ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΜΕΣΩΝ ΜΑΖΙΚΗΣ ΕΝΗΜΕΡΩΣΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΑΥΤΟΜΑΤΗ ΑΝΑΓΝΩΡΙΣΗ ΗΧΩΝ ΣΕ ΡΑΔΙΟΦΩΝΙΚΕΣ ΕΚΠΟΜΠΕΣ

ΟΝΟΜΑΤΕΠΩΝΥΜΟ ΣΠΟΥΔΑΣΤΗ : ΚΑΤΕΛΑ ΑΝΑΣΤΑΣΙΑ.

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΚΟΥΤΡΑΣ ΑΘΑΝΑΣΙΟΣ

ΠΥΡΓΟΣ - 2016

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΜΗ ΛΟΓΟΚΛΟΠΗΣ

Βεβαιώνω ότι είμαι η συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία.

Επίσης, έχω αναφέρει τις οποίες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών η λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες.

Ακόμη δηλώνω ότι αυτή η γραπτή εργασία προετοιμάστηκε από εμένα προσωπικά και αποκλειστικά και ειδικά για την συγκεκριμένη πτυχιακή εργασία ότι θα αναλάβω πλήρως τις συνέπειες εάν η εργασία αυτή αποδειχτεί ότι δεν μου ανήκει.

ΟΝΟΜΑΤΕΠΩΝΥΜΟ ΣΠΟΥΔΑΣΤΗ	ΑΡΙΘ.ΜΗΤΡΩΟΥ	ΥΠΟΓΡΑΦΗ
ΚΑΤΕΛΑ ΑΝΑΣΤΑΣΙΑ	1507	……..

ΠΕΡΙΛΗΨΗ

Ο σκοπός της παρούσας πτυχιακής εργασίας είναι η μελέτη διαφόρων συστημάτων ταξινόμησης ως προς τα αποτελέσματά τους για ακουστικά σήματα από ραδιοφωνικές εκπομπές.

Πιο συγκεκριμένα, συλλέχτηκαν ακουστικά κομμάτια από ραδιοφωνικές εκπομπές τα οποία, μετά την μη αυτοματοποιημένη τμηματοποίηση τους μέσω του εργαλείου PRAAT, συνετέλεσαν την βάση δεδομένων που χρησιμοποιήθηκε. Μέσω του προγράμματος MARSYAS, έγινε η εξαγωγή τριών συνόλων δεδομένων, πάνω στα οποία βασίστηκε η κατηγοριοποίηση μέσω αλγορίθμων ταξινόμησης.

Ο τελικός στόχος αυτής της εργασίας, είναι να αξιολογηθούν τα ποσοστά επιτυχούς ταξινόμησης των αλγορίθμων σε συνδυασμό με το σύνολο των χαρακτηριστικών που εξήχθησαν και να συγκρίνουμε τα αποτελέσματα μεταξύ τους.

Τα αποτελέσματα μας έδειξαν πως ο αλγόριθμος SVM είναι πιο αποδοτικός για όλα τα σύνολα χαρακτηριστικών που εξήχθησαν, ενώ ακολουθεί ο αλγόριθμος C4.5 που χρησιμοποιεί δέντρα απόφασης, με αρκετά μεγάλα ποσοστά επιτυχούς ταξινόμησης. Ο αλγόριθμος Naïve Bayes μας δίνει υψηλά ποσοστά μόνο για κάποια από τα σύνολα χαρακτηριστικών, ενώ ο αλγόριθμος ZeroR κρίνεται ακατάλληλος σε κάθε περίπτωση. Στην συνέχεια, γίνεται εκτενής ανάλυση των μεθόδων που χρησιμοποιήθηκαν, τόσο για την προ-επεξεργασία και την δημιουργία της βάσης δεδομένων, την εξαγωγή των παραμέτρων, τις μεθόδους κατηγοριοποίησης, των εργαλείων ανοιχτού κώδικα που χρησιμοποιήθηκαν, καθώς και η αναλυτική σύγκριση των αποτελεσμάτων που προέκυψαν κατά την διεξαγωγή της πειραματικής διαδικασίας.

Λέξεις Κλειδιά

Αυτόματη αναγνώριση ήχων, τμηματοποίηση ηχητικού σήματος, ταξινόμηση ήχων, χαρακτηριστικά , επεξεργασία σήματος , ραδιοφωνικές εκπομπές, PRAAT, WEKA

ABSTRACT

The purpose of this thesis is to analyze the various classification systems and the results they produce, applied on audio signals from radio broadcasts.

More specifically, audio tracks from radio broadcasts were collected and after a non-automatic segmentation through the PRAAT tool, they contributed to the database used. Three sets of features were extracted, through the MARSYAS program, and they were used for the categorization by the classification algorithms.

The main goal is to analyze and interpret the correct classification rate, using several classification algorithms and comparing the results with each other.

The results demonstrated that the SVM algorithm is more efficient for all sets of features that were exported, while the algorithm C4.5, which uses decision trees, follows at great correct classification rates. The Naïve Bayes algorithm gives high rates for some of the feature sets, while the algorithm ZeroR is considered inefficient in every case. In this essay, there is an extensive analysis of the methods used for the pre-processing of the audio signal, the creation of the database, the exported parameters, the classification methods, the open source tools and the programs that were used, and the detailed comparison of the results obtained during the test procedure.

Keywords:

automatic audio recognition, audio segmentation, audio classification, features, signal processing, Broadcast News, PRAAT, WEKA

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω πρωτίστως την οικογένεια μου και ιδιαίτερα τους γονείς μου, που με στήριξαν με κάθε δυνατό τρόπο κατά την διάρκεια των σπουδών μου, τον επιβλέποντα καθηγητή μου κ. Κούτρα Αθανάσιο για την συνεργασία και την καθοδήγηση καθ' όλη την διάρκεια της πτυχιακής εργασίας και στον σύντροφό μου, Κανελλάκη Νικόλαο, που με ενθάρρυνε και με στήριξε για την επιτυχή ολοκλήρωσή της.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΕΙΣΑΓΩΓΗ

Στην εποχή μας, η τεχνολογία αναπτύσσεται με ταχείς ρυθμούς και οι πληροφορίες που διατίθενται μέσω του Διαδικτύου, καθημερινά αυξάνονται.

Οι τηλεοπτικές και ραδιοφωνικές εκπομπές, οι ηχογραφήσεις συναντήσεων, τα ηχητικά μηνύματα μέσω email, υπάρχουν παντού πλέον. Ωστόσο, αυτός ο συνεχώς αυξανόμενος όγκος ηχογραφημένων δεδομένων, εμποδίζει την διαχείριση, την οργάνωση, την πλοήγηση και την ανάκτησή τους.

Έτσι, δημιουργήθηκε η ανάγκη διαχείρισης μεγάλων δεδομένων και η αυτοματοποίηση κάποιων διαδικασιών. Κάποιες από αυτές, είναι η τμηματοποίηση και η καταχώρηση των σημάτων ήχου σε διάφορες κατηγορίες, ανάλογα το περιεχόμενο τους.

Στην παρούσα εργασία, παρουσιάζεται η αυτόματη αναγνώριση ήχων, οι οποίοι παρουσιάζονται σε ραδιοφωνικές εκπομπές, με χρήση τεχνικών επεξεργασίας σήματος. Στο Κεφάλαιο 1, ορίζεται το πεδίο της Αυτόματης Αναγνώρισης Ήχων και οι εφαρμογές του και γίνεται μία ανασκόπηση στην Διεθνή Βιβλιογραφία, ενώ στο τέλος του κεφαλαίου αναλύονται θεωρητικά τα βασικά είδη των ήχων που συναντάμε στις ραδιοφωνικές εκπομπές. Στο Κεφάλαιο 2 αναλύεται το θεωρητικό υπόβαθρο των χαρακτηριστικών των ακουστικών σημάτων των ραδιοφωνικών εκπομπών, οι αλγόριθμοι ταξινόμησης που χρησιμοποιήθηκαν κατά την πειραματική διαδικασία καθώς και τα μέτρα με τα οποία αξιολογήθηκαν τα αποτελέσματα των αλγορίθμων ταξινόμησης. Στο Κεφάλαιο 3 παρουσιάζεται η μεθοδολογία που ακολουθήθηκε στην παρούσα εργασία ενώ στο Κεφάλαιο 4 παρουσιάζονται τα αποτελέσματα της ταξινόμησης, αναλύονται και σχολιάζονται τα μέτρα αξιολόγησης και προκύπτουν κάποια συμπεράσματα.

ΚΕΦΑΛΑΙΟ 1 ο

1.1. Αυτόματη Αναγνώριση Ήχων

Το ερευνητικό πεδίο της αυτόματης αναγνώρισης ήχων έχει ως στόχο την ανάλυση του περιβάλλοντα χώρου, την καταμέτρηση, τον διαχωρισμό και την αναγνώριση των ηχητικών πηγών, χρησιμοποιώντας μόνο το εισερχόμενο ακουστικό σήμα. (Νταλαμπίρας, 2010) Το πεδίο της αυτόματης αναγνώρισης του γενικευμένου ακουστικού σήματος, είναι τμήμα της Υπολογιστικής Ακουστικής Σκηνικής Ανάλυσης (Computational Auditory Scene Analysis – CASA), (Wang D., 2006).

. Εφαρμογές Αυτόματης Αναγνώρισης Ήχων

Ο τομέας της αυτόματης αναγνώρισης δημιουργήθηκε για να καλύψει διάφορες ανάγκες και έχει πολλαπλές εφαρμογές :

- εφαρμογή σε περιπτώσεις αδυναμίας του χρήστη να παρακολουθεί οθόνη όπως είναι στην περίπτωση των οδηγών, των πιλότων *(ex. Eurofighter Typhoon [1])*, των ιατρών και των χειρούργων *(ex. Dragon Medical Practice Edition [2]).*
- εφαρμογή σε χρήστες με προβλήματα όρασης. *(πχ.Nuance- TALKS and ZOOMS [3])*
- εφαρμογή σε συσκευές για την αποκατάσταση διαφορετικών διαταραχών της ομιλίας όπως είναι η κώφωση- βαρηκοΐα ή άλλες διαταραχές άρθρωσης και φώνησης *(π.χ. συσκευή Hawking, Intel, " ACAT ", 1997)*
- χρήση σε εφαρμογές οι οποίες περιέχουν αλληλεπίδραση μεταξύ ανθρώπου και μηχανής.
- διδασκαλία της προφοράς μιας ξένης γλώσσας *(πχ. Fluenz [4] , TeLL mRubidiume More [5])*
- συστήματα ασφαλείας τα οποία χρησιμοποιούν την αυτόματη αναγνώριση ήχων για να προσδιορίσει την ταυτότητα του ομιλητή από την ανθρώπινη φωνή *(πχ. Rubidium [6]).*
- εφαρμογή σε συστήματα αυτόματης τηλεφωνίας, τα οποία επιτρέπουν στους χρήστες να περιηγηθούν στο μενού χωρίς την χρήση πλήκτρων. *(πχ. cue-me [7])*
- εφαρμογές υπαγόρευσης κειμένου *(πχ. Speechnotes [8]).*
- φωνητικές οδηγίες για περιήγηση στα αρχεία και τους φακέλους ή άνοιγμα κάποιου λογισμικού ή εφαρμογής σε συσκευές όπως προσωπικοί υπολογιστές, tablets και κινητά τηλέφωνα *(πχ. SVoice [9], Cortana [10], Siri [11]).*
- στην ψυχαγωγία, σε εφαρμογές για αναγνώριση μουσικού κομματιού *(πχ. Shazam [12], TrackID [13])*

[1] https://www.eurofighter.com/
[2] http://www.nuance.com/products/dragon-medical-practice-edition/index.htm
[3] http://www.nuance.com/for-individuals/mobile-applications/talks-zooms/index.htm
[4] http://www.fluenz.com/
[5] http://www.effectivelanguagelearning.com/language-course-reviews/tell-me-more-review
[6] http://www.rubidium.com/
[7] http://www.openstream.com/cueme.html
[8] https://speechnotes.co/
[9] http://www.samsung.com/global/galaxy/what-is/s-voice/
[10] https://support.microsoft.com/en-us/help/17214/windows-10-what-is
[11] http://www.apple.com/ios/siri/
[12] http://www.shazam.com/
[13] https://trackid.sonymobile.com/

1.2. Επισκόπηση Διεθνούς Βιβλιογραφίας

Όσον αφορά τα ηχητικά δεδομένα υπάρχει πληθώρα ερευνών τα τελευταία χρόνια στην αυτόματη αναγνώριση ήχων που αφορούν τον εντοπισμό σημάτων ομιλίας (Speech Activity Detection) και κατηγοριοποίηση ως ομιλία-μη ομιλία (speech/non speech) [(J. Sohn, N. S. Kim, W. Sung, 1999) *;* (Y.D. Cho, A. Kondoz,, 2001)], αναγνώριση μουσικών ειδών πχ. Jazz, pop, rock, hip-hop, κλασσική μουσική κλπ (Mark Levy, Mark Sandler, 2008), αναγνώριση μουσικών οργάνων (Antti Eronen, Anssi Klapuri, 2000),αναγνώριση των νοτών σε μουσικά κομμάτια (Music Transcription) (D. FitzGerald, E. Coyle, B. Lawlor, 2002) και αναγνώριση περιβαλλοντολογικών ήχων και άλλων ηχητικών γεγονότων (Context Awareness) (B. Clarkson, N. Sawhney, A. Pentland, 1998).

Πιο συγκεκριμένα, όσον αφορά την περίπτωση των broadcast news και άλλων ραδιοφωνικών εκπομπών, που αποτελούν από μόνα τους μία κατηγορία επεξεργασίας ηχητικού σήματος, έχουν προηγηθεί διάφορες μελέτες για την αυτόματη επεξεργασία οι οποίες είναι σχετικές με την αναγνώριση ήχων *[* (T.Bouko, C. Nadeu, volume 2011), (E. Dogan, M.Sert, A. Yazici, 20-25 July 2009)] , την αναγνώριση ομιλητή (M.Kotti, V. Moschou, C. Kotropoulos, 2008),την αναγνώριση εκφωνητή-anchor (Delphine, 14-19 March 2010), την αναγνώριση ρόλου (role detection) [(B. Bigot, I. Ferrane, J. Pinquier, 2010) *;* (G. Damnati, D. Charlet, 2011) ; (W. Wang, S. Yaman, K. Precoda, C. Richey, 2011)], την αυτόματη αναγνώριση αλλαγής θέματος [(C. H. Wu, C. H. Hsieh, 2009) *;* (L. Xie, Y. Yang, Z. Q. Liu, W. Feng, Z. Liu, 2010)], την ποιότητα και το ύφος του ραδιοφωνικού σταθμού (Cettolo, 2000), την ανίχνευση ηχητικών συμβάντων (H.D. Tran, H. Li, 2011) , την αυτόματη τμηματοποίηση σε αθλητικές εκπομπές και την εξαγωγή των κορυφαίων στιγμιότυπων σε αθλητικά γεγονότα [(J. Huang, Y. Dong, J. Liu, C. Dong, H. Wang, 2009) *;* (M. Baillie, J.M. Jose, 2004) *;* (Y. Itoh, S. Sakaki, K. Kojima, M. Ishigame, 2008) *;* (J. Zhang, B. Jiang, L. Lu, Q. Zhao, 2010)] και την αναγνώριση και κατηγοριοποίηση μη-γλωσσικών ήχων μέσα σε ραδιοφωνικές εκπομπές [(Z. Sun, A. Purohit, K. Yang, N. Pattan, D. Siewiorek, A. Spailagic, I. Lane, P. Zhang, 2011) *;* (S. Petridis, M. Pantic, J.F. Cohn, 2011) *;* (T. Mikami, Y. Kojima, M. Yamamoto, M. Furukawa, 2012) *;* (S. Matos, S.S. Birring, I.D. Pavord, D.H. Evans, 2006) *;* (T. Drugman, J. Urbain, T. Dutoit, 2011)] . Επίσης, η αυτοματοποιημένη ανάλυση των ακουστικών σημάτων, χρησιμοποιείται για την αναγνώριση ομιλίας (I. Mporas, T. Ganchev, O. Kicsis, N. Fakotakis, 2011), την αναγνώριση συναισθήματος του ομιλητή (B. Schuller, G. Rigoll, M. Lang, May 2004), αναγνώριση της γλώσσας του ομιλητή (C.V. Wright, L. Ballard, F. Monrose, G.M. Manson, 2007), την κατηγοριοποίηση των χαρακτηριστικών της μουσικής [(T. Pohle, E. Pampalk, G. Widmer, 2005), (Casey, 2011)] , για την ανίχνευση των Jingles (J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, D. Caseiro, 2008) ή διαφημιστικών διαλειμμάτων (N. Liu, Y. Zhao, Z. Zhu, H. Lu, 2011), καθώς επίσης και για πολλές άλλες εφαρμογές στο ίδιο πεδίο.

1.3. Ανάλυση Ήχου (Speech / Music / Silence / Noise)

Για να φτάσουμε στην αυτόματη αναγνώριση του ήχου, το ηχητικό σήμα πρέπει να αποσυντεθεί σε διακριτές κατηγορίες που αντιπροσωπεύουν τα κύρια ηχητικά γεγονότα-(sound events). Στις ραδιοφωνικές εκπομπές οι κύριες κατηγορίες είναι η ομιλία και η μουσική. Υπάρχουν επίσης άλλες κατηγορίες όπως η σιωπή, ο θόρυβος ή άλλοι μη γλωσσικοί ήχοι που δημιουργούνται από την ηχογράφηση ή την μετάδοση του ακουστικού σήματος.

Ομιλία (Speech)

Η ομιλία αποτελεί το λεκτικό όργανο που χρησιμοποιούμε για την μετάδοση ενός μηνύματος, ώστε να επιτύχουμε την επικοινωνία.

Στο ραδιόφωνο, ως ομιλία, αξιολογούνται στοιχεία όπως οι φωνές των παρουσιαστών, τα λεκτικά μηνύματα από τηλεφωνικές επικοινωνίες με μέλη του κοινού, καθώς επίσης και συνεντεύξεις που πραγματοποιούνται και μεταδίδονται μέσω της ραδιοφωνικής εκπομπής.

Μουσική (Music)

Μια ιδιαίτερη κατηγορία ήχων, που αξιοποιεί η ραδιοφωνία, είναι η Μουσική.

Η μουσική στο ραδιόφωνο και την τηλεόραση, χρησιμοποιείται άλλοτε ως στοιχείο της αισθητικής απόλαυσης και άλλοτε ως βοηθητική λειτουργία , μόνη της ή σε συνδυασμό με ομιλία ή άλλους ήχους.

Ως αντικείμενο απόλαυσης, αποτελεί το βασικό θεμέλιο όλης της ραδιοφωνικής παραγωγής. Μερικά ραδιοφωνικά κανάλια ή ραδιοφωνικές εκπομπές μεταδίδουν αποκλειστικά μουσική. Το 1984, η μουσική στα τέσσερα δίκτυα του BBC, καταλάμβανε το 61,3% ολόκληρου του ραδιοφωνικού προγράμματος. (BBC, 1985)

Η μουσική εμφανίζεται στο ραδιόφωνο σε διάφορες μορφές.

1. Σήματα - Jingles/ Radio Jingles

Τα Jingles είναι συνήθως μικρής διάρκειας τραγούδια ή μουσικά κομμάτια για διαφημιστικούς ή εμπορικούς σκοπούς. (Jingles) Χρησιμοποιούνται για να εδραιώσουν το προϊόν και απομνημονεύονται εύκολα από τον ακροατή. Γενικότερα μπορούμε να πούμε πως τα Jingles είναι ένα είδος ηχητικού branding.

Στο ραδιόφωνο ο όρος «Jingles» χρησιμοποιείται για να δηλώσει τα μουσικά στοιχεία εκείνα που χαρακτηρίζουν τον εκάστοτε ραδιοφωνικό σταθμό ή ραδιοφωνική εκπομπή και τους δίνουν ταυτότητα. Τα Jingles πρέπει να συμφωνούν με το είδος της ραδιοφωνικής εκπομπής (ενημερωτική, ψυχαγωγική, ντοκιμαντέρ κλπ), το κοινό στο οποίο απευθύνεται, αλλά και στην φιλοσοφία του σταθμού και εμφανίζονται ως «μουσικά θέματα» που λειτουργούν ως υπογραφή της ραδιοφωνικής εκπομπής και σηματοδοτούν την αρχή και το τέλος της. Χρησιμοποιούνται περισσότερο στα «λαϊκά» ραδιοφωνικά δίκτυα , ενώ στα «ποιοτικά» χρησιμοποιείται για τον ίδιο σκοπό η φωνή του εκφωνητή, που αναγγέλλει την ροή του προγράμματος, δηλαδή ορίζει το ύφος του σταθμού. (Crisell, 1994). Τα ραδιοφωνικά jingles ηχογραφούνται σε στούντιο και χρησιμοποιούνται κατά την διάρκεια των διαλειμμάτων από τον εκάστοτε σταθμό ή ραδιοφωνική εκπομπή. Στην Ελλάδα, το πρώτο Radio Jingle δημιουργήθηκε την δεκαετία του ’70 από τον Χρήστο Λεοντή, και χρησιμοποιήθηκε σε διαφημιστικό σποτ για μία μάρκα τηλεόρασης. (Radio Jingles - Διαφημιστικά Τραγούδια)

2. Σύνδεσμος – «Γέφυρες»

Η μουσική εμφανίζεται ως σύνδεσμος μεταξύ δύο σκηνών ενός ραδιοφωνικού θεατρικού έργου ή στην μέση ενός προγράμματος. Σε αυτή την περίπτωση, η λειτουργία της μουσικής είναι να γεφυρώνει τις σκηνές του έρχου, είτε την αλλαγή των θεμάτων ενός ραδιοφωνικού προγράμματος, να αποφευχθούν οι σιωπές, καθώς επίσης και να δημιουργηθεί ένα είδος συνέχειας (Crisell, 1994) .

3. "Ατμοσφαιρική» Μουσική

Η «ατμοσφαιρική» μουσική χρησιμοποιείται ως μουσική υπόκρουση κατά την διάρκεια ενός ραδιοφωνικού θεατρικού έργου ώστε να δώσει έμφαση στα λεγόμενα των ηρώων. Οι ήρωες του έργου δεν την ακούν, όμως την ακούν οι ακροατές και έχει συμβολική σημασία. (Crisell, 1994) Υποδηλώνει συνήθως την συναισθηματική κατάσταση των ηρώων και είναι ένδειξη των συναισθημάτων και των σκέψεων τους .

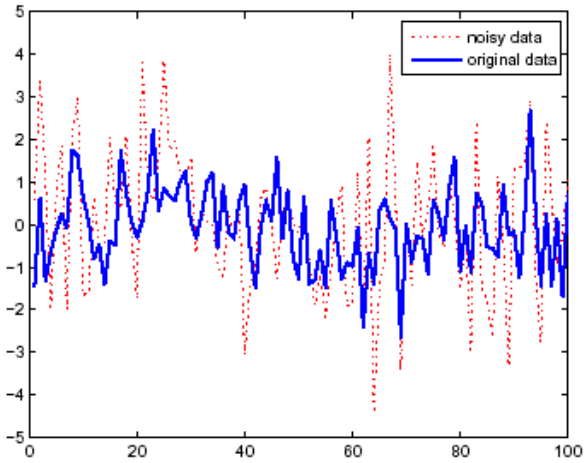
Γενικότερα μπορούμε να πούμε πως η μουσική στο ραδιόφωνο έχει κυρίαρχο ρόλο, (Crisell, 1994) διότι λόγω της τυφλότητας του μέσου, είναι εξαιρετικά χρήσιμο βοήθημα για να δηλώσουμε το ύφος ενός προγράμματος, ενώ ταυτόχρονα είναι υποβλητική και λιγότερο ψυχρή από τις απλές ανακοινώσεις και αναγγελίες.

Θόρυβος (Noise)

Ως θόρυβος χαρακτηρίζεται κάθε ανεπιθύμητος ήχος, η τυχαία αλλοίωση των τιμών ενός ακουστικού σήματος, είτε η παρείσφρηση τυχαίων αντικειμένων. Η έννοια του θορύβου είναι, σε πολλές περιπτώσεις υποκειμενική καθότι, ανάλογα την περίπτωση, κάποιος ήχος μπορεί να θεωρηθεί ως θόρυβος και άλλοτε ως επιθυμητός.

Τα είδη του θορύβου ποικίλουν και μπορεί να προέλθουν από διάφορες πηγές. Στους ραδιοφωνικούς δέκτες, ένα παράδειγμα θορύβου είναι αυτό που εμφανίζεται προκαλώντας το χαρακτηριστικό hiss στα ηχεία του ραδιοφώνου. Επίσης, στις ραδιοφωνικές εκπομπές και κυρίως στις συνεντεύξεις που πραγματοποιούνται στους εξωτερικούς χώρους, λόγω αδυναμίας του ατόμου που δίνει την συνέντευξη να παρευρεθεί στο στούντιο, είναι αναπόφευκτη η παρουσία περιβαλλοντολογικών θορύβων όπως είναι ο ήχος από τα αυτοκίνητα, φωνές πλήθους, ήχοι από διάφορες πηγές στο περιβάλλον , ακόμη και ο ατμοσφαιρικός αέρας. Στο ραδιοφωνικό στούντιο, η ηχομόνωση εξουδετερώνει όλους τους περιβαλλοντολογικούς θορύβους.

Είναι προφανές λοιπόν, ότι πολλά λάθη μπορεί να προκληθούν από την παρουσία θορύβου στο ακουστικό σήμα, αφού μπορεί να εμποδίσει την ομαλή αναγνώριση του σήματος και την καταχώρηση του σε μία κλάση. Για αυτόν τον λόγο, συνήθως, η καταστολή του θορύβου είναι αναγκαία, καθώς αποτελεί έναν από τους πιο καθοριστικούς παράγοντες για την υποβάθμιση του ακουστικού σήματος. Αυτό το επιτυγχάνουμε κανονικοποιώντας (normalize) το ακουστικό σήμα που θέλουμε να μελετήσουμε.



Εικόνα 1.1 Αναπαράσταση σήματος με θόρυβο και χωρίς θόρυβο

Η αλλοίωση του ακουστικού σήματος από τον θόρυβο μετράται από το λόγο του σήματος προς τον θόρυβο (Signal to Noise Ratio - SNR). Αυτό το μέτρο δηλώνει πόσο δυνατό είναι ένα σήμα σε σχέση με τον θόρυβο που αυτό περιέχει (signal-to-noise ratio S/N or SNR) και υπολογίζεται από την εξίσωση:

$$SNR = 20log_{10}(V_s/V_n) \qquad (1.1)$$

όπου V_s δηλώνει την δύναμη του ακουστικού σήματος που αναλύουμε σε microvolts, και V_n την δύναμη του θορύβου που περιέχεται στο αντίστοιχο σήμα, επίσης δοσμένο σε microvolts. Το αποτέλεσμα μας δίνει το signal-to-noise ratio με μονάδα μέτρησης το decibel (db).

Σιωπή (Silence)

Η σιωπή είναι η κατάσταση η οποία επικρατεί απουσία οποιουδήποτε ήχου, συμπεριλαμβανομένου της ομιλίας, της μουσικής και του θορύβου.

Στο ραδιόφωνο οι σιωπές (silences) ακούγονται ως παύσεις της ομιλίας ή της μουσικής , προκειμένου να μπορέσει ο ακροατής να οπτικοποιήσει τις καταστάσεις στο ραδιοφωνικό στούντιο, από όπου μεταδίδεται το ραδιοφωνικό πρόγραμμα. Επίσης, ως σιωπή ακούγεται και ο «νεκρός αέρας» που υποδηλώνει την αρχή και το τέλος των ραδιοφωνικών προγραμμάτων. Στην παρούσα εργασία θεωρήθηκαν ως σιωπές οι μικρές παύσεις των εκφωνητών κατά την ομιλία, τα άηχα σύμφωνα κατά την έναρξη της ομιλίας, καθώς επίσης και οι ανάσες.

ΚΕΦΑΛΑΙΟ 2^{ο} – Θεωρητικό Υπόβαθρο

2.1. Τα χαρακτηριστικά του ήχου

2.1.1. Διαισθητικά Χαρακτηριστικά του Ήχου

Τα χαρακτηριστικά του ήχου και με τα οποία μπορούμε να περιγράψουμε τα ακουστικά σήματα, χωρίζονται σε δύο κατηγορίες (Τσινίκας, 2005), τα φυσικά χαρακτηριστικά, τα οποία είναι εκείνα τα οποία περιγράφουν με μαθηματικό τρόπο τα ακουστικά σήματα, όπως είναι η συχνότητα, τα φασματικά χαρακτηριστικά και το πλάτος, και τα διαισθητικά χαρακτηριστικά, τα οποία είναι βασισμένα στο πώς αντιλαμβάνεται τους ήχους το ανθρώπινο αυτί και είναι πολύ δύσκολο να μετρηθούν με συμβατικές μεθόδους ή όργανα, καθότι βασίζονται κυρίως στην ανθρώπινη αντίληψη του ήχου. Για την μέτρηση των διαισθητικών χαρακτηριστικών, συνήθως χρησιμοποιούνται τεστ από ομάδες ακροατών, με αντίστοιχες γνώσεις, που προσδιορίζουν κατά προσέγγιση αυτά τα χαρακτηριστικά.

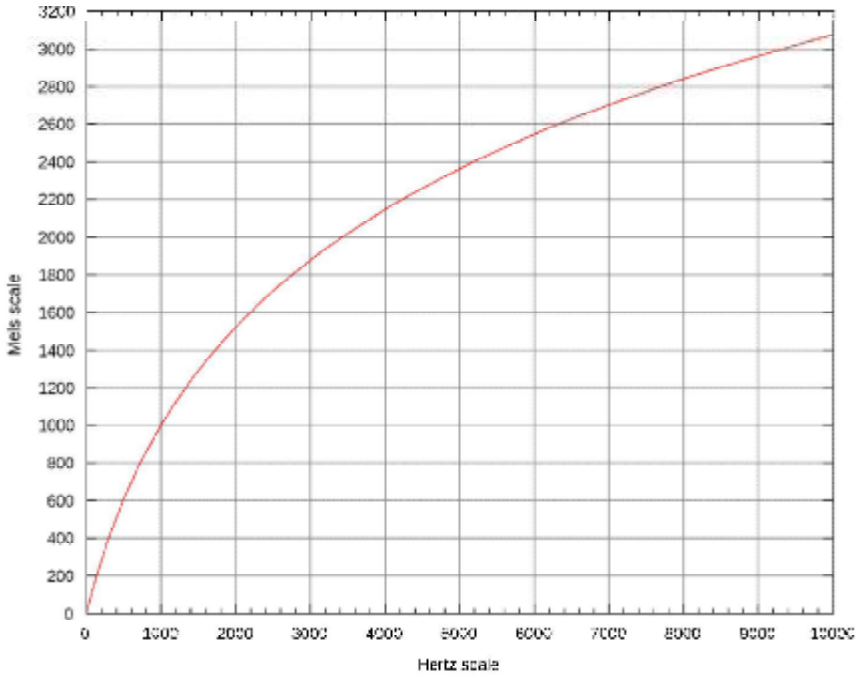
Ακουστότητα: Η ακουστότητα, ή αλλιώς ένταση είναι ένα χαρακτηριστικό που ορίζει πόσο δυνατός είναι ένας ήχος. Η ακουστότητα είναι ένα υποκειμενικό μέγεθος και για την μέτρηση του απαιτείται να ορίσουμε ένα σύστημα αναφοράς , το οποίο ορίστηκε από τους *Flecher & Munson* το 1933, και είναι η συχνότητα 1000Hz. Παράλληλα, απέδειξαν την διαφοροποίηση μεταξύ συχνότητας και έντασης, δηλαδή, ότι εάν δύο ήχοι έχουν την ίδια ένταση αλλά έχουν διαφορετική συχνότητα, το ανθρώπινο αυτί θα ξεχωρίσει αυτόν με την υψηλότερη συχνότητα. Ως μονάδα μέτρησης της στάθμης της ακουστότητας (loudness level - LL) ορίσθηκε από τον Barkhausen το 1926, και είναι το Phon.

Τονικό ύψος: Το τονικό ύψος είναι ένα ακόμη διαισθητικό χαρακτηριστικό των ηχητικών σημάτων και είναι αυτό που διαχωρίζει τους ήχους σε χαμηλούς και υψηλούς. Το τονικό

ύψος εξαρτάται, κατά κύριο λόγο, από την συχνότητα (f), αλλά και από την στάθμη της ηχητικής πίεσης, και έχει ως μονάδα μέτρησης το Mel. Η σχέση μεταξύ της συχνότητας (Hz) και της κλίμακας Mel καθορίστηκε από το πείραμα που διεξήγαν οι *Stevens, Volkmann & Newman*. Στην εικόνα 2.1. παρουσιάζεται η λογαριθμική αντιστοιχία μεταξύ της συχνότητας Hz και της διαισθανόμενης κλίμακας Mel όπως προέκυψε από το πείραμα:

$$mel(f) = 2595log_{10}\left(1 + \frac{f}{700}\right) \qquad (2.1)$$

όπου f η συχνότητα.



Εικόνα 2.1. Σχέση συχνότητας – κλίμακας Mel

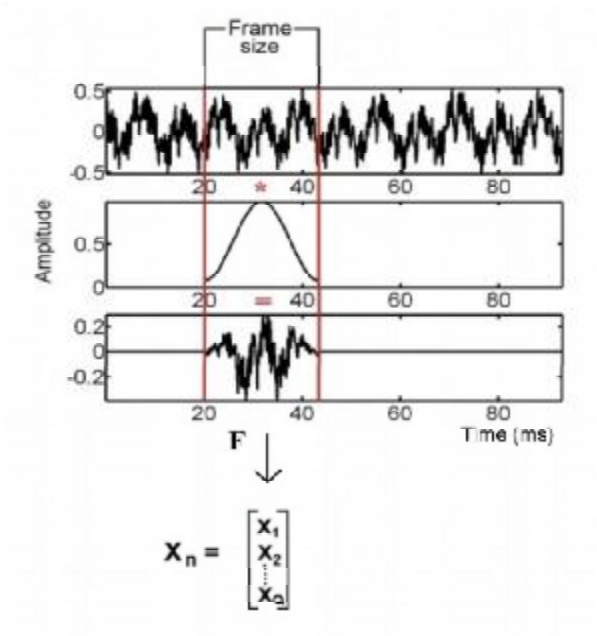
Χροιά: Η χροιά είναι ένα χαρακτηριστικό που κάνει δύο ήχους που έχουν την ίδια ένταση (intensity) και την ίδια ακουστότητα (loudness) να διαφέρουν. Η διαφορετική κατανομή της ηχητικής ενέργειας είναι αυτό που προκαλεί την διαφορετικότητα στην χροιά. Για να περιγραφεί η χροιά χρησιμοποιούνται φασματικά διαστήματα μικρής διάρκειας, περίπου 10s – 40s. (D. Howard, 2009)

2.1.2. Τα χαρακτηριστικά του φάσματος (spectral features - spfe)

Τα φασματικά χαρακτηριστικά υπολογίζονται μέσω του ταχύ μετασχηματισμού Fourier (Fast Fourier Transform - FFT) του ακουστικού σήματος (Peeters, 2004) σε αλληλεπικαλυπτόμενες περιόδους των 20 ms με επικάλυψη 10 ms (50%). Το διάνυσμα χαρακτηριστικών $x_{n,}$ για την χρονική στιγμή n, υπολογίζεται από την εξίσωση

$$x_n = F(wS_n - (N-1), \ldots, w_{N-1}S_n) \qquad (2.2)$$

όπου S_n είναι το ακουστικό σήμα, w η συνάρτηση του παραθύρου και N το μέγεθος του διαστήματος που αναλύουμε (frame)



Εικόνα 2.2. Διαδικασία Εξαγωγής Φασματικών Χαρακτηριστικών από ένα ακουστικό σήμα

Φασματικό κέντρο βάρους (Spectral Centroid - SC)

Αυτό το χαρακτηριστικό είναι το σημείο ισορροπίας του φασματικού πλάτους του τμηματικού μετασχηματισμού Fourier (STFT) και μας δείχνει σε ποιο σημείο του φάσματος είναι συγκεντρωμένη η περισσότερη ενέργεια. Το φασματικό κέντρο βάρους υπολογίζεται από την εξίσωση:

$$SC_t = \frac{\sum_{n=1}^{N} M_t\,[\mathrm{n}]\,n}{\sum_{n=1}^{N} M_t\,[\mathrm{n}]} \qquad (2.3)$$

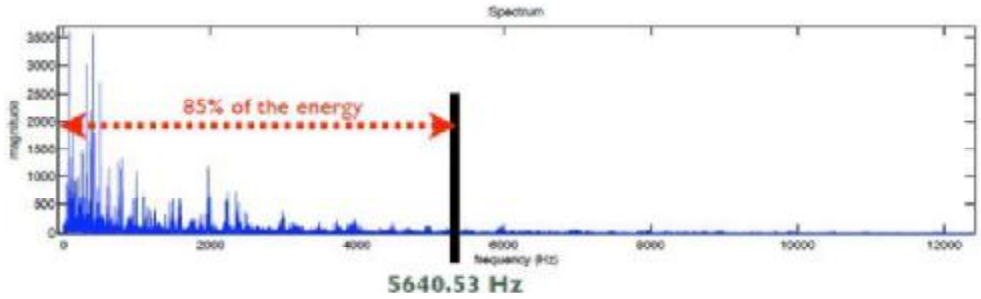
όπου M_t [n], είναι η τιμή του φάσματος μετά τον μετασχηματισμό Fourier στο διάστημα χρόνου t και στην τιμή συχνότητας n. Καθότι είναι άμεσα συνδεδεμένο με την συχνότητα του σήματος, οι χαμηλές τιμές αυτού του χαρακτηριστικού υποδηλώνουν σημαντικές συνιστώσες με χαμηλή συχνότητα και αμελητέες συνιστώσες με υψηλή συχνότητα και επομένως χαμηλή φωτεινότητα (brightness) του ήχου. Η φωτεινότητα είναι άμεσα συνδεδεμένη με την αίσθηση οξύτητας του ήχου αφού πιο υψηλή φωτεινότητα έχουν τα ηχητικά σήματα στα οποία κυριαρχούν υψηλές συχνότητες.

Φασματικό Roll-off (Spectral Roll-off)

Το φασματικό Roll-off αποτελεί ένα φασματικό χαρακτηριστικό που δηλώνει την κατανομή της ενέργειας στις τονικά χαμηλές συχνότητες. Η εξίσωση με την οποία υπολογίζουμε το φασματικό Roll-off είναι:

$$\sum_{n-1}^{Rt} M_t[n] = 0.85 \sum_{n-1}^{N} M_t[n] \qquad (2.4)$$

όπου το R_t αντιπροσωπεύει την συχνότητα κάτω από την οποία βρίσκεται το 85% της συνολικής ενέργειας στις χαμηλές συχνότητες. (McKay, 2005)



Εικόνα 2.3. Σημείο Φασματικού Roll-off

Φασματική Ροή (Spectral Flux – SF)

Η φασματική ροή, που εκτιμάει την ποσότητα της γενικής φασματικής αλλαγής σε ένα σήμα, υπολογίζεται από την εξίσωση:

$$SF_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \qquad (2.5)$$

όπου $N_t[n]$ και $N_{t-1}[n]$ είναι τα μεγέθη του μέτρου για τον τμηματικό μετασχηματισμό Fourier (STFT) στο χρονικό διάστημα t.

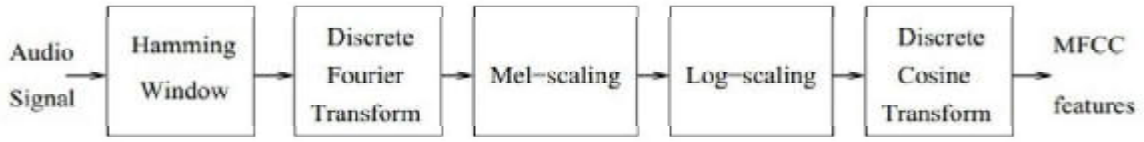
2.1.3. Συντελεστές Mel Frequency Cepstral (Mel-frequency Cepstral Coefficients - MFCC)

Τα Mel-frequency cepstral coefficients χρησιμοποιούνται ευρέως στην αναγνώριση της ομιλίας, αλλά έχουν αποδειχτεί και αρκετά αποτελεσματικά στις εφαρμογές εξόρυξης μουσικής πληροφορίας. Τα MFCC μας δίνουν γενικότερα πληροφορίες για το φασματικό

περιεχόμενο και την φασματική περιβάλλουσα ενός σήματος μικρού χρόνου. Πρόκειται για τον διακριτό μετασχηματισμό συνημίτονου (DCT) του λογαριθμικού φάσματος, μετά από μια μη γραμμική στρέβλωση συχνοτήτων, σε μία αντιληπτή κλίμακα που ονομάζεται Mel. (Ζαρβαδάς, 2013)

Πιο συγκεκριμένα, η διαδικασία περιλαμβάνει την τμηματοποίηση του σήματος σε κομμάτια των 20 - 30 msec τα οποία είναι αλληλεπικαλυπτόμενα με επικάλυψη 50% – 75% και εφαρμόζεται σε αυτά το παράθυρο Hamming. (A.V. Oppenheim, R.W. Schafer, 1999), ώστε να εξομαλυνθούν τα άκρα του ακουστικού σήματος. Στην συνέχεια χρησιμοποιείται ο διακριτός μετασχηματισμός Fourier (Discrete Fourier Transform – DFT), ώστε να υπολογιστεί το φάσμα ισχύος του κάθε τμήματος, και στο οποίο φάσμα απεικονίζεται η κάθε συχνότητα Hz στην κλίμακα Mel, χρησιμοποιώντας μια συστοιχία φίλτρων (filterbanks). Επειδή όμως το ανθρώπινο σύστημα ακοής αντιλαμβάνεται λογαριθμικά την ένταση ενός ηχητικού σήματος, υπολογίζεται ο λογάριθμος των ενεργειών όλων των φίλτρων σε κάθε μία από τις Mel συχνότητες και η διαδικασία ολοκληρώνεται, εφαρμόζοντας ανάστροφο Διακριτό Μετασχηματισμό Συνημίτονου (Discrete Cosine Transform – DCT).

Η διαδικασία εξαγωγής των χαρακτηριστικών MFCC παρουσιάζεται σχηματικά στην εικόνα 2.4.



Εικόνα 2.4. Διαδικασία εξαγωγής MFCC χαρακτηριστικών από ένα ηχητικό σήμα

2.1.4. Πλήθος Μηδενισμού Συνάρτησης (Zero – crossing Rate)

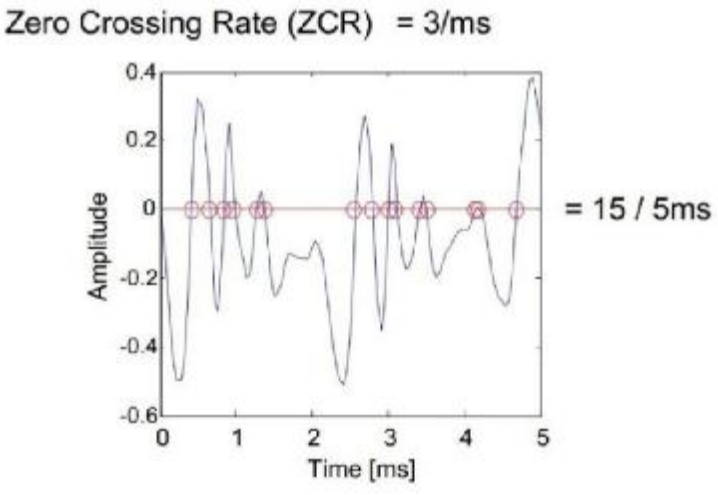
Το πλήθος μηδενισμού συνάρτησης είναι ένα χαρακτηριστικό το οποίο χρησιμοποιείται στα συστήματα αυτόματης αναγνώρισης και κατηγοριοποίησης (T. Lambrou, P. Kudumakis, M. Sandler, R. Speller & A. Linney, 1998) και είναι ουσιαστικά το πλήθος των σημείων στα οποία το σήμα μηδενίζει, δηλαδή έχει διαφορετικό πρόσημο για δύο διαδοχικά στιγμιότυπα (samples). Στα απλά τονικά φάσματα, οι μηδενισμοί μπορούν να αποτελέσουν ένα μέτρο που περιγράφει την θεμελιώδη συχνότητα f_0 του σήματος, (Θεοδωρίδης, 2011), αφού μπορούμε να την υπολογίσουμε διαιρώντας τον αριθμό μηδενισμών με το 2, διότι σε κάθε κύκλο το σήμα μας τέμνει δύο φορές τον άξονα του χρόνου.

Ο ρυθμός μηδενισμών (zero – crossing rate ή ZCR) είναι σημαντικό χαρακτηριστικό για την ακουστική, διότι είναι ένα χαρακτηριστικό που εξαρτάται από τον χρόνο, και επομένως είναι πολύ εύκολο να υπολογιστεί. Στην παρούσα εργασία, επιλέχθηκε αυτό το χαρακτηριστικό διότι χρησιμοποιείται συχνά για τον διαχωρισμό της ομιλίας και της μουσικής αφού όπως έχει παρατηρηθεί σε πειράματα, το ZCR ενός τμήματος μουσικής είναι μεγαλύτερο από το ZCR ενός σήματος ομιλίας, ενώ μας βοηθά να ξεχωρίζουμε και την ύπαρξη ή μη της ομιλίας μέσα στο ακουστικό σήμα, εφόσον στην πρώτη περίπτωση το ZCR είναι πιο χαμηλό από ότι στην δεύτερη περίπτωση.

Ο ρυθμός μηδενισμών υπολογίζεται από την εξίσωση:

$$ZCR_i = \frac{1}{N}\sum_{n=0}^{N-1}|sgn\{x_i(n)\} - sgn\{x_i(n-1)\}| \quad (2.6)$$

όπου N το μήκος του τμήματος ήχου, 1/N υπολογίζει το ποσοστό των μηδενισμών και η μεταβλητή sgn( ) επιστρέφει το πρόσημο της εισόδου.



Εικόνα 2.5. Πλάτος του σήματος με Zero - crossing τιμές

2.1.5. Ενέργεια Βραχέως Χρόνου (Short-time Energy feature)

Η ενέργεια βραχέως χρόνου χρησιμοποιείται ως χαρακτηριστικό συνήθως για τον διαχωρισμό της ύπαρξης του ήχου και της σιωπής. Υπολογίζεται από την εξίσωση :

$$STE_n = \frac{1}{N}\sum_{i=n-N+1}^{n} s_i^2 \quad (2.7)$$

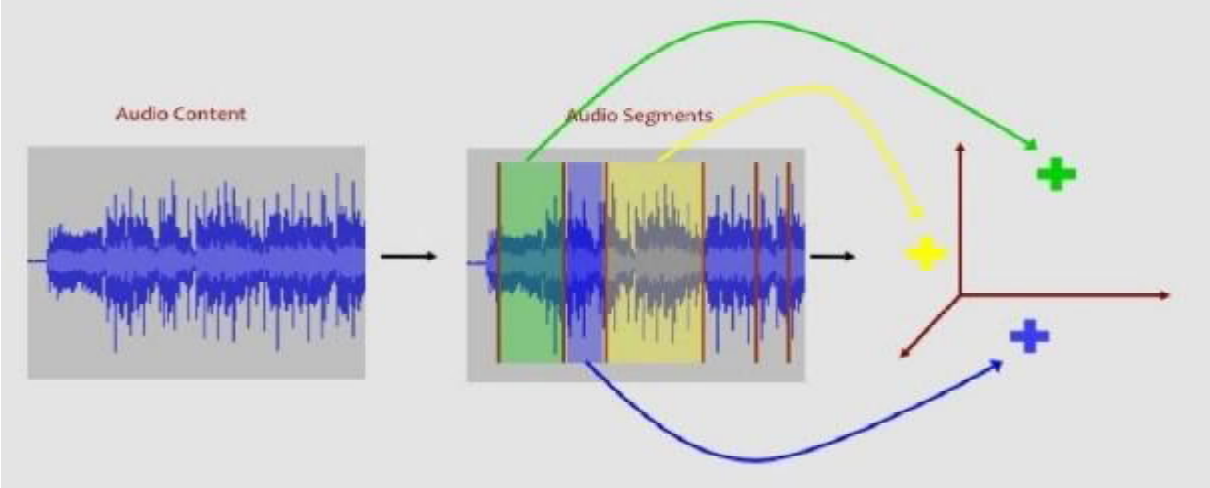
όπου s_i το σήμα στον χρόνο i.

Η ακουστότητα (Loudness) του ήχου είναι συνδεδεμένη με την ένταση (intensity) και με την ενέργεια βραχέως χρόνου του ήχου. (Ζαρβαδάς, 2013) Συνήθως το χαρακτηριστικό της ενέργειας βραχέως χρόνου χρησιμοποιείται ως κριτήριο για τον διαχωρισμό των τμημάτων ενός ακουστικού σήματος σε ομιλία και θόρυβο, αφού στην περίπτωση του θορύβου το STE είναι πολύ μικρότερο από ότι στην περίπτωση της ομιλίας.

2.2. Αλγόριθμοι Κατηγοριοποίησης (Classification Algorithms)

Ο κλάδος της επιστήμης των υπολογιστών που ασχολείται με την αυτόματη κατηγοριοποίηση έχει αναπτυχθεί σημαντικά τα τελευταία χρόνια και έχει κεντρίσει σημαντικά το ενδιαφέρον των ερευνητών.

Στην εικόνα 2.6. αναπαριστάται η διαδικασία της αυτόματης ταξινόμησης ακουστικών σημάτων προσεγγίζοντας τον τρόπο που αυτά κατηγοριοποιούνται από το ανθρώπινο σύστημα ακοής.



Εικόνα 2.6. Διαδικασία Αυτόματης Κατηγοριοποίησης Ακουστικών Σημάτων

Οι λειτουργίες της αναγνώρισης στην αυτόματη κατηγοριοποίηση επιτυγχάνονται χρησιμοποιώντας αλγορίθμους ταξινόμησης (classifiers) οι οποίοι αναγνωρίζουν τα χαρακτηριστικά που εξάγονται από τα δείγματα των ακουστικών σημάτων.

Μερικοί από τους βασικούς αλγορίθμους ταξινόμησης παρουσιάζονται παρακάτω.

ZeroR: Αυτός ο αλγόριθμος αποτελεί την απλούστερη μορφή αλγορίθμου κατηγοριοποίησης και βασίζεται αποκλειστικά στα δεδομένα εισόδου, αγνοεί όλα τα predictors, και συνεπώς δεν έχει καμία δυνατότητα πρόβλεψης. (ZeroR) Ουσιαστικά προβλέπει τα πάντα ως την επικρατέστερη κλάση. Μπορεί να χρησιμοποιηθεί για να κατηγοριοποιήσει την μέγιστη τιμή των κλάσεων, όπως επίσης και για να καθορίσει το ελάχιστο ποσοστό επιτυχίας στην ταξινόμηση σε σύγκριση με άλλες μεθόδους.

Πιο συγκεκριμένα, μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο ZeroR για να καθορίσουμε μία βάση. Για παράδειγμα, εάν έχουμε 10 κλάσεις ίδιου αριθμού δεδομένων στην καθεμία, τότε ο αλγόριθμος θα δώσει ποσοστό επιτυχίας 10%, και επομένως, εάν με έναν άλλο αλγόριθμο πάρουμε 40% τότε αυτό είναι ένα καλό ποσοστό επιτυχίας κατηγοριοποίησης. Αντίθετα, εάν έχουμε αντίστοιχα 2 κλάσεις ίδιου αριθμού δεδομένων, τότε η βάση μας θα είναι το 50% και συνεπώς παίρνοντας ένα ποσοστό επιτυχίας του ύψους 40% με άλλον αλγόριθμο, είναι κάτω το ορίου και χαρακτηρίζεται ως χαμηλό, και ο αλγόριθμος ως μη αποτελεσματικός.

Naive Bayes: Ο αλγόριθμος Μπεϊσιανής κατηγοριοποίησης ή αλλιώς αλγόριθμος Naive Bayes βασίζεται την θεωρία των πιθανοτήτων και προκύπτει από το θεώρημα του Bayes τον οποίο παριστάνεται με την εξίσωση

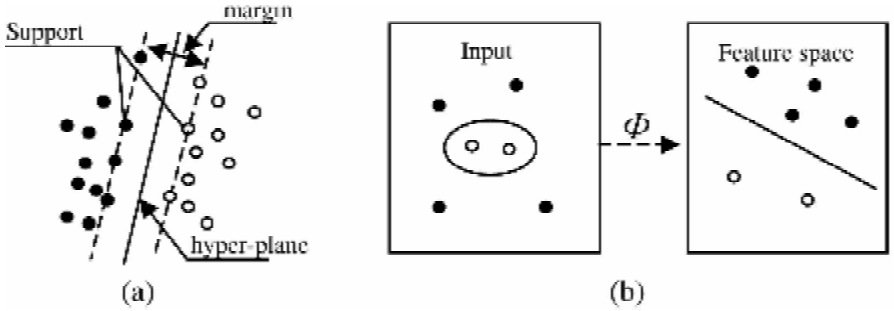
$$P(\omega_j|x) = \frac{P(x|\omega_j) * P(\omega_j)}{P(x)} \qquad (2.8)$$

όπου ωj η πιθανότητα (probability) της κλάσης j και x το διάνυσμα των χαρακτηριστικών (feature vector) και υπολογίζει την πιθανότητα κάθε κατηγορίας της κατηγοριοποίησης. Προϋπόθεση για να είναι ο αλγόριθμος αποτελεσματικός είναι πως τα χαρακτηριστικά – μεταβλητές δοσμένης μιας κλάσης, θα πρέπει να είναι ανεξάρτητα μεταξύ τους. Αυτό ονομάζεται conditional independence.

Αναλύοντας τις παραμέτρους στον αλγόριθμο του Bayes, το P(x|ω_j) συμβολίζει την πιθανότητα του διανύσματος χαρακτηριστικών x δοσμένης της κλάσης j και ονομάζεται likelihood , $P(\omega_j)$ την αρχική πιθανότητα (prior probability) της κλάσης j και P(x) την αρχική πιθανότητα (prior probability) του διανύσματος χαρακτηριστικών x .

Από προηγούμενα πειράματα που έχουν πραγματοποιηθεί, έχει προκύψει ότι ο αλγόριθμος Naive Bayes, παρότι είναι αρκετά απλός στην δομή και την χρήση του, είναι ιδιαίτερα αποτελεσματικός για πολυσύνθετα προβλήματα κατηγοριοποίησης. Έχει αποδειχτεί, επίσης, πως είναι ιδιαίτερα χρήσιμος και αποτελεσματικός στην αυτόματη αναγνώριση ομιλίας. (Π.Ζερβάς, 2007)

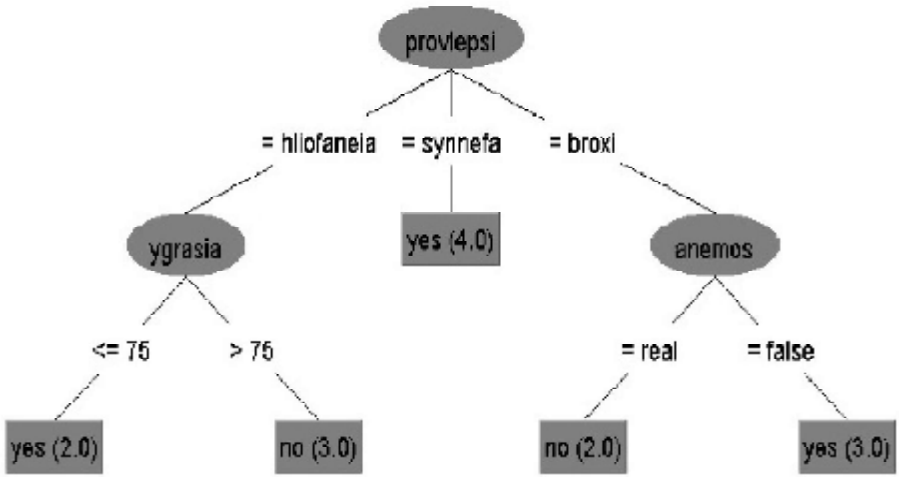
Αλγόριθμος μηχανών διανυσμάτων υποστήριξης (SVM): Τα SVM (Support Vector Machines – Μηχανές Διανυσμάτων Υποστήριξης) [(C. Cortes, V. Vapnik, 1995) *;* (Vapnik, 1998)] θεωρούνται κατά πολλούς η πιο επιτυχημένη μέθοδος κατηγοριοποίησης. Αυτός ο αλγόριθμος, ελαχιστοποιεί την πιθανότητα λανθασμένης ταξινόμησης, καθότι επιλέγοντας στιγμιότυπα που συνορεύουν στον χώρο του προβλήματος, δηλαδή έχουν παρόμοια ή κοινά χαρακτηριστικά, βρίσκει την βέλτιστη διαχωριστική γραμμή ανάμεσα σε ένα σύνολο δεδομένων, και μεγιστοποιεί την απόσταση μεταξύ των διακριτών χαρακτηριστικών τους. Ουσιαστικά, ο αλγόριθμος εκπαιδεύεται να αναγνωρίζει το σημείο που μεταβαίνουμε από την μία κλάση στην άλλη, μεγιστοποιώντας την απόσταση μεταξύ των δύο κλάσεων, χρησιμοποιώντας γραμμικούς διαχωριστές στον χώρο των παραμέτρων. Ένας αλγόριθμος μηχανών διανυσμάτων υποστήριξης μπορεί να είναι είτε γραμμικός (linear support vector machines) είτε μη-γραμμικός (non-linear support vector machines). (Εικόνα 2.7.)



Εικόνα 2.7. (a) γραμμικός αλγόριθμος μηχανών διανυσμάτων υποστήριξης
(b) μη-γραμμικός αλγόριθμος μηχανών διανυσμάτων υποστήριξης

Η μέθοδος Ακολουθιακής Ελάχιστης Βελτιστοποίησης (Sequential Minimal Optimization – SMO), η οποία χρησιμοποιείται στην παρούσα εργασία, έγινε γνωστή από τον Platt το 1999.

C4.5: Αυτός ο αλγόριθμος ανήκει στην κατηγορία των αλγορίθμων ταξινόμησης που χρησιμοποιούν δέντρα απόφασης. Τα δέντρα απόφασης αποτελούνται από κόμβους που αντιστοιχούν σε κάποιον συγκεκριμένο χαρακτηριστικό και διακλαδώσεις εξερχόμενες των κόμβων, οι οποίες αντιστοιχούν σε μία συνθήκη σύμβασης, γνωστή και ως συνθήκη διάσπασης (split criteria) και οδηγούν σε επόμενους κόμβους. Ανάλογα με το αν τηρείται ή όχι αυτή η σύμβαση, η διαδικασία συνεχίζεται στον αντίστοιχο κόμβο, έως ότου όλα τα στιγμιότυπα να ανήκουν σε μία κλάση. Έτσι δομείται το δέντρο απόφασης όπως φαίνεται στην παρακάτω εικόνα. (Εικόνα 2.8.)

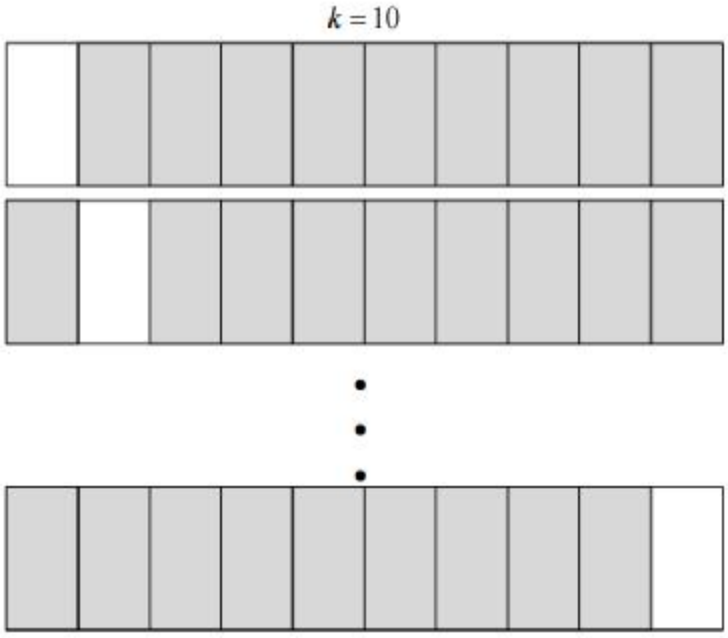


Εικόνα 2.8. Δομή δέντρου απόφασης μέσω του αλγορίθμου C4.5

Στρωματοποιημένη διασταυρωμένη επικύρωση (Cross Validation)

Για την διαδικασία της κατηγοριοποίησης συνηθίζεται να έχουμε δύο βάσεις δεδομένων (datasets). Η πρώτη είναι ένα σύνολο δεδομένων, και χρησιμοποιείται για την εκπαίδευση των αλγορίθμων μάθησης (training dataset), ενώ υπάρχει και μία δεύτερη βάση δεδομένων που χρησιμοποιείται για τον έλεγχο και την αξιολόγηση των αλγορίθμων (testing dataset). Στην παρούσα εργασία, χρησιμοποιήθηκε μία άλλη μέθοδος αξιολόγησης και επικύρωσης των αλγορίθμων μάθησης, η οποία ονομάζεται k – fold Cross Validation, ή αλλιώς Rotation Estimation. Αυτή η μέθοδος χρησιμοποιεί μία μόνο βάση δεδομένων. Αρχικά χωρίζει το σύνολο των δεδομένων, σε δύο υποσύνολα από τα οποία το ένα είναι για εκπαίδευση του αλγορίθμου και το δεύτερο για την αξιολόγηση του. Πιο συγκεκριμένα, η μέθοδος k – fold Cross Validation, διαιρεί τα δεδομένα μας σε k τμήματα ίσου μεγέθους, από τα οποία σε κάθε μία από τις k επαναλήψεις που θα πραγματοποιηθούν τα k-1 τμήματα θα χρησιμοποιούνται για την εκπαίδευση, ενώ το ένα από αυτά θα χρησιμοποιείται για την αξιολόγηση του αλγορίθμου. Σε κάθε μία από τις k επαναλήψεις χρησιμοποιείται διαφορετικό τμήμα για την

αξιολόγηση και το τελικό αποτέλεσμα είναι ο μέσος όρο των αποτελεσμάτων όλων των τμημάτων. (Κωτσιαντής, 2005)



Εικόνα 2.9. k-fold Cross Validation για k=10

Στην παρούσα εργασία, χρησιμοποιείται η τεχνική της στρωματοποιημένης διασταυρωμένης επικύρωσης δέκα σημείων (ten - fold cross validation), έτσι ώστε να προσεγγιστεί η πραγματική απόδοση των αλγορίθμων, όσο το δυνατόν καλύτερα.

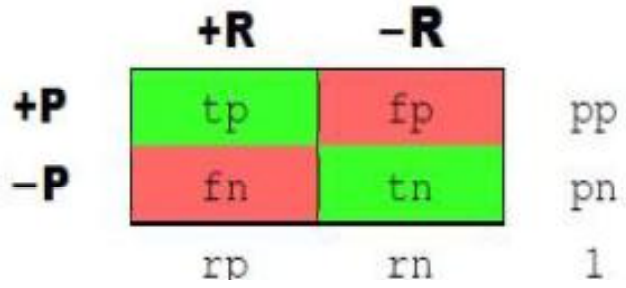
2.3. Αξιολόγηση αλγορίθμων και αποτελεσμάτων

Ο ποιότητα και ο έλεγχος των δεδομένων πραγματοποιείται μέσω των αποτελεσμάτων της κατηγοριοποίησης, και πιο συγκεκριμένα με τον υπολογισμό και την εκτίμηση των ποσοστών επιτυχίας και εσφαλμένων προβλέψεων. Παρακάτω αναφέρονται τα μέτρα με τα οποία μετρήθηκαν και αξιολογήθηκαν τα αποτελέσματα των αλγορίθμων κατηγοριοποίησης στην παρούσα εργασία.

2.3.1. Μέτρα Αξιολόγησης και αποτελεσμάτων (Evaluation Metrics)

Confusion Matrix: Στην παρούσα εργασία, η επίδοση των διαφόρων αλγορίθμων ταξινόμησης περιγράφηκε από πίνακες που ονομάζονται confusion matrix. Τα στοιχεία που περιέχονται στο confusion matrix μας δείχνουν ποια από τα στοιχεία ταξινομήθηκαν σωστά, και ποιά λανθασμένα. Πιο συγκεκριμένα το όνομα προκύπτει από το γεγονός πως από τα στοιχεία που μας δίνει μπορούμε να συμπεράνουμε πότε ένας αλγόριθμος ταξινομεί τα στοιχεία μιας κλάσης σε μιαν άλλη, δηλαδή συγχέει τις κλάσεις μεταξύ τους. Αρχικά θα περιγράψουμε πως λειτουργεί ένας confusion matrix και στην συνέχεια θα αναλύσουμε τα μέτρα αξιολόγησης που αυτός περιέχει.

Εάν μία παρατήρηση είναι θετική και ταξινομηθεί ως θετική τότε προσμετρείται ως *πραγματικά θετική παρατήρηση (tp – true positive)* ενώ εάν η ίδια παρατήρηση ταξινομηθεί εσφαλμένα ως αρνητικό, τότε προσμετρείται ως *λανθασμένα αρνητικό (fn – false negative).* Αντίστοιχα, εν μία παρατήρηση είναι αρνητική και ταξινομηθεί ως αρνητική τότε θεωρούμε πως ανήκει στις *πραγματικά αρνητικές (tn – true negative)* παρατηρήσεις, ενώ εάν ταξινομηθεί ως θετικό τότε υπολογίζεται ως *λανθασμένα θετικό (fp – false positive).*



Εικόνα 2.10. Confusion Matrix

Στην εικόνα 11 απεικονίζεται ένα ενδεικτικό confusion matrix. Οι στήλες +R και –R αντιπροσωπεύουν μια πραγματική (real) θετική παρατήρηση και μία πραγματική (real) αρνητική αντίστοιχα. Οι γραμμές αντιπροσωπεύουν μία προβλεπόμενη κλάση και αντίστοιχα το +P μια θετικά προβλεπόμενη (predicted) παρατήρηση και –P μία αρνητικά προβλεπόμενη παρατήρηση. Με pp και pn συμβολίζονται τα σύνολα των προβλεπόμενων θετικών (predicted positive) και προβλεπόμενων αρνητικών (predicted negative), ενώ με rp (real positive) το σύνολο των πραγματικά θετικών παρατηρήσεων και με rn (real negative) το σύνολο των πραγματικά αρνητικών παρατηρήσεων. Συνεπώς η πράσινη διαγώνιος μας υποδεικνύει τις σωστές προβλέψεις , ενώ η κόκκινη τις εσφαλμένες.

Ανάκληση (Recall) / True Positive Rate : Εκφράζει το ποσοστό των πραγματικά θετικών παρατηρήσεων που ο αλγόριθμος πρόβλεψε ως θετικές.

$$Recall = \frac{Θετικές\ Παρατηρήσεις\ που\ ταξινομήθηκαν\ ως\ θετικές}{Πραγματικά\ Θετικές\ παρατηρήσεις} = \frac{tp}{rp} = \frac{tp}{tp + fn}$$

Ποσοστό Ψευδών Θετικών (False Positive Rate): Εκφράζει το ποσοστό των πραγματικά αρνητικών παραδειγμάτων που ο αλγόριθμος ταξινόμησε εσφαλμένα ως θετικά.

$$False\ Positive\ Rate = \frac{Αρνητικές\ Παρατηρήσεις\ που\ ταξινομήθηκαν\ ως\ θετικές}{Πραγματικά\ Αρνητικές\ Παρατηρήσεις} = \frac{fp}{tn}$$

Ακρίβεια (Precision): Είναι το ποσοστό των παρατηρήσεων που προβλέφθηκαν ως θετικές και είναι πραγματικά θετικές.

$$Precission = \frac{Θετικές\ Παρατηρήσεις\ που\ ταξινομήθηκαν\ ως\ θετικές}{Σύνολο\ Παρατηρήσεων\ που\ προβλευθησαν\ ώς\ θετικές} = \frac{tp}{pp} = \frac{tp}{tp + fp}$$

Accuracy: Εκφράζει το ποσοστό των σωστά ταξινομημένων παρατηρήσεων ως προς το σύνολο των παρατηρήσεων που εξετάζουμε.

$$Accuracy = \frac{Παρατηρήσεις\ που\ ταξινομήθηκαν\ σωστά}{Σύνολο\ Παρατηρήσεων} = \frac{tp + tn}{tp + fp + tn + fn}$$

F-measure: Είναι ο αρμονικός μέσος των Recall και Precision.

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

ΚΕΦΑΛΑΙΟ 3^{ο} – Πειραματική Διαδικασία

3.1. Περιγραφή Βάσης Δεδομένων (Data Description)

Για την εκπαίδευση του συστήματός μας αλλά και την αξιολόγηση χρησιμοποιήθηκαν αρχεία τύπου .wav και αρχεία τύπου .mp3.

Λόγω του ότι δεν υπάρχουν έτοιμες κατάλληλες βάσεις δεδομένων στο πεδίο αναγνώρισης ήχων από ραδιόφωνο, η επιλογή των ακουστικών κομματιών έγινε από ιστοσελίδες ραδιοφωνικών σταθμών που διαθέτουν αρχείο ραδιοφωνικών εκπομπών και (podcasts) ραδιοφωνικών ειδήσεων.

Περιληπτικά, οι ραδιοφωνικές εκπομπές που αποτέλεσαν τα ηχητικά μας δεδομένα, ποικίλουν από 27-34 λεπτά σε διάρκεια και είναι ηχογραφημένα μόνο στην αγγλική γλώσσα, έτσι ώστε να αποφευχθεί σύγχυση στα αποτελέσματα, λόγω διαφορετικής γλώσσας . Οι ραδιοφωνικές εκπομπές που επιλέχτηκαν, χωρίς προηγούμενη γνώση του περιεχομένου τους, είναι ενημερωτικού αλλά και ψυχαγωγικού περιεχομένου, έτσι ώστε να υπάρχει μία σχετική ισορροπία στο περιεχόμενο όσον αφορά τα ποσοστά ομιλίας - μουσικής.

Συγκεντρώσαμε 210 λεπτά ηχητικών δεδομένων από το BBC broadcast news database (BBC News) και πιο συγκεκριμένα από το ενημερωτικό πρόγραμμα “Global News”, ενώ στην βάση δεδομένων μας περιέχονταν και ηχητικά κομμάτια από το κανάλι BBC Radio 6 και την ψυχαγωγική εκπομπή “Screen 6” . Στο Παράρτημα Α περιλαμβάνεται αναλυτικός κατάλογος, με όλες τις εκπομπές καθώς και κάποια χαρακτηριστικά τους (format ήχου, συνολική διάρκεια, κανάλι προέλευσης, είδος εκπομπής).

3.2. Μη – αυτόματη τμηματοποίηση ήχου

Ένα μεγάλο μέρος κατά την διεξαγωγή αυτής της εργασίας ήταν η μη- αυτόματη τμηματοποίηση (segmentation) των αρχείων ήχου με την χρήση του προγράμματος PRAAT (Praat Manual) . Αυτή η διαδικασία, ήταν η πιο δύσκολη και η πιο χρονοβόρα, καθότι απαιτούσε μεγάλη προσοχή στον καθορισμό των ορίων του κάθε τμήματος, με βάση την ακουστική αντίληψη του ακουστικού σήματος αλλά και την εικόνα της κυμματομορφής του. Κατόπιν, ακολούθησε χειροκίνητο labeling , ή αλλιώς annotation σύμφωνα με τις βασικές κατηγορίες ενδιαφέροντος στις εξής κατηγορίες: speech, music, silence, noise και κατά την διάρκεια αυτής της διαδικασίας προέκυψαν και οι μικτές κατηγορίες speech over music, speech over noise.

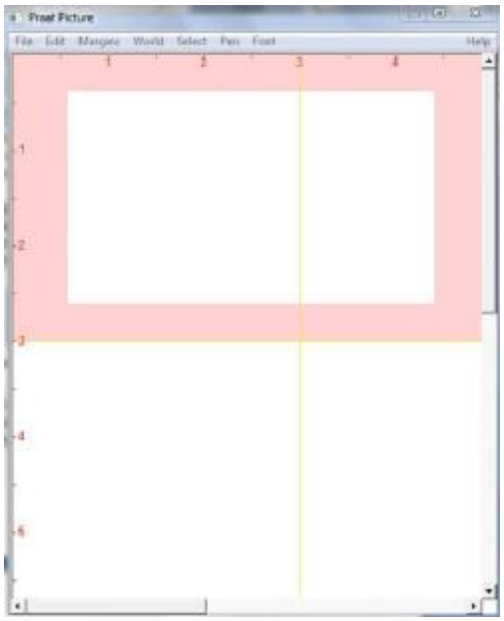
Το πρόγραμμα Praat , είναι ένα πρόγραμμα που μπορεί να δημιουργήσει, να αναπαράγει, να κάνει φωνητική ανάλυση καθώς και επεξεργασία ήχων. (Boersma and Weenink, 1992-2012). Είναι διαθέσιμο για όλες σχεδόν τα λειτουργικά συστήματα (Windows, Macintosh, Unix, Linux) και μπορεί να αποκτηθεί από τους χρήστες από την ηλεκτρονική διεύθυνση (www.praat.org). Στο Praat μπορούμε να ηχογραφήσουμε σε μορφές στέρεο (stereo) αλλά

και μορφές μόνο (mono). Παρόλα αυτά, στο Praat μπορούν να επεξεργαστούν και να αναλυθούν μόνο σήματα mono. (Μπουρνά, 2008). Στα σήματα stereo μπορεί να γίνει ανάλυση και επεξεργασία στο Praat μόνο για τα μεμονωμένα κανάλια του σήματος.

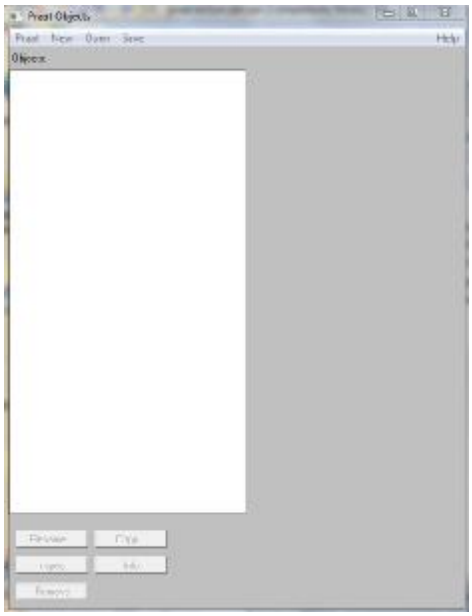
Το Praat ξεκινάει με το άνοιγμα δύο παραθύρων.

Το πρώτο είναι το Praat picture, το οποίο δεν χρησιμοποιήθηκε στην παρούσα εργασία, εμφανίζει τα διαγράμματα που προκύπτουν από την επεξεργασία των ηχητικών σημάτων. (Εικόνα 3.1.)

Το δεύτερο παράθυρο είναι το Praat Objects (Εικόνα 3.2.), στο οποίο μπορούμε να εισάγουμε όλα τα ηχητικά δεδομένα, να τα επιλέξουμε, να τα επεξεργαστούμε και να τα αναπαράγουμε.



Εικόνα 3.1. Παράθυρο Praat picture



Εικόνα 3.2. Παράθυρο Praat Objects

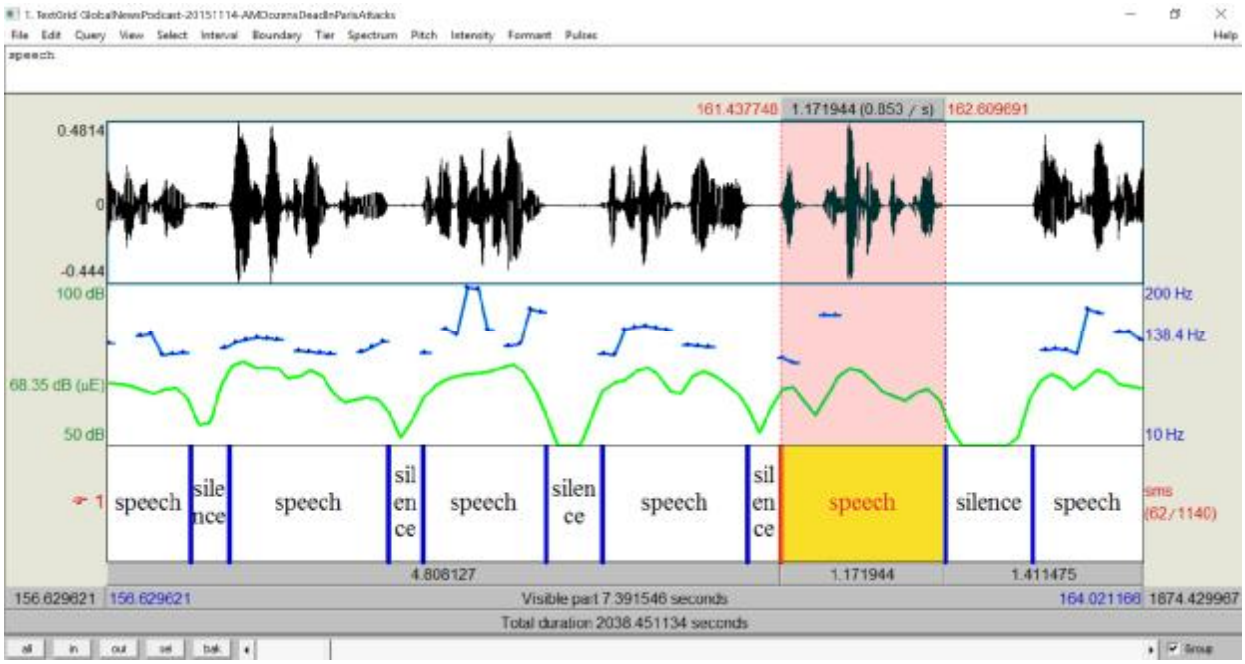
Στο παράθυρο Praat objects, με την επιλογή Menu > Read > Read from file… φορτώνονται τα αρχεία ήχου, που έχουν επιλεχθει προς επεξεργασία.

Μετά την φόρτωση των αρχείων μπορούν να πραγματοποιηθούν διάφορες ενέργειες, χρησιμοποιώντας τα κουμπιά που εμφανίζονται στα δεξιά του παραθύρου praat objects. Πιο συγκεκριμένα μπορούμε να δούμε την κυμματομορφή του ακουστικού σήματος (Edit), να αναπαράγουμε το αρχείο (Play), να σχεδιάσουμε το ηχητικό σήμα (Draw), να το τροποποιήσουμε (Modify), να πάρουμε διάφορες πληροφορίες σχετικά με τον χρόνο και την δειγματοληψία (Query) , να σχολιάσουμε τον επιλεγμενο ήχο και να δημιουργήσουμε αρχείο Textgrid (Annotate) και διάφορες άλλες επιλογές που μπορούν να κάνουν διάφορους υπολογισμους, οι οποίοι θα μας δώσουν πληροφορίες σχετικά με το σήμα που εξετάζουμε και τα χαρακτηριστικά του.

Με την επιλογή Annotate μπορούμε να δημιουργήσουμε και να επεξεργαστούμε αρχεία κειμένου, τα οποία ονομάζονται πλέγματα κειμένου, ή πιο κοινώς καλούμενα Textgrids, στα οποία καταγράφουμε ετικέτες (labeling) των τμημάτων του σήματος.

Στην περίπτωση της παρούσας εργασίας, κάθε τμήμα του σήματος μας που περιέχονταν στο αρχικό αρχείο audio χαρακτηρίστηκε (labeled) ως speech, music, silence και noise, και τα

μικτά είδη, τα οποία είναι εκείνα τα τμήματα τα οποία έχουν δύο ακουστικά επίπεδα (κύριο και background) και τα οποία χαρακτηρίστηκαν ως εξής “speech over music”, “speech over noise”, “music over noise”. Σε αυτό το σημείο αντιμετωπίσαμε δυσκολία στον καθορισμό των ορίων κάποιων τμημάτων καθότι οι κατηγορίες σε πολλές περιπτώσεις είναι υποκειμενικές όπως στην δική μας περίπτωση που θεωρήσαμε ως σιωπή τις μικρές παύσεις (> 0.02sec) του ομιλητή, τα άηχα σύμφωνα κατά την έναρξη την ομιλίας, καθώς επίσης και τις ανάσες παρότι φαινομενικά, σε επίπεδο έντασης, υπάρχει παρουσία ήχου. Τέλος, θεωρούμε ως θόρυβο οποιοδήποτε ανεπιθύμητο ήχο, οπότε στην περίπτωση που ένας ομιλητής μιλούσε πάνω στον άλλο ή υπήρχε παρουσία μη-γλωσσικού θορύβου (βήχας, γέλιο) θεωρήθηκε ως θόρυβος και επομένως τα συγκεκριμένα τμήματα χαρακτηρίστηκαν ως “speech over noise”.



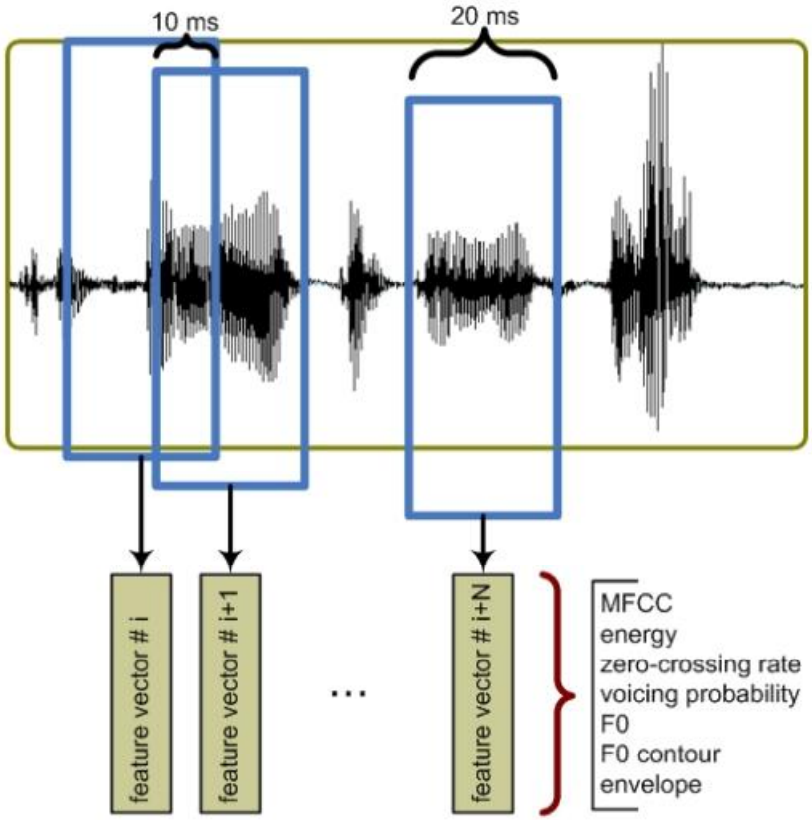
Εικόνα 3.3. Παράδειγμα labeling

Στην παρούσα εργασία επεξεργάστηκαν δεδομένα και αξιολογήθηκαν τα αποτελέσματα των αλγορίθμων κατηγοριοποίησης για τα τμήματα που είναι χαρακτηρισμένα ως ‘Music’, ‘Speech’ ή ‘Speech over Music’, ενώ αγνοήθηκαν τα υπόλοιπα τμήματα ως other. Στο Παράρτημα Β υπάρχει ο πίνακας με τις συνολική διάρκεια αυτών των κατηγοριών για όλα τα αρχεία ήχου που χρησιμοποιήθηκαν.

Μία ακόμη δυνατότητα του προγράμματος Praat είναι αυτή της συγγραφής αρχείων εντολών (scripts), τα οποία επιτρέπουν την πλήρη ή μερική αυτοματοποίηση κάποιων διαδικασιών του. (Μπουρνά, 2008) Τα script files είναι εξαιρετικά χρήσιμα για την επανάληψη μεγάλων ακολουθιών εντολών σε διαφορετικά δεδομένα και χαρακτηρίζονται από την επέκταση .praat. Στην παρούσα εργασία πραγματοποιήθηκε χρήση των Praat scripts για να εξάχθούν τα επιμέρους χαρακτηρισμένα τμήματα του αρχικού ηχητικού αρχείου σε ξεχωριστά .wav αρχεία. Στο Παράρτημα Γ παρουσιάζεται ο κώδικας script όπως εισήχθη και εφαρμόστηκε στα δεδομένα μας.

3.3. Εξαγωγή Χαρακτηριστικών - Feature Extraction

Αυτή η διαδικασία της εξαγωγής δεδομένων, η οποία ονομάζεται και παραμετροποίηση (parametrization) ,αποτελεί ίσως το πιο σημαντικό κομμάτι της διαδικασίας της αυτόματης κατηγοριοποίησης ήχων. Στην εικόνα 3.4. παρουσιάζεται ενδεικτικά η διαδικασία της παραμετροποίησης.



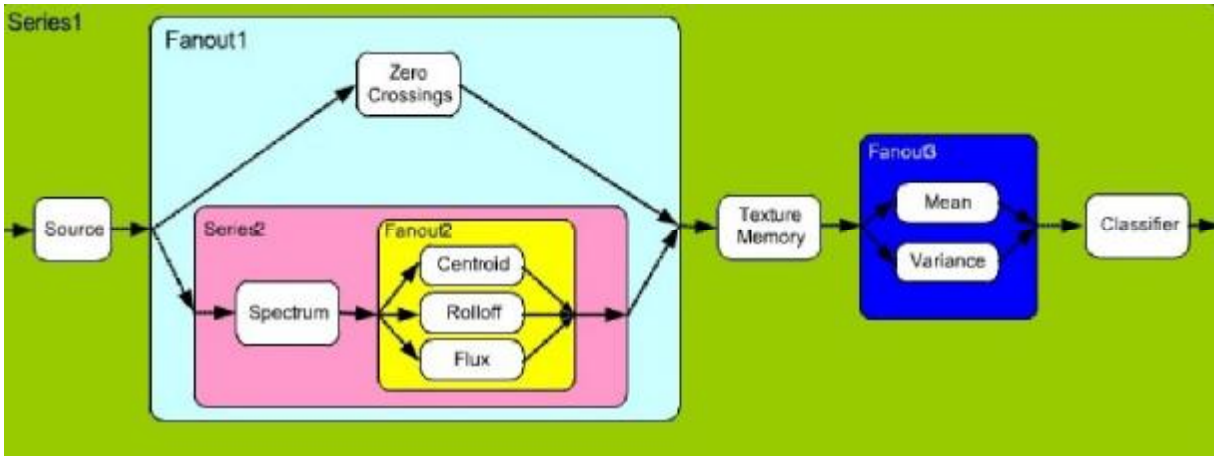
Εικόνα 3.4. Διαδικασία εξαγωγής διανυσμάτων χαρακτηριστικών από ηχητικά σήματα ραδιοφωνικών εκπομπών

Πιο συγκεκριμένα, το ακουστικό σήμα χωρίζεται σε τμήματα των 20 msec αλληλεπικαλυπτόμενα, με επικάλυψη 50% για καλύτερα αποτελέσματα, και από κάθε τμήμα εξάγεται ένα διάνυσμα χαρακτηριστικών το οποίο μας δίνει πληροφορίες για το ηχητικό σήμα.

Η μεθοδολογία που χρησιμοποιήθηκε, βασίζεται στην εξαγωγή ηχητικών χαρακτηριστικών μέσω του MARSYAS v.0.5.0 (Music Analysis Retrieval and SYnthesis for Audio Signals) (Tzanetakis, Music Analysis, Retrieval and Synthesis of Audio Signals MARSYAS, 2009). Το MARSYAS είναι ένα εργαλείο ανοιχτού κώδικα (open source software framework) που λειτουργεί μέσω του περιβάλλοντος γραμμής εντολών του λειτουργικού μας συστήματος (command line – CMD) και μας δίνει την δυνατότητα να εξάγουμε συγκεκριμένα χαρακτηριστικά των ηχητικών σημάτων, που περιέχονται στην βάση δεδομένων που εξετάζουμε. Τα χαρακτηριστικά αυτά, τα οποία εξάγονται και στην συνέχεια συγχωνεύονται σε μορφή κειμένου-διανυσμάτων (feature vectors), μπορούν να βοηθήσουν στην ταξινόμηση των ηχητικών σημάτων, μέσω χρήσης αλγορίθμων μάθησης, στο είδος ήχου στο οποίο ανήκουν. Για την χρήση του προγράμματος χρησιμοποιήθηκε το εγχειρίδιο του MARSYAS

(Tzanetakis, Marsyas User Manual) καθώς επίσης και την αναφορά του Tzanetakis – Marsyas Submissions to Mirex 2009 (Tzanetakis, 2009).

Στο παρακάτω διάγραμμα φαίνεται συμβολικά η διαδικασία εξαγωγής χαρακτηριστικών πριν την ταξινόμηση των ηχητικών σημάτων σε κατηγορίες μέσω του MARSYAS.



Εικόνα 3.5. Ενδεικτικό διάγραμμα εξαγωγής χαρακτηριστικών με το Marsyas

Τα χαρακτηριστικά που εξήχθησαν με την χρήση του εργαλείου Marsyas είναι τα Spectral Features (Spectral Centroid, Spectral Roll Off, Spectral Flux), τα Mel-Frequency Cepstral Coefficients (MFCC – 12 features) ενώ στην συνέχεια προστίθενται σε αυτά σε ξεχωριστή κατηγορία τα Zero-Crossings και το Short-time Energy features τα οποία συγκεντρώνονται σε μία ενιαία κατηγορία που ονομάζεται STFTMFCC ή αλλιώς Timbral Features. Τα παραπάνω σύνολα εξαγόμενων χαρακτηριστικών περιγράφονται αναλυτικά στις παραγράφους 2.1.2, 2.1.3., 2.1.4. και 2.1.5.

Αφού μπούμε στο φάκελο του MARSYAS με την εντολή "cd" στο command line του λειτουργικού μας συστήματος, τρέχουμε το πρόγραμμα με την εντολή marsyas-run.

```
>> cd C:\Users\Anastasia Catela\Desktop\Thesis\marsyas-0.5.0\bin
>> marsyas-run
```

Στην συνέχεια, με την εντολή "mkcollection" δημιουργούμε τρεις συλλογές .mf , μία για την ομιλία, μία για την μουσική και μία για την μικτή κατηγορία, ώστε να μπορούν να χρησιμοποιηθούν από το MARSYAS. Να σημειωθεί, πως η εντολή αυτή διαβάζει μόνο τα αρχεία .wav και τα αρχεία .mp3 που έχουμε στα αντίστοιχα directories, και δεν μας εξασφαλίζει το περιεχόμενο της συλλογής που δημιουργείται αυτόματα.

```
>> mkcollection -c speech.mf -l speech C:\Users\Public\Speech_all
>> mkcollection -c music.mf -l music C:\Users\Public\Music_all
>> mkcollection –c mixed.mf –l mixed C:\Users\Public\Speech_over_Music
```

Τέλος, με την εντολή "bextract" εξάγουμε τα χαρακτηριστικά σε μορφή κειμένου-διανύσματος. Τα αρχεία αυτά έχουν την κατάληξη .arff ώστε να μπορούν να είναι συμβατά, και κατά συνέπεια επεξεργάσιμα, από λογισμικά και εφαρμογές data mining όπως είναι το

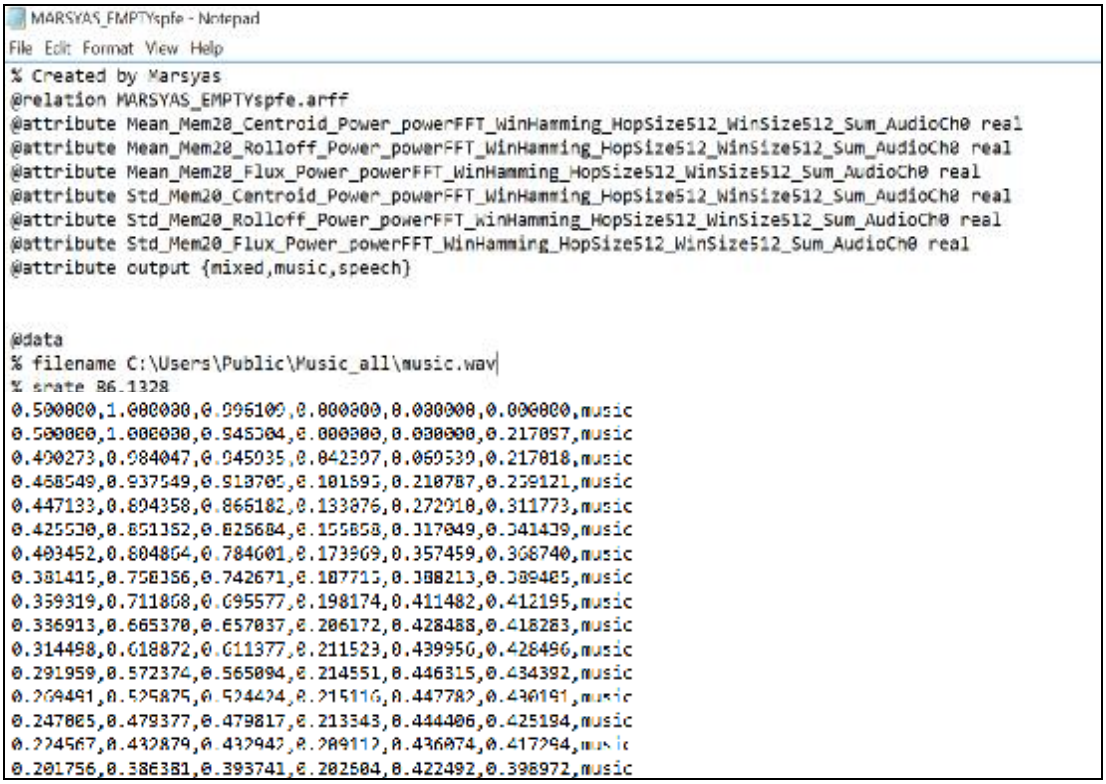
περιβάλλον εφαρμογής WEKA (Weka 3: Data Mining Software in Java). Εάν θέλουμε να εξάγουμε ένα συγκεκριμένο χαρακτηριστικό ή σύνολο χαρακτηριστικών πληκτρολογούμε την εντολή bextract και ορίζουμε τον τύπο των χαρακτηριστικών που θέλουμε να εξάγουμε.

```
bextract –timbral music.mf speech.mf –mixed.mf -w timbral.arff
bextract –spfe music.mf speech.mf –mixed.mf -w spfe.arff
bextract –mfcc music.mf speech.mf mixed.mf –w mfcc.arff
```

Με την εντολή “bextract single vector”, και πιο συγκεκριμένα με την προσθήκη της εντολής –sv κατά την εξαγωγή των χαρακτηριστικών τα χαρακτηριστικά εξάγονται σε ένα και μοναδικό διάνυσμα ανεξάρτητα από το μήκος του ηχητικού σήματος. Με αυτόν τον τρόπο, λόγω των λιγότερων παρατηρήσεων που ροκύπτουν, έχουμε καλύτερα αποτελέσματα και απαιτείται λιγότερος χρόνος για την διαδικασία της ταξινόμησης.

```
bextract –sv –timbral music.mf speech.mf mixed.mf -w timbral_sv.arff
bextract –sv –spfe music.mf speech.mf mixed.mf -w spfe_sv.arff
bextract –sv –mfcc music.mf speech.mf mixed.mf -w mfcc_sv.arff
```

Στην παρακάτω εικόνα φαίνεται ένα παράδειγμα από ένα τμήμα αρχείου .arff (Εικόνα 3.6.)



Εικόνα 3.6. Παράδειγμα αρχείου εξαγόμενων χαρακτηριστικών σε μορφή κειμένου .arff

3.4. Κατηγοριοποίηση – Classification

Για την κατασκευή των μοντέλων κατηγοριοποίησης (classification models), χρησιμοποιήσαμε τους αλγορίθμων μηχανικής μάθησης του εργαλείου WEKA (Weka 3: Data Mining Software in Java). Το WEKA (Waikato Environment Knowledge Analysis) είναι ένα πρόγραμμα εξόρυξης δεδομένων (Data mining) το οποίο είναι γραμμένο σε γλώσσα προγραμματισμού Java και αναπτύχθηκε από το Πανεπιστήμιο του Waikato της Νέα Ζηλανδίας.

Η επιλογή των μεθόδων κατηγοριοποίησης βασίστηκε στις πιο ευρέως γνωστές και χρησιμοποιούμενες μεθόδους στο πεδίο της επεξεργασίας , ήχου, μουσικής και ομιλίας. Πιο συγκεκριμένα, τα αποτελέσματα των αλγορίθμων που αξιολογήθηκαν ήταν με τους αλγόριθμους κατηγοριοποίησης (i) ZeroR, (ii) Naive Bayes (iii) Support Vector Machines (SVM) αλγόριθμο και πιο συγκεκριμένα με τον αλγόριθμο SMO (Sequential Minimal Optimization) και τέλος (iv) με αλγόριθμο τύπου C4.5 που ανήκει στους αλγορίθμους που χρησιμοποιούν τα δέντρα απόφασης, και οι οποίοι στο περιβάλλον του WEKA καλούνται ως J48. Τους αλγόριθμους αυτούς τους αναφέραμε και τους περιγράψαμε αναλυτικά στην ενότητα 2.2.

Η διαδικασία που ακολουθείται είναι να εισάγουμε το αρχείο .arff στο WEKA και στην καρτέλα Classify επιλέγουμε τον αντίστοιχο αλγόριθμο κατηγοριοποίησης που επιθυμούμε να χρησιμοποιήσουμε. Στην συνέχεια, και αφού επιλέξουμε μέθοδο αξιολόγησης και επικυροποίησης 10-fold cross validation πατάμε start. Αφού ολοκληρωθεί η διαδικασία, τα αποτελέσματα της κατηγοριοποίησης που αποτελούνται από τα μέτρα αξιολόγησης που αναφέραμε στην ενότητα 2.3.1. παρουσιάζονται στα δεξιά του παραθύρου. Παρακάτω παρουσιάζονται τα αποτελέσματα της ταξινόμησης όπως προέκυψαν από την πειραματική διαδικασία.

Σημαντικό είναι να τονίσουμε πως όλοι οι αλγόριθμοι εκπαιδεύτηκαν με την ίδια βάση δεδομένων εκπαίδευσης (training dataset) καθώς επίσης και ακολουθήσαμε την ίδια μέθοδο και τα ίδια κριτήρια για την αξιολόγηση των αλγορίθμων, για λόγους άμεσης σύγκρισης των αποτελεσμάτων και των ποσοστών επιτυχίας πρόβλεψης μεταξύ των αλγορίθμων.

=== Classifier model ===

ZeroR predicts class value: speech

=== Summary ===

Correctly Classified Instances	508452	56.2915 %
Incorrectly Classified Instances	394796	43.7085 %
Total Number of Instances	903248	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0,000	0,000	0,000	0,000	0,000	mixed
	0,000	0,000	0,000	0,000	0,000	music
	1,000	1,000	0,563	1,000	0,720	speech
Weighted Avg.	0,563	0,563	0,317	0,563	0,405	

=== Confusion Matrix ===

a	b	c	<-- classified as
0	0	184970	a = mixed
0	0	209826	b = music
0	0	508452	c = speech

Πίνακας 3.1. Αποτελέσματα μεθόδου ZeroR κατηγοριοποίησης στο WEKA

=== Classifier model ===
Naive Bayes Classifier

Attribute	Class		
	mixed (0.2)	music (0.23)	speech (0.56)

=== Summary ===

Correctly Classified Instances	691437	76.5501 %
Incorrectly Classified Instances	211811	23.4499 %
Total Number of Instances	903248	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Class**
	0,607	0,143	0,523	0,607	0,562	mixed
	0,875	0,077	0,776	0,875	0,822	music
	0,778	0,142	0,876	0,778	0,824	speech
Weighted Avg.	0,766	0,127	0,780	0,766	0,770	

=== Confusion Matrix ===

a	b	c	<-- classified as
112352	26045	46573	a = mixed
16663	183673	9490	b = music
85935	27105	395412	c = speech

Πίνακας 3.2. Αποτελέσματα μεθόδου Naïve Bayes στο WEKA με χρήση MFCC χαρακτηριστικών

=== Classifier model ===

Naive Bayes Classifier

Attribute	Class mixed (0.2)	 music (0.23)	 speech (0.56)

=== Summary ===

Correctly Classified Instances	470392	52.0778 %
Incorrectly Classified Instances	432856	47.9222 %
Total Number of Instances	903248	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Class**
	0,179	0,086	0,350	0,179	0,237	mixed
	0,798	0,411	0,370	0,798	0,506	music
	0,531	0,219	0,757	0,531	0,624	speech
Weighted Avg.	0,521	0,236	0,584	0,521	0,517	

=== Confusion Matrix ===

a	b	c	<-- classified as
33196	93199	58575 |	a = mixed
14586	167440	27800 |	b = music
46982	191714	269756 |	c = speech

Πίνακας 3.3. Αποτελέσματα αλγορίθμου Naïve Bayes στο WEKA με χρήση φασματικών χαρακτηριστικών (spfe)

=== Classifier model ===

Naive Bayes Classifier

Attribute	Class mixed (0.2)	music (0.23)	speech (0.56)

=== Summary ===

Correctly Classified Instances	462094	51.1592 %
Incorrectly Classified Instances	441154	48.8408 %
Total Number of Instances	903248	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Class**
	0,843	0,531	0,290	0,843	0,432	mixed
	0,542	0,053	0,756	0,542	0,631	music
	0,379	0,059	0,893	0,379	0,532	speech
Weighted Avg.	0,512	0,154	0,737	0,512	0,534	

=== Confusion Matrix ===

a	b	c	<-- classified as
155882	11362	17726	a = mixed
90690	113693	5443	b = music
290562	25371	192519	c = speech

Πίνακας 3.4. Αποτελέσματα αλγορίθμου Naïve Bayes στο WEKA με χρήση timbral χαρακτηριστικών.

=== Classifier model ===
ZeroR predicts class value: speech

=== Summary ===

Correctly Classified Instances	2086	68.2815 %
Incorrectly Classified Instances	969	31.7185 %
Total Number of Instances	3055	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Class**
	0,000	0,000	0,000	0,000	0,000	mixed
	0,000	0,000	0,000	0,000	0,000	music
	1,000	1,000	0,683	1,000	0,812	speech
Weighted Avg.	0,683	0,683	0,466	0,683	0,554	

=== Confusion Matrix ===

a	**b**	**c**	<-- classified as
0	0	542	| a = mixed
0	**0**	427	| b = music
0	0	**2086**	| c = speech

Πίνακας 3.5. Αποτελέσματα μεθόδου ZeroR κατηγοριοποίησης στο WEKA με single vector

=== Classifier model ===

Naive Bayes Classifier

Attribute	Class mixed (0.18)	music (0.14)	speech (0.68)

=== Summary ===

Correctly Classified Instances	2529	82.7823 %
Incorrectly Classified Instances	526	17.2177 %
Total Number of Instances	3055	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Class**
	0,771	0,136	0,549	0,771	0,642	mixed
	0,941	0,029	0,841	0,941	0,888	music
	0,819	0,110	0,941	0,819	0,876	speech
Weighted Avg.	0,828	0,104	0,858	0,828	0,836	

=== Confusion Matrix ===

a	b	c	<-- classified as
418	23	101	a = mixed
19	402	6	b = music
324	53	1709	c = speech

Πίνακας 3.6. Αποτελέσματα αλγορίθμου Naïve Bayes στο WEKA με χρήση MFCC χαρακτηριστικών (single vector)

=== Classifier model ===
Naive Bayes Classifier

Attribute	Class		
	mixed	music	speech
	(0.18)	(0.14)	(0.68)

=== Summary ===

Correctly Classified Instances	1934	63.3061 %
Incorrectly Classified Instances	1121	36.6939 %
Total Number of Instances	3055	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Class**
	0,485	0,195	0,349	0,485	0,406	mixed
	0,792	0,143	0,473	0,792	0,592	music
	0,639	0,261	0,840	0,639	0,726	speech
Weighted Avg.	0,633	0,233	0,702	0,633	0,650	

=== Confusion Matrix ===

a	b	c	<-- classified as
263	70	209	a = mixed
45	338	44	b = music
446	307	1333	c = speech

Πίνακας 3.7. Αποτελέσματα αλγορίθμου Naïve Bayes στο WEKA με χρήση φασματικών χαρακτηριστικών (single vector)

=== Classifier model ===

Naive Bayes Classifier

Attribute	Class mixed (0.18)	music (0.14)	speech (0.68)

=== Summary ===

Correctly Classified Instances	2513	82.2586 %
Incorrectly Classified Instances	542	17.7414 %
Total Number of Instances	3055	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Class**
	0,485	0,076	0,579	0,485	0,528	mixed
	0,906	0,026	0,851	0,906	0,878	music
	0,893	0,292	0,868	0,893	0,880	speech
Weighted Avg.	0,823	0,217	0,814	0,823	0,818	

=== Confusion Matrix ===

a	b	c	<-- classified as
263	22	257	a = mixed
14	387	26	b = music
177	46	1863	c = speech

Πίνακας 3.8. Αποτελέσματα αλγορίθμου Naïve Bayes στο WEKA με χρήση timbral χαρακτηριστικών (single vector)

=== Classifier model ===

SMO

=== Summary ===

Correctly Classified Instances	2683	87.8232 %
Incorrectly Classified Instances	372	12.1768 %
Total Number of Instances	3055	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Class**
	0,561	0,041	0,747	0,561	0,641	mixed
	0,946	0,014	0,918	0,946	0,932	music
	0,947	0,240	0,894	0,947	0,920	speech
Weighted Avg.	0,878	0,173	0,872	0,878	0,872	

=== Confusion Matrix ===

a	b	c	<-- classified as
304	13	225	a = mixed
15	404	8	b = music
88	23	1975	c = speech

Πίνακας 3.9. Αποτελέσματα αλγορίθμου SVM (SMO) στο WEKA με χρήση MFCC χαρακτηριστικών (single vector)

=== Classifier model ===

SMO

=== Summary ===

Correctly Classified Instances	2322	76.0065 %
Incorrectly Classified Instances	733	23.9935 %
Total Number of Instances	3055	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ClassZ**
	0,000	0,001	0,000	0,000	0,000	mixed
	0,705	0,036	0,760	0,705	0,731	music
	0,969	0,656	0,761	0,969	0,852	speech
Weighted Avg.	0,760	0,453	0,626	0,760	0,684	

=== Confusion Matrix ===

a	b	c	<-- classified as
0	30	512	a = mixed
2	301	124	b = music
0	65	2021	c = speech

Πίνακας 3.10. Αποτελέσματα αλγορίθμου SVM (SMO) στο WEKA με χρήση φαμσατικών χαρακτηριστικών (single vector)

=== Classifier model ===

SMO

=== Summary ===

Correctly Classified Instances	2705	88.5434 %
Incorrectly Classified Instances	350	11.4566 %
Total Number of Instances	3055	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Class**
	0,616	0,045	0,747	0,616	0,675	mixed
	0,956	0,010	0,938	0,956	0,947	music
	0,941	0,217	0,903	0,941	0,922	speech
Weighted Avg.	0,885	0,157	0,880	0,885	0,882	

=== Confusion Matrix ===

a	b	c	<-- classified as
334	8	200	a = mixed
9	408	10	b = music
104	19	1963	c = speech

Πίνακας 3.11. Αποτελέσματα αλγορίθμου SVM (SMO) στο WEKA με χρήση timbral χαρακτηριστικών (single vector)

=== Classifier model ===

J48 pruned tree

Number of Leaves : 139
Size of the tree : 277

=== Summary ===

Correctly Classified Instances	2552	83.5352 %
Incorrectly Classified Instances	503	16.4648 %
Total Number of Instances	3055	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Class**
	0,583	0,084	0,600	0,583	0,591	mixed
	0,885	0,019	0,885	0,885	0,885	music
	0,891	0,251	0,884	0,891	0,888	speech
Weighted Avg.	0,835	0,189	0,834	0,835	0,835	

=== Confusion Matrix ===

a	b	c	<-- classified as
316	15	211	a = mixed
17	378	32	b = music
194	34	1858	c = speech

Πίνακας 3.12. Αποτελέσματα αλγορίθμου C4.5 (J48) στο WEKA με χρήση MFCC χαρακτηριστικών (single vector)

=== Classifier model ===

J48 pruned tree

Number of Leaves : 135
Size of the tree : 269

=== Summary ===

Correctly Classified Instances	2308	75.5483 %
Incorrectly Classified Instances	747	24.4517 %
Total Number of Instances	3055	

=== Detailed Accuracy By Class ===

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Class**
	0,292	0,064	0,497	0,292	0,367	mixed
	0,672	0,045	0,710	0,672	0,691	music
	0,893	0,485	0,799	0,893	0,843	speech
Weighted Avg.	0,755	0,349	0,733	0,755	0,737	

=== Confusion Matrix ===

a	b	c	<-- classified as
158	23	361	a = mixed
31	287	109	b = music
129	94	1863	c = speech

Πίνακας 3.13. Αποτελέσματα αλγορίθμου C4.5 (J48) στο WEKA με χρήση φασματικών χαρακτηριστικών (single vector)

=== Classifier model ===

J48 pruned tree

Number of Leaves : 140
Size of the tree : 279

=== Summary ===

Correctly Classified Instances	2547	83.3715 %
Incorrectly Classified Instances	508	16.6285 %
Total Number of Instances	3055	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0,572	0,081	0,603	0,572	0,587	mixed
	0,869	0,021	0,873	0,869	0,871	music
	0,895	0,258	0,882	0,895	0,888	speech
Weighted Avg.	0,834	0,193	0,831	0,834	0,832	

=== Confusion Matrix ===

a	b	c	<-- classified as
310	21	211	a = mixed
17	371	39	b = music
187	33	1866	c = speech

Πίνακας 3.14. Αποτελέσματα αλγορίθμου C4.5 (J48) στο WEKA με χρήση timbral χαρακτηριστικών (single vector)

ΚΕΦΑΛΑΙΟ 4ο

4.1. Εμπειρικά αποτελέσματα – Αξιολόγηση

Αναλύοντας τα παραπάνω αποτελέσματα παρατηρούμε ότι όταν εξάγουμε διανύσματα χαρακτηριστικών για κάθε τμήμα ακουστικού σήματος, τότε έχουμε πάρα πολλές ξεχωριστές παρατηρήσεις που πρέπει να ταξινομηθούν (903.248), το οποίο είναι αρκετά χρονοβόρο και τα αποτελέσματα που προκύπτουν είναι αρκετά μη-αποτελεσματικά. Πιο συγκεκριμένα, έχοντας ως κατώφλι το 68,28% (Πίνακας 3.1.) που μας δίνει ο αλγόριθμος ZeroR, σαν κατώτατο όριο ποσοστού επιτυχίας ταξινόμησης, παίρνουμε 76,55%,(Πίνακας 3.2.) για τα δεδομένα από MFCC χαρακτηριστικά και χρησιμοποιώντας τον αλγόριθμο Naïve Bayes, ένα ποσοστό ελαφρώς μεγαλύτερο από το όριο που έχουμε θέσει, ενώ για τα φασματικά και τα timbral χαρακτηριστικά, χρησιμοποιώντας τον ίδιο αλγόριθμο, τα ποσοστά επιτυχούς ταξινόμησης που πετυχαίνουμε είναι 52,08% (Πίνακας 3.3.) και 51,16% (Πίνακας 3.4.) αντίστοιχα, αποτελέσματα που κρίνουν αυτούς τους συνδυασμούς εντελώς ακατάλληλους.

Τα πειράματα που διεξήχθησαν εξάγοντας τα χαρακτηριστικά σε single vector, μας δίνουν συνολικά 3055 παρατηρήσεις, απαιτήθηκε σημαντικά λιγότερος χρόνος για την διεξαγωγή των πειραμάτων και μας δίνουν μεγαλύτερο ποσοστό επιτυχίας, ξεκινώντας μάλιστα από το ίδιο το κατώφλι που θέτει ο αλγόριθμος ZeroR και είναι στο 68,28% (Πίνακας 3.5.).

Για τον αλγόριθμο Naïve Bayes τα ποσοστά επιτυχούς ταξινόμησης είναι σημαντικά καλύτερα για όλες τις κατηγορίες χαρακτηριστικών με ποσοστό επιτυχούς ταξινόμησης 82,78% (Πίνακας 3.6.) για την ταξινόμηση που έγινε με την βοήθεια των MFCC χαρακτηριστικών, 63,31% (Πίνακας 3.7.) για τα φασματικά χαρακτηριστικά (spfe) και 82,26% (Πίνακας 3.8.) το ποσοστό της επιτυχούς ταξινόμησης όταν βασιζόμαστε στην εξαγωγή των timbral χαρακτηριστικών. Παρατηρούμε παρόλα αυτά πως το ποσοστό 63,31% είναι χαμηλότερο της βάσης που έχει τεθεί με την βοήθεια του αλγορίθμου ZeroR και συνεπώς ο συνδυασμός του αλγορίθμου Naïve Bayes χρησιμοποιώντας τα φασματικά χαρακτηριστικά κρίνεται ακατάλληλος.

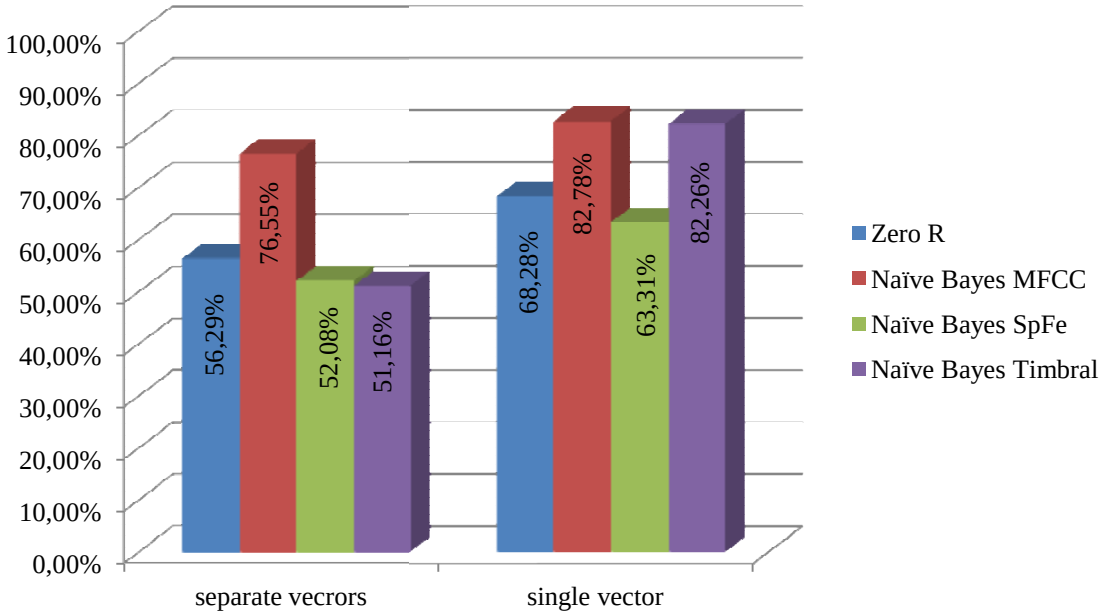
Με την χρήση του αλγορίθμου SMO που ανήκει στην κατηγορία των support vector machines (SVM) παίρνουμε τα καλύτερα αποτελέσματα επιτυχούς ταξινόμησης, με τα ποσοστά επιτυχίας να φτάνουν το 76 % (Πίνακας 3.10.) με την χρήση των φασματικών χαρακτηριστικών, το 87,82% (Πίνακας 3.9.) για τα χαρακτηριστικά MFCC, ενώ με την χρήση των timbral χαρακτηριστικών το ποσοστό αυτό φτάνει το 88,54% (Πίνακας 3.11.), το οποίο είναι το υψηλότερο συνολικό ποσοστό επιτυχούς ταξινόμησης για όλους τους συνδυασμούς χαρακτηριστικών-αλγορίθμων.

Τέλος, με τον αλγόριθμο C4.5 (J48) τα ποσοστά σωστής κατηγοριοποίησης είναι λίγο χαμηλότερα αλλά αρκετά υψηλά για όλες τις κατηγορίες με υψηλότερα ποσοστά επιτυχίας να πετυχαίνονται στο 83,54% (Πίνακας 3.12.) χρησιμοποιώντας τα MFCC χαρακτηριστικά και στο 83,37% (Πίνακας 3.14.) χρησιμοποιώντας τα timbral χαρακτηριστικά, ενώ το χαμηλότερο ποσοστό επιτυχούς ταξινόμησης παρουσιάζουν τα αποτελέσματα της περίπτωσης όπου χρησιμοποιούμε τα φασματικά χαρακτηριστικά και είναι μόλις 75,55% (Πίνακας 3.13.).

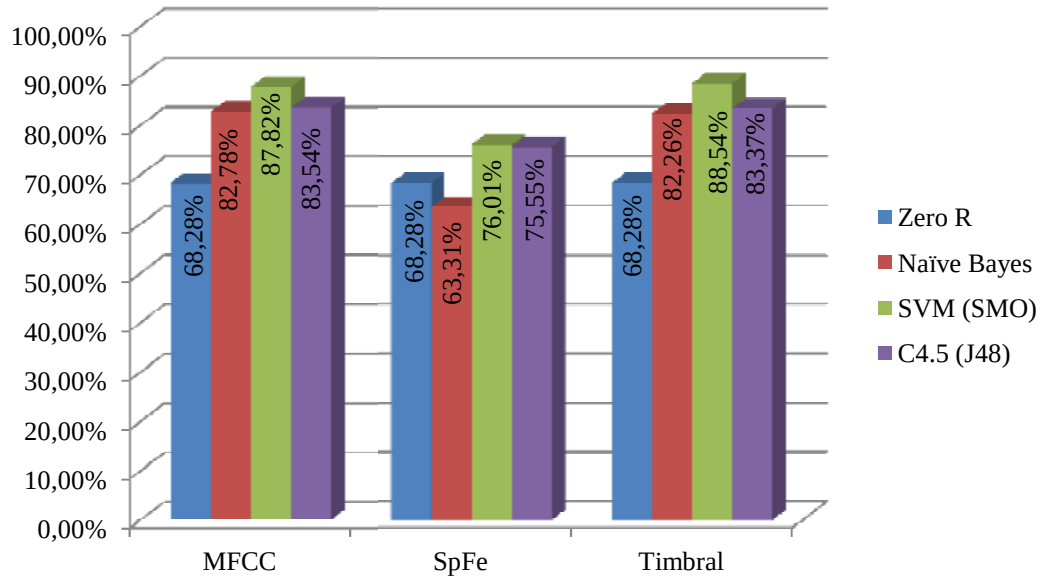
Μεγάλο ενδιαφέρον παρουσιάζουν τα αποτελέσματα του αλγορίθμου SMO, όταν βασίζονται στα φασματικά χαρακτηριστικά (Πίνακας 3.10.), αφού φαίνεται να παραλείπουν την κατηγορία mixed, και ταξινομούν όλες τις παρατηρήσεις στις κατηγορίες Speech και Music. Σε αυτή την περίπτωση, έχουμε και το μεγαλύτερο ποσοστό επιτυχούς ταξινόμησης και για

τις κατηγορίες speech ενώ και για τις δύο κατηγορίες το μεγαλύτερο ποσοστό επιτυχούς ταξινόμησης συναντάται πάλι με την χρήση του αλγορίθμου SMO, σε συνδυασμό με την εξαγωγή των timbral χαρακτηριστικών (Πίνακας 11) και μας δίνει ποσοστά σωστής πρόβλεψης 95,6% για την μουσική και 94,1 για την ομιλία.

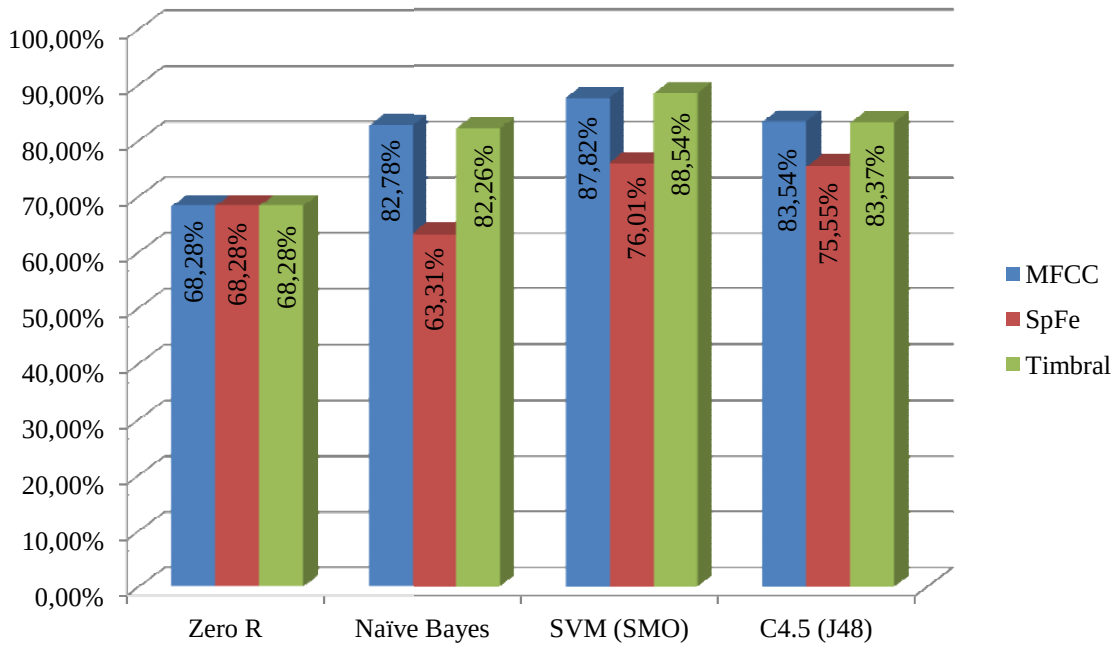
Παρακάτω παρουσιάζονται συνοπτικά στους πίνακες τα ποσοστά επιτυχούς ταξινόμησης όπως περιγράφηκαν παραπάνω.



Διάγραμμα 4.1. Σύγκριση αποτελεσμάτων με την χρήση της εντολής single vector και χωρίς.



Διάγραμμα 4.2. Σύγκριση αποτελεσμάτων επιτυχούς ταξινόμησης με την χρήση διαφόρων συνόλων χαρακτηριστικών (single vector)



Διαγραμμα 4.3. Συγκεντρωτικά ποσοστά επιτυχούς ταξινόμησης για όλους τους αλγορίθμους (single vector)

ΒΙΒΛΙΟΓΡΑΦΙΑ

A.V. Oppenheim, R.W. Schafer. (1999). *Discrete-Time Signal Processing* (2nd εκδ.). Upper Saddle River, New Jersey: Prentice - Hall Inc.

Antti Eronen, Anssi Klapuri. (2000). Musical instrument recognition using cepstral coefficients and temporal features. *IEEE International Conference : Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. 2*, σσ. 753 - 756. IEEE.

Apple Siri. (n.d.). Ανάκτηση από http://www.apple.com/ios/siri/

B. Bigot, I. Ferrane, J. Pinquier. (2010). Exploiting Speaker Segmentation for Automatic Role Detection. An Application to Broadcast News. *2010 International Workshop on Content - Based Multimedia Indexing (CBMI)*, (σσ. 1 - 6).

B. Clarkson, N. Sawhney, A. Pentland. (1998). Auditory Context Awareness via Wearable Computing. *Workshop on Perceptual User In-terfaces*.

B. Schuller, G. Rigoll, M. Lang. (May 2004). Speech Emotion Recognition Combining Acoustic Feature and Language Information in Hybrid Support Vector Machine - Brief Network Architecture. *IEEE Internetional Conference on Acoustics, Speech and Signal Processing, ICASSP '04*, *1*, σσ. 577-580.

BBC. (1985). *BBC 1985: Annual Report and Handbook.* Paperback Edition BBC.

BBC News. (n.d.). Ανάκτηση από http://www.bbc.co.uk/podcasts/series/globalnews

C. Cortes, V. Vapnik. (1995). Machine Learning: Support Vector Networks. σσ. 273 - 297.

C. H. Wu, C. H. Hsieh. (2009). Story Segmentation and Topic Classification of Broadcast News via a Topic - Based Segmental Model and Genetic Algorithm. *IEEE Transactions on Audio, Speech and Language Processings, vol 17, No 8*, σσ. 1612 - 1623.

C.V. Wright, L. Ballard, F. Monrose, G.M. Manson. (2007). Language Identification of Encrypted VoIP Traffic: Alejandra y Roberto or Alice and Bob? *16th USENIX Security Symposium , SS '07.*

Casey, M. (2011, August). General Sound Classification and Similarity in MPEG-7. *Organised Sound Archive, 6*(2), 153-164.

Cettolo, M. (2000). Segmentation, Classification and Clustering of an Italian Broadcast News Corpus. *RIAO*, (σσ. 372-381). Paris, France.

Cortana Windows. (n.d.). Ανάκτηση από https://support.microsoft.com/en-us/help/17214/windows-10-what-is

Crisell, A. (1994). *Understanding Radio*. London: Routledge.

Cue - me. (n.d.). Ανάκτηση από http://www.openstream.com/cueme.html

D. FitzGerald, E. Coyle, B. Lawlor. (2002). Sub-band Independent Subspace Analysis for Drum Transcription. *DAFX'02*, σσ. 65 – 69.

D. Howard, J. A. (2009). *Acoustics and Psychoacoustics.* (4. edition, Επιμ.) Focal Press.

Delphine, C. (14-19 March 2010). Model-Free Anchor Speaker Turn Detection for Automatic Chapter Generation in Broadcast News. *2010 IEEE International Conference on Acoustics Speech and Signal Processing ICASSP*, (σσ. 4966-4969).

Dragon Medical Practice Edition. (n.d.). Ανάκτηση από http://www.nuance.com/products/dragon-medical-practice-edition/index.htm

E. Dogan, M Sert, A. Yazici. (20-25 July 2009). Content-Based Classification and Segmentation of Mixed-Type Audio by Using MPEG-7 Features. *2009 First International Conference on Advances in Multimedia MMEDIA '09*, (σσ. 152-157).

E. Dogan, M.Sert, A. Yazici. (20-25 July 2009). Content-Based Classification and Segmentation of Mixed-Type Audio by Using MPEG-7 Features. *First International Conference on Advances in Multimedia MMEDIA '09*, (σσ. 152-157).

Eurofighter Typhoon. (n.d.). Ανάκτηση από https://www.eurofighter.com/

Fast Fourier Transform. (n.d.). Ανάκτηση από http://mathworld.wolfram.com/FastFourierTransform.html

Fluenz. (n.d.). Ανάκτηση από http://www.fluenz.com/

G. Damnati, D. Charlet. (2011). Robust Speaker Turn Role Labeling of TV Broadcast News Shows. *22011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2011)*, (σσ. 5684 - 5687).

H.D. Tran, H. Li. (2011, August). Sound Event Recognition with Probabilistic Distance SVMs. *IEEE Transactions on Audio, Speech and Language Processing, 19*(6), σσ. 1556-1568.

I. Mporas, T. Ganchev, O. Kicsis, N. Fakotakis. (2011, August). Contex-Adaptive Pre-Processing Scheme for Robust Speech Recognition in Fast-Varying Noise Environment. *Signal Processing, 91*(8), σσ. 2101-2011.

J. Huang, Y. Dong, J. Liu, C. Dong, H. Wang. (2009). Sports Audio Segmentation and Classification. *2009 IEEE International Conference on Network Infrastructure and Digital Content, IC-NIDC 2009*, (σσ. 379 - 383).

J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, D. Caseiro. (2008). Broadcast News Subtitling Systemin Portuguese. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, (σσ. 1561-1564).

J. Sohn, N. S. Kim, W. Sung. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters, 6*(1), σσ. 1 - 3.

J. Zhang, B. Jiang, L. Lu, Q. Zhao. (2010). Audio Segmentation System for Sport Games. *2010 International Conference on Electrical and Control Engineering (ICECE 2010)*, (σσ. 505 - 508).

Jingles. (n.d.). Ανάκτηση από https://en.wikipedia.org/wiki/Jingle

L. Xie, Y. Yang, Z. Q. Liu, W. Feng, Z. Liu. (2010). Interating Acoustic and Lexical Features in Topic Segmentation of Chinese Broadcast News Using Maximum Entropy. *2010 International Conference on Audio Language and Image Processing (ICALIP2010)*, (σσ. 407 - 413).

M. Baillie, J.M. Jose. (2004). An Audio - based Sports Video Segmentation and Event Detection Algorithm. *Conference on Computer Vision and Pattern Recognition Workshop, CVPRW '04*, (σ. 110).

M.Kotti, V. Moschou, C. Kotropoulos. (2008, May). Speaker Segmentation and Clustering. *Signal Processing, Volume 88, Issue 5*, σσ. 1091-1124.

Mark Levy, Mark Sandler. (2008). Structural Segmentation of Musical Audio by Constrained Clustering. *IEEE Transactions on Audio, Speech and Language Processing, 16*(2).

McKay, C. (2005). Automatic music classification and similarity analysis, Course Paper. Université de Montreal, Canada.

N. Liu, Y. Zhao, Z. Zhu, H. Lu. (2011, October). Exploiting Visual-Audio-Textual Characteristics for Automatic TV Commercial Block Detection and Segmentation. *IEEE Transactions on Multimedia, 13*(5), σσ. 961-973.

Peeters, G. (2004). *A large se of audio features for sound description (similarity and classification) in the CUIDADO project, Technical report.* IRCAM.

Praat Manual. (n.d.). Ανάκτηση από http://www.fon.hum.uva.nl/praat/

Radio Jingles - Διαφημιστικά Τραγούδια. (n.d.). Ανάκτηση από Gold Creative Studio: http://www.goldcs.gr/B7723F30.el.aspx

Rubidium. (n.d.). Ανάκτηση από http://www.rubidium.com/

S. Matos, S.S. Birring, I.D. Pavord, D.H. Evans. (2006). Detection of Cough Signals in Continuous Audio Recordings Using Hidden Markov Models. *IEEE Transactions on Biomedical Engineering, Volume: 53, No 6*, σσ. 1078 - 1083.

S. Petridis, M. Pantic, J.F. Cohn. (2011). Prediction-Based Classification for Audiovisual Discrimination Between Laughter and Speech. *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, (σσ. 619 - 626).

Samsung S Voice. (n.d.). Ανάκτηση από http://www.samsung.com/global/galaxy/what-is/s-voice/

Shazam. (n.d.). Ανάκτηση από http://www.shazam.com/

signal-to-noise ratio S/N or SNR. (n.d.). Ανάκτηση από Search Networking: http://searchnetworking.techtarget.com/definition/signal-to-noise-ratio

Speechnotes. (n.d.). Ανάκτηση από https://speechnotes.co/

T. Drugman, J. Urbain, T. Dutoit. (2011). Assessment of Audio Features for Automatic Cough Detection. *19th European Signal Processing Conference (EUSIPCO11).* Barcelona, Spain.

T. Lambrou, P. Kudumakis, M. Sandler, R. Speller & A. Linney. (1998). Classification of audio signals using statistical features on time and wavelet transform domains. *IEEE ICASSP*.

T. Mikami, Y. Kojima, M. Yamamoto, M. Furukawa. (2012). Automatic Classification of Oral / Nasal / Snoring Sounds based on the Acoustic Properties. *2012 IEEE International Conference on Acoustics, Speech and Signal Processings (ICASSP 2012)*, (σσ. 609 - 612).

T. Pohle, E. Pampalk, G. Widmer. (2005). Evaluation of Frequently Used Audio Features for Classification of Music Into Perceptual Categories. *Fourth International Workshop on Content-Based Multimedia Indexing, CBMI '05.*

T.Bouko, C. Nadeu. (volume 2011). Audio Recognition of Broadcast News in the Albayzin-2010 Evaluation: Overview, Results and Discussion. *EURASIP Journal on Audio, Speech and Music 2011*, issue 1.

Talks & Zooms. (n.d.). Ανάκτηση από http://www.nuance.com/for-individuals/mobile-applications/talks-zooms/index.htm

Tell me More. (n.d.). Ανάκτηση από http://www.effectivelanguagelearning.com/language-course-reviews/tell-me-more-review

Theodoros Theodorou, Iosif Mporas, Nikos Fakotakis. (2012). Automatic Sound Classification of Radio Broadcast News. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 5*(1), 37 - 47.

Track ID - Sony . (n.d.). Ανάκτηση από https://trackid.sonymobile.com/

Tzanetakis, G. (2009). Marsyas Submissions to Mirex 2009 . *Music Information Retrieval Evaluation eXchange (MIREX).*

Tzanetakis, G. (2009). Music Analysis, Retrieval and Synthesis of Audio Signals MARSYAS. *17th ACM international conference on Multimedia. ACM.*

Tzanetakis, G. (n.d.). *Marsyas User Manual.* Ανάκτηση από http://www.marsyas.info/pdf/Marsyas0.2_UserManual.pdf

Vapnik, V. N. (1998). *Statistical Learning Theory.* New York: John Wiley & Sons.

W. Wang, S. Yaman, K. Precoda, C. Richey. (2011). Automatic Identification of Speaker Role and Agreement / Disagreement in Broadcast Conversation. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2011)*, (σσ. 5556 - 5559).

Wang D., B. G. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications.* Wiley Blackwell Publishers.

Weka 3: Data Mining Software in Java. (n.d.). Ανάκτηση από Machine Learning Group at the University of Waikato: http://www.cs.waikato.ac.nz/ml/weka/index.html

Wolfram Mathworld. (n.d.). Ανάκτηση από Fast Fourier Transform: http://mathworld.wolfram.com/FastFourierTransform.html

Y. Itoh, S. Sakaki, K. Kojima, M. Ishigame. (2008). Highlight Scene Extraction of Sports Broadcasts Using Sports News Programs. *2008 IEEE 10th Workshop on Multimedia Signal Processing (MMSP 2008)*, (σσ. 646 - 649).

Y.D. Cho, A. Kondoz,. (2001). Analysis and improvement of a statistical model-based voice activity detector. *IEEE Signal Processing Letters, 8*, σσ. 276 – 278.

Z. Sun, A. Purohit, K. Yang, N. Pattan, D. Siewiorek, A. Spailagic, I. Lane, P. Zhang. (2011). CoughLoc: Location Aware Indoor Acoustic Sensing for Non-Intrusive Cough Detection. *International Workshop on Emerging Mobile Sensing Technologies, Systems and Application. Mobisense.* San Francisco, CA.

ZeroR. (n.d.). Ανάκτηση από http://www.saedsayad.com/zeror.htm

Ζαρβαδάς, Ι. (2013). *Αυτόματη Κατηγοριοποίηση ειδών Κρητικής Μουσικής με Χρήση Μεθόδων Μηχανικής Μάθησης.* Ρέθυμνο.

Θεοδωρίδης, Α. (2011). Ανάπτυξη Τεχνικών Επεξεργασίας Ήχου-Ομιλίας στο MATLAB και PRAAT. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Τμήμα Πληροφορικής.

Κωτσιαντής, Σ. (2005). *Ομάδες Ταξινομητών για την Αύξηση της Ακρίβειας των Μεθόδων Μηχανικής Μάθησης και Εξόρυξη Γνώσης, Δοδακτορική Διατριβή*. Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών.

Μαούτσα, Δ.-Δ. (2013). *Μελέτη Μουσικής Ομοιότητας με Χρήση Ευφυών Συστημάτων*. Εθνικό Μετσόβειο Πολυτεχνείο.

Μπουρνά, Β. (2008). *Τεμαχιοποίηση Ομιλίας σε Φωνητικές Ομάδες για Αναγνώριση και Σύνθεση Συναισθηματικής Ομιλίας, Διπλωματική Εργασία.* Πάτρα: Πανεπιστήμιο

Πατρών, Πολυτεχνική Σχολή, Τμήμα Ημεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών.

Νταλαμπίρας, Σ. (2010, Ιούνιος). *Ψηφιακή Επεξεργασία και Αυτόματη Κατηγοριοποίηση Περιβαλλοντικών Ήχων*. Πάτρα.

Π.Ζερβάς. (2007). *Μοντελοποίηση και Ψηφιακή Επεξεργασία Προσωδιακών Φαινομένων της Ελληνικής Γλώσσας με Εφαρμογή στην Σύνθεση Ομιλίας, Διδακτορική Διατριβή*. Πολυτεχνική Σχολή Πανεπιστημίου Πατρών, Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών.

Τσινίκας, Ν. Π. (2005). *Ακουστικός Σχεδιασμός Χώρων: Θέατρα, Αμφιθέατρα, Αίθουσες Συναυλιών, χώροι πολλαπλών χρήσεων* (Τόμ. 2η έκδοση). Θεσσαλονίκη: University Studio Press.

ΠΑΡΑΡΤΗΜΑΤΑ

ΠΑΡΑΡΤΗΜΑ Α

DATASET – Ραδιοφωνικές εκπομπές (table of used audio files)

Title	File Name	Κανάλι Προέλευσης	Είδος Εκπομπής	Total Duration	Format Ήχου
AM: Dozens dead in Paris attacks	GlobalNewsPodcast-20151114-AMDozensDeadInParisAttacks.mp3	Global News	Ενημερωτική Ειδήσεις	00:33:58	.mp3
Bomb threat cancels German football match	GlobalNewsPodcast-20151118-BombThreatCancelsGerman.mp3	Global News	Ενημερωτική Ειδήσεις	00:27:11	.mp3
PM: Life under Brussels lockdown	GlobalNewsPodcast-20151123-PMLifeUnderBrusselsLock.mp3	Global News	Ενημερωτική Ειδήσεις	00:29:11	.mp3
Screen6: With Spike Jonze	20141122-WithSpikeJonze.mp3	6 Music	Ψυχαγωγική	00:28:38	.mp3
Screen6: With The Coen Brothers	20141122-WithTheCoenBrothers.mp3	6 Music	Ψυχαγωγική	00:28:46	.mp3
Screen 6 : With Joe Wright	Screen6WithEdithBowman-20151212-WithJoeWrigh.mp3	6 Music	Ψυχαγωγική	00:28:25	.mp3
Screen 6 : With Sam Mendes	Screen6WithEdithBowman-20151212-WithSamMendes.mp3	6 Music	Ψυχαγωγική	00:33:39	.mp3

ΠΑΡΑΡΤΗΜΑ Β

ΣΥΝΟΛΙΚΗ ΔΙΑΡΚΕΙΑ MUSIC/SPEECH ΑΝΑ ΑΡΧΕΙΟ ΗΧΟΥ

Title	Είδος Εκπομπής	Music Duration	Speech Duration	Mixed Duration	Total Duration
AM: Dozens dead in Paris attacks	Ενημερωτική Ειδήσεις	00:00:11	00:20:33	00:00:31	**00:33:58**
Bomb threat cancels German football match	Ενημερωτική Ειδήσεις	00:00:10	00:22:14	00:00:08	**00:27:11**
PM: Life under Brussels lockdown	Ενημερωτική Ειδήσεις	00:00:31	00:20:07	00:00:25	**00:29:15**
Screen6: With Spike Jonze	Ψυχαγωγική	00:06:09	00:11:37	00:06:48	**00:28:38**
Screen6: With The Coen Brothers	Ψυχαγωγική	00:05:32	00:16:40	00:03:16	**00:28:46**
Screen 6 : With Joe Wright	Ψυχαγωγική	00:13:26	00:01:41	00:10:52	**00:28:25**
Screen 6 : With Sam Mendes	Ψυχαγωγική	00:14:19	00:03:54	00:13:25	**00:33:39**
ΣΥΝΟΛΟ					**03:29:48**

ΠΑΡΑΡΤΗΜΑ Γ

ΚΩΔΙΚΑΣ ΕΞΑΓΩΓΗΣ ΕΠΙΜΕΡΟΥΣ .WAV ΑΡΧΕΙΩΝ ΑΠΟ ΤΟ ΑΡΧΙΚΟ ΣΗΜΑ (Praat Script)

```
form Save intervals to small WAV sound files
     comment Which IntervalTier in the TextGrid would you like to
process?
     integer Tier 1
     comment Starting and ending at which interval of the
TextGrid?
     integer Start_from 1
               comment Exclude some labels with mistakes or emply
     integer End_at_(0=last) 0
     boolean Exclude_empty_labels 1
     boolean Exclude_intervals_starting_with_dot_(.) 1
     comment Give a small margin for the files if you like:
     positive Margin_(seconds) 0.01
     comment Give the folder where to save the sound files:
     sentence Folder C:\Users\Anastasia
Catela\Desktop\Thesis\sounds thesis\BBC Global News\All
     comment Give an optional prefix for all filenames:
     sentence Prefix
     comment Give an optional suffix for all filenames (.wav will
be added anyway):
     sentence Suffix
endform

gridname$ = selected$ ("TextGrid", 1)
soundname$ = selected$ ("LongSound", 1)
select TextGrid 'gridname$'
numberOfIntervals = Get number of intervals... tier

if start_from > numberOfIntervals
     exit There are not that many intervals in the IntervalTier!
endif

if end_at > numberOfIntervals
     end_at = numberOfIntervals
endif

if end_at = 0
     end_at = numberOfIntervals
endif

# Default values for variables
files = 0
intervalstart = 0
intervalend = 0
```

```
interval = 1
intname$ = ""
intervalfile$ = ""
endoffile = Get finishing time


# ask if the user wants to go through with saving all the files:
for interval from start_from to end_at
      xxx$ = Get label of interval... tier interval
      check = 0
      if xxx$ = "" and exclude_empty_labels = 1
            check = 1
      endif
      if left$ (xxx$,1) = "." and
exclude_intervals_starting_with_dot = 1                        check
= 1
      endif
      if check = 0
               files = files + 1
      endif
endfor

interval = 1
pause 'files' sound files will be saved. Continue?

# Loop through all intervals in the selected tier of the TextGrid
for interval from start_from to end_at
      select TextGrid 'gridname$'
      intname$ = ""
      intname$ = Get label of interval... tier interval
      check = 0
      if intname$ = "" and exclude_empty_labels = 1
            check = 1
      endif
      if left$ (intname$,1) = "." and
exclude_intervals_starting_with_dot = 1
            check = 1
      endif
      if check = 0
            intervalstart = Get starting point... tier interval
                  if intervalstart > margin
                        intervalstart = intervalstart – margin
                  else
                        intervalstart = 0
                  endif

            intervalend = Get end point... tier interval
```

```
                if intervalend < endoffile – margin
                      intervalend = intervalend + margin
                else
                      intervalend = endoffile
                endif


           select LongSound 'soundname$'
           Extract part... intervalstart intervalend no
           filename$ = intname$
           intervalfile$ = "'prefix$'" + "'filename$'" +
"'suffix$'" + ".wav"
           indexnumber = 0
           while fileReadable (intervalfile$)
                indexnumber = indexnumber + 1
                intervalfile$ = "'prefix$'" + "'filename$'" +
"'suffix$''indexnumber'" + ".wav"
           endwhile
           Write to WAV file... 'intervalfile$'
           Remove
     endif
endfor
```

ΠΑΡΑΡΤΗΜΑ Δ

ΚΩΔΙΚΑΣ ΓΙΑ ΧΡΗΣΗ MARSYAS ΚΑΙ ΕΞΑΓΩΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ (FEATURE EXTRACTION)

```
### μπαίνω στο φάκελο του Marsyas

 cd C:\Users\Anastasia Catela\Desktop\Thesis\marsyas-0.5.0\bin

### τρέχω το marsyas-run

 marsyas-run

### εμφανίζει
### Usage: marsyas-run <plugin-file> [options]

### δημιουργώ την συλλογή speech (προσοχή στο Path να είναι ενωμένα)

  mkcollection -c speech.mf -l speech C:\Users\Public\Speech_all

### δημιουργώ την συλλογή speech

 mkcollection -c music.mf -l music C:\Users\Public\Music_all

### δημιουργία  της συλλογής mixed που εμπεριέχει τα τμήματα
### 'speech over music'

 mkcollection –c mixed.mf –l mixed C:\Users\Public\Speech_over_Music

### εξάγω τα timbral, spectral και MFCC χαρακτηριστικά των συλλογών music,
### speech και mixed σε ένα αρχείο .arff που διαβαζεται από το weka

 bextract –timbral music.mf speech.mf –mixed.mf -w timbral.arff
  bextract –spfe music.mf speech.mf –mixed.mf -w spfe.arff
  bextract –mfcc music.mf speech.mf mixed.mf –w mfcc.arff

### εξάγω τα timbral, spectral και MFCC χαρακτηριστικά των συλλογών music,
### speech και mixed σε ένα αρχείο .arff χρησιμοποιώντας single vector

 bextract –sv –timbral music.mf speech.mf mixed.mf -w timbral_sv.arff
  bextract –sv –spfe music.mf speech.mf mixed.mf -w spfe_sv.arff
  bextract –sv –mfcc music.mf speech.mf mixed.mf –w mfcc_sv.arff
```