασία εφαρμόζουμε σε αυτά 2 απλούς κανόνες.

1. Συγχωνεύουμε διαδοχικά τμήματα που έχουν κατηγοριοποιηθεί στην ίδια κλάση

2. Αν ένα τμήμα είναι μικρό, δηλαδή μικρότερο από 0.3sec, και τα δύο γειτονικά του είναι μεγάλα, δηλαδή μεγαλύτερα ή ίσα με 0.3sec, τότε διαγράφουμε το μικρό τμήμα και

συγχωνεύουμε τα 2 γειτονικά του (τα οποία σύμφωνα με τον κανόνα 1 θα ανήκουν στην ίδια κλάση)

Πίνακας 5.8: Πληροφορίες σχετικά με τα μήκη (σε sec) των τμημάτων των transcriptions που παράγονται από τους αλγορίθμους Two-Phase και Smoothing για δελτίο μήκους 1 ώρας περίπου, μετά από μετα-επεξεργασία των transcriptions και για κατάταξη σε ομιλία και μη ομιλία

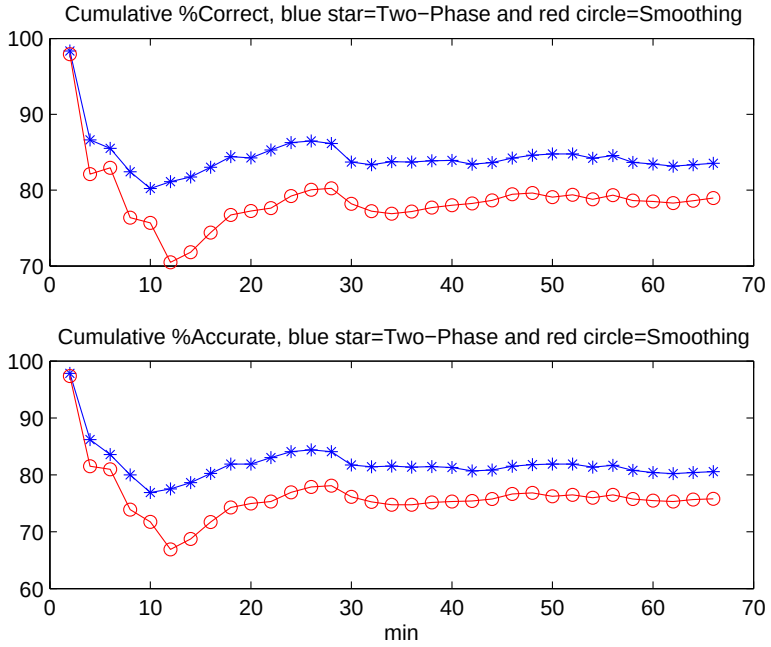
speech-nonspeech		
	Two-Phase	Smoothing
total segments	973	1931
min length	0.38	0.01
max length	143.7	92.31
mean length	4	2
length < 0.1	0	499
length < 0.2	0	738
length < 0.4	6	977
length < 0.8	240	1169

Παρατηρούμε ότι η εφαρμογή των 2 αυτών απλών κανόνων προκάλεσε μεγάλη βελτίωση στην ποιότητα του transcription του αλγορίθμου Two-Phase εξαφανίζοντας σχεδόν τα μικρά τμήματα και μειώνοντας αισθητά την υπερκατάτμηση. Αντίθετα, τα αποτελέσματα του Smoothing δεν παρουσιάζουν αντίστοιχη βελτίωση και παρατηρούμε ότι εξακολουθεί να υπάρχει μεγάλο πλήθος τμημάτων κάτω των 0.2sec. Άρα ο Two-Phase παράγει αποτέλεσμα που προσφέρεται για μετα-επεξεργασία.

Τέλος, στην εικόνα 5.19 φαίνονται οι συγκεντρωτικές καμπύλες Correct και Accurate για τα 2 Transcriptions μετά από εφαρμογή των 2 απλών κανόνων.

Παρατηρούμε ότι η εφαρμογή των 2 αυτών κανόνων έχει αυξήσει ελαφρώς τα ποσοστά επιτυχίας του Transcription του αλγορίθμου Two-Phase και έχει μειώσει ελαφρώς τα ποσοστά επιτυχίας του αλγορίθμου Smoothing. Κατά συνέπεια, η απόδοση του Two-Phase μετά από μεταεπεξεργασία των δεδομένων είναι καλύτερη από την απόδοση του Smoothing.

Κατηγοριοποίηση σε Αντρική Ομιλία, Γυναικεία Ομιλία και Μη Ομιλία Εξετάζουμε την περίπτωση κατηγοριοποίησης σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία.



Εικόνα 5.19: Συγκεντρωτικά ποσοστά success και accurate για τους αλγόριθμους Smoothing και Two-Phase και για την κατηγοριοποίηση σε ομιλία και μη ομιλία, μετά από την εφαρμογή των κανόνων.Με κόκκινους κύκλους φαίνονται τα ποσοστά του Smoothing και με μπλε αστέρια τα ποσοστά του Two-Phase.

Στον πίνακα 5.9 φαίνονται τα ποσοστά Correct και Accurate που βρέθηκαν από την HResults για τα πρώτα 20λεπτά του δελτίου, τα πρώτα 40 λεπτά του δελτίου και για το συνολικό δελτίο.

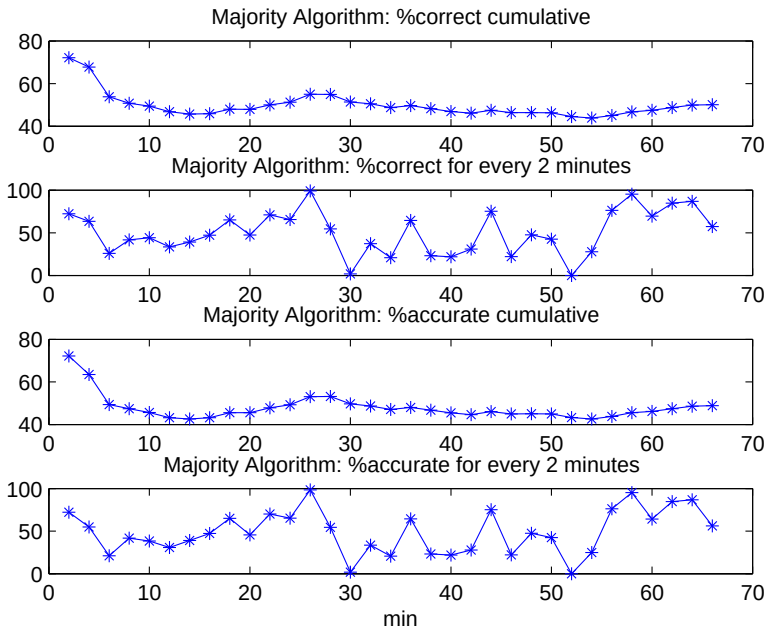
Πίνακας 5.9: Ποσοστά Correct και Accurate για τους 3 αλγόριθμους και για διάφορα μήκη του εξεταζόμενου δελτίου για την κατάταξη σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία

male-female-nonspeech		Majority	Smoothing	Two-Phase
first 20min	%Correct	50.0	55.0	56.14
	%Accurate	47.8	45.2	45.5
first 40min	%Correct	48.2	52.8	54.15
	%Accurate	45.4	42.6	43.65
total	%Correct	50.1	52.8	53.9
	%Accurate	48.85	42.9	43.76

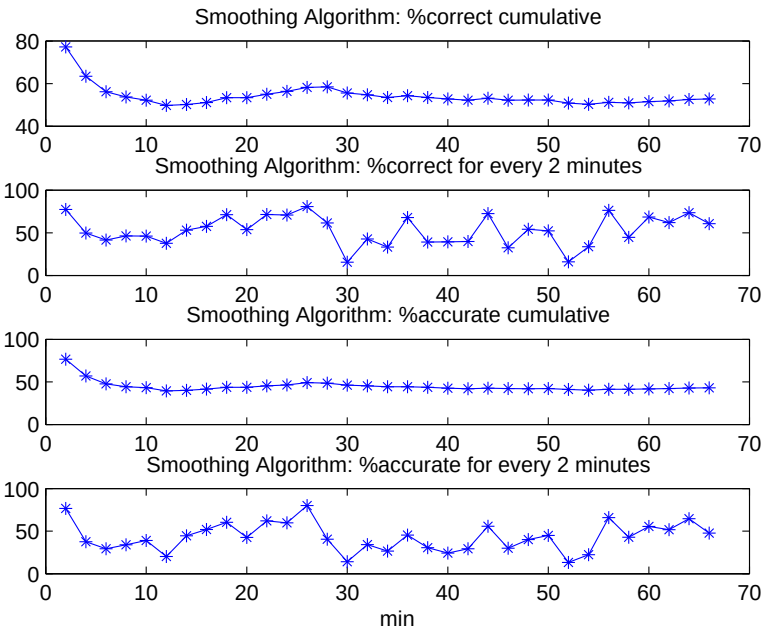
Επίσης στην εικόνα 5.20 φαίνονται τα ποσοστά Correct και Accurate για τον αλγόριθμο Majority. Στην πρώτη γραμμή της εικόνας φαίνονται τα συγκεντρωτικά ποσοστά Success μέχρι την τρέχουσα χρονική στιγμή. Τα ποσοστά μετρώνται κάθε 2 λεπτά. Στην δεύτερη γραμμή φαίνονται τα ποσοστά Success για κάθε δίλεπτο που μετράμε. Ομοίως στην τρίτη γραμμή της εικόνας φαίνονται τα συγκεντρωτικά ποσοστά Accurate μέχρι την τρέχουσα χρονική στιγμή και στην τέταρτη γραμμή φαίνονται τα ποσοστά Accurate για κάθε δίλεπτο που μετράμε. Παρόμοιες γραφικές παραστάσεις παρουσιάζονται στις εικόνες 5.21 και 5.22 για τους αλγορίθμους Smoothing και Two-Phase αντίστοιχα

Παρατηρούμε καταρχήν ότι για το πρόβλημα κατηγοριοποίησης σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία τα ποσοστά είναι πολύ χαμηλότερα και για τους 3 εξεταζόμενους αλγορίθμους. Αυτό συμβαίνει για 2 λόγους. Πρώτον επειδή το πρόβλημα κατηγοριοποίησης σε αντρική και γυναικεία ομιλία είναι πιο δύσκολο από την κατηγοριοποίηση σε ομιλία και μη ομιλία, καθώς οι κλάσεις ομιλίας και μη ομιλίας είναι πολύ πιο διαφορετικές μεταξύ τους. Δεύτερον επειδή για την εκπαίδευση των μοντέλων αντρικής και γυναικείας ομιλίας είχαμε διαθέσιμα τα μισά δεδομένα από ότι για την εκπαίδευση του μοντέλου ομιλίας, καθώς η συνολική διαθέσιμη ομιλία αποτελείται κατά 50% από αντρική και κατά 50% από γυναικεία ομιλία.

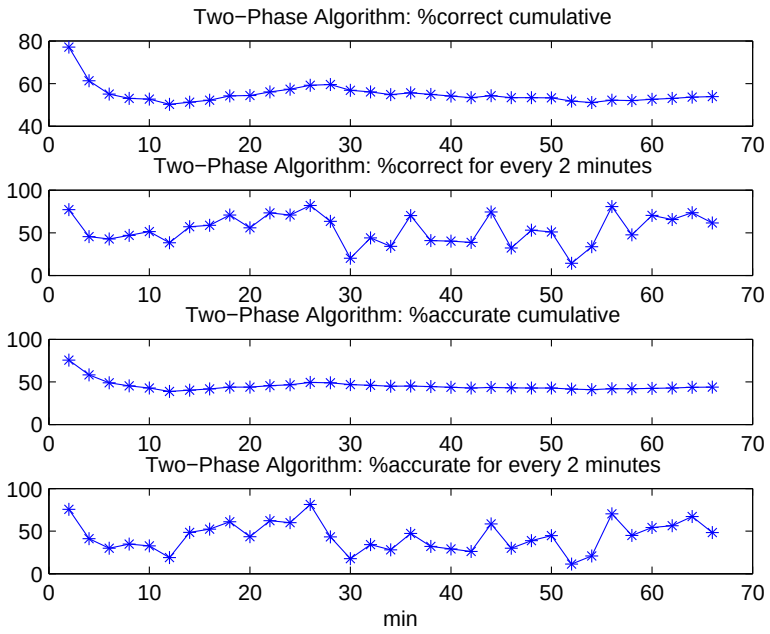
Επιπλέον παρατηρούμε ότι δεν ισχύει πλέον η ανωτερότητα του αλγορίθμου Majority ως



Εικόνα 5.20: Συγκεντρωτικά και στιγμιαία ποσοστά success και accurate για τον αλγόριθμο Majority και για την κατηγοριοποίηση σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία.



Εικόνα 5.21: Συγκεντρωτικά και στιγμιαία ποσοστά success και accurate για τον αλγόριθμο Smoothing και για την κατηγοριοποίηση σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία.



Εικόνα 5.22: Συγκεντρωτικά και στιγμιαία ποσοστά success και accurate για τον αλγόριθμο Two-Phase και για την κατηγοριοποίηση σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία.

προς τους αλγορίθμους Two-Phase και Smoothing. Αντιθέτως ο Majority αποδίδει ελαφρώς χειρότερα από τους άλλους δύο. Αυτό συμβαίνει επειδή δεν ισχύει πλέον η παραδοχή ότι τα τμήματα εισόδου που προέρχονται από τον BIC είναι ομογενή. Πράγματι, οι αλλαγές μεταξύ άντρικής και γυναικείας ομιλίας είναι αλλαγές ομιλητή και έχουμε δείξει στο κεφάλαιο 4 ότι ο BIC δεν βρίσκει με μεγάλη επιτυχία αλλαγές ομιλητή. Κατά συνέπεια πολλά από τα τμήματα εισόδου περιέχουν αλλαγές μεταξύ άντρα και γυναίκας, τις οποίες ο BIC δεν έχει βρει, γι αυτό το τμήμα δεν ανήκει σε μία μόνο κλάση και τα αποτελέσματα του Majority είναι χαμηλά.

Σημειώνουμε επιπλέον ότι τα πιο συνηθισμένα λάθη κατηγοριοποίησης που παρατηρούνται είναι η λανθασμένη κατηγοριοποίηση της αντρικής ομιλίας σε θόρυβο και η λανθασμένη κατηγοριοποίηση της γυναικείας ομιλίας σε θόρυβο ή αντρική ομιλία. Τέλος, συχνό λάθος είναι η λανθασμένη κατηγοριοποίηση της σιωπής σε ομιλία (αντρική ή γυναικεία), για λόγους που έχουν ήδη εξηγηθεί.

Ακολουθεί αναλυτική σύγκριση των αλγορίθμων Smoothing και Two-Phase. Όπως φαίνεται στην εικόνα 5.23 τα συγκεντρωτικά ποσοστά απόδοσης, δηλαδή Correct και Accurate, για τους δύο αλγορίθμους είναι περίπου τα ίδια, ενώ ο Two-Phase φαίνεται να υπερτερεί ελαφρώς. Τα Transcriptions που παράγει ο κάθε αλγόριθμος είναι αρκετά διαφορετικά μεταξύ τους ως προς το πλήθος των τμημάτων και ως προς το μέσο μήκος του κάθε τμήματος. Τέτοιες πληροφορίες φαίνονται στον πίνακα 5.10.

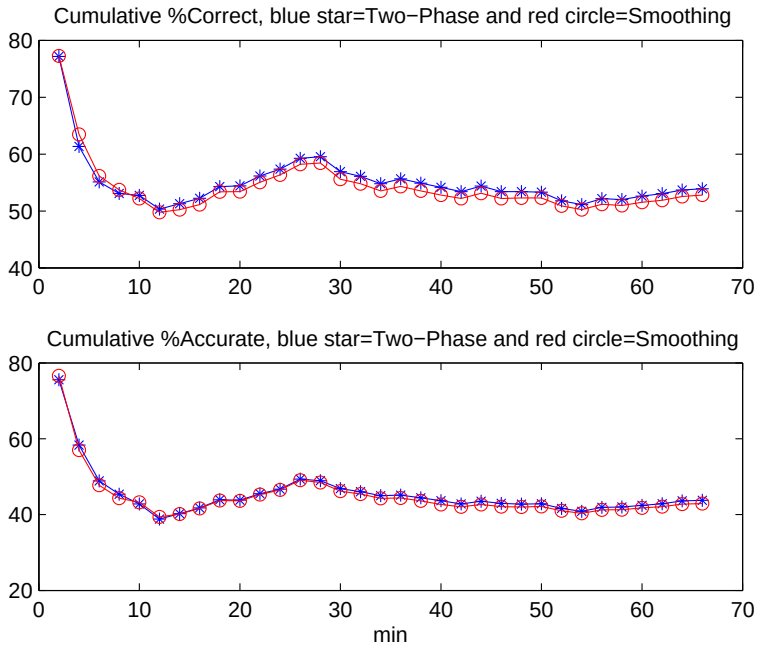
Παρατηρούμε ότι σε αυτό το δυσκολότερο πρόβλημα κατηγοριοποίησης τα ποσοστά μικρών τμημάτων είναι και στις 2 περιπτώσεις αρκετά υψηλά. Εντούτοις, και σε αυτήν την περίπτωση ο αλγόριθμος Two-Phase συμπεριφέρεται πολύ καλύτερα από τον Smoothing. Ο Smoothing δίνει ένα Transcription όπου κυριαρχούν τα μικρά τμήματα και είναι ουσιαστικά ακατάλληλο για περαιτέρω επεξεργασία.

Εφαρμόζουμε 3 απλούς κανόνες ώστε να δούμε πως συμπεριφέρονται τα αποτελέσματα των αλγορίθμων μετά από μία απλή μεταεπεξεργασία. Οι κανόνες είναι:

1. Συγχωνεύουμε διαδοχικά τμήματα που έχουν κατηγοριοποιηθεί στην ίδια κλάση

2. Αν ένα τμήμα είναι μικρό, δηλαδή μικρότερο από 0.3sec, και τα δύο γειτονικά του είναι μεγάλα, δηλαδή μεγαλύτερα ή ίσα με 0.3sec, και ανήκουν στην ίδια κλάση, τότε διαγράφουμε το μικρό τμήμα και συγχωνεύουμε τα 2 γειτονικά του

3. Αν ένα τμήμα είναι μικρό, δηλαδή μικρότερο από 0.3sec, και τα δύο γειτονικά του είναι μεγάλα, δηλαδή μεγαλύτερα ή ίσα με 0.3sec, και ανήκουν σε διαφορετικές κλά-



Εικόνα 5.23: Συγκεντρωτικά ποσοστά success και accurate για τους αλγόριθμους Smoothing και Two-Phase και για την κατηγοριοποίηση σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία.Με κόκκινους κύκλους φαίνονται τα ποσοστά του Smoothing και με μπλε αστέρια τα ποσοστά του Two-Phase.

Πίνακας 5.10: Πληροφορίες σχετικά με τα μήκη (σε sec) των τμημάτων των Transcriptions που παράγονται από τους αλγορίθμους Two-Phase και Smoothing για δελτίο μήκους 1 ώρας περίπου, για την κατηγοριοποίηση σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία

male-female-nonspeech		
	Two-Phase	Smoothing
total segments	2778	5673
min length	0.01	0.01
max length	79	74.13
mean length	1.43	0.7
length <0.1	325	1350
length <0.2	367	2075
length <0.4	525	3171
length <0.8	1297	4245

σεις, τότε διαγράφουμε το μικρό τμήμα και τοποθετούμε το όριο των 2 γειτονικών τμημάτων στο μέσο του διαγραμμένου τμήματος

Αποτελέσματα σχετικά με το πλήθος των τμημάτων των 2 transcriptions και τα μήκη τους φαίνονται στον πίνακα 5.11.

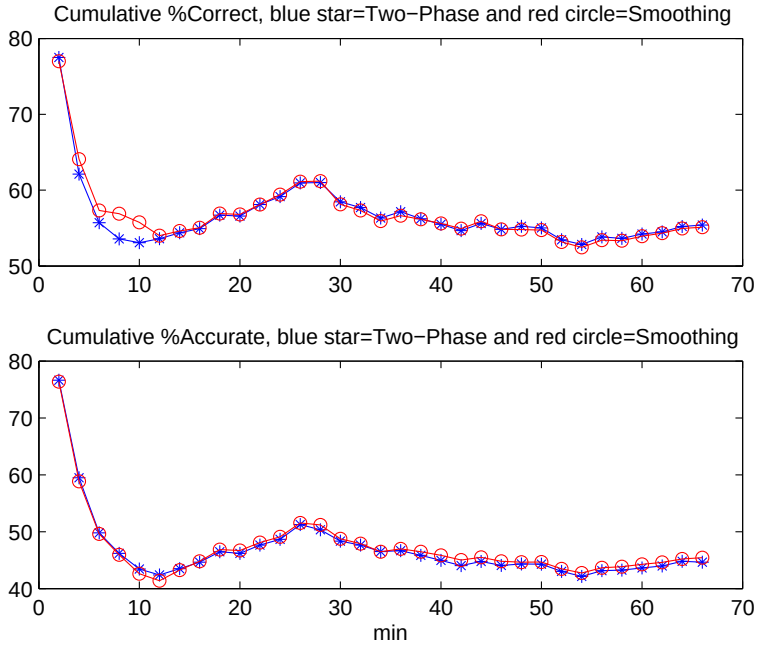
Πίνακας 5.11: Πληροφορίες σχετικά με τα μήκη (σε sec) των τμημάτων των Transcriptions που παράγονται από τους αλγορίθμους Two-Phase και Smoothing, μετά από εφαρμογή των κανόνων, για δελτίο μήκους 1 ώρας περίπου, για την κατηγοριοποίηση σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία

male-female-nonspeech		
	Two-Phase	Smoothing
total segments	1921	2927
min length	0.02	0.01
max length	79.92	74.13
mean length	2.06	1.35
length <0.1	18	576
length <0.2	23	871
length <0.4	67	1275
length <0.8	652	1748

Παρατηρούμε ότι η εφαρμογή των 3 αυτών απλών κανόνων προκάλεσε μεγάλη βελτίωση στην ποιότητα του transcription του αλγορίθμου Two-Phase εξαφανίζοντας σχεδόν τα μικρά τμήματα και μειώνοντας αισθητά την υπερκατάτμηση. Αντίθετα, τα αποτελέσματα του Smoothing παρουσιάζουν μικρότερη βελτίωση και παρατηρούμε ότι εξακολουθεί να υπάρχει μεγάλο πλήθος τμημάτων κάτω των 0.2sec. Άρα ο Two-Phase παράγει αποτέλεσμα που προσφέρεται για μετα-επεξεργασία.

Τέλος, στην εικόνα 5.24 φαίνονται οι συγκεντρωτικές καμπύλες Correct και Accurate για τα 2 Transcriptions μετά από εφαρμογή των 2 απλών κανόνων.

Παρατηρούμε ότι η εφαρμογή των 3 αυτών κανόνων δεν έχει επηρεάσει αισθητά τα ποσοστά correct και accurate για τους αλγορίθμους Smoothing και Two-Phase.



Εικόνα 5.24: Συγκεντρωτικά ποσοστά success και accurate για τους αλγόριθμους Smoothing και Two-Phase και για την κατηγοριοποίηση σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία, μετά από την εφαρμογή των κανόνων.Με κόκκινους κύκλους φαίνονται τα ποσοστά του Smoothing και με μπλε αστέρια τα ποσοστά του Two-Phase.

5.7 Συμπεράσματα

Στην ενότητα αυτή επιχειρήθηκε μία μελέτη του προβλήματος στατιστικής μοντελοποίησης ηχητικών κλάσεων όπως ομιλία, θόρυβος, σιωπή, με χρήση HMM και GMM μοντέλων. Επιδίωξή μας ήταν η κατασκευή ενός συστήματος που θα εκτελεί κατάτμηση και κατηγοριοποίηση ηχητικών τμημάτων με βάση κατάλληλα εκπαιδευμένα μοντέλα. Τελικός στόχος είναι η χρήση του συστήματος σε μία πραγματική εφαρμογή όπως η κατάτμηση και κατηγοριοποίηση δελτίων ειδήσεων σε κατάλληλες κλάσεις. Επίσης, έγινε προσπάθεια να συνδυαστούν τα αποτελέσματα του συστήματος κατάτμησης που αναπτύχθηκε και περιγράφηκε στο κεφάλαιο 4 με τα αποτελέσματα του συστήματος που αναπτύχθηκε στο παρόν κεφάλαιο.

Πιο συγκεκριμένα, αρχικά έγινε μία περιγραφή των εργαλείων που προσφέρει η συλλογή προγραμμάτων HTK για την ανάπτυξη ενός συστήματος βασισμένου σε HMM και GMM μοντέλα. Στη συνέχεια, αναφέρθηκαν τα βασικά βήματα για τη σχεδίαση και την κατασκευή ενός τέτοιου συστήματος και περιγράφηκαν κάποιες βασικές σχεδιαστικές αποφάσεις. Τέτοιες αποφάσεις αφορούν θέματα όπως οι κλάσεις του προβλήματος που θεωρήθηκαν και η δομή των HMM μοντέλων και εξαρτώνται από τη συγκεκριμένη εφαρμογή σε δελτία ειδήσεων αλλά από το πλήθος και την ποιότητα των διαθέσιμων δεδομένων εκπαίδευσης.

Ακολούθως, περιγράφηκε ένα απλό σύστημα που ασχολείται με ένα υποσύνολο του προβλήματος κατάτμησης και κατηγοριοποίησης, συγκεκριμένα μόνο με το πρόβλημα της κατηγοριοποίησης ομογενών ακουστικών τμημάτων με χρήση HMM. Συνεπώς το σύστημα που θα περιγραφεί παρακάτω δέχεται ως είσοδο ένα τμήμα audio-stream το οποίο θεωρείται ομογενές και το κατατάσσει σε κάποια από τις διαθέσιμες κλάσεις, συγκεκριμένα στην κλαση όπου έχει καταταχθεί η πλειοψηφία των frames του. Τα ομογενή τμήματα προέρχονται από το στάδιο εύρεσης αλλαγών σε audio-stream και κατάτμησής του, που έχει περιγραφεί αναλυτικά στο προηγούμενο κεφάλαιο. Ο αλγόριθμος που αναπτύχθηκε αναφέρεται ως Majority.

Στη συνέχεια επεκτείνουμε το σύστημά μας ώστε να εξετάσουμε το συνολικό πρόβλημα κατάτμησης και κατηγοριοποίησης. Το σύστημα που αναπτύχθηκε επιχειρεί όχι μόνο να κατατάξει τα διαθέσιμα τμήματα σε μία από τις υπάρχουσες κλάσεις αλλά επιχειρεί και να κάνει περαιτέρω κατάτμηση τμημάτων που δεν είναι ομογενή. Για την κατάτμηση προτείνονται και εξετάζονται 2 αλγόριθμοι, ο αλγόριθμος Smoothing που βασίζεται σε απλό median filtering και περιγράφεται στην ενότητα 5.5.3 και ο αλγόριθμος Two-Phase που περιγράφεται στην ενότητα 5.5.4. Ο αλγόριθμος Two-Phase είναι μία καινούρια ιδέα που επεκτείνει το απλό median filtering χρησιμοποιώντας την έννοια των καμπύλων ποσοστών που εισά-

γονται στην παρούσα εργασία. Επίσης περιγράφονται αλγόριθμοι για την επεξεργασία των καμπύλων ποσοστών.

Στην ενότητα των πειραματικών αποτελεσμάτων γίνονται πειράματα τόσο σε συνθετικά όσο και σε πραγματικά δεδομένα από δελτία ειδήσεων. Στη περίπτωση των πραγματικών δεδομένων παρουσιάζονται και συγκρίνονται οι επιδόσεις των 3 αλγορίθμων, δηλαδή Majority, Smoothing και Two-Phase. Δίνεται ιδιαίτερο βάρος στη σύγκριση των αλγορίθμων Smoothing και Two-Phase ώστε να διαπιστωθεί αν όντως οι καινούριες ιδέες που προτείνονται επιτυχγάνουν να βελτιώσουν τα αποτελέσματα της κατάτμησης και κατηγοριοποίησης που εκτελεί ο Smoothing. Διαπιστώνεται ότι ο Two-Phase παράγει ένα Transcription με αισθητά καλύτερη ποιότητα, δηλαδή περιέχει λιγότερα υποτμήματα και οι μεγάλη πλειοψηφία των υποτμημάτων του έχουν μήκος αρκετά μεγάλο ώστε να μπορούν να χρησιμοποιηθούν από κάποιο επόμενο σύστημα αναγνώρισης. Αντιθέτως, ο Smoothing παράγει υπερβολικά κατατμημένα Transcriptions όπου η πλειοψηφία των τμημάτων είναι κάτω των 0.4sec ενώ πολλά τμήματα είναι κάτω των 0.1sec. Ουσιαστικά το transcription του Smoothing χρειάζεται σημαντική μετα-επεξεργασία και ομαλοποίηση ώστε να μπορεί να χρησιμοποιηθεί. Όσον αφορά τα ποσοστά απόδοσης %Correct και %Accurate, οι δύο αλγόριθμοι επιτυχγάνουν συγκρίσιμα ποσοστά απόδοσης, με ένα ελαφρό πλεονέκτημα του Two-Phase ιδιαίτερα μετά από κάποια μικρή μεταεπεξεργασία των transcriptions με απλούς κανόνες. Κατά συνέπεια, ο αλγόριθμος Two-Phase φαίνεται ικανός να βελτιώσει την απόδοση του συστήματος κατάτμησης/κατηγοριοποίησης

Επίσης παρατηρούμε ότι η απόδοση του συστήματος για το πρόβλημα κατηγοριοποίησης σε ομιλία και μη ομιλία είναι πολύ καλύτερη από την απόδοση για το πρόβλημα κατηγοριοποίησης σε αντρική ομιλία, γυναικεία ομιλία και μη ομιλία. Αυτό συμβαίνει επειδή το δεύτερο πρόβλημα είναι πιο δύσκολο και επειδή τα διαθέσιμα δεδομένα για το δεύτερο πρόβλημα είναι λιγότερα. Εντούτοις, το σύστημα και οι τεχνικές που περιγράφονται στο παρόν κεφάλαιο θα μπορούσαν να χρησιμοποιηθούν ως βάση ώστε να δημιουργηθεί ένα επεκταμένο και περισσότερο αποδοτικό σύστημα, όταν περισσότερα δεδομένα γίνουν διαθεσιμα.

Εντέλει, το παρόν κεφάλαιο επιτυγχάνει να παρουσιάσει ένα λειτουργικό σύστημα για κατάτμηση και κατηγοριοποίηση δελτίων ειδήσεων βασισμένο σε HMM/GMM μοντέλα. Επίσης εισάγονται νέες ιδέες, όπως οι καμπύλες ποσοστών, και νέοι αλγόριθμοι, που φαίνεται ότι μπορούν να βελτιώσουν την απόδοση του συστήματος.

Κεφάλαιο 6

Περιγραφή του Συνολικού Συστήματος Κατάτμησης και Κατηγοριοποίησης Ηχητικών Τμημάτων

6.1 Σκοπός

Σκοπός του κεφαλαίου αυτού είναι να περιγράψει συνοπτικά ένα σύστημα που να εκτελεί κατάτμηση και κατηγοριοποίηση σε ένα ηχητικό τμήμα εισόδου. Το σύστημα αυτό θα συνδυάζει το υποσύστημα κατάτμησης με χρήση μετρικών κριτηρίων και το υποσύστημα κατάτμησης και κατηγοριοποίησης με χρήση HMM και GMM μοντέλων, τα οποία παρουσιάστηκαν αναλυτικά στα κεφάλαια 4 και 5 αντίστοιχα. Σκοπός του συνολικού συστήματος είναι η παραγωγή μίς κατάτμησης του ηχητικού τμήματος εισόδου και η κατηγοριοποίηση των υποτμημάτων σε μία από τις διαθέσιμες κλάσεις ανάλογα με την εφαρμογή.

Στη συνέχεια θα δώσουμε ένα σχεδιάγραμμα του συστήματος και θα εξηγήσουμε συνοπτικά τις υπομοναδες που το αποτελούν. Εκτός από τα υποσυστήματα κατάτμησης με μετρικά κριτήρια και κατάτμησης/κατηγοριοποίησης με στατιστικά μοντέλα, το συνολικό σύστημα θα περιέχει και ένα τελευταίο υποσύστημα το οποίο θα εφαρμόζει κανόνες για την μετα-επεξεργασία των αποτελεσμάτων των 2 προηγούμενων σταδίων. Ουσιαστικά ο ρόλος του υποσυστήματος εφαρμογής κανόνων είναι να συνδυάσει όλη τη γνώση που είναι διαθέσιμη για το πρόβλημα μέσω των αλγορίθμων των 2 πρώτων σταδίων και να παράγει έξυπνους και γενικούς κανόνες για την βελτίωση του τελικού αποτελέσματος. Η εύρεση τέτοιων γενικών συνδυαστικών κανόνων είναι ένα δύσκολο πρόβλημα και στο παρόν κεφάλαιο θα παρουσιαστούν απλά κάποιες ιδέες προς αυτή την κατεύθυνση.

Τελικά, γίνεται ένας σχολιασμός του γενικού συστήματος και παρουσιάζονται πιθανές μελλοντικές επεκτάσεις του συστήματος και ιδέες για τον αποδοτικότερο συνδυασμό των υποσυστημάτων του και την καλύτερη αξιοποίηση της γνώσης που προσφέρουν οι επιμέρους υπομονάδες.

6.2 Συνοπτική Περιγραφή των Υπομονάδων του Συστήματος

Στην ενότητα αυτή περιγράφονται συνοπτικά οι διάφορες υπομονάδες του συστήματος κατάτμησης και κατηγοριοποίησης ηχητικών τμημάτων

6.2.1 Ένα διάγραμμα του συνολικού συστήματος

Στην εικόνα 6.1 φαίνεται ένα block διάγραμμα του συνολικού συστήματος. Τα διάφορα υποσυστήματα αναλύονται στις ακόλουθες ενότητες.

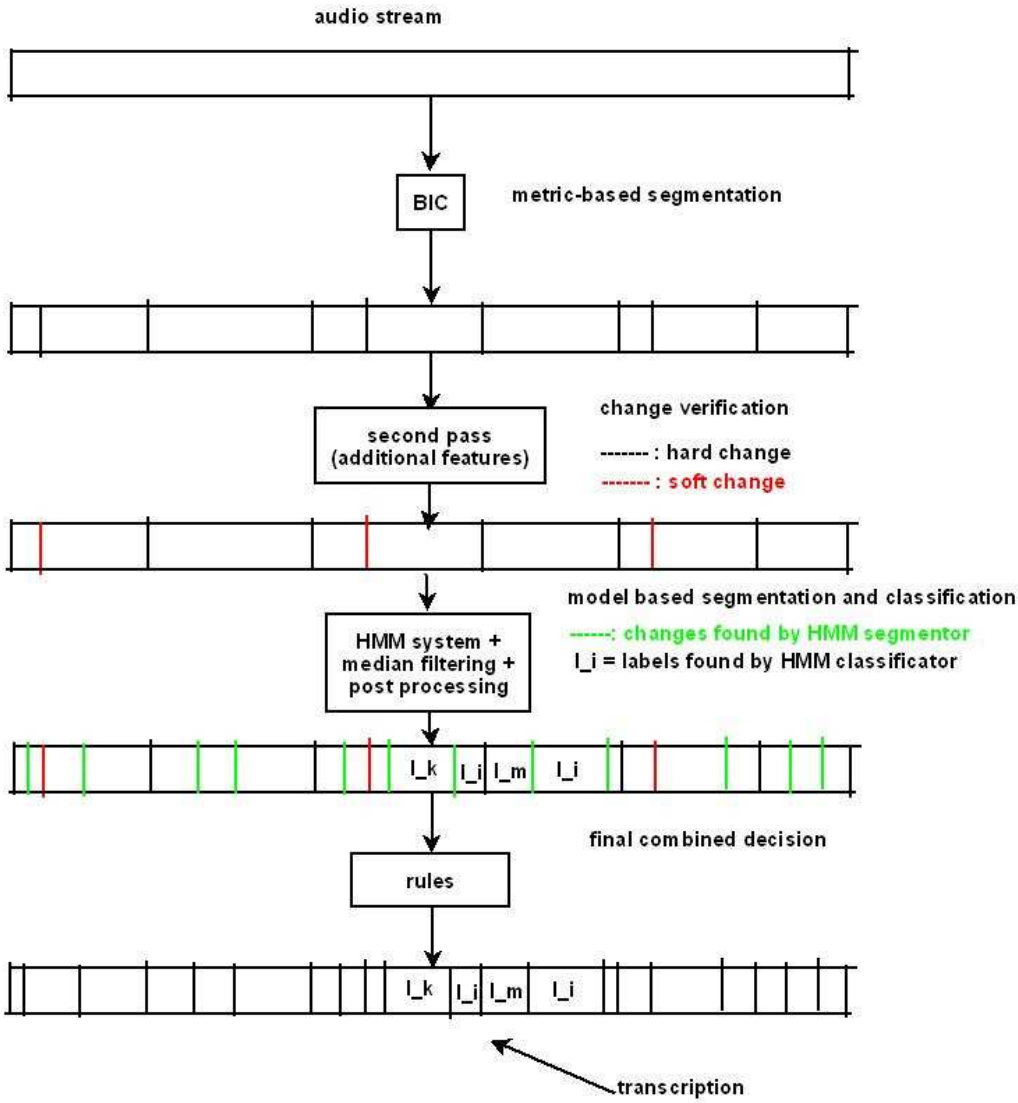
6.2.2 Το υποσύστημα κατάτμησης με χρήση μετρικών κριτηρίων

Το υποσύστημα αυτό περιέχει τις μονάδες BIC και second pass που φαίνονται στο σχήμα 6.1. Είσοδος του υποσυστήματος αυτού είναι το συνολικό audio-stream. Η μονάδα second pass είναι προαιρετική. Με βάση την ανάλυση που έχει γίνει στο κεφάλαιο 4 περιγράφουμε 3 πιθανούς εναλλακτικούς αλγόριθμους για το υποσύστημα αυτό.

1. Χρήση μόνο της μονάδας BIC για την εύρεση των σημείων αλλαγών του audio-stream εισόδου. Η χρήση μόνο του BIC χωρίς το στάδιο second pass αναμένεται να μας δώσει αποτέλεσμα με περισσότερες κατατμήσεις από τις πραγματικές, δηλαδή με υψηλό false alarm. Όλες οι αλλαγές χαρακτηρίζονται ως hard changes.

2. Χρήση της μονάδας BIC για εύρεση αλλαγών και της μονάδας second pass για επικύρωση των αλλαγών που βρέθηκαν. Το στάδιο second pass χρησιμοποιεί τον αλγόριθμο δεύτερου περάσματος με χρήση ενός μονοδιάστατου χαρακτηριστικού, όπως αυτός περιγράφηκε στην ενότητα 4.4.2. Οι αλλαγές του πρώτου περάσματος που δεν περνούν το κριτήριο πιθανότητας του δεύτερου περασματος σημειώνονται ως soft changes ενώ οι αλλαγές που περνούν το κριτήριο πιθανότητας σημειώνονται ως hard changes.

3. Χρήση της μονάδας BIC για εύρεση αλλαγών και της μονάδας second pass για επικύρωση των αλλαγών που βρέθηκαν. Η μόνη διαφορά με την προηγούμενη περίπτωση



Εικόνα 6.1: Block διάγραμμα του συνολικού συστήματος.

είναι ότι το στάδιο second pass χρησιμοποιεί τον γενικευμένο αλγόριθμο δεύτερου περάσματος με χρήση πολλών μονοδιάστατων χαρακτηριστικών. Έτσι ενσωματώνεται πληροφορία πολλών χαρακτηριστικών κατά τη διαδικασία επικύρωσης μίας αλλαγής. Οι αλλαγές του πρώτου περάσματος που δεν περνούν το κριτήριο πιθανότητας του δεύτερου περάσματος σημειώνονται ως soft changes ενώ οι αλλαγές που περνούν το κριτήριο πιθανότητας σημειώνονται ως hard changes.

Παρατηρούμε ότι το δεύτερο πέρασμα, όπου χρησιμοποιείται, δεν απορρίπτει αλλαγές, απλά τις σημειώνει ως soft changes, προσθέτοντας έτσι επιπλέον γνώση σχετικά με αυτές. Αυτό συμβαίνει επειδή δεν θέλουμε να απορρίψουμε αλλαγές τόσο νωρίς στη διαδικασία της κατάτμησης και επιλέγουμε να αφήσουμε την οριστική απόφαση για απόρριψη κάποιας αλλαγής στο τελικό στάδιο. Παρόλα αυτά η σήμανση κάποιας αλλαγής ως soft change σημαίνει ότι η αλλαγή αυτή είναι πιθανόν false alarm ή αλλαγή μεταξύ ομιλητών (κατηγορίας 3). Αυτή η πληροφορία θα μπορούσε να φανεί χρήσιμη σε κάποιο επόμενο στάδιο.

6.2.3 Το υποσύστημα κατάτμησης/κατηγοριοποίησης με χρήση HMM και GMM Μοντέλων

Το υποσύστημα αυτό αποτελείται από τη μοναδα HMM system και τις προαιρετικές μονάδες median filtering και post-proessing που φαίνονται στην εικόνα 6.1. Είσοδος του υποσυστήματος αυτού είναι τα υποτμήματα του τμήματος εισόδου που ορίζονται από κάθε 2 διαδοχικές αλλαγές που βρέθηκαν στο προηγούμενο στάδιο. Οι αλλαγές αυτές μπορεί να είναι είτε soft είτε hard changes, ουσιαστικά δηλαδή είναι όλες οι αλλαγές που βρίσκονται από την υπομονάδα BIC. Με βάση την ανάλυση που έχει γίνει στο κεφάλαιο 5 περιγράφουμε 3 πιθανούς εναλλακτικούς αλγόριθμους για το υποσύστημα αυτό.

1. Χρήση του αλγορίθμου Majority που περιγράφεται στην ενότητα 5.4. Το υποσύστημα στη περίπτωση αυτή δεν επιχειρεί επιπλέον κατάτμηση των υποτμημάτων που δέχεται, αλλά μόνο τα κατηγοριοποιεί στην πιο ταιριαστή διαθέσιμη κλάση. Ο αλγόριθμος αυτός κάνει την παραδοχή ότι τα υποτμήματα που δέχεται είναι ομογενή ως προς τις διαθέσιμες κλάσεις και αν η παραδοχή αυτή δεν ισχύει για το πρόβλημα υπό εξέταση η απόδοση του αλγορίθμου αυτού θα είναι χαμηλή.

2. Χρήση του αλγορίθμου Smoothing που περιγράφεται στην ενότητα 5.5.3. Το υποσύστημα στην περίπτωση αυτή εκτός από κατηγοριοποίηση, επιχειρεί επιπλέον κατάτμηση των υποτμημάτων που δέχεται με βάση τα αποτελέσματα κατάταξης των HMM

μοντέλων. Για την ομαλοποίηση των αποτελεσμάτων κατάταξης των μοντέλων χρησιμοποιείται απλό median filtering. Η εφαρμογή του αλγορίθμου αυτού μπορεί να οδηγήσει σε υπερκατατμημένο αποτέλεσμα με πολλά υποτμήματα μικρού μήκους.

3. Χρήση του αλγορίθμου Two-Phase που περιγράφεται στην ενότητα 5.5.4.Το υποσύστημα στην περίπτωση αυτή εκτός από κατηγοριοποίηση, επιχειρεί επιπλέον κατάτμηση των υποτμημάτων που δέχεται με βάση τα αποτελέσματα κατάταξης των HMM μοντέλων. Για την ομαλοποίηση των αποτελεσμάτων κατάταξης των μοντέλων επεκτείνεται ο αλγόριθμος Smoothing. Συγκεκριμένα χρησιμοποιείται αρχικά median filtering και στη συνέχεια δημιουργούνται και επεξεργάζονται οι καμπύλες ποσοστών σύμφωνα με τις ιδέες και τους αλγορίθμους που παρουσιάζονται στην ενότητα 5.5.4 (post - processing). Ο αλγόριθμος αυτός παράγει ένα πολύ πιο ομαλό transcription από αυτό του αλγορίθμου Smoothing.

Σημειώνουμε τέλος ότι στην περίπτωση που χρησιμοποιούνται οι αλγόριθμοι 2 ή 3, οι επιπλέον αλλαγές που βρίσκονται σημειώνονται ως αλλαγές του HMM system.

6.2.4 Το υποσύστημα εφαρμογής κανόνων

Το υποσύστημα αυτό δέχεται ως είσοδο ένα μία κατάτμηση του ηχητικού τμήματος εισόσου όπου τα κάθε υποτμήμα έχει κατηγοριοποιηθεί σε κάποια κλάση. Επίσης, πέρα από την κατηγοριοποίηση των υποτμημάτων το στάδιο αυτό λαμβάνει επιπλέον πληροφορίες για τα σημεία αλλαγής που δέχεται. Οι πληροφορίες εξαρτώνται από τους αλγόριθμους που χρησιμοποιήθηκαν στα 2 προηγούμενα στάδια και αφορούν το κατά πόσο ένα σημείο αλλαγής αποτελεί soft ή hard change ή αλλαγή που βρέθηκε από το HMM system.

Ο ρόλος του υποσυστήματος είναι να συνδυάσει με κάποιους ευριστικούς κανόνες την πληροφορία που δέχεται από τα προηγούμενα στάδια ώστε να παράγει ένα ομαλό και όσο το δυνατόν περισσότερο σωστό transcription. Λέγοντας ομαλό εννοούμε ότι το transcription εξόδου θα πρέπει να περιέχει σχετικά μεγάλα υποτμήματα και να μην περιέχει υπερβολικά μικρά διαδοχικά υποτμήματα. Ο χαρακτηρισμός μικρό και μεγάλο εξαρτάται από την εκάστοτε εφαρμογή, πάντως για δελτία ειδήσεων τμήματα με μήκος μικρότερο των 0.5sec δεν δεν είναι αρκετά μεγάλα ώστε χρησιμοποιηθούν από ένα επόμενο σύστημα αναγνώρισης ομιλίας.

Πρέπει να σημειωθεί ότι η ευριστικοί κανόνες αυτού του σταδίου πρέπει να είναι όσο το δυνατόν πιο γενικοί ώστε να μπορούν να εφαρμοστούν με ασφάλεια και να μην εξαρτώνται από τις ιδιαιτερότητες κάποιων μεμονωμένων παραδειγμάτων εισόδου.

Στο σημείο αυτό θα αναφέρουμε κάποιους προβληματισμούς σχετικά με τους κανόνες του σταδίου αυτού, κάποιες απλές ιδέες κανόνων και κάποιες γενικές κατευθύνσεις. Το θέμα αποτελεσματικού συνδυασμού της γνώσης που προέκυψε από τα προηγούμενα στάδια είναι δύσκολο και προσφέρεται για μελλοντική μελέτη.

Καταρχήν, ένα ερώτημα που τίθεται είναι κατά πόσο η πληροφορία που μας προσφέρουν τα στάδια σχετικά με τις αλλαγές (soft/hard changes, HMM changes) είναι ή όχι περιττή (redundant) τη στιγμή που έχουμε επιπλέον και μία κατηγοριοποίηση των υποτμημάτων δεξιά και αριστερά μίας αλλαγής. Για παράδειγμα αν κάποια αλλαγή είναι Hard change Θα περιμέναμε τα υποτμήματα γύρω από αυτήν να μην ανήκουν και τα δύο σε κλάση ομιλίας καθότι οι hard αλλαγές είναι με μεγάλη πιθανότητα αλλαγές μεταξύ ομιλίας και μη ομιλίας. Με αυτή την έννοια ο όρος hard change θα μπορούσε να χρησιμοποιηθεί για να επιβεβαιώσει τα αποτελέσματα της κατηγοριοποίησης του HMM system. Αν όμως για *όλες* τις hard αλλαγές που βρίσουμε ισχύει ότι το HMM system κατηγοριοποιεί τα γειτονικά υποτμήματα σε ομιλία και μη ομιλία, τότε η πληροφορία hard change είναι περιττή καθώς το HMM system μας δίνει την ίδια πληροφορία.

Σημειώνουμε ότι οι χρησιμότητα τέτοιων πληροφοριών σχετικά με τις αλλαγές αυξάνει όσο το πρόβλημα κατηγοριοποίησης γίνεται πιο σύνθετο και αυξάνονται οι διαθέσιμες κλάσεις του προβλήματος. Έτσι αν και την περίπτωση κατάταξης μεταξύ ομιλίας και μη ομιλίας τέτοιοι κανόνες μπορεί να φαίνονται περιττοί, αν έχουμε στο πρόβλημά μας κλάσεις όπως αντρική/γυναικεία ομιλία, θορυβώδη/καθαρή ομιλία ή ακόμα και κλάσεις που να αντιστοιχούν στην ομιλία κάποιου κεντρικού ομιλητή ενός δελτίου ειδήσεων, οι κανόνες μπορούν να φανούν χρήσιμοι. Για παράδειγμα αν τα τμήματα γύρω από ένα σημείο hard change έχουν κατηγοριοποιηθεί στη γυναικεία ομιλία, πιθανόν να υπάρχει αλλαγή ομιλητών μεταξύ δύο γυναικών και πιθανόν να υπάρχει αλλαγή στο επίπεδο του θορύβου, έτσι να έχουμε καθαρή ομιλία και θορυβώδη ομιλία στα 2 γειτονικά τμήματα.

Τέλος αναφέρουμε κάποιους απλούν κανόνες για το σύστημα. Κάποιοι από του κανόνες αυτούς βασίζονται στο γεγονός ότι ο αλγόριθμος Two-Phase του δεύτερου σταδίου αφήνει την ελευθερία να οριστούν No decision τμήματα, και έτσι οι κανόνες προσπαθούν να χειριστούν αυτά τα No decision τμήματα.

1. Συγχωνεύουμε διαδοχικά τμήματα που έχουν κατηγοριοποιηθεί σε κλάσεις μη ομιλίας

2. Αν ένα τμήμα είναι μικρότερο από κάποιο κατώφλι και τα δύο γειτονικά του είναι μεγαλύτερα από κάποιο κατώφλι και ανήκουν σε κλάση μη ομιλίας, τότε διαγράφουμε το μικρό τμήμα και συγχωνεύουμε τα 2 γειτονικά του.

3. Αν ένα τμήμα είναι μικρότερο από κάποιο κατώφλι και τα δύο γειτονικά του είναι

μεγαλύτερα από κάποιο κατώφλι και ανήκουν σε κλάσεις ομιλίας, τότε διαγράφουμε το μικρό τμήμα και τοποθετούμε το σημείο αλλαγής των γειτονικών στη μέση του τμήματος που διαγράψαμε.

4. Αν δεξιά και αριστερά ενός ασθενούς σημείου αλλαγής (soft change) οι περιοχές ανήκουν στην ίδια κλάση, τι σημείο εξαφανίζεται και οι περιοχές συγχωνεύονται.

5. Αν από τη μία πλευρά ενός ισχυρού σημείου αλλαγής (hard change) η περιοχή είναι no decision και από την άλλη πλευρά έχουμε θόρυβο ή σιωπή, αναγνωρίζουμε το no decision τμήμα ως ομιλία.

6. Αν από τη μία πλευρά ενός ισχυρού σημείου αλλαγής η περιοχή είναι no decision και από την άλλη πλευρά έχουμε ομιλία, χρησιμοποιούμε κάποιο κατώφλι ενέργειας για να κατατάξουμε την No decision περιοχή σε θόρυβο ή σιωπή

7. Μετακινούμε ελαφρώς τα σημεία αλλαγής, μέσα σε μία μικρή περιοχή, ώστε να βρίσκονται σε σημεία που το σήμα έχει, αν είναι δυνατόν, κάποιο ελάχιστο ενέργειας.

6.3 Συμπεράσματα και Σχόλια

Στην ενότητα αυτή παρουσιάστηκε συνοπτικά ένα λειτουργικό σύστημα που δέχεται ένα audio-stream εισόδου και χρησιμοποιεί μετρικά κριτήρια για να το χωρίσει σε υποτμήματα ενώ χρησιμοποιεί στατιστικά μοντέλα για να επιτύχει περαιτέρω κατάτμηση και κατηγοριοποίηση των τμημάτων.

Το σύστημα αυτό δοκιμάστηκε σε κάποιες από τις παραλλαγές του στην ενότητα των πειραματικών αποτελεσμάτων 5.6.2 και λειτουργεί ικανοποιητικά για απλές κατατάξεις μεταξύ κλάσεων ομιλίας και μη ομιλίας. Για πιο λεπτομερείς κατατάξεις η απόδοση του συστήματος πέφτει, αυτό όμως οφείλεται σε μεγάλο βαθμό και στο μικρό πλήθος των διαθέσιμων δεδομένων.

Οι αλγόριθμοι που περιγράφηκαν θα μπορούσαν να αποτελέσουν τη βάση για την ανάπτυξη ενός εύρωστου συστήματος κατάτμησης και κατηγοριοποίησης που θα εκτελεί λεπτομερή κατάταξη. Για παράδειγμα θα μπορούσαν να προστεθούν επιπλέον στάδια κατάταξης στο σύστημα, ώστε να διαχωρίζει την κλάση της μη ομιλίας σε σιωπή και θόρυβο και την κλάση της ομιλίας σε αντρική και γυναικεία, καθαρή ή θορυβώδη. Επιπλέον, η ύπαρξη μεγάλου πλήθους δεδομένων εκπαίδευσης θα καθιστούσε δυνατή την εκπαίδευση μοντέλων/κλάσεων για κάποιους συχνούς ομιλητές δελτίων ειδήσεων, όπως κάποιοι επιφανείς πολιτικοί ή οι κεντρικοί εκφωνητές των δελτίων ειδήσεων.

Επιπλέον, ένα ζήτημα που αξίζει περαιτέρω μελέτη είναι ο αποτελεσματικός συνδυασμός της γνώσης που μας προσφέρουν οι διάφορες υπομονάδες του συστήματος, μέσω εφαρμογής γενικών ευριστικών κανόνων.

Στόχος του κεφαλαίου αυτού ήταν η συνοπτική παρουσίαση ενός συνολικού συστήματος κατάτμησης και κατηγοριοποίησης audio-stream και η διατύπωση προβληματισμών και ιδεών για την μελλοντική επέκταση και βελτίωση του συστήματος.

Κεφάλαιο 7

Συμπεράσματα και Μελλοντικές Επεκτάσεις του Συστήματος

Στην παρούσα εργασία ορίστηκαν και μελετήθηκαν αναλυτικά τα προβλήματα κατάτμησης ηχητικών αρχείων και κατηγοριοποίησης των υποτμημάτων με χρήση στατιστικών μοντέλων. Σκοπός ήταν να μελετηθούν με πληρότητα οι δύο παραπάνω περιοχές, να υλοποιηθούν κάποιοι από τους αλγόριθμους που προτείνονται στη βιβλιογραφία και να δοκιμαστεί η απόδοση κάποιων παραλλαγών υπαρκτών μεθόδων αλλά και κάποιων νέων ιδεών που προτείνονται στην εργασία αυτή. Τελικός στόχος είναι η υλοποίηση ενός αποδοτικού συστήματος που θα μπορεί να εφαρμοστεί για το πρόβλημα diarization με πραγματικά δεδομένα από δελτία ειδήσεων.

Αρχικά έγινε μία εισαγωγή στο πρόβλημα της επεξεργασίας ηχητικών δελτίων και αναφέρθηκαν οι προκλήσεις, οι δυσκολίες αλλά και οι πρακτικές εφαρμογές που σχετίζονται με αυτή την περιοχή. Ακολούθησε μία σύντομη περιγραφή του θεωρητικού υπόβαθρου που απαιτείται για την κατανόηση της εργασίας αυτής, η οποία επικεντρώθηκε σε θέματα ορισμού στατιστικών μοντέλων όπως τα Κρυφά Μαρκοβιανά Μοντέλα αλλά και περιγραφής των προβλημάτων που σχετίζονται με τα μοντέλα αυτά. Ακολούθως, παρουσιάστηκε το State of the Art του προβλήματος, τόσο για την περιοχή της κατάτμησης ηχητικών αρχείων όσο και για την περιοχή της κατηγοριοποίησής τους με χρήση στατιστικών μοντέλων. Επίσης παρουσιάστηκαν παραδείγματα συστημάτων κατάτμησης και κατηγοριοποίησης που συναντώνται στη βιβλιογραφία.

Στην ενότητα που μελετά την κατάτμηση ηχητικών τμημάτων επιχειρήθηκε μία όσο το δυνατόν πιο πλήρης ανάλυση του προβλήματος της κατάτμησης audio-streams με έμφαση

στη μελέτη χαρακτηριστικών, στη μελέτη μετρικών κριτηρίων κατάτμησης (metric-based) και στην υλοποίηση αλγορίθμων κατάτμησης.

Όσον αφορά τη μελέτη χαρακτηριστικών, παρουσιάστηκαν πλήθος μονοδιάστατων και πολυδιάστατων χαρακτηριστικών που χρησιμοποιούνται στη βιβλιογραφία. Στη συνέχεια παρουσιάστηκαν διάφορα μετρικά κριτήρια για τον εντοπισμό αλλαγών σε audio-streams και παρουσιάστηκαν γραφικά τα αποτελέσματα των κριτηρίων αυτών για κάποιες περιπτώσεις αλλαγών.

Ακολούθως παρουσιάστηκαν με λεπτομέρεια οι 2 αλγόριθμοι εντοπισμού αλλαγών που υλοποιήθηκαν και χρησιμοποιήθηκαν στα πειράματα. Ο αλγόριθμος πρώτου περάσματος χρησιμοποιεί το κριτήριο BIC ή προσεγγίσεις του και προτάθηκε στα [19] και [8]. Ο αλγόριθμος αυτός εντοπίζει πιθανά σημεία αλλαγής.

Ο αλγόριθμος δεύτερου περάσματος είναι μία καινούρια ιδέα που βασίζεται στον υπολογισμό πιθανότητας αλλαγής σε ένα σημείο, που προτείνεται στο [38]. Ο αλγόριθμος αυτός αποφασίζει αν ένα σημείο αλλαγής που βρέθηκε από το πρώτο πέρασμα αντιστοιχεί σε πραγματικό σημείο αλλαγής. Σκοπός του αλγορίθμου είναι να μειωθεί το υψηλό ποσοστό false alarm που συνήθως παίρνουμε από το πρώτο πέρασμα αλλά και να χρησιμοποιηθούν στη διαδικασία εντοπισμού αλλαγών μονοδιάστατα χαρακτηριστικά όπως η Fractal διάσταση και η RMS τιμή, που είναι δύσκολο να συνδυαστούν με τα πολυδιάστατα χαρακτηριστικά του πρώτου περάσματος.

Έγιναν πειράματα που μελετούν τη συμπεριφορά των διάφορων χαρακτηριστικών που μπορούν να χρησιμοποιηθούν τόσο στο πρώτο όσο και στο δεύτερο πέρασμα. Επίσης, τα πειράματα εξέτασαν κατά πόσο το δεύτερο πέρασμα βελτιώνει την απόδοση του πρώτου περάσματος. Τα αποτελέσματα που εκτελέστηκαν υποδεικνύουν ότι το δεύτερο πέρασμα είτε με ένα είτε με πολλά χαρακτηριστικά μπορεί να οδηγήσει σε βελτίωση των ποσοστών Success και False Alarm

Στη συνέχεια, στο κεφάλαιο 5, επιχειρήθηκε μία μελέτη του προβλήματος στατιστικής μοντελοποίησης ηχητικών κλάσεων όπως ομιλία, θόρυβος, σιωπή, με χρήση HMM και GMM μοντέλων. Επιδίωξή μας ήταν η κατασκευή ενός συστήματος ικανού να εκτελεί κατάτμηση και κατηγοριοποίηση ηχητικών τμημάτων με βάση κατάλληλα εκπαιδευμένα μοντέλα. Τελικός στόχος είναι η χρήση του συστήματος σε μία πραγματική εφαρμογή όπως η κατάτμηση και κατηγοριοποίηση δελτίων ειδήσεων σε κατάλληλες κλάσεις.

Πιο συγκεκριμένα, αρχικά έγινε μία περιγραφή των εργαλείων που προσφέρει η συλλογή προγραμμάτων HTK για την ανάπτυξη ενός συστήματος βασισμένου σε HMM και GMM μοντέλα. Στη συνέχεια, αναφέρθηκαν τα βασικά βήματα για τη σχεδίαση και την κατασκευή

ενός τέτοιου συστήματος και περιγράφηκαν κάποιες βασικές σχεδιαστικές αποφάσεις. Τέτοιες αποφάσεις αφορούν θέματα όπως οι κλάσεις του προβλήματος που θεωρήθηκαν και η δομή των HMM μοντέλων και εξαρτώνται από τη συγκεκριμένη εφαρμογή σε δελτία ειδήσεων αλλά από το πλήθος και την ποιότητα των διαθέσιμων δεδομένων εκπαίδευσης.

Ακολούθως, περιγράφηκε ένα απλό σύστημα που ασχολείται με ένα υποσύνολο του προβλήματος κατάτμησης και κατηγοριοποίησης, συγκεκριμένα μόνο με το πρόβλημα της κατηγοριοποίησης ομογενών ακουστικών τμημάτων με χρήση HMM. Συνεπώς το σύστημα που θα περιγραφεί παρακάτω δέχεται ως είσοδο ένα τμήμα audio-stream το οποίο θεωρείται ομογενές και το κατατάσσει σε κάποια από τις διαθέσιμες κλάσεις, συγκεκριμένα στην κλαση όπου έχει καταταχθεί η πλειοψηφία των frames του. Τα ομογενή τμήματα προέρχονται από το στάδιο εύρεσης αλλαγών σε audio-stream και κατάτμησής του, που έχει περιγραφεί αναλυτικά στο προηγούμενο κεφάλαιο. Ο αλγόριθμος που αναπτύχθηκε αναφέρεται ως Majority.

Στη συνέχεια επεκτείναμε το σύστημά μας ώστε να εξετάσουμε το συνολικό πρόβλημα κατάτμησης και κατηγοριοποίησης. Το σύστημα που αναπτύχθηκε επιχειρεί όχι μόνο να κατατάξει τα διαθέσιμα τμήματα σε μία από τις υπάρχουσες κλάσεις αλλά επιχειρεί και να κάνει περαιτέρω κατάτμηση τμημάτων που δεν είναι ομογενή. Για την κατάτμηση προτείνονται και εξετάστηκαν 2 αλγόριθμοι, ο αλγόριθμος Smoothing που βασίζεται σε απλό median filtering και περιγράφεται στην ενότητα 5.5.3 και ο αλγόριθμος Two-Phase που περιγράφεται στην ενότητα 5.5.4. Ο αλγόριθμος Two-Phase είναι μία καινούρια ιδέα που επεκτείνει το απλό median filtering χρησιμοποιώντας την έννοια των καμπύλων ποσοστών που εισάγονται στην παρούσα εργασία. Επίσης περιγράφηκαν αλγόριθμοι για την επεξεργασία των καμπύλων ποσοστών.

Στην ενότητα των πειραματικών αποτελεσμάτων έγιναν πειράματα τόσο σε συνθετικά όσο και σε πραγματικά δεδομένα από δελτία ειδήσεων. Στη περίπτωση των πραγματικών δεδομένων παρουσιάστηκαν και συγκρίθηκαν οι επιδόσεις των 3 αλγορίθμων, δηλαδή Majority, Smoothing και Two-Phase. Δόθηκε ιδιαίτερο βάρος στη σύγκριση των αλγορίθμων Smoothing και Two-Phase ώστε να διαπιστωθεί αν όντως οι καινούριες ιδέες που προτάθηκαν επιτυχγάνουν να βελτιώσουν τα αποτελέσματα της κατάτμησης και κατηγοριοποίησης που εκτελεί ο Smoothing. Διαπιστώθηκε ότι ο Two-Phase παράγει ένα Transcription με αισθητά καλύτερη ποιότητα, δηλαδή περιέχει λιγότερα υποτμήματα και οι μεγάλη πλειοψηφία των υποτμημάτων του έχουν μήκος αρκετά μεγάλο ώστε να μπορούν να χρησιμοποιηθούν από κάποιο επόμενο σύστημα αναγνώρισης. Αντιθέτως, ο Smoothing παράγει υπερβολικά κατατμημένα Transcriptions όπου η πλειοψηφία των τμημάτων είναι πολύ μικρά. Ουσιαστικά το transcription του Smoothing χρειάζεται σημαντική μετα-επεξεργασία και ομαλο-

ποίηση ώστε να μπορεί να χρησιμοποιηθεί. Όσον αφορά την απόδοση, οι δύο αλγόριθμοι επιτυχγάνουν συγκρίσιμα ποσοστά, με ένα ελαφρό πλεονέκτημα του Two-Phase ιδιαίτερα μετά από κάποια μικρή μεταεπεξεργασία των transcriptions με απλούς κανόνες. Κατά συνέπεια, ο νέος αλγόριθμος Two-Phase φαίνεται ικανός να βελτιώσει την απόδοση του συστήματος κατάτμησης/κατηγοριοποίησης

Τελικά, περιγράφηκε συνοπτικά ένα συνολικό σύστημα που εκτελεί κατάτμηση και κατηγοριοποίηση σε ένα ηχητικό τμήμα εισόδου. Το σύστημα αυτό συνδυάζει το υποσύστημα κατάτμησης με χρήση μετρικών κριτηρίων και το υποσύστημα κατάτμησης και κατηγοριοποίησης με χρήση HMM και GMM μοντέλων. Σκοπός του συνολικού συστήματος είναι η παραγωγή μίας κατάτμησης του ηχητικού τμήματος εισόδου και η κατηγοριοποίηση των υποτμημάτων σε μία από τις διαθέσιμες κλάσεις ανάλογα με την εφαρμογή.

Δόθηκε ένα σχεδιάγραμμα του συστήματος και εξηγήθηκαν συνοπτικά οι υπομοναδες που το αποτελούν. Εκτός από τα υποσυστήματα κατάτμησης με μετρικά κριτήρια και κατάτμησης/κατηγοριοποίησης με στατιστικά μοντέλα, το συνολικό σύστημα περιέχει και ένα τελευταίο υποσύστημα το οποίο εφαρμόζει κανόνες για την μετα-επεξεργασία των αποτελεσμάτων των 2 προηγούμενων σταδίων. Ουσιαστικά ο ρόλος του υποσυστήματος εφαρμογής κανόνων είναι να συνδυάσει όλη τη γνώση που είναι διαθέσιμη για το πρόβλημα μέσω των αλγορίθμων των 2 πρώτων σταδίων και να παράγει έξυπνους και γενικούς κανόνες για την βελτίωση του τελικού αποτελέσματος. Η εύρεση τέτοιων γενικών συνδυαστικών κανόνων είναι ένα δύσκολο πρόβλημα και στην παρούσα εργασία παρουσιάστηκαν απλά κάποιες ιδέες προς αυτή την κατεύθυνση.

Συμπερασματικά, η μελέτη των περιοχών κατάτμησης ηχητικών δελτίων και κατηγοριοποίηση τους με χρήση στατιστικών μοντέλων, μας οδήγησε στην υλοποίηση ενός λειτουργικού συστήματος κατάτμησης και κατηγοριοποίησης με εφαρμογή σε πραγματικά δεδομένα από δελτία ειδήσεων. Το σύστημα αυτό έχει ικανοποιητική απόδοση για απλά προβλήματα κατηγοριοποίησης, όπως η κατηγοριοποίηση σε ομιλία και μη ομιλία. Εντούτοις τα σύστημα θα μπορούσε να επεκταθεί ώστε να χειρίζεται αποτελεσματικά δυσκολότερα προβλήματα. Οι αλγόριθμοι και οι νέες ιδέες που περιγράφηκαν στις ενότητες 4 και 5, θα μπορούσαν να αποτελέσουν τη βάση για την ανάπτυξη ενός επεκταμένου συστήματος κατάτμησης και κατηγοριοποίησης που θα εκτελεί λεπτομερέστερη κατάταξη και θα επιτυγχάνει καλύτερη απόδοση.

Βιβλιογραφία

[1] Γ. Καραγιάννης and Γ.Σταινχαουερ. *Μάθηση Μηχανών και Αναγνώριση Προτύπων*. Εκδόσεις ΕΜΠ, 2001.

[2] J. Ajmera, I. McCowan, and H. Bourlard. Robust speaker change detection. *IEEE Signal Processing Letters*, Vol.11:pp.649–651, 2004.

[3] J. Ajmera, I. A. McCowan, and H. Bourlard. Robust hmm-based speech/music segmentation. *Proc. IEEE Int.Conf Acoust.,Speech, Signal Process.*, Vol.1:pp.297–300, 2002.

[4] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Trans. Audio, Speech and Lang. Process.*, Vol.14:pp.1505–1512, 2006.

[5] A. C. Bovik, P. Maragos, and T. F. Quatieri. Am-fm energy detection and separation in noise using multiband energy operators. *IEEE Trans. Sig. Process.*, 41:3245–3265, 1993.

[6] J. P. Campbell. Speaker recognition, a tutorial. *Technical Report, Departmant of Defence Fort Meade MD*, (8), 1997.

[7] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain. Speaker diarization from speech transcripts. *Proc. Int. Conf. Spoken Language Process.*, 2004.

[8] S. S. Chen and P.S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. *Technical Report, IBM Watson Research Center*, 1998.

[9] M. Delakis, G. Gravier, and P. Gros. Multimodal segmental-based modeling of tennis video broadcasts. *ICME*, 2005.

[10] L. Deng, J. Droppo, and A. Acero. Dynamic compensation of hmm variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. *Ieee Transactions On Speech And Audio Processing*, Vol.13:pp.412–421, 2005.

[11] D. Dimitriadis, P. Maragos, and A.Potamianos. Auditory teager energy cepstrum coefficients for robust speech recognition. *Proc. European Conf. on Speech Communication and Technology - Interspeech 2005, Lisbon, Portugal*, pages pp.3013–3016, 2005.

[12] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification, Chapter 3.* John Wiley & Sons, Inc., 2 edition, 1997.

[13] G. Evangelopoulos and P. Maragos. Multiband modulation energy tracking for noisy speech detection. *IEEE Trans. Audio, Speech and Lang. Process.*, Vol.14:pp.2024–2038, 2006.

[14] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. *Intelligent Enterprise Technologies Laboratory,HP Laboratories Palo Alto*, 2003.

[15] M. Figueireido and A. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.24:pp.381–396, 2002.

[16] J.-L. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communication*, Vol.37:p.89–108, 2002.

[17] G. Gravier, G. Potamianos, and C. Neti. Asynchrony modeling for audiovisual speech recognition. *Proc. Human Language Technology Conference, San Diego, California*, 2002.

[18] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland, and S.J. Young. Segment generation and clustering in the htk broadcast news transcription system. *Technical Report, Speech, Vision and Robotics Group, Cambridge University Engineering Department*, 1998.

[19] R. Huang and J. H. L. Hansen. Advances in unsupervised audio classification and segmentation for the broadcast news and ngsw corpora. *IEEE Trans. Audio, Speech and Lang. Process.*, Vol.14:pp.907–919, 2006.

[20] K. Joergensen, L. Moelgaard, and L. K. Hansen. Unsupervised speaker change detection for broadcast news segmentation. *Technical Report, Informatics and Mathematical Modelling, Technical University of Denmark.*

[21] T. C. Justus and J. J. Bharucha. Music perception and cognition. *Stevensŏ Handbook of Experimental Psychology, Volume 1: Sensation and Perception*, Vol.1:pp.453–492, 2002.

[22] A. Katsamanis, V. Pitsikalis, and P. Maragos. Report on the state-of-the-art,event detection, segmentation and classification for audio streams. *Technical Report, MUSCLE-ICCS NTUA*, 2004.

[23] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. *Proc. IEEE Int.Conf Acoust.,Speech, Signal Process.*, 2000.

[24] B. Kostek. Musical instrument classification and duet analysis employing music information retrieval techniques. *Proceedings of the IEEE, Invited Paper*, Vol.92:pp.712–729, 2004.

[25] S. Kwon and S. Narayanan. Speaker change detection using a new weighted distance measure. *ICSLP*, 2002.

[26] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, Vol.9:pp.171ŏ185, 1995.

[27] D. Liu and F. Kubala. Online speaker clustering. *Proc. IEEE Int.Conf. Acoust., Speech, Signal Process.*, 2003.

[28] L. Lu and H.-J. Zhang. Real-time unsupervised speaker change detection. *ICPR*, 2002.

[29] L. Lu, H.-J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *IEEE Trans., Speech, Audio Process.*, Vol.10:pp.504–516, 2002.

[30] I. Magrin-Chagnolleau, G. Gravier, M. Seck, O. Boeffard, R. Blouet, and F. Bimbot. A further investigation on speech features for speaker characterization. *ICSLP*, 2000.

[31] P. Maragos, J. F. Kaiser, and T. F. Quatieri. Energy separation in signal modulations with application to speech signals. *IEEE Transactions on Signal Processing*, Vol.41:pp.3024–3051, 1993.

[32] P. Maragos and A. Potamianos. Fractal dimensions of speech sounds: Computation and application to automatic speech recognition. *J.Acoust.Soc.Amer.*, Vol.105:pp.1925–1932, 1999.

[33] P. Maragos and F.-K. Sun. Measuring the fractal dimension of signals: Morphological covers and iterative optimization. *IEEE Transactions on Signal Processing*, Vol.41:pp.108–121, 1993.

[34] S. Meignier, J.-F. Bonastre, C. Fredouille, and T. Merlin. Evolutive hmm for multispeaker tracking system. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000.

[35] M. N. Murthi and B. D. Rao. All-pole modeling of speech based on the minimum variance distortionless response spectrum. *IEEE Trans., Speech, Audio Process.*, Vol.8:pp.221–239, 2000.

[36] P. Nguyen, L. Rigazio, Y. Moh, and J.-C. Junqua. Rich transcription 2002 site report :panasonic speech technology laboratory (pstl). *Technical Report, Panasonic Speech Technology Laboratory (Pstl)*, 2002.

[37] M. Ostendorf, V. Digalakis, and O.A.Owen. From hmm's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech and Audio Process.*, Vol.4:pp.360–378, 1996.

[38] C. Panagiotakis and G. Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE Transactions On Multimedia*, Vol.7:pp.155–166, 2005.

[39] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audio-visual speech. *Proceedings of the IEEE*, Vol. 91, 2003.

[40] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, Vol.10:pp.19–41, 2000.

[41] A. Samouelian, J. Robert-Ribes, and M. Plumpe. Speech, silence, music and noise classification of tv broadcast material. *ICLSP*, 1998.

[42] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pages pp.1331–1334, 1997.

[43] S.E.Tranter. Who really spoke when?finding speaker turns and identities in audio. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006.

[44] K. Tokuda, T. Kobayashi, and S.Imai T. Masuko. Mel-generalized cepstral analysisΰ a unified approach to speech spectral estimation. *ICSLP*, page 4, 1994.

[45] S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Trans. Audio, Speech Lang. Process.*, Vol.14:pp. 1557–1565, 2006.

[46] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. Speech, Audio Process.*, Vol.10:pp.293–302, 2002.

[47] H. D. Wactlar, A. G. Hauptmann, and M. J. Witbrock. Informedia tm: News-on-demand experiments in speech recognition. *Technical Report, School of Computer Science, Carnegie Mellon University*, 1996.

[48] G. Williams and D. P.W. Ellis. Speech/music discrimination based on posterior probability features. *Eurospeech 1999, Budapest*, 1999.

[49] P.C. Woodland. The development of the htk broadcast news transcription system: An overview. *Speech Communication*, Vol.37:pp.47–67, 2002.

[50] C. Wooters, J. Fung, B. Peskin, and X. Anguera. Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system. *EARS-RT*, 2004.

[51] U. H. Yapanel and J. H.L. Hansen. A new perspective on feature extraction for robust in-vehicle speech recognition. *Eurospeech 2003, Geneva*, pages pp.1281–1284, 2003.

[52] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book (for HTK Version 3.2.1)*. Cambridge University Engineering Department., 2002.