Τεχνολογία Ήχου και Εικόνας 2018
Εργασία 2018-2019

Χριστίνα Θεοδωρίδου - 8055
Φρανκ Μπλάννινγκ - 6698
Αποστόλης Φανάκης - 8261

18 Ιανουαρίου 2019

Περιεχόμενα

1 Εισαγωγή

Το ζητούμενο της εργασίας είναι η ανάπτυξη ενός μοντέλου μηχανικής μάθησης το οποίο, παρέχοντας ένα αρχείο ήχου, θα μπορεί να ξεχωρίσει ανάμεσα στα κομμάτια του χρόνου που περιέχουν ομιλία (speech) και μουσική (music), όπως παρουσιάζεται στον διαγωνισμό MIREX 2018:Music and/or Speech Detection [1] . Η εργασία επικεντρώνεται στην εύρεση των δειγμάτων που περιέχουν είτε ομιλία είτε μουσική και στην ταξινόμησή τους.

Πρόκειται για ένα δυαδικό πρόβλημα ταξινόμησης που είναι σημαντικό καθώς έχει εφαρμογές σε πλατφόρμες κοινωνικών δικτύων για την αναγνώριση περιεχομένου με πνευματικά δικαιώματά, σε συστήματα αυτόματης αναγνώρισης διαφημίσεων, μοντέρνα "έξυπνα" βοηθητικά ακοής κ.α. Η πρόσφατη βιβλιογραφία περιέχει θεματολογία όπου στοχεύει είτε στην ανάπτυξή αλγορίθμων για γρήγορη και φθηνή υπολογιστικά ταξινόμηση, είτε στην αναγνώριση πολύ μεγάλης ακρίβειας. Αυτό διότι αυτή τη στιγμή η αναγνώριση με ποσοστό επιτυχίας γύρω στο 98% είναι κάτι συνηθισμένο.

2 Προηγούμενες υλοποιήσεις

Υπάρχει πληθώρα βιβλιογραφίας σχετική με το θέμα. Έχουν βρεθεί ήδη αρκετές λύσεις, ενώ οι πιο πρόσφατες πετυχαίνουν αξιοσημείωτα αποτελέσματα τόσο όσων αφορά την ταχύτητα του διαχωρισμού όσο και την ακρίβεια των αποτελεσμάτων. Κάποιες από τις δημοσιεύσεις οι οποίες αφορούν το συγκεκριμένο θέμα, καθώς και τα αποτελέσματά τους παρουσιάζονται παρακάτω.

Στο [1] οι συγγραφείς χρησιμοποιούν τα εξής χαρακτηριστικά (features):

1. Διαμόρφωση ενέργειας στα 4Hz του σήματος (4Hz modulation)
2. Διαμόρφωση εντροπίας του σήματος (entropy modulation)
3. Αριθμός των στατικών τμημάτων
4. Διάρκεια των τμημάτων

Παρατηρήθηκε πειραματικά ότι τα πρώτα 3 χαρακτηριστικά δίνουν ξεχωριστά περίπου το ίδιο ποσοστό επιτυχών ταξινομήσεων (περίπου 84%) ενώ η Μπαγιεσιανή προσέγγιση για το χαρακτηριστικό δίαρκειας τμημάτων έδωσε λίγο χαμηλότερο ποσοστό (76.1%).

Για να αυξηθεί το ποσοστό των συνολικών επιτυχών ταξινομήσεων προτάθηκε ένας ιεραρχικός αλγόριθμος ταξινόμησης στον οποίο τα χαρακτηριστικά διαμόρφωσης ενέργειας του σήματος στα 4Hz και διαμόρφωσης εντροπίας του σήματος συγχωνεύονται. Σε περίπτωση που οι 2 ταξινομητές συμφωνούν αποφασίζουν για το αν το τμήμα αποτελεί ομιλία ή όχι, ενώ σε περίπτωση που δεν συμφωνούν, η απόφαση οριστικοποιείται από το χαρακτηριστικό του αριθμού τμημάτων. Αποδεικνύεται ότι τα αποτελέσματα αυτού του αλγορίθμου δίνουν 90.1% σωστές ταξινομήσεις.

Στο [2] το πρόβλημα που δόθηκε αντιμετωπίζεται ως 2 υποπροβλήματα: το πρόβλημα εντοπισμού δειγμάτων και το πρόβλημα κατηγοριοποίησής τους. Για τον εντοπισμό δειγμάτων μουσικής/φωνής εφαρμόστηκε ο αλγόριθμος Random Forest σε 2 εκδοχές του: στην πρώτη, εφαρμόστηκε μαζί με έναν Silence detection αλγόριθμο ενώ στη δεύτερη βασίστηκε μόνο στις πληροφορίες ομοιογένειας (self similarity matrix) και στην λειτουργία του ίδιου του ταξινομητή. Επίσης, για την ταξινόμηση προτάθηκαν 2 εναλλακτικές: στην πρώτη χρησιμοποιήθηκε ένα προ-εκπαιδευμένο μοντέλο ενώ στην δεύτερη η εκπαίδευση γίνεται κατά την αξιολόγηση των δειγμάτων.

Χρησιμοποιήθηκαν τα χαρακτηριστικά (features):

1. RMS ενέργεια
2. ZCR (Zero-Crossing Rate)
3. Spectral rolloff (Συχνότητα Αποκοπής)
4. Spectral flux (Φασματική Ροή)
5. Spectral flatness (Φασματική Επιπεδότητα)
6. Spectral flatness per Band (Φασματική Επιπεδότητα ανά συχνοτικές ομάδες)

[1] https://www.music-ir.org/mirex/wiki/2018:Music_and/or_Speech_Detection

7. MFCCs (Mel Frequency Cepstral Coefficients)

Έγινε ανάλυση κύριων συνιστωσών (Principal component analysis ή PCA) με στόχο να μειωθούν οι διαστάσεις των διανυσμάτων χαρακτηριστικών (feature vectors). Δημιουργήθηκαν οι πίνακες ομοιότητας υπολογίζοντας την ευκλείδεια απόσταση μεταξύ των δειγμάτων ήχου έτσι ώστε να χωριστούν τα τμήματα. Στη συνέχεια τα τμήματα αυτά κατηγοριοποιούνται ενώ ταυτόχρονα εφαρμόζεται ο αλγόριθμος Silence Detection και τα δείγματα αυτά προστίθενται στα προηγούμενα. Για το πρόβλημα της κατηγοριοποίησης χρησιμοποιείται ο ίδιος αλγόριθμος Random Forest για την ταξινόμηση σε επίπεδο (frame) τμημάτων ήχου. Εφόσον για κάθε αρχείο ήχου έχουν εξαχθεί τα παραπάνω χαρακτηριστικά, κάθε τμήμα ήχου ταξινομείται στην κλάση που αποφασίζεται και έπειτα ολόκληρο το αρχείο ταξινομείται στην κλάση στην οποία ταξινομήθηκαν τα τμήματά του κατά πλειοψηφία.

Στο [3] προτείνεται πως τα features μπορεί να μην καλύπτουν χαρακτηριστικά και της φωνής και της μουσικής, αλλά να βασίζονται κυρίως σε χαρακτηριστικά ενός από τα δύο. Ενδιαφέρον παρουσιάζουν τα χαρακτηριστικά της ομιλίας, τα οποία λόγω των μέσων που την παράγουν (τα χείλη, η γλώσσα και οι φωνητικές χορδές) έχουν ιδιαίτερα γνωρίσματα. Η μελέτη αυτών των χαρακτηριστικών και η χρήση τους ως features σε έναν classifier αποδεικνύεται πως μπορεί να αυξήσει την επιτυχία του διαχωρισμού.

Ενδεικτικά, πέρα από το καθιερωμένο feature των 4Hz modulation energy, λόγω του ρυθμού των συλλαβών, κάποια άλλα speech specific features βασίζονται στην αναγνώριση του ήχου που παράγεται στις φωνητικές χορδές κατά την εναλλαγή της προφοράς ενός συμφώνου σε ένα φωνήεν ή στην μελέτη της αυτοσυσχέτησης του σήματος μετά από φιλτράρισμα (Zero Frequency Filtered Signal) όπου εμφανίζονται συγκεκριμένα χαρακτηριστικά μόνο στην ομιλία.

Πέρα από την επιλογή των features, η μέθοδος εκπαίδευσης έχει μεγάλη επίπτωση στην τελική αποτελεσματικότητα του αλγορίθμου. Μερικές φορές χρήση σύνθετων μεθόδων εκπαίδευσης μπορούν να επιφέρουν καλύτερα αποτελέσματα σε μεγαλύτερο ποσοστό διότι επιτρέπουν την έξοδο από τοπικά ελάχιστα. Η σύνθετες μέθοδοι μπορεί να μην είναι συμβατικές ή και να δανείζονται από παρατηρήσεις της φύσης, όπως ο συνδυασμός ενός Support Vector Machine (SVM) με τον Cuckoo Algorithm [4]. Όπου, όπως το πουλί κούκος που γεννάει τα αυγά του σε ξένες φωλιές, στις επαναλήψεις εκπαίδευσης του SVM κάποιες λύσεις πετιούνται και αντικαθίστανται από νέες οι οποίες μπορεί να επιφέρουν καλύτερα αποτελέσματα.

Στο [5] οι συγγραφείς χρησιμοποιούν τα features:

1. MFCCs (Mel Frequency Cepstral Coefficients)
2. ZCR (Zero-Crossing Rate)
3. SC (Spectral Centroid)
4. SR (Spectral Rolloff)
5. SF (Specral Flux)

Τα χαρακτηριστικά MFCC, ZCR και SF ταξινομούν με accuracy 90% το καθένα. Το feature SR με 83%, ενώ το SC με 70%. Ο συνδυασμός όλων των features πετυχαίνει 93.5% σωστή ταξινόμηση, ενώ με χρήση ενός SVM μοντέλου το ποσοστό φτάνει στο 95.68%.

Παρατηρείται ότι η σωστή ταξινόμηση της μουσικής είναι αρκετά δυσκολότερη (με αυτά τα features) σε σχέση με αυτή της ομιλίας. Συγκεκριμένα στην ομιλία επιτυγχάνεται (με το SVM) accuracy 98.25% ενώ στη μουσική 93.1%.

Τέλος, σύμφωνα με το [6], σε εφαρμογές κατηγοριοποίησης όπου δεν επιβάλλεται η λειτουργία σε πραγματικό χρόνο, η χρήση energy features είναι επιθυμητή λόγο της μεγάλης ακρίβειας τους. Συγκεκριμένα η αναζήτηση της Minimum Energy Density δείχνει να υπερέχει από άλλες μεθόδους energy features και στην αποτελεσματικότητα της, και στην απλότητα του υπολογισμού της. Σε συνδυασμό με το χαρακτηριστικό της διαφοράς ενέργειάς στα διάφορα κανάλια μιας πολυκάναλης εισόδου, στο [6] πέτυχαν ακρίβεια 100% στα κομμάτια εισόδου όπου περιείχαν μόνο μουσική ή φωνή και όχι τον συνδυασμό τους (όπως στις ραδιοφωνικές διατιμήσεις).

3 Εργαλεία που χρησιμοποιήθηκαν

Η υλοποίηση αναπτύχθηκε στη γλώσσα Python 3 και χρησιμοποιήθηκε πληθώρα βιβλιοθηκών (modules) όπως η essentia για την εξαγωγή χαρακτηριστικών, η scikit-learn για την προεπεξεργασία δεδομένων και την εκπαίδευση των μοντέλων, η seaborn για την δημιουργία διαγραμμάτων και την οπτικοποίηση των χαρακτηριστικών. Παράλληλα, σε συνδυασμό με όλες αυτές χρησιμοποιήθηκαν και άλλες βιβλιοθήκες όπως η numpy, η pandas, η matplotlib, η multiprocessing, η os, η pyaudio και άλλες. Για την εκπαίδευση, δοκιμάστηκαν τα μοντέλα SVM, Decision Trees, Multilayer Perceptron, Naive Bayes και Random Forest, λεπτομέριες για τα οποία θα αναφερθούν στα επόμενα κεφάλαια.

Το dataset που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου είναι το προτεινόμενο GTZAN dataset [2], το οποίο αποτελείται από 128 αρχεία διάρκειας 30 δευτερολέπτων. Κάθε κλάση (μουσική/φωνή) αποτελείται από 64 αρχεία ενώ δεν υπάρχουν αρχεία που να περιέχουν και τις δύο κλάσεις. Όλα τα αρχεία ήχου είναι δειγματοληπτημένα στα 22050 Hz, μονοκάναλα, με βάθος ήχου 16-bit και σε μορφή WAV.

4 Χαρακτηριστικά

Για την εξαγωγή των χαρακτηριστικών από τα αρχεία ήχου του σετ δεδομένων, αρχικά τμηματήσαμε κάθε σήμα αρχείου σε frames με μέγεθος 6144 δείγματα (~278 ms), το οποίο προέκυψε μετά από επαναλαμβανόμενες δοκιμές. Έπειτα, τα frames, παραθυροποιήθηκαν με παράθυρο τύπου Hamming, ίσου μεγέθους. Στη συνέχεια, έγινε η εξαγωγή των χαρακτηριστικών στο πεδίο του χρόνου, καθώς και στο πεδίο της συχνότητας. Επίσης έγινε εξαγωγή των συντελεστών MFCC. Τα χαρακτηριστικά που εξήχθησαν, τελικά, είναι τα παρακάτω 27 που αναλύονται στη συνέχεια.

4.1 Zero Crossing Rate - ZCR

Είναι ο ρυθμός της αλλαγής πρόσημου κατά τη διάρκεια του σήματος, δηλαδή ο ρυθμός με τον οποίο το σήμα αλλάζει από θετικό σε αρνητικό και αντίστροφα. Σε κάποιο βαθμό, δείχνει την μέση συχνότητα του σήματος ως εξής:

$$\text{ZCR} = \frac{\sum_{n=1}^{N} |sgn\ x(n) - sgn\ x(n-1)|}{2N} \tag{1}$$

όπου $sgn()$ η συνάρτηση πρόσημου και $x(n)$ το διακριτό σήμα ήχου. Στη γενική περίπτωση, το ZCR για την μουσική είναι αρκετά υψηλότερο από ότι στην φωνή.

4.2 Spectral Centroid - SC

Το spectral cendroid ή αλλιώς φασματικό κέντρο, όπως αναφέρεται στο [3], είναι μία μετρική που χρησιμοποιείται ώστε να χαρακτηρίσει ένα φάσμα. Υποδεικνύει πού βρίσκεται το κέντρο του φάσματος. Έχει ισχυρή σύνδεση με την "φωτεινότητα" ενός ήχου, δηλαδή με την χροιά. Συνήθως, το κέντρο του φάσματος της φωνής συγκεντρώνεται σε χαμηλές συχνότητες και έπειτα συμπτύσσεται πολύ γρήγορα στις υψιλότερες συχνότητες ενώ δεν υπάρχει DC συνιστώσα. Αντίθετα, στην μουσική δεν έχει παρατηρηθεί κάποιο συγκεκριμένο σχήμα του φάσματος.

4.3 Roll Off

Το συγκεκριμένο χαρακτηριστικό αναπαριστά την τιμή της συχνότητας, κάτω από την οποία βρίσκεται το 95% της ενέργειας του σήματος. Όπως προαναφέρθηκε, η ενέργεια του μουσικού σήματος συγκεντρώνεται σε υψηλότερες συχνότητες σε σχέση με το σήμα της ομιλίας. Η μαθηματική του έκφραση δίνεται ως:

$$\sum_{k<v} X(k) = 0.95 \cdot \sum_{k} X(k) \tag{2}$$

[2] http://opihi.cs.uvic.ca/sound/music_speech.tar.gz

[3] Speech and Music Classification and Separation: A Review, Abdullah I. Al-Shoshan, Department of Computer Science, College of Computer, Qassim University, Saudi Arabia

όπου το $X(k)$ είναι ο διακριτός μετασχηματιμός Fourier (DFT) του $x(t)$, το αριστερό μέρος της παραπάνω εξίσωσης είναι το άθροισμα της ενέργειας κάτω από την συχνότητα ν, ενώ το δεξί είναι 95% της συνολικής ενέργειας του σήματος στο συγκεκριμένο χρονικό frame.

4.4 Spectral Flux

Το χαρακτηριστικό Spectral Flux ή αλλιώς της φασματικής ροής, όπως αναφέρεται στο [3] μετράει την φασματική διαφορά ανάμεσα στα frames. Η μουσική έχει μεγαλύτερο ρυθμό διαφοράς ενώ έχει πιο δραστικές αλλαγές ανάμεσα στα frames από ότι η ομιλία. Σημειώνεται ότι η μουσική εναλλάσσεται ανάμεσα σε περιόδους μετάβασης και στατικές περιόδους ενώ η ομιλία, γενικότερα, έχει έναν πιο σταθερό ρυθμό εναλλαγών. Ως αποτέλεσμα, η τιμή της φασματικής ροής είναι υψηλότερη για την μουσική σε σχέση με την ομιλία.

4.5 Envelope

Το envelope είναι μία ομαλή καμπύλη που καλύπτει το περίγραμμα ενός ταλαντούμενού σήματος. Εκφράζει τις χρονικές αλλαγές στο πλάτος του σήματος. Οι αλλαγές αυτές είναι υπεύθυνες για πολλές πτυχές της ακουστικής αντίληψης, συμπεριλαμβανομένου της έντασης, της χροιάς, της οξύτητας και τις χωρικής ακουστότητας.

4.6 Flatness

Το flatness ή αλλιώς επιπεδότητα του ήχου, είναι μία μετρική η οποία χρησιμοποιείται στην ανάλυση ψηφιακών σημάτων για να χαρακτηρίσει το φάσμα ενός ηχητικού σήματος. Συνήθως μετριέται σε decibels (dB), και αποτελεί έναν τρόπο ποσοτικοποίησης του πόσο κοντά είναι ένας ήχος σε θόρυβο και πόσο σε τονικότητα. [4] Η αναφορά στην τονικότητα γίνεται με την έννοια του αριθμού των κορυφών σε ένα φάσμα συχνοτήτων που θα υπήρχαν λόγω των πολλαπλών ημίτονων σε αντίθεση με το επίπεδο φάσμα του λευκού θορύβου. Τα μουσικά σήματα, τείνουν να αποτελούνται από πολλαπλούς τόνους, ο καθένας με την δική του κατανομή αρμονικών ενώ στην ομιλία δεν εμφανίζεται αυτό.

4.7 Perceptual attack time

Αυτό το χαρακτηριστικό αναφέρεται στην χρονική διάρκεια ανάμεσα στη χρονική στιγμή που το σήμα γίνεται ακουστικά αντιληπτό μέχρι τη χρονική στιγμή που φτάνει την μέγιστη έντασή του.

4.8 Sound Decay

Η προοδευτική μείωση του πλάτους ενός σήματος με την πάροδο του χρόνου. Αυτή η συμπεριφορά ξεκινάει μόλις το perceptual attack time φτάσει στο μέγιστό του. Σε αυτήν την φάση το πλάτος του σήματος μειώνεται μέχρι να φτάσει σε ένα συγκεκριμένο πλάτος στο οποίο διατηρείται μέχρι να αρχίσει να σβήνει.

4.9 Spectral Complexity

Το spectral complexity ή αλλιώς η φασματική πολυπλοκότητα, βασίζεται στον αριθμό των κορυφών του φάσματος του σήματος.

4.10 Mel Frequency Cepstral Coefficients - MFCC

Στην επεξεργασία ήχου, το cepstrum συχνοτήτων Mel (Mel frequency cepstrum - MFC) είναι μια αναπαράσταση του βραχυπρόθεσμου φάσματος έντασης ενός ήχου, βασισμένου σε έναν γραμμικό μετασχηματισμό συνημιτόνου του λογαριθμισμένου φάσματος έντασης σε μια μη γραμμική κλίμακα της συχνότητας (μη γραμμικής κλίμακας Mel). Οι συντελεστές του cepstrum συχνότητας Mel (MFCCs

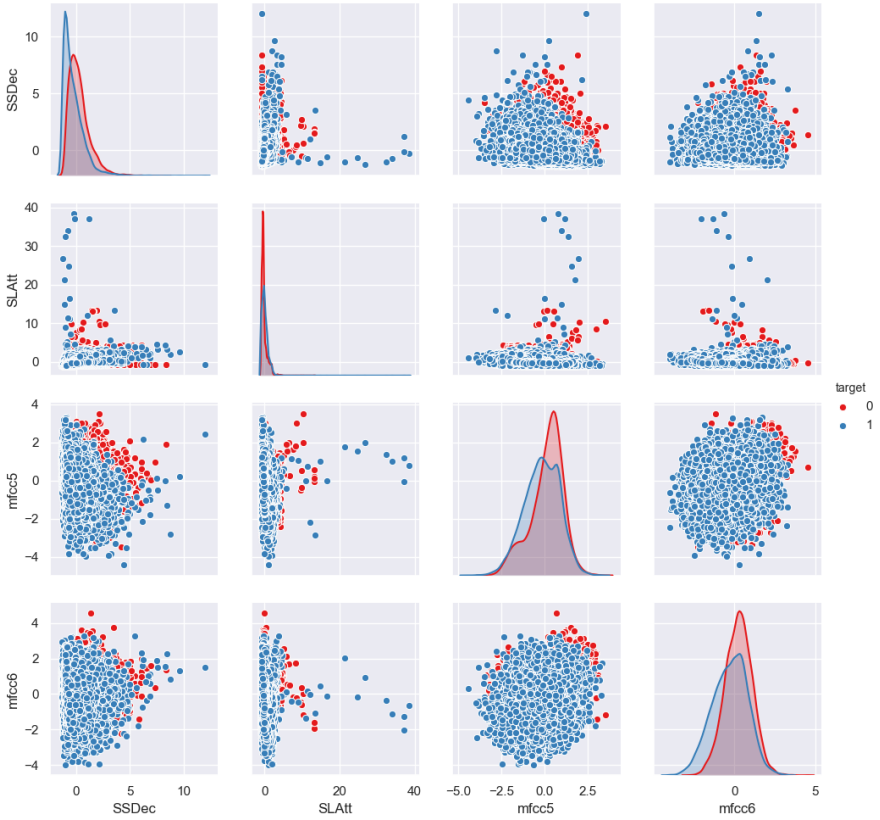
[4] https://en.wikipedia.org/wiki/Spectral_flatness

– Mel Frequency Cepstrum Coefficients) είναι οι συντελεστές εκείνοι που αποτελούν στο σύνολο τους το φάσμα MFC.

4.11 4Hz Energy Modulation

Τα σήματα ομιλίας έχουν χαρακτηριστικό μέγιστο στη διαμόρφωση ενέργειας γύρω στα 4Hz του ρυθμού συλλαβών. Για να μοντελοποιηθεί αυτή η ιδιότητα ακολουθείται η παρακάτω διαδικασία[5]: το σήμα τμηματοποιείται σε frames, εξάγονται οι συντελεστές Mel Frequency Spectrum και υπολογίζεται η ενέργεια σε 40 κανάλια αντίληψης. Αυτή η ενέργεια έπειτα φιλτράρεται με ένα ζωνοδιαβατό φίλτρο, κεντραρισμένο στα 4Hz. Η ενέργεια αθροίζεται για όλα τα κανάλια, και κανονικοποιείται με βάση τη μέση ενέργεια του κάθε frame. Η διαμόρφωση δίνεται από το κανονικοποιημένο άθροισμα της φιλτραρισμένης ενέργειας. Η φωνή περιέχει περισσότερη διαμόρφωση από την μουσική.

Στα διαγράμματα [1] και [2] φαίνονται ενδεικτικά κάποια από τα χαρακτηριστικά που υπολογίστηκαν και το πόσο αποτελεσματικά είναι στον διαχωρισμό των κλάσεων.



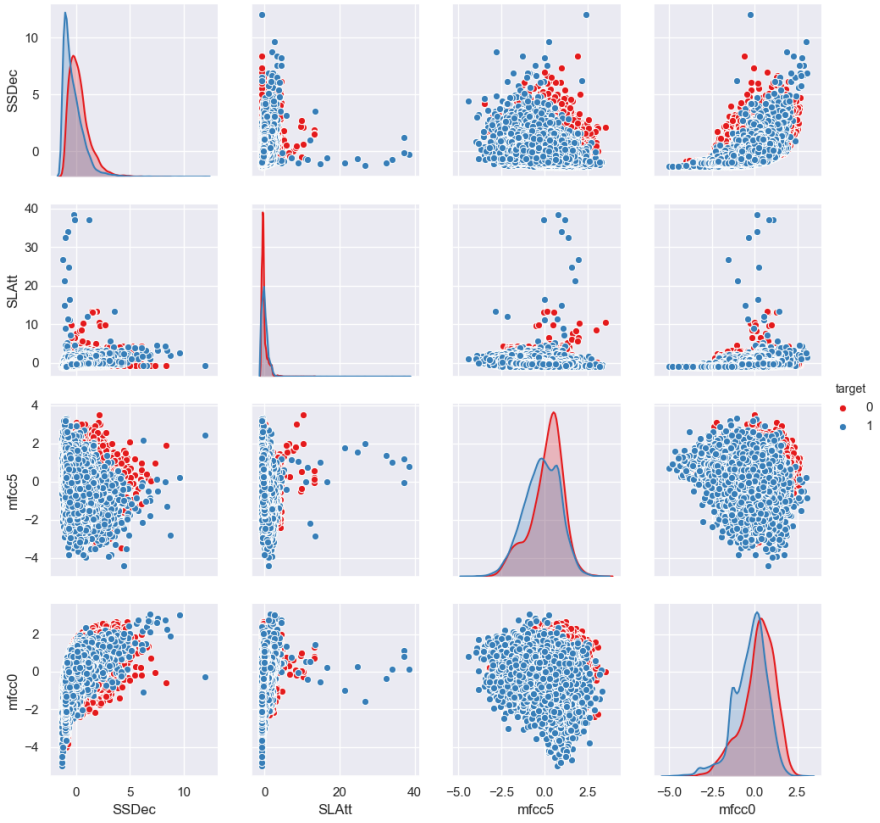
Σχήμα 1: Διαγράμματα συνδυασμών τεσσάρων χαρακτηριστικών ανά δύο

5 Προεπεξεργασία δεδομένων

Κατά την προεπεξεργασία των δεδομένων, δοκιμάστηκαν διάφορες τεχνικές έτσι ώστε να βρεθούν η βέλτιστη επιλογή χαρακτηριστικών και συνδυασμός μεθόδων. Οι μέθοδοι που δοκιμάστηκαν είναι η κλιμακοποίηση, κανονικοποίηση και ο συνδυασμός τους. Επίσης για την επιλογή των χαρακτηριστικών δοκιμάστηκαν το κατώφλι με βάση τη διακύμανση (VarianceThreshold) και η εκατοστιαία επιλογή (PercentileSelection), ενώ έγινε δοκιμή διάφορων τιμών των παραμέτρων αυτών των αλγορίθμων.

Τα αποτελέσματα παρουσιάζονται στον πίνακα 1. Στην κλιμακοποίηση όλα τα χαρακτηριστικά έχουν μηδενικό μέσο και απόκλιση ίση με τη μονάδα ενώ στην κανονικοποίηση μετατοπίζονται οι

[5] https://www.irit.fr/recherches/SAMOVA/FeaturesExtraction.html#me4hz



Σχήμα 2: Διαγράμματα συνδυασμών τεσσάρων χαρακτηριστικών ανά δύο

τιμές τους ώστε να ανήκουν στο εύρος $[0, 1]$. Τα VarianceThreshold και PercentileSelection, που είναι συναρτήσεις του sub-module feature_selection της scikit-learn, αποκλείουν χαρακτηριστικά μέσω μετρικών αξιολόγησής τους. Στο VarianceThreshold, η μείωση του αριθμού των χαρακτηριστικών γίνεται με όρους διακύμανσης, δηλαδή αφαιρούνται όλα τα χαρακτηριστικά των οποίων η διακύμανση δεν ξεπερνά κάποιο κατώφλι, ενώ η PercentileSelection κατατάσσει τα χαρακτηριστικά με βάση τη διακριτική του ικανότητα και καταργεί όλα τα χαρακτηριστικά τα οποία βρίσκονται κάτω από ένα ποσοστό των καλύτερων καθορισμένο από το χρήστη.

Μοντέλο	**Μέθοδοι προεπεξεργασίας**	**Ακρίβεια**
SVM	Χωρίς προεπεξεργασία	0.49
	Scaling	0.89
	Κανονικοποίηση	0.49
	Scaling και κανονικοποίηση	0.78
	VarThreshold και scaling	0.88
	PercenSel και scaling	0.81
	VarThreshold, scaling και gamma=scale	0.88
	VarThreshold, scaling και sigmoid kernel	0.58
	VarThreshold, scaling και polynomial kernel ($5^{ου}$ βαθμού)	0.84
Decision Tree	VarThreshold και scaling	0.75
Multi-Layer Perceptron	VarThreshold, scaling και rndState = 2	0.86
Naive Bayes	VarThreshold και scaling	0.65

Πίνακας 1: Μέθοδοι προεπεξεργασίας για διάφορα μοντέλα. Εδώ για συντομία όπου VarThreshold εννοείται VarianceThreshold, PercenSel εννοείται PercentileSelection και scaling εννοείται κλιμακοποίηση.

Εν τέλει, αποφασίστηκε να χρησιμοποιηθεί μόνο η κλιμακοποίηση, επειδή η κανονικοποίηση δεν

είχε κανένα αποτέλεσμα και το κέρδος σε ταχύτητα ταξινόμησης των παραπάνω τρόπων μείωσης του αριθμού χαρακτηριστικών δεν ήταν αρκετό συγκριτικά με την μείωση της ακρίβειας ώστε να δικαιολογήσει τη χρήση τους στην υλοποίηση. Βοήθησαν παρ' όλα αυτά στον προσδιορισμό των χαρακτηριστικών που υπερτερούν.

Στη συνέχεια, σε μία προσπάθεια περαιτέρω κατανόησης και κατάταξης των χαρακτηριστικών με βάση τη διακριτική τους ικανότητα, απομονώθηκαν όλα και ελέγχθηκε η ακρίβειά τους ένα ένα. Τα αποτελέσματα έδειξαν ότι, τελικά, κανένα χαρακτηριστικό από μόνο του δεν είναι ικανό να δώσει ικανοποιητικό ποσοστό ακρίβειας. Ακόμα και αν πάρουμε το καλύτερο σε όρους ακρίβειας και το δοκιμάσουμε σε συνδυασμό με τα επόμενα καλύτερα, φαίνεται ότι η ακρίβεια αυξάνεται λίγο αλλά όχι αρκετά. Τέλος, αν επαναληφθεί ακόμα μία φορά η διαδικασία, φαίνεται ότι έχουμε και πάλι μια μικρή αύξηση στην ακρίβεια, η οποία όμως είναι αρκετά μακρυά από την ακρίβεια που επιτυγχάνεται χρησιμοποιώντας όλα τα χαρακτηριστικά.

Accuracy	**Individually**	**With best first[1]**	**With best second[1]**
4Hz Mod	0.58	0.66	0.73
Flat	0.63	0.71	0.75
HFC	0.58	0.65	0.72
LAtt	0.62	0.71	0.75
SC	0.59	0.66	0.73
Scomp	0.57	0.66	0.73
SDec	0.63	0.65	0.72
SEFlat	0.51	0.65	0.72
SF	0.55	0.69	0.75
SFlat	0.57	0.66	0.72
SLAtt	0.63	0.71	0.74
SR	0.60	0.66	0.72
SSDec	**0.65**	-	-
ZCR	0.58	0.65	0.72
mfcc0	0.61	0.66	0.73
mfcc1	0.58	0.67	0.73
mfcc2	0.52	0.66	0.73
mfcc3	0.56	0.69	0.76
mfcc4	0.54	0.67	0.74
mfcc5	0.57	0.70	0.75
mfcc6	0.61	**0.72**	-
mfcc7	0.57	0.68	0.75
mfcc8	0.55	0.67	0.74
mfcc9	0.54	0.67	0.73
mfcc10	0.54	0.65	0.73
mfcc11	0.51	0.66	0.73
mfcc12	0.54	0.67	0.73

Πίνακας 2: Ακρίβεια μεμονωμένων χαρακτηριστικών και συνδυασμών τους.

1: όπου "With best first" εννοείται ο συνδυασμός του χαρακτηριστικού με το καλύτερο, ενώ "With best second" εννοείται ο συνδυασμός του χαρακτηριστικού με τον καλύτερο συνδυασμό που προέκυψε στο προηγούμενο βήμα

Άρα, γίνεται προφανές ότι δεν υπάρχει κάποιο συγκεκριμένο, μοναδικό χαρακτηριστικό το οποίο ευθύνεται για το μεγαλύτερο ποσοστό της ακρίβειας του μοντέλου αλλά είναι ο συνδυασμός τους.

6 Machine Learning Model

Στη συνέχεια αναφέρεται συνοπτικά η λειτουργία των μοντέλων που χρησιμοποιήθηκαν για την εκπαίδευση των δεδομένων (Οι ορισμοί είναι σύμφωνα με την ιστοσελίδα της analytics vidhya [6]) ενώ στο τέλος παρατίθεται ένας πίνακας στο οποίο φαίνονται οι διάφορες μέθοδοι και οι ακρίβειες που επιτεύχθησαν.

6.1 Support Vector Machine - SVM

Τα SVMs ανήκουν στα μοντέλα επιβλεπόμενης μάθησης, και ο σκοπός τους είναι η εύρεση ενός γραμμικού υπερεπιπέδου (σύνορο απόφασης) το οποίο θα διαχωρίσει τα δεδομένα. Σε αυτόν τον αλγόριθμο, σχεδιάζουμε κάθε δεδομένο ως ένα σημείο σε έναν ν-διάστατο χώρο (όπου ν είναι ο αριθμός των features) με την τιμή κάθε feature να είναι η τιμή της εκάστοτε συντεταγμένης. Έπειτα, κατηγοριοποιούμε βρίσκοντας ένα υπερεπίπεδο το οποίο διαχωρίζει τις 2 κλάσεις καλύτερα.

6.2 Decision Trees

Τα δένδρα απόφασης ή decision trees ανήκουν στα μοντέλα επιβλεπόμενης μάθησης και εφαρμόζονται τόσο σε κατηγορικά όσο και συνεχή δεδομένα. Σε αυτόν τον αλγόριθμο, χωρίζουμε τα δείγματα σε πιο ομοιογενείς υποομάδες βασιζόμενοι στο χαρακτηριστικό που τα διαχωρίζει καλύτερα κάθε φορά.

6.3 Multilayer Perceptron

Ένα perceptron, μπορεί να κατανοηθεί ως οτιδήποτε δέχεται πολλαπλές εισόδους και παράγει μία έξοδο. Ο τρόπος όμως με τον οποίο συσχετίζεται η είσοδος την έξοδο εμφανίζει ενδιαφέρον. Αρχικά σε κάθε είσοδο προστίθεται ένα βάρος, το οποίο σημαίνει ουσιαστικά το πόσο σημασία να δοθεί σε κάθε μία ενώ στην έξοδο ένα κατώφλι.Τέλος, προστίθεται και μία πόλωση η οποία μπορεί να θεωηθεί ως το ποσό ευελιξίας του perceptron.Για λόγους απόδοσης, χρησιμοποιούνται πολλά perceptrons σε layers, τα οποία είναι πλήρως συνδεδεμένα μεταξύ τους.

6.4 Naive Bayes

Είναι μία τεχνική ταξινόμησης η οποία βασίζεται στο θεώρημα του Bayes[7]με την υπόθεση ανεξαρτησίας ανάμεσα στους προβλέπτες. Με απλά λόγια, ο ταξινομητής Naive Bayes, υποθέτει ότι η ύπαρξη ενός συγκεκριμένου feature σε μια κλάση είναι ασυσχέτιστη με την υπάρξη οποιουδήποτε άλλου.

6.5 Random Forest

Ο Random Forest είναι ένας αλγόριθμος τύπου Bootstrap, με δένδρα απόφασης. Αυτό που προσπαθεί να κάνει, είναι να φτιάξει δίαφορα δένδρα με διαφορετικά δείγματα και διαφορετικές αρχικές τιμές. Επαναλαμβάνει την διαδικασία και κάνει μια τελική πρόβλεψη για κάθε παρατήρηση, η οποία είναι συνάρτηση όλων των προβλέψεων.

7 Αξιολόγηση μοντέλων

Για την αξιολόγηση των μοντέλων, δοκιμάστηκαν πολλοί τρόποι για τις διάφορες μεθόδους ταξινόμησης μέχρι να βρεθεί ο βέλτιστος. Στον πίνακα 3 φαίνονται ενδεικτικά για καποιους ταξινομητές οι τρόποι αξιολόγησης που δοκιμάστηκαν και οι αντίστοιχες ακρίβειες τους.

[6] https://www.analyticsvidhya.com/

[7] https://en.wikipedia.org/wiki/Bayes%27_theorem

Μέθοδος	**Τρόπος αξιολόγησης**	**Ακρίβεια**
SVM	Κλιμακοποίηση + τυχαίο split	0.94
	Κλιμακοποίηση + k-fold cross validation	**0.96** (best fold)
	Κλιμακοποίηση + PCA + τυχαίο split	0.92
	Κλιμακοποίηση + PCA + k-fold cross validation	0.93 (best fold)
Random Forest	Κλιμακοποίηση + k-fold cross validation	0.95 (best fold)
	Κλιμακοποίηση + PCA + k-fold cross validation	0.94 (best fold)

Πίνακας 3: Τρόποι αξιολόγησης για διάφορα μοντέλα

Είναι φανερό ότι η καλύτερη ακρίβεια που επιτυγχάνεται είναι 0.96 με κλιμακοποίηση και k-fold cross validation. Συνεπώς, επιλέχθηκαν για την τελική υλοποίηση ενώ χρησιμοποιήθηκε για το k-fold cross validation, k=4.

8 Συμπεράσματα

Παρατίθεται στη συνέχεια ο πίνακας στον οποίο φαίνονται οι ακρίβειες των μοντέλων για την ταξινόμηση.

Method	**Accuracy**
SVM	***96.06***
Decision Tree	86.51
MultiLayer Perceptron	90.34
Naive Bayes	70.25
Random Forest	95.49
SVM (PCA(10))	90.02

Πίνακας 4: Ακρίβεια ταξινομητών

Όπως φαίνεται, η καλύτερες μέθοδοι είναι τα Support Vector Machines και ο αλγόριθμος Random Forest με 96% και 95% ακρίβεια αντίστοιχα, ενώ κοντά βρίσκεται και ο αλγόριθμος του Multilayer perceptron. Παράλληλα, βλέπουμε ότι ο χειρότερος είναι ο Naive Bayes με περίπου 70% ακρίβεια.Τέλος, η εφαρμογή του PCA είναι φανερό ότι μείωσε αρκετά την ακρίβεια του μοντέλου και για αυτόν τον λόγο συνίσταται μόνο στην περίπτωση που υπάρχει κάποιος χρονικός περιορισμός καθώς, λόγω της μείωσης των χαρακτηριστικών από 27 σε 10, η υλοποίηση θα εκτελείται πιο γρήγορα.

Αναφορές

[1] J. L. R. Julien PINQUIER and R. egine ANDRE-OBRECHT, "Robust speech / music classification in audio documents," 7^{th} *International Conference on Spoken Language Processing*, 2002.

[2] C. D. Nikolaos Tsipas, Lazaros Vrysis and G. Papanikolaou, "Mirex 2015: Methods for speech/music detection and classification," *MIREX 2015 Conference*, 2015.

[3] B. K. Khonglah and S. M. Prasanna, "Speech / music classification using speech-specific features," *Digital Signal Processing 48*, 2016.

[4] W. Shi and X. Fan, "Speech classification based on cuckoo algorithm and support vector machines," *2nd IEEE International Conference on Computational Intelligence and Applications*, 2017.

[5] K. Wang, Y. Yang, and Y. Yang, "Speech/music discrimination using hybrid-based feature extraction for audio data indexing," in *2017 International Conference on System Science and Engineering (ICSSE)*, pp. 515–519, July 2017.

[6] B. C. Stanisław Kacprzak and B. Ziółko, “Speech/music discrimination for analysis of radio stations,” *2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2017.