Τεχνολογία Ήχου και Εικόνας 2018
Παραδοτεό 1

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

4 Νοεμβρίου 2018

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).

Πρόκειται για ένα 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 διάρκειας 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 προεκπαιδευμένο 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)

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 Σχεδιασμός υλοποίησης

Μετά από μελέτη των προηγούμενων υλοποιήσεων και πειραματισμό με την εξαγωγή διάφορων χαρακτηριστικών (features) στη Matlab αποφασίσαμε να ακολουθήσουμε την παρακάτω πορεία αντιμετώπισης του προβλήματος.

3.1 Παραθυροποίηση

Για την παραθυροποίηση του σήματος θα γίνει χρήση Hamming παραθύρων με επικάλυψη 50%. Η τελική χρονική διάρκεια των παραθύρων αναμένεται να είναι στο πεδίο του μισού με ενός δευτερολέπτου (0.5-1 sec) και θα καθοριστεί στη πορεία μέσω trial and error τεχνικών.

3.2 Χαρακτηριστικά

Τα χαρακτηριστικά που έχουν επιλεγεί είναι τα εξής:

1. MFCCs (Mel Frequency Cepstral Coefficients)
2. Silence ratio
3. ZCR (Zero-Crossing Rate)
4. SC (Spectral Centroid)
5. SR (Spectral Rolloff)
6. SF (Specral Flux)
7. 4Hz modulation
8. Minimum Energy Density (MED)

Τα συγκεκριμένα χαρακτηριστικά εμφανίζουν τις μεγαλύτερες ακρίβειες στη ταξινόμηση ενώ ταυτόχρονα έχουν μικρή ετεροσυσχέτιση. Άλλα χαρακτηριστικά μπορεί να προστεθούν στη πορεία μετά από αναλυτικότερη έρευνα της βιβλιογραφίας.

3.3 Μοντέλο ταξινόμησης

Από την βιβλιογραφική έρευνα διαπιστώθηκε ότι οι διαφορετικές επιλογές χαρακτηριστικών επηρεάζουν την ακρίβεια των μοντέλων. Έτσι με μία συγκεκριμένη επιλογή χαρακτηριστικών μπορεί τα πιθανοτικά μοντέλα (Naive Bayes, GMM, κ.α.) να είναι αποτελεσματικότερα των SVM ή των νευρωνικών. Αλλά με επιλογή διαφορετικών features να ισχύει το αντίθετο. Για τον λόγο αυτό είναι απαραίτητο, αφού αποφασιστεί το σετ των χαρακτηριστικών να γίνει εκπαίδευση και testing πολλών μοντέλων πριν την τελική επιλογή. Έτσι η πρόταση μας είναι η δοκιμή των περισσότερων ευραίως διαδεδομένων μοντέλων, όπως: Decision trees, Bayesian networks, Gaussian mixture model, Hidden Markov Model, SVMs, Artificial Neural networks, Genetic Algorithms.

3.4 Preprocessing, άλλες τεχνικές

Περισσότερες και πιο εξεζητημένες τεχνικές θα χρησιμοποιηθούν στο πρακτικό κομμάτι που θα υλοποιηθεί αργότερα. Κατά το preprocessing των δεδομένων μέθοδοι όπως data rescaling, data standardization, data binarization, data cleaning, data integration, data transformation ενδέχεται να φανούν χρήσιμες. Ακόμα, κατά την εκπαίδευση διάφορες γνωστοί μέθοδοι validation όπως το k-fold cross-validation, leave one out, bootstrap, hold out θα δοκιμαστούν.

3.5 Stack

Τόσο για την εξερεύνηση του χώρου των χαρακτηριστικών όσο και για την εκπαίδευση και τον έλεγχο του μοντέλου θα χρησιμοποιηθεί το προγραμματιστικό περιβάλλον της R. Το περιβάλλον αυτό είναι ειδικά σχεδιασμένο για στατιστικούς υπολογισμούς (statistical computing) και αποτελεί (μαζί με την python) το στάνταρ της βιομηχανίας μηχανικής μάθησης. Επίσης παρέχεται αφθονία βιβλιοθηκών έτοιμων machine learning αλγορίθμων από τις οποίες θα χρησιμοποιηθούν μεταξύ άλλων οι: 'e1071', 'rpart', 'nnet', 'random forest'.

Σε διάφορα στάδια της εργασίας ενδέχεται να χρησιμοποιηθεί και η γλώσσα Matlab λόγω της ευκολίας που προσφέρει στους μαθηματικούς υπολογισμούς.

Αναφορές

[1] J.-L. R. Julien Pinquier and R. André-Obrecht, "Robust speech/music classification in audio documents," *7th International Conference on Spoken Language Processing [ICSLP2002]*, 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] Y.-M. Y. Kun-Ching Wang and Y.-R. Yang, "Speech/music discrimination using hybrid-based feature extraction for audio data indexing," *2017 International Conference on System Science and Engineering (ICSSE)*, 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.