Review 1(koolagudi)

## Εισαγωγη

Τα περισσότερα συστήματα επεξεργάζονται ουδέτερη ομιλία αποτελεσματικά χωρίς να κάνουν αναγνώριση του συναισθήματος της ομιλίας. Αυτό οφείλεται κυρίως στο γεγονός ότι υπάχει δυσκολία στο να μοντελοποιηθεί το συναίσθημα, όπως αναφέρθηκε και παραπάνω.

Σε μια συζήτηση η κατανόηση της κατάστασης του ομιλητή παίζει σημαντικό ρόλο στην ερμηνεία του μηνύματος που εκείνος θέλει να περάσει. Για παράδειγμα η λέξη “ΝΑΙ” μπορεί να έχει πολλές ερμηνείες ανάλογα με τον τρόπο που θα εκφωνηθεί. Άρα η λέξη από μόνη της πολλές φορές δεν φανερώνει την λανθάνουσα σημασιολογία.

Για να αναγνωρίσουμε την σημασιολογία του κειμένου αλλά και την συναισθηματική κατάσταση του ομιλητή κατ' επέκταση μπορούμε να παρατηρήσουμε:

1. Τις εκφράσεις του προσώπου όταν αναλύουμε βίντεο,
2. τις εκφράσεις των συναισθημάτων στην ομιλία όταν αναλύουμε ήχο και
3. την στίξη όταν αναλύουμε γραπό κείμενο

Στην προκειμένη εργασία θα ασχοληθούμε μόνο με ομιλία οπότε θα παρατηρούμε μόνο τις εκφράσεις όσον αφορά την ομιλία κάποιου πομπού.

Η ομιλία είναι μια βασική φυσική ιδιότητα της επικοινωνίας ανθρώπου-μηχανής, οπότε για να φτάσουν οι μηχανές σε ένα επιθυμητό επίπεδο απόδοσης πρέπει να είναι ικανές να αναγνωρίζουν και να συνθέτουν ομιλία που εμπεριέχει συναισθήματα.

Η υλοποίηση τέτοιων συστημάτων, δηλαδή αναγνώρισης ομιλίας, βρίσκει εφαρμογή σε πολλές περιπτώσεις. Αναφέρονταςς μερικά παραδείγματα είναι :

* η αναγνώριση της κατάστασης ενός οδηγού οχήματος ώστε να τον ειδοποιεί κατά την διάρκεια της οδήγησης αλλά και να τον αποτρέπει να πάρει τον έλεγχο του οχήματος εάν χρειαστεί
* η ανάλυση της ομιλίας πελατών σε ένα τηλεφωνικό κέντρο μπορεί να βελτιώσει την ποιότητα της εξυπηρέτησης του πελάτη
* είναι δυνατόν οι ιατροί να διαγνώσουν διαταραχές των ασθενών τους αξιοποιώντα το συναισθηματικό περιεχόμενο της φωνής τους
* η αναγνώρηση των συναισθημάτων,εγκλιματιών μπορεί να βοηθήσει στην εξέλιξη μιας εγκληματολογικής έρευνας
* η ανάλυση των επίγοντων κλήσεων μπορεί να βοηθήσει στην εγκυρότητα της κλήσεις ώστε να ακυρωνονται λανθασμένες αιτήτσεις

## Βάσεις δεδομένων

Για να υλοποιήσουμε ένα σύστημα είτε αναγνώρισης είτε σύνθεσης συναισθημάτων πρέπει πρώτα να εκπαιδεύσουμε το σύστημα με κατάλληλα δεδομένα δηλαδή μια καλή βάση δεδομένων ώστε να γίνεται διαχωρισμός των συναισθημάτων. Οι βάσεις δεδομένων που χρησιμοποιούνται για συστήματα που επεξεργάζονται συναισθήματα ομιλίας μπορούν να χωριστούν σε τρεις κατηγορίες:

1. Βάσεις δεδομένων με ηθοποιούς που υποκρίνονται συναισθήματα (Simulated)
2. Βασεις δεδομένων όπου οι ομιλιτές διεγείρονται ώστε να εκφράσουν μια συναισθηματική κατάσταση (Elicited)
3. Βάσεις δεδομένων φυσικών εκφράσεων όπου οι ομιλητές δεν γνωρίζουν ότι καταγράφονται επομένως οι εκφράσεις τους είναι αμιγείς (Natural)

Στην πρώτη κατηγορία, όπου είναι και η πιο διαδεδομένη (60% των βάσεων είναι τέτοιου τύπου), ζητείται από ηθοποιούς να παράξουν ομιλία συγκεκριμένης συναισθηματικής φόρτισης. Τις περισσότερες φορες η εκφραστηκότητα σε τέτοιες περιπτώσεις είναι πιο έντονη σε σχέση με την πραγματική έκφραση των συναισθημάτων οπότε γίνεται και πιο αξιόπιστος διαχωρισμός των συναισθημάτων.

Στην διεγερμένη ομιλία καταγράφουμε ομιλητές οι οποίοι ωθούνται στην παραγωγή ομιλίας που περιέχει το εκάστοτε συναίσθημα αλλά χωρίς να ξέρουν τον σκοπό της ηχογράφησης. Αυτές οι βάσεις δεδομένων τείνουν να είναι πιο φυσικές αλλά οι καταγραφόμενοι μπορεί να μην έχουν την κατάλληλη εκφραστηκότητα ώστε να γίνει σαφής διαχωρισμός των συναισθημάτων που εκφράζουν αφού ξέρουν ότι ηχογραφούνται οπότε και υπάρχει περίπτωση να μην αποκρίνονται φυσικά στην κάθε διέγερση.

Στην πραγματικότητα η ομιλία κάποιου μπορεί να μην φανερώσει τα συναισθήματά του διότι η έκφρασή του είναι πιο ήπια. Αυτά τα συναισθήματα είναι τα υποκείμενα που απαντούνται και συχνότερα στην καθημερινή ομιλία. Ο σκοπός της τρίτης κατηγορίας είναι να εκπαιδεύσει συστήματα που μπορούν να αναγνωρίσουν τέτοια συναισθήματα. Η καταγραφή των δεδομένων μπορεί να γίνει από τηλεφωνικά κέντρα, από ομιλίες πιλότων σε ασυνήθηστες καταστάεις ή από συζητήσεις ιατρών-ασθενών και συζητήσεις σε δημόσια μέρη εν γένει. Εδώ υπάρχει ξεκάθαρο πρόβλημα κατηγοριοποίησης των συναισθημάτων και είναι ένα εμπόδιο στην εκπαίδευση του συστήματος. Άλλο ένα πρόβλημα είναι αυτό των νομικών θεμάτων που προκύπτουν με την καταγραφή ομιλητών εν αγνοία τους, που είναι η παραβίαση προσωπικών δεδομένων και πνευματικής ιδιοκτησίας, οπότε τέτοιες βάσεις δεδομένων εκτός του ότι είναι δύσκολο να υλοποιηθούν, αποφεύγεται και η χρησιμοποίησή τους.

Τελικά η επιλογή της κατηγορίας της βάσης που θα χρησιμοποιηθεί εξαρτάται από τους σκοπούς της. Επίσης πρέπει να ληφθεί υπόψιν, αν γίνεται επιλογή από τις υπάρχουσες βάσεις, η γλώσσα, οι ομιλιτές και ο αριθμός των συναισθημάτων που πρόκειται να μοντελοποιηθούν. Η επιλογή των καταγραφών, δηλαδή αν θέλουμε υποκρισία η φυσικότητα κρίνει και την ποιότητα της εφαρμογής που θα υλοποιηθεί. Επίσης το μέγεθος της βάσης είναι ένα ζήτημα από το οποίο εξαρτάται η αξιοπιστία και η γενικότητα όσο αναφορα την αναγνώριση συναισθημάτων.

Η μέση απόδοση της αναγνώρισης συναισθημάτων από ακροατές είναι περίπου 80% για τις υπάρχουσες βάσεις δεδομένων ενώ γύρω στο 70% περιλαμβάνουν 4-5 κατηγοριοποιημένα συναισθήματα, δηλαδή αυτά που περιγράφτηκαν και παραπάνω, θυμός, λύπη, χαρά, ουδέτερο κλπ.

## Δεδομένα αναγνώρισης – χαρακτηριστικα ομιλίας

Η επιλογή της σωστής πληροφορίας είναι μια κρίσιμη επιλογή. Διαφορετικά χαρακτηριστικά της ομιλίας μας δίνουν διαφορετική πληροφορία όπως π.χ. ο ομιλητής το συναίσθημα η ηλικία κ.α. Τα χαρακτηριστικά της ομιλίας μπορούν να χωριστούν σε:

* χαρακτηριστικά πηγής διέγερσης (excitation source features)
* χαρακτηριστικά φωνητικού καναλιού (vocal tract features)
* χαρακτηριστικά προσωδίας της φωνής (prosodic features)

Το πρώτο είδος πληροφορίας, έχει να κάνει ουσιαστικά με το σήμα που λαμβάνουμε αφού αφαιρέσουμε το φιλτράρισμα που προκαλεί το φωνητικό κανάλι. Η γνωστή μέθοδος που πετυχαίνει διαχωρισμό της πληροφορίας που μας δίνει η πηγή διέγερσης (δηλαδή η γλωττίδα) από την πληροφορία του φωνητικού καναλιού, είναι η ανάλυση γραμμικής πρόγνωσης (LP analysis – linear prediction analysis) όπου πρώτα υπολογίζουμε τους συντελεστές της γραμμικής πρόγνωσης (linear prediction coefficients) για να προβλέψουμε το σήμα του φωνητικού καναλιού και στην συνέχεια αφαιρούμε το αποτέλεσμα από το αρχικό μας σήμα. Αυτό

που βρίσκουμε είναι το σήμα της πηγής διέγερσης και η διαδικασία με την οποία το υπολογίσαμε είναι η γραμμική πρόγνωση υπολοίπου (linear prediction residual).

Τα χαρακτηριστικά της δραστηριότητας της γλωττίδας, τα οποία λαμβάνουμε παραμετροποιημένα με την ανάλυση της γραμμικής πρόγνωσης υπολοίπου, μπορούν να μας δώσουν πληροφορία σχετικά με το συναίσθημα του ομιλητή. Βέβαια στην μέχρι τώρα έρευνα λίγες είναι οι περιπτώσεις που έχουν χρησιμοποιηθεί αποκλειστικά αυτά τα χαρακτηριστικά και τούτο διότι το σήμα της διέγερσης με την ανάλυση LP φαίνεται σαν σήμα σφάλματος, τα φασματικά χαρακτηριστικά (spectral features) είναι πάρα πολλά και τέλος,, τo LP υπόλοιπο περιέχει συσχετισμούς μεγαλύτερης τάξης όπου δεν είναι διαδεδομένη η χρησιμοποίησή τους.

Παρ' όλο λοιπόν που είναι δύσκολο να παραμετροποιηθεί το LP υπόλοιπο, περιέχει καλή πληροφορία για την πρώιμη διέγερση του φωνητικού καναλιού στην παραγωγή ομιλίας. Έρευνες έχουν δείξει ότι τα χαρακτηριστικά της πηγής διέγερσης εμπεριέχουν ένα μεγάλο κομμάτι της πληροφορίας δηλαδή την ταυτότητα του ομιλητή, το μήνυμα, την γλώσσα αλλά και την πληροφορία σχετικά με το συναίσθημα.

## Βάση Δεδομένων “AIBO”

4 Κύρια συναισθήματα: Θυμός, Εμφατικό, Ουδέτερο, Μητρικό