Αναγνώριση Συναισθημάτων σε Περιβάλλοντα Θορύβου

Θεόδωρος Κωστούλας
Υπ. Διδάκτωρ
tkost@wcl.ee.upatras.gr

Ιωσήφ Μπόρας
Υπ. Διδάκτωρ
imporas@wcl.ee.upatras.gr

Νίκος Φακωτάκης
Καθηγητής
fakotaki@wcl.ee.upatras.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία παρουσιάζεται η υλοποίηση ενός συστήματος αναγνώρισης συναισθημάτων. Συγκεκριμένα εξετάζεται η συμπεριφορά του συστήματος σε διαφορετικές συνθήκες θορύβου.

Για το σκοπό αυτό, τρία διαφορετικά σενάρια υλοποιούνται. Αρχικά εξετάζουμε τη συμπεριφορά του συστήματος χωρίς την παρουσία θορύβου. Στη συνέχεια, το σύστημα εκπαιδεύεται και δοκιμάζεται με δεδομένα που εμπεριέχουν θόρυβο σε αναλογία 10dB. Τέλος, το σύστημα δοκιμάζεται με δεδομένα που εμπεριέχουν θόρυβο σε αναλογία 10dB ενώ η εκπαίδευσή του έχει γίνει με δεδομένα που δεν περιέχουν θόρυβο.

Για περιβάλλον χωρίς θόρυβο η απόδοση του συστήματος είναι 85.40%. Η απόδοση αυτή πέφτει κατά 2.32% κατά την εκτέλεση του δεύτερου σεναρίου. Τέλος, κατά την εκτέλεση του τρίτου σεναρίου όπου εξετάζεται η ανεξαρτήτου θορύβου απόδοση του συστήματος, αυτή πέφτει στο 72.39%.

Emotion Recognition in Noise Environments

ABSTRACT

In this paper an implementation of an emotion recognition framework is presented. Within this work, we focus on a speaker-dependent emotion recognition evaluation in different additive noise environments.

For this purpose three experimental setups are exploited. Initially, we trained and tested our system with clean speech data. Secondly, the system is trained and tested with data affected by noise at a global SNR of 10 dB. Finally, the system is trained with data which are not affected by any noise while the test data are affected by noise at the same global SNR of 10 dB.

For examining the system's behavior, an utterance-level analysis was applied, showing an accuracy of 85.40% for un-noisy environment. This performance was reduced by the mean value of 2.32% when both train and test data were affected by noise. In contrast, the accuracy of the system was much lower, though much promising, for noise-independent evaluation, reaching an overall accuracy of 72.39%.

Εισαγωγή

Με τη ραγδαία πρόοδο της τεχνολογίας, η χρήση φιλικών διεπαφών χρήστη είναι απαραίτητη [1]. Καθ' όλη την διάρκεια αλληλεπίδρασης του τελικού χρήστη με τη μηχανή, η εξασφάλιση ενός ευχάριστου περιβάλλοντος είναι αναγκαία. Επίσης, είναι γνωστό ότι καθένας ως συναισθηματικό πλάσμα αρέσκεται στο να αλληλεπιδρά με συναισθηματικούς οργανισμούς. Συνεπώς, η πληροφορία που αφορά την συναισθηματική κατάσταση του τελικού χρήστη κατά τη διάρκεια αλληλεπίδρασής του με μία μηχανή είναι εξαιρετικά χρήσιμη.

Πληθώρα εφαρμογών έχουν αναπτυχθεί με στόχο τη δημιουργία φιλικών διεπαφών, χρησιμοποιώντας την ομιλία ως μέσο επικοινωνίας με τον χρήστη. Ωστόσο, η πλειονότητα των διαλογικών συστημάτων χρησιμοποιούν ένα συγκεκριμένο πλάνο προκειμένου να εξυπηρετήσουν τον χρήστη με τον οποίο αλληλεπιδρούν, μη λαμβάνοντας υπόψη την συναισθηματική κατάσταση του τελευταίου. Επιπλέον, σε πολλά έξυπνα σπίτια, ο χρήστης δυσαρεστείται από τη συμπεριφορά του συστήματος, καθώς οι διαλογικές εφαρμογές αγνοούν τα συναισθήματά του. Συνεπώς, η παρουσία ενός συστήματος αναγνώρισης συναισθημάτων είναι κάτι παραπάνω από αναγκαία.

Το πρόβλημα της αυτόματης αναγνώρισης συναισθημάτων είναι στην ουσία ένα πρόβλημα αναγνώρισης προτύπων, το οποίο χαρακτηρίζεται κατά βάση από: α) τις εξαγόμενες παραμέτρους, β) τον χρησιμοποιούμενο κατηγοροποιητή, γ) τη βάση δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση του κατηγοροποιητή και δ) τις κατηγορίες συναισθημάτων που επιλέχθηκαν ανάλογα με την εφαρμογή για την οποία προορίζεται το σύστημα.

Ένας υπολογιστής έχει τη δυνατότητα να αντιλαμβάνεται την ανθρώπινη συμπεριφορά μέσω αισθητήρων που καταγράφουν τόσο οπτική όσο και ηχητική πληροφορία.

Εξετάζοντας τις περιπτώσεις εκείνες όπου μόνο ηχητική πληροφορία χρησιμοποιήθηκε για την εξαγωγή της συναισθηματικής κατάστασης του ομιλητή, ο Yacoub [2] αναφέρθηκε στην αυτόματη αναγνώριση συναισθημάτων από ομιλία, πραγματοποιώντας εξαγωγή παραμέτρων από μικρές προτάσεις, οι οποίες χρησιμοποιούνται σε συστήματα αυτόματης απόκρισης IVR (Interactive Voice Response), εξετάζοντας τη χρήση διαφόρων κατηγοροποιητών όπως είναι τα νευρωνικά δίκτυα, SVM (Support Vector Machines), KNN (K-Nearest Neighbors) και δέντρα αναζήτησης. Τα παραπάνω πειράματα οδήγησαν στο συμπέρασμα ότι μπορούσε να διαχωριστεί ο οξύς θυμός από την ουδέτερη συναισθηματική κατάσταση με 94.00% πιθανότητα. Οι Hojcan και Kacic [3] εξέτασαν την εξαρτημένου ομιλητή αναγνώριση συναισθημάτων σε διαφορετικές γλώσσες. Χρησιμοποιώντας ένα μεγάλο σύνολο από στατιστικές παραμέτρους κατέληξαν σε μια μέση βελτίωση στην αναγνώριση του συναισθήματος της τάξης του 18% και μία μέγιστη της τάξης του 44.99%. Ο Yu [4] χρησιμοποίησε SVM και HMM προκειμένου να αναγνωρίσει 5 συναισθηματικές καταστάσεις. Για εξαρτημένου ομιλητή ανέφερε ακρίβεια 75.00%.

Εξετάζοντας την περίπτωση οπού εκτός από ακουστική πληροφορία από το σήμα ομιλίας εξάγεται και γλωσσολογική πληροφορία, ο Sculler [5] παρουσίασε μια σύγκριση συγκεκριμένων τεχνικών για ανάλυση ακουστικής και γλωσσολογικής πληροφορίας αναφέροντας μέση απόδοση συστήματος 90.30%. Για γλωσσολογική

ανάλυση χρησιμοποιήθηκε η τεχνική «Bag of Words text representation» ενώ πραγματοποιήθηκε η εξαγωγή 276 ακουστικών παραμέτρων.

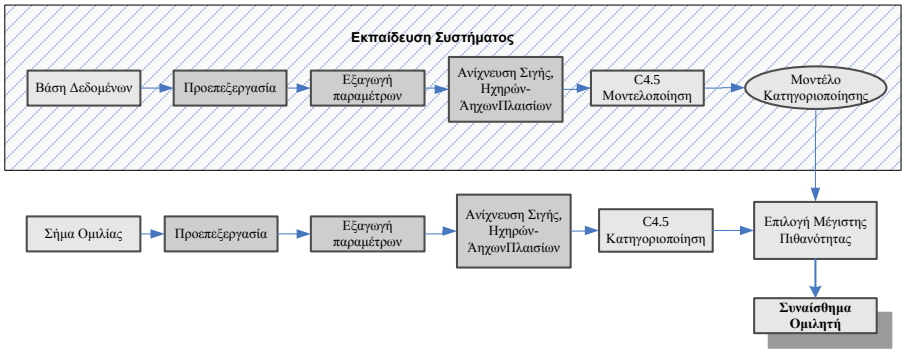
Οι Wang & Guan [6] εξέτασαν την περίπτωση που χρησιμοποιείται και οπτική πληροφορία για την πρόβλεψη της συναισθηματικής κατάστασης του τελικού χρήστη. Για την περιγραφή του σήματος ομιλίας πραγματοποιήθηκε εξαγωγή MFCC παραμέτρων και των συντονισμών του φωνητικού καναλιού, ενώ για την οπτική πληροφορία χρησιμοποιήθηκαν τα Gabor wavelet Features. Το μοντέλο κατηγοριοποίησης ήταν βασισμένο στην Fisher's Linear Discriminant Analysis και η μέση απόδοση του συστήματος είναι 82.14%.

Στην παρούσα εργασία παρουσιάζεται ένα σύστημα αναγνώρισης συναισθημάτων και εξετάζεται η συμπεριφορά αυτού σε διαφορετικές συνθήκες θορύβου. Το σύστημα, αναλύει πληροφορίες που περιέχονται μόνο από το σήμα ομιλίας, ενώ για την πρόβλεψη του αποτελέσματος χρησιμοποιήθηκε ο C4.5 κατηγοριοποιητής. Για την εξέταση της συμπεριφοράς του συστήματος χρησιμοποιήθηκε η βάση δεδομένων LDC2002S28 για τα δεδομένα καθαρής ομιλίας και η βάση δεδομένων NOISEX-92 για τα δεδομένα θορύβου.

Η παρούσα εργασία είναι οργανωμένη ως εξής: Στο πρώτο μέρος περιγράφεται η αρχιτεκτονική του συστήματος. Στο δεύτερο μέρος περιγράφονται οι βάσεις δεδομένων που χρησιμοποιήθηκαν και οι κατηγορίες συναισθημάτων που επιλέχθηκαν. Επίσης, περιγράφεται η πειραματική διαδικασία που ακολουθήθηκε και παρουσιάζονται τα πειραματικά αποτελέσματα.

1. Αρχιτεκτονική του συστήματος

Η αρχιτεκτονική του συστήματος αναγνώρισης συναισθημάτων φαίνεται στο σχήμα 1.1.



Σχήμα 1.1: Αρχιτεκτονική Συστήματος Αναγνώρισης Συναισθημάτων

Στο σύστημα αυτό, δύο βασικές διεργασίες λαμβάνουν χώρα. Η μία είναι η εκπαίδευση του συστήματος και η άλλη η κατηγοριοποίηση ενός οποιουδήποτε σήματος ομιλίας.

Η βάση δεδομένων που χρησιμοποιείται είναι στη ουσία τα δεδομένα εκπαίδευσης του συστήματος. Αυτά τα δεδομένα ανάλογα με το αν ενδιαφερόμαστε να εκπαιδεύσουμε το σύστημα σε περιβάλλον θορύβου ή όχι, έχουν διαμορφωθεί με

ή χωρίς θόρυβο. Στη συνέχεια, σε κάθε σήμα ομιλίας εφαρμόζεται προεπεξεργασία, εξαγωγή παραμέτρων και ανίχνευση σιγής και ηχηρών / άηχων πλαισίων ομιλίας.

1.1 Προεπεξεργασία και εξαγωγή παραμέτρων

Οι παράμετροι που επιλέχθηκαν να εξαχθούν ήταν τέτοιες ώστε να μην απαιτείται αναγνώριση φωνημάτων. Με αυτήν την επιλογή, το σύστημα είναι δυνατόν να χρησιμοποιηθεί σε οποιαδήποτε εφαρμογή πραγματικού χρόνου χωρίς την παρουσία αναγνωριστή φωνημάτων πραγματικού χρόνου.

Η συναισθηματική κατάσταση ενός ομιλητή εκφράζεται άμεσα μέσω της θεμελιώδους ταλάντωσης του φωνητικού καναλιού και της ενέργειας του σήματος ομιλίας. Για παράδειγμα, το ύψος φωνής και η ενέργεια ενός χαρούμενου ή θυμωμένου ατόμου είναι σε γενικές γραμμές υψηλότερη από αυτές ενός ατόμου που εκφράζει λύπη. Οι παράμετροι που επιλέχθηκαν είναι: Η θεμελιώδης ταλάντωση (pitch), οι 13 πρώτοι φασματικοί συντελεστές Mel (MFCC, Mel Frequency Cepstrum Coefficients), οι πρώτες 4 συχνότητες συντονισμού, η ενέργεια και η αρμονικότητα.

Ο υπολογισμός της θεμελιώδους ταλάντωσης, των παραμέτρων MFCC, των συχνοτήτων συντονισμού, και της αρμονικότητας έγινε με χρήση του προγράμματος επεξεργασίας ομιλίας Praat [7]. Τόσο η θεμελιώδης ταλάντωση όσο και η αρμονικότητα υπολογίστηκε με τον αλγόριθμο του Boersma [8].

Η αρμονικότητα HNR (HNR, Harmonics-to-noise-ratio) ορίζεται ως εξής:

$$HNR = 10 \cdot \log \frac{r'_{\chi}(\tau_{\max})}{1 - r'_{\chi}(\tau_{\max})} \tag{1.1}$$

Η αρμονικότητα αναπαριστά ένα μέτρο της περιοδικότητας του σήματος ομιλίας. Για ιδανικά περιοδικό σήμα η αρμονικότητα είναι άπειρη.

Ανίχνευση σιγής, ηχηρών-άηχων πλαισίων

Προκειμένου να πραγματοποιηθεί ανίχνευση σιγής και ηχηρών-άηχων πλαισίων πραγματοποιείται βραχύχρονη ανάλυση της θεμελιώδους συχνότητας και της αρμονικότητας. Από ελάχιστη ως καθόλου πληροφορία που να αφορά την συναισθηματική κατάσταση ενός ομιλητή περιλαμβάνεται στα άηχα πλαίσια του σήματος ομιλίας, εφόσον δεν υπάρχει ταλάντωση των φωνητικών χορδών. Συνεπώς, καθορίζεται ένα κατώφλι για το ύψος φωνής, ώστε όλα τα άηχα πλαίσια να απορρίπτονται. Επίσης, κατά τον υπολογισμό της αρμονικότητας καθορίζεται ένα ενεργειακό κατώφλι.

1.3 Κατηγοριοποίηση

Κατά τη διαδικασία κατηγοριοποίησης δύο βασικές διεργασίες λαμβάνουν χώρα: Η δημιουργία του μοντέλου κατηγοριοποίησης συναισθημάτων και η κατηγοριοποίηση του σήματος ομιλίας εισόδου. Η πρώτη διεργασία είναι η εκπαίδευση του συστήματος. Κατά τη δεύτερη διεργασία κάθε διάνυσμα που αντιστοιχεί σε κάθε πλαίσιο ομιλίας εισόδου κατηγοριοποιείται με χρήση του μοντέλου κατηγοριοποίησης που δημιουργήθηκε κατά τη διαδικασία της εκπαίδευσης του συστήματος. Η τελική απόφαση για το συναίσθημα που

εκφράζεται στην εκάστοτε πρόταση λαμβάνεται υπολογίζοντας την μέγιστη πιθανότητα.

Ο κατηγοριοποιητής που χρησιμοποιήθηκε είναι ο C4.5. Ο αλγόριθμος υλοποιήθηκε με χρήση της βιβλιοθήκης μηχανικής εκμάθησης WEKA [9]. Ο C4.5 είναι ένας αλγόριθμος γένεσης δέντρων αποφάσεων βασισμένος στον ID3 αλγόριθμο [10].

2. Πειραματική διαδικασία

2.1 Βάσεις Δεδομένων

Για την αξιολόγηση της συμπεριφοράς του συστήματος χρησιμοποιήθηκε η γνωστή βάση δεδομένων LDC 2002S28 (Emotional Prosody Speech and Transcripts) [11]. Η συγκεκριμένη βάση δεδομένων αποτελείται από 30 ηχογραφήσεις σε διαμόρφωση sphere και τα transcript αρχεία τους. Προκειμένου να εξαχθούν από τις ηχογραφήσεις αυτές καθαρές προτάσεις ομιλίας κατασκευάστηκε μια ειδική μηχανή. Σαν αποτέλεσμα προέκυψαν προτάσεις από 8 ηθοποιούς κατά τις οποίες εκφράζονται τα εξής συναισθήματα: Ουδετερότητα, οξύς θυμός, ψυχρός θυμός, χαρά, λύπη, αποστροφή, πανικός, αγωνία, απελπισία, ενθουσιασμός, ενδιαφέρον, ντροπή, ανία, περηφάνια και περιφρόνηση.

Για την προσθήκη θορύβου στο καθαρό σήμα ομιλίας χρησιμοποιήθηκε η βάση δεδομένων NOISEX-92 η οποία περιέχει διάφορους θορύβους όπως θόρυβος από περιβάλλον γραφείου, από εργοστάσιο, HF θόρυβος ραδιοφωνικού καναλιού, ροζ θόρυβος, λευκός θόρυβος. Επιπλέον περιέχονται σήματα θορύβου από στρατιωτικά πεδία όπως από μαχητικά αεροπλάνα (Buccaneer, F16), καταστροφικοί θόρυβοι (από δωμάτιο μηχανής-επιχειρήσεων), από τανκ (Leopard, M109), από πυροβόλο όπλο. Τέλος εμπεριέχεται θόρυβος από αυτοκίνητο (Volvo340).

Ορισμός Πειραματικής διαδικασίας

Πολλές θεωρίες υπάρχουν που προσπαθούν να εξηγήσουν την κατηγοριοποίηση των συναισθημάτων [12]. Σύμφωνα με μία επικρατούσα θεωρία, κάποια είναι τα βασικά συναισθήματα και όλα τα υπόλοιπα είναι απλά συνδυασμός ή τροποποίηση των βασικών. Υιοθετώντας αυτή τη θεωρία επιλέχθηκαν 5 βασικές συναισθηματικές καταστάσεις: Οξύς Θυμός, λύπη, χαρά, πανικός (που προσεγγίζει τον φόβο) και ουδετερότητα. Όσον αφορά τις κατηγορίες θορύβου, επιλέχθηκαν: λευκός θόρυβος, θόρυβος γραφείου, θόρυβος από πιλοτήριο αεροσκάφους, HF θόρυβος ραδιοφωνικού καναλιού και θόρυβος πυροβόλου όπλου.

Εξετάζοντας την εξαρτημένου ομιλητή αναγνώριση συναισθημάτων, εφαρμόσαμε για κάθε ηθοποιό-ομιλητή την τεχνική leave-one-out ώστε να εκμεταλλευτούμε με τον καλύτερο τρόπο όλα τα διαθέσιμα δεδομένα. Κατά την διάρκεια της προεπεξεργασίας, σε περίπτωση που επιθυμούμε να διαμορφώσουμε σήμα ομιλίας με θόρυβο, χρησιμοποιούμε το αντίστοιχο σήμα θορύβου σε αναλογία 10dB. Κάθε σήμα που προκύπτει πλαισιοποιείται σε παράθυρα των 25 msec με βήμα 10 msec.

2.3 Πειραματικά αποτελέσματα

Στον παρακάτω πίνακα φαίνονται τα πειραματικά αποτελέσματα για την περίπτωση α) που υπάρχει θόρυβος αλλά το σύστημα έχει εκπαιδευτεί ανάλογα, β) που υπάρχει θόρυβος αλλά το σύστημα είναι εκπαιδευμένο χωρίς θόρυβο. Σε κάθε περίπτωση, όταν το σύστημα είναι εκπαιδευμένο, προβλέπει καλύτερα το συναίσθημα του ομιλητή. Εξαίρεση αποτελεί η περίπτωση του θορύβου από πυροβόλο όπλο όπου λόγω των ιδιαίτερων φασματικών χαρακτηριστικών του το συναίσθημα είναι αναγνωρίσιμο εξίσου αξιόπιστα.

Είδος θορύβου	**Απόδοση (α)**	**Απόδοση (β)**
F16	82.12 %	66.19 %
HF ραδιοφωνικού καναλιού	84.12 %	71.88 %
Πυροβόλου όπλου	83.73 %	83.74 %
Γραφείου	82.39 %	72.75 %
Λευκός θόρυβος	83.02 %	67.38 %

3. Αναφορές

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J.G. Taylor, «Emotion recognition in human-computer interaction,» IEEE Signal Processing magazine, Vol. 18, no. 1, pp. 32-80, (2001)

[2] S. Yacoub, S. Simske, X. Lin, J Burns, «Recognition of emotions in interactive voice response systems,» Procedings Eurospeech 2003, pp. 729-732, (2003)

[3] V. Hojzan and Z. Kacic, «Improved emotion recognition with large set of statistical features», Proceedings Eurocspeech 2003, pp. 133-136, (2003)

[4] C. Yu, P. M. Aoki, A. Woodruff, "Detecting user engagement in everyday conversations," ICSLP 2004, vol. 2: pp. 1329-1332, 2004

[5] B. Schuller, R. Müller, M. Lang and G. Rigoll, «Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles,» Proceedings Interspeech 2005, pp. 805-808, 2005

[6] Y. Wang and L. Guan, «Recognizing human emotion from audiovisual information,» Proceedings ICASP 2005, pp. 1125-1128, 2005

[7] P. Boersma, D. Weenink, "Praat: Doing phonetics by computer (Version 4.3.28)," Computer program, http://www.praat.org/, 2005

[8] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," Proceedings 17 IFA 1993, pp.97-110, 1993

[9] I. H. Witten, E. Frank, "Data Mining: Practical machine learning tools and techniques," 2nd Edition, Morgan Kaufmann, San Francisco, 2005

[10] R. Quinlan, "C4.5: Programs for machine learning," Morgan Kaufmann Publishers Inc., 1993

[11] Linguistic Data Consortium, “Emotional Prosody Speech,” www.ldc.uppen.edu/Catalog/CatalogEntry.jsp?cataloId=LDC2002S28, University of Pennsylvania

[12] A. Ortony, G. L. Clore, A. Collins, “The cognitive structure of emotions,” Cambridge University Press, 1988