ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΜΜΕ

**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ**

# «ΑΝΑΠΤΥΞΗ ΣΥΣΤΗΜΑΤΟΣ ΑΥΤΟΜΑΤΟΥ ΤΕΜΑΧΙΣΜΟΥ ΚΑΙ ΑΝΑΓΝΩΡΙΣΗΣ ΗΧΗΤΙΚΩΝ ΚΑΤΗΓΟΡΙΩΝ ΑΠΟ ΡΑΔΙΟΦΩΝΙΚΕΣ ΕΚΠΟΜΠΕΣ»

**ΕΥΣΤΑΘΙΟΥ ΓΕΩΡΓΙΟΣ-ΠΑΝΑΓΙΩΤΗΣ**

**ΕΠΟΠΤΕΥΩΝ ΚΑΘΗΓΗΤΗΣ: ΚΟΥΤΡΑΣ ΑΘΑΝΑΣΙΟΣ**

**ΠΥΡΓΟΣ, 2018**

# ΠΙΣΤΟΠΟΙΗΣΗ

Πιστοποιείται ότι η πτυχιακή εργασία με θέμα:

**«ΑΝΑΠΤΥΞΗ ΣΥΣΤΗΜΑΤΟΣ ΑΥΤΟΜΑΤΟΥ ΤΕΜΑΧΙΣΜΟΥ ΚΑΙ ΑΝΑΓΝΩΡΙΣΗΣ ΗΧΗΤΙΚΩΝ ΚΑΤΗΓΟΡΙΩΝ ΑΠΌ ΡΑΔΙΟΦΩΝΙΚΕΣ ΕΚΠΟΜΠΕΣ»**

του φοιτητή του Τμήματος ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΜΜΕ

**ΕΥΣΤΑΘΙΟΥ ΓΕΩΡΓΙΟΥ-ΠΑΝΑΓΙΩΤΗ**

**Α.Μ.: 0552**

παρουσιάστηκε δημόσια και εξετάσθηκε στο Τμήμα ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΜΜΕ στις

15/10 / 2018

Ο ΕΠΙΒΛΕΠΩΝ — Ο ΠΡΟΕΔΡΟΣ ΤΟΥ ΤΜΗΜΑΤΟΣ

ΚΟΥΤΡΑΣ ΑΘΑΝΑΣΙΟΣ — ΚΟΥΓΙΑΣ ΙΩΑΝΝΗΣ

# ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΜΗ ΛΟΓΟΚΛΟΠΗΣ

Βεβαιώνω ότι είμαι 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 όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Ακόμα δηλώνω ότι αυτή η γραπτή εργασία προετοιμάστηκε από εμένα προσωπικά και αποκλειστικά και ειδικά για την συγκεκριμένη πτυχιακή εργασία και ότι θα αναλάβω πλήρως τις συνέπειες εάν η εργασία αυτή αποδειχθεί ότι δεν μου ανήκει.

| **ΟΝΟΜΑΤΕΠΩΝΥΜΟ ΣΠΟΥΔΑΣΤΗ** | **ΑΜ** | **ΥΠΟΓΡΑΦΗ** |
|---|---|---|
| ΕΥΣΤΑΘΙΟΥ ΓΕΩΡΓΙΟΣ-ΠΑΝΑΓΙΩΤΗΣ | 0552 |  |

# ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τους γονείς μου για τη στήριξή τους καθ'όλη τη διάρκεια των σπουδών μου, τον επιβλέποντα καθηγητή μου κ.Κούτρα Αθανάσιο για την συνεργασία, την εμπιστοσύνη που μου έδειξε και την πλήρη καθοδήγησή του.

Τέλος, θα ήθελα να ευχαριστήσω τη σύντροφό μου που με στήριξε και με ενθάρρυνε με σκοπό την ολοκλήρωση των σπουδών και της πτυχιακής μου εργασίας.

# ΠΡΟΛΟΓΟΣ

Η παρούσα Πτυχιακή Εργασία με τίτλο «Ανάπτυξη συστήματος τεμαχισμού αναγνώρισης ηχητικών κατηγοριων από ραδιοφωνικές εκπομπές» εκπονήθηκε στα πλαίσια της ολοκλήρωσης των προϋποθέσεων, για τη λήψη του πτυχίου μου από το ΤΕΙ Δυτικής Ελλάδας, Τμήμα Πληροφορικής & ΜΜΕ με έδρα τον Πύργο Ηλείας, με επιβλέποντα καθηγητή τον κ.Κούτρα Αθανάσιο.

# ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσας πτυχιακής εργασίας είναι η ανάπτυξη ενός συστήματος τεμαχισμού και αναγνώρισης ηχητικών κατηγοριών από ραδιοφωνικές εκπομπές.

Συγκεκριμένα, χρησιμοποιήθηκαν εκπομπές-δελτία ειδήσεων από τη Φωνή της Αμερικής (Voice of America), οι οποίες κατατμήθηκαν χειροκίνητα μέσω του εργαλείου PRAAT και αποτελούν τη βάση δεδομένων που χρησιμοποιήθηκε.

Ακολούθως, έγινε η εξαγωγή χαρακτηριστικών μέσω του MARSYAS και κατηγοριοποίηση μέσω WEKA. Ο αλγόριθμος που χρησιμοποιήθηκε ήταν οι Support Vector Machines (SVM), ο οποίος αποδείχθηκε αποτελεσματικός, ενώ πραγματοποιήθηκε εξαγωγή παραμέτρων, με τους Mel Frequency Cepstral Coefficients (MFCC) να δίνουν τα υψηλότερα ποσοστά στην κατηγοριοποίηση των κλάσεων. Έγινε εκτενής ανάλυση των μεθόδων που χρησιμοποιήθηκαν, τόσο στην προεπεξεργασία της βάσης δεδομένων, όσο και για την εξαγωγή των παραμέτρων και στην κατηγοριοποίηση. Στο τέλος, γίνεται αναλυτική παρουσίαση των αποτελεσμάτων που προέκυψαν από την πειραματική διαδικασία με πίνακες και στατιστικά, καθώς και σύγκριση μεταξύ τους, ώστε να προκύψει το τελικό συμπέρασμα.

# ABSTRACT

The aim of this thesis is to develop a system for the fragmentation and recognition of sound sequences from radio broadcasts.
In particular, Voice of America news broadcasts were used, which were manually split through the PRAAT tool and used as the database used.
Subsequently, attributes were extracted through MARSYAS and categorized by WEKA software. The algorithm used was Support Vector Machines (SVM), which proved to be effective, while parameter extraction was performed, with Mel Frequency Cepstral Coefficients (MFCCs) giving the highest rankings in class categorization. An extensive analysis was made of the methods used both in the pre-processing of the database, as well as in the extraction of the parameters and in the categorization.Finally, the results of the experimental process are presented with tables and statistics, as well as a comparison between them, in order to reach the final conclusion.

# ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

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

ΠΕΡΙΕΧΟΜΕΝΑ

# ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

# ΕΙΣΑΓΩΓΗ

Η παρούσα Πτυχιακή Εργασία περιγράφει τη διαδικασία σύμφωνα με την οποία επεξηγείται το πως μπορούν να απομονωθούν τα δομικά στοιχεία του ήχου που εκπέμπεται σε μια ραδιοφωνική εκπομπή. Περιγράφεται η διαδικασία της κατάτμησής τους, μέσω πειραματικής διαδικασίας, ώστε να είναι δυνατή η περαιτέρω χρήση και αξιοποίηση των αποτελεσμάτων αυτών στην επιστημονική έρευνα γύρω από τα ζητήματα της κατάμησης ήχου και η εκπαίδευση συστημάτων που θα μπορούν να πραγματοποιήσουν την ίδια διαδικασία πλέον αυτόματα.

Η διαδικασία της χειροκίνητης κατάτμησης ήχου αποσκοπεί στην δημιουργία αρχείων προς επεξεργασία, έτσι ώστε ο εκάστοτε χρήστης, με τη χρήση ειδικού λογισμικού να πραγματοποιήσει εκμάθηση σε κάποιο υπολογιστικό σύστημα, προκειμένου αυτό με τη σειρά του να μπορεί να πραγματοποιεί αυτόματη κατάτμηση ήχου από ραδιοφωνικό περιβάλλον.

Ο Ηλεκτρονικός Υπολογιστής θα είναι σε θέση να μπορεί να αντιλαμβάνεται πλέον ξεχωριστά ποια τμήματα του ήχου είναι μουσική, ομιλία, θόρυβος, σιωπή ή και συνδυασμός αυτών.

# 1 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

## 1.1 ΕΠΙΣΚΟΠΗΣΗ ΔΙΕΘΝΟΥΣ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Στην περίπτωση των ραδιοφωνικών εκπομπών και ειδικά στην κατηγορία των **broadcast news**, που αποτελούν μια ξεχωριστή κατηγορία στην επεξεργασία σήματος στον ήχο, με σκοπό την ανάλυσή του και την κατηγοριοποίησή του, έχουν υπάρξει αρκετές μελέτες σε παγκόσμιο επίπεδο στο παρελθόν, οι οποίες σχετίζονται με την αναγνώριση ήχων **[** (E . Dogan, M. Sert, & A. Yazici, 20-25 July 2009)**;** (Bouko & Nadeu, volume 2011)**]**, την αναγνώριση του βασικού εκφωνητή-anchor speaker (Delphine, 14-19 March 2010), την αναγνώριση ρόλου-Role Detection (B. Bigot, I. Ferrane, & J. Pinquer, 2010), την αυτόματη τμηματοποίηση και κατηγοριοποίηση ήχου σε αθλητικές εκπομπές και την εξαγωγή κορυφαίων στιγμιοτύπων από αθλητικές μεταδόσεις [ (J. Huang, Y. Dong, J. Liu, C. Dong, & H. Wang, 2009)**;** (Y. Itoh, S. Sakaki, K. Kojima, & M. Ishigame, 2008)**]**

## 1.2 ΑΥΤΟΜΑΤΗ ΑΝΑΓΝΩΡΙΣΗ ΗΧΩΝ

Το ερευνητικό πεδίο της αναγνώρισης του ακουστικού σήματος στοχεύει στην ανάλυση του περιβάλλοντα χώρου, την καταμέτρηση, τον διαχωρισμό και την αναγνώριση των ηχητικών πηγών, χρησιμοποιώντας μόνο το εισερχόμενο ακουστικό σήμα, το οποίο προέρχεται από διάφορες πηγές.( (Νταλαμπίρας & Νταλαμπίρας, Ιούνιος 2010)).

Το πεδίο της αυτόματης αναγνώρισης ήχων καλύπτει ένα ευρύ φάσμα εφαρμογών, καθώς έχει χρησιμοποιηθεί με σκοπό να καλύψει ένα μεγάλο πλήθος αναγκών, όπως:

- Εφαρμογή σε χρήστες που αντιμετωπίζουν προβλήματα όρασης (πχ. (Nuance)[1])
- Αναγνώριση μουσικών κομματιών μέσω εφαρμογών (πχ. (TrackID)[2], (Shazam)[3])
- Εφαρμογή υπαγόρευσης και καταγραφής προφορικού λόγου σε κείμενο (πχ. (Speechnotes)[4])
- Εφαρμογή σε συστήματα ασφαλείας που χρησιμοποιούν την αυτόματη αναγνώριση ήχων με σκοπό τον προσδιορισμό της ταυτότητας του ομιλούντος από τη φωνή (πχ. (Rubidium)[5])
- Εφαρμογή σε συστήματα τηλεφωνίας, όπου ο χρήστης μπορεί να πλοηγηθεί στο μενού μέσω της φωνής, χωρίς την χρήση πλήκτρων (πχ. (Cue-me)[6])

[1] http://www.nuance.com/for-individuals/mobile-applications/talks-zooms/index.htm
[2] https://trackid.sonymobile.com/
[3] http://www.shazam.com/
[4] https://speechnotes.co/
[5] http://www.rubidium.com/
[6] http://www.openstream.com/cueme.html

## 1.3 Ο ΗΧΟΣ ΣΤΗ ΡΑΔΙΟΦΩΝΙΚΗ ΠΑΡΑΓΩΓΗ

Το ραδιόφωνο αποτελεί ένα μέσο μετάδοσης ηχητικής πληροφορίας μέσω ραδιοκυμάτων. Η συγκεκριμένη πληροφορία, συλλέγεται από όλες τις πηγές εισόδου που υπάρχουν σε ένα ραδιοφωνικό στούντιο, καταλήγοντας στον πομπό και εν συνεχεία στους ραδιοφωνικούς δέκτες ή σε ηλεκτρονικούς υπολογιστές (στην περίπτωση της διαδικτυακής μετάδοσης).

Με όποιο τρόπο και αν μεταδίδεται το ραδιοφωνικό σήμα, αυτό που έχει σημασία είναι το πως δημιουργείται και από που προέρχεται. Σε μια ραδιοφωνική εκπομπή οι πηγές εισόδου και ο ήχος που διέρχεται μέσα από αυτές συνθέτουν το τελικό ηχητικό αποτέλεσμα, το οποίο λαμβάνει ο εκάστοτε ακροατής ως σύνολο.

Τα τμήματα του ήχου που εισέρχονται από τις διαφορετικές πηγές ενώνονται μεταξύ τους δημιουργώντας το τελικό αποτέλεσμα, το οποίο και αναλύεται στην παρούσα εργασία, με σκοπό τον εντοπισμό των επιμέρους κλάσεων και την μετέπειτα επεξεργασία τους.

Συνήθως, ένας ραδιοφωνικός σταθμός είναι εξοπλισμένος με το στούντιο για την ζωντανή μετάδοση, στούντιο ηχογραφήσεων και ενίοτε με ένα βοηθητικό. Κατά πλειονότητα οι περισσότεροι ραδιοφωνικοί σταθμοί διαθέτουν τα δύο πρώτα.

Το πρώτο συνήθως είναι το “on-air” studio, το οποίο χρησιμοποιείται για τις καθημερινές ζωντανές (live) εκπομπές. Το άλλο είναι το στούντιο παραγωγής, το οποίο χρησιμοποιείται για τον προγραμματισμό του ηχητικού υλικού που έχει καταγραφεί για αναπαραγωγή σε μεταγενέστερο χρόνο. Με άλλα λόγια είναι ό,τι δεν μεταδίδεται ζωντανά. Περιλαμβάνει στοιχεία, όπως διαφημίσεις, ανακοινώσεις, ή σποτάκια για την διαφημιστική προώθηση του ραδιοφωνικού σταθμού (promos).

Ο εξοπλισμός που φαίνεται στο σχεδιάγραμμα είναι ο αντιπροσωπευτικός που μπορεί να υπάρχει σε ένα τυπικό στούντιο.

Η εικόνα 1.1 δείχνει έναν απλοποιημένο χάρτη ενός τυπικού στούντιο παραγωγής. Ξεκινώντας με διάφορες πηγές ήχου, όπως τη φωνή του εκφωνητή, ένα CD, ένα audio recorder (εγγραφέας), φάινεται η τελική διαδρομή του ήχου, μέχρι να μεταδοθεί ζωντανά ή να μαγνητοσκοπηθεί. Αυτό συχνά καλείται «ακουστική αλυσίδα» (audio chain), επειδή τα διάφορα κομμάτια του εξοπλισμού συνδέονται μεταξύ τους. (Reese, Gross, & Gross, Έκδοση 5, Αναθεωρημένη 2012)

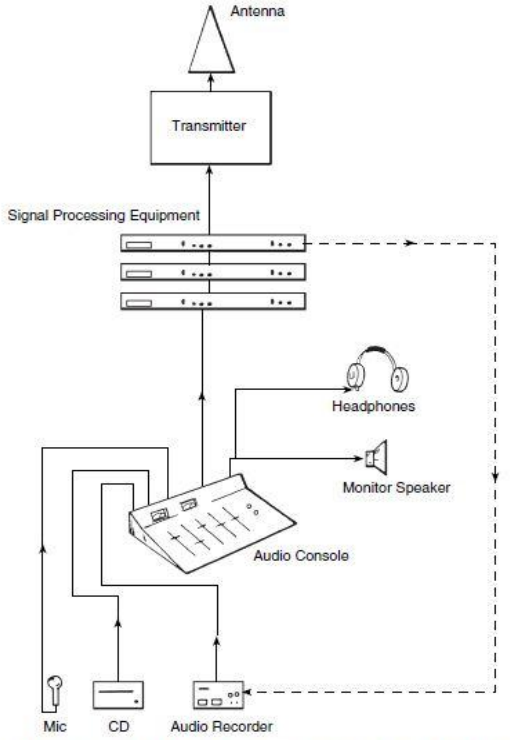



**FIGURE 1.2** The audio chain shows how sound moves through the broadcast equipment that is linked together in the production studio.

Εικόνα 1-1: Η αλυσίδα του ήχου μέσα στο στούντιο.

Το ταξίδι του ήχου εντός του ραδιοφωνικού στούντιο μπορεί να είναι περίπλοκο, δεδομένου ότι ο ήχος μπορεί κάποιες φορές να αλλάξει πορεία. Για παράδειγμα, μπορεί να αντιγραφεί από CD σε MiniDisc ή μπορεί να ισοσταθμιστεί (διαδικασία η οποία είναι μια μορφή επεξεργασίας σήματος).

Οι συμπαγείς γραμμές δείχνουν τον ήχο που αποστέλλεται στην κονσόλα ήχου από μια ποικιλία ηχητικών πηγών. Στη συνέχεια ο ήχος περνάει από τον εξοπλισμό επεξεργασίας σήματος, και τέλος στο σύστημα μετάδοσης.

Ένα μικρόφωνο μετατρέπει τη φωνή του εκφωνητή σε σήμα ήχου. Δεν είναι καθόλου ασυνήθιστο για μια μονάδα παραγωγής να έχει ένα ή παραπάνω βοηθητικά μικρόφωνα για τις εργασίες παραγωγής που απαιτούν δύο ή περισσότερες φωνές.

Ο ήχος που παράγεται στην παραγωγή ραδιοφωνικών εκπομπών, συνήθως αποτελείται από επιμέρους τμήματα, αναλόγως τις πηγές εισόδου του ήχου που χρησιμοποιούνται κάθε φορά, όπως τα μικρόφωνα για την ομιλία ή άλλες πηγές ήχου, όπως CD Players και Ηλεκτρονικοί Υπολογιστές για την αναπαραγωγή μουσικής, ηχητικών εφέ, κλπ.

Οι βασικές κατηγορίες ήχου στο περιβάλλον μιας ραδιοφωνικής εκπομπής (εν προκειμένω σε εκπομπή ειδησεογραφικού χαρακτήρα), είναι κατά κύριο λόγο η ομιλία και η μουσική. Δευτερευόντως, μπορεί να υπάρξουν διαστήματα με σιωπή ή με παραγωγή θορύβου από εξωγενείς παράγοντες που μπορούν να καταγραφούν ενίοτε στην εγγραφή ή την μετάδοση του ραδιοφωνικού σήματος, όπως για παράδειγμα ο ήχος μιας πόρτας, ο ήχος από αυτοκίνητα ή ακόμα και φωνές τρίτων που δεν θα έπρεπε να εμπεριέχονται στο εκπεμπόμενο σήμα.

Αναλυτικά οι κατηγορίες τις οποίες μελετάμε στην παρούσα εργασία και προσπαθούμε μέσω της πειραματικής διαδικασίας που περιγράφεται παρακάτω, να τις κατηγοριοποιήσουμε είναι οι:

- **Μουσική (Music)**
- **Ομιλία (Speech)**
- **Θόρυβος (Noise)**
- **Σιωπή (Silence)**

## Μουσική (Music)

Πρόκειται για μια σημαντική και ιδιαίτερη κατηγορία ήχων που αξιοποιεί η ραδιοφωνία. Η μουσική στις ραδιοφωνικές εκπομπές χρησιμοποιείται ως το βασικό στοιχείο το οποίο παίζει σημαντικό ρόλο, καθώς λειτουργεί παράλληλα με την ομιλία των παρουσιαστών ή των συμμετεχόντων, ως «μουσικό χαλί» κατά τη διάρκεια των εκπομπών, αλλά και ανεξάρτητα ως ο κύριος πυλώνας κάλυψης του ραδιοφωνικού προγράμματος, όπως επίσης και ως γέφυρα μεταξύ διαφόρων καταστάσεων.

Στην περίπτωση του δελτίου ειδήσεων του “Voice Of America”, η μουσική λειτουργεί με όλους τους προαναφερθέντες τρόπους, αλλά και ως εργαλείο για την ένωση ή των διαχωρισμό της θεματολογίας.

## Ομιλία (Speech)

Η ομιλία επίσης αποτελεί μια πολύ βασική κατηγορία ήχου στις ραδιοφωνικές εκπομπές. Μπορεί να παραχθεί μέσα στο στούντιο μέσα από το μικρόφωνο των παρουσιαστών, διά μέσου τηλεφωνικής επικοινωνίας, ή διαφόρων συμμετεχόντων.

## Θόρυβος (Noise)

Ως θόρυβος στη ραδιοφωνική παραγωγή μπορεί να θεωρηθεί οποιοσδήποτε μη επιθυμητός ήχος, ο οποίος ουσιαστικά δεν θα πρέπει να ανήκει στο σύνολο του εκπεμπόμενου ραδιοφωνικού σήματος. Ο θόρυβος μπορεί να εμφανιστεί στο σήμα του ήχου κατά τη διαδικασία της καταγραφής και ψηφιοποίησής του, αλλά και από τυχαίους ήχους που προέρχονται από το περιβάλλον του στούντιο ή από άλλους εξωτερικούς παράγοντες.

Η παρουσία του θορύβου στο ραδιοφωνικό σήμα μπορεί να προκαλέσει λάθη στην αναγνώριση των κατηγοριών, καθώς μπορεί να εντοπιστεί σε τμήματα ήχου που περιέχουν μουσική ή ομιλία.

Για το λόγο αυτό τα ραδιοφωνικά στούντιο είναι επενδεδυμένα με υλικά ηχομόνωσης, με σκοπό την εξουδετέρωση όλων των ανεπιθύμητων ήχων.

Η αλλοίωση ενός σήματος ήχου από θόρυβο μπορεί να μετρηθεί από το λόγο του σήματος προς το θόρυβο (Signal-to-Noise Ratio - SNR), με τη βοήθεια του οποίο μπορεί να υπολογιστεί το πόσο δυνατό είναι ένα σήμα σε σχέση με το θόρυβο που αυτό περιέχει..

Υπολογίζεται από την εξίσωση:

$$SNR = 20log_{10}(\frac{V_s}{V_n})$$

Εάν $V_s$=$V_n$, τότε ο λόγος σήματος προς τον θόρυβο SNR είναι 0.

## Σιωπή (Silence)

Ως σιωπή ορίζονται τα διαστήματα εκείνα κατά τα οποία επικρατεί η απόλυτη απουσία των προαναφερθέντων κατηγοριών. Σε μια ραδιοφωνική εκπομπή μπορούμε να αναφερθούμε στα τμήματα του ήχου στα οποία υπάρχουν παύσεις στην ομιλία ή την μουσική αλλά και στα κενά που υπάρχουν κατά την έναρξη ή τη λήξη ενός ραδιοφωνικού προγράμματος.

# 2 ΕΠΕΞΕΡΓΑΣΙΑ ΣΗΜΑΤΟΣ ΣΤΗΝ ΡΑΔΙΟΦΩΝΙΚΗ ΠΑΡΑΓΩΓΗ

Ο κλάδος της επιστήμης των ηλεκτρονικών υπολογιστών έχει αναπτυχθεί σημαντικά τα τελευταία χρόνια στα ζητήματα της αυτόματης κατηγοριοποίησης.

Ένα μικρόφωνο στους υπολογιστές μετατρέπει τη μηχανική ταλάντωση του σήματος ενός ήχου, μετατρέποντάς το σε ψηφιακό, με σκοπό να μεταδοθεί.

Έτσι, κατά τον τρόπο που το ανθρώπινο αυτί αντιλαμβάνεται τους ήχους αναλύοντάς τους με βάση την ακουστότητα, τη χροιά και το τονικό τους ύψος, παρομοίως ένας υπολογιστής θα πρέπει να αναλύσει τα φασματικά του χαρακτηριστικά προκειμένου να τον κατανοήσει και εν συνεχεία να τον μοντελοποιήσει, ώστε να μπορεί να εξάγει τα χαρακτηριστικά εκείνα που θα είναι ικανά να περιγράψουν το είδος του ήχου που ανήκουν.

## 2.1 ΥΠΟΚΕΙΜΕΝΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΟΥ ΗΧΟΥ

Ο ήχος σε μια ραδιοφωνική εκπομπή, γίνεται αντιληπτός από τον κάθε άνθρωπο με διαφορετικό τρόπο, καθώς το αυτί ως αισθητήριο όργανο αντιλαμβάνεται τον ήχο με βάση τα υποκειμενικά χαρακτηριστικά, τα οποία είναι η ακουστότητα (loudness), το τονικό ύψος (pitch) και η χροιά (timbre).

### 2.1.1 Ακουστότητα (Loudness)

Η ακουστότητα είναι το υποκειμενικό χαρακτηριστικό του ήχου, σύμφωνα με το οποίο το ανθρώπινο αυτί μπορεί να καταλάβει το πόσο δυνατός είναι ένας ήχος. Για να χαρακτηριστεί ένας ήχος όσον αφορά την ακουστότητά του, οι Fletcher & Munson καθιέρωσαν το 1933 ένα σύστημα αναφοράς σύμφωνα με το οποίο συσχετίζεται ο μετρούμενος ήχος με αυτόν που αντιλαμβάνεται το ανθρώπινο αυτό. Ως μέτρο σύγκρισης ορίστηκε η συχνότητα των 1000 Hz, αποδεικνύοντας πως όταν δύο ήχοι έχουν ίδια ένταση και διαφορετική συχνότητα, το ανθρώπινο αυτί πρόκειται να ξεχωρίσει εκείνον που θα έχει την υψηλότερη συχνότητα,

Ως μονάδα μέτρησης της στάθμης ακουστότητας ορίστηκε το phon από τον Γερμανό Φυσικό Heinrich Georg Barkhausen το 1926.

### 2.1.2 Τονικό Ύψος (Pitch)

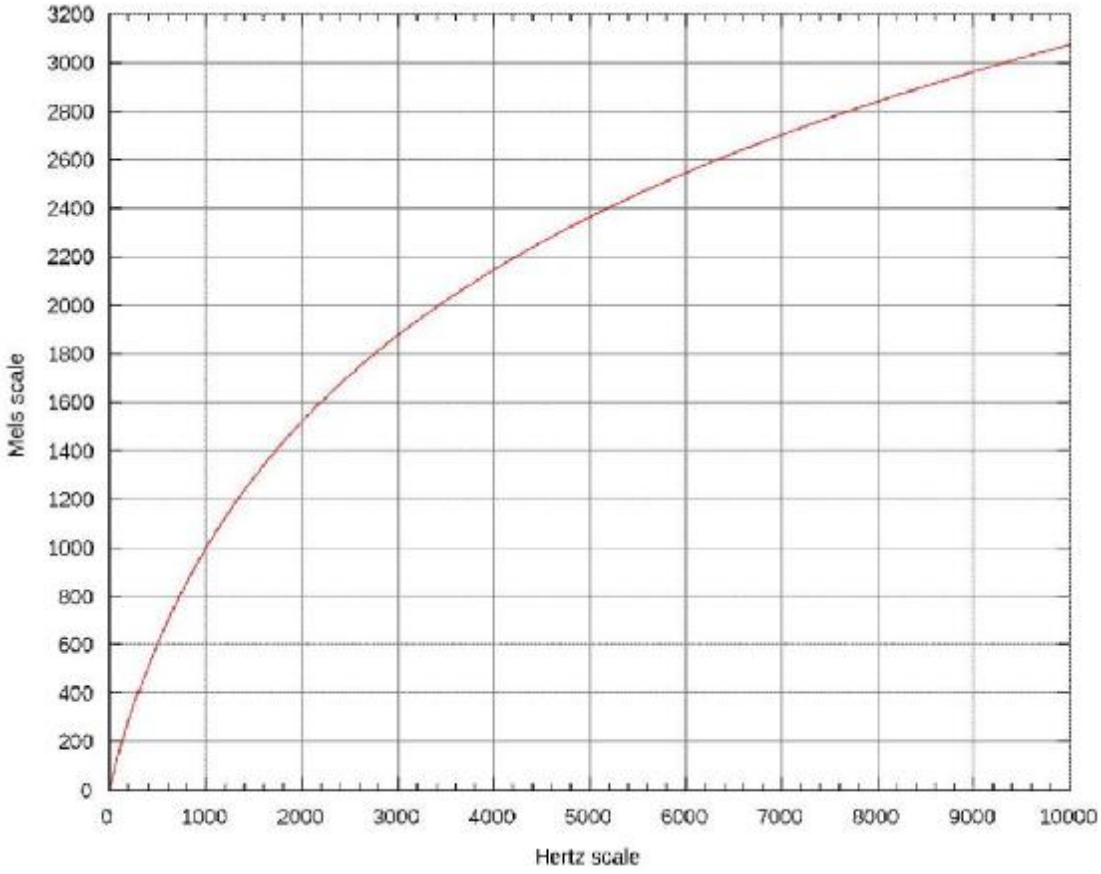
Ως τονικό ύψος ορίζεται το υποκειμενικό χαρακτηριστικό του ήχου σύμφωνα με το οποίο κατατάσσονται σε οξείς, μέσους και βαρείς. Ουσιαστικά ορίζεται ως η υποκειμενική απόκριση του ανθρώπινου αυτιού στη συχνότητα (f), καθώς σχετίζεται με αυτήν, ωστόσω η σχέση που συνδέει τα δύο μεγέθη δεν είναι γραμμική, αφού όταν μεταβάλλεται η συχνότητα δεν σημαίνει πως μεταβάλλεται με τον ίδιο τρόπο και η αίσθηση του ύψους του ήχου.

Σπουδαίο ρόλο στην καθορισμό του τονικού ύψους παίζουν η κυματομορφή του και η στάθμη του ήχου.

Ως μονάδα μέτρησης του ύψους έχει οριστεί η μονάδα Mel από τους Stevens, Volkmann & Newman το 1937 και υπολογίζεται από τον τύπο:

$$\boldsymbol{mel(f) = 2595log_{10}(1 + \frac{f}{700})}$$

Όπου f η συχνότητα του ήχου




Εικόνα 2-1: Σχέση συχνότητας-κλίμακας mel

### 2.1.3 Χροιά (Timbre)

Η χροιά είναι το υποκειμενικό χαρακτηριστικό του ήχου το οποίο σχετίζεται με το είδος της ηχητικής πηγής. Κάνει δυνατό το διαχωρισμό δύο τόνων της ίδιας έντασης και θεμελιώδους συχνότητας αλλά διαφορετικών κυματομορφών.

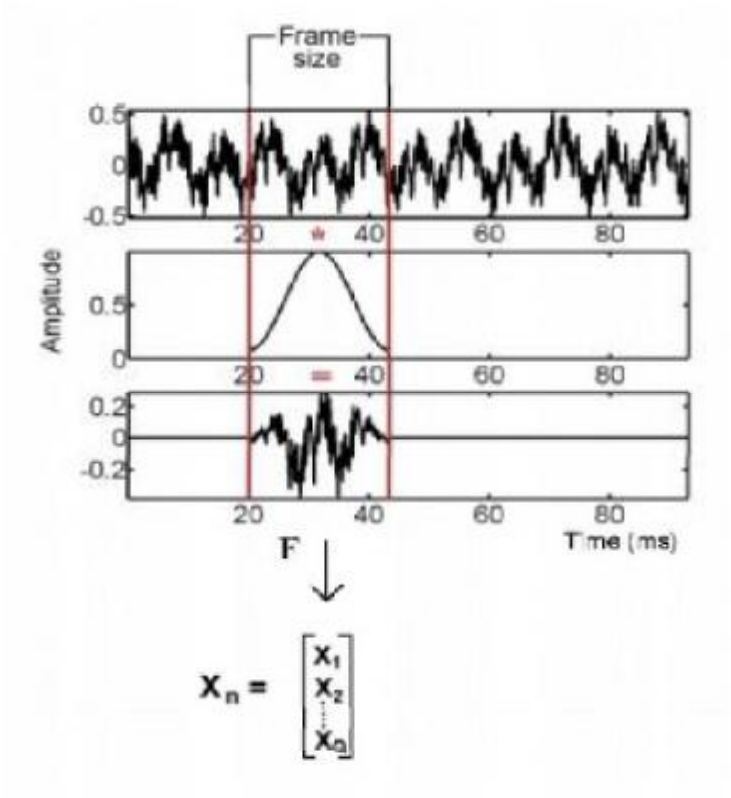
## 2.2 ΦΑΣΜΑΤΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΟΥ ΗΧΟΥ

Τα φασματικά χαρακτηριστικά του ήχου υπολογίζονται από τον γρήγορο μετασχηματισμό Fourier (Fast Fourier Transform-FFT) .

Το διάνυσμα $x_n$για τη χρονική στιγμή n υπολογίζεται με την εξίσωση:

$$x_n = F(ws_n - 1), \dots, w_{n-1}s_n$$

Όπου $s_n$είναι το ακουστικό σήμα, w η συνάρτηση του παραθύρου και N το μέγεθος του διαστήματος (frame).




**Εικόνα 2-2: Εξαγωγή φασματικών χαρακτηριστικών**

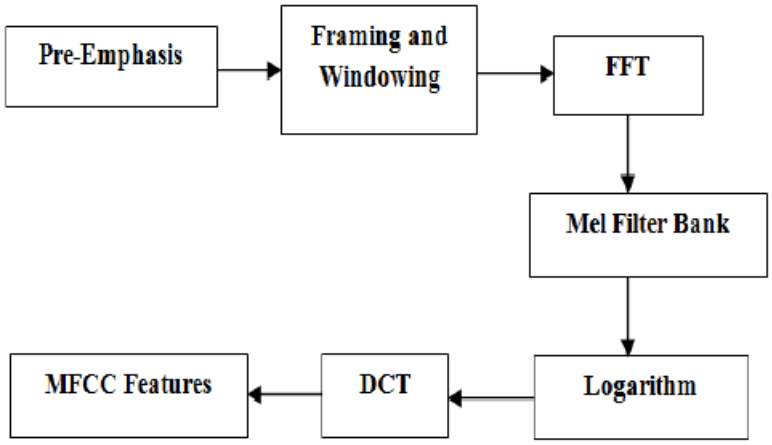
### 2.2.1 Mel-Frequency Cepstral Coefficients (MFCC)

Η εξαγωγή και η επιλογή της καλύτερης παραμετρικής αναπαράστασης των ακουστικών σημάτων είναι πολυ σημαντικά στο σχεδιασμό οποιουδήποτε συστήματος αναγνώρισης ομιλίας.

Επηρεάζει σημαντικά την απόδοση αναγνώρισης. Μία συμπαγής αναπαράσταση θα παρέχεται από ένα σύνολο συντελεστών cepstrum (MFCC), οι οποίοι είναι τα αποτελέσματα ενός μετασχηματισμού συνημίτονου του πραγματικού λογαρίθμου του βραχυπρόθεσμου ενεργειακού φάσματος που εκφράζεται σε μια κλίμακα συχνότητας mel.

Τα MFCC αποδείχθηκαν πιο αποτελεσματικά στο διαχωρισμό ομιλίας από μουσική, ωστόσο στην παρούσα εργασία χρησιμοποιήθηκαν και για τον διαχωρισμό του θορύβου και της σιωπής, γεγονός που αποδεικνύεται από το μεγάλο ποσοστό επιτυχίας στα αποτελέσματα της πειραματικής διαδικασίας.

**Ο υπολογισμός του MFCC περιλαμβάνει τα ακόλουθα βήματα:**




Εικόνα 2-3: Η διαδικασία εξαγωγής των MFCC χαρακτηριστικών από ένα ηχητικό σήμα

### 2.2.2 Φασματικό Κέντρο Βάρους (Spectral Centroid)

Το Φασματικό Κέντρο Βάρους περιγράφει το σημείο ισορροπίας του φασματικού πλάτους του τμηματικού μετασχηματισμού Fourier (STFT).

Δείχνει σε ποιο σημείο του φάσματος είναι συγκεντρωμένη η περισσότερη ενέργεια.

Υπολογίζεται από την εξίσωση:

$$SC_{t=} \frac{\sum_{n=1}^{N} M_t[n]n}{\sum_{n=1}^{N} M_t[n]}$$

Όπου $M_t$[n] η τιμή του φάσματος μετά τον μετασχηματισμό Fourier στο διάστημα χρόνου t και στην τιμή συχνότητας n.

### 2.2.3 Φασματικό Roll-off (Spectral Roll-off)

Το φασματικό Roll-off αποτελεί το χαρακτηριστικό εκείνο το οποίο δηλώνει την κατανομή της ενέργειας ενός σήματος στις τονικά χαμηλές συχνότητες.

Υπολογίζεται από την εξίσωση:

$$\sum_{n-1}^{Rt} M_t[n] = 0.85 \sum M_t[n]_{n-1}^{N}$$

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

### 2.2.4 Φασματική Ροή (Spectral Flux)

Η Φασματική Ροή (Spectral Flux) είναι το χαρακτηριστικό εκείνο το οποίο υπολογίζει και εκτιμάει την ποσότητα της γενικής φασματικής αλλαγής σε ένα σήμα.

Υπολογίζεται από την εξίσωση:

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

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

### 2.2.5 Zero-Crossing Rate (ZCR)

Στο πεδίο των σημάτων διακριτού χρόνου, τυχαίνει να συμβαίνει το zero-crossing εάν διαδοχικά δείγματα έχουν διαφορετικά αλγεβρικά σημάδια. Ο ρυθμός με τον οποίο συμβαίνουν μηδενικές διασταυρώσεις είναι ένα απλό μέτρο της περιεκτικότητας σε συχνότητα του ακουστικού σήματος.

Το Πλήθος Μηδενισμού Συνάρτησης περιγράφει τον αριθμό των φορών που διέρχεται το σήμα από το σημείο 0, δηλαδή το πλήθος των σημείων που μηδενίζει.

Ο ρυθμός με τον οποίο συμβαίνουν στα σήματα διακριτού χρόνου μηδενικές διασταυρώσεις.

Ο ρυθμός μηδενικής διέλευσης (zero-crossing rate) μας βοηθά στο να μπορέσουμε να διαχωρίσουμε με μεγαλύτερη ευκολία την ομιλία από την μουσική, καθώς έχει διαπιστωθεί μέσω πειραμάτων πως τα zero-crossings σε τμήματα που περιέχουν μουσική είναι μεγαλύτερα, καθώς μέρος της ενέργειας του σήματος βρίσκεται σε υψηλότερες συχνότητες σε σύγκριση με εκείνα που περιέχουν ομιλία και μέρος της ενέργειάς τους βρίσκεται σε χαμηλότερες συχνότητες. (Bachu, Adapa, & Barkana, March 2008)

**Το Zero-Crossing Rate υπολογίζεται από την εξίσωση:**

$$\boldsymbol{Z}\mathbf{n} = \sum_{\boldsymbol{m}=-\infty}^{\infty} |\, \boldsymbol{sgn}[\boldsymbol{x}(\boldsymbol{m})] - \boldsymbol{sgn}[\boldsymbol{x}(\boldsymbol{m}-\mathbf{1})]|\, \boldsymbol{w}(\boldsymbol{n}-\boldsymbol{m})$$

όπου

$$\text{Sgn[x(n)} = \begin{cases} 1, x(n) \geq 0 \\ -1, x(n) < 0 \end{cases}$$

Και

$$\text{W(n)} = \begin{cases} \frac{1}{2N} \; \gamma\iota\alpha \; 0 \leq \nu \leq N-1 \\ 0 \; \gamma\iota\alpha \; \acute{\alpha}\lambda\lambda o \end{cases}$$

### 2.2.6 Ενέργεια Βραχέως Χρόνου (Short-Time Energy)

Η ενέργεια βραχέως χρόνου χρησιμοποιείται κατά κύριο λόγο με σκοπό να εντοπιστούν τα τμήματα του σήματος που εμπεριέχουν ήχο ή σιωπή.

Το χαρακτηριστικό Short-Time Energy υπολογίζεται με βάση την εξίσωση:

$$STE_n = \frac{1}{N} \sum_{i=n-N+1}^{n} s^2\mathrm{i}$$

Όπου $s_i$ το σήμα στην τιμή του χρόνου.

Συμπερασματικά, η ακουστότητα (loudness) ενός ήχου συνδέεται στενά με την ένταση του σήματος και με την ενέργεια του μικρού χρόνου το ήχου (Ζαρβαδάς, 2013).

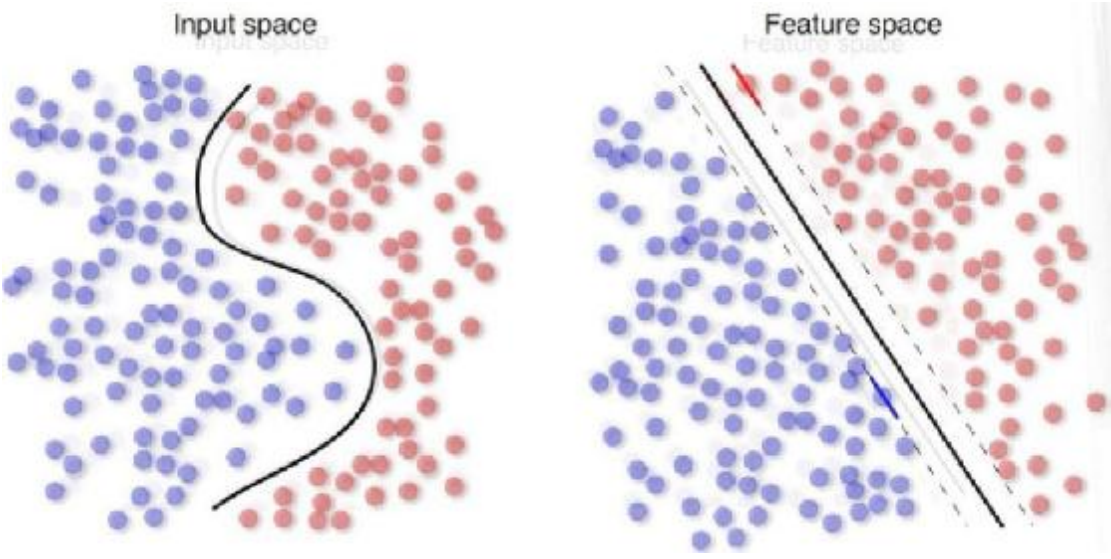
### 2.2.7 Αλγόριθμος Κατηγοριοποίησης SVM (SVM Classification Algorithm)

Ο αλγόριθμος που χρησιμοποιήθηκε στην παρούσα εργασία ήταν οι Support Vector Machine (SVM), καθώς θεωρείται ως ένας από τους πιο αξιόπιστους αλγόριθμους που επιτυγχάνουν υψηλά ποσοστά στην κατηγοριοποίηση.

Στον συγκεκριμένο αλγόριθμο επιλέγεται ένας αριθμός στιγμιοτύπων από κάθε κλάση που συνορεύουν στο χώρο του προβλήματος με στιγμιότυπα άλλων κλάσεων.

Ο ταξινομητής πρέπει να κάνει διάκριση μεταξύ δύο τύπων δεδομένων με ετικέτα με μη γραμμική διακριτική λειτουργία (χώρος εισόδου). Τα δεδομένα είναι ανυψωμένα σε ένα υψηλότερο χώρο διαστάσεων όπου τα δεδομένα μπορούν να διακριθούν με ένα υπερπλαστή (χώρος χαρακτηριστικών]. Η διακεκομμένη γραμμή αντιπροσωπεύει τους φορείς υποστήριξης. Ο αλγόριθμος SVM προσπαθεί να μεγιστοποιήσει την απόσταση μεταξύ φορέων υποστήριξης αντίθετων τάξεων. (T. Jehan, 2005)

Στην παρακάτω εικόνα φαίνεται η λειτουργία τους:




Εικόνα 2-4: Η τυπική αναπαράσταση λειτουργίας ενός SVM (Jehan,2005)

Ουσιαστικά πρόκειται για έναν απλό δυαδικό ταξινομητή, ο οποίος «μαθαίνει» το σύνορο απόφασης ανάμεσα σε δύο κλάσεις.

Για να μπορέσει να εντοπίσει το σύνορο αυτό, μεγιστοποιεί την απόσταση αυτή ανάμεσα σε δύο κλάσεις, επιλέγοντας γραμμικούς διαχωριστές στο χώρο των παραμέτρων.

Μια συνάρτηση τύπου kernel χρησιμοποιείται με σκοπό να προβάλει τα δεδομένα από το χώρο εισόδου στο χώρο των παραμέτρων, με γραμμικό ή με μη γραμμικό τρόπο, ανάλογα με τη μορφή των συνόρων απόφασης. (D. Michie, D.J. Spiegelhalter, & C. Taylor, 1994)

Η ικανότητα των Support Vector Machines να μπορούν να παράγουν και μη-γραμμικές επιφάνειες απόφασης, τις καθιστά πολύ σημαντικές από άποψη υπολογιστικής ικανότητας, ώστε να επιλύουν έναν μεγάλο όγκο προβλημάτων ταξινόμησης, τα οποία δεν θα ήταν εύκολο να υπολογιστούν με μοντέλα γραμμικού τύπου.

# 3 ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

Στην πειραματική διαδικασία σκοπός ήταν η επεξεργασία των αρχείων ήχου της Βάσης Δεδομένων

## 3.1 ΠΕΡΙΓΡΑΦΗ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ (DATABASE DESCRIPTION)

Για την ανάλυση του ήχου σε μια ραδιοφωνική εκπομπή είναι απαραίτητη η δημιουργία μιας Βάσης Δεδομένων στην οποία έχουν αναλυθεί όλοι οι ήχοι που διατίθεται προς επεξεργασία.

Η Βάση Δεδομένων περιέχει 300 αρχεία από το Δελτίο Ειδήσεων της Φωνής της Αμερικής (Voice Of America) από τις οποίες προκύπτουν 2394 διαφορετικές καταστάσεις ήχου (ομιλία, μουσική, ομιλία+μουσική, σιωπή+θόρυβος) συνολικά.

Αναλυτικά, τα αρχεία που χρησιμοποιήθηκαν και κατατμήθηκαν ανά κατηγορία είναι:

**Μουσική (Music) -736 δείγματα**

Μικρότερη Διάρκεια: 0,989 δευτερόλεπτα
Μεγαλύτερη Διάρκεια: 8 λεπτά και 20,839 δευτερόλεπτα / 500,839 δευτερόλεπτα.
Μέση Διάρκεια: 144,840 δευτερόλεπτα.

**Ομιλία (Speech) – 429 δείγματα**

Μικρότερη Διάρκεια: 0,341 δευτερόλεπτα
Μεγαλύτερη Διάρκεια: 302,324 δευτερόλεπτα
Μέση Διάρκεια: 199,802 δευτερόλεπτα

**Μουσική + Ομιλία (Music+Speech) – 486 δείγματα**

Μικρότερη Διάρκεια: 1.376 δευτερόλεπτα
Μεγαλύτερη Διάρκεια: 19,493 δευτερόλεπτα
Μέση Διάρκεια: 9,738 δευτερόλεπτα

**Θόρυβος (Noise) – 113 δείγματα**

Μικρότερη Διάρκεια: 0,417 δευτερόλεπτα
Μεγαλύτερη Διάρκεια: 13,843 δευτερόλεπτα
Μέση Διάρκεια: 3,577 δευτερόλεπτα

**Σιωπή (Silence) – 617 δείγματα**

Μικρότερη Διάρκεια: 0,091 δευτερόλεπτα
Μεγαλύτερη Διάρκεια: 51,779 δευτερόλεπτα
Μέση Διάρκεια: 2,929 δευτερόλεπτα

## 3.2 ΜΗ ΑΥΤΟΜΑΤΗ ΤΜΗΜΑΤΟΠΟΙΗΣΗ ΗΧΟΥ (MANUAL SOUND SEGMENTANTION) - PRAAT

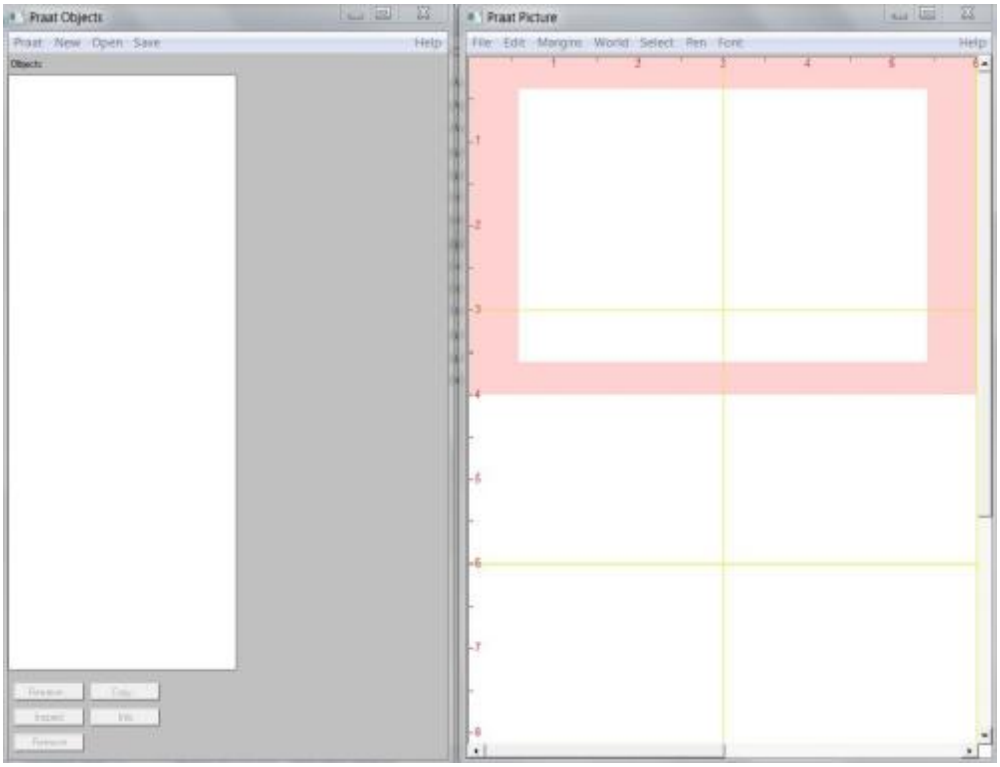
Ένα μεγάλο μέρος αυτής της εργασίας αφιερώθηκε χρονικά στην μη-αυτόματη τμηματοποίηση των ραδιοφωνικών εκπομπών σε επιμέρους κατηγορίες ήχων με τη χρήση του λογισμικού Praat. Το εν λόγω πρόγραμμα χρησιμοποιήθηκε σε 300 αρχεία ήχου τύπου .wav σε μονοφωνικό κανάλι (mono), όλα στην Ελληνική γλώσσα, με σκοπό να εξαχθούν καλύτερα απότελέσματα κατά την πειραματική διαδικασία και για την ορθότερη εξαγωγή παραμέτρων.

Η αναγνώριση των ηχητικών κατηγοριών (μουσική-ομιλία-σιωπή-θόρυβος) έγινε χειροκίνητα μέσω του **λογισμικού Praat**, όπου η κυματομορφή του εκάστοτε αρχείου ήχου επισημάθηκε με labels, ώστε να γίνει η κατηγοριοποίησή τους (classification).

Το λογισμικό (Praat) είναι ένα ευέλικτο πρόγραμμα που πραγματοποιεί ανάλυση στον ήχο. Προσφέρει ένα ευρύ φάσμα τυπικών και μη τυποποιημένων διαδικασιών, συμπεριλαμβανομένης της φασματογραφικής ανάλυσης, της αρθρωτικής σύνθεσης και των νευρωνικών δικτύων.

Κατά την είσοδο στο πρόγραμμα, εμφανίζονται δύο παράθυρα. Στα δεξιά εμφανίζεται το «Praat Picture», το οποίο ουσιαστικά δεν χρησιμοποιήθηκε στην πειραματική διαδικασία της συγκεκριμένης εργασίας.

Στα αριστερά εμφανίζεται το παράθυρο «Praat Objects», όπου εμφανίζονται τα αρχεία ήχου που φορτώνονται

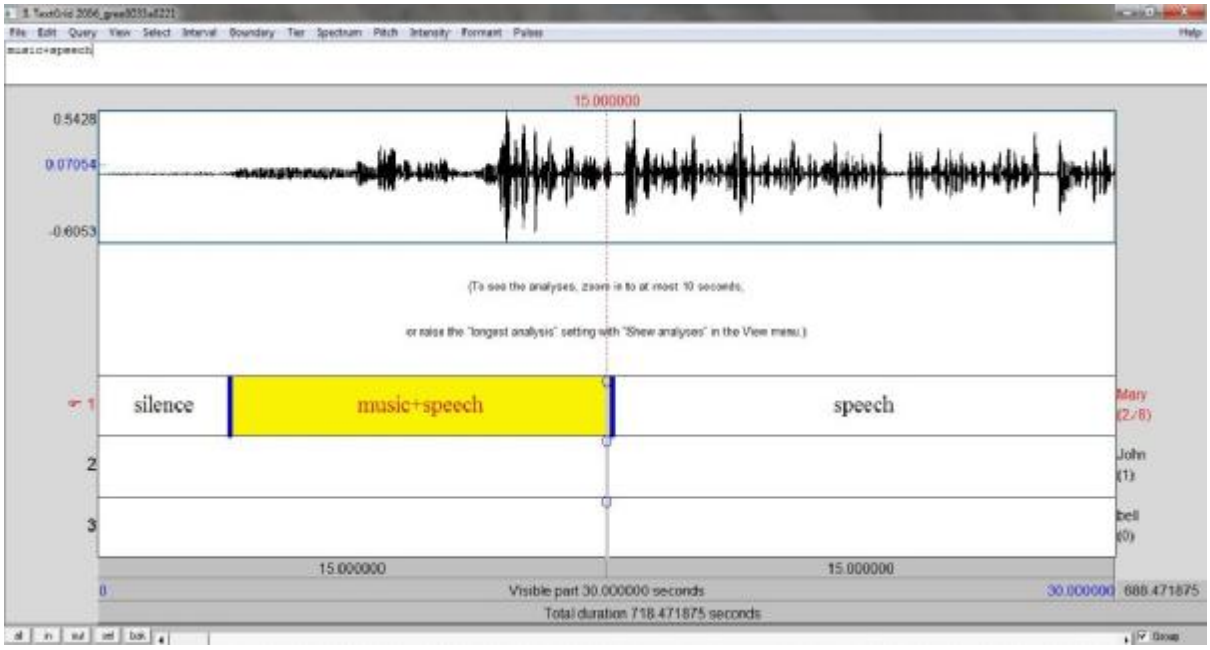



Εικόνα 3-1: Τα παράθυρα που ανοίγουν κατά την εκκίνηση του Praat

Αφού ανοίξουμε το αρχείο από την επιλογή του menu “Open long sound file”, με την επιλογή “Annotate” δημιουργούνται αρχεία κειμένου, που ονομάζονται “TextGrids” στα οποία καταγράφονται οι ετικέτες που δημιουργούνται από τη διαδικασία της χειροκίνητης κατάτμησης του ήχου.

Στην παρακάτω εικόνα εμφανίζεται ενδεικτικά το παράθυρο επεξεργασίας του ήχου στο οποίο πραγματοποιείται η εν λόγω διαδικασία. Κάθε αρχικό αρχείο που χρησιμοποιήθηκε ονοματίστηκε ως “speech”, “music”, “silence”, “noise”, αλλά και μεικτά, καθώς αρκετά σημεία παρουσίαζαν παραπάνω από ένα είδος συνδυαστικά, γεγονός το οποίο καθιστούσε τη διαδικασία ακόμη πιο δύσκολη, καθώς με υποκειμενικά κριτήρια έγινε η συνολική κατάτμηση.

Μέσα στο πρόγραμμα, οι ήχοι εμφανίζονται με τη μορφή κυματομορφής η οποία διαχωρίζεται τμηματικά, ανάλογα με το είδος του ήχου στο οποίο αναφερόμαστε.




Εικόνα 3-2: Παράθυρο επεξεργασίας στο Praat-Παράδειγμα labeling

Αφού ολοκληρωθεί η διαδικασία, γίνεται αποθήκευση σε TextGrid, το οποίο είναι το αρχείο εκείνο στο οποίο εξάγονται τα χαρακτηριστικά από την κατάτμηση και το labeling.

## 3.3 ΕΞΑΓΩΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ (FEATURE EXTRACTION) - MARSYAS

Η διαδικασία της μετατροπής του ακουστικού σήματος σε μία ακολουθία ακουστικών διανυσμάτων με χαρακτηριστικά που μεταφέρουν πληροφορίες για τα χαρακτηριστικά του ήχου, ονομάζεται εξαγωγή παραμέτρων (Feature Extaction) και πρόκειται ίσως για το σημαντικότερο στάδιο της πειραματικής διαδικασίας.

Η μεθοδολογία που ακολουθήθηκε για την εξαγωγή παραμέτρων στην παρούσα εργασία γίνεται με τη βοήθεια του εκτελέσιμου υποπρογράμματος **bextract** που ανήκει στην πλατφόρμα ανοιχτού κώδικα **MARSYAS** (**M**usic **A**nalysis **R**etrieval and **SY**nthesis for **A**udio **S**ignals) (Tzanetakis, Music Analysis, Retrieval and Synthesis of Audio Signals MARSYAS, 2009) και μπορεί να εκτελεστεί μέσω της γραμμής εντολών του λειτουργικού συτήματος (Command Line –CMD)

Μπορεί να χρησιμοποιηθεί προκειμένου να εκτελέσει πλήρη εξαγωγή παραμέτρων και κατηγοριοποίηση από πολλαπλά αρχεία με απώτερο σκοπό τον εύκολο διαχωρισμό μεταξύ ομιλίας, μουσικής, θορύβου και σιωπής από αρχεία ήχων που έχουμε στη βάση δεδομένων, για την αναλυτική εξαγωγή των χαρακτηριστικών του ήχου, δημιουργώντας feature vectors με σκοπό την ταξινόμηση των ηχητικών σημάτων.

Το πρόγραμμα αυτό εκτελείται σε γραμμή εντολών σε λογισμικά Windows, σε Mac-OS αλλά και σε Linux. Προηγουμένως με παρόμοιο τρόπο έχουμε δημιουργήσει συλλογές με τα είδη των ήχων που υπάρχουν ήδη στη Βάση Δεδομένων, χρησιμοποιώντας το εκτελέσιμο πρόγραμμα -mkcollection, μέσω του οποίου δημιουργούνται αρχεία με κατάληξη ".mf".

Τα ".mf" αρχεία είναι αυτά τα οποία χρησιμοποιούνται στο bextract για την εξαγωγή παραμέτρων. Από εκεί δημιουργείται ένα αρχείο ".arff".

Ο κώδικας που γράφεται σε γραμμή εντολών στο MARSYAS για τη δημιουργία ενός τέτοιου αρχείου για κάθε κατηγορία είναι:

***bextract music.mf –w music.arff***
***bextract speech.mf –w speech.arff***
***bextract noise.mf –w noise.arff***
***bextract silence.mf –w silence.arff***

καθώς και μίξη αυτών όπως είναι πχ.:

***bextract music_speech.mf –w music_speech.arff***

Μετέπειτα, το ".arff" αρχείο χρησιμοποιείται στο στάδιο του model training από τα προγράμματα kea από το MARSYAS και το Weka, όπως θα δούμε και στο παρακάτω κεφάλαιο.

Στον ήχο μπορούμε να διακρίνουμε κάποια χαρακτηριστικά τα οποία μπορούμε να χρησιμοποιήσουμε κατα τη διαδικασία της εξαγωγής παραμέτρων και θα μας βοηθήσουν στο να μπορέσουμε με μεγαλύτερη ευκολία να διαχωρίσουμε τα ηχητικά σήματα που ανήκουν στην κατηγορία της μουσικής, στην κατηγορία της ομιλίας ή σε κάποια άλλη κατηγορία, π.χ. θόρυβος ή συνδυασμός ομιλίας με μουσική, κλπ., δημιουργώντας τις κλάσεις music, speech, noise και silence.

Στην πειραματική διαδικασία που ακολουθήθηκε πραγματοποιήθηκε η εξαγωγή των εξής χαρακτηριστικών:

- Beat Histogram
- LPCC (Linear Prediction Cepstral Coefficients)
- LSP (Line Spectral Pair)
- MFCC (Mel-Frequency Cepstral Coefficients)
- SMFCF
- STFT (Short-Time Fourier Transform)
- STFTMFCC (Short-Time Fourier Transform & Mel Frequency Cepstral Coefficients)

## 3.4 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ (CLASSIFICATION) - WEKA

Βασικό στάδιο στην πειραματική διαδικασία ήταν το στάδιο της κατηγοριοποίησης (Classification), διαδικασία στην οποία χρησιμοποιήθηκε το πρόγραμμα (WEKA) από το Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία, το οποίο είναι ένα λογισμικό ανοιχτού κώδικα γραμμένο σε γλώσσα προγραμματισμού Java και διαθέτει μια συλλογή από αλγόριθμους εκμάθησης μηχανής και εξόρυξης δεδομένων (data-mining). Περιέχει εργαλεία για την προετοιμασία δεδομένων, την ταξινόμησή τους, την ομαδοποίηση αλλά και την οπτικοποίησή τους με τη μορφή γραφημάτων.

Η κατηγοριοποίηση πραγματοποιήθηκε χρησιμοποιώντας τις βασικότερες μεθόδους και τεχνικές με σκοπό να γίνει ο πιο αποτελεσματικός διαχωρισμός ανάμεσα στα είδη του ήχου που διαχρίστηκαν από τις προηγούμενες διαδικασίες.

Ο αλγόριθμος που χρησιμοποιήθηκε ήταν οι Support Vector Machines (SVM), του οποίου η λειτουργία αναφέρεται στο Κεφάλαιο 2.

Αφού έχει εισαχθεί το αρχείο .arff στο WEKA, επιλέγουμε στην καρτέλα Classsify τον αλγόριθμο LibSVM με 10-fold cross validation.

Αφού έχουν εισαχθεί όλα τα .arff αρχεία που έχουν προκύψει από τα προηγούμενα στάδια, πλέον έχουμε εξάγει όλα τα αποτελέσματα που μας δείχνουν τα ποσοστά κατηγοριοποίησης των ήχων και κατά πόσο επιτυχημένα διαχωρίστηκαν από τον ηλεκτρονικό υπολογιστή οι επι μέρους κατηγορίες σε μια ραδιοφωνική εκπομπή.

Σε αυτό το κεφάλαιο παρουσιάζονται αναλυτικά όλα τα αποτελέσματα από τη διαδικασία του Machine Learning, η οποία πραγματοποιήθηκε με το λογισμικό WEKA.

Τα παρακάτω δείχνουν με απόλυτη λεπτομέρεια όλα τα χαρακτηριστικά καθώς και τους πίνακες confusion matrix οι οποίοι δείχνουν με ακρίβεια τον διαχωρισμό των εκπομπών ήχου στις διαφορετικές κατηγορίες που έχουν ορισθεί εξ’αρχής μέσω του λογισμικού Praat και με την διαδικασία εξαγωγής χαρακτηριστικών από το MARSYAS.

Έτσι, κατά την ίδια σειρά με την οποία εκτελέστηκαν τα πειράματα, παρουσιάζονται τα αποτελέσματα όπως εμφανίστηκαν τελικά στο παράθυρο του λογισμικού, καθώς οι σχετικοί πίνακες που αναφέρονται σε αυτά.

Επομένως, έχουμε τα:

**All Default**

Σε αυτό σετ χρησιμοποιήθηκαν όλες οι προκαθορισμένες παράμετροι του Weka, χωρίς να γίνει κάποια αλλαγή στις ρυθμίσεις.

Η διαδικασία έβγαλε σαν αποτέλεσμα επιτυχίας 91.893%, και αποτυχίας 8.107%

Σύμφωνα με τον πίνακα που προκύπτει παρακάτω (confusion matrix), μπορούμε να διαπιστώσουμε με ακρίβεια ποιους ήχους αντιλαμβάνεται σωστά το σύστημα, με βάση το πως τους έχουμε ήδη εμείς ονοματίσει από την αρχή (labeling).

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1
Relation: all_default.arff
Instances: 2393
Attributes: 125
[list of attributes omitted]
Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 3.96 seconds

*=== Stratified cross-validation ===*
*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 2199 | 91.893 % |
| Incorrectly Classified Instances | 194 | 8.107 % |
| Kappa statistic | 0.8896 | |
| Mean absolute error | 0.0405 | |
| Root mean squared error | 0.2013 | |
| Relative absolute error | 11.0164 % | |
| Root relative squared error | 46.9398 % | |
| Coverage of cases (0.95 level) | 91.893 % | |
| Mean rel. region size (0.95 level) | 25% | |
| Total Number of Instances | 2393 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.951 | 0.048 | 0.897 | 0.951 | 0.923 | 0.951 | **Music** |
| | 0.819 | 0.031 | 0.871 | 0.819 | 0.884 | 0.894 | **Music+Speech** |
| | 0.987 | 0.002 | 0.996 | 0.987 | 0.991 | 0.992 | **Silence+Noise** |
| | 0.86 | 0.026 | 0.876 | 0.86 | 0.868 | 0.917 | **Speech** |
| **Weighted Avg.** | 0.919 | 0.026 | 0.919 | 0.919 | 0.918 | 0.946 | |

Δηλαδή διαπιστώνεται πως τα κάτωθι:

- a= Music
- b= Music+Speech
- c= Silence+Noise
- d= Speech

στο διαγώνιο νοητό άξονα του πίνακα με τα bold φαίνεται ο αριθμός των τμημάτων ήχου που έχουν ονοματιστεί σωστά και τα οποία αναγνωρίζονται από το σύστημα.

Για τη μουσική (music), αναγνωρίζονται σωστά 700, ενώ μπερδεύονται τα 44 ως music+speech, τα 6 ως silence+noise και 30 ως speech.
Για το music+speech, αναγνωρίζονται σωστά τα 398, τα 31 ως music, κανένα ως silence+noise και 28 ως speech.
Για το silence+noise, αναγνωρίζονται σωστά τα 732, 1 μόνο ως music, κανένα ως music+speech και 2 ως speech.
Τέλος για το speech αναγνωρίζονται σωστά 369, 4 μπερδεύονται ως music, 44 ως music+speech και 4 ως speech.

*=== **Confusion Matrix** ===*

| **a** | **b** | **c** | **d** | **ç classified as** |
|---|---|---|---|---|
| **700** | 31 | 1 | 4 | **a=music** |
| 44 | **398** | 0 | 44 | **b=music+speech** |
| 6 | 0 | **732** | 4 | **c=silence+noise** |
| 30 | 28 | 2 | **369** | **d=speech** |

Συμπεραίνοντας, για το 1$^{ο}$ δοκιμαστικό set, διαπιστώνουμε πως βάσει του ποσοστού επιτυχίας που βγάζει και αξιολογώντας τον πίνακα του confusion matrix, προκύπτει ένα αρκετά καλό αποτέλεσμα, το οποίο μας δείχνει πως δεν μπερδεύονται τόσο εύκολα οι κλάσεις του ήχου.

**All Default – HopSamples 256**

Σε αυτό το σετ χρησιμοποιήθηκαν ακριβώς τα ίδια χαρακτηριστικά με το 1ο πείραμα με τη βασική, όμως, διαφορά ότι στο Marsyas, κατά τη διαδικασία του Feature Extraction, το μέγεθος των HopSamples ορίστηκε στα 256 και όχι default (δηλαδή 512).

Επιτυχώς ταξινομήθηκε το 67.6507% και λανθασμένα το 32.3493%. μέσα σε χρονικό διάστημα 12.47 δευτερόλεπτα.

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1
Relation: ALLCombination_HopSamples256.arff
Instances: 2405
Attributes: 377
[list of attributes omitted]
Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 12.47 seconds

*=== Stratified cross-validation ===*
*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 1627 | 67.6507 % |
| Incorrectly Classified Instances | 778 | 32.3493 % |
| Kappa statistic | 0.5382 | |
| Mean absolute error | 0.1617 | |
| Root mean squared error | 0.4022 | |
| Relative absolute error | 43.9869 % | |
| Root relative squared error | 93.796 % | |
| Coverage of cases (0.95 level) | 67.6507 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2405 | |

*=== Detailed Accuracy By Class ===*

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|---|
| | 0.99 | 0.451 | 0.492 | 0.99 | 0.657 | 0.77 | **music** |
| | 0.241 | 0.007 | 0.893 | 0.241 | 0.379 | 0.617 | **music+speech** |
| | 0.967 | 0.002 | 0.995 | 0.967 | 0.98 | 0.982 | **silence+noise** |
| | 0.121 | 0.004 | 0.881 | 0.121 | 0.213 | 0.559 | **speech** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Weighted Avg.** | 0.677 | 0.141 | 0.8 | 0.677 | 0.623 | 0.768 | |

Το ποσοστό 67.6507% δικαιολογείται εάν αναλύσουμε τα αποτελέσματα του παρακάτω πίνακα, όπως έγινε και στο πρώτο πείραμα.

Για το music αναγνωρίζονται σωστά τα 729, ενώ 369 αναγνωρίζονται λανθασμένα ως music+speech, 24 ως silence+noise και 366 ως speech.
Για το music+speech αναγνωρίζονται σωστά 117, τα 4 μπερδεύονται ως music, κανένα ως silence+noise, και 10 μπερδεύονται με το speech.
Το silence+noise λειτουργεί καλύτερα σε αυτήν την περίπτωση, μιας και έχει 729 αναγνωρισμένα σωστά, 3 μόνο τα αναγνωρίζει ως music, κανένα ως music+speech και 1 μόνο ως speech.

*=== Confusion Matrix ===*

| **a** | **b** | **c** | **d** | **ç classified as** |
|---|---|---|---|---|
| **729** | 4 | 3 | 0 | **a=music** |
| 363 | **117** | 0 | 6 | **b=music+speech** |
| 24 | 0 | **729** | 1 | **c=silence+noise** |
| 366 | 10 | 1 | **52** | **d=speech** |

**Beat Histogram Features (BEAT)**

Στο συγκεκριμένο set έχουμε ακριβώς τα ίδια αποτελέσματα με το default σετ ($1^{ο}$). Δηλαδή το ποσοστό των σωστά ταξινομημένων τμημάτων είναι 91.893% και των λανθασμένα ταξινομημένων είναι 8.107%.

Ως εκ τούτου, και τα συμπεράσματα που προκύπτουν από τον πίνακα confusion matrix, είναι ακριβώς τα ίδια.

Άρα:

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1
Relation: beat_default.arff
Instances: 2393
Attributes: 125
[list of attributes omitted]
Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 4.78 seconds

*=== Stratified cross-validation ===*
*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 2199 | 91.893 % |
| Incorrectly Classified Instances | 194 | 8.107 % |
| Kappa statistic | 0.8896 | |
| Mean absolute error | 0.0405 | |
| Root mean squared error | 0.2013 | |
| Relative absolute error | 11.0164 % | |
| Root relative squared error | 46.9398 % | |
| Coverage of cases (0.95 level) | 91.893 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2393 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.951 | 0.048 | 0.897 | 0.951 | 0.923 | 0.951 | **Music** |
| | 0.819 | 0.031 | 0.871 | 0.819 | 0.844 | 0.894 | **Music+Speech** |
| | 0.987 | 0.002 | 0.996 | 0.987 | 0.991 | 0.992 | **Silence+Noise** |
| | 0.86 | 0.026 | 0.876 | 0.86 | 0.868 | 0.917 | **Speech** |
| **Weighted Avg.** | 0.919 | 0.026 | 0.919 | 0.919 | 0.918 | 0.946 | |

*=== Confusion Matrix ===*

| **a** | **b** | **c** | **d** | **ç classified as** |
|---|---|---|---|---|
| **700** | 31 | 1 | 4 | **a=music** |
| 44 | **398** | 0 | 44 | **b=music+speech** |
| 6 | 0 | **732** | 4 | **c=silence+noise** |
| 30 | 28 | 2 | **369** | **d=speech** |

**Beat Histogram – HopSamples 256**

Παρ'ότι το default σετ του Beat Histogram είναι ίδιο με το Default σετ, δε θα λέγαμε το ίδιο και για το σετ Beat Histogram – Hopsamples 256 σε σχέση με το Default – Hopsamples 256.

Εδώ, το ποσοστό των επιτυχώς ταξινομημένων τμημάτων ήχου είναι 92.183%, ενώ αποτυχημένα ταξινομήθηκε το 7.817%.

Παρακάτω, εξηγείται με τη βοήθεια του πίνακα confusion matrix αυτό το ποσοστό.

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1
Relation: beat_HopSamples256.arff
Instances: 2405
Attributes: 125
[list of attributes omitted]
Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 5.91 seconds

*=== Stratified cross-validation ===*
*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 2217 | 92.183 % |
| Incorrectly Classified Instances | 188 | 7.817 % |
| Kappa statistic | 0.8935 | |
| Mean absolute error | 0.0391 | |
| Root mean squared error | 0.1977 | |
| Relative absolute error | 10.6292 % | |
| Root relative squared error | 46.1077 % | |
| Coverage of cases (0.95 level) | 92.183 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2405 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.959 | 0.048 | 0.897 | 0.951 | 0.923 | 0.951 | **Music** |
| | 0.819 | 0.031 | 0.871 | 0.819 | 0.844 | 0.894 | **Music+Speech** |
| | 0.993 | 0.002 | 0.996 | 0.987 | 0.991 | 0.992 | **Silence+Noise** |
| | 0.848 | 0.026 | 0.876 | 0.86 | 0.868 | 0.917 | **Speech** |
| **Weighted Avg.** | 0.919 | 0.026 | 0.919 | 0.919 | 0.918 | 0.946 | |

*=== Confusion Matrix ===*

Εξηγώντας, λοιπόν τους παρακάτω αριθμούς, έχουμε:

Για τη μουσική έχουν ταξινομηθεί σωστά 706 τμήματα ήχου, ενώ 31 μπερδεύονται ως music+speech, 5 ως silence+noise και 30 ως speech.
Για το music+speech, σωστά έχουν ταξινομηθεί 398 τμήματα ήχου, ενώ λανθασμένα 27 ως music, κανένα ως silence+noise και 34 ως speech.
Για το silence+noise αναγνωρίζονται σωστά 749 τμήματα ήχου, 1 ως music, κανένα ως music+speech και 1 ως speech.

Για το speech ταξινομούνται σωστά 364 τμήματα ήχου, ενώ αντιθέτως 2 μπερδεύονται ως music, 57 ως music+speech και κανένα ως silence+noise.

| a | b | c | d | ç classified as |
|---|---|---|---|---|
| **706** | 27 | 1 | 2 | **a=music** |
| 31 | **398** | 0 | 57 | **b=music+speech** |
| 5 | 0 | **749** | 0 | **c=silence+noise** |
| 30 | 34 | 1 | **364** | **d=speech** |

**LPCC Default (LPC derived Cepstral Coefficients)**

Στο πείραμα με τα LPCC, δεν προκύπτει καλό ποσοστό επιτυχημένα ταξινομημένων τμημάτων ήχου. Σε αυτό το σετ, το ποσοστό επιτυχίας είναι 58.7965%, ενώ το ποσοστό αποτυχίας 41.2035%.

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1
Relation: LPCC_default.arff
Instances: 2393
Attributes: 49
Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 3.68 seconds

*=== Stratified cross-validation ===*
*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 1407 | 58.7965 % |
| Incorrectly Classified Instances | 986 | 41.2035 % |
| Kappa statistic | 0.4044 | |
| Mean absolute error | 0.206 | |
| Root mean squared error | 0.4539 | |
| Relative absolute error | 55.9904 % | |
| Root relative squared error | 105.8228 % | |
| Coverage of cases (0.95 level) | 58.7965 % | |
| Mean rel. region size (0.95 level) | 25% | |
| Total Number of Instances | 2393 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.999 | 0.593 | 0.428 | 0.999 | 0.599 | 0.703 | **music** |
| | 0 | 0 | 0 | 0 | 0 | 0.5 | **music+speech** |
| | 0.906 | 0.001 | 0.997 | 0.906 | 0.949 | 0.952 | **silence+noise** |
| | 0 | 0.001 | 0 | 0 | 0 | 0.499 | **speech** |
| **Weighted Avg.** | 0.588 | 0.183 | 0.441 | 0.588 | 0.479 | 0.703 | |

*=== Confusion Matrix ===*

Πραγματικά, όπως αποδεικνύεται από τον παρακάτω πίνακα, η μοναδική κατηγορία που ταξινομείται σωστά, χωρίς κανένα σχεδόν λάθος, είναι η κατηγορία silence+noise.

Αναλυτικά, στην κατηγορία music ταξινομούνται σωστά 735 τμήματα, όμως 486 μπερδεύονται ως music+speech, 68 ως silence+noise, και 428 ως speech.
Στην κατηγορία music+speech δεν υφίσταται καμία ταξινόμηση σε καμία από τις τέσσερις κατηγορίες που έχουν ορισθεί.
Στην κατηγορία silence+noise ταξινομούνται σωστά 672 τμήματα ήχου, ενώ λανθασμένα έχει ταξινομηθεί μόλις 1 για την κατηγορία music, κανένα για την κατηγορία music+speech, και 1 για την κατηγορία speech.
Στην κατηγορία speech δεν ταξινομείται τίποτα στις κατηγορίες music, music+speech, όπως και στην κατηγορία speech (στην οποία κανονικά θα έπρεπε να ταξινομηθούν περισσότερα τμήματα ήχου), ενώ αντιθέτως, ταξινομήθηκαν 2 αποσπάσματα ήχου στην κατηγορία silence+noise.

| **a** | **b** | **c** | **d** | **ç classified as** |
|---|---|---|---|---|
| **735** | 0 | 1 | 0 | **a=music** |
| 486 | **0** | 0 | 0 | **b=music+speech** |
| 68 | 0 | **672** | 2 | **c=silence+noise** |
| 428 | 0 | 1 | **0** | **d=speech** |

**LPCC – HopSamples 256**

Στο πείραμα LPCC – HopSamples 256 εξάγονται τα ίδια ακριβώς χαρακτηριστικά με τη βασική διαφορά ότι το μέγεθος των HopSamples είναι 256 σε αντίθεση με το default που είναι 512.

Το ποσοστό των επιτυχημένα ταξινομημένων αποσπασμάτων ήχου είναι ελαφρώς καλύτερο από το προηγούμενο πείραμα, κάτι το οποίο φαίνεται από το ποσοστό επιτυχίας που επιτυγχάνεται, το οποίο είναι 59.1684%, δηλαδή κατά 0.3719% καλύτερα από το προηγούμενο πείραμα.

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1
Relation: LPCC_HopSamples256.arff
Instances: 2405
Attributes: 49
Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 3.3 seconds

*=== Stratified cross-validation ===*
*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 1423 | 59.1684 % |
| Incorrectly Classified Instances | 982 | 40.8316 % |
| Kappa statistic | 0.4099 | |
| Mean absolute error | 0.2042 | |
| Root mean squared error | 0.4518 | |
| Relative absolute error | 55.5208 % | |
| Root relative squared error | 105.3781 % | |
| Coverage of cases (0.95 level) | 59.1684 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2405 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.999 | 0.588 | 0.428 | 0.999 | 0.6 | 0.705 | **music** |
| | 0.004 | 0 | 1 | 0.004 | 0.008 | 0.502 | **music+speech** |
| | 0.908 | 0.001 | 0.999 | 0.908 | 0.951 | 0.954 | **silence+noise** |
| | 0.002 | 0 | 1 | 0.002 | 0.005 | 0.501 | **speech** |
| **Weighted Avg.** | 0.592 | 0.18 | 0.825 | 0.592 | 0.484 | 0.706 | |

*=== Confusion Matrix ===*

Όπως φαίνεται από τον παρακάτω πίνακα οι κατηγορίες ήχου ταξινομούνται με σχεδόν ίδιο τρόπο όπως και στο προηγούμενο πείραμα.
Στην κατηγορία music ταξινομούνται σωστά 735 τμήματα ήχου, ενώ ταξινομούνται λάθος 484 ως music+speech, 69 ως silence+noise και 428 ως speech.
Στην κατηγορία music+speech ταξινομούνται σωστά μόνο 2, ενώ δεν ταξινομείται κανένα άλλο από τις υπόλοιπες κατηγορίες.
Στην κατηγορία silence+noise παρατηρείται σωστή ταξινόμηση των κατηγοριών, αφού έχουν ταξινομηθεί σωστά 685 τμήματα ήχου, ενώ 1 μόνο έχει ταξινομηθεί ως music και κανένα ως music+speech και ως speech.

Στην κατηγορία speech ταξινομείται σωστά μόνο 1 ενώ αντιθέτως στις υπόλοιπες κατηγορίες δεν ταξινομείται κανένα.

| a | b | c | d | ç classified as |
|---|---|---|---|---|
| **735** | 0 | 1 | 0 | **a=music** |
| 484 | **2** | 0 | 0 | **b=music+speech** |
| 69 | 0 | **685** | 0 | **c=silence+noise** |
| 428 | 0 | 0 | **1** | **d=speech** |

**LSP-Default**

Περίπου επίσης το ίδιο αποτέλεσμα με το προηγούμενο πείραμα βγάζει το πείραμα LSP με default χαρακτηριστικά.

Εδώ το ποσοστό επιτυχίας είναι 57.0414%.

Παρακάτω στον πίνακα “Confusion Matrix” που δημιουργείται από το λογισμικό WEKA, εξηγείται αναλυτικά το ποσοστό αυτό.

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1

Relation: LSP_default.arff

Instances: 2393

Attributes: 73

Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 5.05 seconds

*=== Stratified cross-validation ===*

*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 1365 | 57.0414 % |
| Incorrectly Classified Instances | 1028 | 42.9586 % |
| Kappa statistic | 0.379 | |
| Mean absolute error | 0.2148 | |
| Root mean squared error | 0.4635 | |
| Relative absolute error | 58.3754 % | |
| Root relative squared error | 108.0531 % | |

Coverage of cases (0.95 level) 57.0414 %

Mean rel. region size (0.95 level) 25 %

Total Number of Instances 2393

=== *Detailed Accuracy By Class* ===

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 1 | 0.619 | 0.418 | 1 | 0.59 | 0.691 | **music** |
| | 0 | 0 | 0 | 0 | 0 | 0.5 | **music+speech** |
| | 0.848 | 0.002 | 0.995 | 0.848 | 0.916 | 0.923 | **silence+noise** |
| | 0 | 0 | 0 | 0 | 0 | 0.5 | **speech** |
| **Weighted Avg.** | 0.87 | 0.191 | 0.437 | 0.57 | 0.465 | 0.69 | |

=== ***Confusion Matrix*** ===

Όπως και στο προηγούμενο πείραμα (LPCC – HopSamples 256), ο πίνακας παρουσιάζει σχεδόν την ίδια μορφή.

Τα δείγματα ήχου είναι ανακατεμένα και ταξινομημένα λάθος στη στήλη a, η οποία αντιστοιχεί στην κατηγορία “music”.

Συγκεκριμένα, ταξινομούνται σωστά 736 στην κατηγορία music, ενώ λάθος ταξινομούνται 486 στην κατηγορία music+speech, 113 στην κατηγορία silence+noise και 426 στην κατηγορία speech.

Κανένα δεν ταξινομείται στην κατηγορία speech.

Όλα αυτά τα αριθμητικά στοιχεία που υπάρχουν στον πίνακα φανερώνουν και αποδεικνύουν το ποσοστό 57.0414% που προαναφέρθηκε προηγουμένως.

| **a** | **b** | **c** | **d** | **ç classified** |
|---|---|---|---|---|
| **736** | 0 | 0 | 0 | **a=music** |
| 486 | **0** | 0 | 0 | **b=music+speech** |
| 113 | 0 | **629** | 0 | **c=silence+noise** |
| 426 | 0 | 3 | **0** | **d=speech** |

**LSP – HopSamples 256**

Πραγματοποιώντας στο WEKA το ίδιο πείραμα, χρησιμοποιώντας τις ίδιες ακριβώς παραμέτρους, αλλά μειώνοντας το μέγεθος των HopSamples από 512 σε 256, προκύπτει το ποσοστό επιτυχίας 57,3389%, δηλαδή κατά 0,2975% καλύτερα από το Default LSP set, που είδαμε προηγουμένως.

Η ταξινόμηση των κατηγοριών εξηγείται από τον πίνακα confusion matrix, που βρίσκεται παρακάτω.

=== Run information ===

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1

Relation: LSP_HopSamples256.arff

Instances: 2405

Attributes: 73

Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 5.86 seconds

*=== Stratified cross-validation ===*

*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 1379 | 57.3389 % |
| Incorrectly Classified Instances | 1026 | 42.6611 % |
| Kappa statistic | 0.3835 | |
| Mean absolute error | 0.2133 | |
| Root mean squared error | 0.4619 | |
| Relative absolute error | 58.0085 % | |
| Root relative squared error | 107.713 % | |
| Coverage of cases (0.95 level) | 57.3389 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2405 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 1 | 0.614 | 0.418 | 1 | 0.59 | 0.693 | **music** |
| | 0 | 0 | 0 | 0 | 0 | 0.5 | **music+speech** |
| | 0.853 | 0.001 | 0.998 | 0.853 | 0.92 | 0.926 | **silence+noise** |
| | 0 | 0 | 0 | 0 | 0 | 0.5 | **speech** |
| **Weighted Avg.** | 0.573 | 0.188 | 0.441 | 0.573 | 0.469 | 0.693 | |

*=== Confusion Matrix ===*

Σε αυτόν τον πίνακα το συμπέρασμα που μπορεί να βγεί είναι ότι ο αλγόριθμος LSP με 256 HopSamples, είναι εξαιρετικά λειτουργικός στο να απομονώσει τα κομμάτια ήχου τα οποία έχουν το label «silence+noise», κάτι το οποίο σημαίνει ότι τα τμήματα ήχου τα οποία αποτελούνται από σιωπή ή θόρυβο διαχωρίστηκαν με μεγαλύτερη επιτυχία από τις υπόλοιπες κατηγορίες.

Συγκεκριμένα, στην κατηγορία music δεν ταξινομούνται σωστά οι κατηγορίες ήχου.

Με σωστό τρόπο ταξινομούνται ως “music” 736 τμήματα ήχου, ενώ το σύστημα «μπερδεύεται» και ταξινομεί 486 τμήματα ήχου ως “music+speech”, 111 ως “silence+noise” και 428 ως “speech”.

Για την κατηγορία “music+speech” δεν ταξινομείται καθόλου κανένα ηχητικό απόσπασμα.

Ο ταξινομητής δουλεύει σωστά για την κατηγορία “silence+noise” μιας και ταξινομούνται σωστά 643 τμήματα ήχου ως “silence+noise” ενώ μόλις 1 μπερδεύεται και ταξινομείται λανθασμένα ως “speech”.

Τέλος, στην κατηγορία “speech”, συμβαίνει το ίδιο ακριβώς με την κατηγορία “music+speech”, δηλαδή δεν ταξινομείται κανένα απόσπασμα ήχου.

| a | b | c | d | ç classified as |
|---|---|---|---|---|
| **736** | 0 | 0 | 0 | **a=music** |
| 486 | **0** | 0 | 0 | **b=music+speech** |
| 111 | 0 | **643** | 0 | **c=silence+noise** |
| 428 | 0 | 1 | **0** | **d=speech** |

**MFCC – Default**

Σε αυτό το πείραμα, εξάγονται τα MFCC χαρακτηριστικά του ήχου (Mel Frequency Cepstral Coefficients) και παρατηρούμε πως το ποσοστό επιτυχίας που προκύπτει είναι εξαιρετικό. Το ποσοστό σε αυτή την περίπτωση αγγίζει το 98,412%, το οποίο είναι κατά πολύ καλύτερο από τα ποσοστά που προέκυψαν από τα προηγούμενα πειράματα.

Το αποτέλεσμα αυτό προκύπτει από τον πίνακα Confusion Matrix παρακάτω, και στον οποίο διαβάζουμε αναλυτικά τον τρόπο με τον οποίο ταξινομούνται τα τμήματα ήχου με βάση το “labeling” που τους έχει δοθεί.

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1

Relation: MFCC_Default.arff

Instances: 2393

Attributes: 53

Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 1.24 seconds

*=== Stratified cross-validation ===*

*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 2355 | 98.412 % |
| Incorrectly Classified Instances | 38 | 1.588 % |
| Kappa statistic | 0.9784 | |
| Mean absolute error | 0.0079 | |
| Root mean squared error | 0.0891 | |
| Relative absolute error | 2.1578 % | |
| Root relative squared error | 20.7746 % | |
| Coverage of cases (0.95 level) | 98.412 % | |
| Mean rel. region size (0.95 level) | 25% | |
| Total Number of Instances | 2393 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.988 | 0.004 | 0.992 | 0.988 | 0.99 | 0.992 | **music** |
| | 0.967 | 0.006 | 0.975 | 0.967 | 0.971 | 0.98 | **music+speech** |
| | 0.999 | 0.002 | 0.995 | 0.999 | 0.997 | 0.998 | **silence+noise** |
| | 0.972 | 0.008 | 0.963 | 0.972 | 0.968 | 0.982 | **speech** |
| **Weighted Avg.** | 0.984 | 0.005 | 0.984 | 0.984 | 0.984 | 0.99 | |

*=== Confusion Matrix ===*

Στον πίνακα Confusion Matrix, η ταξινόμηση των ηχητικών κατηγοριών γίνεται ως ακολούθως:

Για την κατηγορία “music”, ταξινομούνται σωστά 727 τμήματα ήχου, ενώ λανθασμένα ταξινομούνται 4 ως “music+speech”, κανένα ως “silence+noise” και 2 ως “speech”.

Για την κατηγορία “music+speech” ταξινομούνται σωστά 470 τμήματα ήχου, ενώ λανθασμένα ταξινομούνται 5 ως “music”, κανένα ως “silence+noise” και 7 ταξινομούνται ως “speech”.

Για την κατηγορία “silence+noise” ταξινομούνται σωστά 741 τμήματα ήχου, ενώ λανθασμένα ταξινομούνται 1 ως “music”, κανένα ως “music+speech” και 3 ως “speech”.

Για την κατηγορία “speech” ταξινομούνται σωστά 417 τμήματα ήχου, ενώ λανθασμένα ταξινομούνται 3 ως “music”, 12 ως “music”, και 1 ως “silence+noise”.

Επομένως από τα 2393 τμήματα, ταξινομούνται σωστά τα 2355 και λανθασμένα μόνο τα 38.

| a | b | c | d | ç classified as |
|---|---|---|---|---|
| **727** | 5 | 1 | 3 | **a=music** |
| 4 | **470** | 0 | 12 | **b=music+speech** |
| 0 | 0 | **741** | 1 | **c=silence+noise** |
| 2 | 7 | 3 | **417** | **d=speech** |

**MFCC – HopSamples 256**

Στο πείραμα αυτό (MFCC-HopSamples 256), το αποτέλεσμα το οποίο προκύπτει είναι ελαφρώς χειρότερο από το πείραμα “MFCC”, στο οποίο τα HopSamples είχαν διπλάσιο μέγεθος (512). Εδώ το ποσοστό επιτυχίας που επιτυγχάνεται είναι το 98.2536%, έναντι του 98.412%.

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1

Relation: MFCC_HopSamples256.arff

Instances: 2405

Attributes: 53

Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 1.13 seconds

*=== Stratified cross-validation ===*

*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 2363 | 98.2536 % |
| Incorrectly Classified Instances | 42 | 1.7464 % |
| Kappa statistic | 0.9763 | |
| Mean absolute error | 0.0087 | |
| Root mean squared error | 0.0934 | |
| Relative absolute error | 2.3746 % | |
| Root relative squared error | 21.7931 % | |
| Coverage of cases (0.95 level) | 98.2536 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2405 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.988 | 0.004 | 0.992 | 0.988 | 0.99 | 0.992 | **music** |
| | 0.967 | 0.006 | 0.975 | 0.967 | 0.971 | 0.98 | **music+speech** |
| | 0.999 | 0.002 | 0.995 | 0.999 | 0.997 | 0.998 | **silence+noise** |
| | 0.972 | 0.008 | 0.963 | 0.972 | 0.968 | 0.982 | **speech** |
| **Weighted Avg.** | 0.984 | 0.005 | 0.984 | 0.984 | 0.984 | 0.99 | |

*=== Confusion Matrix ===*

Στον πίνακα Confusion Matrix, εξηγείται το ποσοστό επιτυχίας που επιτυγχάνεται με το πείραμα MFCC-HopSamples 256, δηλαδή το 98.2536%

Συγκεκριμένα, από τα 2405 τμήματα ήχου ταξινομούνται σωστά τα 2363.

Για την κατηγορία “music” ταξινομούνται σωστά 726 τμήματα ήχου, ενώ λανθασμένα ταξινομούνται 4 ως “music+speech”, κανένα ως “silence+noise” και 1 ως “speech”.

Για την κατηγορία “music+speech” ταξινομούνται σωστά 463 τμήματα ήχου, ενώ λανθασμένα ταξινομούνται 5 ως “music”, κανένα ως “silence+noise” και 3 ως “speech”.

Για την κατηγορία “silence+noise”, ταξινομούνται σωστά 753 τμήματα ήχου, ενώ λανθασμένα ταξινομούνται 3 ως “music”, κανένα ως “music+speech” και 4 ως “speech”.

Για την κατηγορία “speech” ταξινομούνται σωστά 421 τμήματα ήχου, ενώ λανθασμένα ταξινομούνται 2 ως “music”, 19 ως “music+speech”, και 1 ως “silence+noise”.

| a | b | c | d | ç classified as |
|---|---|---|---|---|
| **726** | 5 | 3 | 2 | **a=music** |
| 4 | **463** | 0 | 19 | **b=music+speech** |
| 0 | 0 | **753** | 1 | **c=silence+noise** |
| 1 | 3 | 4 | **421** | **d=speech** |

**SFMSCF – Default**

Στο πείραμα SFMSCF – Default παρατηρούμε πως το ποσοστό επιτυχίας 95,8629% είναι αρκετά μεγάλο, κάτι το οποίο επαληθεύεται από τον πίνακα Confusion Matrix παρακάτω.

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1

Relation: SFMSCF_Default.arff

Instances: 2393

Attributes: 193

Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 1.99 seconds

*=== Stratified cross-validation ===*

*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 2294 | 95.8629 % |
| Incorrectly Classified Instances | 99 | 4.1371 % |
| Kappa statistic | 0.9437 | |
| Mean absolute error | 0.0207 | |
| Root mean squared error | 0.1438 | |
| Relative absolute error | 5.6218 % | |
| Root relative squared error | 33.5319 % | |
| Coverage of cases (0.95 level) | 95.8629 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2393 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.981 | 0.012 | 0.973 | 0.981 | 0.977 | 0.984 | **music** |
| | 0.897 | 0.012 | 0.95 | 0.897 | 0.923 | 0.943 | **music+speech** |
| | 0.987 | 0.012 | 0.975 | 0.987 | 0.981 | 0.988 | **silence+noise** |
| | 0.942 | 0.019 | 0.916 | 0.942 | 0.929 | 0.961 | **speech** |
| **Weighted Avg.** | 0.959 | 0.013 | 0.959 | 0.959 | 0.958 | 0.973 | |

*=== Confusion Matrix ===*

Στον πίνακα Confusion Matrix για το πείραμα SFMSCF-Default ταξινομούνται σωστά 2294 από τα 2393 τμήματα ήχου.

Συγκεκριμένα:

Για την κατηγορία “music” ταξινομούνται σωστά 722 τμήματα ήχου ενώ λανθασμένα ταξινομούνται, 14 ως “music+speech”, 6 ως “silence+noise”, και κανένα ως “speech”.

Για την κατηγορία “music+speech”“music+speech” ταξινομούνται σωστά 436 τμήματα

| **a** | **b** | **c** | **d** | **ç classified as** |
|---|---|---|---|---|
| **722** | 11 | 2 | 1 | **a=music** |
| 14 | **436** | 1 | 35 | **b=music+speech** |
| 6 | 3 | **732** | 1 | **c=silence+noise** |
| 0 | 9 | 16 | **404** | **d=speech** |

**SFMSCF – HopSamples 256**

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1

Relation: SFMSCF_HopsSamples256.arff

Instances: 2405

Attributes: 193

Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 2.39 seconds

*=== Stratified cross-validation ===*
*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 2316 | 96.2994 % |
| Incorrectly Classified Instances | 89 | 3.7006 % |
| Kappa statistic | 0.9497 | |
| Mean absolute error | 0.0185 | |
| Root mean squared error | 0.136 | |
| Relative absolute error | 5.0319 % | |
| Root relative squared error | 31.7241 % | |
| Coverage of cases (0.95 level) | 96.2994 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2405 | |

=== Detailed Accuracy By Class ===

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.985 | 0.004 | 0.992 | 0.985 | 0.988 | 0.991 | **music** |
| | 0.905 | 0.011 | 0.954 | 0.905 | 0.929 | 0.947 | **music+speech** |
| | 0.992 | 0.01 | 0.979 | 0.992 | 0.986 | 0.991 | **silence+noise** |
| | 0.939 | 0.023 | 0.898 | 0.939 | 0.918 | 0.958 | **speech** |
| **Weighted Avg.** | 0.963 | 0.011 | 0.963 | 0.963 | 0.963 | 0.976 | |

*=== Confusion Matrix ===*

| **a** | **b** | **c** | **d** | **ç classified as** |
|---|---|---|---|---|
| **725** | 7 | 3 | 1 | **music** |
| 5 | **440** | 0 | 41 | **music+speech** |
| 1 | 1 | **748** | 4 | **silence+noise** |
| 0 | 13 | 13 | **403** | **speech** |

**STFT – Default**

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1
Relation: STFT_Default.arff

Instances: 2393
Attributes: 13
Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 1.39 seconds

*=== Stratified cross-validation ===*
*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 1331 | 55.6206 % |
| Incorrectly Classified Instances | 1062 | 44.3794 % |
| Kappa statistic | 0.3606 | |
| Mean absolute error | 0.2219 | |
| Root mean squared error | 0.4711 | |
| Relative absolute error | 60.3061 % | |
| Root relative squared error | 109.8255 % | |
| Coverage of cases (0.95 level) | 55.6206 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2393 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.999 | 0.614 | 0.42 | 0.999 | 0.591 | 0.692 | **music** |
| | 0.004 | 0.02 | 0.049 | 0.004 | 0.008 | 0.492 | **music+speech** |
| | 0.791 | 0.002 | 0.993 | 0.791 | 0.881 | 0.894 | **silence+noise** |
| | 0.016 | 0.001 | 0.778 | 0.016 | 0. 032 | 0.508 | **speech** |
| **Weighted Avg.** | 0.556 | 0.194 | 0.586 | 0.556 | 0.462 | 0.681 | |

*=== Confusion Matrix ===*

| **a** | **b** | **c** | **d** | **ç classified as** |
|---|---|---|---|---|
| **735** | 1 | 0 | 0 | **a=music** |
| 480 | **2** | 2 | 2 | **b=music+speech** |

| | | | | |
|---|---|---|---|---|
| 152 | 3 | **587** | 0 | **c=silence+noise** |
| 385 | 35 | 2 | **7** | **d=speech** |

**STFT – HopSamples 256**

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1

Relation: STFT_HopSamples256.arff

Instances: 2405

Attributes: 13

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 1.22 seconds

*=== Stratified cross-validation ===*

*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 1382 | 57.4636 % |
| Incorrectly Classified Instances | 1023 | 42.5364 % |
| Kappa statistic | 0.3915 | |
| Mean absolute error | 0.2127 | |
| Root mean squared error | 0.4612 | |
| Relative absolute error | 57.8389 % | |
| Root relative squared error | 107.5554 % | |
| Coverage of cases (0.95 level) | 57.4636 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2405 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 1 | 0.56 | 0.44 | 1 | 0.612 | 0.72 | **music** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0.021 | 0.035 | 0.128 | 0.021 | 0.035 | 0.493 | **music+speech** |
| | 0.779 | 0.002 | 0.995 | 0.779 | 0.874 | 0.888 | **silence+noise** |
| | 0.114 | 0.009 | 0.742 | 0.114 | 0.198 | 0.553 | **speech** |
| **Weighted Avg.** | 0.575 | 0.181 | 0.605 | 0.575 | 0.503 | 0.697 | |

*=== Confusion Matrix ===*

| **a** | **b** | **c** | **d** | **ç classified as** |
|---|---|---|---|---|
| **736** | 0 | 0 | 0 | **a=music** |
| 469 | **10** | 1 | 6 | **b=music+speech** |
| 148 | 8 | **587** | 11 | **c=silence+noise** |
| 318 | 60 | 2 | **49** | **d=speech** |

**STFTMFCC – Default**

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1

Relation: STFTMFCC_Default.arff

Instances: 2393

Attributes: 65

Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 1.55 seconds

*=== Stratified cross-validation ===*

*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 2349 | 98.1613 % |
| Incorrectly Classified Instances | 44 | 1.8387 % |
| Kappa statistic | 0.975 | |
| Mean absolute error | 0.0092 | |
| Root mean squared error | 0.0959 | |
| Relative absolute error | 2.4986 % | |
| Root relative squared error | 22.3546 % | |

Coverage of cases (0.95 level) 98.1613 %
Mean rel. region size (0.95 level) 25 %
Total Number of Instances 2393

*=== Detailed Accuracy By Class ===*

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|---|
| | 0.986 | 0 .003 | 0.993 | 0.986 | 0.99 | 0.992 | **music** |
| | 0.955 | 0 .006 | 0.975 | 0.955 | 0.965 | 0.974 | **music+speech** |
| | 0.999 | 0 .003 | 0.993 | 0.999 | 0.996 | 0.998 | **silence+noise** |
| | 0.974 | 0 .011 | 0.95 | 0.974 | 0.962 | 0.982 | **speech** |
| **Weighted Avg.** | 0.982 | 0 .005 | 0.982 | 0.982 | 0.982 | 0.988 | |

*=== Confusion Matrix ===*

| a | b | c | d | ç classified as |
|---|---|---|---|---|
| **726** | 6 | 1 | 3 | **a=music** |
| 4 | **464** | 0 | 18 | **b=music+speech** |
| 0 | 0 | **741** | 1 | **c=silence+noise** |
| 1 | 6 | 4 | **418** | **d=speech** |

**STFTMFCC – HopSamples 256**

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1
Relation: STFTMFCC_HopSamples256.arff
Instances: 2405
Attributes: 65
Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 1.35 seconds

*=== Stratified cross-validation ===*

*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 2359 | 98.0873 % |
| Incorrectly Classified Instances | 46 | 1.9127 % |
| Kappa statistic | 0.974 | |
| Mean absolute error | 0.0096 | |
| Root mean squared error | 0.0978 | |
| Relative absolute error | 2.6008 % | |
| Root relative squared error | 22.8073 % | |
| Coverage of cases (0.95 level) | 98.0873 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2405 | |

*=== Detailed Accuracy By Class ===*

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|---|
| | 0.985 | 0.003 | 0.993 | 0.985 | 0.989 | 0.991 | **music** |
| | 0.947 | 0.005 | 0.981 | 0.947 | 0.963 | 0.971 | **music+speech** |
| | 0.999 | 0.004 | 0.991 | 0.999 | 0.995 | 0.997 | **silence+noise** |
| | 0.981 | 0.013 | 0.944 | 0.981 | 0.962 | 0.984 | **speech** |
| **Weighted Avg.** | 0.981 | 0.005 | 0.981 | 0.981 | 0.981 | 0.988 | |

*=== Confusion Matrix ===*

| a | b | c | d | ç classified as |
|---|---|---|---|---|
| **725** | 6 | 3 | 2 | **a=music** |
| 4 | **460** | 0 | 22 | **b=music+speech** |
| 0 | 0 | **753** | 1 | **c=silence+noise** |
| 1 | 3 | 4 | **421** | **d=speech** |

**Combination of ALL Above - Default**

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1

Relation: ALLCombination_Default.arff

Instances: 2393
Attributes: 377
Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 12.47 seconds

*=== Stratified cross-validation ===*
*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 1573 | 65.7334 % |
| Incorrectly Classified Instances | 820 | 34.2666 % |
| Kappa statistic | 0.5091 | |
| Mean absolute error | 0.1713 | |
| Root mean squared error | 0.4139 | |
| Relative absolute error | 46.5641 % | |
| Root relative squared error | 96.5045 % | |
| Coverage of cases (0.95 level) | 65.7334 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2393 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.995 | 0.486 | 0.476 | 0.995 | 0.644 | 0.754 | **music** |
| | 0.154 | 0.004 | 0.915 | 0.154 | 0.264 | 0.575 | **music+speech** |
| | 0.964 | 0.003 | 0.993 | 0.964 | 0.978 | 0.98 | **silence+noise** |
| | 0.119 | 0.001 | 0.962 | 0.119 | 0.212 | 0.559 | **speech** |
| **Weighted Avg.** | 0.657 | 0.151 | 0.813 | 0.657 | 0.593 | 0.753 | |

*=== Confusion Matrix ===*

| **a** | **b** | **c** | **d** | **ç classified as** |
|---|---|---|---|---|
| **732** | 1 | 3 | 0 | **a=music** |
| 409 | **75** | 0 | 2 | **b=music+speech** |
| 27 | 0 | **715** | 0 | **c=silence+noise** |

| | | | | |
|---|---|---|---|---|
| 370 | 6 | 2 | **51** | **d=speech** |

**Combination of ALL Above – HopSamples 256**

*=== Run information ===*

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1

Relation: ALLCombination_HopSamples256.arff

Instances: 2405

Attributes: 377

Test mode: 10-fold cross-validation

*=== Classifier model (full training set) ===*

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 12.47 seconds

*=== Stratified cross-validation ===*

*=== Summary ===*

| | | |
|---|---|---|
| Correctly Classified Instances | 1627 | 67.6507 % |
| Incorrectly Classified Instances | 778 | 32.3493 % |
| Kappa statistic | 0.5382 | |
| Mean absolute error | 0.1617 | |
| Root mean squared error | 0.4022 | |
| Relative absolute error | 43.9869 % | |
| Root relative squared error | 93.796 % | |
| Coverage of cases (0.95 level) | 67.6507 % | |
| Mean rel. region size (0.95 level) | 25 % | |
| Total Number of Instances | 2405 | |

*=== Detailed Accuracy By Class ===*

| | **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **ROC Area** | **Class** |
|---|---|---|---|---|---|---|---|
| | 0.99 | 0.451 | 0.492 | 0.99 | 0.657 | 0.77 | **music** |
| | 0.241 | 0.007 | 0.893 | 0.241 | 0.379 | 0.617 | **music+speech** |
| | 0.967 | 0.002 | 0.995 | 0.967 | 0.98 | 0.982 | **silence+noise** |
| | 0.121 | 0.004 | 0.881 | 0.121 | 0.213 | 0.559 | **speech** |
| **Weighted** | 0.677 | 0.141 | 0.8 | 0.677 | 0.623 | 0.768 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Avg.** | | | | | | | |

*=== Confusion Matrix ===*

| a | b | c | d | ç classified as |
|---|---|---|---|---|
| **729** | 4 | 3 | 0 | **a=music** |
| 363 | **117** | 0 | 6 | **b=music+speech** |
| 24 | 0 | **729** | 1 | **c=silence+noise** |
| 366 | 10 | 1 | **52** | **d=speech** |

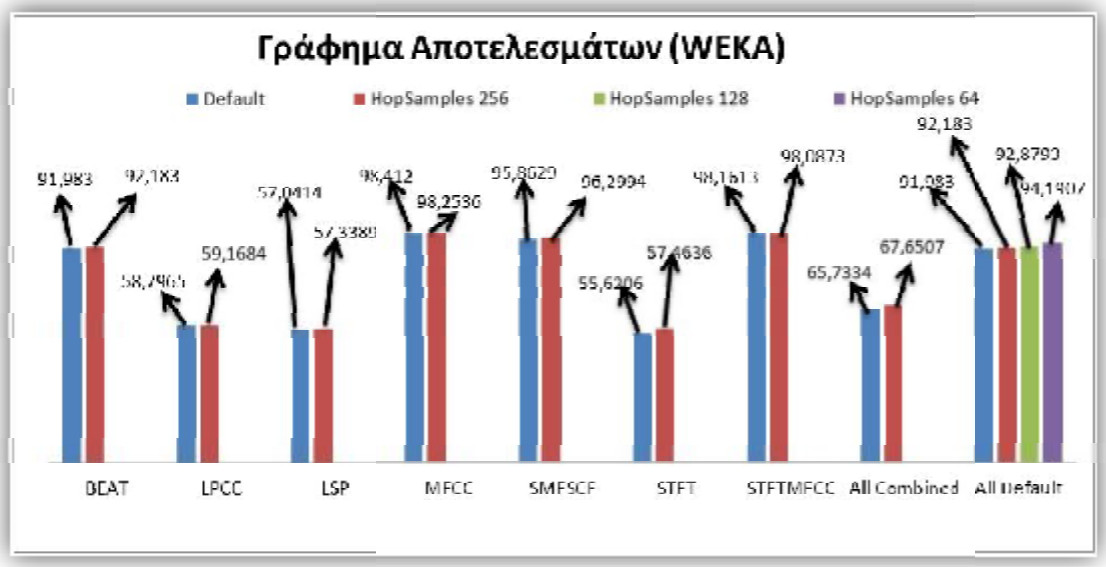
# 4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε αυτό το κεφάλαιο παρουσιάζονται αναλυτικά όλα τα συμπεράσματα τα οποία εξήχθησαν βάσει της πειραματικής διαδικασίας, η οποία ακολουθήθηκε με σκοπό να γίνει η κατάτμηση του ήχου με τον πιο εύκολο, πιο σύντομο και τον λιγότερο ενεργοβόρο τρόπο.

Αποδείχθηκε πως μια ραδιοφωνική εκπομπή που περιλαμβάνει μουσική, ομιλία, θόρυβο και διαστήματα σιωπής, μπορεί να απομονωθεί μέσω του προγράμματος Praat, να εξαχθούν τα επιμέρους χαρακτηριστικά του ήχου και να αναλυθούν μέσω του προγράμματος ανοιχτού κώδικα Marsyas και τέλος να γίνει Data-Mining μέσω του λογισμικού WEKA, με τη βοήθεια του οποίου μπορούμε να εκπαιδεύσουμε ένα υπολογιστικό σύστημα το οποίο θα αντιλαμβάνεται ως αποτέλεσμα τα διαφορετικά είδη που απαρτίζουν τον ήχο σε μια ραδιοφωνική εκπομπή.

Τα στοιχεία που εξάγονται, η Βάση Δεδομένων που δημιουργήθηκε, καθώς και τα τελικά αρχεία και αποτελέσματα μπορούν να αποτελέσουν το «θεμέλιο λίθο» για περαιτέρω έρευνα και αξιοποίησή τους από την επιστημονική κοινότητα, με σκοπό την ανάπτυξη συστημάτων για την αναγνώριση ηχητικών κατηγοριών στο περιβάλλον του ραδιοφώνου με αυτόματο τρόπο.

Στο παρακάτω γράφημα παρουσιάζονται τα αποτελέσματα από την πειραματική διαδικασία συνοπτικά, όπου μπορεί να διαπιστωθεί ποια μέθοδος έδωσε τα καλύτερα αποτελέσματα στην κατηγοριοποίηση των ήχων.




Εικόνα 4-1: Γράφημα αποτελεσμάτων

# ΑΝΑΦΟΡΕΣ


B. Bigot, I. Ferrane, & J. Pinquer. (2010, January ). Speaker Role Recognition to help Spontaneous Conversational Speech Detection. σσ. 1-7.

Bachu, R., Adapa, B., & Barkana, B. D. (March 2008). Separation of Voiced and Unvoiced Speech Signals using Energy and Zero Crossing Rate. *ASEE Regional Conference.*

Bouko, T., & Nadeu, C. (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.

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

D. Michie, D.J. Spiegelhalter, & C. Taylor. (1994). *Machine Learning, Neural and Statistical.* New York: Ellis Horwood.

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).

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).

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* (σσ. 379 - 383). IC-NIDC 2009.

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

Reese, D., Gross, L., & Gross, B. (Έκδοση 5, Αναθεωρημένη 2012). *Radio Production Worktext: Studio And Equipment.*

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

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

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

T. Jehan. (2005). *Creating Music by Listening.* Massachusetts: Massachusetts Institute of Technology.

TrackID. (n.d.). *TrackID*. Ανάκτηση από https://trackid.sonymobile.com.

Transform, F. F. (n.d.). Ανάκτηση από http://mathworld.wolfram.com/FastFouri.

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 Ανάκτηση από http://www.marsyas.info/pdf/Marsyas0.2_UserManual.pdf.

WEKA. (n.d.). *http://www.cs.waikato.ac.nz/ml/weka/index.html*. Ανάκτηση από Weka 3.7: Data Mining Software in Java.

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)*, 646 - 649.

Zooms, T. &. (n.d.). *http://www.nuance.com/for-individuals/mobile-applications/talks-zooms/index.htm*.

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

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