ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Γιακουμάκη Θεοδώρα

Μεταπτυχιακή Φοιτήτρια

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: Καθηγητής, Ιωάννης Στυλιανού

Πέμπτη, 2 Απριλίου 2015, 12:00

Αίθουσα Κ206, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Ανάλυση και Ταξινόμηση Εκφραστική Ομιλίας βασισμένη σε προσαρμόσιμα Ημιτονοειδή μοντέλα"

ΠΕΡΙΛΗΨΗ

Η εκφραστική (ή αγχωμένη/ συναισθηματική) ομιλία μπορεί να ορισθεί ως το είδος ομιλίας το οποίο παράγεται από έναν ομιλητή ο οποίος είναι συναισθηματικά φορτισμένος. Ομιλητές οι οποίοι αισθάνονται λυπημένοι, θυμωμένοι, χαρούμενοι ή ουδέτεροι προσθέτουν ένα συγκεκριμένο βάρος στην ομιλία τους, το οποίο συνήθως χαρακτηρίζεται ως συναίσθημα. Η επεξεργασία της εκφραστικής ομιλίας θεωρείται μια από τις πιο απαιτητικές διεργασίες για μοντελοποίηση, αναγνώριση και ταξινόμηση συναισθήματος. Η συναισθηματική κατάσταση ενός ομιλιτή μπορεί να αποκαλυφθεί από την ανάλυση της ομιλίας του, και μια τέτοιου είδους γνώση θα ήταν χρήσιμη σε καταστάσεις εκτάκτου ανάγκης, σε εφαργμογές υγείας, καθώς και μεταξύ άλλων ως ένα στάδιο επεξεργασίας σε συστήματα αναγνώρισης και ταξινόμησης του συναισθήματος.

Η ακουστική ανάλυση της ομιλίας η οποία παράγεται κάτω από διάφορες συναισθηματικές καταστάσεις αποκαλύπτει έναν εξαιρετικά μεγάλο αριθμό χαρακτηριστικών τα οποία ποικίλουν ανάλογα με τον είδος της συναισθηματικής

κατάστασης του ομιλητή. Ως εκ τούτου αυτά τα χαρακτηριστικά θα μπορούσαν να χρησιμοποιηθούν για αναγνώριση και/ή ταξινόμηση διαφόρων συναισθηματικών καταστάσεων. Υπάρχει πολύ μικρή έρευνα πάνω στις παραμέτρους του Ημιτονοειδούς Μοντέλου (SM), (οι οποίες είναι το πλάτος, η συχνότητα και η φάση) ως γνωρίσματα για τον διαχωρισμό των ειδών ομιλίας. Ωστόσο, η εκτίμηση αυτών των παραμέτρων υπόκειται σε έναν πολύ σημαντικό περιορισμό: εξάγονται με την παραδοχή της ``τοπικής στασιμότητας '', ότι δηλαδή το σήμα φωνής θεωρείται στάσιμο μέσα σε ένα παράθυρο ανάλυσης. Όμως, είδη ομιλίας τα οποία χαρακτηρίζονται γρήγορα ή θυμωμένα ίσως να μην συμφωνούν με αυτή την παραδοχή. Προσφάτως, αυτό το πρόβλημα το χειρίζονται με επιτυχία τα προσαρμόσιμα Ημιτονοειδή Μοντέλα (aSMs), προβάλλοντας το σήμα επάνω σε ένα σύνολο συναρτήσεων βάσης μεταβλητής συχνότητας και πλάτους μέσα σε ένα παράθυρο ανάλυσης. Ως εκ τούτου, οι ημιτονοειδείς παράμετροι εκτιμούνται με περισσότερη ακρίβεια σε σχέση με τα συνήθη ημιτονοειδή μοντέλα.

Σε αυτή την εργασία, προτείνουμε την χρήση ενός προσαρμόσιμου Ημιτονοειδούς Μοντέλου (aSM), το εκτεταμένο προσαρμόσιμο Σχεδόν - Αρμονικό Μοντέλο (eaQHM), για ανάλυση και ταξινόμηση συναισθηματικής ομιλίας. Το (eaQHM) προσαρμόζει το πλάτος και την φάση των συναρτήσεων βάσης στα τοπικά χαρακτηριστικά του σήματος. Αρχικά, το (eaQHM) καλείται να αναλύσει την εκφραστική ομιλία με πιο ακριβείς, αξιόπιστες, συνεχόμενες, χρονικά - μεταβαλλόμενες παραμέτρους (πλάτη και συχνότητες). Αποδεικνύεται ότι οι παράμετροι αυτοί μπορούν να αναπαραστήσουν το εκφραστικό περιεχόμενο της ομιλίας με επάρκεια και ακρίβεια σε σχέση με τα συνήθη ημιτονοειδή μοντέλα. Χρησιμοποιώντας μια πολύ διαδεδομένη βάση δεδομένων προ-επισημασμένης στενής ζώνης εκφραστικής ομιλίας (SUSAS) και την εκφραστική βάση δεδομένων του Βερολίνου (EmoDB), δείχνουμε ότι μπορούμε να επιτύχουμε πολύ υψηλή αναλογία σφάλματος σήματος ως προς το σφάλμα ανακατασκευής (SRER), σε σύγκριση με το κλασσικό Ημιτονοειδές Μοντέλο (SM). Συγκεκριμένα, το (eaQHM) ξεπερνά το (SM) κατά 93\% μέσο όρο (SRER). Επιπλέον, έγιναν επίσημα ακουστικά τέστ, σε μια δεύτερη ευρείας ζώνης βάση δεδομένων με ομιλία, τα οποία δείχνουν ότι το (eaQHM) ξεπερνά το (SM) σε ότι αφορά την ποιότητα ανακατασκευής. Οι βάσεις δεδομένων της (SUSAS) και (EmoDB) χρησιμοποιήθηκαν επίσης για την κατασκευή δύο χωριστών Διανυσματικών Κβαντιστών (VQ) για ταξινόμηση, ένα για τα πλάτη και ένα για τις συχνότητες ως γνωρίσματα. Τέλος, προτείνουμε ένα συνδυαστικό σχήμα ταξινόμησης με πλάτη και συχνότητες. Τα αποτελέσματα δείχνουν ότι τόσο για τα απλά γνωρίσματα όσο και για τα συνδυαστικά επιτυγχάνεται καλύτερη απόδοση χρησιμοποιώντας το (eaQHM) αντί του (SM).

Giakoumaki Theodora

M.Sc. Thesis

Computer Science Department

University of Crete

Master's Thesis Supervisor: Professor Yannis Stylianou

Thursday, 2/4/2015, 12:00

Room K206, Computer Science dept., University of Crete

"Expressive Speech Analysis and Classification using adaptive Sinusoidal Modeling"

ABSTRACT

Emotional (or stressed/expressive) speech can be defined as the speech style produced by an emotionally charged speaker. Speakers that feel sad, angry, happy and neutral put a certain stress in their speech that is typically characterized as emotional. Processing of emotional speech is assumed among the most challenging speech styles for modelling, recognition, and classifications. The emotional condition of speakers may be revealed by the analysis of their speech, and such knowledge could be effective in emergency conditions, health care applications, and as pre-processing step in recognition and classification systems, among others.

Acoustic analysis of speech produced under different emotional conditions reveals a great number of speech characteristics that vary according to the emotional state of the speaker. Therefore these characteristics could be used to identify and/or classify different emotional speech styles. There is little research on the parameters of the Sinusoidal Model (SM), namely amplitude, frequency, and phase as features to separate different speaking styles. However, the estimation of these parameters is subjected to an important constraint; they are derived under the assumption of local stationarity, that is, the speech signal is assumed to be stationary inside the analysis window. Nonetheless, speaking styles described as fast or angry may not hold this assumption. Recently, this problem has been handled by the adaptive Sinusoidal Models (aSMs), by projecting the signal onto a set of amplitude and frequency varying basis functions inside the analysis window. Hence, sinusoidal parameters are more accurately estimated.

In this thesis, we propose the use of an adaptive Sinusoidal Model (aSM), the extended adaptive Quasi-Harmonic Model (eaQHM), for emotional speech analysis and classification. The eaQHM adapts the amplitude and the phase of the basis functions to the local characteristics of the signal. Firstly, the eaQHM is employed to analyze emotional speech in accurate, robust, continuous, time-varying parameters (amplitude and frequency). It is shown that these parameters can adequately and accurately represent emotional speech content. Using a well known database of pre-labeled narrowband expressive speech (SUSAS) and the emotional database of Berlin, we show that very high Signal to Reconstruction Error Ratio (SRER) values can be obtained, compared to the standard Sinusoidal Model (SM). Specifically, eaQHM outperforms SM in average by 90\% in SRER. Additionally, formal listening tests,on a wideband custom emotional speech database of running speech, show that eaQHM outperforms SM from a perceptual resynthesis quality point of view. The SUSAS and Berlin databases were also used in order to develop two separate Vector Quantizers (VQs) for the classification, one for amplitude and one for frequency features. Finally, we suggest a combined amplitude-frequency classification scheme. Experiments show that both single and combined classification schemes achieve higher performance when the features are obtained from eaQHM.