ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Γιαννικάκη Σοφία- Ελπινίκη

Μεταπτυχιακή Φοιτήτρια

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: Καθηγητής, Ιωάννης Στυλιανού

Πέμπτη, 2 Απριλίου 2015, 14:00

Αίθουσα Κ206, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Ανίχνευση φωνής σε ερασιτεχνικές καταγραφές μουσικών σεμιναρίων υπό πραγματικές συνθήκες"

ΠΕΡΙΛΗΨΗ

Μία από τις σημαντικότερες ικανότητες που έχει ο άνθρωπος είναι η ομιλία, η οποία αποτελεί και το βασικό τρόπος επικοινωνίας με τον υπόλοιπο κόσμο. Τα τελευταία χρόνια το ενδιαφέρον πολλών έχει επικεντρωθεί στην ανάπτυξη εφαρμογών, οι οποίες βασίζονται στη φωνή. Σε τέτοιου είδους εφαρμογές, μας δίδεται ένα σήμα εισόδου από το οποίο χρησιμοποιούμε μόνο τα κομμάτια που περιέχουν φωνή. Με άλλα λόγια, αναλύοντας το σήμα εντοπίζουμε τα κομμάτια φωνής, τα οποία και κρατάμε, ενώ τα υπόλοιπα (θόρυβος, ησυχία κλπ) τα αγνοούμε. Η διαδικασία αυτή ονομάζεται ανίχνευση φωνής (Voice Detection). Με τη διαδικασία αυτή μειώνεται δραματικά ο όγκος της πληροφορίας που πρόκειται να επεξεργαστούμε, κάτι το οποίο είναι πολύ χρήσιμο.

Η διαδικασία της ανίχνευσης της φωνής σχετίζεται στενά με την ταξινόμηση σε ομιλία και μη ομιλία. Επίσης, τόσο η ανίχνευση τραγουδιού όσο και η διάκριση ομιλίας/μουσικής μπορούν να θεωρηθούν υποκατηγορίες της ανίχνευσης φωνής. Σε όλες αυτές τις περιπτώσεις μας δίδεται ένας σήμα εισόδου το οποίο και επεξεργαζόμαστε. Συνήθως η ανάλυση του σήματος γίνεται σε μικρότερα κομμάτια, από τα οποία εξάγουμε χαρακτηριστικά. Η διάρκεια των κομματιών κυμαίνεται περίπου μεταξύ 0.02 και 3 δευτερολέπτων και ορίζεται ανάλογα με το πρόβλημα που έχουμε κληθεί να λύσουμε. Μπορεί επίσης να εξαρτάται από το είδος των χαρακτηριστικών που θέλουμε να εξάγουμε. Μέχρι τώρα έχουν προταθεί πλήθος χαρακτηριστικών, κάποια από τα οποία είναι εφικτό να παράγουν αποτελέσματα χρησιμοποιώντας μικρά κομμάτια του σήματος. Αντίθετα, υπάρχουν χαρακτηριστικά τα οποία απαιτούν περισσότερη πληροφορία με αποτέλεσμα η διάρκεια των κομματιών να πρέπει να είναι μεγάλη. Τα χαρακτηριστικά μπορούν να χωριστούν σε δύο κατηγορίες, σε αυτά του πεδίου του χρόνου και σε εκείνα του πεδίου των συχνοτήτων. Στο πεδίο του χρόνου ευρέως διαδεδομένα είναι η ενέργεια, ο ρυθμός διέλευσης από το μηδενικό άξονα και χαρακτηριστικά που βασίζονται στην αυτοσυσχέτιση. Από την άλλη, στο πεδίο των συχνοτήτων ένα μεγάλο ποσοστό των χαρακτηριστικών εξάγεται από το Cepstrum (επέκταση του φάσματος). Αυτό συμβαίνει διότι εκεί υπάρχει χρήσιμη πληροφορία για τη φωνή. Συγκεκριμένα, το πιο διαδεδομένο χαρακτηριστικό στην ανίχνευση τραγουδιού και στη διάκριση ομιλίας/μουσικής είναι οι Mel-frequency Cepstral συντελεστές. Υποστηρίζεται ότι το χαρακτηριστικό αυτό δίνει τα καλύτερα αποτελέσματα στην πλειοψηφία των περιπτώσεων.

Στην εργασία αυτή παρουσιάζεται ένας αλγόριθμος ανίχνευσης φωνής πάνω σε πραγματικές καταγραφές από μαθήματα μουσικής. Καθώς η φύση των ηχογραφήσεων είναι τέτοια, στόχος είναι να εντοπίζεται τόσο η ομιλία όσο και το τραγούδι. Ένα κλασικό σύστημα χρησιμοποιεί τους MFC συντελεστές ως χαρακτηριστικό διαχωρισμού "φωνής"/"μη φωνής" και μία μηχανή διανυσματικής υποστήριξης (Support Vector Machine) για την ταξινόμηση. Βάση ενός τέτοιου συστήματος λοιπόν, ορίζουμε τους MFC συντελεστές ως το κύριο χαρακτηριστικό και προσθέτουμε άλλα τρία, τη ροή του Cepstrum, τη Σαφήνεια και την Αρμονικότητα. Τα δύο τελευταία βασίζονται στην αυτοσυσχέτιση του σήματος στο πεδίο του χρόνου. Ο σκοπός είναι να βελτιωθεί η απόδοση ενός συστήματος, που χρησιμοποιεί μόνο τους MFC συντελεστές. Εξετάζουμε 5 διαφορετικούς συνδυασμούς των χαρακτηριστικών που προαναφέρθηκαν με τους MFC συντελεστές. Σήματα διάρκειας 3 δευτερολέπτων αναλύονται σε κομμάτια των 0.03 δευτερολέπτων, έχοντας 0.02 δευτερόλεπτα επικάλυψη μεταξύ τους. Σε κάθε τέτοιο κομμάτι εξάγονται τα χαρακτηριστικά και αποθηκεύονται σε ένα διάνυσμα. Έπειτα, εφαρμόζεται ένα 10-fold cross-validation πάνω στα σήματα διάρκειας 3 δευτερολέπτων, για να ταξινομηθούν σε "φωνή" και "μη φωνή". Η βάση που χρησιμοποιήθηκε για την εκπαίδευση και τον έλεγχο του συστήματος αποτελείται από 3 σεμινάρια. Δύο από αυτά σχετίζονται με τη λύρα στην παραδοσιακή κρητική μουσική, ενώ το τρίτο αφορά το λαούτο. Σημειώνεται ότι η κάθε ηχογράφηση έχει πραγματοποιηθεί κάτω από διαφορετικές συνθήκες.

Η απόδοση του αλγορίθμου αξιολογήθηκε βάσει των Detection Error Tradeoff (DET) και Receiver Operating Characteristic (ROC) καμπυλών. Παράλληλα, υπολογίστηκε και το ποσοστό ίσου σφάλματος (Equal Error Rate), το μέτρο Αποδοτικότητας και το εμβαδό της καμπύλης ROC. Πραγματοποιήθηκε αξιολόγηση του κάθε σεμιναρίου χωριστά και όλων μαζί. Επίσης, έγινε συνδυασμός δεδομένων εκπαίδευσης και ελέγχου του συστήματος από δύο διαφορετικά σεμινάρια. Με τον τρόπο αυτό παρέχουμε πιο αξιόπιστα αποτελέσματα. Καταλήγουμε ότι η χρήση των επιπλέον χαρακτηριστικών βελτιώνει αισθητά την απόδοση του κλασικού αλγορίθμου που χρησιμοποιεί μόνο τους MFC συντελεστές. Τέλος, παρατηρήθηκε ότι τρεις από τους πέντε συνδυασμούς ξεχωρίζουν, μειώνοντας κατά 20% την πιθανότητα του να χάσουμε ένα κομμάτι "φωνής", δεδομένης μιας πιθανότητας ίση με 5%, να χαρακτηρίσουμε ως "φωνή" κάποιο κομμάτι που στην πραγματικότητα δεν είναι.

“Voicing detection in spontaneous and real-life recordings from music lessons”

ABSTRACT

Speech is one of the most important abilities that we have, since it is one of the principal ways of communication with the world. In the past few years a lot of interest has been shown in developing voice-based applications. Such applications involve the isolation of speech from an audio file. The algorithms that achieve this are called Voice Detection algorithms. From the analysis of a given input audio signal, the parts containing voice are kept while the other parts (noise, silence, etc) are discarded. In this way a great reduction of the information to be further processed is achieved.

The task of Voice Detection is closely related with Speech/Nonspeech Classification. In addition, Singing Voice Detection and Speech/Music Discrimination can be seen as subclasses of what we generally call Voice Detection. When dealing with such tasks, an audio signal is given as an input to a system and is then processed. The signal is usually analysed in small parts called frames, from which features are extracted. The frame length can vary from about 0.02 up to 3 seconds. The duration depends mostly on the application and sometimes on the features being used. Many features have been proposed until now. Some of them work well when the frame size is small (local information). However, other features require more information, therefore the frame length is chosen to be bigger. There are two categories in which the features could be divided, time domain and frequency domain features. In time domain the short time energy, the zero-crossing rate and autocorrelation based features are most often used. In frequency domain cepstral features are most frequently used, due to the useful information about speech presence.

To be more specific, in Singing Voice Detection and in Speech/Music Discrimination the state-of-the-art feature are the Mel-Frequency Cepstral Coefficients. It has been reported, that this particular feature provides the best performance in the majority of the cases.

In this thesis an algorithm is developed that performs voice detection in spontaneous and real-life recordings from music lessons. The content of the recordings was such that the proposed algorithm was challenged to discriminate both speech and singing voice from music and other noises. A classic approach for this problem would use MFCCs as the discrimination feature and an SVM classifier for the classification into "speech" or "nonspeech". In our work the methodology of this approach is expanded by preserving the MFCCs as the main feature and incorporating three other features namely, the Cepstral Flux, the Clarity and the Harmonicity. Cepstral Flux is extracted from the Cepstrum, while Clarity and Harmonicity are time-domain autocorrelation-based features. The goal is to improve with these additional features the performance of the system that uses only the MFCCs. So, different combination of the three additional features with the MFCCs were examined and evaluated. A 3 second lasting input signal, is processed in 30ms long frames, with an 20ms overlap. In each frame the features are extracted and put into a vector. A 10-fold cross-validation then is applied, on the 3 second lasting segments, which are labelled as "speech" or "nonspeech". The database used for the training and the testing purposes of our algorithm consists of three seminars. Two of them concern traditional Cretan music classes with lira and the third one traditional Cretan music classes with lute. Each recording has been carried out under different environmental conditions.
Performance evaluation was conducted using the Detection Error Tradeoff (DET) and Receiver Operating Characteristic (ROC) curves as a visual evaluation tool. Also, the Equal Error Rate (EER), the Efficiency and the Area Under the Curve (AUC) were computed in each case. Each seminar was evaluated separately, as well as all together. A combination of training and testing sets from different seminars was also done, to be able to provide reliable results. It is shown that the use of the additional features significantly enhances the performance of the classic algorithm that uses only the MFCCs. Finally, it is observed that three out of the five combinations stand out, by reducing about 20% the miss probability given a false alarm probability equal to 5%.