https://doi.org/10.26219/heal.aueb.4940

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ

(MSc)

στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Αναγνώριση συναισθημάτων σε προφορικούς διαλόγους»

Δημήτριος Παππάς

M13018

Επιβλέπων Καθηγητής: Ίων Ανδρουτσόπουλος

Συν-επιβλέπων: Χάρης Παπαγεωργίου

ΑΘΗΝΑ, Φεβρουάριος 2015

https://doi.org/10.26219/heal.aueb.4940

Αφιερώσεις

Στην οικογένειά μου Βασίλη, Σοφία και Γιάννη Παππά για την αμέριστη συμπαράστασή τους.

https://doi.org/10.26219/heal.aueb.4940

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

https://doi.org/10.26219/heal.aueb.4940

1. Εισαγωγή

1.1 Αντικείμενο της εργασίας

Η αναγνώριση συναισθημάτων σε οπτικοακουστικό υλικό είναι ένας επίκαιρος και απαιτητικός κλάδος της τεχνολογίας φυσικής γλώσσας. Η έρευνα που έχει γίνει μέχρι σήμερα έχει βρει εφαρμογές σε πολλούς τομείς της βιομηχανίας (π.χ. στον εντοπισμό δυσαρεστημένων πελατών από τις τηλεφωνικές συνομιλίες τους). Έχουν προταθεί πολλές διαφορετικές προσεγγίσεις, στις οποίες χρησιμοποιούνται διαφορετικοί αλγόριθμοι μηχανικής μάθησης, συνδυασμοί τους και διαφορετικά χαρακτηριστικά (features), χωρίς όμως να έχει προκύψει μία καθιερωμένη προσέγγιση.

Στη παρούσα διπλωματική εργασία εξετάζεται η περίπτωση κατά την οποία έχουμε πρόσβαση μόνο στη φωνή και όχι στην εικόνα του ομιλητή. Μελετάται εάν το συναίσθημα και πιο ιδιαίτερα ο θυμός μπορεί να αναγνωρισθεί χρησιμοποιώντας μόνο ακουστικά χαρακτηριστικά. Δοκιμάστηκαν διαφορετικά σύνολα χαρακτηριστικών, τα οποία εξάγονται χρησιμοποιώντας τη βιβλιοθήκη λογισμικού OPENSmile[1]. Δοκιμάστηκε, επίσης, ο μετασχηματισμός των χαρακτηριστικών μέσω Ανάλυσης Κυρίων Συνιστωσών (Principal Component Analysis ή PCA). Μελετήθηκε η επίδοση διαφορετικών αλγορίθμων μάθησης και ειδικότερα των Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines ή SVMs) και της Λογιστικής Παλινδρόμησης (Logistic Regression). Όλα τα πειράματα πραγματοποιήθηκαν χρησιμοποιώντας πραγματικές ηχογραφημένες συνομιλίες από ένα τηλεφωνικό κέντρο.

1.2 Κίνητρο

Η αναγνώριση συναισθήματος είναι ένα ενδιαφέρον ερευνητικά πρόβλημα, ενώ παράλληλα έχει αυξανόμενη ζήτηση στη βιομηχανία των υπολογιστικών συστημάτων, μιας και η πληροφορία για τα συναισθήματα τόσο των πελατών όσο και των υπαλλήλων μιας επιχείρησης συνεισφέρει στην επιχειρηματική ευφυΐα της εταιρίας. Ένα ακόμη κίνητρο για την παρούσα εργασία είναι ότι αφενός δεν υπάρχουν πολλά συστήματα τα οποία να έχουν εκπαιδευτεί και αξιολογηθεί σε πραγματικά δεδομένα και αφετέρου ακόμα και μεταξύ εκείνων που ήδη υπάρχουν, δεν υπάρχει κάποιο στο οποίο να χρησιμοποιούνται δεδομένα στην Ελληνική γλώσσα και μάλιστα από πραγματικές συνομιλίες μέσω τηλεφώνου.

Για να κατανοηθεί η δυσκολία του προβλήματος αλλά και η δυνατότητα συνεισφοράς των αλγόριθμων μηχανικής μάθησης στη λύση του, αναφέρεται το εξής πείραμα των van derWal κ.ά. [6]. Ζητήθηκε από $\mathbf{k}$ ανθρώπους να ηχογραφήσουν μια ομιλία τους και έπειτα να επισημειώσουν ο καθένας την ηχογραφημένη ομιλία του με τα συναισθήματα που εξέφρασε. Ύστερα τους ζητήθηκε να ακούσουν τις ηχογραφημένες ομιλίες των υπολοίπων $k-1$ ομιλητών και να επισημειώσουν τα συναισθήματα που εξέφρασαν εκείνοι. Με αυτό τον τρόπο, δημιουργήθηκαν για κάθε ομιλία $\mathbf{k}$ επισημειώσεις, μία από τον ίδιο τον ομιλητή και $\mathbf{k-1}$ από από τους υπόλοιπους. Θεωρώντας ως

[1] Βλ. http://openaudio.eu/

σωστές τις επισημειώσεις του ίδιου του ομιλητή, διαπιστώθηκε ότι αλγόριθμοι μάθησης κατάφεραν να επισημειώσουν τα συναισθήματα με μεγαλύτερη επιτυχία από ότι οι $\mathbf{k-1}$ υπόλοιποι άνθρωποι.

Στο Κεφάλαιο 2 εξετάζονται προηγούμενες έρευνες που έχουν γίνει στον τομέα της αυτόματης αναγνώρισης συναισθημάτων. Τα ποσοστά επιτυχίας που αναφέρονται στη σχετική βιβλιογραφία παρουσιάζουν περιθώρια βελτίωσης και πράγματι η παρούσα εργασία καταφέρνει ως ένα βαθμό να τα βελτιώσει. Σημειώνεται, όμως, ότι τα χαρακτηριστικά που χρησιμοποιούνται σε αρκετές προηγούμενες έρευνες δεν είναι αμιγώς ακουστικά (μπορεί να περιλαμβάνουν π.χ. τη σχέση μεταξύ των συνομιλητών ή χαρακτηριστικά εικόνας). Επομένως δεν μπορεί κανείς να γνωρίζει πώς θα συμπεριφέρονταν τα συστήματα εκείνων των ερευνών μόνο με ακουστικά χαρακτηριστικά.

1.3 Συνεισφορά

Στην παρούσα εργασία ασχολούμαστε με την αναγνώριση συναισθήματος και ιδιαίτερα του θυμού σε ηχογραφημένες συνομιλίες, στην Ελληνική γλώσσα, μεταξύ των πελατών και των υπαλλήλων ενός τηλεφωνικού κέντρου. Η παρούσα εργασία είναι η πρώτη του είδους της για τα Ελληνικά και τα πειραματικά της αποτελέσματα είναι εφάμιλλα των αποτελεσμάτων αντιστοίχων ερευνών που έχουν γίνει με συνομιλίες σε άλλες γλώσσες. Το σύστημα της εργασίας επιτυγχάνει ποσοστό ορθότητας (accuracy) της τάξης του 70-80%, ανάλογα με τον τρόπο εκπαίδευσης και αξιολόγησής του (εξηγούνται παρακάτω), αποτελέσματα που είναι αρκετά ικανοποιητικά, δεδομένου ότι χρησιμοποιούνται μόνο ακουστικά χαρακτηριστικά. Σημειώνεται, επίσης, ότι το σύστημα της εργασίας μπορεί να εντοπίσει συγκεκριμένα τμήματα της ομιλίας, όπου εκφράζεται (έντονο) συναίσθημα και ιδιαίτερα θυμός.

Το σύστημα της εργασίας μπορεί να χρησιμοποιηθεί ως αφετηρία προς ένα βελτιωμένο σύστημα που θα εντοπίζει περισσότερα συγκεκριμένα συναισθήματα (όχι μόνο θυμό), λαμβάνοντας υπόψη και λεξιλογικά χαρακτηριστικά (μέσω αναγνώρισης ομιλίας) ή χαρακτηριστικά εικόνας.

1.4 Διάρθρωση Εργασίας

Στο Κεφάλαιο 2 (Υπόβαθρο) αναφέρονται οι αλγόριθμοι μηχανικής μάθησης και μείωσης διαστάσεων που χρησιμοποιήθηκαν, μαζί με παραπομπές σε βιβλία που τους περιγράφουν αναλυτικότερα. Επίσης, παρουσιάζονται οι τρόποι διαχείρισης της ανισοκατανομής των κατηγοριών των δεδομένων (class imbalance) που υιοθετήθηκαν.

Στο Κεφάλαιο 3 (Δεδομένα) περιγράφονται εκτενώς τα δεδομένα των πειραμάτων της εργασίας, ο τρόπος χειρωνακτικής επισημείωσής τους και το εργαλείο ELAN που χρησιμοποιήθηκε κατά την επισημείωση.

Στο Κεφάλαιο 4 (Συστήματα) παρουσιάζεται το σύστημα που δημιουργήθηκε καθώς και τα επί μέρους στοιχεία του.

Στο Κεφάλαιο 5 (Πειράματα) περιγράφονται τα μέτρα αξιολόγησης που χρησιμοποιήθηκαν και τα πειραματικά αποτελέσματα διαφόρων μορφών του συστήματος της εργασίας.

https://doi.org/10.26219/heal.aueb.4940

Στο Κεφάλαιο 6 (Συμπεράσματα και μελλοντική έρευνα) συνοψίζονται τα συμπεράσματα της εργασίας και προτείνονται τρόποι μελλοντικής βελτίωσής του συστήματος.

1.5 Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Ίωνα Ανδρουτσόπουλο και τον Ερευνητή του Ερευνητικού Κέντρου «Αθηνά» και συν-επιβλέποντα κ. Χάρη Παπαγεωργίου για τις πολύτιμες συμβουλές τους κατά την εκπόνηση της εργασίας και το χρόνο που επένδυσαν ώστε να διαμορφωθεί η ερευνητική μου σκέψη. Οφείλω επίσης να ευχαριστήσω το Μάκη Μαλακασιώτη, μέλος της Ομάδας Επεξεργασίας Φυσικής Γλώσσας του Οικονομικού Πανεπιστημίου Αθηνών, για τις σημαντικές συμβουλές, ιδέες και παρατηρήσεις του. Τέλος θα ήθελα να ευχαριστήσω τα μέλη του Ερευνητικού Κέντρου «Αθηνά» κ. Μαρία Κουτσομπόγερα και κ. Δημήτρη Γαλάνη για τις συμβουλές τους, τις βάσεις και το υλικό που παρείχαν για τη δημιουργία του συστήματος αυτής της εργασίας.

https://doi.org/10.26219/heal.aueb.4940

2. Υπόβαθρο

2.1 Αλγόριθμοι μηχανικής μάθησης

Στην παρούσα εργασία χρησιμοποιήθηκαν αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης και πιο συγκεκριμένα κυρίως λογιστική παλινδρόμηση (Logistic Regression, LR) και μηχανές διανυσμάτων υποστήριξης (Support Vector Machines, SVMs) με γραμμικό πυρήνα. Οι αλγόριθμοι αυτοί χρησιμοποιήθηκαν διότι είναι πολύ διαδεδομένοι και επιτυγχάνουν πολύ καλά αποτελέσματα σε πολλά προβλήματα κατηγοριοποίησης. Εξετάστηκαν επίσης εν μέρει Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models, HMMs) και Υπό Όρους Τυχαία Πεδία γραμμικής αλυσίδας (linear-chain Conditional Random Fields, CRFs). Όλοι αυτοί οι αλγόριθμοι περιγράφονται σε καθιερωμένα βιβλία μηχανικής μάθησης (βλ. π.χ. [40], [41]).

2.2 Μείωση διαστάσεων με Ανάλυση Κυρίων Συνιστωσών

Η Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis ή PCA) απεικονίζει ένα σύνολο από πιθανώς συσχετισμένες τυχαίες μεταβλητές σε ένα χώρο όπου οι (νέες) μεταβλητές είναι γραμμικά ανεξάρτητες. Για να το επιτύχει αυτό χρησιμοποιεί τις ιδιοτιμές και τα ιδιοδιανύσματα των δεδομένων (δείγματος τιμών των τυχαίων μεταβλητών) που παίρνει ως είσοδο. Στην περίπτωση της επιβλεπόμενης μηχανικής μάθησης, οι τυχαίες μεταβλητές είναι χαρακτηριστικά (features) και η PCA χρησιμοποιείται για να δημιουργηθεί ένα νέο σύνολο (συνήθως λιγότερων) και γραμμικά ανεξάρτητων νέων χαρακτηριστικών. Η PCA και η χρήση της στη μηχανική μάθηση περιγράφονται εκτενώς στη βιβλιογραφία (βλ. π.χ. [38]) .

Σε γενικές γραμμές, και όπως αναφέρεται στο [38], η PCA εφαρμόζεται ως εξής:

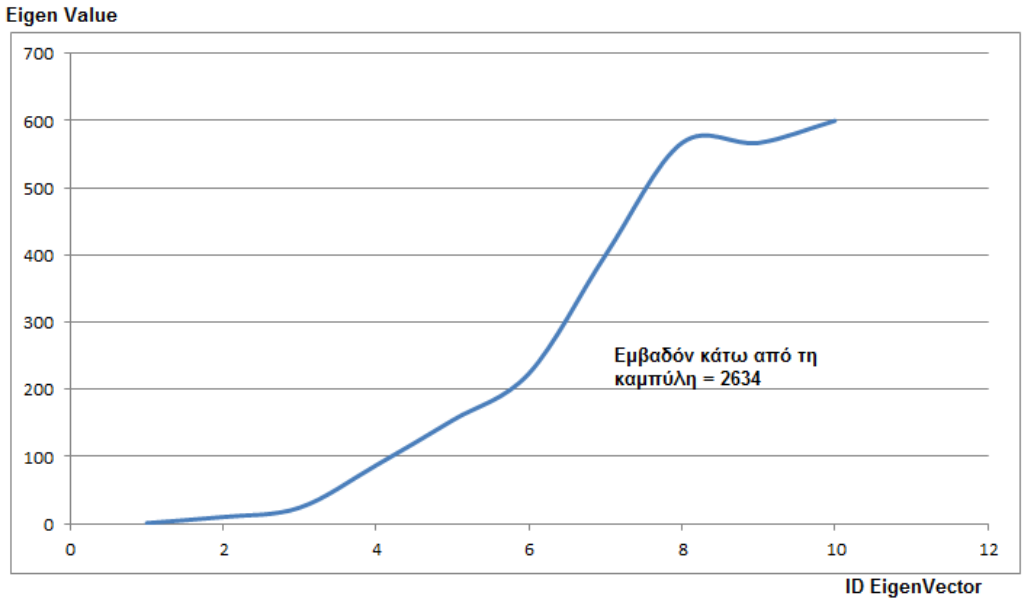
1. Λαμβάνουμε κάποια δεδομένα εκπαίδευσης (πίνακας A).
2. Αφαιρούμε το μέσο όρο από κάθε χαρακτηριστικό.
3. Υπολογίζουμε τον τετραγωνικό πίνακα συνδιακύμανσης (Σ).
4. Υπολογίζουμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα Σ.
5. Ταξινομούμε τις ιδιοτιμές (και τα αντίστοιχα ιδιοδιανύσματα) κατά φθίνουσα σειρά.
6. Επιλέγουμε τα k πρώτα ιδιοδιανύσματα, όπου k το πλήθος των χαρακτηριστικών που θέλουμε να έχει το νέο σύνολο χαρακτηριστικών, και δημιουργούμε ένα πίνακα $FeatVec = (ιδιοδιάνυσμα_1, \dots, ιδιοδιάνυσμα_k)$.
7. Πολλαπλασιάζουμε την αντιμετάθεση του $FeatVec$ με την αντιμετάθεση του πίνακα των αρχικών μας δεδομένων και την αντιμετάθεση των δεδομένων αξιολόγησης.

Σημειώνεται ότι η PCA είναι ευαίσθητη στο εύρος τιμών των αρχικών μεταβλητών, γι' αυτό και οι τιμές των μεταβλητών του δείγματος αρχικά συνήθως κανονικοποιούνται (βλ. παρακάτω).

https://doi.org/10.26219/heal.aueb.4940

Για να επιλέξουμε το $\mathbf{k}$, δηλαδή το πλήθος των νέων χαρακτηριστικών, σε αυτή την εργασία ακολουθούμε την εξής διαδικασία:

1. Υπολογίζουμε τις ιδιοτιμές και τα ιδιοδιανύσματα όπως παραπάνω.
2. Ταξινομούμε τις ιδιοτιμές (και αντίστοιχα τα ιδιοδιανύσματα) κατά αύξουσα σειρά. Διαισθητικά, οι ιδιοτιμές μας δείχνουν την πληροφορία του αρχικού συνόλου δεδομένων εκπαίδευσης την οποία περιέχουν τα αντίστοιχα ιδιοδιανύσματα, πιο συγκεκριμένα τη διαφορά της πληροφορίας του εκάστοτε ιδιοδιανύσματος από τα υπόλοιπα.

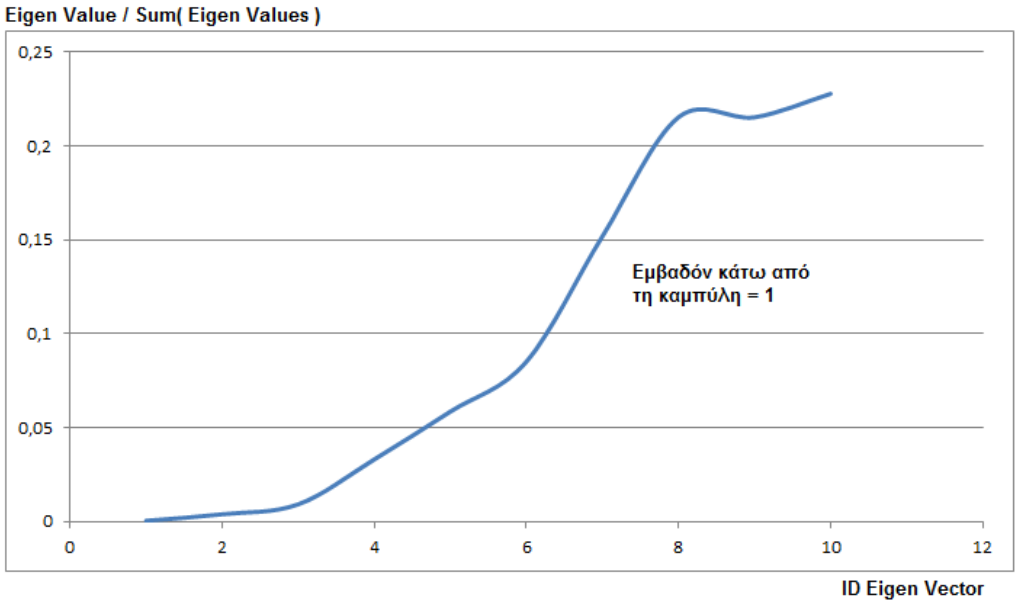


Εικόνα 1: Καμπύλη ιδιοτιμών σε αύξουσα σειρά

3. Έπειτα υπολογίζουμε για κάθε ιδιοτιμή X_i την τιμή $\frac{X_i}{\sum_{j=1}^{n} X_j}$ όπου n το πλήθος των ιδιοτιμών.

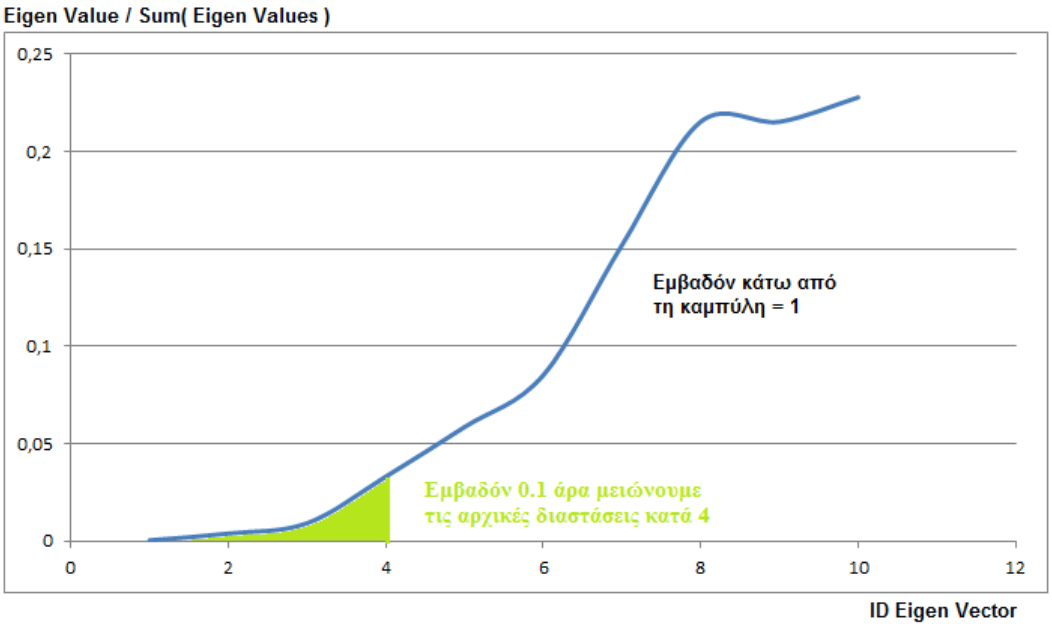
   Καταλήγουμε έτσι, σε μία νέα καμπύλη με τιμές στο διάστημα (0,1) και εμβαδόν ένα.

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 2: Κανονικοποιημένη καμπύλη ιδιοτιμών με εμβαδόν 1

4. Τέλος επιλέγουμε να χαθεί πληροφορία που να αντιστοιχεί π.χ. στο 10% του εμβαδού. Στο παράδειγμα αφαιρούνται 4 ιδιοδιανύσματα από τα συνολικά 10, οπότε ο νέος χώρος έχει 6 διαστάσεις). Το ποσοστό του συνολικού εμβαδού που αφαιρείται είναι παράμετρος του συστήματος.



Εικόνα 3: Επιλογή χαρακτηριστικών με χρήση της καμπύλης των ιδιοτιμών

2.3 Διαχείριση ανισοκατανομής κατηγοριών

Ως ανισοκατανομή κατηγοριών (class imbalance) εννοούμε την περίπτωση όπου το πλήθος των στιγμιοτύπων εκπαίδευσης μιας κατηγορίας ξεπερνά κατά πολύ το πλήθος των στιγμιοτύπων εκπαίδευσης μίας ή περισσοτέρων άλλων κατηγοριών. Αυτό αποτελεί πρόβλημα, γιατί συχνά οι αλγόριθμοι μάθησης, έχοντας πολύ περισσότερα παραδείγματα από μία κατηγορία, καταλήγουν να

μεροληπτούν εις βάρος των υπολοίπων. Έχουν προταθεί πολλές τεχνικές αντιμετώπισης του προβλήματος αυτού (βλ. π.χ. [8]):

- Υπερδειγματοληψία της μειονεκτικής κατηγορίας με τυχαία δειγματοληψία (ROS). Στην περίπτωση αυτή προσθέτουμε στα δεδομένα εκπαίδευσης αντίγραφα τυχαία επιλεγμένων στιγμιοτύπων εκπαίδευσης της μειονεκτικής κατηγορίας, ώστε οι δύο κατηγορίες να καταλήξουν να έχουν το ίδιο πλήθος στιγμιοτύπων εκπαίδευσης.
- Υποδειγματοληψία της πλεονεκτικής κατηγορίας με τυχαία δειγματοληψία (RUS). Στην περίπτωση αυτή αφαιρούμε από τα δεδομένα εκπαίδευσης τυχαία επιλεγμένα στιγμιότυπα της μειονεκτικής κατηγορίας, ώστε οι δύο κατηγορίες να καταλήξουν πάλι να έχουν το ίδιο πλήθος στιγμιοτύπων εκπαίδευσης.
- Από κοινού τυχαία υπερδειγματοληψία της μειονεκτικής και υποδειγματοληψία της πλεονεκτικής κατηγορίας (ROS και RUS).
- Συνθετική υπερδειγματοληψία της μειονεκτικής κατηγορίας (SMOT, SMOTE, κ.α.) [8][9].
- Δημιουργία αλγορίθμων μάθησης που χρησιμοποιούν κάποιο κόστος (ή βάρος) υψηλότερο για την μειονεκτική κατηγορία και χαμηλότερο για την πλεονεκτική.
- Υπερδειγματοληψία χρησιμοποιώντας ομαδοποίηση (clustering) δεδομένων, αφαιρώντας τα περιττά ή θορυβώδη δείγματα [8].
- Άλλοι αλγόριθμοι ενίσχυσης (BABoost κ.α.) [8].

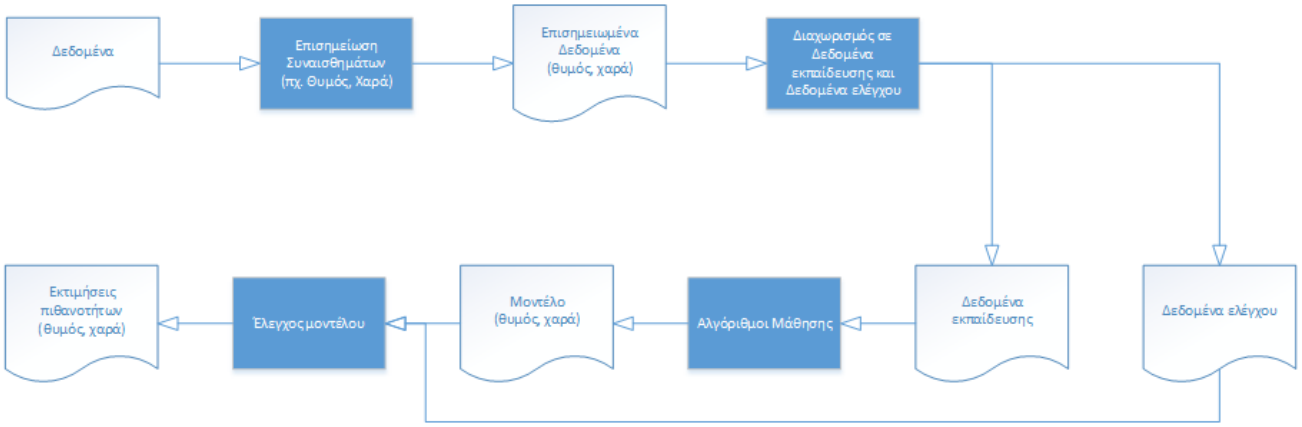
Στην παρούσα εργασία, χρησιμοποιούνται οι τεχνικές ROS και RUS (οι κατηγορίες είναι πάντα δύο).

2.4 Σχετική έρευνα

Οι προηγούμενες έρευνες που μελετήθηκαν στη διάρκεια της παρούσας εργασίας τοποθετούνται χρονικά από το 2001 μέχρι το 2014. Ο μεγαλύτερος προβληματισμός κατά την αναγνώριση συναισθήματος φαίνεται να είναι ο σαφής διαχωρισμός των κατηγοριών. Κάποιοι ξεχωρίζουν το συναίσθημα σε τρεις κατηγορίες: καθόλου, λίγο ή πολύ [13]. Άλλοι ξεχωρίζουν το συναίσθημα σε δύο κατηγορίες: χωρίς συναίσθημα, με συναίσθημα [29] ή αρνητικά και θετικά συναισθήματα [12]. Παρατηρείται, επίσης, ότι οι άνθρωποι-επισημειωτές συχνά διαφωνούν ως προς τη σωστή επισημείωση τμημάτων ομιλίας, όταν έχουν πια καθοριστεί οι δυνατές κατηγορίες. Για αυτό το λόγο χρησιμοποιείται συχνά ο συντελεστής Kappa (Cohen's kappa coefficient) [37], ένα μέτρο για την μέτρηση της συμφωνίας μεταξύ επισημειωτών (inter-annotator agreement).

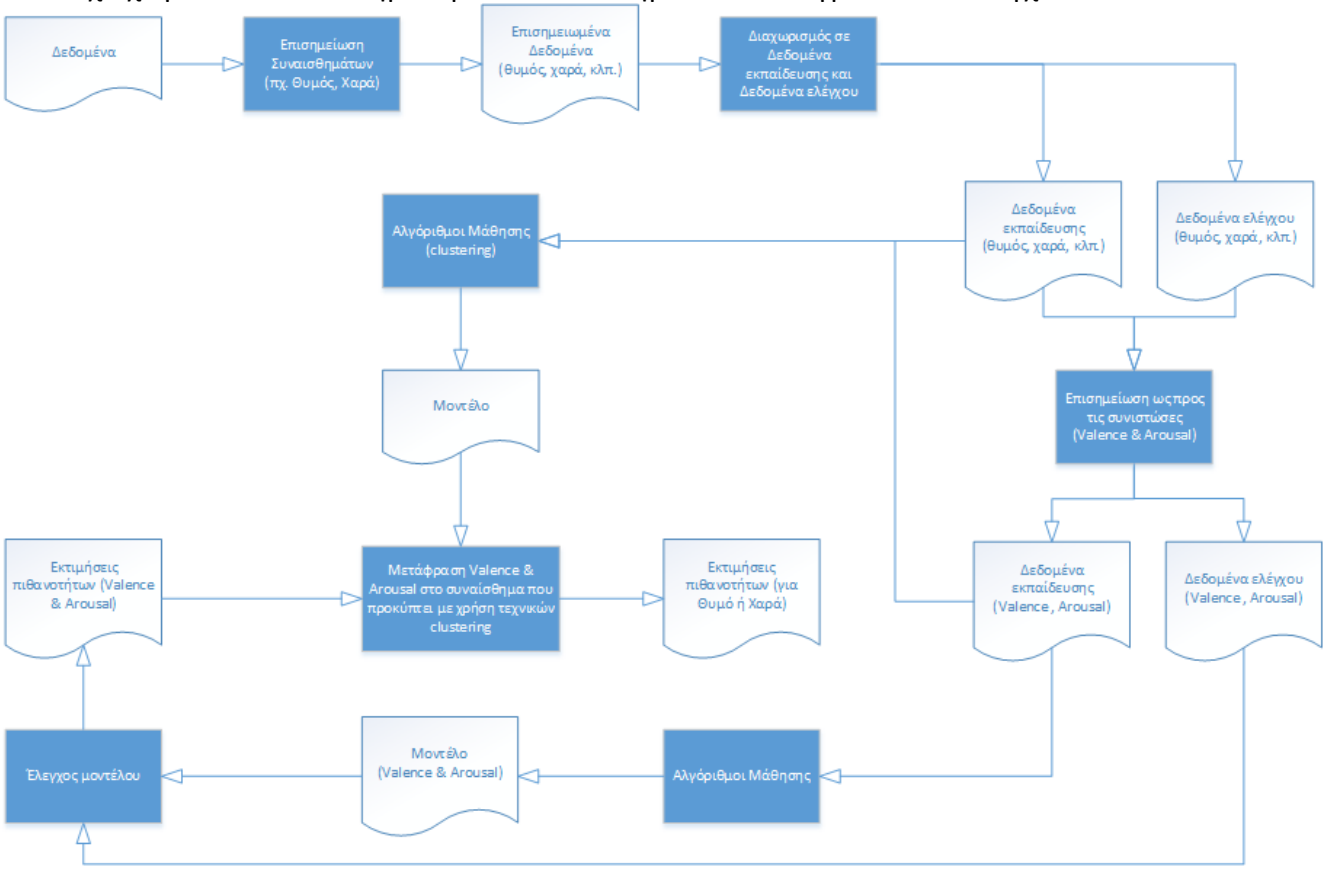
Οι ερευνητές του τομέα της ανάλυσης συναισθήματος προσεγγίζουν το πρόβλημα από δύο κατευθύνσεις. Σύμφωνα με τη πρώτη προσέγγιση, δημιουργούμε χειρωνακτικά επισημειωμένα δεδομένα (π.χ. τμήματα ομιλίας), τα χωρίζουμε σε δεδομένα εκπαίδευσης (training data) και δεδομένα ελέγχου (test data), εκπαιδεύουμε ένα μοντέλο στα δεδομένα εκπαίδευσης μέσω ενός αλγορίθμου μάθησης και κατόπιν αξιολογούμε το μοντέλο στα δεδομένα ελέγχου (Εικόνα 14).

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 4: Μοντέλο συστημάτων χωρίς μετασχηματισμό των συναισθημάτων

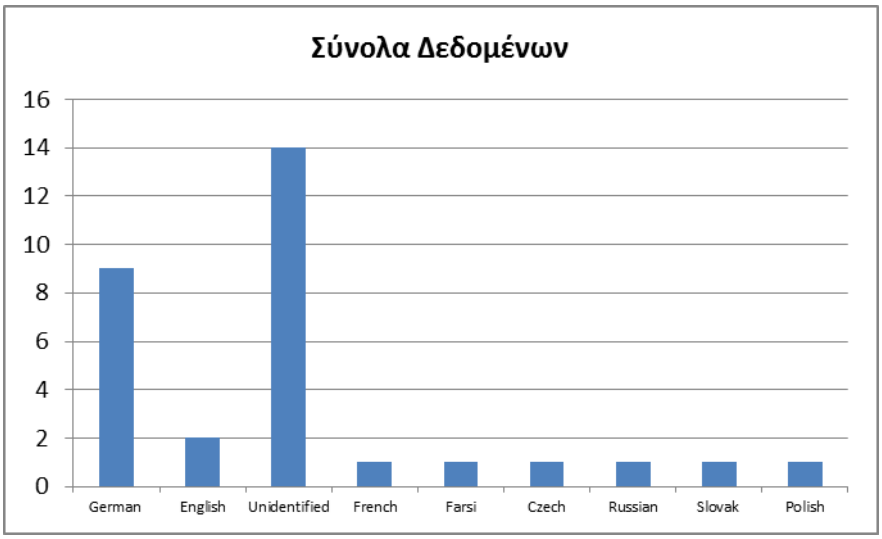
Κατά τη δεύτερη προσέγγιση γίνονται δύο επισημειώσεις για κάθε συνομιλία. Μία επισημείωση για το συναίσθημα (χαρά, φόβος, κλπ.) και μία επισημείωση για δύο συνήθως συνιστώσες: valence και arousal (ή άλλες συνιστώσες που ορίζονται ανάλογα με το πρόβλημα [16]). Ένα μοντέλο εκπαιδεύεται να προβλέπει τις συνιστώσες (βλ. Εικόνα 5). Για κάθε στιγμιότυπο των δεδομένων ελέγχου, το εκπαιδευμένο μοντέλο επιστρέφει μία τιμή ανά συνιστώσα. Οι τιμές αυτές μετατρέπονται σε εκτιμήσεις συναισθήματος (χαρά, φόβος, κλπ.) συγκρίνοντάς τις με τις αντίστοιχες τιμές των χειρωνακτικά επισημειωμένων στιγμιοτύπων εκπαίδευσης, για τα οποία είναι διαθέσιμα τόσο τα συναισθήματα όσο και οι τιμές των συνιστωσών (valence, arousal). Η σύγκριση γίνεται συχνά εφαρμόζοντας ομαδοποίηση (clustering) στα χειρωνακτικά επισημειωμένα στιγμιότυπα εκπαίδευσης. Για κάθε στιγμιότυπο ελέγχου για το οποίο έχουμε μια εκτίμηση των συνιστωσών (valence, arousal), βρίσκουμε π.χ. την ομάδα (cluster) των στιγμιοτύπων εκπαίδευσης με τις πιο κοντινές τιμές συνιστωσών και κατατάσσουμε το στιγμιότυπο ελέγχου στο πιο συχνό συναίσθημα (χαρά, φόβος κλπ.) της ομάδας. Οι προβλέψεις συναισθήματος για τα στιγμιότυπα ελέγχου συγκρίνονται τελικά με τα αντίστοιχα χειρωνακτικά επισημειωμένα συναισθήματα των στιγμιοτύπων ελέγχου.



Εικόνα 5: Μοντέλο συστημάτων με μετασχηματισμό συναισθημάτων σε Valence Arousal

https://doi.org/10.26219/heal.aueb.4940

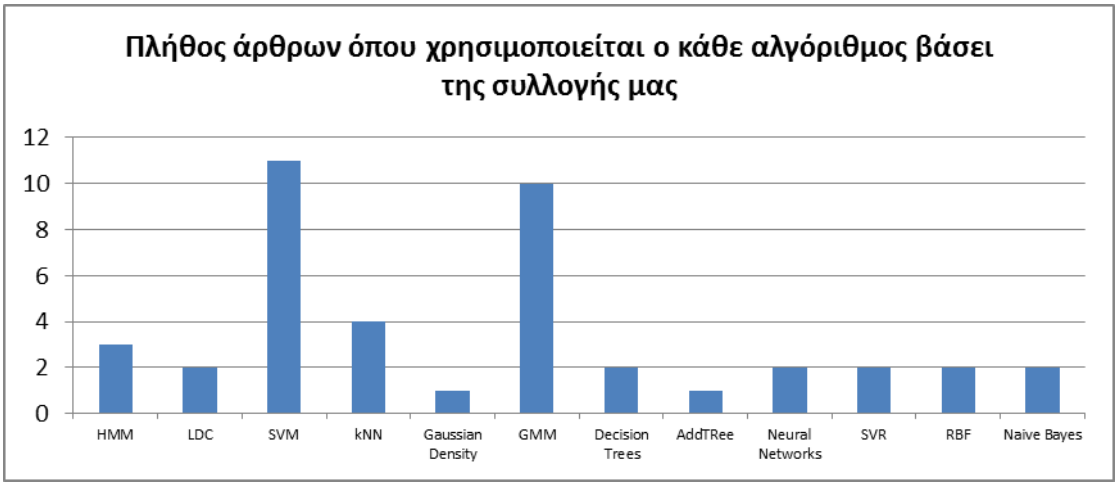
Στη βιβλιογραφία τα δεδομένα που χρησιμοποιούνται είναι κατά κύριο λόγο ηχογραφήσεις ηθοποιών που τους ζητείται να υποκριθούν συναισθήματα (acted data) [15] και σε ορισμένες περιπτώσεις δεδομένα από συνομιλίες με διαλογικά συστήματα [25] ή υπαλλήλους [30]. Τα περισσότερα σχετικά σύνολα δεδομένων που εντοπίστηκαν στη διάρκεια της παρούσας εργασίας περιέχουν ηχογραφήσεις στη Γερμανική γλώσσα (Εικόνα 6).



Εικόνα 6: Γράφημα πλήθους συνόλων δεδομένων ανά γλώσσα

Τα χαρακτηριστικά που εξάγονται στα πειράματα της βιβλιογραφίας είναι ακουστικά, λεξικολογικά και οπτικά. Τα ακουστικά χαρακτηριστικά είναι εκείνα τα οποία προέρχονται από τη κυματομορφή του σήματος, όπως διάρκεια, ενέργεια, τόνος ή συχνότητα. Τα λεξικολογικά είναι εκείνα που προέρχονται είτε από αυτόματη αναγνώριση ομιλίας είτε από αυτόματη αναγνώριση συγκεκριμένων λεξικολογικών όρων, όπως λέξεις, φωνήματα, διγράμματα. Τα οπτικά είναι εκείνα που λαμβάνουμε ύστερα από καταγραφή του ομιλητή με οπτικά μέσα και συνήθως αφορούν τα χαρακτηριστικά του προσώπου του, τις κινήσεις του ή τη στάση του σώματός του.

Οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται στη βιβλιογραφία συνοψίζονται στο Παράρτημα 1. Ο παρακάτω πίνακας συνοψίζει τη συχνότητα χρήσης κάθε αλγορίθμου μάθησης.



Εικόνα 7: Γράφημα πλήθους αναφορών σε άρθρα ανά αλγόριθμο

Παρατηρείται λοιπόν ότι κατά κύριο λόγο στην αναγνώριση συναισθημάτων στη φωνή χρησιμοποιούνται SVMs και Gaussian Mixture Models (GMMs) [42].

Παραθέτουμε παρακάτω περισσότερες πληροφορίες για τις έρευνες που είναι πιο συναφείς με τη δική μας:

Αρχικά οι Min Lee κ.ά. [11] κατηγοριοποίησαν καταγραφές από ένα τηλεφωνικό κέντρο, που χρησιμοποιεί Intelligent Virtual Agents για την εξυπηρέτηση των πελατών, σε αρνητικές και μη. Στη συνέχεια εξήγαγαν χαρακτηριστικά pitch και ενέργειας. Χρησιμοποιώντας LDC, SVMs και kNN καθώς και PCA για μείωση διαστάσεων πέτυχαν ποσοστό ορθότητας (accuracy) 72,46%. Τα δεδομένα μας διαφέρουν, καθώς στην περίπτωσή μας υπάρχουν δύο άνθρωποι συνομιλητές.

Οι Vidrascu κ.ά. [18] πέτυχαν ορθότητα 73,00% προσπαθώντας να ξεχωρίσουν αρνητικές συνομιλίες από ουδέτερες. Χρησιμοποίησαν δεδομένα στη Γαλλική γλώσσα από μια διαδικτυακή εταιρία αγοραπωλησίας μετοχών και εξήγαγαν χαρακτηριστικά ενέργειας, διάρκειας συνομιλίας, f0, φασματικά (spectral) και προβλημάτων άρθρωσης (disfluency). Οι αλγόριθμοι μάθησης που εξέτασαν είναι SVM, δέντρα απόφασης και ψηφοφορίας (voting). Οι ίδιοι [19] κατάφεραν να ξεχωρίσουν αρνητικά από θετικά συναισθήματα με ορθότητα 83,49% σε δεδομένα από το τηλεφωνικό κέντρο ενός ιατρικού κέντρου. Χρησιμοποίησαν SVM και Logistic Model Trees και χαρακτηριστικά που αφορούν την προσωδία, φασματικά χαρακτηριστικά, παύσεις, θόρυβο κλπ. (non-verbal events), προβλήματα άρθρωσης, την ενέργεια, f0 και τη διάρκεια της συνομιλίας. Το πρόβλημα που εξετάζουμε διαφέρει από την άποψη ότι εμείς προσπαθούμε να ανακαλύψουμε εάν υπάρχει συναίσθημα και έπειτα να βρούμε ένα συγκεκριμένο αρνητικό συναίσθημα (θυμός).

Οι Burkhardt κ.ά. [25] χρησιμοποίησαν δεδομένα από ένα Voice Portal όπου ο πελάτης συνομιλούσε με ένα αυτόματο διαλογικό σύστημα. Εξήγαγαν χαρακτηριστικά προσωδίας, διάρκειας, pitch, intensity, και MFCC [44] για να ξεχωρίσουν συνομιλίες που περιείχαν το στοιχείο του θυμού από εκείνες που δεν το περιείχαν. Έπειτα από μείωση διαστάσεων με PCA και χρησιμοποιώντας τους αλγορίθμους μάθησης GMM και SVM πέτυχαν ποσοστό ορθότητας 70,00%.

Οι Neiberg κ.ά. [26] χρησιμοποίησαν δεδομένα από μια Σουηδική εταιρεία τηλεφωνίας για να εξετάσουν το πρόβλημα αυτόματης αναγνώρισης του θυμού σε αυθόρμητο λόγο. Χώρισαν τα δεδομένα σε τρεις κατηγορίες (αρνητικό, εμφατικό ή έντονο, ουδέτερο λόγο) σύμφωνα με την επισημείωση ενός έμπειρου ερευνητή φωνής, χρησιμοποίησαν GMMs για την ταξινόμηση και MFCC χαρακτηριστικά. Τα αποτελέσματα που παρουσιάζουν δείχνουν ποσοστό ορθότητας 76.50 %.

Το Παράρτημα 1 συνοψίζει τα σχετικά άρθρα που μελετήθηκαν στη διάρκεια της εργασίας, τα δεδομένα που χρησιμοποιούν, τους αλγορίθμους μάθησης, τα μέτρα αξιολόγησης, τα αποτελέσματά τους κλπ. Ο πίνακας είναι ταξινομημένος σύμφωνα με το έτος δημοσίευσης, κατά αύξουσα σειρά.

https://doi.org/10.26219/heal.aueb.4940

3. Δεδομένα

3.1 Πηγές δεδομένων

Τα δεδομένα τα οποία εξετάζει η βιβλιογραφία μπορούν να χωριστούν σε τρεις κατηγορίες. Η πρώτη κατηγορία (Acted), αφορά δεδομένα (ομιλία) τα οποία έχουν ηχογραφηθεί σε εργαστηριακές, ελεγχόμενες συνθήκες με ηθοποιούς στους οποίους είχε δοθεί ένα σενάριο, βάσει του οποίου έπρεπε να εκφράσουν συγκεκριμένα συναισθήματα. Η δεύτερη κατηγορία (Semi-acted) αφορά δεδομένα τα οποία, όπως και η πρώτη κατηγορία, έχουν ηχογραφηθεί σε εργαστηριακές, ελεγχόμενες συνθήκες. Σε αυτή τη κατηγορία, όμως, προσπαθούμε να εκμαιεύσουμε συναισθήματα από ανυποψίαστους ανθρώπους, προβάλλοντάς τους κάποια ερεθίσματα που μπορεί να επιφέρουν το αντίστοιχο συναίσθημα. Στη τρίτη κατηγορία (Real), τα δεδομένα συλλέγονται από συνομιλίες σε πραγματικό περιβάλλον, από ομιλητές οι οποίοι εκφράζουν συναισθήματα για πραγματικά γεγονότα, σε πραγματικές συνθήκες. Τέτοια δεδομένα αφορούν συνήθως συνομιλίες οι οποίες καταγράφονται σε τηλεφωνικά κέντρα, όπου οι πελάτες μιας εταιρίας συνομιλούν είτε με κάποιον υπάλληλο, είτε με αυτόματα συστήματα ερωταποκρίσεων. Στη περίπτωσή μας εξετάσαμε δεδομένα της τρίτης κατηγορίας (Real).

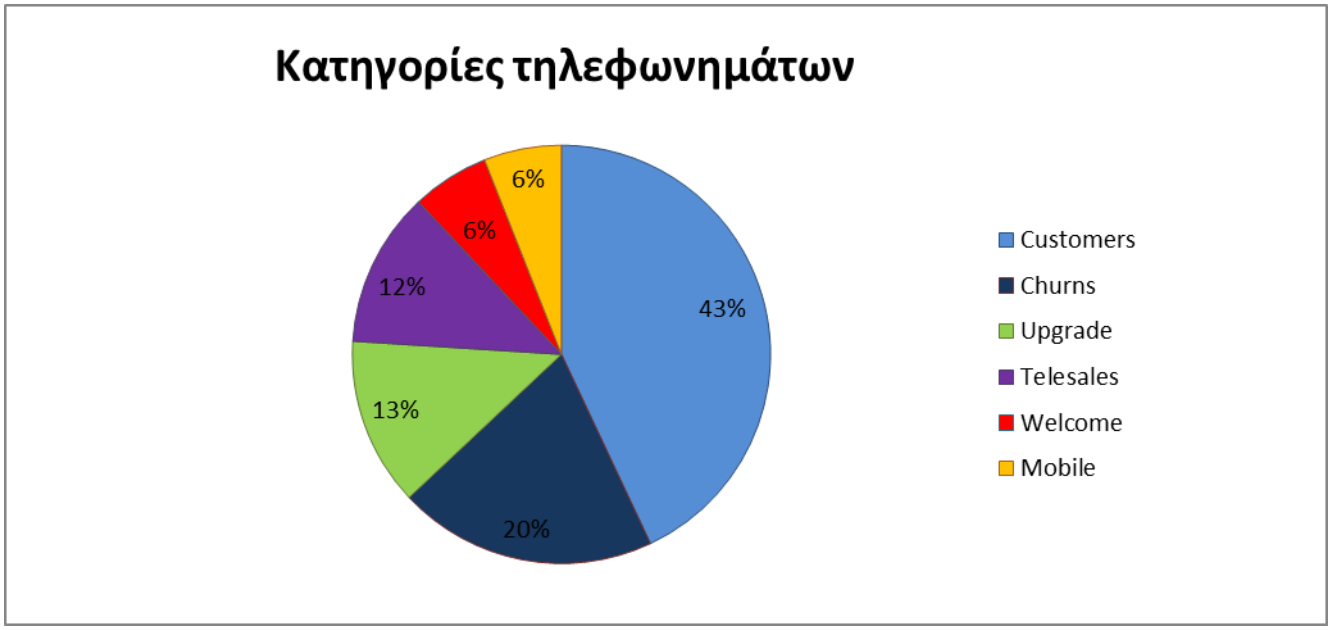
Τα δεδομένα που χρησιμοποιήθηκαν στη παρούσα εργασία προέρχονται από ηχογραφήσεις στο τηλεφωνικό κέντρο μίας εταιρίας που παρέχει υπηρεσίες τηλεφωνίας. Είναι ηχογραφημένα σε ένα κανάλι, 16 bits ανά δείγμα και ρυθμό δειγματοληψίας 8000 Hz. Πρόκειται για συνολικά 137 αρχεία ήχου τα οποία χωρίζονται σε 6 κατηγορίες:

- από την εταιρία προς τον πελάτη για τη προώθηση κινητής τηλεφωνίας (Mobile),
- από την εταιρία προς τον πελάτη για αλλαγή πακέτου ομιλίας (Upgrade),
- από την εταιρία προς τον πελάτη για να γίνει νέος συνδρομητής (Telesales),
- από την εταιρία προς τον πελάτη για να τον καλωσορίσει στην εταιρία (Welcome),
- από τον πελάτη προς την εταιρία για να ζητήσει διακοπή (Churns),
- από τον πελάτη προς την εταιρία για να ενημερωθεί (Customers).

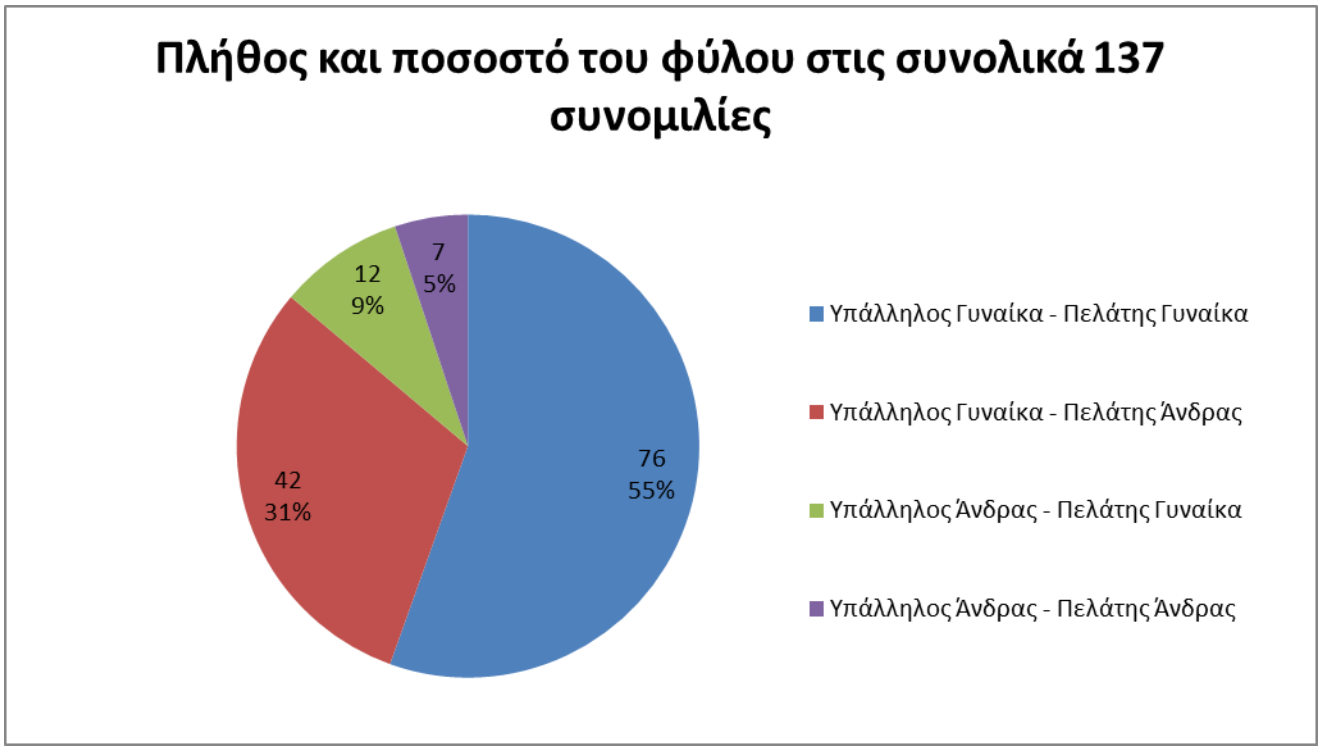
Η Εικόνα 8 δείχνει το ποσοστό του συνολικού πλήθους των κλήσεων που υπάρχουν στις παραπάνω κατηγορίες.

Μαζί με τα αρχεία ήχου η εταιρία προσέφερε και στοιχεία για τον κωδικό αριθμό (id) του υπαλλήλου που μιλάει, καθώς και το φύλο του. Σε κάθε συνομιλία από τις συνολικά 137 υπάρχουν δύο ομιλητές. Καταγράψαμε τις 4 διαφορετικές περιπτώσεις στην κατανομή των φύλων συνομιλητών (άνδρας υπάλληλος-άνδρας πελάτης, άνδρας υπάλληλος- γυναίκα πελάτης, γυναίκα υπάλληλος-άνδρας πελάτης, γυναίκα υπάλληλος-γυναίκα πελάτης,). Η Εικόνα 9 δείχνει τα αποτελέσματα.

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 8: Τα είδη των συνομιλιών στα δεδομένα της εργασίας



Εικόνα 9: Πλήθος και ποσοστό του φύλου στις συνολικά 137 συνομιλίες

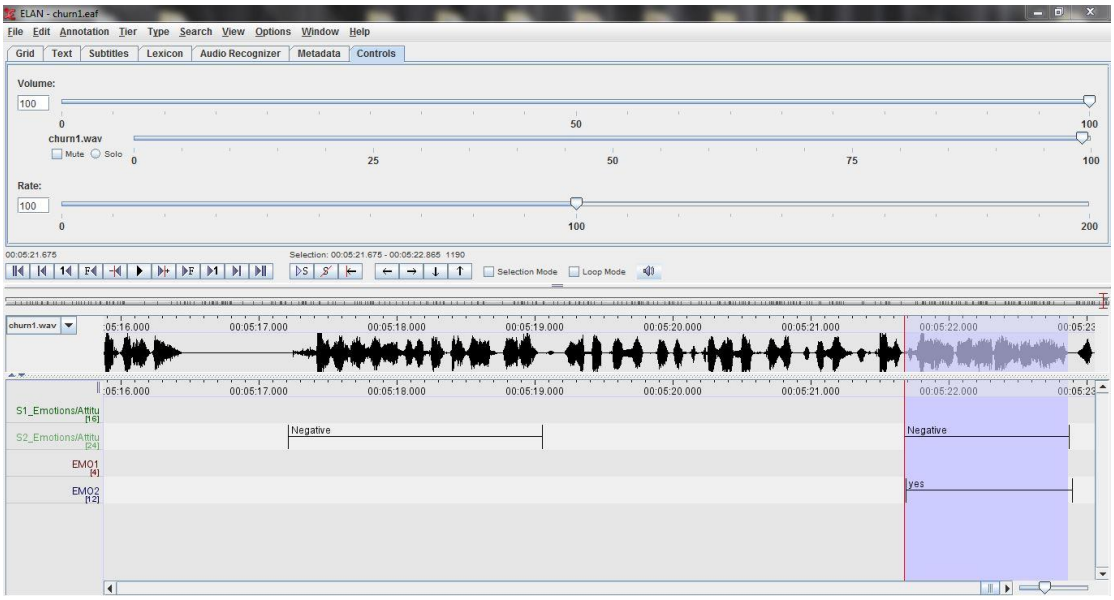
4.2 Επισημείωση

Για να γίνει χειρωνακτική επισημείωση σε κάθε κομμάτι ήχου, χρησιμοποιήθηκε το εργαλείο ELAN, έκδοση 4.7.2, στιγμιότυπο του οποίου παρουσιάζεται στην Εικόνα 10.[2] Το συγκεκριμένο εργαλείο επιτρέπει την επισημείωση σε αρχεία ήχου και βίντεο, οργανωμένη σε επίπεδα ιεραρχίας (βλ. Εικόνα 11), με ευελιξία στα όριά της έως και σε επίπεδο msec. Το γραφικό του περιβάλλον επιτρέπει την επιλογή οποιουδήποτε κομματιού από το αρχείο, την απόκρυψη και εμφάνιση επιπέδων για καλύτερη επισημείωση και τη χρήση προκαθορισμένων όρων, για να οριστεί μια ενιαία μορφή

[2] Βλ. https://tla.mpi.nl/tools/tla-tools/elan/

https://doi.org/10.26219/heal.aueb.4940

επισημειώσεων από όλους όσοι ασχολούνται με την επισημείωση σε ένα συγκεκριμένο πείραμα. Τέλος επιτρέπει την προβολή στατιστικών στοιχείων αλλά και τη σύγκριση επισημειωτών.

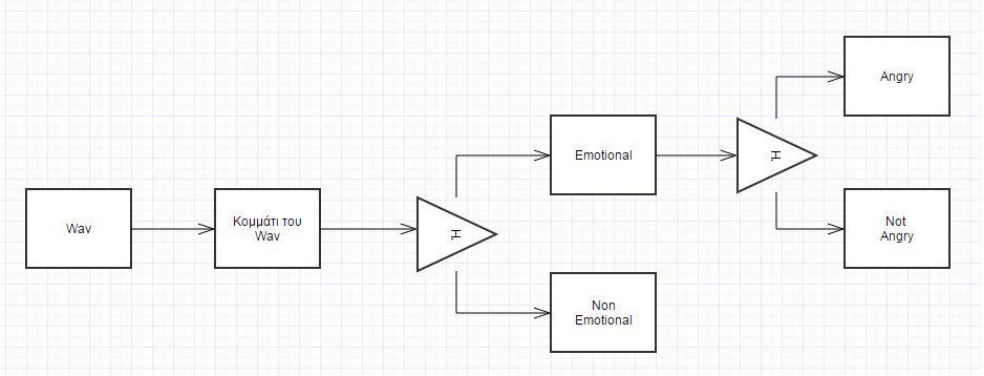


Εικόνα 10: Στιγμιότυπο εργαλείου ELAN

Το ELAN χρησιμοποιήθηκε για την επισημείωση τμημάτων ομιλίας όπου:

- ομιλεί ο υπάλληλος
- ομιλεί ο πελάτης
- υπάρχει ή δεν υπάρχει οποιοδήποτε συναίσθημα
    - υπάρχει θετικό ή αρνητικό συναίσθημα
- υπάρχει θυμός.

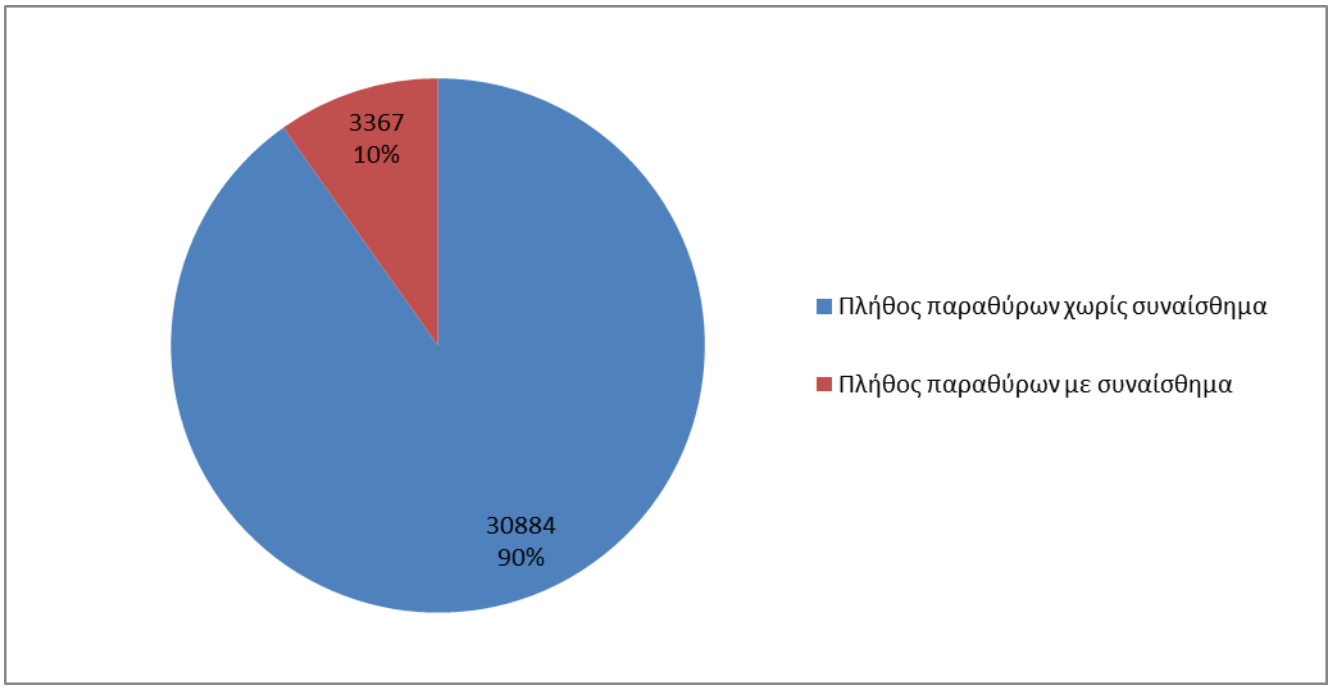
Τα συναισθήματα που ο επισημειωτής (ο συγγραφέας αυτής της εργασίας) έπρεπε να εντοπίσει είναι εκείνα που ορίστηκαν από τον Paul Ekman το 1972 [1], τα οποία είναι: χαρά, έκπληξη, αηδία, φόβος, θυμός και λύπη. Εάν σε κάποιο τμήμα υπήρχε οποιοδήποτε από αυτά τα συναισθήματα, ο επισημειωτής σημείωνε το τμήμα ως Emotional και έπειτα εάν ο ομιλητής ήταν θυμωμένος, το επισημείωνε ως Angry αλλιώς ως Not Angry (Εικόνα 11).



Εικόνα 11: Ιεραρχία επισημείωσης

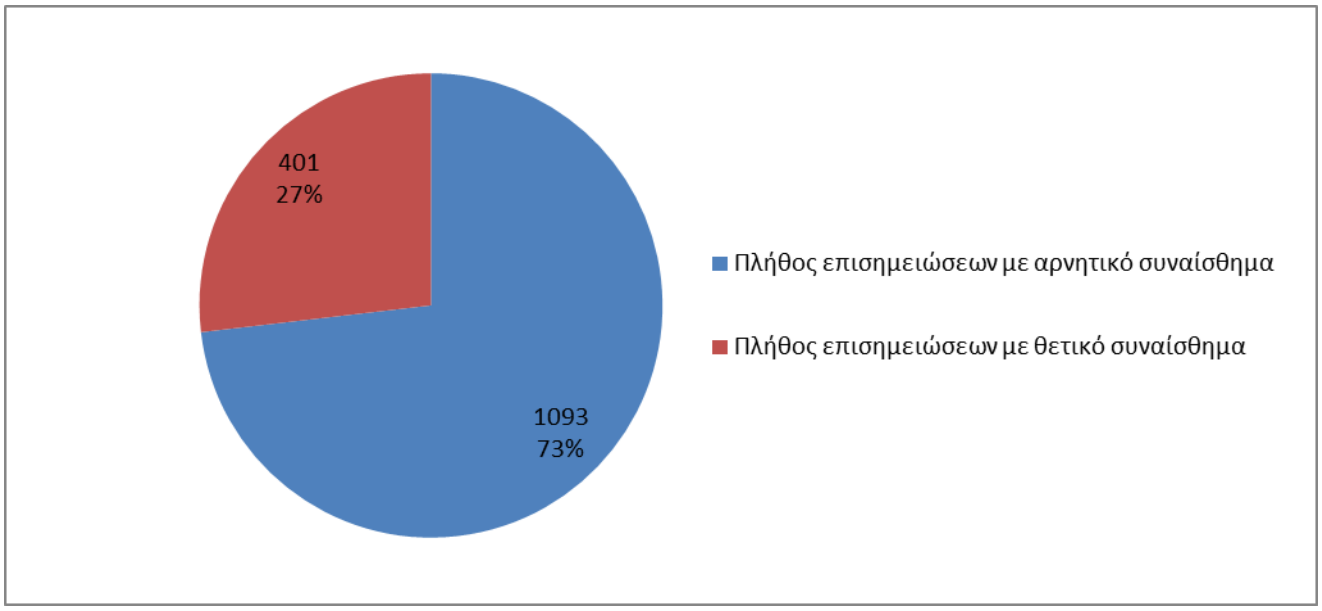
https://doi.org/10.26219/heal.aueb.4940

Στην Εικόνα 12 παρουσιάζεται η αναλογία του πλήθους των παραθύρων που παρουσιάζουν συναίσθημα με εκείνα που δεν παρουσιάζουν. Συνολικά υπάρχουν 34681 παράθυρα διάρκειας ενός δευτερολέπτου σε 137 συνομιλίες.



Εικόνα 12: Αναλογία παραθύρων με και χωρίς συναίσθημα

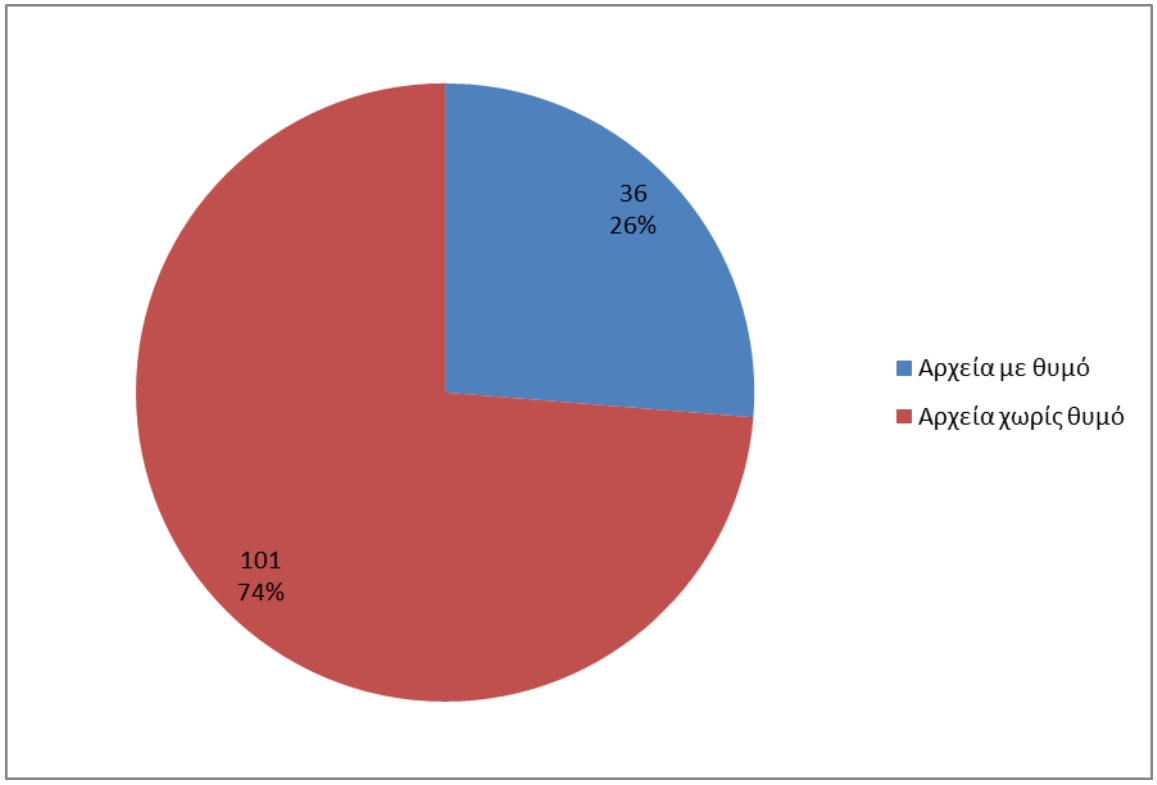
Στην Εικόνα 13 παρουσιάζεται το ποσοστό του πλήθους των επισημειωμένων τμημάτων ήχου που περιέχουν αρνητικό και θετικό συναίσθημα. Παρατηρείται ότι τα περισσότερα επισημειωμένα δεδομένα είναι αρνητικά, το οποίο αποδίδεται στο ότι οι περισσότεροι ομιλητές εκφράζουν παράπονα.



Εικόνα 13: Αναλογία αρνητικών-θετικών συναισθημάτων

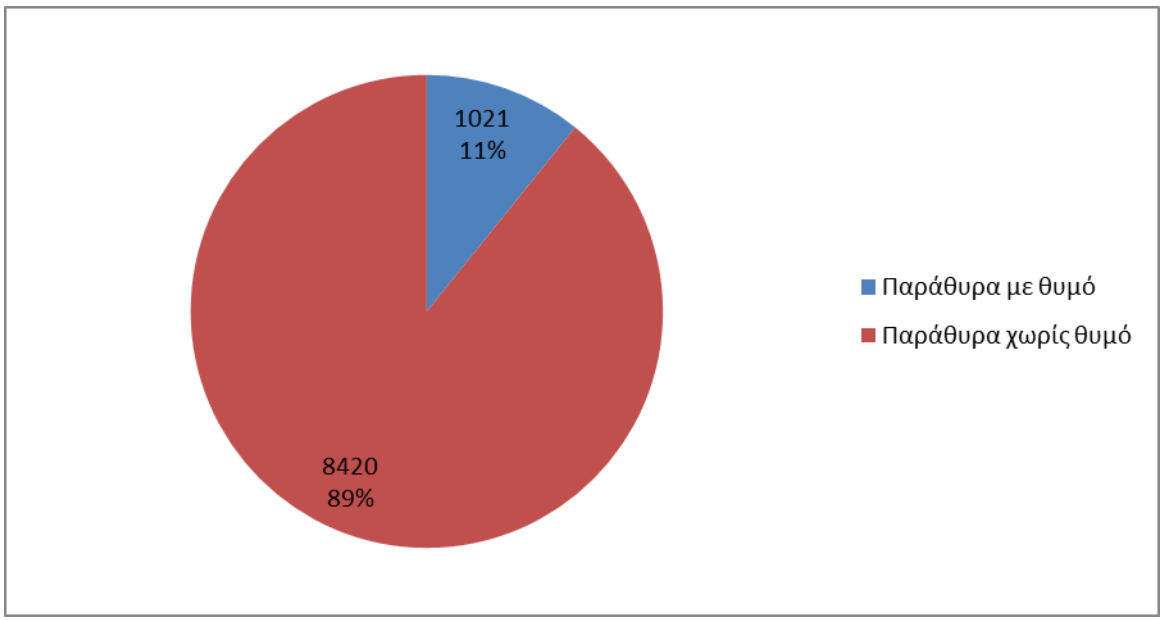
Από τα συνολικά 137 αρχεία ήχου, μόνο τα 36 αρχεία παρουσίασαν στοιχεία θυμού κατά την τελική επισημείωση, ενώ από τα 36 αρχεία μόνο το 11% της συνολικής διάρκειας παρουσιάζει θυμό. Η Εικόνα 14 δείχνει την αναλογία του πλήθους των αρχείων που παρουσιάζουν θυμό σε κάποιο σημείο με εκείνα τα οποία δεν περιέχουν καθόλου το στοιχείο του θυμού.

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 14: Πλήθος αρχείων που παρουσιάζουν θυμό

Η εικόνα 15 παρουσιάζει την αναλογία παραθύρων διάρκειας ενός δευτερολέπτου που περιέχουν το στοιχείο του θυμού και εκείνων που δεν περιέχουν το στοιχείο του θυμού. Το διάγραμμα αφορά μόνο τα αρχεία που παρουσίασαν θυμό δηλαδή το 26% των αρχείων όπως φαίνεται στην Εικόνα 14.



Εικόνα 15: Ποσοστό εμφάνισης θυμού σε παράθυρα 1s

Τα δεδομένα παρουσίασαν ανισορροπία κατηγοριών τόσο για την επισημείωση συναισθήματος όσο και για την επισημείωση του θυμού. Για αυτό το λόγο, χρειάστηκε πριν από τα πειράματα ανίχνευσης θυμού να γίνει υποδειγματοληψία (undersampling) ή υπερδειγματοληψία (oversampling) στα δεδομένα.

https://doi.org/10.26219/heal.aueb.4940

4. Συστήματα

Η διαδικασία εκπαίδευσης και ελέγχου (αξιολόγησης) του συστήματος αποτελείται από 10 βασικά βήματα:

Βήμα 1: Χειρωνακτική επισημείωση

Βήμα 2: Εξαγωγή παραθύρων ήχου

Βήμα 3: Εξαγωγή χαρακτηριστικών

Βήμα 4: Διαχωρισμός στιγμιοτύπων σε κατηγορίες

Βήμα 5: Κανονικοποίηση δεδομένων

Βήμα 6: Μείωση διαστάσεων του χώρου των δεδομένων

Βήμα 7: Υποδειγματοληψία ή υπερδειγματοληψία

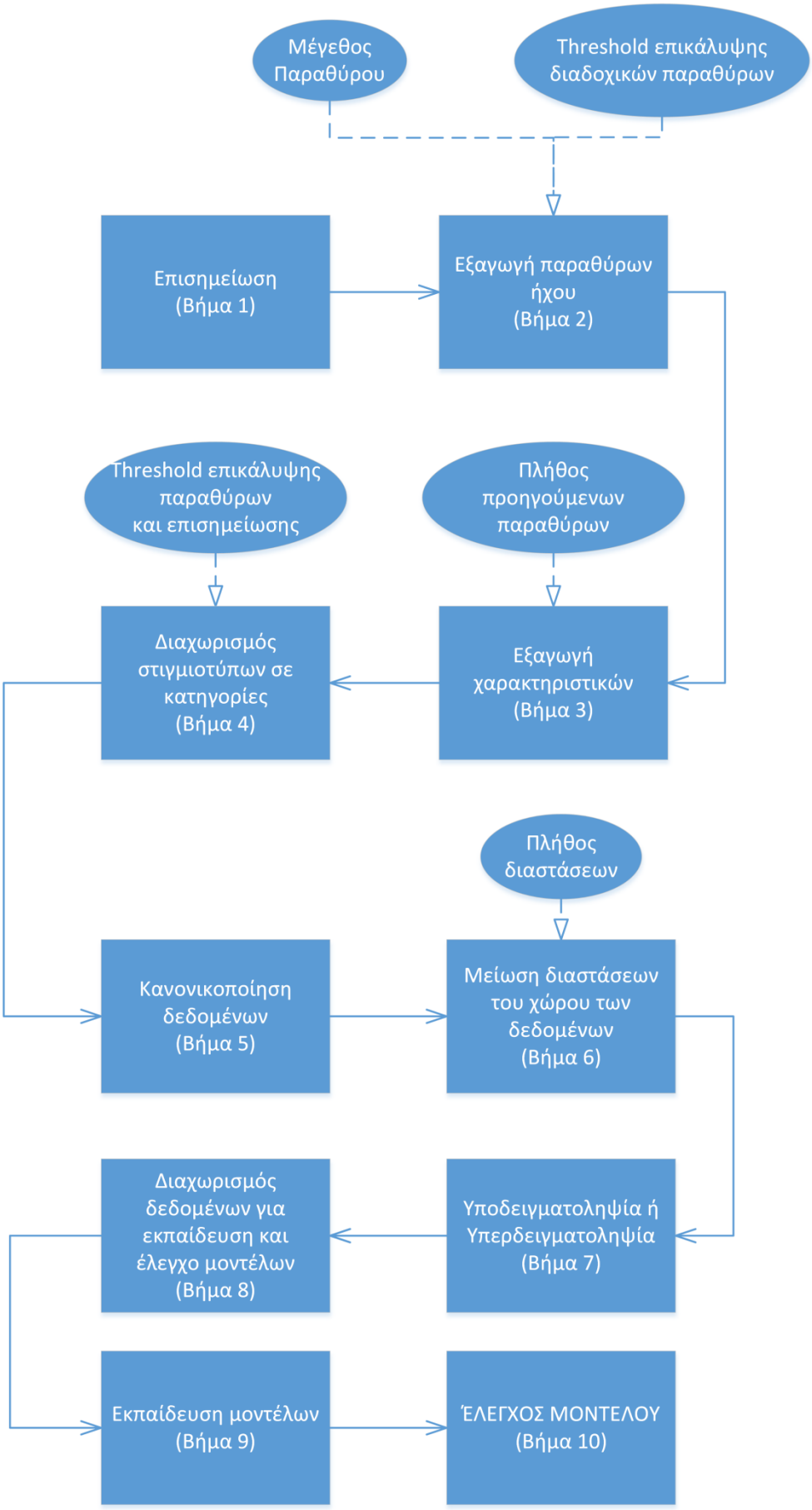
Βήμα 8: Διαχωρισμός δεδομένων για εκπαίδευση και έλεγχο μοντέλων

Βήμα 9: Εκπαίδευση μοντέλων

Βήμα 10: Έλεγχος μοντέλων

Όλα τα βήματα γίνονται αυτόματα, εκτός του πρώτου. Η διαδικασία παρουσιάζεται και στην Εικόνα 16. Οι μπλε κύκλοι είναι σημεία στα οποία παράμετροι του συστήματος πρέπει να ρυθμιστούν, ενώ τα τετράγωνα είναι υπο-διαδικασίες.

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 16: Διαδικασία εκπαίδευσης και αξιολόγησης του συστήματος.

https://doi.org/10.26219/heal.aueb.4940

Κάποια βήματα μπορούν να παραλειφθούν (π.χ. μείωση διαστάσεων). Έτσι μπορούμε να διακρίνουμε την επίδραση που έχουν στην επίδοση του συνολικού συστήματος. Επίσης, το ίδιο σύστημα μπορεί να χρησιμοποιηθεί για την ανίχνευση και άλλων συγκεκριμένων συναισθημάτων (π.χ. άγχος, χαρά, λύπη, αντί του θυμού), αρκεί να υπάρχουν κατάλληλα επισημειωμένα δεδομένα.

Όπως φαίνεται και στην Εικόνα 25 στο σύστημα υπάρχουν πέντε παράμετροι οι οποίες επεξηγούνται παρακάτω. Οι παράμετροι είναι:

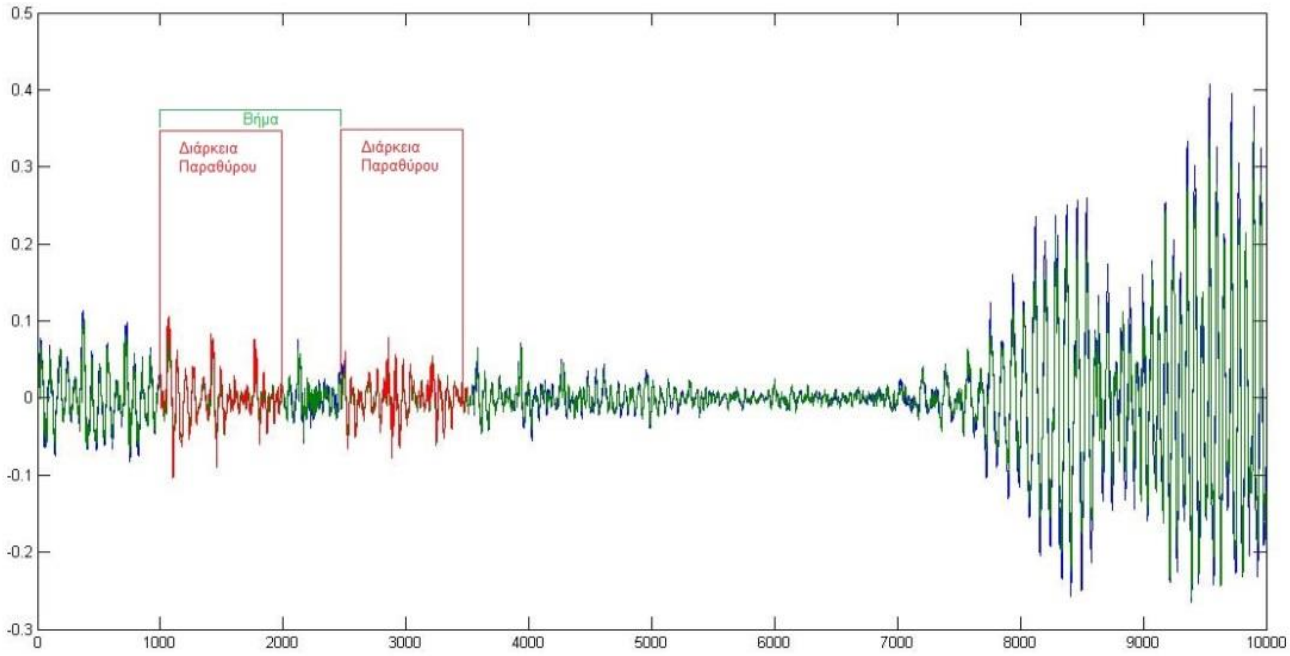
- Το χρονικό μέγεθος των παραθύρων.
- Η επικάλυψη διαδοχικών παραθύρων.
- Το ποσοστό επικάλυψης μεταξύ παραθύρου και επισημείωσης που πρέπει να ξεπεραστεί ώστε να θεωρηθεί ότι η σωστή κατηγορία του παραθύρου είναι εκείνη της επισημείωσης.
- Το πλήθος των προηγούμενων παραθύρων των οποίων τα χαρακτηριστικά προστίθενται στο διάνυσμα χαρακτηριστικών του υπό εξέταση παραθύρου.
- Το επιθυμητό πλήθος των διαστάσεων του νέου χώρου στην περίπτωση μείωσης διαστάσεων (π.χ. με PCA).

Η επισημείωση του Βήματος 1 εξηγήθηκε στο κεφάλαιο 3. Τα Βήματα 9 και 10 παρουσιάζονται στο Κεφάλαιο 5 όπου εξηγείται εκτενώς η εκπαίδευση και ο έλεγχος των μοντέλων. Στο υπόλοιπο του παρόντος κεφαλαίου εξετάζονται τα Βήματα 2 έως και 8.

4.2 Εξαγωγή παραθύρων ήχου (Βήμα 2)

Η πρώτη παράμετρος είναι το μήκος (η χρονική διάρκεια) του παραθύρου. Η δεύτερη είναι το βήμα (step), το οποίο ορίζει την απόσταση του πρώτου msec ενός παραθύρου από το πρώτο msec του επόμενου. Αυξομειώνοντας το βήμα και το μέγεθος του παραθύρου μπορεί κάποιος να ορίσει την επικάλυψη μεταξύ των διαδοχικών παραθύρων και το διάστημα μεταξύ τους. Στην Εικόνα 16 βλέπουμε δύο εντελώς τυχαία παράθυρα του 1 sec το καθένα και το βήμα που τα χωρίζει. Τα συγκεκριμένα παράθυρα έχουν 0% επικάλυψη, αφού το βήμα που τα χωρίζει είναι μεγαλύτερο από το μέγεθος των παραθύρων. Στα πειράματά μας ορίσαμε μέγεθος παραθύρου ίσο με 1 sec και βήμα ίσο με 1 sec, δηλαδή 0% επικάλυψη και καθόλου κενό μεταξύ δύο διαδοχικών παραθύρων.

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 17: Διαδοχικά παράθυρα του ηχητικού σήματος

4.3 Εξαγωγή Χαρακτηριστικών (Βήμα 3)

Η εξαγωγή των χαρακτηριστικών έγινε χρησιμοποιώντας το OpenSmile[3][44], ένα ευέλικτο και γρήγορο εργαλείο ανοικτού πηγαίου κώδικα (open source) για την εξαγωγή χαρακτηριστικών ήχου, γραμμένο αποκλειστικά σε C++, το οποίο λειτουργεί σε διάφορες πλατφόρμες, όπως Linux, Windows και MacOS.

Τα πλεονεκτήματα της χρήσης του OpenSmile είναι:

- Υποστηρίζει την ανάγνωση και την εγγραφή διαφόρων μορφών δεδομένων που συνήθως χρησιμοποιούνται στο πεδίο της εξόρυξης δεδομένων και της μηχανικής μάθησης (CSV, ARFF, κ.λπ.).
- Χρησιμοποιεί το gnuplot[4], παρέχοντας τη δυνατότητα απεικόνισης των χαρακτηριστικών που εξάγονται.
- Σχεδόν όλα τα ενδιάμεσα δεδομένα που δημιουργούνται κατά τη διαδικασία εξαγωγής των χαρακτηριστικών μπορούν να αποθηκευτούν σε αρχεία για απεικόνιση ή περαιτέρω επεξεργασία.
- Μπορεί να χρησιμοποιηθεί σε πολλά λειτουργικά συστήματα, όπως προαναφέρθηκε.

Το συγκεκριμένο εργαλείο χρησιμοποιεί ως είσοδο ένα αρχείο configuration και ένα αρχείο ήχου για να εξαγάγει από τον ήχο εκείνα τα χαρακτηριστικά που υποδεικνύει το configuration αρχείο. Χρησιμοποιήθηκαν τέσσερα configuration files για την εξαγωγή των χαρακτηριστικών.

[3] http://www.audeering.com/research/opensmile
[4] http://www.gnuplot.info/

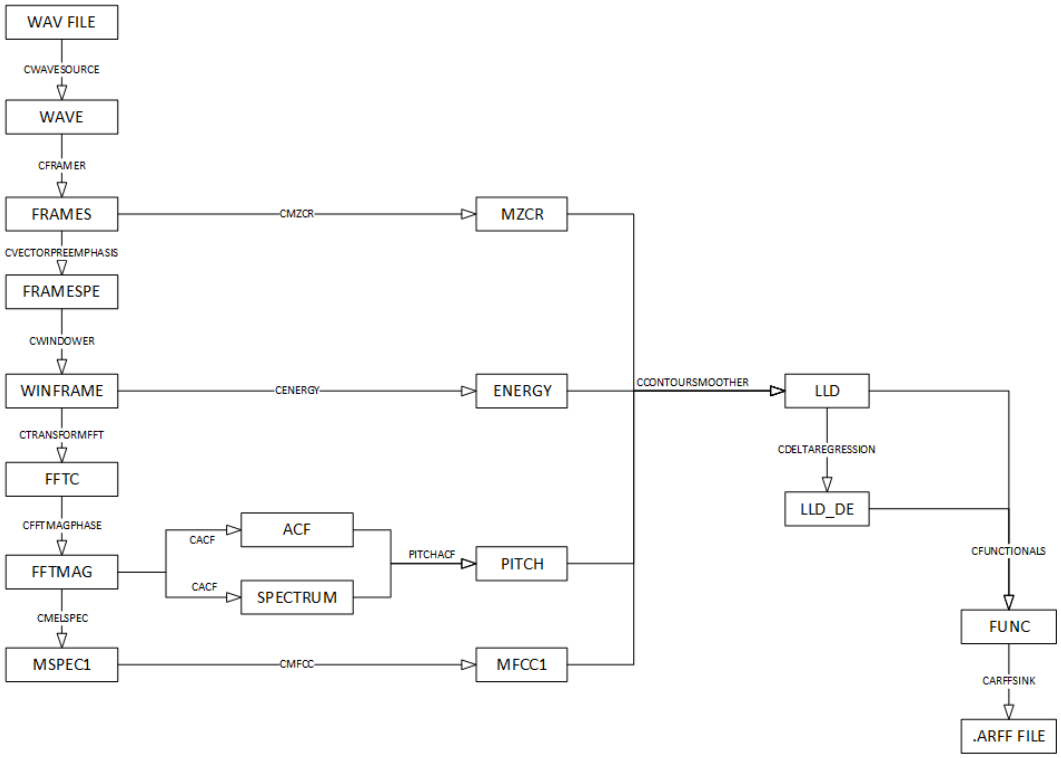
https://doi.org/10.26219/heal.aueb.4940

Τα χαρακτηριστικά που εξάγονται είναι τα χαρακτηριστικά MFCC (Mel Frequency Cepstral Coefficents ή MFCCs), που είναι καθιερωμένα στην αναγνώριση ομιλίας, το τρέμουλο (jitter), η λάμψη (shimmer), φασματικά και cepstal (αντίστροφου φάσματος) χαρακτηριστικά, η ενέργεια του σήματος, περιγραφείς χαμηλού επιπέδου (low-level descriptors ή LLDs) και συναρτησιακά χαρακτηριστικά (functionals) των παραθύρων, όπως αριθμητικός μέσος, τυπική απόκλιση κ.ά.

Στις Εικόνες 26-29 παρουσιάζονται ως παραλληλόγραμμα τα αρχεία και τα ενδιάμεσα χαρακτηριστικά που παράγονται από το OpenSmile, ενώ ως ακμές παρουσιάζονται οι διαδικασίες του προγράμματος που τις παράγουν. Σε αυτή την εργασία χρησιμοποιούνται εικοσιέξι συνολικά διαδικασίες του OpenSmile. Οι διαδικασίες αυτές περιγράφονται αναλυτικά στο βιβλίο [5].

1. **384 Χαρακτηριστικά (FS384 για χάρη συντομίας)**
   Σύνολο χαρακτηριστικών από το INTERSPEECH 2009 Emotion Challenge [7].

   Χαρακτηριστικά που εξάγονται:
   - Ενέργεια
   - Pitch
   - MFCC
   - Frame functionals



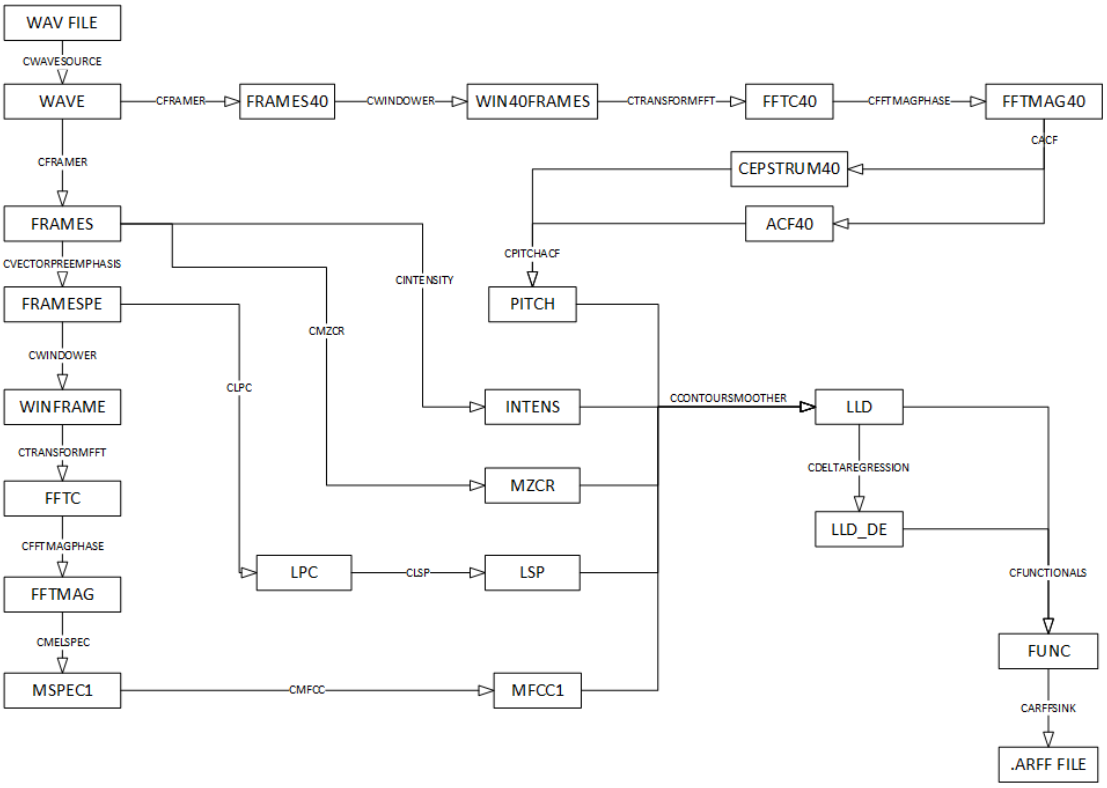
Εικόνα 18: Διαδικασία εξαγωγής πρώτου συνόλου χαρακτηριστικών

2. **988 Χαρακτηριστικά (FS988 για χάρη συντομίας)**
   Βασικό σύνολο 988 χαρακτηριστικών

   Χαρακτηριστικά που εξάγονται:
   - Ενέργεια

https://doi.org/10.26219/heal.aueb.4940

- Pitch
- MFCC
- Frame functionals
- Intensity
- LSP (Line Spectral Pairs)

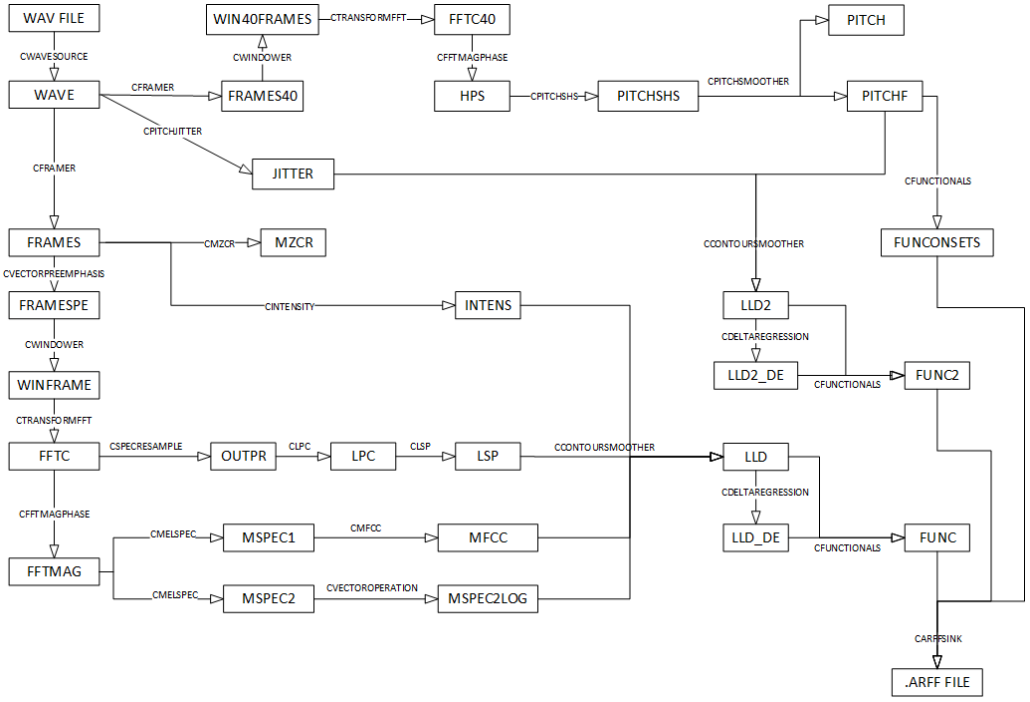


Εικόνα 19: Διαδικασία εξαγωγής δευτέρου συνόλου χαρακτηριστικών

3. **1582 Χαρακτηριστικά (FS1582 για χάρη συντομίας)**
   Βασισμένο στο INTERSPEECH 2010 paralinguistic challenge [4].

   Χαρακτηριστικά που εξάγονται:
   - Ενέργεια
   - Pitch
   - MFCC
   - Frames functionals
   - Intensity
   - LSP (Line Spectral Pairs)
   - Jitter

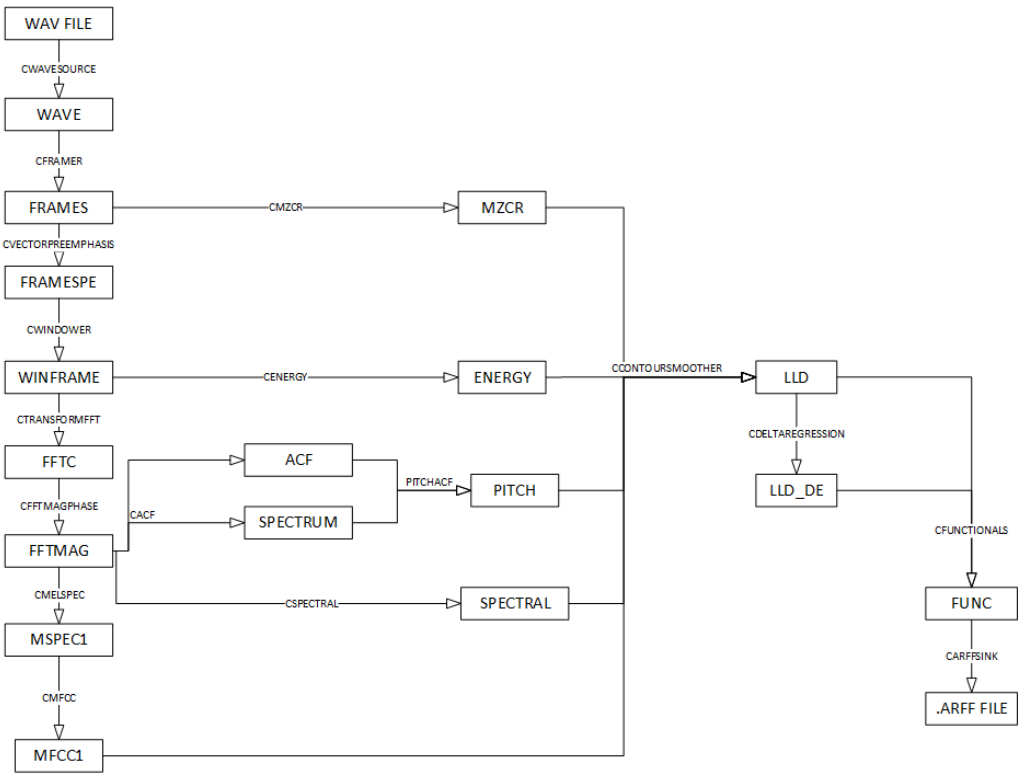
https://doi.org/10.26219/heal.aueb.4940



Εικόνα 20: Διαδικασία εξαγωγής τρίτου συνόλου χαρακτηριστικών

4. **6552 Χαρακτηριστικά (FS6552 για χάρη συντομίας)**
   Το μεγαλύτερο σύνολο από τα τέσσερα.

   Χαρακτηριστικά που εξάγονται:
   - Ενέργεια
   - Pitch
   - Frames functionals
   - Spectral



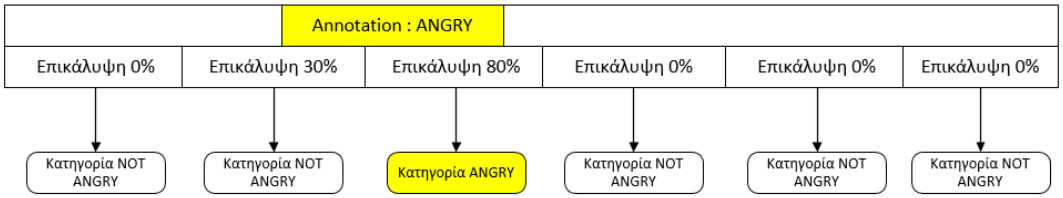
Εικόνα 21: Διαδικασία εξαγωγής τετάρτου συνόλου χαρακτηριστικών

https://doi.org/10.26219/heal.aueb.4940

4.4 Διαχωρισμός στιγμιοτύπων σε κατηγορίες (Βήμα 4)

Για τον προσδιορισμό της ορθής κατηγοριοποίησης των παραθύρων (ground truth), εάν ένα παράθυρο επικαλύπτεται από ένα χειρωνακτικά επισημειωμένο τμήμα ομιλίας που έχει (χειρωνακτική) επισημείωση Angry, τότε:

1. εάν η (χρονική) επικάλυψη είναι μεγαλύτερη ή ίση με 50%, τότε η σωστή κατηγορία του παραθύρου είναι Angry, αλλιώς
2. εάν η επικάλυψη είναι μικρότερη από 50%, τότε η σωστή κατηγορία του παραθύρου είναι Not Angry.



Εικόνα 22: Διαχωρισμός παραθύρων σε δύο κατηγορίες

Με εντελώς αντίστοιχο τρόπο καθορίζεται η σωστή κατηγορία των παραθύρων στην περίπτωση που προσπαθούμε να εντοπίσουμε συναίσθημα γενικά (πρώτο επίπεδο επισημείωσης της Εικόνας 11), όπου οι σωστές κατηγορίες είναι Emotional ή Not Emotional.

Γενικότερα, το λογισμικό της εργασίας μπορεί να ρυθμιστεί (μέσω παραμέτρου) ώστε να χρησιμοποιείται οποιοδήποτε ποσοστό επικάλυψης, αντί του 50% που αναφέρθηκε παραπάνω. Στα πειράματα της εργασίας, όμως, το ποσοστό επικάλυψης ήταν 50% και οι σωστές κατηγορίες ήταν Emotional και Not Emotional (στο πρώτο επίπεδο επισημείωσης της Εικόνας 11) ή Angry και Not Angry (στο δεύτερο επίπεδο).

Τέλος ορίστηκαν 3 ηλικιακές ομάδες (μικρότερη των 40 ετών, μεταξύ 40 και 60 ετών, μεγαλύτερη των 60 ετών). Ο συγγραφέας της εργασίας κατέταξε τους ομιλητές σε αυτές τις κατηγορίες ακούγοντας τις φωνές τους. Προέκυψαν έτσι δύο επιπλέον χαρακτηριστικά για κάθε συνομιλία: η ηλικία του υπαλλήλου και η ηλικία του πελάτη. Βέβαια σε ένα πραγματικό σύστημα αυτή η πληροφορία θα έπρεπε να εξάγεται αυτόματα. Όπως φαίνεται όμως από τα πειράματα, η ηλικία των ομιλητών δεν βελτιώνει τις επιδόσεις του συστήματος.

4.5 Κανονικοποίηση δεδομένων (Βήμα 5)

Στο στάδιο της προετοιμασίας των δεδομένων, κανονικοποιήσαμε κάθε τιμή x_i του χαρακτηριστικού X στην τιμή $x'_i = \frac{x_i - \min(X)}{\max(X) - \min(X)}$ περιορίζοντας έτσι τις τιμές όλων των χαρακτηριστικών στο διάστημα $[0,1]$ για τις τιμές του . Οι τιμές min και max εκφράζουν την ελάχιστη και τη μέγιστη αντίστοιχα τιμή για όλες τις τιμές του X στα δεδομένα εκπαίδευσης. Στη συνέχεια για κάθε τιμή x_k στα δεδομένα ελέγχου υπολογίστηκε η αντίστοιχη τιμή x'_k η οποία για τις αρνητικές τιμές στρογγυλοποιήθηκε στο μηδέν και για τις τιμές που ήταν μεγαλύτερες του ένα στρογγυλοποιήθηκε στο

https://doi.org/10.26219/heal.aueb.4940

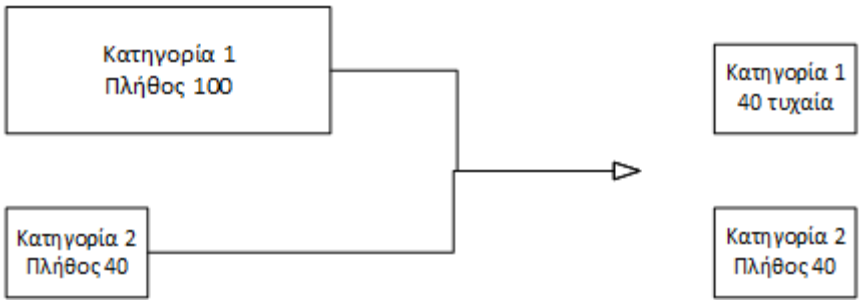
ένα. Η κανονικοποίηση χρειάζεται, γιατί τόσο η μείωση των διαστάσεων στο Βήμα 6 όσο και η εκπαίδευση μοντέλων στο Βήμα 9 χρειάζονται κανονικοποιημένα δεδομένα, ώστε να μη δίνουν μεγαλύτερη σημασία σε χαρακτηριστικά που απλά έχουν μεγαλύτερα εύρη τιμών. Στο μέλλον θα μπορούσαν να δοκιμαστούν και άλλες μέθοδοι κανονικοποίησης (π.χ., $\frac{x_i-\mu}{3\sigma}$ θεωρώντας ότι το $\mathbf{X}$ ακολουθεί κανονική κατανομή).

4.6 Μείωση του χώρου διαστάσεων (Βήμα 6)

Για τη μείωση του χώρου διαστάσεων χρησιμοποιήθηκε Ανάλυση Κυρίων Συνιστωσών (PCA, βλ. Ενότητα 2.3.1). Το πλήθος των νέων χαρακτηριστικών (του νέου χώρου) επιλέγεται όπως στην Ενότητα 2.3.1.7. Παρουσιάζεται, λοιπόν, ακόμα μια παράμετρος στο σύστημά μας, που καθορίζει το ποσοστό του εμβαδού της Εικόνας 7 (διαισθητικά το ποσοστό της πληροφορίας) που θα διατηρηθεί μετά το μετασχηματισμό PCA.

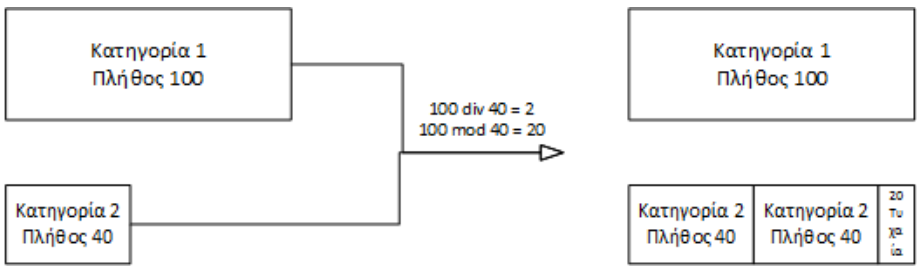
4.7 Υποδειγματοληψία ή υπερδειγματοληψία (Βήμα 7)

Για να αντιμετωπιστεί το πρόβλημα της ανισοκατανομής των στιγμιοτύπων στις κατηγορίες (Ενότητα 2.3), χρησιμοποιήθηκε τυχαία υποδειγματοληψία (RUS) ή τυχαία υπερδειγματοληψία (RUS).



Εικόνα 23: Υποδειγματοληψία

Κατά την τυχαία υπερδειγματοληψία (ROS), η κατηγορία με τα λιγότερα στιγμιότυπα πολλαπλασιάστηκε τόσες φορές (προστέθηκαν αντίγραφα των στιγμιοτύπων της) όσο είναι το πηλίκο της διαίρεσης του μεγέθους της μεγαλύτερης κατηγορίας με το μέγεθος της μικρότερης. Έπειτα επιλέχθηκαν τυχαία από την μειονεκτική κατηγορία τόσα στιγμιότυπα όσο το υπόλοιπο της προηγούμενης διαίρεσης και διπλασιάστηκαν (προστέθηκαν πάλι αντίγραφά τους).



Εικόνα 24: Υπερδειγματοληψία

https://doi.org/10.26219/heal.aueb.4940

4.8 Διαχωρισμός δεδομένων για εκπαίδευση και έλεγχο μοντέλων (Βήμα 8)

Για να εκπαιδεύσουμε και να αξιολογήσουμε το σύστημα, χωρίζουμε τα δεδομένα σε δεδομένα εκπαίδευσης (training) και δεδομένα ελέγχου (test). Ο διαχωρισμός αυτός έγινε με δύο τρόπους.

Στη πρώτη περίπτωση, χωρίζουμε τα δεδομένα ώστε να μην υπάρχουν ομιλίες του ίδιου ομιλητή στα δεδομένα εκπαίδευσης και ελέγχου (speaker-out). Παράλληλα προσπαθούμε να επιτύχουμε το 80% των στιγμιοτύπων (παραθύρων) να χρησιμοποιούνται για εκπαίδευση και το 20% για έλεγχο.

Εάν ορίσουμε ως M το συνολικό πλήθος των διαθεσίμων στιγμιοτύπων, ως επιθυμητό πλήθος των στιγμιοτύπων εκπαίδευσης το $T = L \cdot M$, $L \in (0,1)$, και ως O_i τα αρχεία του ομιλητή i, τότε όσο $T > 0$, βρίσκουμε τον ομιλητή που ελαχιστοποιεί το $||O_i| - T|$ και προσθέτουμε τα αρχεία του στα δεδομένα εκπαίδευσης. Έπειτα αφαιρούμε το μέγεθος $|O_i|$ από το επιθυμητό μέγεθος T. Όσες ομιλίες απομείνουν χρησιμοποιούνται ως δεδομένα ελέγχου.

Για παράδειγμα, εάν έχουμε τα παρακάτω στοιχεία για τους ομιλητές

Ομιλητής	Ομ.1	Ομ.2	Ομ.3	Ομ.4	Ομ.5	Ομ.6	Ομ.7	Ομ.8	Ομ.9	Ομ.10
Πλήθος Ομιλιών	2	14	25	2	3	12	7	1	2	9

Ταξινομούμε σε φθίνουσα σειρά τον πίνακα

Ομιλητής	Ομ.3	Ομ.2	Ομ.6	Ομ.10	Ομ.7	Ομ.5	Ομ.1	Ομ.4	Ομ.9	Ομ.7
Πλήθος Ομιλιών	25	14	12	9	7	3	2	2	2	1

Βρίσκουμε το 80% του συνολικού πλήθους το οποίο είναι και το μέγεθος των δεδομένων εκπαίδευσης:
$77 \cdot 0{,}8 = 61{,}6$

Βήμα 1:

Θέτω $T = 61{,}6$

Ομιλητές στα δεδομένα εκπαίδευσης: []

$$\min(|O_i - T|) = 36{,}6 \; και \; i = 3$$

Βήμα 2:

Θέτω $T = 61{,}6 - 25 = 36{,}6$

Ομιλητές στα δεδομένα εκπαίδευσης: [3]

$\min(|O_i - T|) = 22{,}6 \; και \; i = 2$

Βήμα 3:

https://doi.org/10.26219/heal.aueb.4940

Θέτω $T = 36{,}6 - 14 = 22{,}6$

Ομιλητές στα δεδομένα εκπαίδευσης: [3,2]

$\min(|O_i - T|) = 10{,}6 \; και \; i = 6$

Βήμα 4:

Θέτω $T = 22{,}6 - 12 = 10{,}6$

Ομιλητές στα δεδομένα εκπαίδευσης: [3,2,6]

$\min(|O_i - T|) = 1{,}6 \; και \; i = 10$

Βήμα 5:

Θέτω $T = 10{,}6 - 9 = 1{,}6$

Ομιλητές στα δεδομένα εκπαίδευσης : [3,2,6,10]

$\min(|O_i - T|) = -0{,}4 \; και \; i = 1$

Βήμα 6:

Θέτω $T = 1{,}6 - 2 = -0{,}4$

Ομιλητές στα δεδομένα εκπαίδευσης: [3,2,6,10,1]

$T = -0{,}4$ Έξοδος

Άρα στα δεδομένα εκπαίδευσης θα τοποθετηθούν οι ομιλητές 1, 2, 3, 6, 10 ενώ στα δεδομένα ελέγχου οι ομιλητές 4, 5, 7, 8, 9, δηλαδή τα δεδομένα εκπαίδευσης θα έχουν 62 αρχεία (80,52% του συνολικού πλήθους των αρχείων) ενώ τα δεδομένα ελέγχου 15 αρχεία (19,48% του συνολικού πλήθους των αρχείων). Χάριν συντομίας θα αναφερόμαστε στη παραπάνω μέθοδο ως LSSO (Leave Some Speakers Out).

Ο δεύτερος τρόπος που χρησιμοποιείται για να διαχωρίσουμε τα δεδομένα εκπαίδευσης από τα δεδομένα ελέγχου είναι μια μορφή διασταυρωμένης επικύρωσης (cross-validation). Επαναλαμβάνει το πείραμα τόσες φορές όσες είναι οι ομιλητές. Σε κάθε επανάληψη, ξεχωρίζει τα αρχεία με τέτοιο τρόπο ώστε τα αρχεία μόνο ενός ομιλητή (διαφορετικού σε κάθε επανάληψη) να χρησιμοποιούνται ως δεδομένα ελέγχου και τα δεδομένα όλων των υπολοίπων ως δεδομένα εκπαίδευσης. Με αυτό τον τρόπο όλα τα δεδομένα χρησιμοποιούνται τελικά για έλεγχο. Χάριν συντομίας θα αναφερόμαστε στη μέθοδο αυτή ως LOSO (Leave One Speaker Out).

Για την αξιολόγηση χρησιμοποιήθηκαν οι μέθοδοι micro-averaging και macro-averaging. Εάν ορίσουμε ως N τον αριθμό των διαφορετικών ομιλητών και k_i τον αριθμό των στιγμιοτύπων ελέγχου του ομιλητή i τότε :

$$MacroAveraged\ Accuracy = \frac{1}{N} * \sum_{i=1}^{N} Acc_i$$

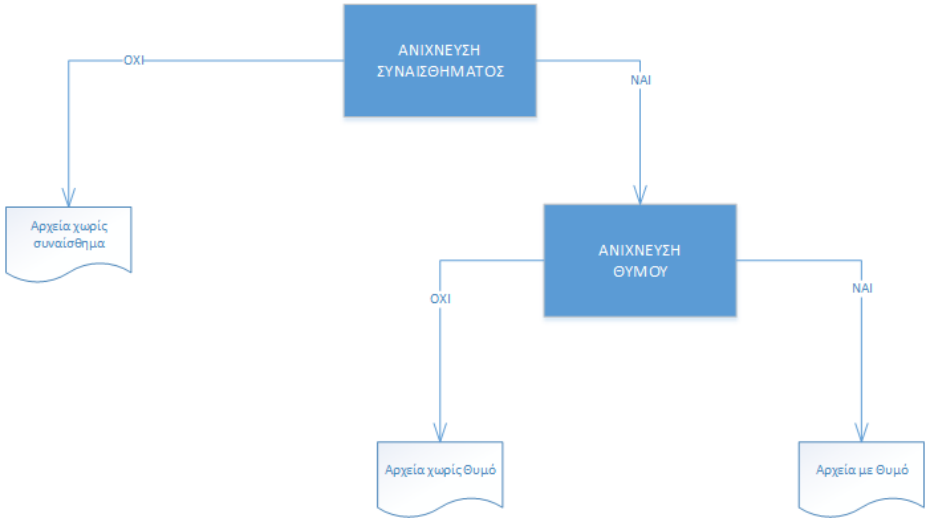
https://doi.org/10.26219/heal.aueb.4940

$$MicroAveraged\ Accuracy\ = \frac{1}{\sum_{i=1}^{N} k_i} * \sum_{i=1}^{N} Acc_i * k_i$$

όπου Acc_i είναι το ποσοστό ορθότητας που λαμβάνουμε χρησιμοποιώντας ως δεδομένα ελέγχου τα στιγμιότυπα του ομιλητή i.

Ομοίως μπορούμε να υπολογίσουμε $MacroAveraged\ Precision$ και $MacroAveraged\ Recall$.

Στόχος είναι να δημιουργηθούν δύο συστήματα τα οποία θα λειτουργούν είτε ξεχωριστά είτε σχηματίζοντας δέντρο αποφάσεων (Εικόνα 25). Το πρώτο σύστημα θα ξεχωρίζει τα παράθυρα ήχου τα οποία περιέχουν συναίσθημα και το δεύτερο θα ελέγχει τα παράθυρα που το πρώτο σύστημα θεώρησε ότι περιέχουν συναίσθημα και θα τα ταξινομεί σε εκείνα που έχουν θυμό και εκείνα που δεν έχουν.



Εικόνα 25: Δενδρικός συνδυασμός ταξινομητών

https://doi.org/10.26219/heal.aueb.4940

5. Πειράματα

5.1 Μέτρα επίδοσης

Για κάθε ένα υπό εξέταση στιγμιότυπο x_i και κατηγορία K_j, οι αλγόριθμοι που χρησιμοποιούμε επιστρέφουν μία πιθανότητα να ανήκει το x_i στην K_j. Ορίζοντας ένα κατώφλι στην πιθανότητα αυτή (προεπιλογή 0.5) μπορούμε να θεωρήσουμε ότι εάν η πιθανότητα ξεπερνάει το κατώφλι, τότε πράγματι το x_i ανήκει στην K_j. Με αυτό τον τρόπο εξάγονται πίνακες σύγχυσης (confusion matrices), οι οποίοι μας δείχνουν το πλήθος των στιγμιοτύπων που κατατάσσονται σωστά ή λανθασμένα στην κάθε κατηγορία.

Εάν εξετάσουμε μια κατηγορία (π.χ. Angry), μπορούμε να ορίσουμε ως TP (true positives) τα στιγμιότυπα που το σύστημα ταξινόμησε ορθώς ως στιγμιότυπα αυτής της κατηγορίας, TN (true negatives) τα στιγμιότυπα που το σύστημα ταξινόμησε ορθώς ως στιγμιότυπα (οποιασδήποτε) διαφορετικής κατηγορίας, FP (false positives) τα στιγμιότυπα που το σύστημα ταξινόμησε λανθασμένα ως στιγμιότυπα αυτής της κατηγορίας, και FN (false negatives) τα στιγμιότυπα που το σύστημα ταξινόμησε λανθασμένα ως στιγμιότυπα (οποιασδήποτε) διαφορετικής κατηγορίας.

Μπορούμε πλέον να ορίσουμε τα μέτρα επίδοσης: ορθότητα (accuracy), ακρίβεια (precision), ανάκληση (recall). Για κάθε κατηγορία, τα μέτρα αυτά ορίζονται ως εξής:

$$\text{Accuracy: } \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{Precision: } \frac{TP}{TP+FP}$$

$$\text{Recall: } \frac{TP}{TP+FN}$$

Η ορθότητα (accuracy) είναι ο λόγος των σωστών αποφάσεων δια των συνολικών αποφάσεων κατάταξης στιγμιοτύπων. Η ανάκληση (recall) μας δείχνει πόσα από τα στιγμιότυπα που ο αλγόριθμος έπρεπε να κατατάξει στην υπό εξέταση κατηγορία τα τοποθέτησε πραγματικά στην υπό εξέταση κατηγορία. Η ακρίβεια (precision) μας δείχνει πόσα από τα στιγμιότυπα που ο αλγόριθμος κατέταξε στην υπό εξέταση κατηγορία ανήκουν πραγματικά στην υπό εξέταση κατηγορία.

Για διαφορετικές τιμές του κατωφλίου πιθανότητας, παίρνουμε (για τον ίδιο αλγόριθμο και την ίδια κατηγορία) διαφορετικά ζεύγη τιμών ανάκλησης και ακρίβειας. Απεικονίζοντας αυτά τα ζεύγη σε ένα διάγραμμα με οριζόντιο άξονα την ανάκληση (recall) και κατακόρυφο την ακρίβεια (precision), λαμβάνουμε ένα διάγραμμα ανάκλησης-ακρίβειας (βλ. π.χ. Εικόνα 38 παρακάτω). Όσο πιο πάνω βρίσκεται η καμπύλη μιας μεθόδου, τόσο καλύτερη είναι η μέθοδος.

https://doi.org/10.26219/heal.aueb.4940

5.2 Αποτελέσματα

Στις στήλες των παρακάτω πινάκων παρουσιάζονται συνοπτικά τα καλύτερα αποτελέσματα για τις επιλογές που αναφέρονται. Τα αποτελέσματα είναι αξιόλογα, δεδομένου ότι πρόκειται για πειράματα που εκτελέστηκαν σε πραγματικά δεδομένα. Συνολικά για την αναγνώριση συναισθήματος (γενικά, όχι μόνο θυμού) εξετάστηκαν 96 παραλλαγές του συστήματος. Για την αναγνώριση θυμού εξετάστηκαν 288 παραλλαγές, συνυπολογίζοντας και τα συστήματα που χρησιμοποιούν τα χαρακτηριστικά για το φύλο και την ηλικία. Από όλα αυτά τα συστήματα παρουσιάζονται τα 15 με τα καλύτερα αποτελέσματα. Ως «context» εννοείται η επιπλέον χρήση των χαρακτηριστικών του προηγούμενου παραθύρου. Με έντονα μαύρα γράμματα φαίνονται τα καλύτερα συστήματα που προκύπτουν μελετώντας τις καμπύλες μάθησης και τις καμπύλες ακρίβειας-ανάκλησης. Για την αναγνώριση του θυμού επιλέχθηκε ως καλύτερο το σύστημα που χρησιμοποιεί τεχνική LOSO γιατί σύμφωνα με αυτή τη τεχνική ελέγχονται όλοι οι ομιλητές ξεχωριστά, δίνοντας με αυτό τον τρόπο μια καλή εκτίμηση για την ορθότητα των συστημάτων.

Υπερδειγματοληψία / Υποδειγματοληψία	Χαρακτη-ριστικά	Διαχωρισμός Train/test	Μείωση Διαστάσεων	Αλγόριθμος Μάθησης	Context	Ορθότητα
RUS	FS384	LSSO	-	SVM	Όχι	64,20%
RUS	FS384	LSSO	PCA	SVM	Όχι	65,60%
RUS	**FS384**	**LSSO**	**PCA**	**SVM**	**Ναι**	**68,00%**
RUS	FS1582	LSSO	-	LR	Όχι	64,40%
RUS	FS1582	LSSO	PCA	LR	Όχι	65,20%

Πίνακας 1: Αποτελέσματα για την αναγνώριση συναισθήματος

Υπερδειγματοληψία / Υποδειγματοληψία	Χαρακτη-ριστικά	Διαχωρισμός Train/test	Μείωση Διαστάσεων	Αλγόριθμος Μάθησης	Context	Ορθότητα
RUS	FS384	LSSO	-	LR	Όχι	79,63%
RUS	FS384	LSSO	-	SVM	Όχι	79,63%
ROS	FS384	LSSO	-	LR	Όχι	82,74%
ROS	FS384	LSSO	-	SVM	Όχι	83,04%
RUS	**FS384**	**LOSO**	**PCA**	**LR**	**Όχι**	**70,31%**
RUS	FS1582	LOSO	PCA	SVM	Όχι	68,47%
RUS	FS6552	LOSO	PCA	LR	Όχι	71,00%

Πίνακας 2: Αποτελέσματα για την αναγνώριση θυμού

Υπερδειγματοληψία / Υποδειγματοληψία	Χαρακτη-ριστικά	Διαχωρισμός Train/test	Μείωση Διαστάσεων	Αλγόριθμος Μάθησης	Context	Επιπλέον χαρακτηριστικά	Ορθότητα

https://doi.org/10.26219/heal.aueb.4940

RUS	FS384	LSSO	PCA	LR	Όχι	Όχι	75,93%
RUS	**FS384**	**LSSO**	**PCA**	**LR**	**Όχι**	**Φύλο**	**75,93%**
RUS	FS384	LSSO	PCA	LR	Όχι	Φύλο και ηλ.ομάδα	62,96%

Πίνακας 3: Αποτελέσματα για την αναγνώριση θυμού με προσθήκη χαρακτηριστικών φύλου και ηλ.ομάδας

Το PCA σε γενικές γραμμές κάνει το τελικό (εκπαιδευμένο) σύστημα γρηγορότερο και μειώνει τις απαιτήσεις μνήμης. Επίσης βελτιώνει σε μερικές τουλάχιστον περιπτώσεις (αλλά όχι πάντα) τις επιδόσεις του συστήματος (ορθότητα, ανάκληση, ακρίβεια, βλ. π.χ. Πίνακα 1). Το σύνολο χαρακτηριστικών FS384 προσφέρει τη μεγαλύτερη ορθότητα. Τέλος, όπως φαίνεται στον Πίνακα 1, η προσθήκη χαρακτηριστικών από τα προηγούμενα παράθυρα ήχου βελτιώνει τα αποτελέσματα της αναγνώρισης συναισθήματος αλλά όχι την αναγνώριση θυμού.

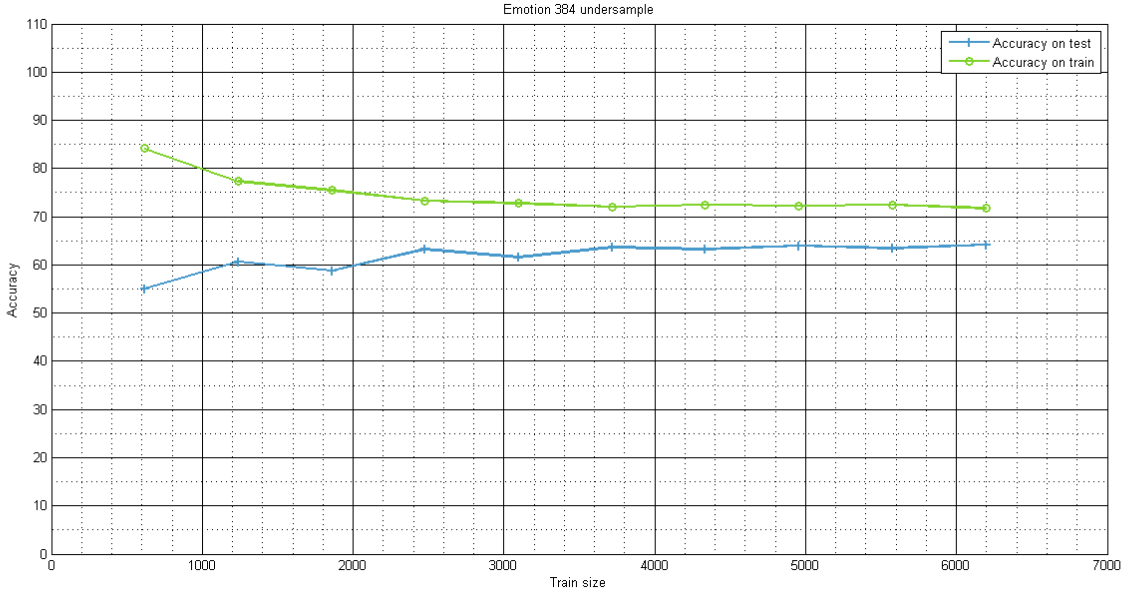
Τα (σημαντικότερα) πειράματα που εκτελέστηκαν στη διάρκεια της εργασίας περιγράφονται αναλυτικότερα στις επόμενες σελίδες. Αρχικά παρουσιάζονται τα αποτελέσματα των τριών καλύτερων συστημάτων για την αναγνώριση συναισθήματος. Στη συνέχεια παρουσιάζονται τα έξι καλύτερα συστήματα για αναγνώριση θυμού (τρία για τη μέθοδο LSSO και τρία για τη μέθοδο LOSO). Τέλος παρουσιάζονται τρία συστήματα τα οποία δείχνουν τη συμβολή των χαρακτηριστικών του φύλου και της ηλικιακής ομάδας.

Σε κάθε πείραμα εφαρμόζουμε και στα δεδομένα εκπαίδευσης και στα δεδομένα αξιολόγησης τυχαία υποδειγματοληψία ή υπερδειγματοληψία (RUS ή ROS). Ως σύστημα αφετηρίας (baseline), κατατάσσουμε τα στιγμιότυπα σε μία (τυχαία, πάντα την ίδια) κατηγορία. Η ορθότητα ενός τέτοιου συστήματος είναι πάντα 50%. Επειδή θα ήταν αδύνατο να περιγραφούν όλα τα διαφορετικά συστήματα που δημιουργήθηκαν στο πλαίσιο της εργασίας, θα αναφερθούν εκείνα που παρουσιάζουν τα πιο αξιόλογα αποτελέσματα.

5.2.1 Αποτελέσματα στην αναγνώριση συναισθήματος

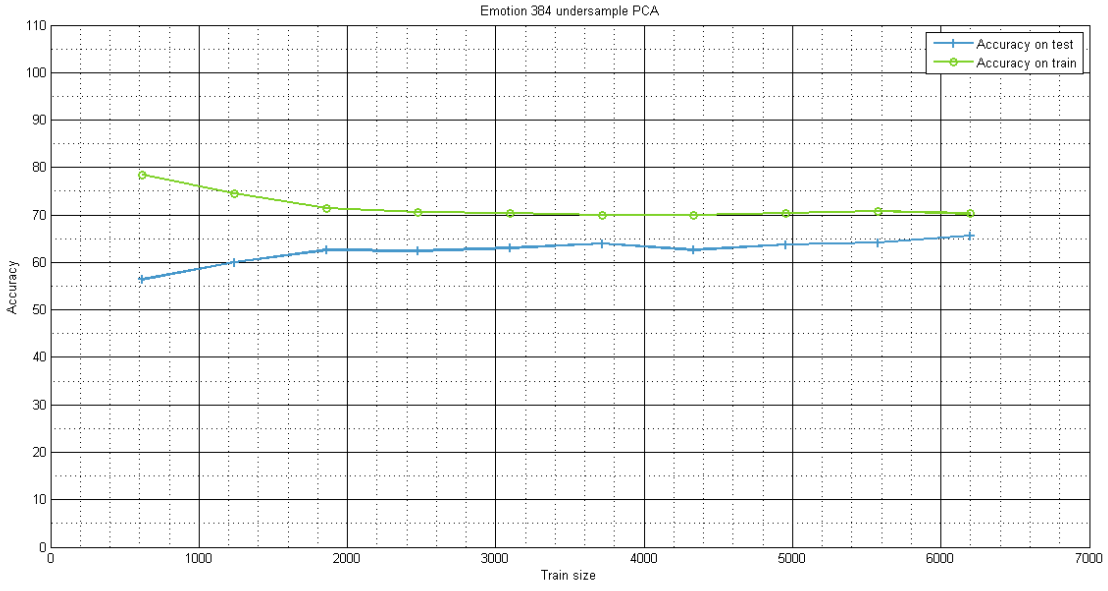
Στο αρχικό μας πείραμα (Σύστημα 1, Εικόνα 26) προσπαθούμε να ξεχωρίσουμε τα κομμάτια (παράθυρα) ήχου που περιέχουν συναίσθημα από εκείνα που δεν περιέχουν. Χρησιμοποιήθηκαν τα χαρακτηριστικά FS384 (όπως περιγράφονται στην Ενότητα 4.3) και SVM με γραμμικό πυρήνα. Στα δεδομένα εφαρμόστηκε RUS.

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 26: Καμπύλη μάθησης Συστήματος 1 (SVM, 384 χαρακτηριστικά, υποδειγματοληψία, LSSO)

Το σύστημα (Σύστημα 1, Εικόνα 26) φαίνεται στη καλύτερη των περιπτώσεων να επιτυγχάνει 71,58 % ορθότητα στα δεδομένα εκπαίδευσης. Η καμπύλη της ορθότητας στα δεδομένα εκπαίδευσης μπορεί να θεωρηθεί άνω φράγμα για την καμπύλη της ορθότητας στα δεδομένα ελέγχου, αφού ένας αλγόριθμος μάθησης συνήθως έχει καλύτερες επιδόσεις στα δεδομένα εκπαίδευσης και χειρότερες στα δεδομένα ελέγχου. Προσθέτοντας στιγμιότυπα εκπαίδευσης το σύστημα δείχνει να μαθαίνει, επιτυγχάνοντας τελικά ορθότητα 64,00% στα δεδομένα ελέγχου.



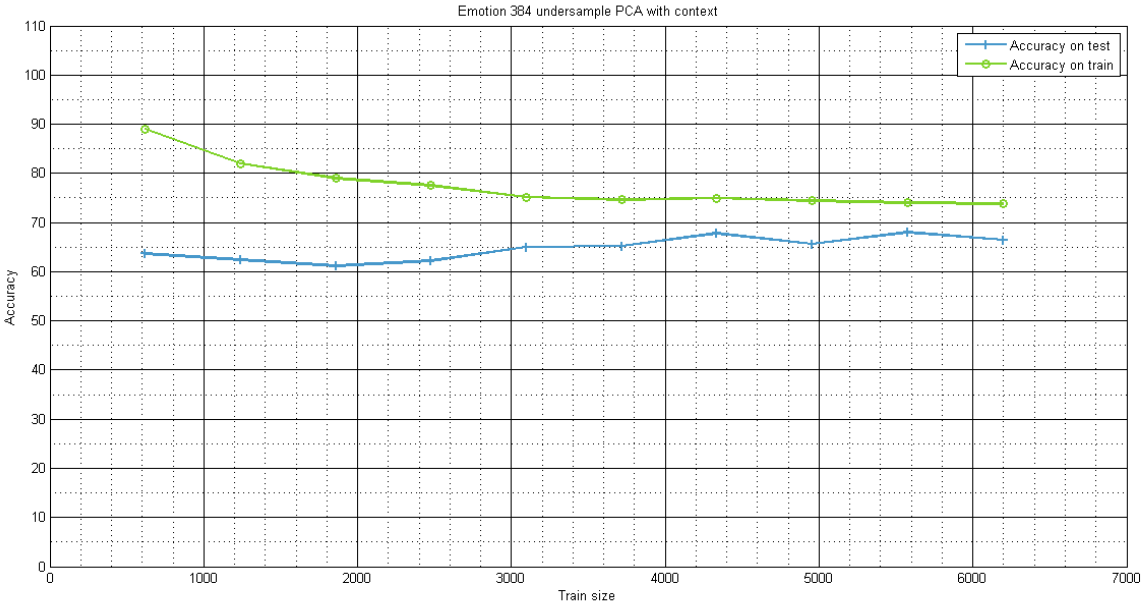
Εικόνα 27: Καμπύλη μάθησης Συστήματος 2 (SVM, 384 χαρακτηριστικά, υποδειγματοληψία, PCA, LSSO)

Στη συνέχεια, ελέγξαμε το ίδιο σύστημα με χρήση PCA (Σύστημα 2, Εικόνα 27). Η παράμετρος του PCA ήταν 0.01, ώστε να αποφασίσει ο αλγόριθμος πόσα χαρακτηριστικά να κρατήσει χάνοντας 1% πληροφορία (βλ. Ενότητα 2.2). Το PCA επέστρεψε 203 νέα χαρακτηριστικά που κατασκευάστηκαν από τα αρχικά 384. Παρατηρείται ότι στο Σύστημα 2 οι δύο καμπύλες συγκλίνουν ελαφρά πιο γρήγορα (μικρότερη υπερ-εφαρμογή, over-fitting) και υπάρχει μία μικρή βελτίωση στην τελική ορθότητα σε

https://doi.org/10.26219/heal.aueb.4940

δεδομένα ελέγχου (64,60%, αντί 64,00% στην Εικόνα 26). Όπως φαίνεται 3800 στιγμιότυπα εκπαίδευσης αρκούν για να φτάσει το σύστημα σε ένα σημείο κορεσμού, όπου προσθέτοντας νέα στιγμιότυπα εκπαίδευσης το σύστημα δεν βελτιώνεται. Το άνω φράγμα της ορθότητας σε δεδομένα ελέγχου (δηλαδή η ορθότητα στα δεδομένα εκπαίδευσης) είναι 70,43 %.

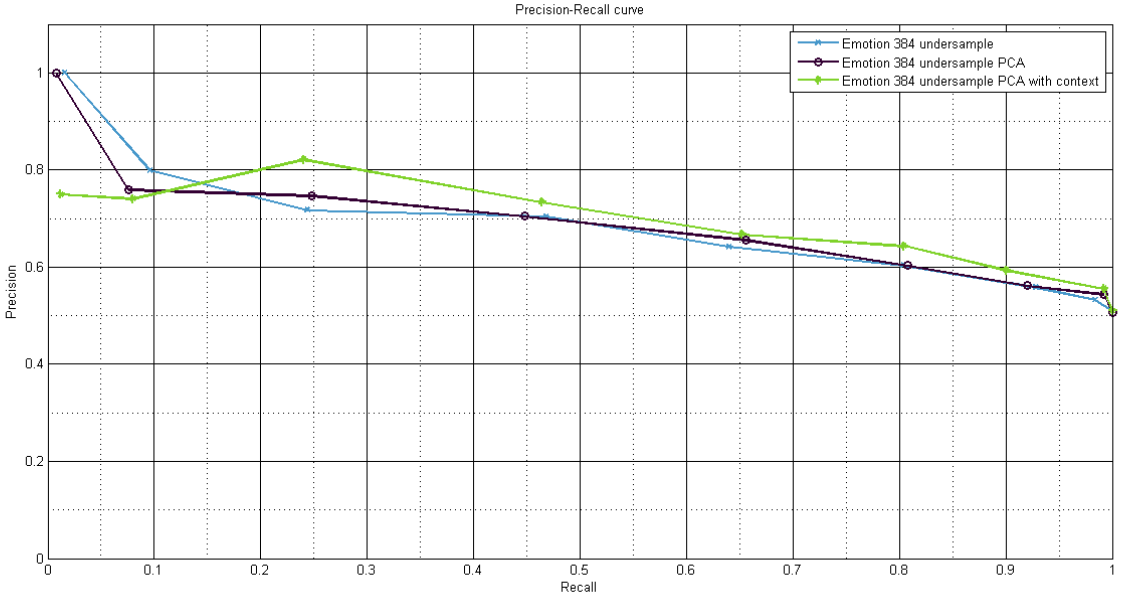
Στη συνέχεια (Σύστημα 3), προσθέσαμε σε κάθε στιγμιότυπο και τα χαρακτηριστικά του προηγούμενου στιγμιοτύπου (δηλ. του προηγούμενου χρονικά παραθύρου ήχου) προσπαθώντας με αυτό τον τρόπο να δώσουμε στον αλγόριθμο γνώση και για τη προηγούμενη κατάσταση. Με χρήση PCA (και απώλεια πληροφορίας πάλι 1%) δημιουργήθηκαν 403 χαρακτηριστικά αντί των αρχικών 768. Στα δεδομένα εφαρμόστηκε RUS και χρησιμοποιήθηκε SVM με γραμμικό πυρήνα για τη κατηγοριοποίηση. Στο Σύστημα 3 (Εικόνα 28) οι δύο καμπύλες συγκλίνουν πιο αργά από τα προηγούμενα δύο συστήματα, κάτι που είναι αναμενόμενο αφού υπάρχουν περισσότερα χαρακτηριστικά (και άρα μεγαλύτερος κίνδυνος υπερ-εφαρμογής). Η ορθότητα σε στιγμιότυπα ελέγχου είναι ελαφρά μεγαλύτερη (μέγιστη τιμή 68,00%) και το άνω φράγμα της είναι περίπου 74,03%.



Εικόνα 28: Καμπύλη μάθησης Συστήματος 3 (SVM, 384 χαρακτηριστικά, υποδειγματοληψία, PCA, και context, LSSO)

Στην Εικόνα 29 παρουσιάζονται οι καμπύλες ακρίβειας-ανάκλησης των τριών προηγούμενων συστημάτων. Παρατηρείται ότι το Σύστημα 3 είναι γενικά καλύτερο. Η χρήση PCA δεν έχει σημαντική επίπτωση. Σημειώνεται όμως ότι η χρήση PCA μειώνει το πλήθος των χαρακτηριστικών αρκετά, επιτρέποντας την πιο γρήγορη ταξινόμηση νέων στιγμιοτύπων.

https://doi.org/10.26219/heal.aueb.4940

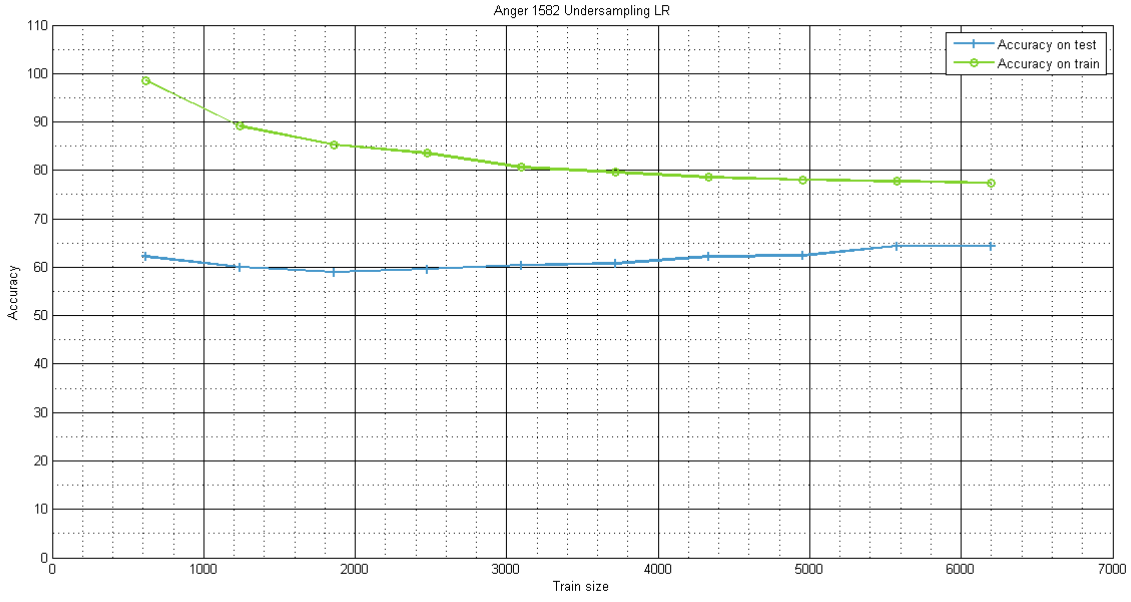


Εικόνα 29: Καμπύλη Ακρίβειας-Ανάκλησης Συστημάτων1-3

Στο παραπάνω σχήμα φαίνεται ότι μια καλή επιλογή για την αναγνώριση συναισθήματος είναι το Σύστημα 3, θεωρώντας ως συναισθηματικά τα στιγμιότυπα που κατατάσσονται σε αυτή την κατηγορία με πιθανότητα μεγαλύτερη του 0,4 καθώς σε αυτό το σημείο παρατηρείται ανάκληση 80,40% και ακρίβεια 64,42%.

Στη συνέχεια εξετάστηκαν τα χαρακτηριστικά FS1582 (Σύστημα 4). Στα πειράματα με αυτά τα χαρακτηριστικά ο αλγόριθμος της Λογιστικής Παλινδρόμησης παρήγαγε καλύτερα αποτελέσματα από τον αλγόριθμο SVM. Επομένως παρουσιάζονται μόνο τα συγκεκριμένα αποτελέσματα.

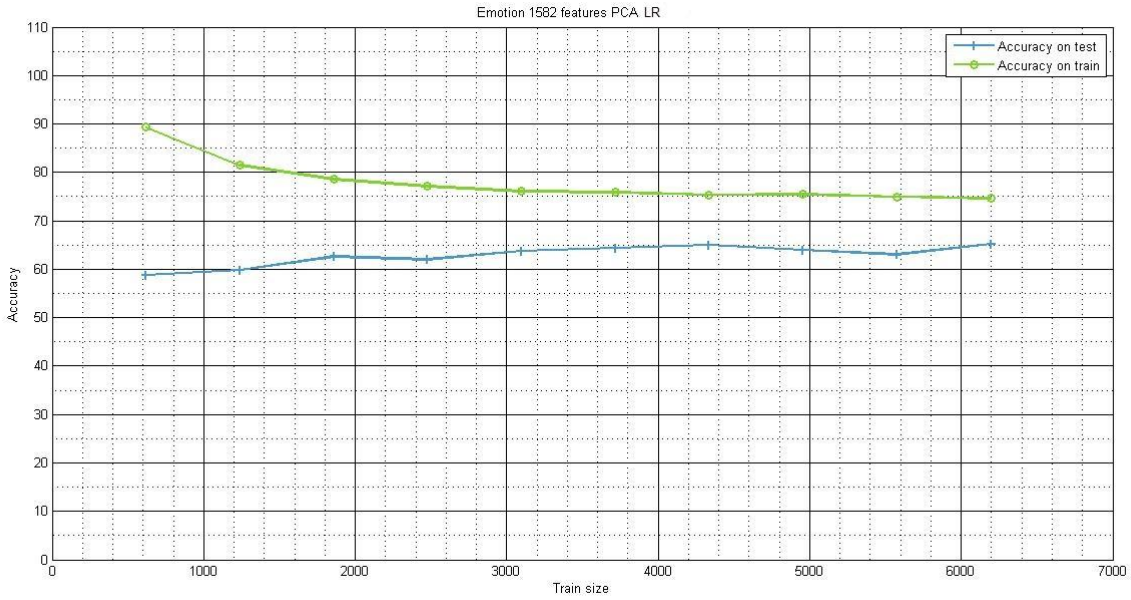
Αρχικά εξετάζεται το σύστημα που χρησιμοποιεί Λογιστική Παλινδρόμηση (Logistic Regression, LR), αφού αρχικά έχει γίνει υποδειγματοληψία RUS.



Εικόνα 30: Καμπύλη μάθησης Συστήματος 4 (Λογιστική Παλινδρόμηση, 1582 χαρακτηριστικά, Υποδειγματοληψία , LSSO)

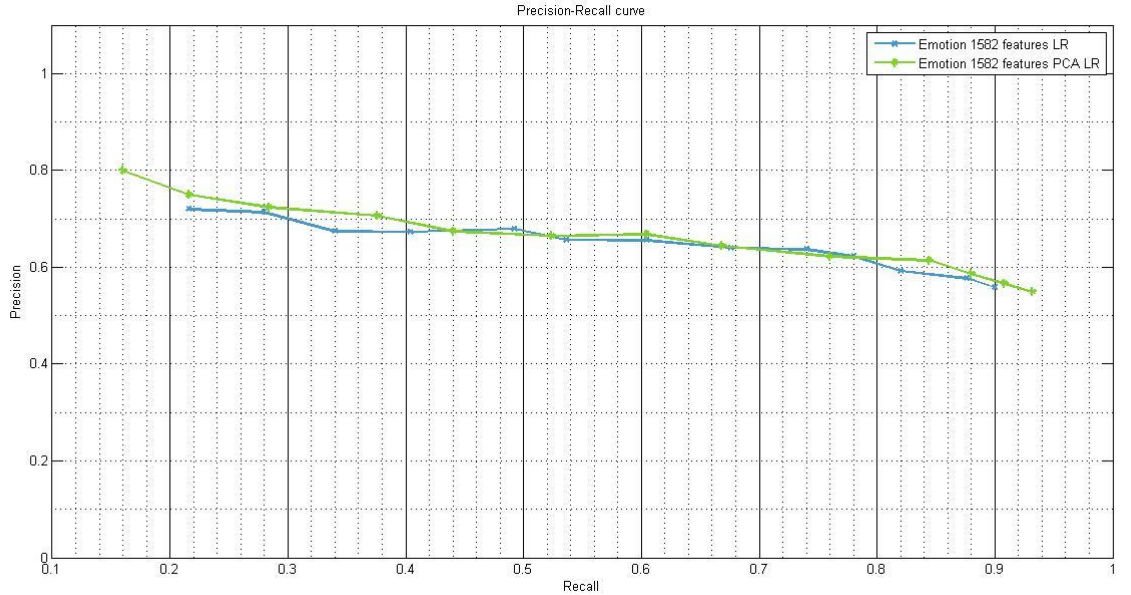
https://doi.org/10.26219/heal.aueb.4940

Το σύστημα 4 (Εικόνα 30) δείχνει να μαθαίνει αλλά η απόσταση των δύο καμπυλών παραμένει μεγάλη, μαρτυρώντας υπερ-εφαρμογή και αφήνοντας ανοιχτό το ενδεχόμενο τα αποτελέσματα να βελτιωθούν εάν αυξηθούν τα δεδομένα εκπαίδευσης.



Εικόνα 31: Καμπύλη μάθησης Συστήματος 5 (Λογιστική Παλινδρόμηση, 1582 χαρακτηριστικά, υποδειγματοληψία, PCA, LSSO)

Έπειτα εξετάζεται το ίδιο σύστημα εάν παράλληλα χρησιμοποιείται PCA (Σύστημα 5). Το σύστημα (Εικόνα 31) δείχνει να μαθαίνει αλλά αρκετά αργά. Η απόσταση των δύο καμπυλών είναι τώρα μικρότερη (λιγότερη υπερ-εφαρμογή). Χρειάζονται όμως πάλι περισσότερα δεδομένα εκπαίδευσης, ώστε οι καμπύλες να συγκλίνουν. Η σύγκριση των δύο παραπάνω συστημάτων φαίνεται στο παρακάτω διάγραμμα ανάκλησης-ακρίβειας (Εικόνα 32). Η χρήση PCA βελτιώνει ελάχιστα τις επιδόσεις (από 64,40% ορθότητα στα δεδομένα εκπαίδευσης στο Σύστημα 4 σε 65,20% Σύστημα 5), αλλά μειώνει το πλήθος των χαρακτηριστικών από 1582 σε 627.

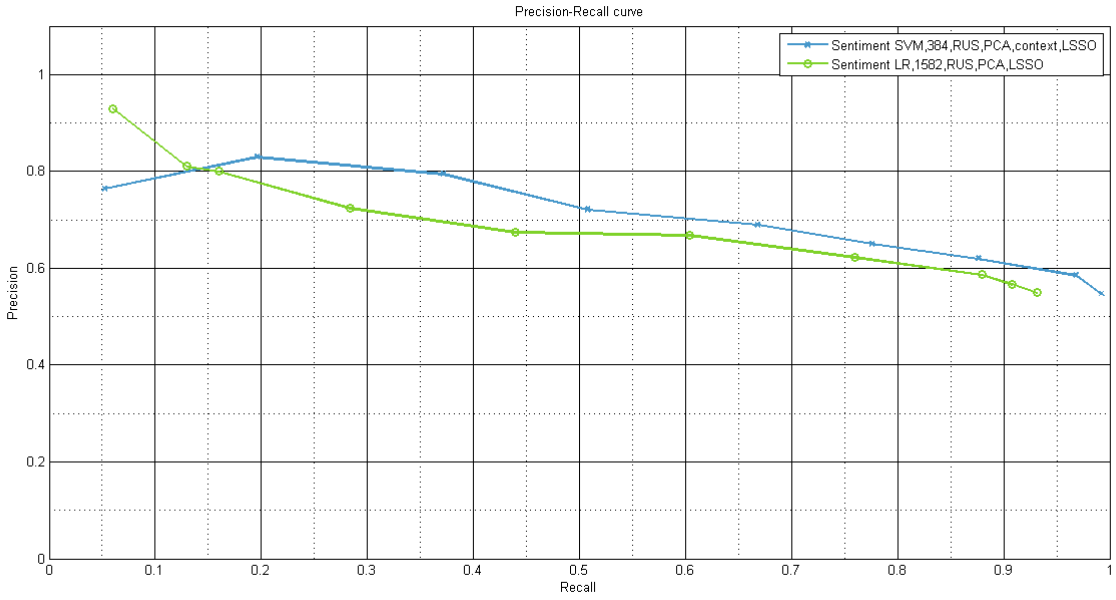


Εικόνα 32: Καμπύλη Ακρίβειας - Ανάκλησης Συστημάτων 4 και 5

https://doi.org/10.26219/heal.aueb.4940

Τα συμπεράσματα που εξάγονται παρατηρώντας το παραπάνω σχήμα είναι ότι για το σύνολο χαρακτηριστικών FS1582 μια καλή επιλογή για την αναγνώριση συναισθήματος είναι το Σύστημα 5, θεωρώντας ως συναισθηματικά τα στιγμιότυπα που κατατάσσονται σε αυτή την κατηγορία με πιθανότητα μεγαλύτερη του 0.35 καθώς σε αυτό το σημείο παρατηρείται 84,40% ανάκληση και 61,52% ακρίβεια.

Ολοκληρώνοντας τα πειράματα για την αναγνώριση του συναισθήματος παρουσιάζεται η καμπύλη μάθησης (Εικόνα 33) των δύο καλύτερων συστημάτων (Σύστημα 3 Εικόνα 28 και Σύστημα 5 31).



Εικόνα 33: Καμπύλη Ακρίβειας - Ανάκλησης Συστημάτων 3 και 5

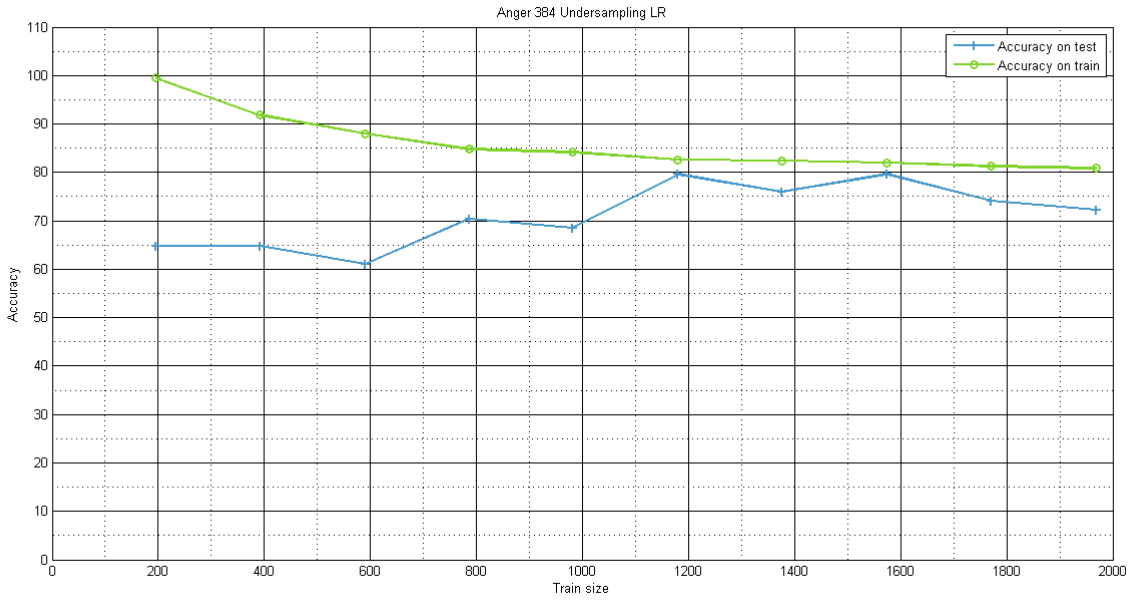
Το Σύστημα 3 φαίνεται να επιτυγχάνει καλύτερο ποσοστό ακρίβειας και ανάκλησης. Το συμπέρασμα που προκύπτει είναι ότι για την εκτίμηση της ύπαρξης συναισθήματος σε προφορικό διάλογο το καλύτερο σύστημα από εκείνα που δοκιμάσαμε είναι το Σύστημα 3.

5.2.2 Αποτελέσματα στην αναγνώριση θυμού

5.2.2.1 Αποτελέσματα μεθόδου LSSO για αναγνώριση θυμού

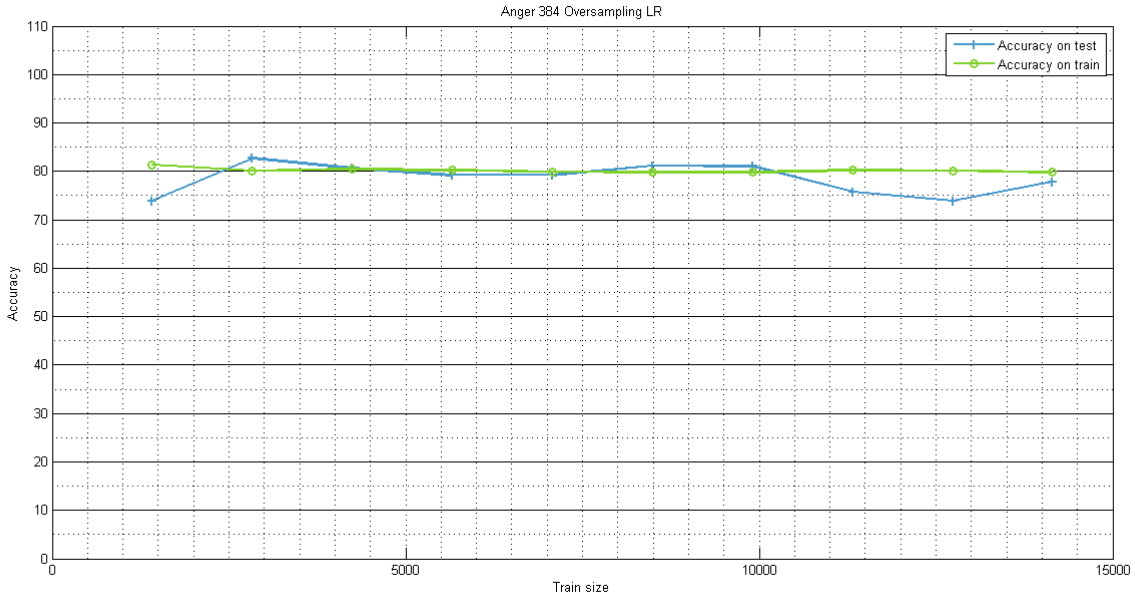
Στα επόμενα πείραμα εξετάζουμε την αναγνώριση θυμού. Ασχολούμαστε πάλι με όλα τα παράθυρα και όχι μόνο με όσα επέστρεψε η αναγνώριση συναισθημάτων, ώστε το αντίστοιχο υποσύστημα να μην εξαρτάται από την αναγνώριση συναισθήματος, αλλά να μπορεί να χρησιμοποιηθεί και αυτόνομα. Με διαχωρισμό δεδομένων πάντα μέσω LSSO, παρουσιάζονται τέσσερις μέθοδοι που παρουσιάζουν αξιοσημείωτα αποτελέσματα. Αρχικά εξετάστηκαν τα χαρακτηριστικά FS384, αφού πρώτα επιλέχθηκαν στιγμιότυπα με τη μέθοδο RUS (Σύστημα 6, Εικόνα 35). Ο αλγόριθμος μάθησης που χρησιμοποιήθηκε ήταν ο LR.

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 34: Καμπύλη μάθησης Συστήματος 6 (Λογιστική Παλινδρόμηση, 384 χαρακτηριστικά, υποδειγματοληψία, LSSO)

Το μοντέλο δείχνει να μαθαίνει μέχρι τα 1200 στιγμιότυπα εκπαίδευσης (Εικόνα 34) αλλά φαίνεται ότι περισσότερα στιγμιότυπα εκπαίδευσης δεν προσθέτουν παραπάνω πληροφορία. Αυτό φαίνεται και με τη δεύτερη μέθοδο, όπου στο ίδιο σύστημα έγινε υπερδειματοληψία ROS για να εξετάσουμε τι συμβαίνει εάν το πλήθος των παραδειγμάτων αυξηθεί (Σύστημα 7, Εικόνα 35).

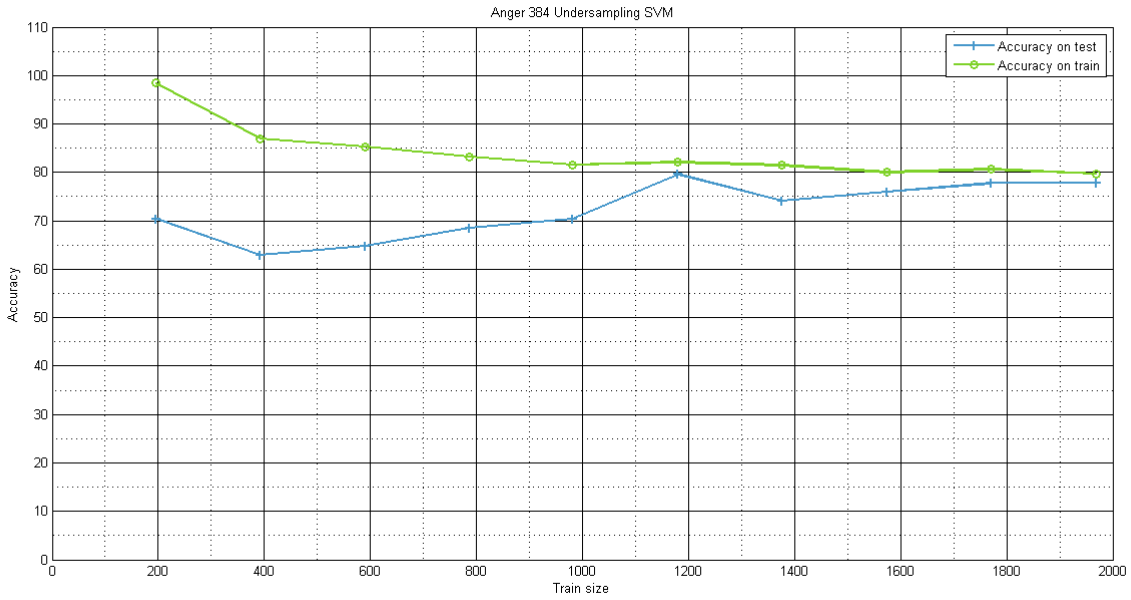


Εικόνα 35: Καμπύλη μάθησης Συστήματος 7 (Λογιστική Παλινδρόμηση, 384 χαρακτηριστικά, Υπερδειγματοληψία, LSSO)

Στην Εικόνα 35 παρατηρείται ότι για το σύστημα 7 το καλύτερο ποσοστό ορθότητας παράγεται για 2500 στιγμιότυπα. Φαίνεται επίσης να διατηρεί αυτό το ποσοστό μέχρι τα 10000 στιγμιότυπα. Για αυτό το λόγο συμπεραίνουμε ότι το μοντέλο χρειάζεται περίπου 2500 στιγμιότυπα.

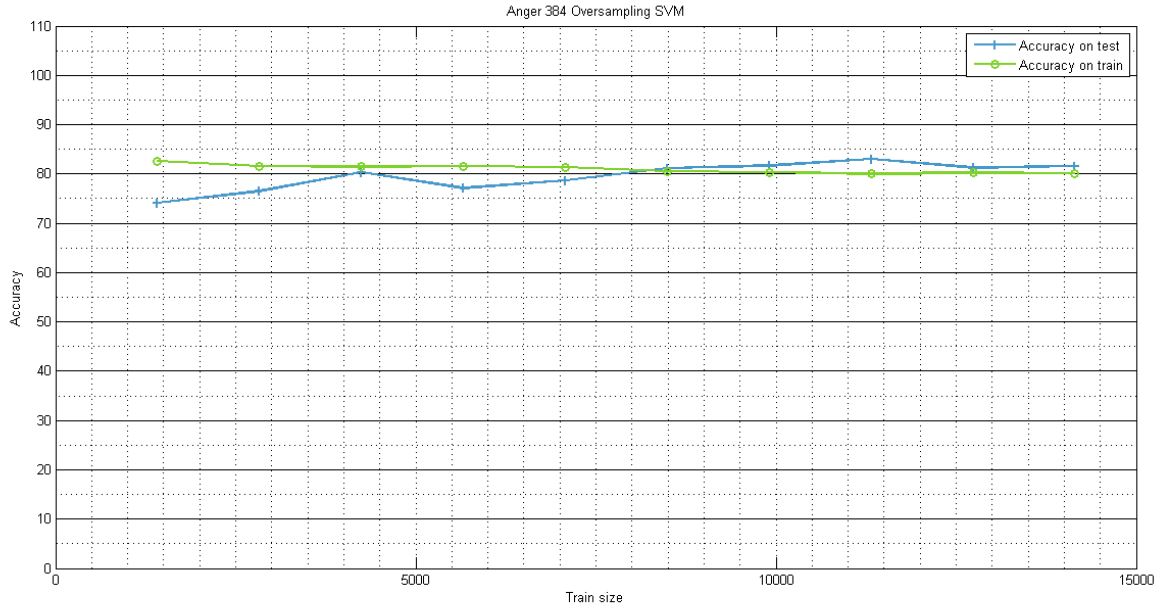
Στη συνέχεια δοκιμάστηκε το σύστημα με τα χαρακτηριστικά FS384, αλλά τώρα με SVM γραμμικού πυρήνα και υποδειγματοληψία RUS (Σύστημα 8).

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 36: Καμπύλη μάθησης Συστήματος 8 (SVM, 384 χαρακτηριστικά, Υποδειγματοληψία, LSSO)

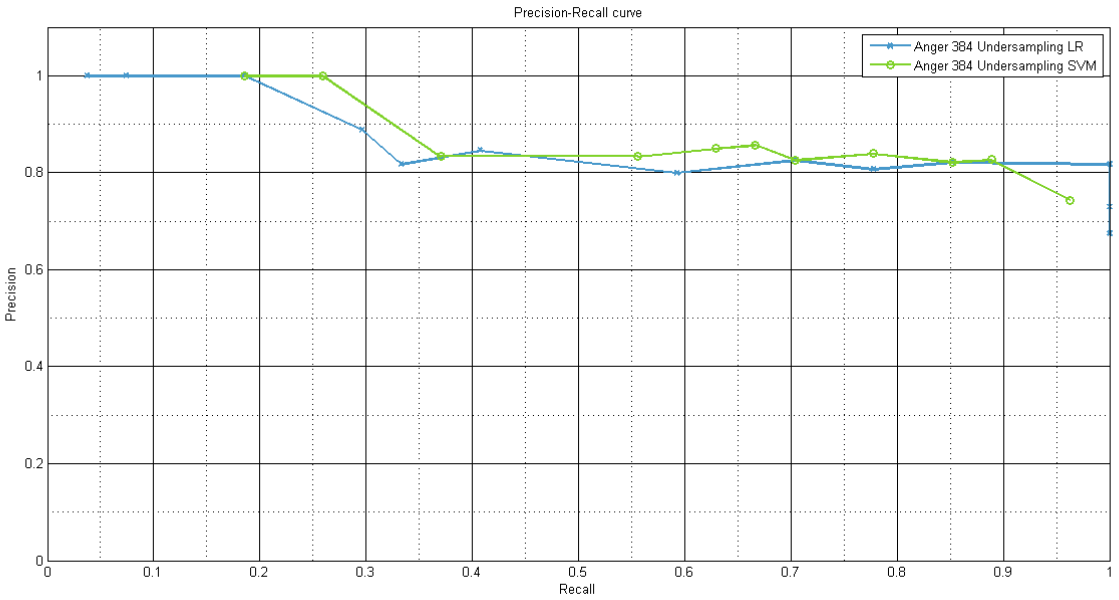
Όπως και στο Σύστημα 6 (Εικόνα 34) έτσι και στο Σύστημα 8 (Εικόνα 36) παρατηρείται ότι αρκούν 1200 στιγμιότυπα εκπαίδευσης για να φτάσουμε στο άνω φράγμα της ορθότητας στιγμιοτύπων ελέγχου. Αυτό φαίνεται και κατά την υπερδειγματοληψία ROS του ίδιου συστήματος (Σύστημα 9, Εικόνα 37) καθώς οι δύο καμπύλες για περισσότερα παραδείγματα εκπαίδευσης σχεδόν συμπίπτουν.



Εικόνα 37: Καμπύλη μάθησης Συστήματος 9 (SVM, 384 χαρακτηριστικά, Υπερδειγματοληψία, LSSO)

Στο παρακάτω σχήμα (Εικόνα 38) φαίνονται οι καμπύλες ανάκλησης-ακρίβειας των Συστημάτων 6 και 8, αφού εκείνα δείχνουν πιο υποσχόμενα. Το σύστημα που χρησιμοποιεί SVM (Σύστημα 8) φαίνεται να έχει ελαφρά καλύτερες επιδόσεις από το σύστημα που χρησιμοποιεί LR. Αλλά εάν αποφασίσουμε να κατατάξουμε στη κατηγορία του θυμού τα στιγμιότυπα για τα οποία το σύστημα επιστρέφει πιθανότητα μεγαλύτερη του 0.35 το Σύστημα 6 έχει ανάκληση 85,20% και ακρίβεια 82,14% ενώ το Σύστημα 8 έχει ανάκληση 77,78% και ακρίβεια 84,00%.

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 38: Καμπύλη Ακρίβειας - Ανάκλησης Συστημάτων 6 και 8

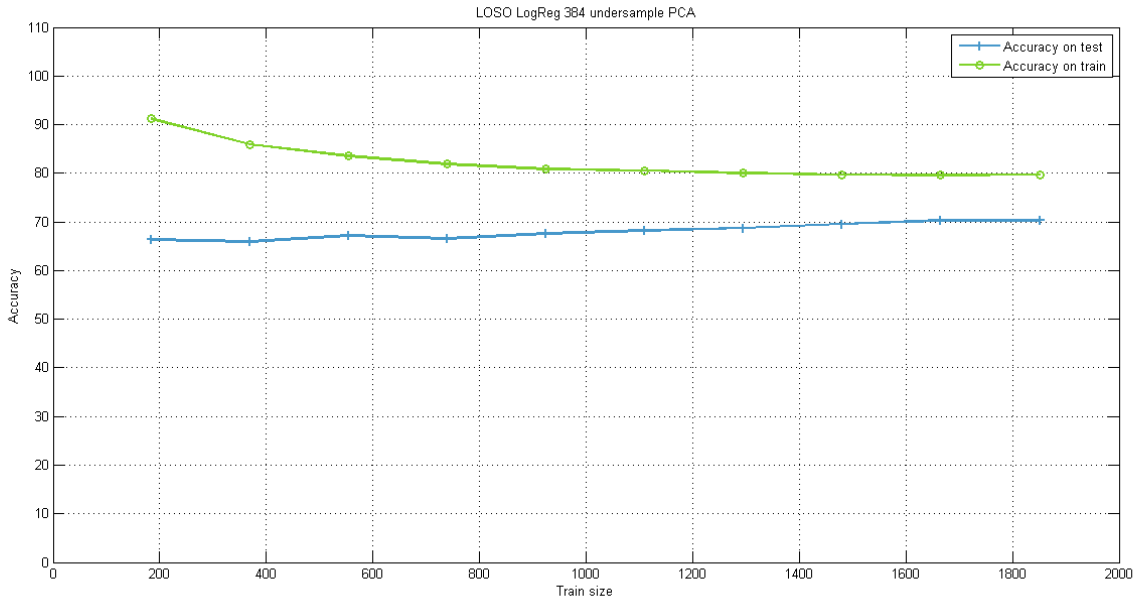
Τα συμπεράσματα που εξάγονται παρατηρώντας το παραπάνω σχήμα είναι ότι μια καλή επιλογή για την αναγνώριση θυμού (στην περίπτωση που χρησιμοποιείται LSSO) είναι το Σύστημα 8 (Εικόνα 36) γιατί παρατηρείται ότι στις περισσότερες περιπτώσεις η καμπύλη της ακρίβειας – ανάκλησης για το σύστημα 8 ξεπερνάει εκείνη του συστήματος 6 (Εικόνα 34).

5.2.2.2 Αποτελέσματα μεθόδου LOSO για αναγνώριση θυμού

Με τη μέθοδο LOSO εξετάζονται τώρα τρία συστήματα που τα αποτελέσματά τους είναι αξιοσημείωτα. Αντίθετα με τα προηγούμενα πειράματα, χρησιμοποιήθηκε η τεχνική micro-averaging (Ενότητα 4.8) στα μέτρα ορθότητα, ακρίβεια και ανάκληση, διότι με τη μέθοδο LOSO το σύστημα επιστρέφει αποτελέσματα για διαφορετικά μεγέθη δεδομένων ελέγχου στις επαναλήψεις της διασταυρωμένης επικύρωσης και η micro-averaging τεχνική βοηθάει να αποδοθούν βάρη στα αποτελέσματα ανάλογα με το πλήθος των δεδομένων ελέγχου κάθε επανάληψης. Για τη σύγκριση των συστημάτων παρουσιάζονται τα τρία καλύτερα συστήματα που προέκυψαν ύστερα από δοκιμές στα χαρακτηριστικά, τη δειγματοληψία, τη μείωση των διαστάσεων και τους αλγορίθμους μάθησης.

Αρχικά χρησιμοποιήθηκαν τα χαρακτηριστικά FS384 και ο αλγόριθμος LR. Στα δεδομένα εφαρμόστηκε υποδειγματοληψία RUS και ο αλγόριθμος PCA για τη μείωση των διαστάσεων (Σύστημα 10).

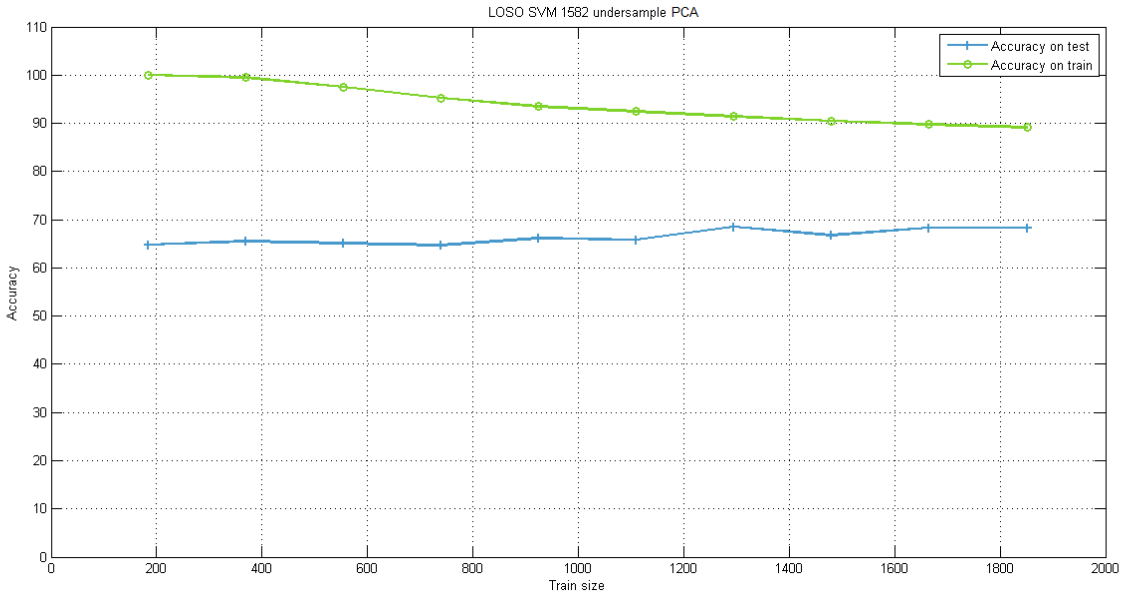
https://doi.org/10.26219/heal.aueb.4940



Εικόνα 39: Καμπύλη μάθησης Συστήματος 10 (Λογιστική Παλινδρόμηση, 384 χαρακτηριστικά, PCA, Υποδειγματοληψία, LOSO)

Παρατηρείται ότι το Σύστημα 10 δεν μπορεί να ξεπεράσει ποτέ το φράγμα του 81,81% για την ορθότητα που ορίζεται από το αποτέλεσμα του συστήματος όταν κατηγοριοποιεί τα στιγμιότυπα εκπαίδευσης. Φαίνεται ότι για τα δεδομένα ελέγχου η ορθότητα είναι 69,17% και ότι το σύστημα σταματάει να μαθαίνει στα 1700 στιγμιότυπα.

Στη συνέχεια εξετάστηκε η περίπτωση χρήσης των χαρακτηριστικών FS1582 με χρήση SVM αφού πρώτα έγινε υποδειγματοληψία RUS στα δεδομένα. Πριν την εκπαίδευση των μοντέλων εκτελέστηκε PCA (Σύστημα 11).

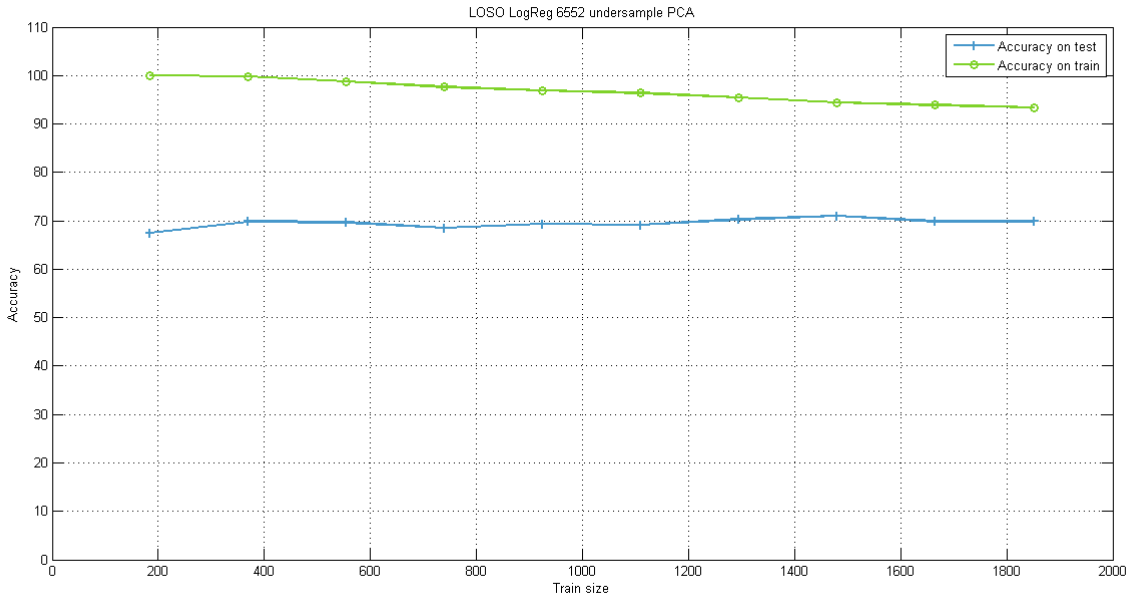


Εικόνα 40: Καμπύλη μάθησης Συστήματος 11 (SVM, 1582 χαρακτηριστικά, PCA, Υποδειγματοληψία, LOSO)

Το Σύστημα 11 μαθαίνει πολύ αργά και αργεί να συγκλίνει. Ενδεχομένως, βέβαια, να τα πήγαινε καλύτερα από το Σύστημα 10 αν υπήρχαν περισσότερα στιγμιότυπα εκπαίδευσης.

https://doi.org/10.26219/heal.aueb.4940

Τέλος παρουσιάζεται η περίπτωση όπου χρησιμοποιήθηκε το μεγαλύτερο σύνολο χαρακτηριστικών FS6552. Όπως και πριν, εφαρμόσθηκε RUS και PCA στα δεδομένα ενώ ο αλγόριθμος που χρησιμοποιήθηκε για τη ταξινόμηση είναι ο αλγόριθμος LR (Σύστημα 12).

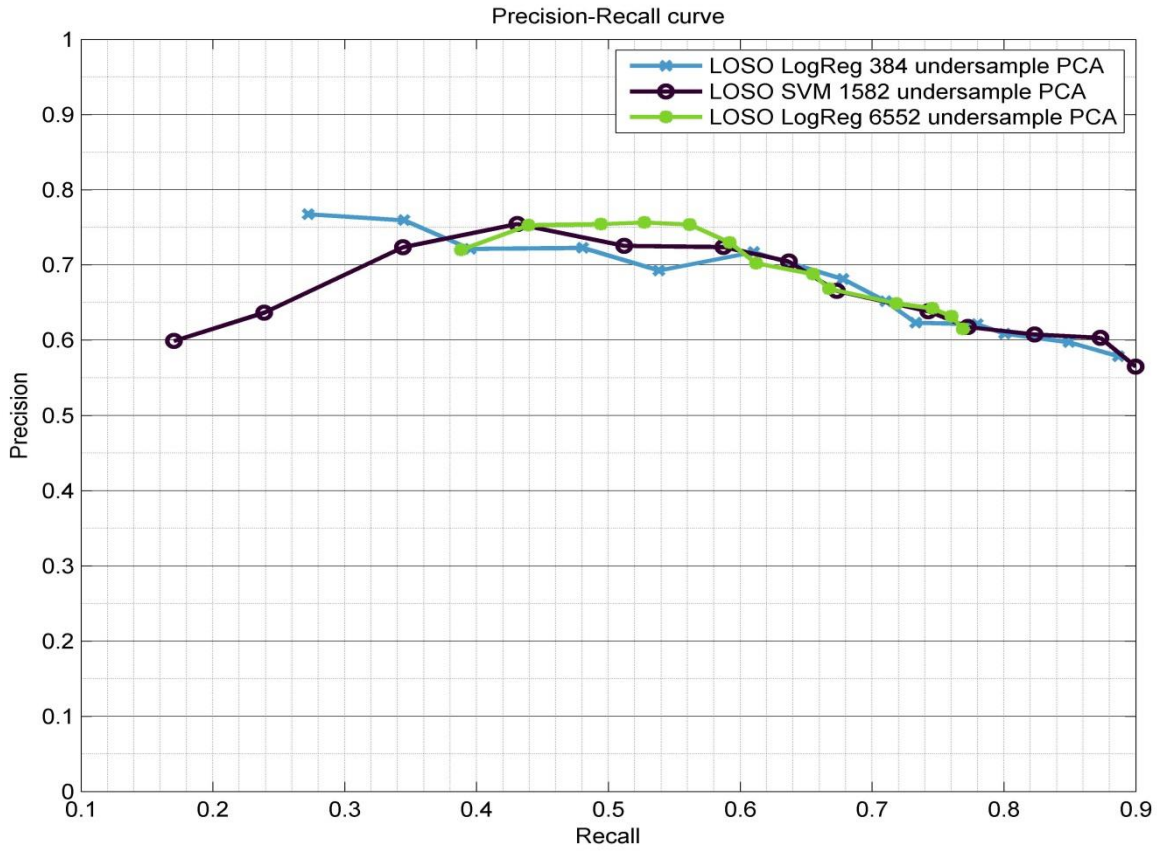


Εικόνα 41: Καμπύλη μάθησης Συστήματος 12 (Λογιστική Παλινδρόμηση, 6552 χαρακτηριστικά, PCA, Υποδειγματοληψία, LOSO)

Το σύστημα 12 φαίνεται ότι δεν μαθαίνει ή μαθαίνει πολύ αργά. Ενδεχομένως να χρειάζεται επιλογή χαρακτηριστικών πριν την μείωση των διαστάσεων με PCA, ώστε να εξαιρεθούν χαρακτηριστικά που προσθέτουν θόρυβο στο σύστημα.

Για να συγκρίνουμε τα παραπάνω συστήματα δημιουργήσαμε γραφική παράσταση Ακρίβειας-Ανάκλησης (Εικόνα 42).

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 42: Καμπύλη Ακρίβειας-Ανάκλησης Συστημάτων 10, 11 και 12

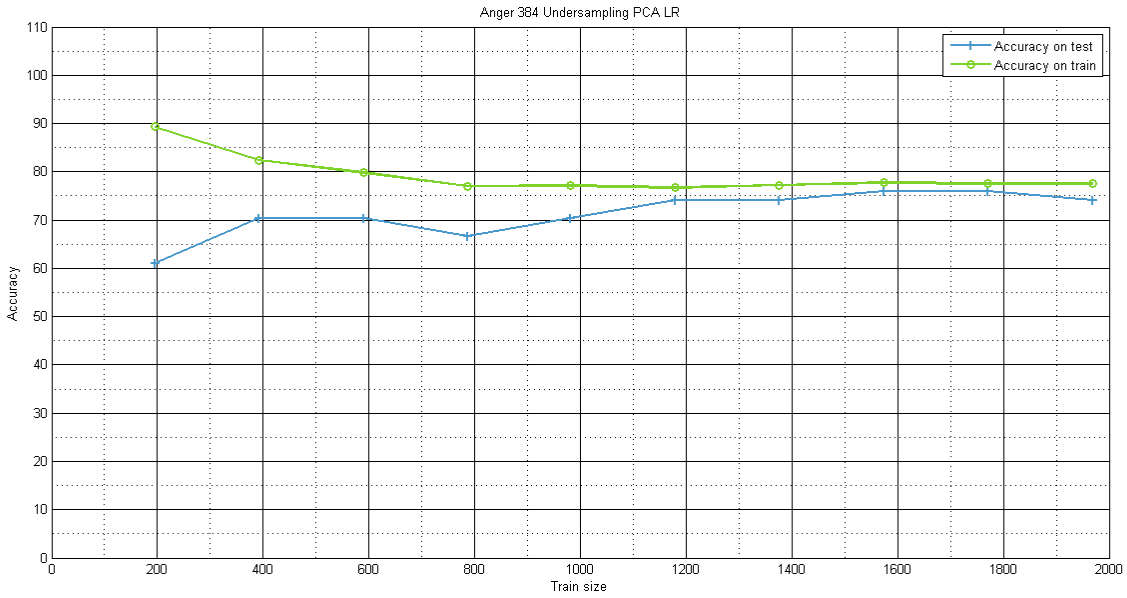
Όπως φαίνεται, τα τρία συστήματα δεν απέχουν αρκετά αν και φαίνεται να είναι καλύτερο εκείνο με τα 6552 χαρακτηριστικά (Σύστημα 12). Όμως από άποψη χρόνου εκτέλεσης, από άποψη υπολογιστικής ισχύος, και αν αναλογιστεί κανείς ότι το Σύστημα 12 δεν φαίνεται να εκπαιδεύεται, μπορεί κανείς να επιλέξει το σύστημα με τα 384 χαρακτηριστικά (Σύστημα 10).

Αξίζει να σημειωθεί επίσης ότι η μείωση διαστάσεων μέσω PCA για το σύνολο χαρακτηριστικών FS384 επέστρεψε ένα μόνο χαρακτηριστικό. Αυτό σημαίνει ότι ανακαλύφθηκε ένα χαρακτηριστικό το οποίο μπορεί και μετράει το θυμό στη συνομιλία. Το αποτέλεσμα εξηγείται εύλογα, αν σκεφτεί κανείς ότι στα συστήματα που χρησιμοποιούν τη μέθοδο Valence-Arousal το συναίσθημα του θυμού σχεδόν συμπίπτει με τον άξονα του Arousal. Επομένως ενδέχεται το χαρακτηριστικό που επιστρέφεται να εκφράζει το Arousal.

5.2.2.4 Αποτελέσματα με χαρακτηριστικά φύλου και ηλικίας στην αναγνώριση θυμού

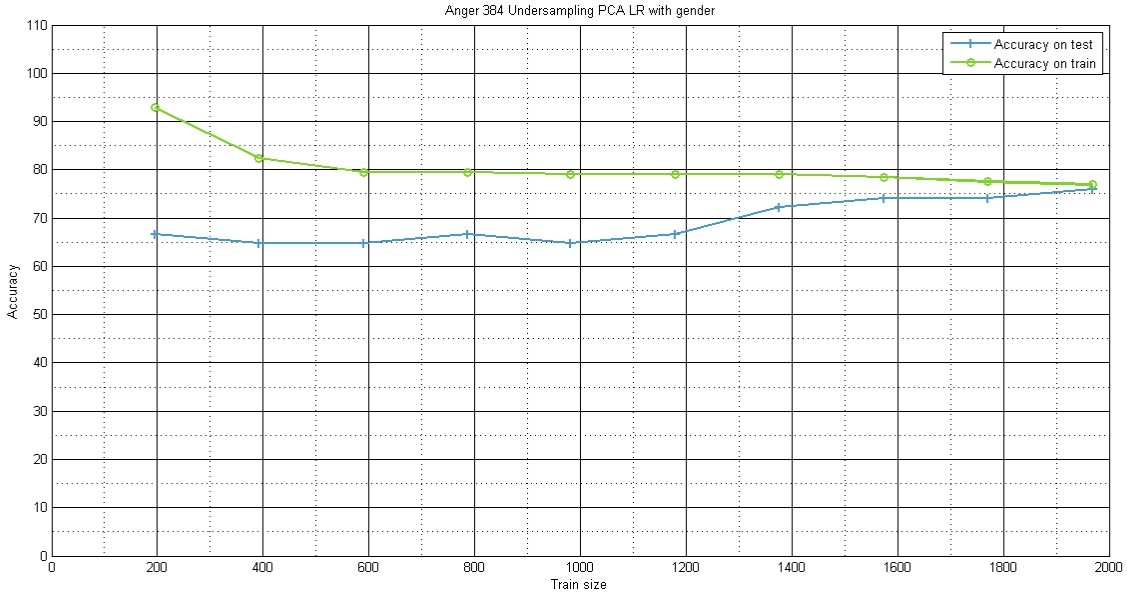
Στα τελευταία τρία πειράματα που παρουσιάζονται εξετάζεται η συνεισφορά της γνώσης του φύλου και της ηλικιακής ομάδας των ομιλητών στην αναγνώριση θυμού. Το σύστημα που εξετάζεται χρησιμοποιεί τα χαρακτηριστικά FS384 με λογιστική παλινδρόμηση και μείωση διαστάσεων με PCA, ενώ στα δεδομένα εφαρμόσθηκε RUS (Σύστημα 13).

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 43: Καμπύλη μάθησης Συστήματος 13 (Λογιστική Παλινδρόμηση, 384 χαρακτηριστικά, PCA, Υποδειγματοληψία, LSSO)

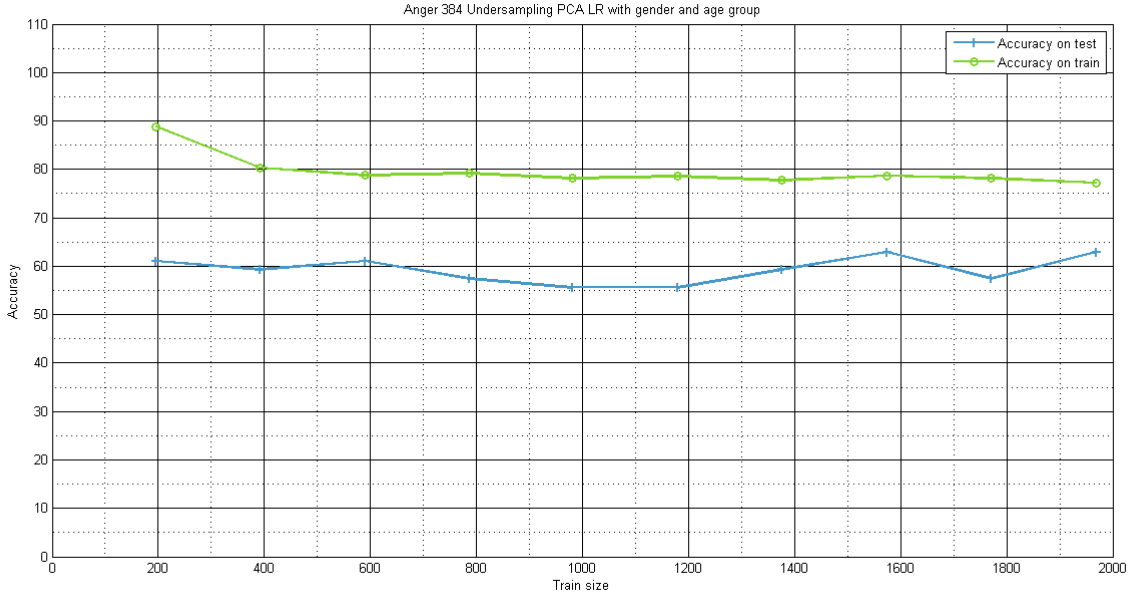
Αρχικά παρουσιάζεται η συνεισφορά της πληροφορίας για το φύλο (Σύστημα 14). Στα 384 χαρακτηριστικά προστέθηκαν δύο χαρακτηριστικά που εκφράζουν το φύλο των ομιλητών.



Εικόνα 44: Καμπύλη μάθησης Συστήματος 14 (Λογιστική Παλινδρόμηση, 384 χαρακτηριστικά, PCA, Υποδειγματοληψία, LSSO, με φύλο)

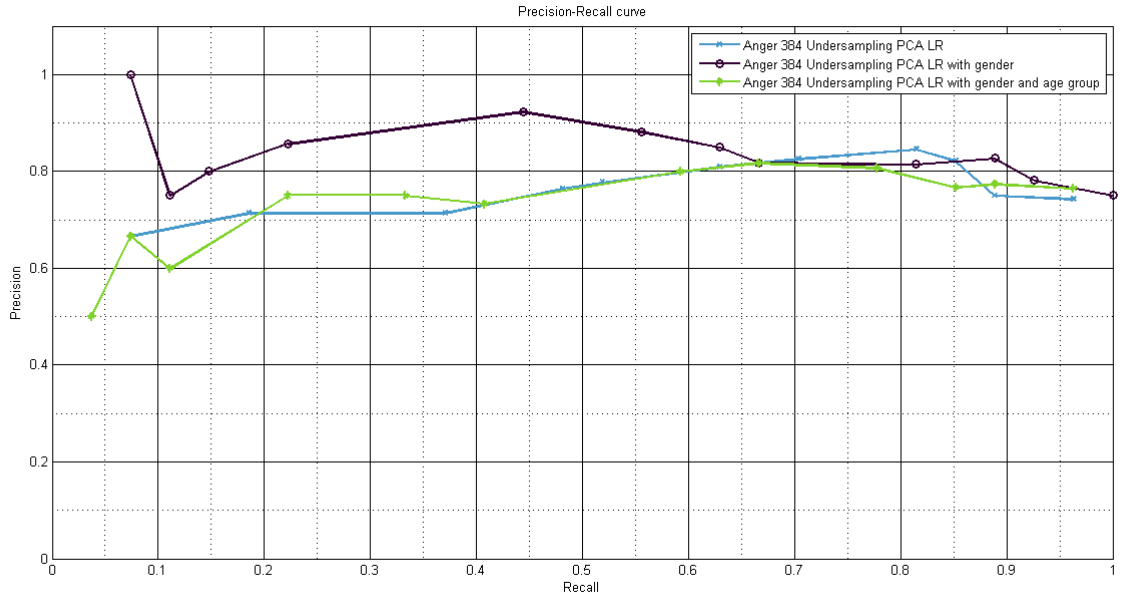
Το σύστημα φαίνεται να μαθαίνει και η καμπύλες συγκλίνουν στο 75,93% ορθότητα. Φαίνεται ότι αρκούν 2000 στιγμιότυπα εκπαίδευσης για το συγκεκριμένο μοντέλο. Έπειτα παρουσιάζεται η συνεισφορά της πληροφορίας του φύλου αλλά και της ηλικιακής ομάδας (Σύστημα 15). Μαζί με τα χαρακτηριστικά του φύλου των ομιλητών προστέθηκαν ακόμα δύο που υποδεικνύουν την ηλικιακή ομάδα των ομιλητών.

https://doi.org/10.26219/heal.aueb.4940



Εικόνα 45: Καμπύλη μάθησης Συστήματος 15 (Λογ. Παλινδρόμηση, 384 χαρ., PCA, Υποδειγματοληψία, LSSO, με φύλο και ηλ.ομάδα)

Παρατηρείται ότι το τελευταίο σύστημα δεν μαθαίνει. Η πληροφορία για την ηλικιακή ομάδα προσθέτει θόρυβο στο σύστημα, επομένως χρειάζεται να εξαιρεθεί. Για να συγκριθούν τα τρία συστήματα δημιουργήθηκε γράφημα Ακρίβειας- Ανάκλησης.



Εικόνα 46: Καμπύλη Ακρίβειας-Ανάκλησης Συστημάτων 13, 14 και 15

Παρατηρείται ότι το σύστημα βελτιώνεται με χρήση της πληροφορίας του φύλου των ομιλητών. Η χρήση αυτής της πληροφορίας βέβαια προϋποθέτει την ύπαρξη κάποιου συστήματος που επισημειώνει το φύλο των ομιλητών χρησιμοποιώντας τη φωνή. Από τις καμπύλες μάθησης προκύπτει ότι το σύστημα συγκλίνει πιο γρήγορα χωρίς τη χρήση των πληροφοριών αυτών. Οπότε υπάρχουν δύο εναλλακτικές επιλογές. Η πρώτη είναι να μην χρησιμοποιήσουμε πληροφορία για το φύλο, το οποίο οδηγεί σε 81,48% ανάκληση και 84,62% ακρίβεια, κατατάσσοντας στην κατηγορία των στιγμιοτύπων που περιέχουν θυμό τα στιγμιότυπα για το οποία το σύστημα παράγει εκτίμηση πιθανότητας μεγαλύτερη του 0,35. Η δεύτερη προϋποθέτει τη χρήση της πληροφορίας για το φύλο (είτε με

χειρωνακτική επισημείωση είτε με έναν ταξινομητή που θα προσθέτει όμως σφάλμα). Σε ένα τέτοιο σύστημα εάν κατατάξουμε στην κατηγορία των στιγμιοτύπων που περιέχουν θυμό τα στιγμιότυπα για το οποία το σύστημα παράγει εκτίμηση πιθανότητας θυμού μεγαλύτερη του 0,35 τότε έχουμε 88,89% ανάκληση και 82,76% ακρίβεια.

https://doi.org/10.26219/heal.aueb.4940

6. Συμπεράσματα και μελλοντική έρευνα

6.1 Συμπεράσματα

Τα προβλήματα που μελετήθηκαν στην εργασία είναι δύο: (i) η αναγνώριση τμημάτων ήχου μίας τηλεφωνικής συνομιλίας τα οποία περιέχουν συναισθηματικό φορτίο και (ii) η αναγνώριση τμημάτων ήχου που περιέχουν θυμό. Δημιουργήθηκαν και εξετάστηκαν συστήματα τα οποία χρησιμοποιούν μια μέθοδο κυλιόμενου «παραθύρου» για να εξάγουν διανύσματα χαρακτηριστικών (feature vectors) από μικρά διαδοχικά τμήματα («παράθυρα») ενός διαλόγου. Χρησιμοποιήθηκε το εργαλείο OpenSmile για την εξαγωγή των χαρακτηριστικών, ενώ για τη μείωση των διαστάσεων του διανυσματικού χώρου χρησιμοποιήθηκε PCA. Τέλος χρησιμοποιήθηκαν οι αλγόριθμοι μάθησης της Λογιστικής Παλινδρόμησης και των Μηχανών Διανυσμάτων Υποστήριξης για την κατάταξη των παραθύρων σε κατηγορίες.

Ένα πρόβλημα που προέκυψε είναι το μικρό πλήθος παραδειγμάτων (παραθύρων) εκπαίδευσης. Κάποιοι αλγόριθμοι μάθησης δεν είχαν το απαιτούμενο πλήθος παραδειγμάτων εκπαίδευσης, με αποτέλεσμα να έχουν περιθώριο για καλύτερα αποτελέσματα εάν προσθέσουμε νέα παραδείγματα. Χρειάζεται λοιπόν να επισημειωθούν και να προστεθούν νέα παραδείγματα.

Ένα ακόμη θέμα υπό συζήτηση αποτελούν οι πέντε παράμετροι του συστήματος: το μέγεθος του παραθύρου σε msec, η επικάλυψη διαδοχικών παραθύρων, το ποσοστό επικάλυψης με μία χειρωνακτική επισημείωση το οποίο πρέπει να ξεπεράσει ένα παράθυρο για να επισημειωθεί αντίστοιχα, το πλήθος των προηγούμενων παραθύρων που λαμβάνονται υπόψη κατά την κατάταξη του τρέχοντος παραθύρου, το πλήθος των διαστάσεων του νέου χώρου στον οποίο απεικονίζονται τα διανύσματα χαρακτηριστικών μέσω PCA. Χρειάζεται λοιπόν να ρυθμιστούν οι παράμετροι αυτές, ώστε να βελτιωθούν ενδεχομένως τα αποτελέσματα.

Για την αναγνώριση συναισθήματος (πρόβλημα i), το καλύτερο αποτέλεσμα προήλθε από το Σύστημα 3 (Ενότητα 5.2.1) το οποίο χρησιμοποιεί 768 χαρακτηριστικά (από το υπό εξέταση παράθυρο και το προηγούμενό του) και περιλαμβάνουν την ενέργεια, το pitch, τα χαρακτηριστικά MFCC και συναρτησιακά χαρακτηριστικά (functionals) των παραθύρων. Χρησιμοποιεί PCA για μείωση διαστάσεων και υποδειγματοληψία για την αντιμετώπιση της ανισοκατανομής. Τέλος χρησιμοποιεί SVM με γραμμικό πυρήνα για ταξινόμηση των στιγμιοτύπων. Κατατάσσοντας ως θετικά (θεωρώντας ότι περιέχουν συναίσθημα) τα παράθυρα που κατατάσσονται στη θετική κατηγορία με εκτίμηση πιθανότητας μεγαλύτερη του 0.4, λαμβάνουμε ανάκληση 80,40% και ακρίβεια 64,42%.

Για την αναγνώριση θυμού (πρόβλημα ii), το καλύτερο αποτέλεσμα προήλθε από το Σύστημα 10 (Ενότητα 5.2.3) το οποίο χρησιμοποιεί 384 χαρακτηριστικά (από το υπό εξέταση παράθυρο μόνο) και περιλαμβάνουν την ενέργεια, το pitch, τα MFCCs και συναρτησιακά χαρακτηριστικά των παραθύρων. Χρησιμοποιεί PCA για μείωση διαστάσεων και υποδειγματοληψία για την αντιμετώπιση της ανισοκατανομής. Τέλος χρησιμοποιεί Λογιστική Παλινδρόμηση για ταξινόμηση των στιγμιοτύπων.

https://doi.org/10.26219/heal.aueb.4940

Κατατάσσοντας ως θετικά (με θυμό) τα παράθυρα που κατατάσσονται στη θετική κατηγορία με εκτίμηση πιθανότητας μεγαλύτερη του 0.35, λαμβάνουμε ανάκληση 77,89% και ακρίβεια 62,13%.

6.2 Μελλοντική Έρευνα

Κατά τη χειρωνακτική επισημείωση των δεδομένων, παρατηρήσαμε ότι ο επισημειωτής δεν είναι πάντα βέβαιος για την ορθή επισημείωση (π.χ. αν ένα τμήμα ήχου περιέχει ή όχι συναίσθημα ή ειδικότερα θυμό). Συχνά η άποψή του για το πότε ένα παράθυρο πρέπει ή όχι να επισημειώνεται ως παράθυρο συναισθήματος ή θυμού ειδικότερα αλλάζει καθώς ακούει νέα τμήματα ήχου και νέες συνομιλίες και αυτό το θέμα χρειάζεται να διερευνηθεί περαιτέρω (π.χ. μετρώντας τη συμφωνία μεταξύ κριτών και δίνοντάς τους οδηγίες που να καθοδηγούν την επισημείωση).

Επίσης, εάν ο επισημειωτής επισημειώνει συνομιλίες σε γλώσσα που καταλαβαίνει (όπως η μητρική του) ακούσια χρησιμοποιεί και το νόημα των λεγομένων. Είναι επομένως «άδικο» να εκπαιδεύονται αλγόριθμοι με χρήση μόνο ακουστικών χαρακτηριστικών. Χρειάζεται να εξετασθούν και άλλα, μη ακουστικά χαρακτηριστικά, για παράδειγμα λεξικολογικά, αλλά ενδεχομένως και οπτικά (π.χ. από το πρόσωπο ή τις χειρονομίες), κατεύθυνση που η βιβλιογραφία δείχνει πως είναι γόνιμη (Min Lee et al., [17]). Για να γίνει λεξικολογική επεξεργασία, απαιτείται επεξεργασία των δεδομένων ώστε να αφαιρεθεί ο θόρυβος, να διαχωριστούν τα ηχητικά σήματα των ομιλητών και να γίνει επιτυχής αναγνώριση ομιλίας (speech recognition).

7. Παράρτημα 1

Συγγραφείς	Έτος	Εργασία	Σκοπός	Δεδομένα	Γλώσσα	Κατηγορίες	Μέτρο Αξιολόγησης	Αποτέλεσμα	Αλγόριθμος Μάθησης	Επιλογή Χαρακτηριστικών Ή Μείωση Διαστάσεων Δεδομένων	Χαρακτηριστικά
Lay Nwe, Say Wei, De Silva	2001	[10]	Emotion Classification	acted emotion		7 (Anger, Dislike, Fear, Happiness, Sadness, Surprise, Neutral)	Accuracy	72.22%	HMM		power
Min Lee, Narayanan, Pieraccini	2001	[11]	Emotion Detection	(SpeechWorks call center application) Real data recording		2 (non negative, negative)	Error	27.54% (SVM)	LDC SVMs kNN	PCA	Pitch energy
Lee, Narayanan, Pieraccini	2001	[12]	Recognition of Negative Emotions	real users from telephone company call center with a machine agent		fear, sadness, happiness, surprise, anger, disgust	Error	20%	kNN	PCA	Pitch energy
Burkhardt, Ballegooy, Englert, Huber	2003	[13]	Anger Detection	recordings (acted)		3 (not angry, low anger, high anger)	Accuracy	55% για 3 κατηγορίες (71% για 2)	Gaussian Density		Pitch Energy Duration
Schuller, Rigoll, Lang	2003	[14]	Emotion Recognition	Acted in Isolated room	German,English	7 (anger, disgust, fear, surprise, joy, neutral, sadness)	Accuracy	68.5 (anger) 84.7 (disgust) 95.4 (fear) 73.5 (surprise) 68.0 (joy) 79.6 (neutral) 75.1 (sadness) (not phrase based)	GMM		Pitch Energy
Ververidis, Kotropoulos	2005	[15]	Emotion Classification	DES Database (acted data)		5 (anger, happiness, sadness, surprise, neutral)	Accuracy	67.3%	GMM	SFFS (Sequential Floating Forward Selection)	Pitch Energy formants
Inanoglu, Caneel	2005	[16]	Emotion Detection	PhoneCell Voicemail messages, CallHome English Speech Corpus, Oasis Database	English	8 (Happy, Sad, Excited, Calm, Urgent, Not Urgent, Formal, Informal), 4 (Valence, Arousal, Urgency, Formality)	Accuracy	87% Happy 44% Sad 68% Excited 79% Calm 66% Urgent 92% Not Urgent 73% Formal 68% Informal	GMM, HMM		Pitch Loudness segment duration
Min Lee, Narayanan	2005	[17]	Emotion Detection	Real data recording		2 (negative , non negative)	Error	7.95 (Accoustic and Language)	LDC, kNN	Forward Selection, PCA	F0 Duration Energy Formants Lexical Features Discourse information
Vidrascu, Devillers	2005	[18]	Emotion Detection	A web-based stock exchange Customer service center	French	2 (Negative, Neutral) , 2 (Anger, Fear)	Accuracy	73% (Negative,Neutral) 60%(Anger,Fear)	Decision Tree, SVM, Voting Algorithms (AddTRee And AdaBoost)		Energy Duration spectral features disfluency f0 features
Vidrascu, Devillers	2005	[19]	Emotion Detection	call center medical advice services		2 (positive, negative)	Accuracy	83.49%	SVM, Logistic Model Tree	weka SVM, OneR, CfsSubset,GainRatio	Prosodic Spectral non-verbal events disfluency energy f0 duration
Neiberg, Elenius, Karlsson, Laskowski	2006	[20]	Emotion Recognition	2 corpora: Swedish Voice Provider Company Recordings, ISL Meeting Corpus		3 (neutral, emphasized, negative (VP)), 3 (negative, neutral, positive(ISL))	Accuracy	93% (using priors VP) 82% (using priors ISL)	GMM		MFCC MFFC-low pitch
Vidrascu, Devillers	2007	[21]	Emotion Detection	CEMO Corpus (Medical emergency call center)		5 (Fear, Anger, Sadness, Relief, Neutral)	Accuracy	53% (Fear) 60% (Anger) 58% (Sadness) 68% (Relief) 63% (Neutral)	SVM Gaussian Classifier	SVM	speaker personality age gender relationship to the patient F0 MFCC Energy Microprosody Formants Duration spectral
Soltani, Noor Ainon	2007	[22]	Emotion Detection	Berlin Database of Emotional Speech (acted)	German	6 (happy, angry, anxious, fearful, bored,	Accuracy	82.3%(happy) 76.5% (angry) 70.1% (anxious) 75.4% (fearful) 74.1% (bored)	2 Layer Back propagation neural network	manual forward selection followed by backward elimination	F0 Energy Formant voicing rate

						neutral)		83.4% (neutral)			
Li, Tao, Johnson, Soltis, Savage, Leong, Newman	2007	[23]	Stress and Emotion Classificaton	SUSAS Dataset, Rhesus Dataset, African Elephant Emotional Arousal Dataset		2 (Arousal, Non-Arousal)	Accuracy	72.7% (speaker independent)	Hmm (each state 4 GMMs)		Jitter Shimmer MFCCs Energy GFCCs
Grimm, Kroschel, Narayanan	2007	[24]	Ποσοτική μέτρηση valence, activation, dominance	VAM Corpus	German	3 (valence, activation, dominance)	Error	0.13(valence) 0.15(activation) 0.14(dominance)	SVR-RBF, SVR-Poly, SVR-Linear, KNNS	Sequential forward selection (SFS)	Pitch MFCCs Prosodic Energy 1st and 2nd derivatives
Neiberg, Elenius	2008	[26]	Anger Recognition	Swedish voice provider company recordings		3 (neutral, emphasized, negative)	Recall	71% (for speaker independent classification)	GMM	Incremental addition of features	MFCCs
Lugger, Yang	2008	[27]	Emotion Classification	Berlin Emotional Database	German	6 (happy, bored, neutral, sad, angry, anxious)	Accuracy	82.2(happy) 94.6 (bored) 87.4 (neutral) 86.7 (sad) 94.1 (angry) 86.7 (anxious)	Naive Bayes	SFFS Sequential Floating Forward Selection	Prosody MFCCs Voice Quality Parameters
Burkhardt, Polzehl, Stegmann, Metze	2008	[25]	Anger Detection	voice portal (customers report problems with phone connection to automated voice dialog)	German	2 (angry, non angry)	Accuracy	70%	GMM, weka SVM	PCA	Prosodic Duration Pitch Intensity MFCC
Batliner, Steidl, Schuller, Seppi, Vogt, Wagner, Devillers, Vidrascu, Aharonson, Kessous, Amir	2009	[28]	Emotion Distinction	Aibo Emotion Corpus (children)	German	4 (Neutral, Emphatic, Angry, Motherese)	Fmeasure	63.4% (0,629842539 Accuracy)	SVMs		Duration Energy Pitch Spectrum Cepstrum voice quality wavelets bag of words part of speech higher semantics varia sequential and combinatorial extremes means percentiles higher statistical moments specific functions not attributable
Schmitt, Pieraccini, Polzehl	2010	[29]	Anger Detection	calls from automated internet agent		2 (anger, non_anger)	f-score	73,74% for anger	SVM	Information Gain Ratio	Acoustic Prosodic MFCC Pitch Loudness Intensity harmonics-to-noise ratio formants call quality frustration history
Hicham Atassi, Zdenek Smékal, Anna Esposito	2012	[30]	Emotion Recognition	call center agent-client phone calls (real data)	Czech, Russian, Slovak, Polish	5 (neutral, surprise, happiness, sadness, anger) mapped to 2 (Valence, Activation)	mean absolure error	1.1 (neutral) 0.87 (surprise) 1.79 (happiness) 0.86 (sadness) 1.05 (anger) 0.37 (Valence) 0.39 (Activation)	SVR, RBF, ANN	(minimum Redundancy Maximum Relevance), FFS	MFCC HFCC LFCC MELBS HFBS LFBS PLP
Sheikhan, Bejani, Gharavian	2012	[31]	Emotion Recognition	FARSDAT Corpus	Farsi	3 (Happiness, Anger, Neutral)	Accuracy	76.3%	Modular neural - SVM model	ANOVA	Formant Pitch LE and 1st and 2nd derivatives MFCCs
Rozgić, Anantha krishnan, Saleem, Kumar, Prasad	2012	[32]	Emotion Recognition	USC-IEMOCAP (acted data)		5 (Anger, Happiness, Excitement, Sadness, Neutral)	Accuracy	60.9% (acoustic) 67.4%(acoustic and lexical) 69.4%(acoustic, lexical and visual)	SVM Trees		Acoustic (MFCC Jitter Shimmer F0 Intensity formants), Lexical (number of words Words word stems) Visual (Facial Animation Parameters)
Pohjalainen, Alku	2013	[33]	Anger Detection	EMO-DB (Berlin Database)	German	7 (anger, disgust, fear, joy, sadness, boredom, neutral)	Equal Error Rate (EER)	7.7%	GMM		MFCC LP (linear prediction)
Chaspari, Dimitriadis, Maragos	2013	[34]	Emotion Classification	EmoDB	German	7 (Anger, Boredom, Disgust, Fear, Happiness, Sadness, Neutral)	Accuracy	72.6% (Leave One Speaker Out)	GMM, SVM	Fisher Discriminant Ratio criterion	MFCCs Prosody
Vinay, Shilpi Gupta, Anu Mehra	2014	[35]	Emotion Recognition	SAVEE Database (acted)		7 (anger, surprise, disgust, fear, happiness, sadness, neutral)	Accuracy	78.28% 81.85%(with gender)	Naive Bayes	PCA	Energy Pitch MFCCs gender

https://doi.org/10.26219/heal.aueb.4940

9. Αναφορές

[1]. T. Dalgleish, M.J. Power, *Handbook of cognition and emotion*, Wiley Online Library, 1999.

[2]. C. Chang, C. Lin, "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):27, 2011.

[3]. H. Brugman, A. Russel, X. Nijmegen, Annotating multi-media/multi-modal resources with ELAN, LREC, Lisbon, Portugal, 2004.

[4]. B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C.A. Muller, S.S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge.", INTERSPEECH, Makuhari, Japan, 1(1):2794--2797, 2010.

[5]. F. Eyben, M. Woellmer, B. Schuller, Speech and music interpretation by large-space extraction, Technische Universitaet Muenchen, 2009.

[6]. C.N. van der Wal, W. Kowalczyk, Detecting changing emotions in human speech by machine and humans, Applied intelligence, Springer, 39(4): 675-691, 2013.

[7]. B. Schuller, S. Steidl, A. Batliner, The INTERSPEECH 2009 emotion challenge., INTERSPEECH, Brighton, UK, 2009: 312-315, 2009.

[8]. V. Ganganwar, An overview of classification algorithms for imbalanced datasets, International Journal of Emerging Technology and Advanced Engineering (IJETAE), 2(4): 42-47, 2012.

[9]. O. Maimon, L. Rokach, Data mining and knowledge discovery handbook, 2, Springer, 2005.

[10]. T.L. Nwe, F.S. Wei, L.C. De Silva, Speech based emotion classification, TENCON (2001: Singapore), 1(1):297-301, 2001.

[11]. C. M. Lee, S. S. Narayanan, R. Pieraccini, Classifying emotions in human-machine spoken dialogs, ICME 2002, Rochester, NY, 1:737-740,2002.

[12]. C. M. Lee, S. S. Narayanan, R. Pieraccini, Recognition of negative emotions from the speech signal, IEEE workshop on automatic speech recognition and understanding, 1: 240-243, 2001

[13]. F. Burkhardt, M. Van Ballegooy, R. Englert, R. Huber, An emotion-aware voice portal, ESSP, Burkhardt, 2005.

[14]. B. Schuller, G. Rigoll, M. Lang, Hidden Markov model-based speech emotion recognition, ICASSP 2003, Hong Kong, 2:1,2003.

[15]. D. Ververidis, C. Kotropoulos, Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm, ICME 2005, NY, USA, 1: 1500-1503, 2005.

[16]. Z. Inanoglu, R. Caneel, Emotive alert: HMM-based emotion detection in voicemail messages, IUI 2005, San Diego, USA, 1: 251-253, 2005.

[17]. C.M. Lee, S. S. Narayanan, Toward detecting emotions in spoken dialogs, IEEE ACM Transactions, 13(2):293-303, 2005.

[18]. L. Vidrascu, L. Devillers, Annotation and detection of blended emotions in real human-human dialogs recorded in a call center, ICME 2005, NY, USA, 1841-1845, 2005.

[19]. L. Vidrascu, L. Devillers, Detection of real-life emotions in call centers, INTERSPEECH 2005, Lisbon, Portugal, 2005(10): 1841-1844, 2005.

[20]. D. Neiberg, K. Elenius, I. Karlsson, K. Laskowski, Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics 101 Working Papers, 52:101-104, 2006.

[21]. L. Vidrascu, L. Devillers, Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features, ParaLing 2007, Saarbrücken, Germany, 2007.

[22]. K. Soltani, R. N. Ainon, Speech emotion detection based on neural networks, ISSPA, Sharjah UAE, 1: 1-3, 2007.

[23]. X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, J. D. Newman, Stress and emotion classification using jitter and shimmer features, ICASSP 2007, Honolulu, Hawai'i, 4: 1081, 2007.

[24]. M. Grimm, K. Kroschel, S. Narayanan, Support vector regression for automatic recognition of spontaneous emotions in speech, ICASSP, Honolulu, Hawai'i, 4: 1085, 2007.

[25]. F. Burkhardt, T. Polzehl, J. Stegmann, F. Metze, R. Huber, Detecting real life anger, ICASSP, Taipei, Taiwan, 4761-4764, 2009.

[26]. D. Neiberg, K. Elenius, Automatic recognition of anger in spontaneous speech, INTERSPEECH 2008, Brisbane, Australia, 2755-2758, 2008.

[27]. M. Lugger, B. Yang, Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters, ICASSP 2008, Las Vegas, USA, 4945-4948, 2008.

[28]. A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, Whodunnit-searching for the most important feature types signalling emotion-related user states in speech, Computer Speech & Language, 25(1):4-28, 2011.

[29]. A. Schmitt, R. Pieraccini, T. Polzehl, *"For Heaven's Sake, Gimme a Live Person!" Designing Emotion-Detection Customer Care Voice Applications in Automated Call Centers*, Advances in Speech Recognition, 191-219, Springer, 2010.

[30]. H. Atassi, A. Esposito, Z. Sme kal, Emotion recognition from spontaneous Slavic speech, CogInfoCom, Kosice, Slovakia, 389-394, 2012.

[31]. M. Sheikhan, M. Bejani, D. Gharavian, Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method, Neural Computing and Applications, 23 (1):215--227, Springer, 2013.

[32]. V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, R. Prasad, Speech Language & Multimedia Technology, APSIPA ASC 2012, California, USA, 1-4, 2012.

[33]. J. Pohjalainen, P. Alku, Automatic detection of anger in telephone speech with robust autoregressive modulation filtering, ICASSP 2013, Vancouver, BC, Canada, 7537-7541, 2013.

[34]. T. Chaspari, D. Dimitriadis, P. Maragos, Emotion classification of speech using modulation features, EUSIPCO, 2013, Palais de Congres, Marrakech, 1552-1556, 2013.

[35]. S. Gupta, A. Mehra, Viney, Gender specific emotion recognition through speech signals, SPIN 2014, Beijing, China, 727-733, 2014.

[36]. J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, Language resources and evaluation, 39(2-3):165-210, 2005.

[37]. J. Carletta, Assessing agreement on classification tasks: the kappa statistic, Computational linguistics, 22 (2):249-254, 1996

[38]. L. I. Smith, A tutorial on principal components analysis, Cornell University, USA, 51:52, 2002

[39]. N. Halko, P. Martinsson, J. A. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, SIAM review, 5, (2):217-288, 2011.

[40]. T. M. Mitchell, Machine learning, McGraw-Hill, Inc., 1997.

[41]. K. P. Murphy, Machine learning: a probabilistic perspective, MIT press, 2012.

[42]. D. Barber, Bayesian reasoning and machine learning, Cambridge University Press, 2012.

[44]. F. Eyben, M. Wollmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, ACM Multimedia Conference, Firenze, Italy, 1459-1462, 2010.

[45]. P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, H. Sloetjes, Elan: a professional framework for multimodality research, LREC, Genoa, Italy, 2006: 5, 2006