Τμηματοποίηση ήχου και κατηγοριοποίηση σε μουσική και ομιλία

Κώστας Παναγιωτάκης και Γιώργος Τζιρίτας

Τεχνική αναφορά
CSD-TR-2000-02

24 Νοεμβρίου 2000

A Speech/Music discriminator based on RMS and zero-crossings

C. Panagiotakis and G. Tziritas

Tecnical Report
CSD-TR-2000-02

24th November 2000

Περίληψη

Τα τελευταία χρόνια γίνεται μεγάλη προσπάθεια να εξαχθεί πληροφορία από οπτικοακουστικά μέσα, ώστε να είναι δυνατή η περιγραφή του περιεχομένου των. Μ' αυτό τον τρόπο μπορούν να καταχωρηθούν σε βάσεις δεδομένων και να ανακαλούνται αυτόματα με βάση το περιεχόμενό των.

Στην παρούσα εργασία αντιμετωπίζεται ο χαρακτηρισμός ενός ηχητικού σήματος που είτε αποτελεί μέρος ενός οπτικοακουστικού προγράμματος, είτε υφίσταται αυτόνομα για παράδειγμα καταγραμμένο σ' ένα ακουστικό ψηφιακό δίσκο. Σκοπός μας ήταν να αναπτυχθεί ένα σύστημα πρώτα τμηματοποίησης του ηχητικού σήματος, και έπειτα κατηγοριοποίησης σε δύο κύριες κατηγορίες: ομιλία και μουσική. Μεταξύ των απαιτήσεων συμπεριλαμβάνεται η ταχύτητα της επεξεργασίας και η απόκριση του συστήματος σε πραγματικό χρόνο. Λόγω του περιορισμού σε δύο μόνο κλάσεις τα χαρακτηριστικά που εξάγονται περιορίζονται σημαντικά και επιπλέον δεν απαιτούν πολύπλοκους υπολογισμούς. Ο πειραματικός έλεγχος έδειξε ότι οι επιδόσεις είναι εξαιρετικές, χωρίς να θυσιασθεί η απόδοση του συστήματος.

Η τμηματοποίηση βασίζεται στην κατανομή του πλάτους του σήματος. ενώ στην ταξινόμηση έγινε επιπλέον χρήση ενός χαρακτηριστικού που σχετίζεται με την συχνότητα του σήματος. Ο ταξινομητής μπορεί να χρησιμοποιηθεί είτε σε συνδυασμό με την τμηματοποίηση, οπότε επιβεβαιώνει ή διαψεύδει μία αλλαγή τύπου μουσική/ομιλία ή ομιλία/μουσική, είτε αυτόνομα, σε δοσμένα τμήματα ήχου. Τα βασικά χαρακτηριστικά υπολογίζονται σε διαστήματα 20 msec, με αποτέλεσμα τα όρια των τμημάτων να προσδιορίζονται με ακρίβεια 20 msec. Η ελάχιστη διάρκεια των τμημάτων τίθεται στο ένα δευτερόλεπτο.

Οι αλγόριθμοι τμηματοποίησης και κατηγοριοποίησης δοκιμάσθηκαν σε μία μεγάλη βάση από δεδομένα, με ποσοστά επιτυχούς τμηματοποίησης που ανέρχονται σε 97% και επιτυχούς ταξινόμησης κοντά στο 95%.

Abstract

Over the last years major efforts have been made to develop methods for extracting information from audio-visual media, in order that they may be stored and retrieved in databases automatically, based on their content.

In this work we deal with the characterization of an audio signal, which may be part of a larger audiovisual system or may be autonomous, as for example in the case of an audio recording stored digitally on disk. Our goal was to first develop a system for segmentation of the audio signal, and then classification into one of two main categories: speech or music. Among the system's requirements are its processing speed and its ability to function in a real time environment. Because of the restriction to two classes, the characteristics that are extracted are considerably reduced and moreover the required computations are straightforward. Experimental results show that efficiency is exceptionally good, without sacrificing performance.

Segmentation is based on mean signal amplitude distribution, whereas classification utilizes an additional characteristic related to the frequency. The classification algorithm may be used either in conjunction with the segmentation algorithm, in which case it verifies or refutes a music-speech or speech-music change, or autonomously, with given audio segments. The basic characteristics are computed in 20 msec intervals, resulting in the segments' limits being specified within an accuracy of 20 msec. The smallest segment length is one second.

The segmentation and classification algorithms were benchmarked on a large data set, with correct segmentation about 97% of the time and correct classification about 95%.

Κεφάλαιο 1

Εισαγωγή

1.1 Ορισμός προβλήματος

Τα προβλήματα της τμηματοποίησης και της ταξινόμησης είναι ανάμεσα στα πιο ενδιαφέροντα προβλήματα της ανάλυσης σημάτων και της αναγνώρισης προτύπων. Σε πολλές εφαρμογές ενδιαφέρει να τμηματοποιηθούν και να χαρακτηριστούν ανάλογα με το περιεχόμενο τους σήματα ήχου. Ένας αρχικός διαχωρισμός θα ήταν σε μουσική ή ομιλία ή κενά διαστήματα. Η ταξινόμηση θα μπορούσε ιεραρχικά να συνεχισθεί σε είδη μουσικής ή σε ομιλητές κατά περίπτωση. Στην παρούσα εργασία περιοριζόμεθα στο πρώτο επίπεδο της ιεραρχίας, όπου και η τμηματοποίηση και η κατηγοριοποίηση νοούνται σε αναφορά με τις γενικές κλάσεις μουσικής και ομιλίας. Αν και για την τμηματοποίηση θα μπορούσε να προκύψει λεπτότερος τεμαχισμός οφειλόμενος στη δυνατότητα περαιτέρω διαχωρισμού. Αποτυχία θα θεωρείται σ' αυτό το πλαίσιο η αδυναμία διάκρισης μίας μετάβασης τύπου ομιλίας/μουσικής και αντίστροφα, ή η αδυναμία ορθής κατάταξης σε μία από τις δύο αυτές κλάσεις.

Συστήματα για τμηματοποίηση και/ή κατηγοριοποίηση ήχου έχουν αναπτυχθεί και δοκιμασθεί σε διάφορες εφαρμογές [4], [5], [7], [6], [9], [11]. Τα συστήματα αυτά θα παρουσιασθούν αναλυτικότερα παρακάτω και θα συγκριθούν με το προτεινόμενο σ' αυτή την εργασία. Κατά κανόνα η τμηματοποίηση και η ταξινόμηση γίνονται σε δύο διαδοχικά στάδια εξασφαλίζοντας και καλή απόδοση και καλές επιδόσεις.

Ο J. Saunders [5] πρότεινε μία τεχνική για το διαχωρισμό σε ομιλία και μουσική, χρησιμοποιώντας το περίβλημα της ενέργειας και την κατανομή των διελεύσεων από το μηδέν. Η τεχνική εφαρμόσθηκε σε ραδιοφωνικά σήματα με δειγματοληψία 16 kHz, και έγιναν κατατάξεις για τμήματα διάρκειας 2.4 δευτερολέπτων όπου τα βασικά χαρακτηριστικά υπολογίσθηκαν σε διαστήματα 16 msec. Μετρήθηκε η πιθανότητα εμφάνισης ελαχίστων τιμών της ενέργειας κάτω από ένα όριο. Αυτή η πιθανότητα είναι υψηλή στην ομιλία, λόγω της διάκρισης σε συλλαβές και σε λέξεις. Χρησιμοποιήθηκαν ακόμα τέσσερα μέτρα σχετικά με την ασυμμετρία της κατανομής των διελεύσεων από το μηδέν. Αναφέρεται απόδοση ορθής κατάταξης της τάξης του 90% με χρήση μόνο των διελεύσεων από το μηδέν, και 98% όταν χρησιμοποιήθηκε και το μέτρο που εξάχθηκε από την ενέργεια.

Οι E. Scheirer και M. Slaney [7] χρησιμοποιούν συνολικά δεκατρία χαρακτηριστικά από τα οποία τα οκτώ εξάγονται από το φάσμα ισχύος. Τέσσερις μέθοδοι ταξινόμησης χρησιμοποιήθηκαν και αναφέρεται ποσοστό επιτυχίας 94.2% σε διαστήματα διάρκειας 20 msec, και 98.6% σε διαστήματα 2.4 sec. Τα σήματα ήχου προέρχονταν από ένα ραδιοφωνικό δέκτη με συχνότητα δειγματοληψίας 22050 Hz. Αναφέρεται μεγάλη ποικιλία ως προς το περιεχόμενο, την πηγή των σημάτων και το επίπεδο του θορύβου.

Οι G. Tzanetakis και P. Cook [9] επρότειναν ένα πλαίσιο τμηματοποίησης και ταξινόμησης για την ανάλυση του ήχου. Η τμηματοποίηση χρησιμοποιώντας τα εξαγόμενα χαρακτηριστικά βασίζεται στην ανίχνευση αλλαγών μέσω παραγώγισης. Στο πλαίσιο αυτό αξιοποιούνται διάφορα χαρακτηριστικά ενέργειας, φάσματος ισχύος, διελεύσεων από το μηδέν, συντελεστές πρόβλεψης, κλπ. Πολλοί γνωστοί ταξινομητές έχουν ενσωματωθεί και δοκιμασθεί στο προτεινόμενο πλαίσιο. Το

αναφερόμενο ποσοστό επιτυχίας είναι της τάξης του 90% σε διαστήματα των 20 msec.

Στο [11] παρουσιάζεται ένα σύστημα ταξινόμησης, αναζήτησης και ανάκλησης σημάτων ήχου με βάση το περιεχόμενό τους. Η ανάλυση χρησιμοποιεί την ισχύ του σήματος, τον τόνο, την κεντρική συχνότητα, το εύρος των συχνοτήτων και την αρμονικότητα. Το σύστημα αξιοποιείται κύρια σε βάσεις με δεδομένα ακουστικών σημάτων. Μια επισκόπηση θεμάτων σχετικών με την ανάκληση ακουστικής πληροφορίας δίδεται στο [1].

Στις εργασίες [4] και [6] χρησιμοποιείται η ανάλυση του φάσματος ισχύος (*cepstrum coefficients*). Οι Moreno και Rifkin [4] μοντελοποιούν τα δεδομένα μέσω μίξεων κατανομών Gauss και τα ταξινομούν με ένα νευρωνικό δίκτυο διανυσμάτων στήριξης (*Support Vector Machine*). Σε σύνολο 173 ωρών τυχαία συλλεγμένων από το Διαδίκτυο ηχητικών σημάτων αναφέρεται ποσοστό 81.8% επιτυχούς κατηγοριοποίησης. Οι Seck, Bimbot, Zugah και Delyon χρησιμοποιούν επίσης μίξεις κατανομών Gauss, και στη συνέχεια το λόγο πιθανοφάνειας για την τμηματοποίηση σε δύο κλάσεις. Αναφέρεται ποσοστό 80% ορθής κατάταξης για τμήματα διάρκειας 26 msec.

Μια παρατήρηση που μπορεί να γίνει στις μέχρι τώρα εργασίες είναι ότι χρησιμοποιούν μεγάλο αριθμό χαρακτηριστικών, που συχνά υπαγορεύεται από το πλήθος των κλάσεων. Μία δεύτερη παρατήρηση αφορά στη σύνδεση της χρησιμοποιούμενης τεχνικής κατηγοριοποίησης με τα εξαγόμενα χαρακτηριστικά. Πρώτη φροντίδα στην παρούσα εργασία υπήρξε η στατιστική ανάλυση των χαρακτηριστικών, ώστε αφενός η μέθοδος ταξινόμησης να τα χρησιμοποιεί με βέλτιστο τρόπο, και αφετέρου να χρησιμοποιούνται μόνο εφόσον κομίζουν πρόσθετη πληροφορία. Ετσι στην παρούσα εργασία για την τμηματοποίηση χρησιμοποιείται μόνο ένα χαρακτηριστικό, η ενέργεια ή το πλάτος του σήματος, και για την ακρίβεια η κατανομή του μέσου πλάτους του σήματος σε ένα ορισμένης διάρκειας διάστημα. Αρκούν μόνο δύο χαρακτηριστικά για να διαχωρισθεί η μουσική από την ομιλία: το μέσο πλάτος και το πλήθος των διελεύσεων από το μηδέν. Το δεύτερο χαρακτηριστικό αντιστοιχεί στη μέση συχνότητα χωρίς να απαιτείται ο υπολογισμός του φάσματος ισχύος.

Στον αλγόριθμο που προτείνουμε τα δύο βασικά χαρακτηριστικά υπολογίζονται ανά 20 msec και η τμηματοποίηση γίνεται με καθυστέρηση 4 sec σήματος. Κάθε νέο τμήμα που δημιουργείται ταξινομείται στη μία από τις δύο δυνατές κλάσεις ανάλογα με τις τιμές που έχουν τα χαρακτηριστικά ισχύος και συχνότητας. Ετσι με 4 sec καθυστέρηση ολοκληρώνεται η τμηματική κατηγοριοποίηση. Λόγω της απλότητας των υπολογισμών, πέραν της μεθοδολογικής αυτής καθυστέρησης, η απόκριση του συστήματος λαμβάνεται σε πραγματικό χρόνο. Αφού περιγράψουμε το ηχητικό σήμα εισόδου και τις λοιπές προδιαγραφές του συστήματος, στα επόμενα κεφάλαια θα δώσουμε τους αλγόριθμους τμηματοποίησης και ταξινόμησης.

1.2 Μορφή σήματος και προδιαγραφές συστήματος

Το σήμα εισόδου του συστήματος είναι ένα μονοφωνικό ή στερεοφωνικό ηχητικό σήμα. Βέβαια η πολυκαναλική πληροφορία σχετίζεται με τη γένεση του ηχητικού συμβάντος και δεν έχει σχέση με τον χαρακτήρα του σήματος. Από τη φύση της εξαρτάται από τη θέση του δέκτη. Αφού λοιπόν δεν μας ενδιαφέρει η τοπολογία του ηχητικού φαινομένου, αλλά μόνο το ίδιο το φαινόμενο, μετατρέπουμε στις περιπτώσεις που έχουμε πάνω από ένα κανάλι ήχου σε μονοφωνικό σήμα παίρνοντας απλά τη μέση τιμή των καναλιών. Εναλλακτικά θα μπορούσαμε να χαρακτηρίσουμε ξεχωριστά τα σήματα των δύο καναλιών, αλλά στη μεγάλη πλειονότητα των περιπτώσεων οι χαρακτηρισμοί θα συμπίπτουν.

Οσον αφορά τη συχνότητα δειγματοληψίας του σήματος έγιναν δοκιμές με συχνότητα από 11025 Hz έως και 44100 Hz, ενώ η ένταση μπορούσε να ποικίλει από ηχογράφηση σε ηχογράφηση. Βέβαια τα παραπάνω είναι ανεξάρτητα από το περιεχόμενο του σήματος και επομένως δεν θα πρέπει να επηρεάζουν τον αλγόριθμο τμηματοποίησης, εξαγωγής χαρακτηριστικών και κατηγοριοποίησης.

Η τμηματοποίηση θα βασισθεί σε ανίχνευση μεταβολών των χαρακτηριστικών. Προς τούτο απαιτείται όπως ορισθεί η μονάδα του χρόνου όπου θα εκτιμώνται τα χαρακτηριστικά και στη συνέχεια θα ελέγχεται η διατήρηση, ή ομοιότητα, ως προς την αλλαγή, ή ανομοιότητα. Γίνεται επιλογή το χρονικό διάστημα για την εύρεση των κατανομών των χαρακτηριστικών να είναι 1 sec, δηλαδή ουσιατικά βρίσκουμε την ομοιότητα ανάμεσα σε διαστήματα διάρκειας 1 sec. Κατά συνέπεια

μια πρώτη προϋπόθεση της προτεινόμενης τεχνικής είναι οι αλλαγές να απέχουν τουλάχιστον 1 sec. Τα διαστήματα βέβαια θα μπορούσαν να είχαν μεγαλύτερη διάρκεια, οπότε οι αλλαγές που θα εντοπίζαμε θα είχαν μεγαλύτερη αξιοπιστία, αλλά δεν θα μπορούσαμε να εντοπίσουμε αλλαγές που βρίσκονται σε χρονική απόσταση μικρότερη από το μέγεθος του διαστήματος. Ομως, ακόμα και στην περίπτωση που η μέθοδος δώσει λάθος αλλαγή τα δύο τμήματα που θα προκύψουν, θα ενωθούν σε ένα αφού ακολουθεί η κατάταξη και η αλλαγή διαγράφεται, εάν η ταξινόμηση δώσει την ίδια κατηγορία για δύο διαδοχικά διαστήματα. Αρα δεν ενοχλούν ιδιαίτερα οι λαθεμένες ανιχνεύσεις αλλαγών, αφού μπορούν να διορθωθούν μέσω της κατηγοριοποίησης. Η επιλογή για την τιμή 1 sec έγινε διότι, όπως έδειξαν και τα πειράματα, δεν προκύπτουν πολλές λαθεμένες μετρήσεις, ενώ είναι σχεδόν απίθανο να έχουμε δύο αλλαγές σε τόσο μικρό διάστημα.

Για τις αλλαγές, όπως αναφέραμε, χρησιμοποιούνται οι κατανομές των μετρούμενων χαρακτηριστικών. Απαιτείται επομένως άλλη μία χρονική σταθερά, ένα διάστημα, όπου θα εκτιμώνται τα ''στιγμιαία" χαρακτηριστικά του ηχητικού σήματος. Αυτή η διάρκεια ορίζεται σε 20 msec, ώστε να διατίθενται 50 μετρήσεις στο διάστημα του 1 sec. Ο ορισμός της ''στιγμής" δίνει ταυτόχρονα την ακρίβεια της προτεινόμενης τεχνικής. Η ακρίβεια αυτή δείχνει αρκετά υψηλή, αφού ακόμα και το ανθρώπινο αυτί δεν μπορεί να πετύχει μεγαλύτερη ακρίβεια, επομένως μεγαλύτερη ακρίβεια θα μπορούσε δύσκολα να ελεγχθεί. Από την άλλη πλευρά το διάστημα αυτό δεν θα μπορούσε να είναι μεγαλύτερο, διότι τα σήματα φωνής δεν διακρίνονται από στασιμότητα, παρά μόνο για βραχέα διαστήματα μέχρι 20 msec [8]. Επιπλέον στα φωνήεντα υπάρχει περιοδικότητα και αρμονία, ενώ στα σύμφωνα το εύρος συχνοτήτων του σήματος είναι μεγάλο και το φάσμα ισχύος είναι τύπου θορύβου.

1.3 Περιγραφή βασικών χαρακτηριστικών

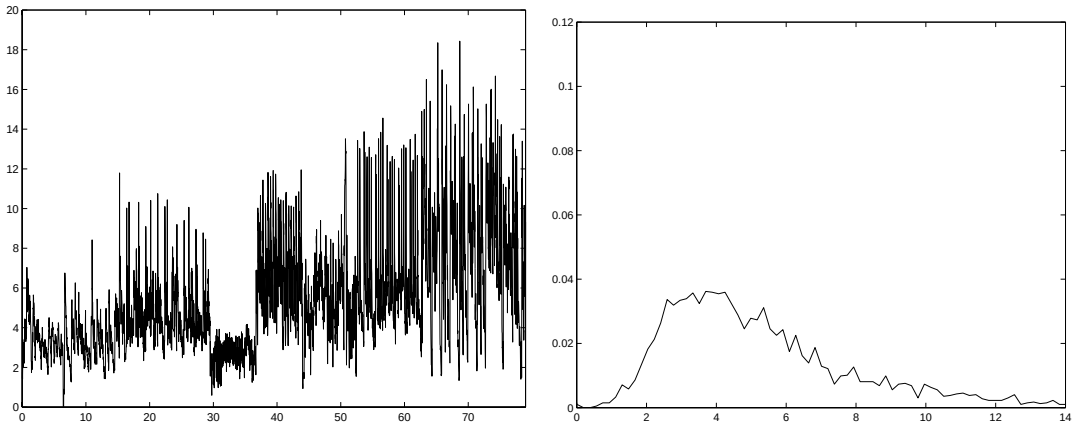
Τα χρησιμοποιούμενα χαρακτηριστικά είναι δύο: το πλάτος του σήματος (*Root Mean Square*) και η συχνότητά του, εκτιμώμενη μέσω της πυκνότητας των διελεύσεων από το μηδέν (*Zero Crossings*). Τα χαρακτηριστικά υπολογίζονται ανά 20 msec. Από αυτά εκτιμώνται τα χαρακτηριστικά συνόλου, που αφορούν ένα ολόκληρο τμήμα και χρησιμοποιούνται στην κατάταξη. Μόνο το πλάτος του σήματος χρησιμοποιείται για την αρχική τμηματοποίηση.

1.3.1 Πλάτος σήματος, *RMS*

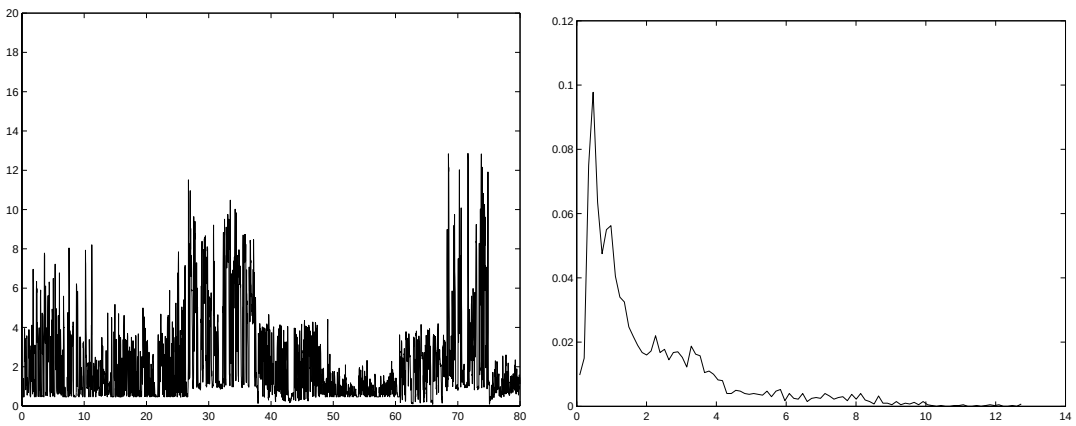
Το πλάτος του σήματος, *RMS*, ορίζεται ως η τετραγωνική ρίζα της ενέργειας του σήματος στο συγκεκριμένο χρονικό διάστημα. Είναι ουσιαστικά ένα μέτρο για την ένταση του σήματος. Ο υπολογισμός του είναι άμεσος από τις τιμές του σήματος. Θα πρέπει σε αυτό το σημείο να πούμε πως η τιμή της έντασης, επειδή εξαρτάται από την απόσταση της πηγής και του δέκτη δε δίνει ουσιαστική πληροφορία ως απόλυτο μέγεθος, αλλά κύρια η κατανομή της και οι μεταβολές της στο χρόνο παρουσιάζουν ενδιαφέρουν. Βέβαια για την τμηματοποίηση αρκεί η αλλαγή της έντασης για να μιλήσουμε για πιθανό νέο τμήμα. Η τιμή του πλάτους του σήματος, A, για N δείγματα του σήματος $x(n)$ δίδεται από την ακόλουθη εξίσωση

$$\mathrm{A} = \sqrt{\sum_{n=1}^{N} x^2(n)} \tag{1.1}$$

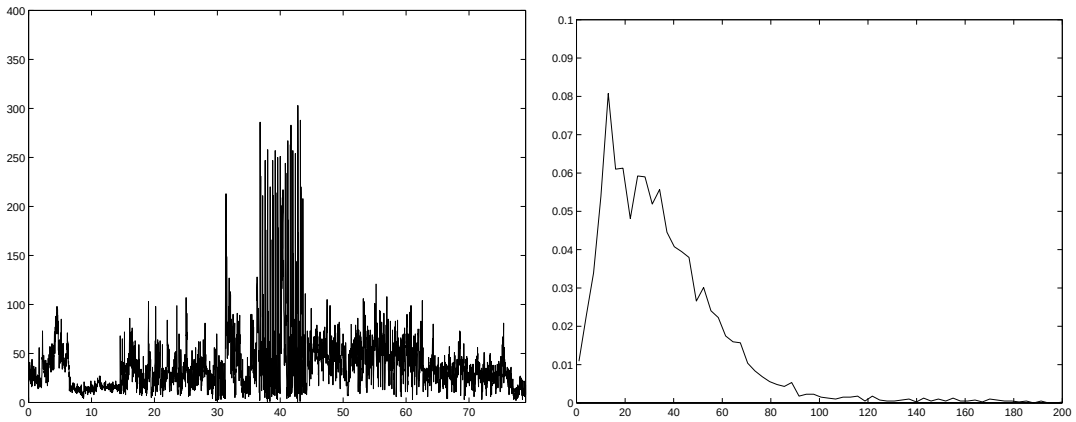
Από τα σχήματα 1.1 και1.2, που αντιστοιχούν στο πλατος σήματος μουσικής και ομιλίας αντίστοιχα, μπορούμε να δούμε πως τα ιστογράμματά τους διαφέρουν σημαντικά. Οπότε σε μία μετάβαση από μουσική σε ομιλία θα αλλάξει η κατανομή του πλάτους του σήματος, γεγονός που χρησιμοποιούμε και στην τμηματοποίηση και στην κατηγοριοποίηση.



Σχήμα 1.1: Αριστερά εικονίζεται το πλάτος ενός σήματος μουσικής ως συνάρτηση του χρόνου, ενώ δεξιά δίδεται το ιστόγραμμά του.



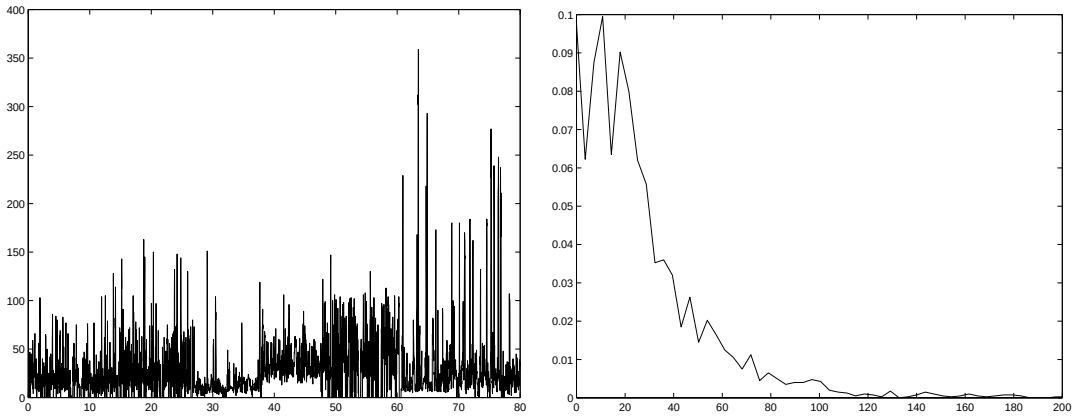
Σχήμα 1.2: Αριστερά εικονίζεται το πλάτος ενός σήματος ομιλίας ως συνάρτηση του χρόνου, ενώ δεξιά δίδεται το ιστόγραμμά του.



Σχήμα 1.3: Αριστερά εικονίζεται το ZC ως συνάρτηση του χρόνου σε αρχείο με μουσική, ενώ δεξιά δίδεται το ιστόγραμμά του.

1.3.2 Διελεύσεις από το μηδέν, *ZC*

Το πλήθος των διελεύσεων από το μηδέν αποτελεί μέτρο της μέσης συχνότητας του σήματος. Είναι ο αριθμός των εναλλαγών προσήμου στις τιμές του σήματος. Πολύ μικρή τιμή, περίπου μηδενική, της πυκνότητας των διελεύσεων από το μηδέν σε ένα διάστημα, σημαίνει ουσιαστικά απουσία σήματος, δηλαδή σιωπή ή μικρή παύση. Στα σχήματα 1.3 και 1.4 δίδονται δύο παραδείγματα μετρήσεων των διελεύσεων από το μηδέν για σήματα μουσικής και φωνής αντίστοιχα.



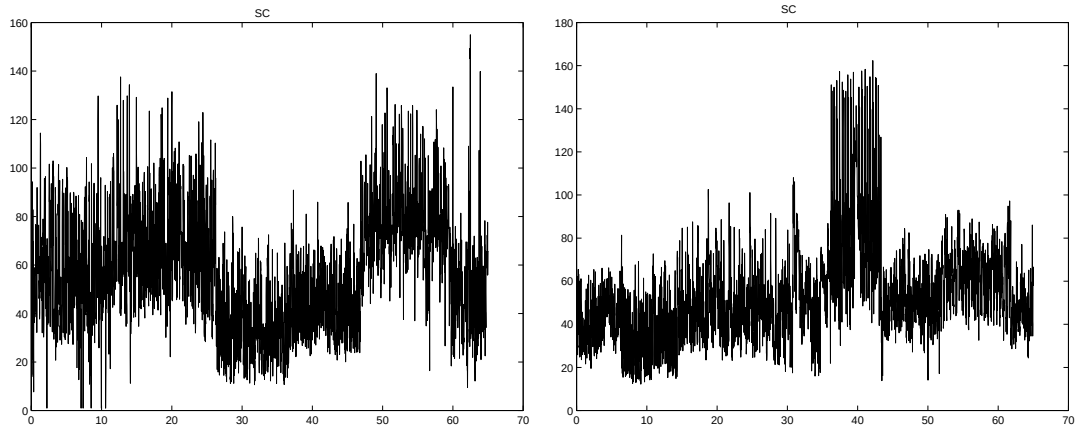
Σχήμα 1.4: Αριστερά εικονίζεται το ZC ως συνάρτηση του χρόνου σε αρχείο με ομιλία, ενώ δεξιά δίδεται το ιστόγραμμά του.

1.3.3 Αλλα χαρακτηριστικά

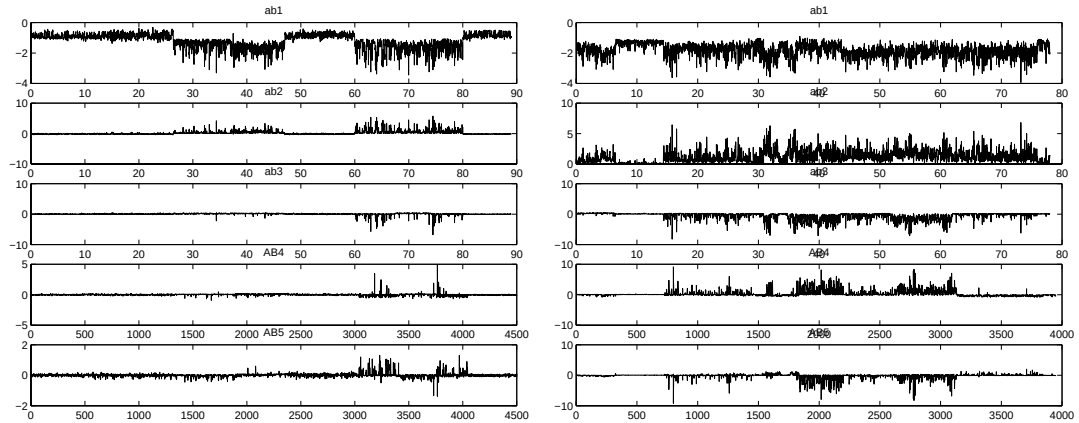
Οι G. Tzanetakis και P. Cook [9] έχουν χρησιμοποιήσει επιπλέον χαρακτηριστικά που τα περισσότερα προκύπτουν από την ανάλυση του φάσματος ισχύος του σήματος, που υπολογίζεται μέσω του διακριτού μετασχηματισμού Fourier. Τέτοιο χαρακτηριστικό είναι η μέση τιμή του φάσματος ισχύος

$$f_1 = \frac{\int_0^{1/2} f|X(f)|^2 df}{\int_0^{1/2} |X(f)|^2 df}$$

Στο Σχήμα 1.5 δίδεται η μέση συχνότητα σε διαδοχικά διαστήματα 20 msec για ένα σήμα ομιλίας και ένα σήμα μουσικής. Η μέση συχνότητα είναι εξαιρετικά συσχετισμένη με το μέσο πλήθος διελεύσεων από το μηδέν. Αλλα χαρακτηριστικά είναι η συχνότητα που συγκεντρώνει το 95% της ενέργειας του σήματος, οι συντελεστές φίλτρων γραμμικής πρόβλεψης, *LPC* (Σχήμα 1.6) και οι συντελεστές ανάλυσης του λογαρίθμου του φάσματος ισχύος, *MFCC*, που χρησιμοποιούνται για ανάλυση φωνής, αλλά και η ενέργεια στα 4 Hz που είναι η συχνότητα συλλαβών για τον άνθρωπο. Οι D. Krubsack και R. Niederjonh [3] χρησιμοποιούν τον τόνο, *pitch*, για την αναγνώριση ομιλίας. Για τον καθορισμό του τόνου έχουν αναπτυχθεί αρκετές μέθοδοι. Εκτος από τον ίδιο τον τόνο μπορούν να χρησιμοποιηθούν και συντελεστές σφάλματος που έχουν σχέση με αυτόν. Βέβαια για την τμηματοποίηση, όπως θα φανεί στη συνέχεια, αρκεί το πλάτος του σήματος, ενώ για την κατάταξη θα δούμε πως από το συνδυασμό του με τη μέση συχνότητα μπορούν να προκύψουν χαρακτηριστικά ικανά για επιτυχή ταξινόμηση.



Σχήμα 1.5: Μέση τιμή του φάσματος ισχύος για ένα σήμα ομιλίας (αριστερά) και για ένα σήμα μουσικής (δεξιά)



Σχήμα 1.6: Οι πέντε πρώτοι συντελεστές ανάλυσης LPC για σήμα ομιλίας (δεξιά) και μουσικής (αριστερά).

1.3.4 Ανεξαρτησία RMS - ZC

Σε αυτό το σημείο είναι σημαντικό να εξετάσουμε την ανεξαρτησία πλάτους και συχνότητας του σήματος. Καταρχήν και διαισθητικά μπορούμε να πούμε ότι τα δύο αυτά χαρακτηριστικά

είναι ανεξάρτητα, ή τουλάχιστον ασυσχέτιστα. Βέβαια για να δούμε αν τα χαρακτηριστικά είναι ανεξάρτητα στα δεδομένα μουσικής και ομιλίας θα πρέπει να εξετάσουμε κάποιους συντελεστές συσχέτισης σε μεγάλο αριθμό από δεδομένα. Για τον υπολογισμό της εξάρτησης χρησιμοποιήσαμε το μέτρο του Blomquist [2], που ορίζεται ως ακολούθως

$$V = \frac{|n_1 - n_2|}{n} \tag{1.2}$$

όπου n είναι το πλήθος όλων των διατιθεμένων ζευγών δεδομένων, n_1 είναι το πλήθος των ζευγών που έχουν το ίδιο πρόσημο συγκρινόμενα με τις αντίστοιχες μεσαίες τιμές, ενώ n_2 είναι το πλήθος των ζευγών που έχουν αντίθετο πρόσημο. Σε μεγάλο τμήμα από δεδομένα πήραμε τιμή γύρω στο 0.1, που δείχνει σχετικά καλή ανεξαρτησία, αφού η τιμή της πλήρους εξάρτησης είναι 1. Η χρήση του λόγου της αμοιβαίας πληροφορίας ως προς την εντροπία

$$I = 2\frac{\sum\sum P_{ij} \log \frac{P_{ij}}{P_i Q_j}}{\sum P_i \log \frac{1}{P_i} + \sum Q_j \log \frac{1}{Q_j}} \tag{1.3}$$

έδωσε τιμή 0.05, δηλαδή πλησιάζει την πλήρη ανεξαρτησία. Στην παραπάνω εξίσωση P_i είναι οι πιθανότητες για ένα σύνολο τιμών της μίας μεταβλητής, Q_j αντίστοιχα για την άλλη μεταβλητή και P_{ij} η από κοινού πιθανότητα για τα διάφορα ζεύγη τιμών των δύο μεταβλητών. Επομένως μπορούμε να μιλήσουμε για ανεξαρτησία μεταξύ πλάτους και συχνότητας που μάλιστα στη μουσική εμφανίζεται κάπως πιο έντονη, αφού οι αντίστοιχες τιμές που υπολογίστηκαν στα μουσικά κομμάτια βρέθηκαν κατα 10% μικρότερες από αυτές σε ομιλία. Τα παραπάνω δείχνουν πως η χρήση και των δύο χαρακτηριστικών, αλλά και ο συνδυασμός τους μπορεί να προσφέρει νέα πληροφορία γεγονός που έχουμε εκμεταλλευτεί στην κατάταξη.

Κεφάλαιο 2

Τμηματοποίηση

Η τμηματοποίηση γίνεται πρακτικά σε πραγματικό χρόνο και στηρίζεται στο χαρακτηριστικό RMS. Καθώς διαβάζεται το αρχείο χωρίζεται σε διαστήματα 1 sec για κάθε ένα από τα οποία υπάρχουν 50 τιμές RMS από τις οποίες υπολογίζονται η μέση τιμή τους και η διασπορά τους. Ο αλγόριθμος μπορεί να χωριστεί σε δύο στάδια. Στο πρώτο στάδιο έχουμε εντοπισμό του διαστήματος που υπάρχει αλλαγή, ενώ στο δεύτερο γίνεται εντοπισμός της αλλαγής σε μια γειτονιά του διαστήματος με ακρίβεια 20 msec. Με αυτόν τον τρόπο επιταχύνεται η όλη διεργασία, αφού το πλέον χρονοβόρο βήμα του αλγορίθμου, όπου η αναζήτηση γίνεται ανά 20 msec, εκτελείται μόνο εφόσον στο πρώτο βήμα εντοπίζεται πιθανή αλλαγή.

2.1 Ανίχνευση αλλαγών σε διαδοχικά διαστήματα

Σκοπός του μέρους αυτού είναι ο εντοπισμός διαστημάτων που πιθανότατα θα υπάρχει κάποια αλλαγή. Οπότε ένα τέτοιο διάστημα θα έχει την ιδιότητα ότι το προηγούμενο και το επόμενο από αυτό θα διαφέρουν μεταξύ τους, άρα θα υπάρχει μεγάλη αλλαγή στην κατανομή των τιμών του χαρακτηριστικού RMS. Στην ομιλία επειδή υπάρχουν μικρά κενά ανάμεσα στις συλλαβές και στις λέξεις το πλάτος του σήματος θα εμφανίζει αρκετές διακυμάνσεις, ενώ στην μουσική εμφανίζει μεγαλύτερη σταθερότητα. Δηλαδή ο λόγος της τυπικής απόκλισης ως προς τη μέση τιμή αναμένεται να είναι διαφορετικός, ανάλογα με την κλάση. Βέβαια είναι αρκετά πιθανό να λάβουμε με αυτόν τον τρόπο και αλλαγές που τελικά να μην είναι αποδεκτές. Ομως στην συνέχεια ακολουθεί η κατάταξη που θα συνενώσει τμήματα που ανήκουν στην ίδια κλάση. Επομένως θα μπορούσε να γίνει μια σύγκριση ιστογραμμάτων του πλάτους A, ώστε να βρεθούν τα υποψήφια για μετάβαση διαστήματα.

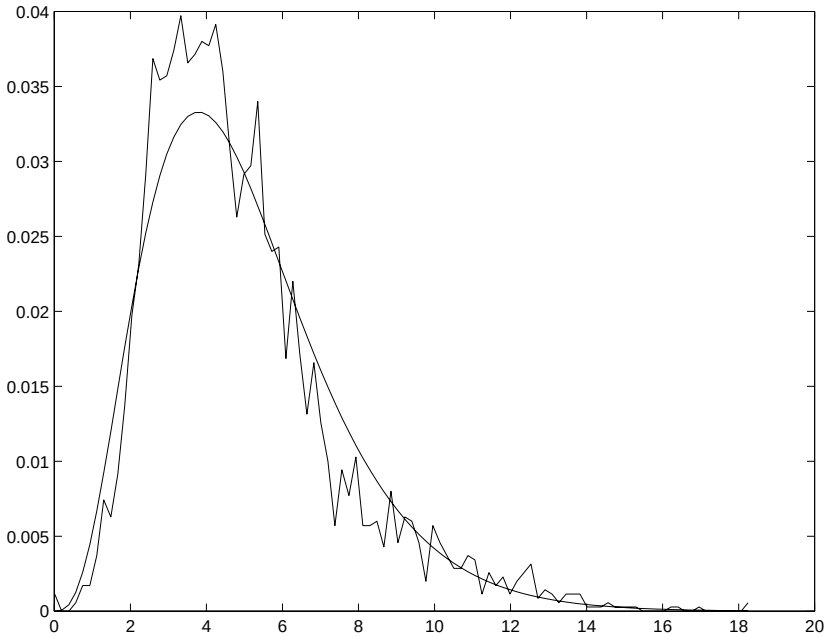
Ο υπολογισμός και η απόσταση των ιστογραμμάτων απαιτούν αρκετό χρόνο, αλλά η στατιστική ανάλυση του πλάτους δείχνει ότι προσεγγίζεται πολύ καλά από την κατανομή χ^2 τόσο στην μουσική όσο και στην ομιλία, γεγονός που φαίνεται πολύ καλά στα Σχήματα 2.1 και 2.2. Η κατανομή χ^2 δίνεται από τη σχέση

$$p(x) = \frac{x^a e^{-bx}}{b^{a+1}\Gamma(a+1)}, \quad x \geq 0 \tag{2.1}$$

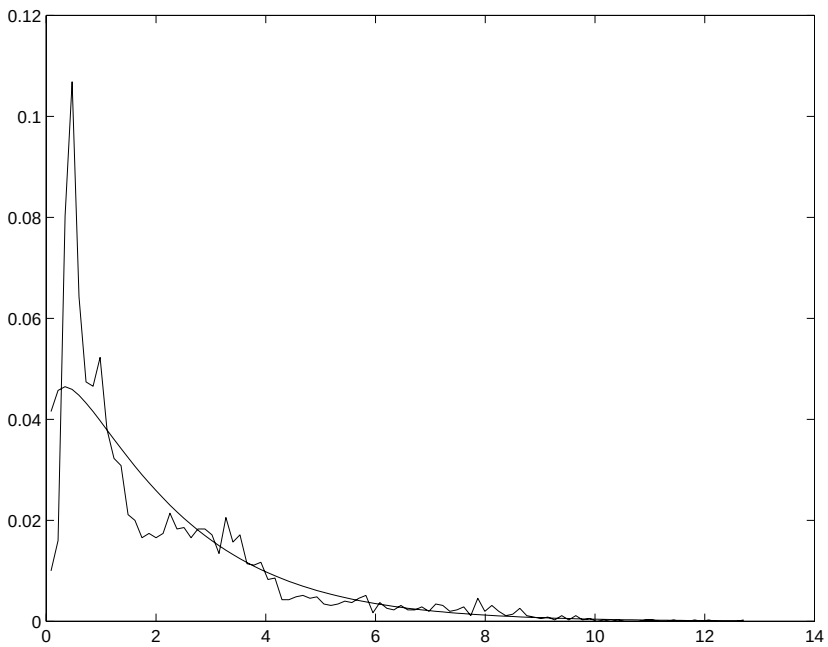
Τα a, b εξαρτώνται από τη μέση τιμή μ και διασπορά σ^2 του πλάτους του σήματος, όπως φαίνεται στην ακόλουθη σχέση

$$a = \frac{\mu^2}{\sigma^2} - 1 \quad \text{και} \quad b = \frac{\sigma^2}{\mu} \tag{2.2}$$

Επομένως αρκεί να γνωρίζουμε τη μέση τιμή και τη διασπορά των τιμών του RMS σε δύο διαστήματα για να υπολογίσουμε την ομοιότητά τους. Για τον υπολογισμό της ομοιότητας έγινε



Σχήμα 2.1: Το ιστόγραμμα του RMS σε συλλογή από μουσικά κομμάτια και η προσέγγισή του από την κατανομή χ^2.



Σχήμα 2.2: Το ιστόγραμμα του RMS σε συλλογή από σήματα ομιλίας και η προσέγγισή του από την κατανομή χ^2.

χρήση του ακόλουθου μέτρου [12]

$$\rho(p_1, p_2) = \int \sqrt{p_1(x)p_2(x)}dx \tag{2.3}$$

Η ομοιότητα λαμβάνει τιμές από 0 έως και 1, με το 1 να εμφανίζεται όταν οι δύο κατανομές είναι ίδιες και το 0 όταν δεν έχουν κανένα σημείο τομής. Το μέτρο ομοιότητας σχετίζεται με την πιθανότητα λαθεμένης κατάταξης P_e. Ακριβέστερα η πιθανότητα λαθεμένης κατάταξης στην περίπτωση δύο ισοπίθανων υποθέσεων φράσσεται ως εξής

$$P_e \leq \frac{\rho(p_1, p_2)}{2} \tag{2.4}$$

Για την κατανομή χ^2 μπορεί να αποδειχθεί πως το μέτρο ομοιότητας μεταξύ δύο κατανομών $\rho(p_1, p_2)$, με συντελεστές a_1, b_1 και a_2, b_2 αντίστοιχα έχει την μορφή

$$\rho(p_1, p_2) = \frac{\Gamma(\frac{a_1+a_2}{2} + 1)}{\sqrt{\Gamma(a_1 + 1)\Gamma(a_2 + 1)}} \frac{2^{\frac{a_1+a_2}{2}+1} b_1^{\frac{a_2+1}{2}} b_2^{\frac{a_1+1}{2}}}{(b_1 + b_2)^{\frac{a_1+a_2}{2}+1}} \tag{2.5}$$

Το μέτρο της ομοιότητας μπορεί να χρησιμοποιηθεί για να βρεθεί αν υπάρχει αλλαγή σε κάποιο διάστημα. Ετσι για ένα διάστημα i υπολογίζουμε την τιμή $D(i)$ που δίνει το βαθμό μεταβολής γύρω απ' αυτό το διάστημα. Βασικά αν υπάρχει αλλαγή στο διάστημα i, τότε οι τιμές του χαρακτηριστικού στα διαστήματα $i - 1$ και $i + 1$ θα πρέπει να διαφέρουν, και ο παράγοντας $\rho(p_{i-1}, p_{i+1})$ θα είναι κοντά στο 0, ενώ στην αντίθετη περίπτωση που δεν θα υπάρχει αλλαγή ο παράγοντας $\rho(p_{i-1}, p_{i+1})$ θα είναι περίπου 1. Η απόσταση ορίζεται ως το συμπλήρωμα της ομοιότητας, και είναι γνωστή ως απόσταση Matusita [12]

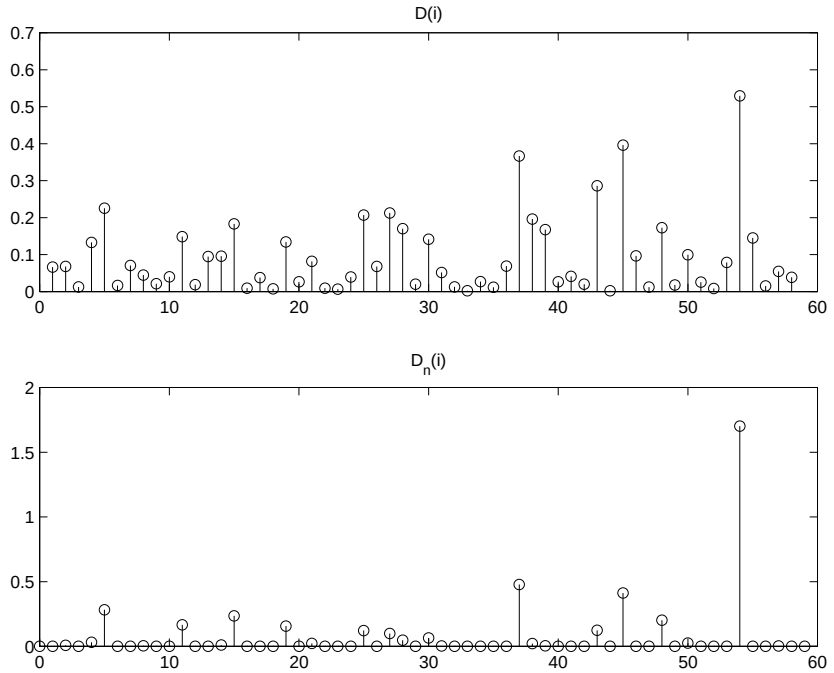
$$D(i) = 1 - \rho(p_{i-1}, p_{i+1}) \tag{2.6}$$

Οπότε αλλαγές από ομιλία σε μουσική ή απότομες μεταβολές στην ένταση του ήχου θα μεγιστοποιούν τοπικά το $D(i)$ και θα μπορούσαν να εντοπιστούν με την χρήση κάποιου φράγματος. Ομως, φαίνεται να χρειάζεται μία κανονικοποίηση, διότι μεγάλες αποστάσεις αναμένονται και στα γειτονικά του διαστήματα. Ακόμα η τιμή του φράγματος δεν είναι σταθερή για όλα τα σήματα ήχου, αλλά, για παράδειγμα, είναι δυνατόν για σήματα μουσικής να υπερβαίνει πάντοτε το φράγμα και να χαρακτηρίζονται όλα τα διαστήματα ως υποψήφια αλλαγής. Λογικό επομένως θα ήταν να απαιτούμε όχι μόνο η απόσταση $D(i)$ να έχει υψηλή τιμή, αλλά επίσης να είναι μεγαλύτερη από τις αντίστοιχες τιμές των γειτονικών διαστημάτων. Η παρακάτω σχέση δίνει την κανονικοποιημένη διαφορά $D_n(i)$ που έχει τις παραπάνω ιδιότητες

$$D_n(i) = \frac{D(i)V(i)}{D_M(i)} \tag{2.7}$$

Το $V(i)$ είναι η θετική διαφορά του $D(i)$ από τη μέση τιμή των γειτονικών διαστημάτων. Το μέγεθος της γειτονιάς που χρησιμοποιήθηκε στην παρούσα εργασία είναι 2 διαστήματα (2 sec) πριν και μετά από το διάστημα i. Αν η διαφορά είναι αρνητική τότε το $V(i)$ τίθεται 0. Το $D_M(i)$ είναι η μέγιστη τιμή των αποστάσεων $D(j)$ στη γειτονιά του διαστήματος i. Ως διαστήματα με πιθανή αλλαγή επιλέγονται τα τοπικά μέγιστα του $D_n(i)$ εφόσον υπερβαίνουν ένα κατώφλι που καθορίζεται με βάση τη μέση διακύμανση της ομοιότητας. Αν η τυπική απόκλιση της ομοιότητας είναι μικρή (αντίστοιχα, μεγάλη), τότε το κατώφλι είναι μικρό (αντίστοιχα, μεγάλο), δηλαδή το κατώφλι είναι ευθέως ανάλογο της μέσης διακύμανσης της ομοιότητας. Με τη χρήση της κανονικοποιημένης απόστασης αποφεύγεται ταυτόχρονα ο κατακερματισμός του σήματος που θα επιβράδυνε την εκτέλεση του αλγορίθμου. Η σύγκριση της απόστασης $D(\cdot)$ και της κανονικοποιημένης $D_n(\cdot)$ φαίνεται στα Σχήματα 2.3 και 2.5.

Η καθυστέρηση 4 sec που έχουμε από τον πραγματικό χρόνο λήψης του σήματος οφείλεται στις παραπάνω διεργασίες, μιας και για να αποφασίσουμε μια αλλαγή στο διάστημα i πρέπει να γνωρίζουμε το $D_n(i + 1)$ που χρειάζεται να έχουμε $D(i + 3)$ το οποίο προϋποθέτει το p_{i+4}.



Σχήμα 2.3: Επάνω δίδεται η απόσταση $D(i)$ και κάτω η κανονικοποιημένη απόσταση $D_n(i)$ με βάση την οποία γίνεται η ανίχνευση των αλλαγών. Στο σήμα που εξετάζεται υπάρχουν δύο αλλαγές από ομιλία σε μουσική και δύο από μουσική σε ομιλία.

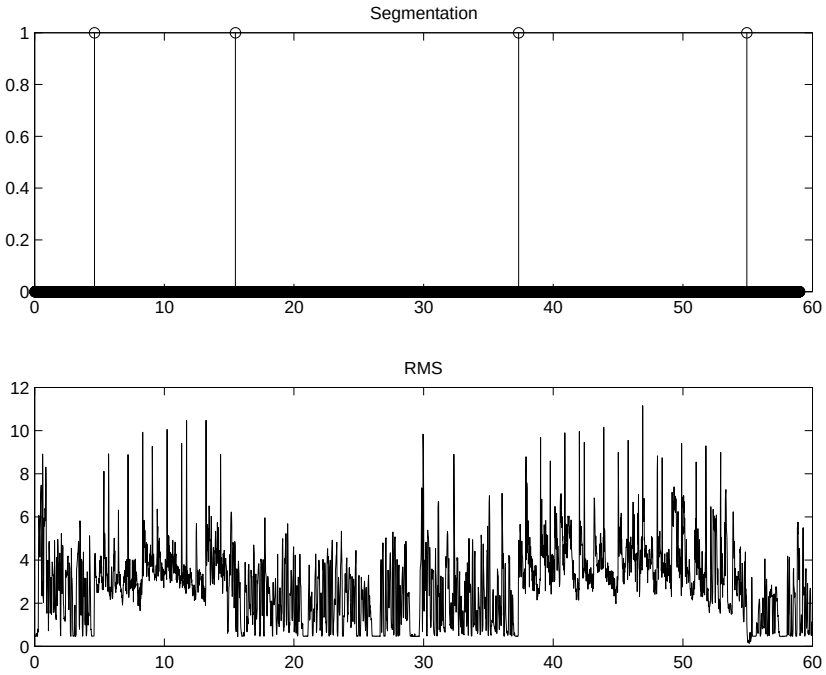
2.2 Στιγμιαίος εντοπισμός αλλαγής

Στο δεύτερο μέρος της τμηματοποίησης πρέπει να εντοπισθεί η αλλαγή, εφόσον υπάρχει, με ακρίβεια 20 msec μέσα στο υποψήφιο διάστημα που ανιχνεύθηκε στο πρώτο μέρος. Θεωρούμε δύο διαδοχικά διαστήματα που μετατοπίζουμε και για τα οποία ψάχνουμε να βρούμε τη θέση μέσα στο υποψήφιο διάστημα αλλαγής όπου διαφέρουν περισσότερο. Ετσι τα μετατοπίζουμε στις 50 δυνατές θέσεις μέσα στο διάστημα (20 msec ακρίβεια) και υπολογίζουμε με βάση τη σχέση (2.6) τη θέση όπου οι κατανομές στα δύο διαστήματα διαφέρουν περισσότερο. Η αναζήτηση μπορεί να επεκταθεί και σε κοντινές θέσεις στα γειτονικά διαστήματα για την περίπτωση που η αλλαγή βρίσκεται αρκετά κοντά στα σύνορα, οπότε η αρχική τμηματοποίηση πιθανόν να μην έδωσε το ακριβές διάστημα αλλαγής. Τελικά παίρνουμε σαν θέση εκείνη με την μεγαλύτερη τιμή στο $D(i)$, ενώ απαιτούμε και η μέση τιμή όλων των υπολογιζόμενων των $D(\cdot)$ να ξεπερνάει κάποιο φράγμα, με αυτόν τον τρόπο αποφεύγουμε κάποιες περιπτώσεις που έχουμε στιγμιαία αλλαγή στην ένταση και όχι πραγματική αλλάγη. Τελικά το τμήμα που θα προκύψει θα είναι ομοιογενές δηλαδή θα ανήκει σε μια κλάση η οποία θα προσδιορισθεί στην κατάταξη.

2.3 Αποτελέσματα τμηματοποίησης

Στα Σχήματα 2.7 και 2.8 παρουσιάζονται αποτελέσματα ανίχνευσης μεταβάσεων από ομιλία σε μουσική και από μουσική σε ομιλία αντίστοιχα. Δείχνεται τόσο ο εντοπισμός του διαστήματος διάρκειας 1 δευτερολέπτου που περιέχει τη μετάβαση, όσο και ο στιγμιαίος εντοπισμός. Στα δύο αυτά παραδείγματα η ακρίβεια εντοπισμού της αλλαγής είναι πολύ καλή.

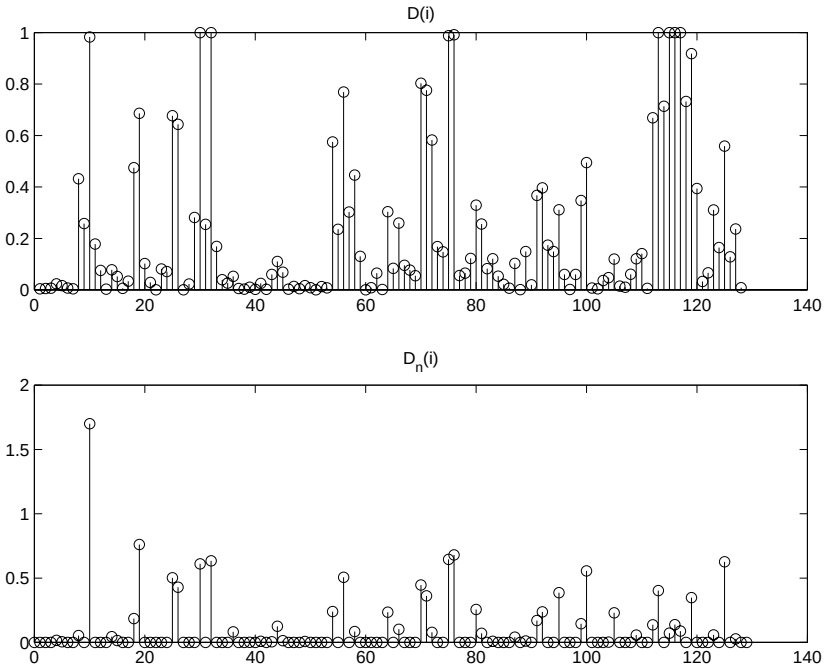
Γενικά η τμηματοποίηση του ήχου σε ομοιογενή τμήματα επιτυγχάνεται με μεγάλη αξιοπιστία. Μπορούμε να ορίσουμε ως μέτρο επιτυχίας το ποσοστό των πραγματικών αλλαγών που ανιχνεύονται.



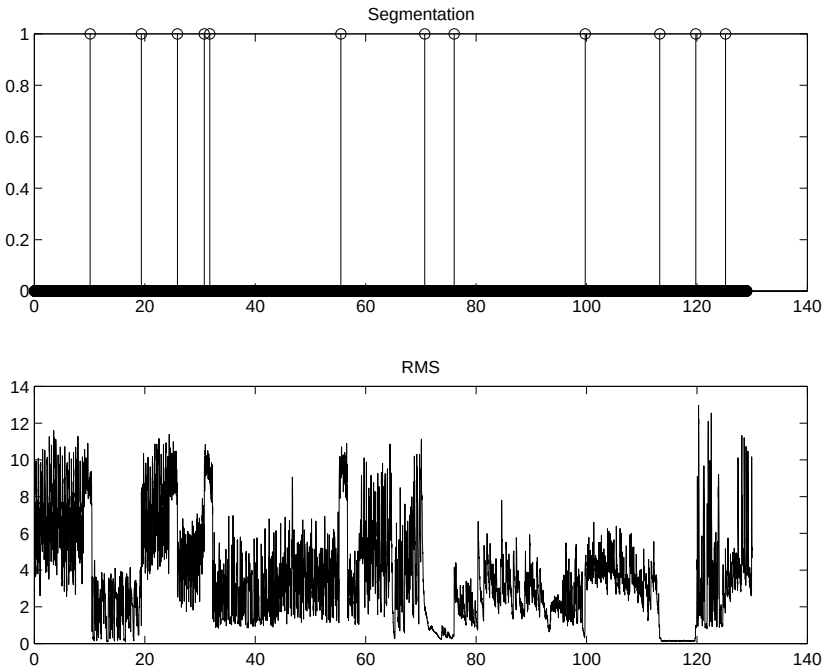
Σχήμα 2.4: Επάνω δίδονται τα διαστήματα όπου εντοπίζονται οι αλλαγές και κάτω το πλάτος του σήματος

Στο σύνολο των δοκιμών που κάναμε είχαμε ποσοστό επιτυχών ανιχνεύσεων περί το 97%. Μπορεί ενδεχόμενα να ανιχνεύονται αλλαγές που δεν είναι πραγματικές, όμως αυτό δεν ενοχλεί, διότι η κατάταξη που ακολουθεί μπορεί να ενώσει τα όμοια τμήματα. Οσον αφορά την ακρίβεια είδαμε πως μπορεί να φτάσει μέχρι τα 20 msec. Βέβαια υπάρχουν περιπτώσεις που το σημείο αλλαγής παρουσιάζει απόκλιση ως προς την ακριβή θέση. Σχεδόν πάντα το σφάλμα είναι μικρότερο από 0.5 sec, ενώ δεν έχει παρατηρηθεί να ξεπερνάει το 1.5 sec. Τα παραπάνω εξαρτώνται από την μετάβαση που έχουμε, οι αλλαγές σε κενά τμήματα εντοπίζονται με τη μέγιστη ακρίβεια, ενώ οι αλλαγές από ομιλία σε μουσική ή αντίστροφα είναι δυνατόν να εμφανίζονται μετατοπισμένες, κάτι που συνήθως εξαρτάται από τη μορφή της μουσικής. Συνήθως το σφάλμα είναι γύρω στα 0.2 sec, αλλά αν έχουμε απαλή μουσική (σταθερή ένταση) ή οι μέσες εντάσεις διαφέρουν, τότε το σφάλμα εντοπισμού μηδενίζεται. Πάντως λόγω της διαφορετικής κατανομής του RMS στην ομιλία και στην μουσική η αλλαγή είναι σχεδόν βέβαιο ότι θα εντοπισθεί, απλώς η ακρίβεια εντοπισμού διαφέρει.

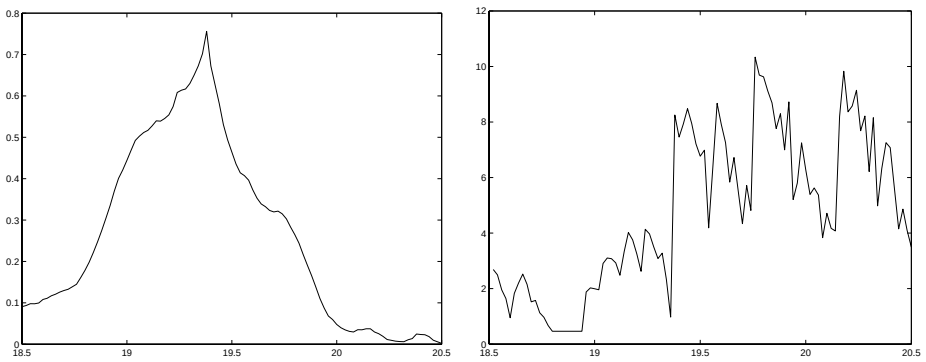
Παραδείγματα ανιχνεύσεων αλλαγών μαζί με το αντίστοιχο πλάτος σήματος που χρησιμοποιείται για την τμηματοποίηση δίδονται στα Σχήματα 2.4, 2.6 και 2.9. Το Σχήμα 2.4 (αντίστοιχα 2.6) αφορά στα ενδιάμεσα αποτελέσματα αποστάσεων που δίδονται στο 2.3 (αντίστοιχα 2.5). Η αλλαγή που φαίνεται ότι πιθανόν να υπάρχει στο δευτερόλεπτο 45 στο Σχήμα 2.3 δεν επαληθεύεται κατά τη λεπτομερή αναζήτηση και γι' αυτό δεν υπάρχει ανίχνευση αλλαγής (Σχήμα 2.4).



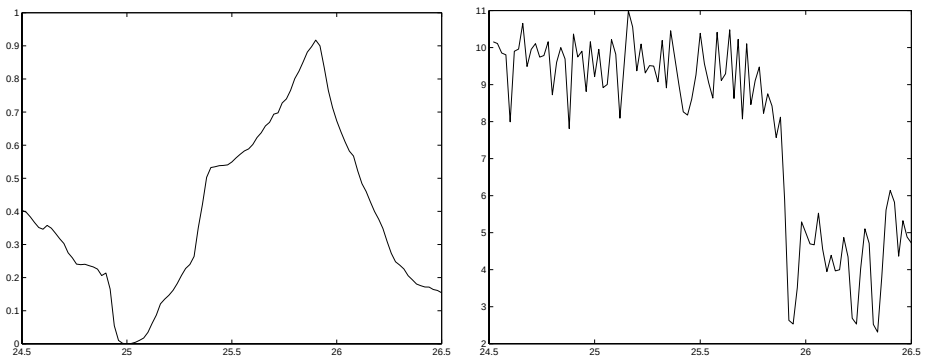
Σχήμα 2.5: Η απόσταση $D(i)$ και η κανονικοποιημένη απόσταση $D_n(i)$ για ένα σήμα με πολλές αλλαγές.



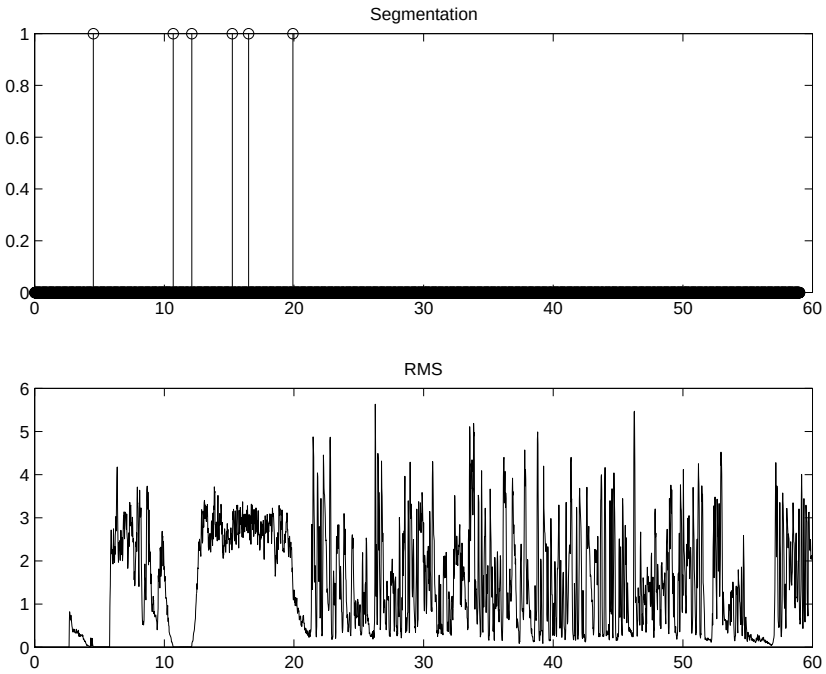
Σχήμα 2.6: Επάνω δίδονται τα διαστήματα όπου εντοπίζονται οι αλλαγές και κάτω το πλάτος του σήματος.



Σχήμα 2.7: Αριστερά εικονίζεται το $D(i)$ ως συνάρτηση του χρόνου με μονάδα 20 msec. Ως στιγμή αλλαγής λαμβάνεται η θέση που μεγιστοποιείται το $D(i)$. Το σήμα από ομιλία μεταβαίνει σε μουσική. Δεξιά επαληθεύεται η επιλογή από την καμπύλη του RMS για το αντίστοιχο διάστημα. Τελικά ο υπολογισμός της στιγμής αλλαγής έχει γίνει με μεγάλη ακρίβεια.



Σχήμα 2.8: Ανίχνευση μετάβασης από μουσική σε ομιλία. Το σφάλμα εντοπισμού της αλλαγής είναι πολύ μικρό.



Σχήμα 2.9: Επάνω δίδονται οι θέσεις όπου εντοπίζονται αλλαγές και κάτω το χρησιμοποιούμενο πλάτος σήματος. Ουδεμία απώλεια μετάβασης υπήρξε, ενώ υπερτμήθηκαν κάποια τμήματα.

Κεφάλαιο 3

Κατηγοριοποίηση

3.1 Περιγραφή πραγματικών χαρακτηριστικών

Οπως έχουμε αναφέρει για κάθε τμήμα που εξάγεται απο την τμηματοποίηση υπολογίζονται κάποια χαρακτηριστικά και στη συνέχεια από τις τιμές τους θα καθορισθεί το είδος του τμήματος. Τα χαρακτηριστικά αυτά αποτελούν τα πραγματικά χαρακτηριστικά του αλγορίθμου. Το σύνολό τους προκύπτει από τα δύο βασικά χαρακτηριστικά του πλάτους και της συχνότητας, οπότε ουσιαστικά δεν αυξάνουν καθόλου το κόστος των υπολογισμών.

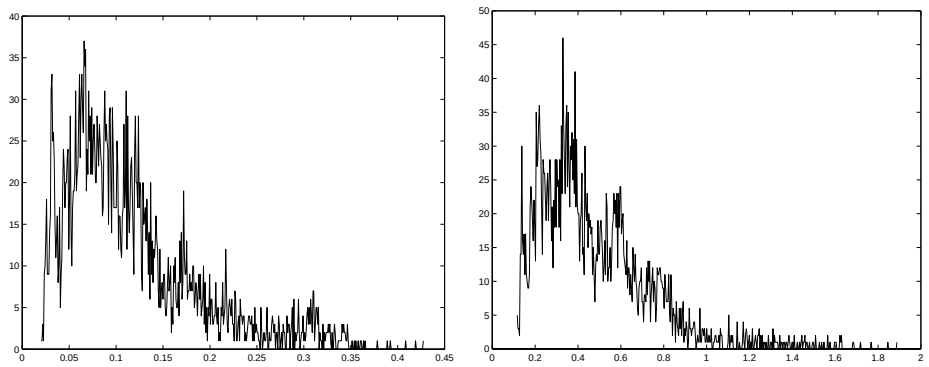
Η επιλογή τους έγινε με σκοπό να μπορούν να διαχωρίσουν τις κλάσεις, δηλαδή να εμφανίζουν διακριτή συμπεριφορά ανάλογα με την κλάση, οπότε με βάση τις τιμές που παίρνει το χαρακτηριστικό για το τμήμα να μπορεί να αποφασισθεί αν ανήκει στην κλάση αυτή. Εκτός από την διαχωρισιμότητα, η επιλογή στηρίχθηκε και στην ανεξαρτησία μεταξύ τους, δηλαδή στο γεγονός της εισαγωγής νέας πληροφορίας με την επιπλέον χρήση κάποιου χαρακτηριστικού μαζί με τα ήδη υπάρχοντα, που μεταφράζεται σε αύξηση του ποσοστού επιτυχίας του αλγορίθμου.

Η ουσιαστική διαφορά που διαχωρίζει την ομιλία από την μουσική είναι οι μικρές παύσεις που υπάρχουν στην ομιλία ανάμεσα στις συλλαβές, ενώ στη μουσική συνήθως δεν έχουμε τέτοια φαινόμενα. Ετσι τα περισσότερα χαρακτηριστικά που έχουμε χρησιμοποιήσει στηρίζονται στην παραπάνω ιδιότητα.

3.1.1 Κανονικοποιημένη διασπορά του RMS

Η κανονικοποιημένη διασπορά του RMS, σ_A^2, ορίζεται σαν το πηλίκο της διασποράς του RMS προς το τετράγωνο της μέσης τιμής του. Το παραπάνω χαρακτηριστικό υπολογίζεται σε διάστημα 1 sec και τελικά λαμβάνεται η μέση τιμή του για όλο το τμήμα. Δηλαδή σχετίζεται στενά με τον συντελεστή a της κατανομής χ^2. Η διαίρεση με την μέση τιμή γίνεται για να πετύχουμε το αναλλοίωτο του χαρακτηριστικού με την ένταση του σήματος, ιδιότητα που όπως είπαμε πρέπει να έχουν τα χρησιμοποιούμενα χαρακτηριστικά.

Ανάλογο χαρακτηριστικό έχει χρησιμοποιηθεί από τους E. Scheirer and M. Slaney [7] οι οποίοι μετρούσαν το ποσοστό των τιμών του RMS που είναι κάτω από το 50% της μέσης τιμής του σε διαστήματα 1 δευτερολέπτου. Τα πειράματα έδειξαν ότι στο 88% της ομιλίας εμφανίζονται τιμές του σ_A^2 πάνω από το 0.24 ενώ στο 84% της μουσικής συνήθως παίρνει τιμές κάτω από 0.24. Ιστογράμματα του χαρακτηριστικού αυτού για τυπικά σήματα φωνής και μουσικής δίδονται στο Σχήμα 3.1. Παρατηρούμε ότι τα δύο ιστογράμματα έχουν πολύ μικρή αλληλεπικάλυψη, άρα οι δύο κλάσεις μπορούν ναδιακριθούν πολύ καλά με αυτό το χαρακτηριστικό. Επιπλέον οι εμπειρικές κατανομές πιθανοτήτων μπορούν να προσεγγισθούν από τη γενικευμένη κατανομή χ^2. Αρα μπορεί να γίνει χρήση του λόγου πιθανοφανειών για την ταξινόμηση και την εύρεση του αναγκαίου κατωφλιού, που όπως αναφέρθηκε παραπάνω τέθηκε στο 0.24. Επομένως ακόμα και μόνο του το παραπάνω χαρακτηριστικό θα μπορούσε να δώσει καλά αποτελέσματα. Γι' αυτό το λόγο, όπως θα δούμε στον



Σχήμα 3.1: Ιστογράμματα της κανονικοποιημένης διασποράς του πλάτους του σήματος για μουσική (αριστερά) και ομιλία (δεξιά).

αλγόριθμο κατάταξης, η κανονικοποιημένη διασπορά του πλάτους του σήματος χρησιμοποιείται στον τελικό έλεγχο, δηλαδή για τα τμήματα που δεν έχουν κάποια ξεχωριστή ιδιότητα σε κάποιο από τα υπόλοιπα χαρακτηριστικά για να τα ταξινομήσουμε απευθείας σε κάποια κλάση, με υψηλή απόδοση.

3.1.2 Χρήση διελεύσεων από το μηδέν

Οπως είδαμε στην παράγραφο 1.3.2 το πλήθος των διελεύσεων από το μηδέν σχετίζεται με τη συχνότητα για το εξεταζόμενο τμήμα. Ετσι αν ένα μικρό τμήμα των 20 msec είναι κενό τότε ο αριθμός τών διελεύσεων από το 0 του σήματος θα είναι 0. Εάν το ποσοστό των μικρών τμημάτων των 20 msec που δεν έχουν διελεύσεις από το μηδέν υπερβαίνει κάποιο όριο, τότε έχουμε σίγουρα ομιλία. Το παραπάνω χαρακτηριστικό ($ZC0$) χρησιμοποιείται βοηθητικά για το χαρακτηρισμό τμημάτων ως ομιλία, όταν το ποσοστό αυτό ξεπερνάει το 0.1. Ωστόσο λόγω της πιθανής ύπαρξης θορύβου ή γρήγορης ομιλίας το ποσοστό της ομιλίας που ξεπερνάει το παραπάνω φράγμα είναι γύρω στο 40%, ενώ κανένα μουσικό κομμάτι δεν μπορεί να το υπερβεί.

Στα ιστογράμματα του μέσου πλήθους διελεύσεων από το μηδέν σε μουσική και ομιλία που εικονίζονται στα Σχήματα 1.3 και 1.4 φαίνεται ότι στην ομιλία το ποσοστό μικρών τιμών για το πλήθος διελεύσεων από το μηδέν είναι υψηλό, ενώ στην μουσική το αντίστοιχο ποσοστό είναι κατά μία τάξη μεγέθους μικρότερο. Αρα το παραπάνω κριτήριο που παίρνει υψηλές τιμές όταν το ZC είναι 0, ή κοντά στο 0, θα μπορούσε με κάποιο κατώφλι να βοηθήσει στο διαχωρισμό των δύο κλάσεων.

3.1.3 Συνδυασμένη χρήση συχνότητας και πλάτους σήματος

Ενα επιπλέον χαρακτηριστικό που θα μπορούσε να χρησιμοποιηθεί είναι το γινόμενο του πλάτους με τη συχνότητα (ZC*RMS). Αυτό σαν χαρακτηριστικό είναι σταθερό, αφού σε τμήματα ομιλίας που υπάρχουν μικρές παύσεις το ZC, όπως είδαμε, είναι κοντά στο 0, αλλά και το RMS επίσης είναι κοντά στο 0. Στη μουσική παρατηρείται πως το ZC είναι πιο ανεξάρτητο με το RMS, έτσι ακόμα και όταν το ZC είναι κοντά στο 0 το RMS μπορεί να παίρνει οποιεσδήποτε τιμές. Η παρακάτω σχέση χρησιμοποιήθηκε για τον υπολογισμό του χαρακτηριστικού

$$C_Z = e^{-\psi C}, \ \ \psi \in [1-5]$$

όπου

$$C = \frac{\sum_{i=1}^{N} A(i)z(i)}{2A_x - A_n - A_m}$$

με $A_x = \max\{A(i) : 1 \leq i \leq N\}$, $A_n = \min\{A(i) : 1 \leq i \leq N\}$ και $A_m = \text{median}\{A(i) : 1 \leq i \leq N\}$. Η κανονικοποίηση με το $2A_x - A_n - A_m$ έχει επιλεγεί, διότι στην ομιλία ο παρονομαστής συνήθως έχει

σχετικά υψηλή τιμή μιας και η μεσαία τιμή και η ελάχιστη τιμή του RMS είναι μικρές λόγω του τύπου του σήματος. Το C και το C_Z υπολογίζονται σε κάθε τμήμα. Το χαρακτηριστικό αυτό χρησιμοποιείται ως διορθωτικό, δηλαδή αν βρεθεί το $C_Z > \tau_1$, τότε αυτόματα αποφασίζουμε για ομιλία.

Ακόμα το χαρακτηριστικό αυτό θα μπορούσε επίσης να χρησιμοποιηθεί και ως ανιχνευτής θορύβου σε ομιλία με το ίδιο κατώφλι και με ψ κοντά στο 5. Δηλαδή για τα τμήματα που έχουν ήδη χαρακτηριστεί ως ομιλία, θα μπορούσε να υπολογισθεί η τιμή του C_Z και να τους δοθεί ενδεχόμενα επιπλέον ο χαρακτηρισμός ότι υπάρχει κάποιος θόρυβος μαζί με την ομιλία ανάλογα με την τιμή του C_Z.

3.1.4 Συχνότητα κενών τμημάτων

Η συχνότητα εμφάνισης κενών τμημάτων (F_v) είναι ένα από τα πραγματικά χαρακτηριστικά που μπορούν να διαχωρίσουν ικανοποιητικά την μουσική από την ομιλία. Αρχικά πρέπει να εντοπισθούν τα κενά διαστήματα. Μιας και έχει γίνει ο διαχωρισμός των 20 msec χρησιμοποιούνται τα ήδη μικρά αυτά διαστήματα και ανάμεσα τους εντοπίζονται εκείνα που μπορούν να χαρακτηρισθούν κενά. Αυτό γίνεται με βάση την τιμή των RMS και ZC σύμφωνα με την παρακάτω συνθήκη

$$(RMS < \mathrm{T}_1) \vee (RMS < 0.1 \max(RMS) \wedge RMS < T_2) \vee (ZC = 0)$$

Το $\max(RMS)$ αναφέρεται στη μέγιστη τιμή του RMS για όλο το τμήμα για αποφάσεις που λαμβάνονται σε κάθε διάστημα. Τελικά τα κενά τμήματα δημιουργούνται με τη συνένωση των μικρών γειτονικών διαστημάτων. Ετσι από το αρχικό σήμα δημιουργείται ένα δυαδικό σήμα με τιμή 1, όπου υπάρχει πληροφορία, αλλιώς είναι 0 στα τμήματα που υπάρχει σιωπή. Ο αριθμός των κενών τμημάτων που υπάρχουν σε ένα τμήμα για ταξινόμηση προς την χρονική διάρκεια όλου του τμήματος ορίζεται ως συχνότητα κενών τμημάτων (F_v). Το παραπάνω χαρακτηριστικό μετράει τη συχνότητα συλλαβών και λέξεων στην ομιλία. Η μουσική πολύ σπάνια εμφανίζει διακοπές και μάλιστα με συχνότητα εμφάνισης πολύ μικρότερη από την ομιλία όπου σχεδόν συνέχεια έχουμε διακοπές.

Στην ομιλία ισχύει πάντα $F_v > 0.6$, αφού είναι αδύνατο να μην υπάρχουν συλλαβές ή λέξεις. Αντίθετα στη μουσική συνήθως $F_v \approx 0$ και η πλειψηφία (65%) έχει $F_v < 0.6$. Ακόμα στα πειράματα έχει φανεί πως στην ομιλία το σύνολο των κομματιών έχει $F_v < 4.5$, ενώ στη μουσική υπάρχουν περιπτώσεις (πολύ λίγες) που το παραπάνω φράγμα έχει ξεπερασθεί. Τα παραπάνω επιβεβαιώνονται και από το Σχήμα 3.2.

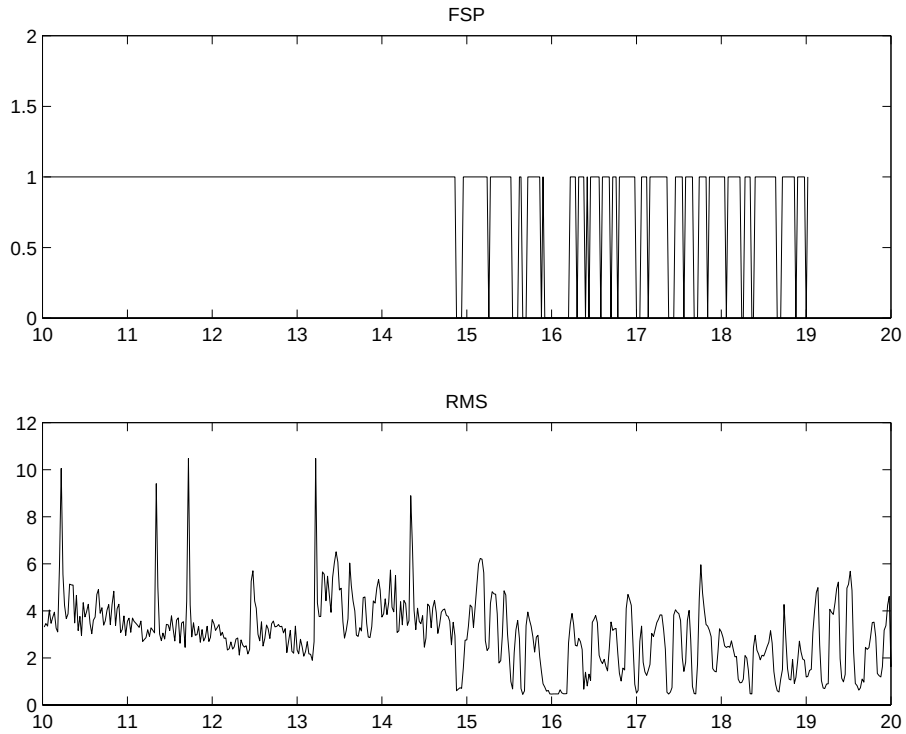
3.1.5 Μέγιστη συχνότητα

Η μέγιστη συχνότητα ($\max(ZC)$) προκύπτει από την μέγιστη τιμή των ZC από τις περιοχές του σήματος που έχουμε σχετικά υψηλή ένταση. Οι παραπάνω περιοχές επιλέγονται με ένα όριο στην ένταση, ώστε να μην μετρήσουμε περιοχές με θόρυβο όπου κατά κανόνα κυριαρχούν υψηλές συχνότητες. Στην ομιλία αυτή η τιμή φράσσεται από μια μέγιστη συχνότητα (2.4 kHz), ενώ στην μουσική φαίνεται να μπορεί να πάρει και μεγαλύτερες τιμές (20 kHz).

Το ποσοστό της μουσικής που έχει πάνω από 2.4 kHz συχνότητα είναι σχετικά μικρό ($\approx 2\%$) όμως λόγω των μεγάλων και συχνών μεταβολών του με βάση τα προηγούμενα χαρακτηριστικά δινόταν συνήθως ως ομιλία, γεγονός που διορθώνεται με την προσθήκη του παραπάνω χαρακτηριστικού.

3.2 Αλγόριθμος κατάταξης

Κάθε νέο τμήμα που προκύπτει από την τμηματοποίηση θα πρέπει να κατηγοριοποιηθεί σε μία από τις τρεις κλάσεις: σιωπή, ομιλία, μουσική. Γι' αυτό υπολογίζονται τα πραγματικά χαρακτηριστικά από ολόκληρο το τμήμα, και με βάση τις τιμές των χαρακτηριστικών λαμβάνεται η απόφαση κατάταξης. Αρχικά ελέγχεται η περίπτωση της σιωπής.



Σχήμα 3.2: Μετάβαση από ομιλία σε μουσική. Στο κάτω σχήμα φαίνεται το RMS, ενώ στο πάνω εικονίζεται η ανίχνευση κενών τμημάτων (τιμή '0' για σιωπή). Παρατηρούμε πως στην ομιλία υπάρχουν αρκετά τμήματα σιωπής σε αντίθεση με τη μουσική. Η συχνότητα εμφάνισης των κενών δίδει το F_v για κάθε τμήμα.

3.2.1 Αναγνώριση κενών τμημάτων

Αφού γίνει η τμηματοποίηση για το κάθε τμήμα που προκύπτει, αποφασίζεται αν είναι μουσική ή ομιλία ή κενό. Αρχικά αποφασίζεται εάν το τμήμα είναι κενό με χρήση φράγματος στην παρακάτω έκφραση που εξαρτάται από την ισχύ του σήματος.

$$\mathrm{E} = 0.7A_m + \frac{0.3}{N}\sum_{i=1}^{N} A(i) \tag{3.1}$$

Ο παραπάνω συνδυασμός μέσης και μεσαίας τιμής είναι αναγκαίος, αφού η μέση τιμή που φαίνεται με μια πρώτη ματιά να αρκεί, επηρεάζεται σε μεγάλο βαθμό από τις μεγάλες τιμές του RMS που, αν και μπορεί να είναι λίγες σε πλήθος μπορούν να μετατοπίσουν αρκετά το E. Οι μεγάλες τιμές του RMS μπορεί να προκύψουν από ένα μικρό σφάλμα στην τμηματοποίηση ή από στιγμιαία ισχυρό θόρυβο. Ετσι η χρήση μόνο της μέσης τιμής θα έδινε κάτι ασταθές. Η χρήση της μεσαίας τιμής δίνει μια ευστάθεια στο E και ουσιαστικά χαρακτηρίζει το τμήμα με βάση την πλειοψηφία των μετρήσεων, δηλαδή αν το μεγαλύτερο μέρος του τμήματος έχει ισχύ κάτω από την οριακή ισχύ κενού τότε χαρακτηρίζεται ως κενό. Βέβαια υπάρχουν περιπτώσεις που το παραπάνω δεν είναι σωστό όπως για παράδειγμα σε ομιλία με πολλά κενά. Για τους παραπάνω λόγους, αλλά και με βάση τα πειραματικά αποτελέσματα, έγινε φανερό πως ο παραπάνω συνδυασμός με σχετικά υψηλότερο βάρος στη μεσαία τιμή αποτελεί μια σωστή και σταθερή λύση.

Χαρακτηριστικά	Απόδοση για μουσική	Απόδοση για ομιλία
$ZC0$	90%	60%
σ_A^2	84%	88%
C_Z	90%	60%
σ_A^2, $ZC0$	80%	97%
σ_A^2, C_Z	82%	97%
C_Z, σ_A^2	80%	97%
$ZC0$, σ_A^2	70%	97%
F_v, σ_A^2	88%	92%
F_v, C_Z, max(ZC), $ZC0$, σ_A^2	92%	97%

Πίνακας 3.1: Απόδοση χαρακτηριστικών ξεχωριστά και συνδυασμένα.

3.2.2 Αλγόριθμος κατάταξης σε μουσική ή ομιλία

Αν διαπιστωθεί ότι δεν έχουμε σιωπή, εξετάζουμε αν η τιμή του χαρακτηριστικού F_v είναι μικρότερη από 0.6 για να διαπιστωθεί με σχεδόν βεβαιότητα εάν έχουμε μουσική. Με αυτόν το έλεγχο το 50% περίπου της μουσικής κατατάσσεται σωστά, ενώ δεν υπάρχει καμία κατάταξη από ομιλία σε μουσική. Στη συνέχεια εξετάζεται αν οι τιμές των $RMS * ZC$ ή $ZC0$ υπερβαίνουν κάποιο ανώτατο όριο, οπότε το τμήμα χαρακτηρίζεται αυτόματα ομιλία. Ακόμα ελέγχεται αν το maxZC υπερβαίνει κάποιο όριο οπότε έχουμε σίγουρα μουσική. Τελικά θα μείνει ένα ποσοστό ομιλίας και μουσικής που δεν έχουν ταξινομηθεί ακόμη, και η απόφαση γι' αυτά θα ληφθεί με ένα φράγμα στην κανονικοποιημένη διασπορά γύρω στο 0.24. Στον τελευταίο ουσιαστικά έλεγχο θα φθάσουν τα τμήματα ομιλίας και μουσικής που δεν ξεχωρίζουν από τις οριακές τιμες των πραγματικών χαρακτηριστικών, άρα αυτά που διαχωρίζονται δύσκολα και οι λάθος κατατάξεις είναι σχεδον σίγουρο ότι οφείλονται στον τελευταίο έλεγχο που υφίσταται περίπου το 40% της μουσικής και το 60% της ομιλίας.

3.3 Αποτελέσματα κατάταξης για κάθε χαρακτηριστικό

Στον πίνακα 3.1 δίδονται τα ποσοστά επιτυχίας διαφόρων χρήσεων των χαρακτηριστικών, είτε χωριστά είτε συνδυασμένα, για κατάταξη στις κλάσεις μουσικής και ομιλίας. Η σειρά που γράφονται είναι εκείνη που χρησιμοποιούνται. Οι έλεγχοι εκτελούνται διαδοχικά με τιμές για τα φράγματα που έχουν βελτιστοποιηθεί για να μεγιστοποιηθεί η απόδοση. Οπως μπορούμε να διαπιστώσουμε σημαντική είναι η συμβολή του F_v για την αύξηση του ποσοστού στη μουσική. Βέβαια ο συνδυασμός όλων δίνουν το καλύτερο αποτέλεσμα δηλαδή 97% για ομιλία και 92% για μουσική. Φαίνεται πως κάθε χαρακτηριστικό δρα συμπληρωματικά και έτσι το 84% με 88% που πετυχαίνει μόνο η κανονικοποιημένη διασπορά του πλάτους του σήματος εμφανίζει άνοδο 8% για κάθε κλάση. Η χρήση όλων των χαρακτηριστικών δεν αυξάνει την πολυπλοκότητα του αλγορίθμου, αφού όλα προκύπτουν αρκετά απλά από τα δύο βασικά χαρακτηριστικά ZC και RMS.

Αντί για τη σταδιακή εφαρμογή ελέγχων θα μπορούσε να χρησιμοποιηθεί κάποιο εκπαιδευμένο νευρωνικό δίκτυο, ή κάποια απόσταση όπως η Mahalanobis. Τα παραπάνω αποφεύχθηκαν για λόγους απλότητας, γιατί τα χρησιμοποιούμενα βασικά χαρακτηριστικά είναι σχεδόν ανεξάρτητα και επιπλέον δεν ακολουθούν την κατανομή Gauss για την οποία θα ήταν κατάλληλη η απόσταση Mahalanobis.

Κεφάλαιο 4

Αποτελέσματα

4.1 Πειραματικά αποτελέσματα

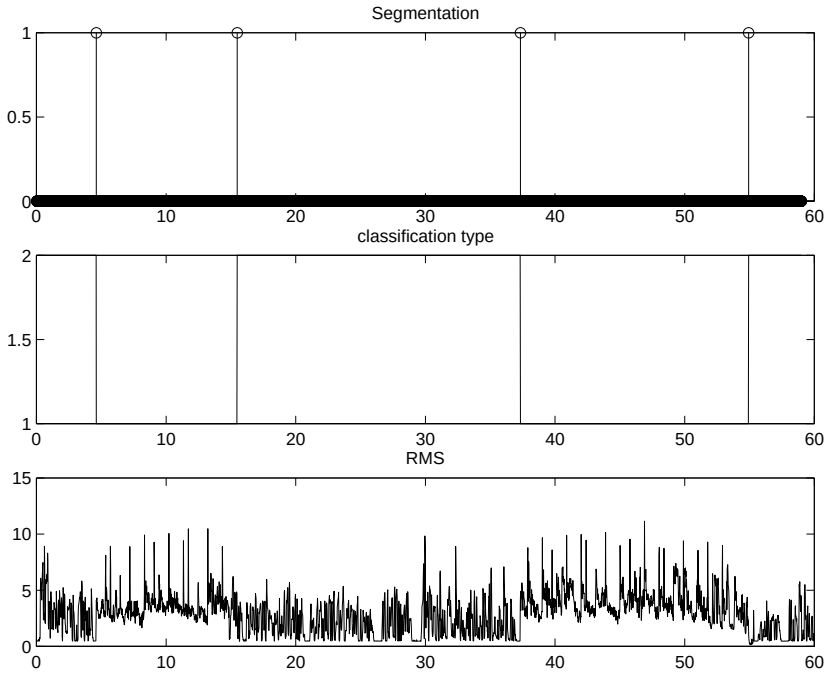
Η δοκιμή του αλγορίθμου έγινε σε αρχεία ήχου δειγματοληψίας από 11025 Hz έως και 44100 Hz χωρίς να επηρεάζεται καθόλου η απόδοσή του. Η βάση που χρησιμοποιήθηκε περιείχε αρχεία που προήλθαν είτε απο ηχογραφήσεις (15%), είτε από το Διαδίκτυο (15%), είτε από συλλογές CD (70%) με μουσική ή ομιλία.

Για τον αλγόριθμο της ταξινόμησης δοκιμάστηκαν 92 κομμάτια ομιλίας διάρκειας 11328 sec που χωρίστηκαν με τον αλγόριθμο της τμηματοποίησης σε 800 περίπου τμήματα και το ποσοστό επιτυχίας έφτασε το 97%. Οσον αφορά την απόδοση για την μουσική έφτασε το 92% σε 80 κομμάτια διάρκειας 3131 sec που χωρίστικαν σε 400 τμήματα. Τα περισσότερα κομμάτια περιείχαν θόρυβο αλλά και κενά διαστήματα τα οποία αναγνωρίζονταν όλα στην κατάταξη. Τα ποσοστά προκύπτουν από την χρονική διάρκεια αν και τα ίδια είναι περίπου αν υπολογιστούν με βάση τα τμήματα. Οι δοκιμές έγιναν στο μαθηματικό προγραμματιστικό περιβάλλον MATLAB, ενώ ο αλγόριθμος έχει υλοποιηθεί και σε γλώσσα C.

Αποτελέσματα δίδονται στα Σχήματα 4.1, 4.2 και 4.3. Κάθε σχήμα αποτελείται από τρία μέρη, που φαίνονται σα συνάρτηση του χρόνου: (α) το αποτέλεσμα της τμηματοποίησης, (β) το αποτέλεσμα της κατηγοριοποίησης, όπου η τιμή 1 αντιστοιχεί στη μουσική, η τιμή 2 στην ομιλία και η τιμή 3 στη σιωπη, και (γ) το πλάτος του σήματος. Δίδονται επίσης δύο αποτελέσματα λαθεμένης κατάταξης (Σχήματα 4.4 και 4.5). Και στις δύο περιπτώσεις πρόκειται για μουσική με σημαντικές αυξομειώσεις στην ένταση με αποτέλεσμα να έχει όμοια χαρακτηριστικά με την ομιλία.

4.2 Πολυπλοκότητα και ταχύτητα αλγορίθμου

Ενας από τους βασικούς σκοπούς της εργασίας μας ήταν η επίτευξη επιδόσεων πραγματικού χρόνου. Η προτεινόμενη μέθοδος έχει πραγματικά αυτή την ιδιότητα, αφού η τμηματοποίηση επιτυγχάνεται με καθυστέρηση μόλις 4 sec από το λήψη του σήματος, ενώ η ταξινόμηση γίνεται για κάθε νέο τμήμα που εμφανίζεται σε πραγματικό χρόνο. Πράγματι και οι υπολογισμοί αλλά και η πλοκή του αλγορίθμου είναι κατάλληλες γι' αυτό το στόχο. Οσον αφορά τις αριθμητικές και λογικές πράξεις όπως είδαμε είναι αρκετά απλές. Το πλάτος και η συχνότητα του σήματος που υπολογίζονται ανά 20 msec, δεν περιέχουν κάποιο πολύπλοκο υπολογισμό, ενώ τα πραγματικά χαρακτηριστικά του υποκεφαλαίου 3.1 υπολογίζονται απευθείας από τα βασικά που ορίσθησαν στην Εισαγωγή. Ακόμα ο επειδή στον αλγόριθμο τμηματοποίησης χρησιμοποιούνται δύο στάδια, μειώνονται κατά πολύ οι υπολογισμοί, κατά 50 φορές περίπου λιγότεροι από το αν γίνονταν σε ένα. Η πλοκή είναι $O(N)$, όπου N η διάρκεια του σήματος, και είναι η ελάχιστη δυνατή, αφού είναι αναγκαίο να διαβαστεί ολόκληρο το σήμα. Οι περισσότερες μέθοδοι έχουν πλοκή πάνω απο O($N \log N$) αφού χρησιμοποιούν FFT κάτι που έχουμε αποφύγει με τη χρήση του ZC.



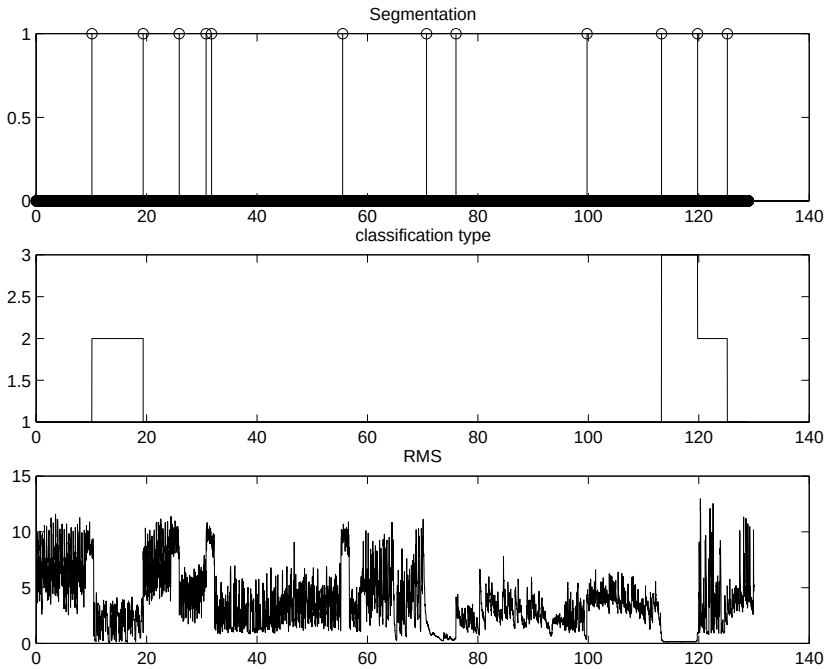
Σχήμα 4.1: Αποτέλεσμα κατηγοριοποίησης χωρίς σφάλματα. Το δεύτερο και το τέταρτο τμήμα είναι μουσική, ενώ τα υπόλοιπα είναι ομιλία.

Τελικά ο απαιτούμενος χρόνος υπολογισμών μπορεί να είναι μικρότερος από τη χρονική διάρκεια του σήματος. Η υλοποίηση σε C η οποία περιείχε και αποκωδικοποίηση αφού έπαιρνε είσοδο σε μορφή MP3 έτρεχε 10 φορές ταχύτερα από την διάρκεια του κομματιού σε Pentium III 450 MHz, χωρίς την αποκωδικοποίηση γινόταν ακόμα και 50 με 100 φορές ταχύτερο.

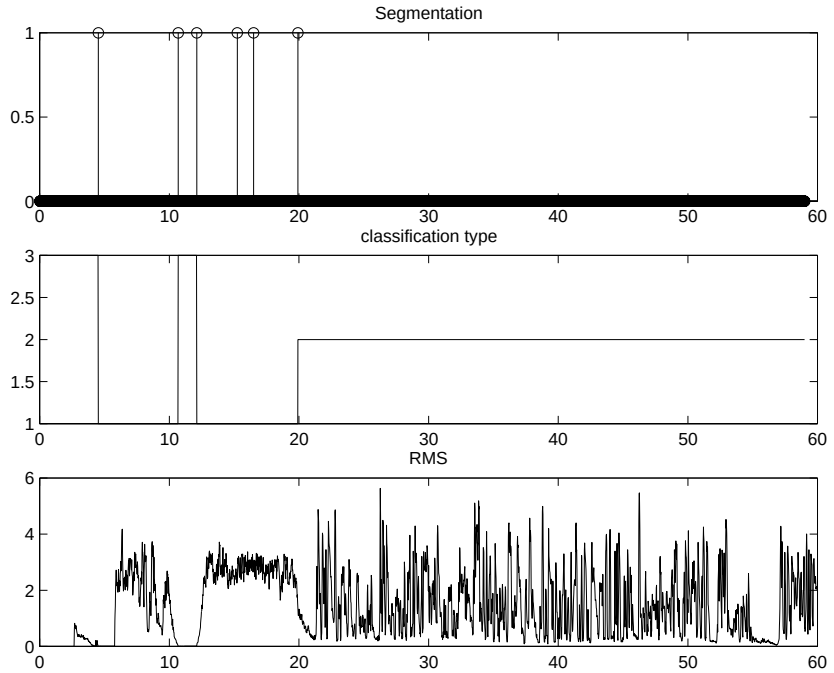
4.3 Συγκρίσεις με άλλες μεθόδους

Τελικά μπορούμε να πούμε πως το ποσοστό επιτυχίας της μεθόδου μας σταθεροποιείται πάνω από 95% με απόδοση 92% για μουσική 97% για ομιλία και σχεδόν 100% για σιωπή. Ενώ η τμηματοποίηση μόνη της φτάνει στο 97%. Οι P. Moreno και R. Rifkin [4] σχεδίασαν έναν αλγόριθμο για ταξινόμηση ηχητικών σημάτων σε 3 κλάσεις ομιλία, μουσική και ‘‘άλλο” χρησιμοποιώντας μοντέλα μίξεων κατανομών Gauss για την εξαγωγή χαρακτηριστικών και μηχανές διανυσμάτων υποστήριξης για τον ταξινομητή. Το ποσοστό επιτυχίας τους έφτασε το 81.8%. Ο J. Foote [1] πέτυχε ποσοστό 81% στην ταξινόμηση σε 7 κλάσεις (μουσική, ομιλία, ομιλία με μουσική , σιωπή, τηλέφωνο, θόρυβος και ομιλία με θόρυβο) χρησιμοποιώντας παράθυρα 0.5 sec. Οι G. Tzanetakis και P. Cook [10] ανέπτυξαν έναν αλγόριθμο τμηματοποίησης χρησιμοποιώντας 5 χαρακτηριστικά, παράθυρα 1 sec και την απόσταση Mahalanobis για την έυρεση σημείων αλλαγής. Η απόδοση του έφτασε το 87%.

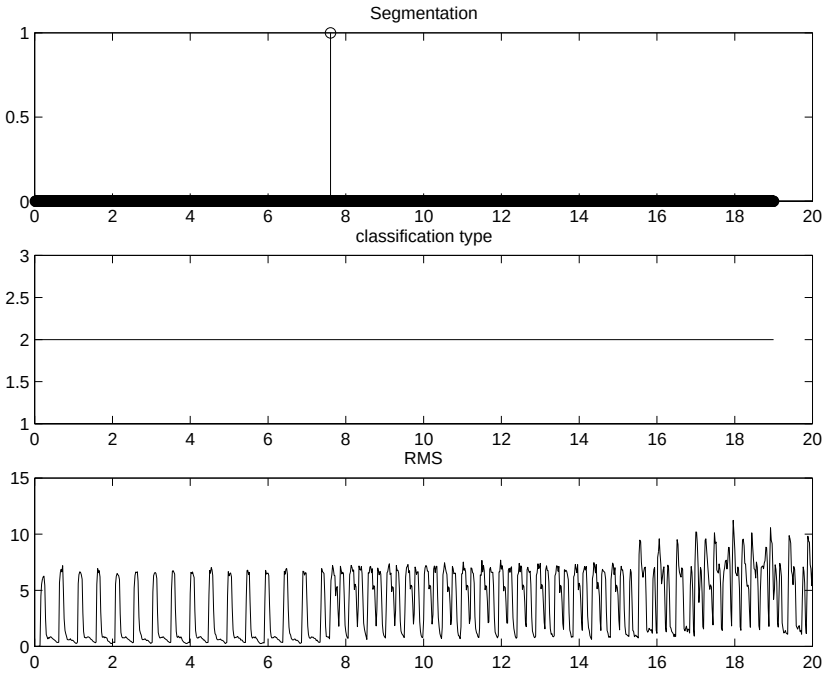
Οι περισσότεροι από τους αλγόριθμους αυτούς χρησιμοποιούν περισσότερα χαρακτηριστικά από τη δική μας μέθοδο. Η δική μας εργασία είχε σκοπό να εντοπίσει τα μεγέθη εκείνα που διαφοροποιούνται όταν αλλάζουμε κλάση και σε αυτό οφείλονται η υψηλή απόδοσή της.



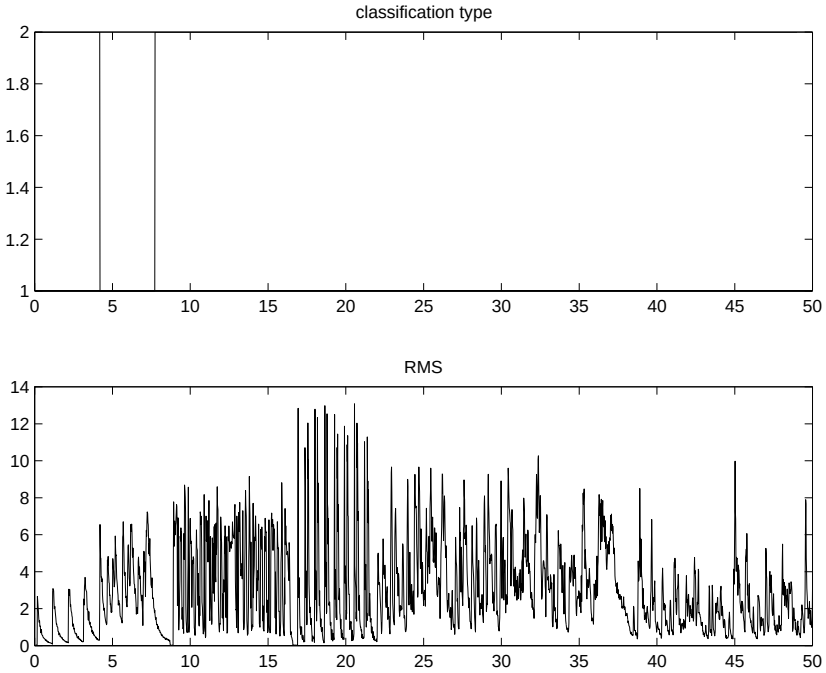
Σχήμα 4.2: Το αρχείο αυτό υπερ-τμηματοποιήθηκε, αλλά η ταξινόμηση έδωσε σωστές κατατάξεις.



Σχήμα 4.3: Αποτέλεσμα ορθής ταξινόμησης σε δύο τμήματα μουσικής που διαχωρίζονται με κενά και ακολουθεί ομιλία.



Σχήμα 4.4: Περίπτωση λαθεμένης κατάταξης μουσικής σε ομιλία.



Σχήμα 4.5: Περίπτωση λαθεμένων κατατάξεων λόγω αυξομειώσεων της έντασης και ύπαρξης μικρών διακοπών στη μουσική.

Κεφάλαιο 5

Επίλογος

5.1 Συμπεράσματα

Στην εργασία αυτή αναπτύχθηκε ένας γρήγορος και αποτελεσματικός αλγόριθμος που τμηματοποιεί και ταξινομεί σε πραγματικό χρόνο ηχητικά σήματα σε 3 κλάσεις: ομιλία, μουσική και σιωπή. Τελικά φαίνεται να αρκεί η κατανομή της ενέργειας σε μικρά τμήματα για να γίνει η τμηματοποίηση, ενώ χωρίς επιπλέον κόστος με τη βοήθεια της κύριας συχνότητας μπορεί να γίνει και η ταξινόμηση. Σκοπός μας ήταν να ελαχιστοποιήσουμε το κόστος, ώστε ο αλγόριθμος να μπορεί να εκτελείται σε πραγματικό χρόνο, αλλά και να αυξήσουμε την απόδοση. Οπως είδαμε τα παραπάνω μπορούν να συνυπάρξουν αρκεί να εντοπίσουμε τα κατάλληλα εκείνα χαρακτηριστικά που θα διαφοροποιούν τις δύο κλάσεις και θα είναι σχετικά ανεξάρτητα μεταξύ τους, ώστε να έχουμε νέα πληροφορία με το συνδυασμό τους. Τέτοια είναι το μέσο πλάτος και η μέση συχνότητα.

Μια από τις χρήσεις της μεθόδου λόγω της μεγάλης ταχύτητάς της είναι η τμηματοποιήση και ο χαρακτηρισμός βάσης δεδομένων από ηχητικά αρχεία και ο διαχωρισμός τους στις 3 κλάσεις. Ενα τέτοιο σύστημα έχει υλοποιηθεί στο πλαίσιο του έργου ΠΑΝΟΡΑΜΑ, όπου ο αλγόριθμος που περιγράψαμε έχει ενσωματωθεί. Αλλά θα μπορούσε να χρησιμοποιηθεί και σε ένα οποιοδήποτε πραγματικού χρόνου σύστημα λόγω των επιδόσεών του, όπως για παρακολούθηση ενός ραδιοφωνικού προγράμματος και ο διαχωρισμός σε τμήματα ομιλίας και μουσικής.

5.2 Μελλοντικές επεκτάσεις

Η μέθοδος θα μπορούσε να επεκταθεί αλλά και να εφαρμοστεί σαν αρχικό στάδιο για επιπλεόν επεξεργασία. Καταρχήν θα μπορούσαν να προστεθούν επιπλέον χαρακτηριστικά, βασικά και πραγματικά, ώστε να αυξηθεί ο αριθμός των κλάσεων που αναγνωρίζει. Επίσης με προσθήκη κατάλληλων χαρακτηριστικών θα μπορούσε να χρησιμοποιηθεί για αναγνώριση φωνής ή μουσικής από μια βάση δεδομένων, αφού προηγηθεί η τμηματοποίηση. Αλλά και να εντοπίσει ομιλητές και να δώσει στατιστικά για τον χρόνο και τον ρυθμό ομιλίας τους. Το γεγονός ότι εκτελείται σε πραγματικό χρόνο χωρίς υψηλό κόστος, αλλά και το υψηλό ποσοστό επιτυχίας της δίνουν την δυνατότητα να χρησιμοποιηθεί για προεπεξεργασία αναγνώρισης ομιλίας στα τμήματα που θα χαρακτηριστούν ως ομιλία με τη μέθοδο αυτή. Ακόμα θα μπορούσε να συνδυασθεί με τμηματοποιητή βίντεο για να βοηθήσει στην τεμαχισμό του βίντεο σε σκηνές, αφού συχνά όταν αλλάζει η σκηνή αλλάζει και ο ήχος.

Βιβλιογραφία

[1] J. Foote. An Overview of audio information retrieval. *Multimedia Systems*, pp. 2--10, 1999.

[2] P.R. Krishnaiah and P.K. Sen, *Handbook of statistics: Nonparametric methods*, North-Holland, 1984.

[3] D. Krubsack and R. Niederjonh. An Autocorrelation Pitch Detector and Voicing Decision with Confidence Measures Developed for Noise-Corrupted Speech. *IEEE Trans. on Signal Processing*, Vol. 39, No. 2, Feb. 1991.

[4] P. Moreno and R. Rifkin. Using the Fisher kernel method for web audio classification. In *Proc. IEEE Conf. on Acoustics,Speech and Signal Processing*, pp. 1921--1924, 2000.

[5] J. Saunders. Real-time discrimination of broadcast speech/music.In *Proc. IEEE Intern. Conf on Acoustics, Speech and Signal Processing*, pp. 993--996, 1996.

[6] M. Seck, F. Bimbot, D. Zugah, and B. Delyon. Two-class signal segmentation for speech/music detection in audio tracks. In *Proc. Eurospeech*, Vol. 6, pp. 2801-2804, Sept. 1999.

[7] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discrimination. In *Proc. IEEE Conf. on Acoustics,Speech and Signal Processing*, 1997.

[8] A. Spanias. Speech coding: a tutorial review. *Proc. of the IEEE*, Vol. 82, pp. 1541--1582, Oct. 1994.

[9] G. Tzanetakis and P. Cook. A framework for audio analysis based on classification and temporal segmentation. In *Proc. 25th Euromicro Conference. Workshop on Music Techonology and Audio Processing*, 1999.

[10] G. Tzanetakis and P. Cook. Multiframe audio for browsing and annotation. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, Oct. 1999.

[11] E. Wold, T. Blum, D. Keistar, and J. Wheaton. Content based classification, search, and retrieval of audio. *IEEE Multimedia Magazine*, pp. 27--36, 1996.

[12] T. Young and K.-S.Fu, eds. *Handbook of pattern recognition and image processing*, Academic Press, 1986.