ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι. ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή εργασία

«Έρευνα χρήσεων τεχνολογιών αναγνώρισης λόγου και φωνητικών εντολών»

Του Φοιτητή
Οικονόμου Μιχαήλ
Α.Μ. 03/2333

Επιβλέπων Καθηγητής
Χρήστος Κουρουπέτρογλου

Θεσσαλονίκη 2010

Πρόλογος

Η αναγνώριση ομιλίας είναι η διαδικασία μετατροπής ενός ηχητικού σήματος, το οποίο έχει προέλθει από μία συσκευή όπως ένα μικρόφωνο ή ένα τηλέφωνο, σε μια σειρά λέξεων. Οι λέξεις αυτές μπορούν να χρησιμοποιηθούν από εφαρμογές ως εντολές ελέγχου ή ως δεδομένα εισόδου σε κάποια άλλη εφαρμογή, για παράδειγμα σε ένα έγγραφο κειμένου. Η αναγνώριση ομιλίας, κάνει ευκολότερη την παραγωγή και την χρήση πληροφοριών. Τα τελευταία χρόνια, τα συστήματα αναγνώρισης ομιλίας αρχίζουν να γίνονται δημοφιλή σε τομείς όπως τα τμήματα εξυπηρέτησης πελατών μεγάλων εταιριών ή σε εταιρίες που παρέχουν τηλεφωνικές υπηρεσίες στους πελάτες τους, όπως οι αεροπορικές. Συστήματα αναγνώρισης ομιλίας που μπορούν να αναγνωρίσουν λίγες λέξεις, όπως για παράδειγμα αριθμούς ή μονολεκτικές απαντήσεις, επιτρέπουν στους χρήστες τους να εισάγουν δεδομένα χωρίς την χρήση πληκτρολογίου. Έτσι αποκλείεται η περίπτωση λανθασμένης πληκτρολόγησης, όμως υπάρχει πάντα ο κίνδυνος λανθασμένης αναγνώρισης. Σήμερα υπάρχουν συστήματα στα οποία ο χρήστης μπορεί να ομιλεί με φυσικό τρόπο, και μπορούν να χρησιμοποιηθούν από χρήστες με προβλήματα όρασης, από χρήστες που θέλουν να υπαγορεύουν κείμενο, εξοικονομώντας έτσι χρόνο αλλά και αποφεύγοντας τον κόπο να το πληκτρολογήσουν, κ.ά. Τα συστήματα αυτά είναι αρκετά ακριβή, όμως υπάρχουν περιθώρια βελτίωσης. Η επιστημονική φαντασία στις ταινίες «Star Trek» και «2001: Η Οδύσσεια του Διαστήματος», όπου ένας υπολογιστής είχε την ικανότητα να καταλαβαίνει την ομιλία των ανθρώπων και να απαντά, σήμερα πλέον είναι πραγματικότητα.

Περίληψη

Η παρούσα εργασία αναφέρεται στην τεχνολογία της αναγνώρισης ανθρώπινης ομιλίας από ένα ηλεκτρονικό υπολογιστή. Στο κείμενο αναλύονται οι χρήσεις της αναγνώρισης ομιλίας, ο τρόπος λειτουργίας των συστημάτων αναγνώρισης ομιλίας, καθώς και άλλα ζητήματα σχετικά με αυτή την τεχνολογία.

Αναλυτικότερα, από το 1960 έως και σήμερα υπάρχει μία αξιοσημείωτη εξέλιξη των τεχνολογιών και των τεχνικών που απετέλεσαν ορόσημα για την βελτίωση των συστημάτων αναγνώρισης ομιλίας. Κάθε μία από αυτές τις εξελίξεις όπως η ανάπτυξη του δυναμικού προγραμματισμού και των Hidden Markov μοντέλων ήταν ένα «λιθαράκι» για την ανάπτυξη των σημερινών συστημάτων.

Υπάρχουν διάφορες κατηγορίες συστημάτων που έχουν ως είσοδο την ανθρώπινη ομιλία, όπως τα συστήματα αναγνώρισης γλώσσας, αναγνώρισης ομιλητή, αναγνώρισης φυσικής ομιλίας κ.ά. Οι εφαρμογές της αναγνώρισης ομιλίας από αυτά τα συστήματα χρησιμοποιούνται σε πολλούς και διαφορετικούς μεταξύ τους τομείς όπως για στρατιωτικούς σκοπούς, από άτομα με αναπηρίες (ΑμεΑ) που δεν μπορούν να χρησιμοποιήσουν εύκολα το πληκτρολόγιο και το ποντίκι, ή για την υπαγόρευση κειμένων που διαφορετικά θα χρειαζόταν περισσότερος κόπος και χρόνος.

Για να μπορέσει μία μηχανή να αναγνωρίσει την ανθρώπινη ομιλία είναι σκόπιμο να εξεταστούν όροι σχετικοί με την ανθρώπινη ομιλία όπως τα φωνήματα, που είναι οι δομικές μονάδες της προφορικής γλώσσας. Επίσης, έννοιες σχετικές με την αναγνώριση ομιλίας, όπως η ακρίβεια και η απόδοση, είναι πολύ βασικές, καθώς από αυτές εξαρτάται αν ένα σύστημα είναι αρκετά καλό ώστε να χρησιμοποιηθεί για εμπορική χρήση. Ενδεικτικά αναφέρεται ότι ένα σύστημα αναγνώρισης ομιλίας, κάτω από ιδανικές συνθήκες μπορεί να έχει ακρίβεια πάνω από 98%. Οι ιδανικές συνθήκες που ακόμα και τα σημερινά συστήματα απαιτούν, είναι η απουσία θορύβου στο χώρο που γίνεται η αναγνώριση, και από τη πλευρά του χρήστη, η καθαρή διατύπωση των λέξεων και των προτάσεων.

Τέλος, η αναγνώριση ομιλίας γίνεται με την χρήση στατιστικών μοντέλων, συγκεκριμένα με Hidden Markov μοντέλα. Η αναγνώριση γίνεται με τον τεμαχισμό του σήματος ομιλίας σε πολύ μικρά τμήματα, μεγέθους 1/100 του δευτερολέπτου, και στη συνέχεια την συνένωση αυτών των τμημάτων ήχου με βάση στατιστικές

πληροφορίες που τα Hidden Markov μοντέλα έχουν μάθει κατά την εκπαίδευσή τους, που είναι το αποτέλεσμα της αναγνώρισης.

Abstract

The present work constitutes a thesis on the technology of recognition of human speech from a computer. In the text the uses of recognition of speech are analyzed, the way of operation of recognition of speech systems, as well as other questions with regard to this technology.

More analytically, from 1960 until today scientists have made a remarkable development of technologies and techniques that constitute landmarks for the improvement of speech recognition systems. Every one of these developments, like the growth of dynamic programming or Hidden Markov Models was very significant for the growth of current systems.

There exist various categories of systems that have as entry the human speech, as the systems that recognize language, the speaker, natural speech and others. The applications of these systems are used in a lot of and different between them sectors as for military aims, for persons with special needs that cannot use easily the keyboard and the mouse, or for the dictation of texts that otherwise would need more labour and time.

In order to a machine be able to recognize the human speech it is deliberate to examine terms relative with the human speech as the phonemes, that are the structural units of oral language. Also, significances relative with speech recognition, as the accuracy and the performance, are very basic, because from them it depends if a system is good enough to be used for commercial use. Indicatively it is reported that a speech recognition system, under ideal conditions can have accuracy above 98%. The ideal conditions that even the current systems require are the absence of noise in the space where the recognition is done, and from the side of user, the clean formulation of words and proposals.

Finally, the recognition of speech becomes with the use of statistical models, concretely the Hidden Markov Models. This is done with the partition of signal of speech in very small departments, size of 1/1000 of second, and then the conjunction of these departments of sound with base statistical information that the Hidden Markov models has learned during their training, and that conjunction is the result of recognition.

Ευχαριστίες

Με την ολοκλήρωση της εργασίας αυτής, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κ. Κουρουπέτρογλου για την ανάθεση του θέματος της μελέτης, καθώς και για την πολύτιμη βοήθεια που μου προσέφερε κατά τη διάρκεια της εκπόνησής της.

Ευχαριστώ ολόψυχα όλους όσους με την υπομονή και την πολύτιμη συμπαράστασή τους, με βοήθησαν στην ολοκλήρωση αυτής της μελέτης.

ΠΕΡΙΕΧΟΜΕΝΑ

Ευρετήριο σχημάτων

Ευρετήριο Πινάκων

Εισαγωγή

Ο σκοπός αυτής της εργασίας είναι να διερευνηθούν οι χρήσεις και ο τρόπος λειτουργίας των συστημάτων αναγνώρισης ομιλίας.

Οι στόχοι που τίθενται για την ικανοποίηση αυτού του σκοπού είναι η δημιουργία λογισμικού που θα παρουσιάζει την μεθοδολογία και τις διαδικασίες που χρειάζονται για να πραγματοποιηθεί αναγνώριση ομιλίας από ένα browser. Ακόμα, να αναφερθούν οι εφαρμογές της αναγνώρισης ομιλίας και πως μπορούν κάποιες εφαρμογές να είναι χρήσιμες σε άτομα με αναπηρίες. Τέλος, στόχος της εργασίας είναι να αναλυθεί ο τρόπος λειτουργίας των συστημάτων αναγνώρισης ομιλίας με όσο το δυνατό κατανοητότερο τρόπο.

Στο πρώτο κεφάλαιο γίνεται μία αναδρομή στην εξέλιξη της τεχνολογίας της αναγνώρισης ομιλίας από την δεκαετία του ’60 μέχρι σήμερα. Παρατίθενται για κάθε δεκαετία οι τεχνολογικές ανακαλύψεις που υιοθετήθηκαν για να μπορεί να γίνει πράξη η αναγνώριση ομιλίας από μηχανές. Παράλληλα, παρατίθενται και οι ικανότητες των συστημάτων καθώς η τεχνολογία εξελισσόταν. Ακόμα, αναφέρονται έρευνες που γίνονται με σκοπό την βελτίωση της τεχνολογίας αναγνώρισης ομιλίας. Τέλος, παρατίθενται και κάποια πιθανά σενάρια για τις χρήσεις της αναγνώρισης ομιλίας στο μέλλον.

Υπάρχουν αρκετά συστήματα που χρησιμοποιούν την ανθρώπινη ομιλία ως είσοδο για να παράξουν αποτελέσματα όπως: αναγνώριση ομιλίας, αναγνώριση χρήστη, αναγνώριση συναισθήματος, αναγνώριση γλώσσας. Όλες αυτές οι κατηγορίες αναφέρονται στο δεύτερο κεφάλαιο, μαζί με μία σύντομη περιγραφή για τον τρόπο λειτουργίας τους καθώς και για τους τομείς που χρησιμοποιούνται.

Στο τρίτο κεφάλαιο, αναφέρονται πολλές από τις χρήσεις που η αναγνώριση ομιλίας έχει σήμερα. Χρησιμοποιείται σε τηλεφωνικά κέντρα, από άτομα με ειδικές ανάγκες, για στρατιωτικούς σκοπούς, και γενικότερα για έλεγχο λογισμικού. Σε πολλές από τις χρήσεις αναφέρονται και παραδείγματα που η χρήση αναγνώριση ομιλίας είχε οφέλη.

Στο τέταρτο κεφάλαιο, γίνεται μία αναφορά σε ζητήματα σχετικά με τα συστήματα αναγνώρισης ομιλίας, όπως τα λεξικά, η εκπαίδευση, η εξάρτηση από τους χρήστες και η ακρίβεια. Γενικά, σε αυτό το κεφάλαιο αναφέρονται

χαρακτηριστικά των συστημάτων. Τέλος, αναφέρονται και διάφοροι παράγοντες που επηρεάζουν την ακρίβεια.

Στο πέμπτο κεφάλαιο, παρουσιάζεται ο τρόπος λειτουργίας των συστημάτων αναγνώρισης ομιλίας. Για να γίνει αυτό, επεξηγείται η έννοια των φωνημάτων και αναφέρονται βασικές μαθηματικές τεχνικές, όπως τα Hidden Markov Models, που η κατανόηση τους αποτελεί βασικό παράγοντα για την κατανόηση του τρόπου λειτουργίας των συστημάτων αναγνώρισης ομιλίας. Ακόμα, χρησιμοποιείται μαθηματικός φορμαλισμός, για την κομψότερη εξήγηση των διαδικασιών που παίρνουν μέρος στην αναγνώριση, καθώς γίνεται και εκτενής αναφορά στον τρόπο που ένα σύστημα επεξεργάζεται την ανθρώπινη ομιλία για να την αναγνωρίσει.

Κεφάλαιο 1. Ιστορική εξέλιξη των συστημάτων αναγνώρισης ομιλίας

Εισαγωγή

Η επιθυμία του ανθρώπου να κατασκευάσει μία μηχανή ικανή να αναγνωρίζει ομιλία δεν είναι σύγχρονη. Τουλάχιστον πριν από 120 χρόνια, οι άνθρωποι ονειρεύονταν μία τέτοια μηχανή. Για παράδειγμα, το 1881, ο *Alexander Graham Bell*, μαζί με τον εξάδελφο του *Chichester Bell* και τον *Charles Sumner Tainter* εφηύραν μία συσκευή η οποία χρησιμοποιούσε ένα περιστρεφόμενο κύλινδρο, επάνω στον οποίο υπήρχε μία επίστρωση από κερί. Μία γραφίδα παρόμοια με αυτή ενός γραμμοφώνου δημιουργούσε αυλακώσεις επάνω στο κερί με βάση τον ήχο που ακουγόταν. Από τότε μέχρι σήμερα, η τεχνολογία της αναγνώρισης ομιλίας από μία μηχανή έχει σημειώσει αλματώδη εξέλιξη. Στο πρώτο κεφάλαιο αυτής της εργασίας, παρουσιάζεται η εξέλιξη της τεχνολογίας της αναγνώρισης ομιλίας από το 1960, όταν και η προσπάθεια αυτή έγινε πιο συστηματική, έως και σήμερα και οι έρευνες που γίνονται με σκοπό την εξέλιξη της.

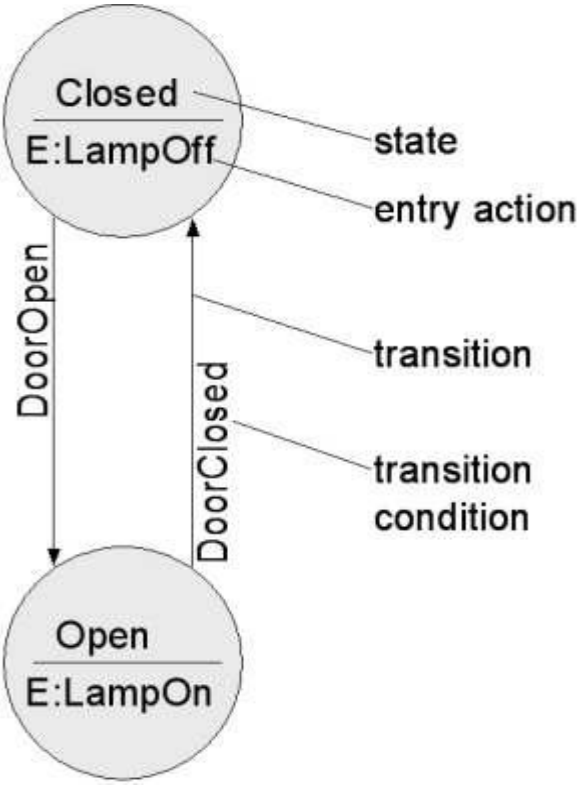
1.1. Εξέλιξη της τεχνολογίας αναγνώρισης ομιλίας

Η εικόνα 2 (Huang and Rabiner, 2004) παρουσιάζει την ιστορία της εξέλιξης της τεχνολογίας αναγνώρισης ομιλίας. Βλέπουμε ότι την δεκαετία του 1960 τα συστήματα είχαν μικρά λεξικά αποτελούμενα από 10-100 εκφράσεις. Αναγνώριση μπορούσε να γίνει μόνο σε απομονωμένες εκφράσεις, και όχι σε συνδυασμό εκφράσεων. Η αναγνώριση βασιζόταν σε απλές ακουστικές ιδιότητες των ήχων των λέξεων. Οι τεχνολογίες που αναπτύχθηκαν εκείνη την περίοδο και βοήθησαν στην εξέλιξη της αναγνώρισης ομιλίας είναι η ανάλυση *filter-bank*, με την οποία ένα ηχητικό σήμα χωρίζεται σε τμήματα με την βοήθεια φίλτρων ακουστικών συχνοτήτων, και η μεθοδολογία του δυναμικού προγραμματισμού (*dynamic programming*), που επιτρέπει την επίλυση ενός σύνθετου προβλήματος, κατακερματίζοντας το σε μικρότερα.

Στην δεκαετία του 1970 τα συστήματα αναγνώρισης ομιλίας ήταν ικανά να αναγνωρίσουν εκφράσεις από μεσαίου μεγέθους λεξικά, αποτελούμενα από 100-1000 εκφράσεις. Επίσης, μπορούσαν να αναγνωρίσουν αριθμούς. Πλέον στο αποτέλεσμα της αναγνώρισης γινόταν επεξεργασία. Αναπτύχθηκαν διάφορες τεχνολογίες που βοήθησαν σε αυτή την εξέλιξη. Η αναγνώριση προτύπων *(pattern recognition*), που ταξινομεί δεδομένα με βάση στατιστικές πληροφορίες που έχει εξαγάγει από άλλα πρότυπα, χρησιμοποιήθηκε ώστε να ταξινομούνται ακουστικές εκφράσεις. Η εισαγωγή της τεχνικής *Linear Predictive Coding* (LPC), για την αναπαράσταση του ακουστικού φάσματος συχνοτήτων ενός ψηφιακού σήματος φωνής, έδωσε την δυνατότητα δημιουργίας φασματογράφων. Η τεχνική *Pattern Clustering*, έδινε την δυνατότητα ομαδοποίησης προτύπων που είχαν σχέση μεταξύ τους. Αυτό βοήθησε στην δημιουργία συστημάτων αναγνώρισης ομιλίας που δεν ήταν εξαρτημένα από τον ομιλητή.

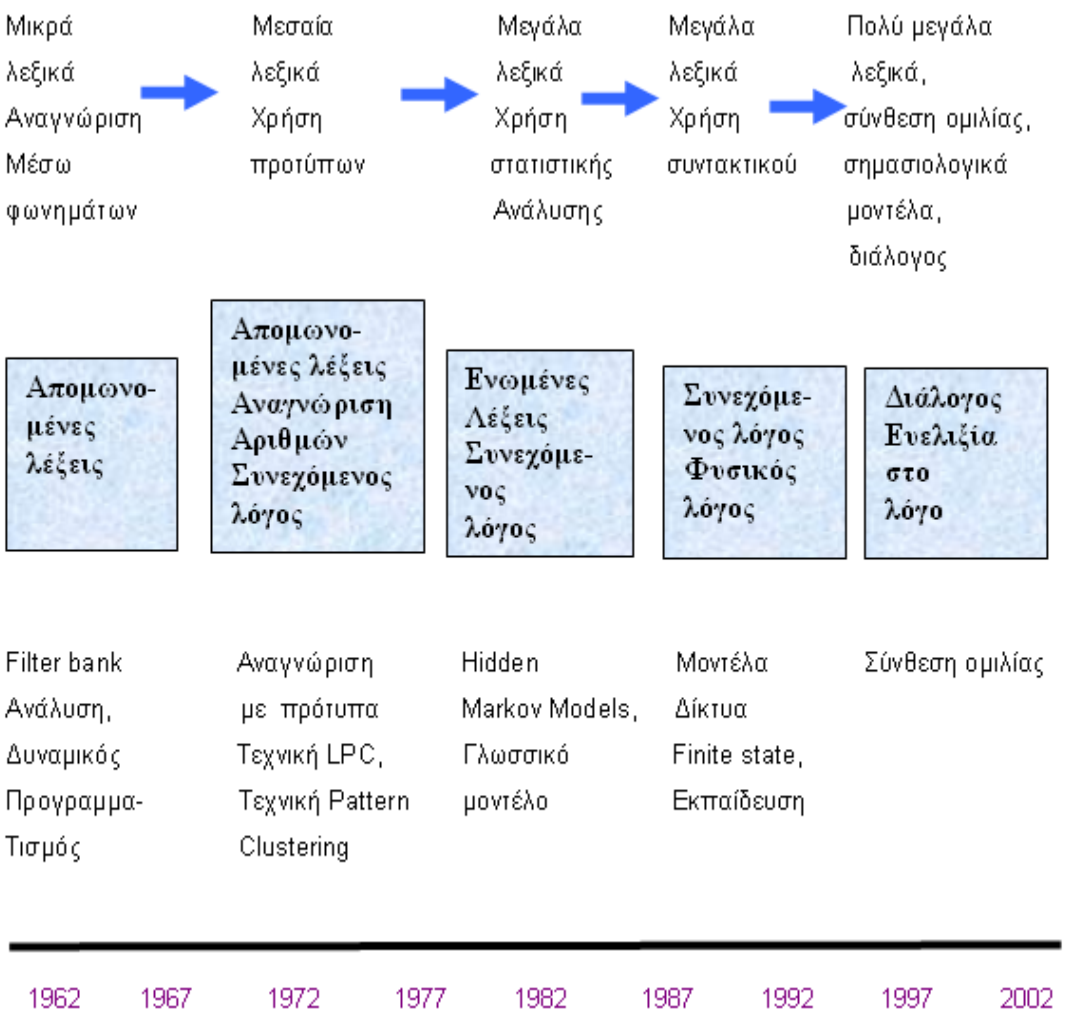
Στην δεκαετία του 1980 τα συστήματα αναγνώρισης ομιλίας μπορούσαν να διαχειριστούν λεξικά που το μέγεθος τους ήταν μεγαλύτερο από 1000 εκφράσεις. Σε αυτό συνέβαλε η χρησιμοποίηση των *Hidden Markov Models* (HMM) και η ανακάλυψη της μεθόδου του γλωσσικού μοντέλου (*language model*), ο συνδυασμός των οποίων προσφέρει μία πολύ ισχυρή μεθοδολογία ικανή να χειριστεί αποτελεσματικά και έγκυρα το ζήτημα της αναγνώρισης ομιλίας. Οι τεχνικές αυτές αναλύονται σε επόμενο κεφάλαιο.

Την δεκαετία του 1990 τα συστήματα αναγνώρισης ομιλίας μπορούσαν να χειριστούν μεγάλα λεξικά με πολλές χιλιάδες εκφράσεις να περιέχονται σε αυτά. Εκείνη την περίοδο αναπτύχθηκε η τεχνολογία της εκπαίδευσης του γλωσσικού και του ακουστικού μοντέλου. Επίσης η ανάπτυξη τεχνικών που χρησιμοποιούσαν *Finite-state* δίκτυα και οι μέθοδοι για ελαχιστοποίηση των καταστάσεων αυτών των δικτύων, συντέλεσαν στην δημιουργία μεγάλων λεξικών. Τα δίκτυα Finite-state περιγράφουν την συμπεριφορά ενός συστήματος, και αποτελούνται από ένα πεπερασμένο αριθμό καταστάσεων (*states*), οι οποίες περιέχουν πληροφορίες σχετικά με το τι έχει συμβεί μέχρι εκείνη την στιγμή στο σύστημα, μεταβάσεων (*transitions*) από μία κατάσταση σε μία άλλη, και δράσεων (*actions*) που πραγματοποιούνται στον κατάλληλο χρόνο. Η εικόνα 1 (finite state machine, www.stateworks.com/technology/finite_state_machine) παρουσιάζει ένα Finite-state δίκτυο. Τα δίκτυα Finite-state χρησιμοποιούνται στο γλωσσικό μοντέλο.



Εικόνα 1. Ένα finite-state δίκτυο

Σήμερα, τα λεξικά είναι πολύ μεγάλου μεγέθους σε συνδυασμό με σημασιολογικά μοντέλα που περιγράφουν τις εκφράσεις που περιέχουν. Ακόμα, τα συστήματα αναγνώρισης ομιλίας έρχονται μαζί με την ικανότητα σύνθεσης φωνής (*Text to Speech*). Τα συστήματα σήμερα υποστηρίζουν δυνατότητα διαλόγου ο οποίος περιέχει διαφορετικές κλίσεις ρημάτων, εγκλίσεις (φαινόμενο κατά το οποίο μερικές μονοσύλλαβες συνηθισμένες λέξεις προφέρονται τόσο στενά με την προηγούμενή τους, ώστε ο τόνος τους ή χάνεται εντελώς ή μετακινείται στη λήγουσα της προηγούμενης λέξης π.χ. σαν να-τανε), πράγμα που τα κάνει εύκολα στη χρήση και ικανά να χρησιμοποιηθούν σε απαιτητικά περιβάλλοντα. Η ικανότητα αντίληψης φυσικού λόγου (δηλαδή του τρόπου ομιλίας ενός ανθρώπου με έναν άλλο άνθρωπο), η δυνατότητα εκπαίδευσης για την βελτίωση του αποτελέσματος, καθώς και η δυνατότητα ελέγχου λογισμικού όταν ο χρήστης το επιλέγει, κάνουν τα συστήματα αναγνώρισης ομιλίας πιο προσιτά από ποτέ.



Εικόνα 2. Ορόσημα στην τεχνολογία αναγνώρισης ομιλίας

1.2. Ερευνητικά θέματα που βρίσκονται σε εξέλιξη

Υπάρχουν ζητήματα στην τεχνολογία της αναγνώρισης ομιλίας που χρήζουν περαιτέρω μελέτης και έρευνας. Διάφορα φιλόδοξα ερευνητικά προγράμματα θα δώσουν ώθηση στην τεχνολογία της κατανόησης και αναγνώρισης ομιλίας δημιουργώντας νέες εφαρμογές. Τα πεδία της έρευνας που πραγματοποιείται (Baker J, et al., 2007) αναφέρονται παρακάτω.

1.2.1. Αναγνώριση σε συνηθισμένο ηχητικό περιβάλλον

Οι άνθρωποι έχουν συνηθίσει να αναγνωρίζουν ομιλία σε περιβάλλοντα όπου υπάρχει θόρυβος από την ομιλία άλλων ανθρώπων, από ηλεκτρικές συσκευές,

από τον δρόμο κ.ά. Η ακρίβεια των σημερινών συστημάτων αναγνώρισης ομιλίας σε ένα τέτοιο περιβάλλον, μειώνεται σημαντικά.

Γίνονται έρευνες ώστε να δημιουργηθούν προγράμματα αναγνώρισης ομιλίας που θα είναι ανθεκτικότερα και ευπροσάρμοστα σε ηχητικά περιβάλλοντα και να μην επηρεάζονται από αντηχήσεις, εξωτερικούς θορύβους, θορύβους που δημιουργούνται από μαγνητικά κύματα όπως των κινητών τηλεφώνων, από τα χαρακτηριστικά της φωνής του ομιλητή (χροιά, τρόπος ομιλίας, συναισθηματική φόρτιση), και χαρακτηριστικά της γλώσσας (διάλεκτος, λεξιλόγιο κ.ά.). Τα προγράμματα αναγνώρισης ομιλίας θα μπορούν να προσαρμόζονται στο ηχητικό περιβάλλον που βρίσκονται.

1.2.2. Εύκολη ενσωμάτωση μη δημοφιλών γλωσσών

Σήμερα, τα συστήματα αναγνώρισης ομιλίας δουλεύουν με την δημιουργία πολύπλοκων ακουστικών και γλωσσικών μοντέλων χρησιμοποιώντας μεγάλες συλλογές από ομιλίες και κείμενα κάποιας συγκεκριμένης γλώσσας. Έτσι, κάποια γλώσσα η οποία δεν χρησιμοποιείται από μεγάλο τμήμα του πληθυσμού, δεν μπορεί να χρησιμοποιηθεί στα συστήματα αναγνώρισης ομιλίας.

Οι έρευνες που είναι σε εξέλιξη εξετάζουν τον τρόπο με τον οποίο θα μπορεί ένα σύστημα αναγνώρισης ομιλίας να ενσωματώνει εύκολα μία γλώσσα. Για να γίνει αυτό χρειάζονται να μελετηθούν οι γλωσσικές και ακουστικές μονάδες που έχουν κοινές οι γλώσσες μεταξύ τους.

1.2.3. Δυνατότητα προσαρμογής σε μία γλώσσα

Τα συστήματα αναγνώρισης ομιλίας βασίζονται σε στατιστικά μοντέλα, που δημιουργούνται από κείμενα και λεξικά που περιέχουν την προφορά των λέξεων. Πολλές φορές αυτή η γνωστική βάση δεδομένων χρειάζεται επέκταση, για να καλύψει εκφράσεις που δεν έχουν συμπεριληφθεί από την αρχή. Για να γίνει αυτό, χρειάζεται το σύστημα να εκπαιδευτεί ξανά από τον χρήστη. Αυτό έρχεται σε αντίθεση με την ικανότητα εκμάθησης νέων λέξεων, ιδιωματισμών στην γλώσσα, διαφορετικών προφορών από τον άνθρωπο, που ενσωματώνει αυτή την γνώση σε όλη την διάρκεια της ζωής του με ευκολία.

Οι έρευνες που γίνονται αποσκοπούν στην ανάπτυξη συστημάτων που θα ενσωματώνουν νέες εκφράσεις με αυτόματο τρόπο. Θα μπορεί να υπάρχει η ικανότητα εκπαίδευσης από αρχεία video, ή από νέα και άγνωστα αρχεία κειμένου.

1.2.4. Αναγνώριση μίας έγκυρης λέξης

Τα σημερινά συστήματα αναγνώρισης ομιλίας αντιμετωπίζουν δυσκολίες όταν πρέπει να αναγνωρίσουν μία λέξη η οποία δεν βρίσκεται στο λεξικό τους. Το αποτέλεσμα της αναγνώρισης είναι λέξεις που ηχούν παρόμοια.

Ο σκοπός της έρευνας που γίνεται είναι να δημιουργηθούν συστήματα που να αναγνωρίζουν πότε δεν ξέρουν μία σωστή και έγκυρη λέξη. Ένα στοιχείο για την αναγνώριση τέτοιων λέξεων είναι όταν δεν υπάρχει αντιστοιχία μεταξύ της αναγνώρισης του ηχητικού σήματος και των υποθέσεων που παράγει το γλωσσικό μοντέλο. Σε αυτή την έρευνα θα πρέπει να δημιουργηθούν ακριβή μοντέλα μέτρησης της αβεβαιότητας, βασισμένα στην αναντιστοιχία μεταξύ των δεδομένων που λαμβάνονται από την αναγνώριση μίας λέξης, και της γνωστικής βάσης του συστήματος.

Επίλογος

Σε πρώτο κεφάλαιο έγινε μία ανασκόπηση της εξέλιξης της τεχνολογίας της αναγνώρισης ομιλίας. Είδαμε ότι στην δεκαετία του ’60 άρχισε μία συστηματική προσπάθεια για την δημιουργία μίας μηχανής ικανής να αναγνωρίζει την ανθρώπινη ομιλία. Τα συστήματα τότε ήταν ικανά να αναγνωρίσουν πολύ λίγες λέξεις, 10-100 το πολύ, και δεν μπορούσαν να αναγνωρίσουν συνδυασμούς λέξεων. Την δεκαετία του ‘70 εισήχθη η τεχνική της αναγνώρισης προτύπων, πράγμα που επέτρεψε την δημιουργία μεγαλύτερων λεξικών, έως και 1000 λέξεων. Την δεκαετία του ΄80 για την αναγνώριση προτύπων χρησιμοποιήθηκαν τα Hidden Markov Models, πράγμα που επέτρεψε την αναγνώριση λέξεων από λεξικά μεγαλύτερα των 1000 λέξεων. Ακόμα, η ανακάλυψη της χρήσης του γλωσσικού μοντέλου, έκανε τα συστήματα ικανά να αναγνωρίσουν ευκολότερα συνδυασμούς λέξεων. Αυτές είναι πολύ σημαντικές τεχνικές που αποτελούν ακόμα και σήμερα τον πυρήνα των συστημάτων αναγνώρισης ομιλίας. Την δεκαετία του

’90, τα λεξικά μεγάλωσαν και άλλο, δημιουργήθηκαν συστήματα που ήταν ικανά να αναγνωρίσουν φυσικό λόγο και που μπορούσαν να εκπαιδευτούν από τον χρήστη. Τα σημερινά συστήματα είναι πιο ακριβή από ποτέ, διαθέτουν πολύ μεγάλα λεξικά, είναι ικανά να αναγνωρίσουν ιδιωματισμούς, και είναι καλύτερα στην αναγνώριση φυσικού λόγου από τους προκατόχους τους.

Επειδή όμως η τεχνολογία της αναγνώρισης ομιλίας δεν έχει ακόμα επιτύχει την μέγιστη ακρίβεια, γίνονται διάφορες έρευνες που την βελτιώσουν. Οι έρευνες που γίνονται αποσκοπούν στο να κάνουν τα συστήματα αναγνώρισης ικανά να αναγνωρίζουν ομιλία σε περιβάλλοντα με συνηθισμένο θόρυβο όπως ένα γραφείο, πράγμα που ο μέσος άνθρωπος κάνει με ευκολία. Ακόμα, γίνονται έρευνες ώστε ένα σύστημα αναγνώρισης ομιλίας να μπορεί να διακρίνει όπως και ένας άνθρωπος αν μία λέξη είναι έγκυρη ή όχι, και ώστε να μπορούν να επεκτείνουν τις λέξεις που γνωρίζουν χωρίς να χρειάζεται κάποια επίπονη διαδικασία για τον χρήστη, όπως η εκπαίδευση, ακριβώς όπως και ένας άνθρωπος μαθαίνει με σχετική ευκολία νέες λέξεις. Αυτές οι έρευνες, ίσως καταφέρουν να κάνουν την αναγνώριση ομιλίας ακόμα πιο δημοφιλή στα χρόνια που έρχονται.

Τα συστήματα αναγνώρισης ομιλίας είναι δημοφιλή σήμερα, και μπορούν να χρησιμοποιηθούν σε ένα μεγάλο εύρος εφαρμογών, ανάλογα με το περιβάλλον που χρησιμοποιούνται. Οι κατηγορίες συστημάτων που χρησιμοποιούν ομιλία παρουσιάζονται στο δεύτερο κεφάλαιο.

Κεφάλαιο 2. Κατηγορίες συστημάτων που έχουν ως είσοδο την ανθρώπινη ομιλία

Εισαγωγή

Τα συστήματα αναγνώρισης ομιλίας, αλλά και άλλα συναφή συστήματα που επεξεργάζονται το ακουστικό διάνυσμα της ανθρώπινης ομιλίας, το οποίο δημιουργείται μετά από ειδική επεξεργασία, ανάλογα με την χρήση τους στο περιβάλλον στο οποίο βρίσκονται, και τον τρόπο λειτουργίας τους χωρίζονται σε διάφορες κατηγορίες. Σε αυτό το κεφάλαιο παρουσιάζονται οι εφαρμογές των συστημάτων που έχοντας ως είσοδο την ανθρώπινη ομιλία, με κατάλληλη επεξεργασία, παράγουν αποτελέσματα διαφορετικά μεταξύ τους.

2.1. Συστήματα αναγνώρισης ομιλίας

Τα συστήματα αναγνώρισης ομιλίας χωρίζονται σε δύο κατηγορίες. Σε αυτά που μπορούν να αναγνωρίσουν φυσικό λόγο, δηλαδή τον τρόπο που επικοινωνεί ένας άνθρωπος με έναν άλλο, και αυτά που για να αναγνωρίσουν μία έκφραση, ο χρήστης πρέπει να την προφέρει με ειδικό τρόπο. Ακόμα, υπάρχει και ένας ακόμη διαχωρισμός. Υπάρχουν συστήματα που αποδίδουν καλύτερα με ένα συγκεκριμένο χρήστη, ενώ άλλα μπορούν να αποδώσουν και με διαφορετικούς ομιλητές. Όλα αυτά εξηγούνται παρακάτω.

2.1.1. Συστήματα αναγνώρισης απομονωμένων λέξεων

Τα συστήματα αναγνώρισης απομονωμένων λέξεων (isolated words recognition systems) απαιτούν συνήθως μεταξύ κάθε φωνήματος να υπάρχει ησυχία (έλλειψη ηχητικού σήματος) .

Συχνά, αυτά τα συστήματα διαχωρίζονται σε καταστάσεις Ενεργό/μη Ενεργό (Listen/not Listen), και απαιτούν ο ομιλητής να περιμένει μεταξύ των εκφράσεων (συνήθως επεξεργάζονται το φώνημα κατά τη διάρκεια των μικρών διακοπών).

Είναι συστήματα που έχουν ξεπεραστεί τεχνολογικά, αφού πλέον δεν υπάρχει ανάγκη για παύσεις μεταξύ των εκφράσεων.

2.1.2. Συστήματα αναγνώρισης φυσικής ομιλίας

Τα συστήματα αναγνώρισης φυσικής ομιλίας ή φυσικού λόγου είναι συστήματα που αναγνωρίζουν φυσικό ήχο και όχι ειδικά εκφερόμενο. Ένα σύστημα με αναγνώριση φυσικής ομιλίας πρέπει να είναι σε θέση να χειριστεί ποικίλα φυσικά λεκτικά χαρακτηριστικά γνωρίσματα όπως λέξεις που ακούγονται ή προφέρονται παρόμοια. Τα τρέχοντα συστήματα αναγνώρισης αντιμετωπίζουν κάποια προβλήματα με την αναγνώριση φυσικής ομιλίας. Αυτά τα προβλήματα οφείλονται σε φτωχή άρθρωση, αυξημένη συνάρθρωση, μεγάλη μεταβολή στον ρυθμό ομιλίας καθώς και πολλούς άλλους παράγοντες όπως δισταγμοί, λανθασμένα ξεκινήματα και διορθώσεις. Συνήθως λοιπόν, ο ομιλητής μπορεί να βελτιώσει την ακρίβεια τους, απλά προσέχοντας τον τρόπο ομιλίας του. Λύσεις στα προβλήματα αυτά ίσως να απαιτήσουν ουσιώδεις επεκτάσεις στις υπάρχουσες τεχνικές αναγνώρισης.

2.1.2.1. Εξάρτηση από τους χρήστες

Τα συστήματα αναγνώρισης φυσικής ομιλίας διαιρούνται σε κατηγορίες ανάλογα με το αν εξαρτώνται ή όχι από τον χρήστη. Τα συστήματα που είναι σχεδιασμένα να αποδίδουν με έναν συγκεκριμένο ομιλητή (speaker dependent systems) είναι ακριβέστερα για το σωστό ομιλητή, αλλά πολύ λιγότερο ακριβή για άλλους ομιλητές. Τα συστήματα αυτά «μαθαίνουν» τα ιδιαίτερα χαρακτηριστικά της φωνής ενός συγκεκριμένου ομιλητή και έτσι προσαρμόζονται σε αυτόν. Ένας νέος χρήστης πρέπει να εκπαιδεύσει πρώτα το σύστημα μιλώντας του, έτσι ώστε να αναλυθεί ο τρόπος που ομιλεί. Αυτό σημαίνει ότι ο χρήστης πρέπει να διαβάσει μερικές σελίδες κειμένου στο σύστημα πριν το χρησιμοποιήσει.

Στα συστήματα αναγνώρισης ομιλίας που μπορούν να λειτουργήσουν με διαφορετικούς ομιλητές (speaker independent systems) δεν είναι απαραίτητη η διαδικασία της εκπαίδευσης. Το μειονέκτημα τους σε σχέση με τα συστήματα που εξαρτώνται από τον χρήστη είναι η μικρότερη ακρίβεια τους. Τα συστήματα αυτά

συνήθως έχουν μικρότερα λεξικά. Η χρήση μικρότερων λεξικών κάνει πιο πιθανό να είναι σωστό το αποτέλεσμα της αναγνώρισης (βέβαια υστερούν στο ότι ο χρήστης μπορεί να πει μία έκφραση που να μην υπάρχει στο λεξικό).

Τα συστήματα που εξαρτώνται από τον ομιλητή χρησιμοποιούνται περισσότερο σε εφαρμογές όπου γίνεται υπαγόρευση (dictation), επειδή είναι σημαντική η ακρίβεια καθώς και η ανάγκη για λεξικά που περιέχουν πολλές εκφράσεις. Τα συστήματα που δεν εξαρτώνται από τον ομιλητή χρησιμοποιούνται σε εφαρμογές που χρησιμοποιούν πολλοί διαφορετικοί χρήστες, και δεν είναι δυνατή η εκπαίδευση, όπως φωνητικές υπηρεσίες που παρέχονται μέσω τηλεφώνου.

2.2. Συστήματα αναγνώρισης ομιλητή

Η αναγνώριση του ομιλητή (speaker recognition) είναι η διαδικασία κατά την οποία αναγνωρίζεται ο ομιλητής με βάση τα μοναδικά χαρακτηριστικά της φωνής του. Τα χαρακτηριστικά τα οποία εξάγονται μπορούν να δώσουν πληροφορίες όχι μόνο για τον τρόπο ομιλίας (στυλ ομιλίας, τόνος της φωνής, χροιά) αλλά και για την ανατομία του ομιλητή όπως το μέγεθος και το σχήμα της στοματικής κοιλότητας.

Υπάρχουν δύο κατηγορίες συστημάτων αναγνώρισης ομιλητή. Αυτά που χρησιμοποιούνται για αναγνώριση (identification), και αυτά που χρησιμοποιούνται για ταυτοποίηση (authentication). Η αναγνώριση είναι η διαδικασία κατά την οποία αναγνωρίζεται η ταυτότητα ενός άγνωστου ομιλητή. Για παράδειγμα, η αναγνώριση ενός κακοποιού με βάση την φωνή του που έχει καταγραφεί από κλειστό κύκλωμα. (Αν υποθέσουμε ότι η αστυνομία διαθέτει βάσεις δεδομένων με φωνές κακοποιών). Η ταυτοποίηση είναι η διαδικασία κατά την οποία εξακριβώνεται η ταυτότητα ενός χρήστη, ο οποίος δηλώνει ότι είναι δική του. Για παράδειγμα, σε ένα σύστημα διαχείρισης τραπεζικού λογαριασμού, ο χρήστης δηλώνει τα στοιχεία του και στην συνέχεια ελέγχεται η φωνή του για να εξακριβωθεί αν είναι όντως ο κάτοχος του λογαριασμού.

Τα συστήματα αναγνώρισης ομιλητή χρησιμοποιούνται σε εφαρμογές όπως έλεγχος και χρήση τραπεζικού λογαριασμού (telephone banking), αγορές μέσω τηλεφώνου (telephone shopping), υπηρεσίες που απαιτούν αυθεντικοποίηση ώστε

να επιτραπεί η πρόσβαση σε βάσεις δεδομένων, φωνητικό ηλεκτρονικό ταχυδρομείο (voice mail) και απομακρυσμένο έλεγχο υπολογιστών με χρήση φωνητικών εντολών.

2.3. Συστήματα αναγνώρισης γλώσσας

Τα συστήματα αναγνώρισης της γλώσσας του ομιλητή (language identification), αναγνωρίζουν την γλώσσα του ομιλητή με βάση την ομιλία του. Πολλές προσεγγίσεις για να επιτευχθεί η αναγνώριση γλώσσας βασίζονται σε τεχνικές που χρησιμοποιούνται στα συστήματα αναγνώρισης ομιλίας και συγκεκριμένα σε αυτά που δεν εξαρτώνται από τον ομιλητή. Μία δημοφιλής προσέγγιση αποτελείται από παραλλαγές των εξής δύο βημάτων. Πρώτο, δημιουργία μίας μηχανής αναγνώρισης ομιλίας για κάθε γλώσσα, και δεύτερο, σύγκριση της πιθανότητας που δίνει κάθε μηχανή αναγνώρισης, και επιλογή αυτής με την μεγαλύτερη. Δηλαδή, αυτό που γίνεται είναι να περνά το ακουστικό διάνυσμα από κάθε μηχανή αναγνώρισης, και η μηχανή αναγνώρισης που επιστρέφει την μεγαλύτερη πιθανότητα εκφράζει τη γλώσσα του ομιλητή.

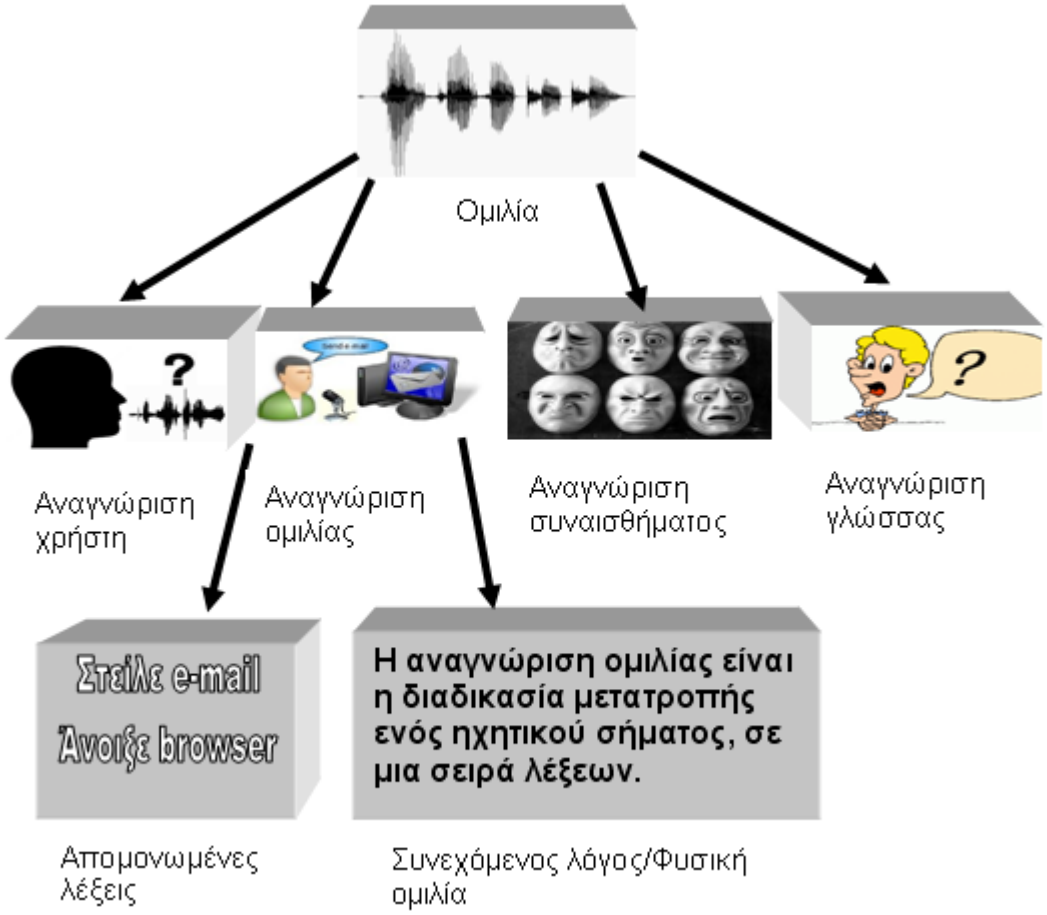
Σήμερα τα καλύτερα συστήματα αναγνώρισης γλώσσας μπορούν να επιτύχουν ακρίβεια 79%, μεταξύ 11 γλωσσών. (Muthusamy and Spitz, 1997)

Τα συστήματα αναγνώρισης γλώσσας μπορούν να χρησιμοποιηθούν σε τηλεφωνικά κέντρα, όπου μπορεί να αναγνωρίζεται η γλώσσα του καλούντα, και να προωθείται στον κατάλληλο τηλεφωνητή. Ακόμα, ένα σύστημα που κάνει μετάφραση των λεγομένων του χρήστη σε μία άλλη γλώσσα μπορεί να αναγνωρίσει την γλώσσα του χρήστη, και να κάνει την μετάφραση στη γλώσσα που επιθυμεί.

2.4. Συστήματα αναγνώρισης του συναισθήματος

Τα συστήματα αυτόματης αναγνώρισης συναισθήματος (automatic emotion identification), χρησιμοποιούν πληροφορίες από την φωνή του ομιλητή, όπως την ένταση, το ύψος, δισταγμούς στην ομιλία, π.χ. για τα ελληνικά: χμμ...., εεεε.... για να αναγνωρίσουν την συναισθηματική κατάσταση του ομιλητή.

Τα συστήματα αυτά είναι χρήσιμα διότι με τη συνεχώς αυξανόμενη παρουσία αυτόματων συστημάτων στην καθημερινότητά μας, εισέρχεται και το βάρος της αλληλεπίδρασης με αυτά τα συστήματα εξαιτίας της έλλειψης συναισθηματικής νοημοσύνης από την πλευρά των μηχανών. Η συναισθηματική πληροφορία που μεταδίδεται μέσω της ανθρώπινης ομιλίας αποτελεί σημαντικό παράγοντα στις ανθρώπινες επικοινωνίες και αλληλεπιδράσεις. (Μιχαλέτου, 2008)



Εικόνα 3. Κατηγορίες συστημάτων ομιλίας

Επίλογος

Στο δεύτερο κεφάλαιο παρουσιάστηκαν οι διαφορετικές κατηγορίες συστημάτων που χρησιμοποιούν ως είσοδο το ακουστικό διάνυσμα της ανθρώπινης ομιλία για να παράξουν αποτελέσματα. Τα συστήματα αυτά είναι τα συστήματα που αναγνωρίζουν ομιλία (speech recognition), και ο σκοπός τους είναι να επιστρέφουν το αποτέλεσμα της αναγνώρισης της ομιλίας του χρήστη. Τα

συστήματα αναγνώρισης ομιλίας χωρίζονται σε αυτά που είναι ικανά να αναγνωρίσουν φυσική ομιλία, δηλαδή τον τρόπο που οι άνθρωποι ομιλούν μεταξύ τους, και σε αυτά που μπορούν να αναγνωρίσουν απομονωμένες λέξεις. Επίσης, ένα περαιτέρω διαχωρισμός των συστημάτων αναγνώρισης ομιλίας είναι αν το σύστημα είναι σχεδιασμένο να αποδίδει καλύτερα για ένα ομιλητή (speaker dependent), ή για πολλούς και διαφορετικούς ομιλητές (speaker independent). Ακόμα, με βάση την ανθρώπινη ομιλία, υπάρχουν συστήματα που αναγνωρίζουν τον ομιλητή (speaker recognition), όπως ένας άνθρωπος αναγνωρίζει κάποιον άλλο άνθρωπο με βάση τα ιδιαίτερα χαρακτηριστικά της φωνής του. Άλλα συστήματα είναι ικανά να αναγνωρίσουν την γλώσσα του ομιλητή (language identification). Τέλος, με βάση τον τρόπο ομιλίας, υπάρχουν συστήματα που μπορούν να αναγνωρίσουν το συναίσθημα του ομιλητή (automatic emotion identification). Στο επόμενο κεφάλαιο, θα εξεταστούν κυρίως οι χρήσεις μίας κατηγορίας συστημάτων που χρησιμοποιούν ομιλία, αυτά που αναγνωρίζουν ομιλία, που είναι και τα πιο δημοφιλή.

Κεφάλαιο 3. Χρήσεις της αυτόματης αναγνώρισης ομιλίας

Εισαγωγή

Ο λόγος που χρησιμοποιείται η τεχνολογία της αυτόματης αναγνώρισης ομιλίας συνήθως είναι το ότι ένας άνθρωπος μπορεί να εκφέρει μία έκφραση πολύ πιο γρήγορα από το να την γράψει. Ένας σχετικά γρήγορος δακτυλογράφος γράφει περίπου 50 λέξεις το λεπτό. Συνεπώς, για ένα e-mail αποτελούμενο από 300 λέξεις, θα χρειαστεί 6 λεπτά. Χρησιμοποιώντας ένα σύστημα αναγνώρισης ομιλίας ένας άνθρωπος μπορεί υπαγορεύοντας 140 με 160 λέξεις το λεπτό, χωρίς λάθη κατά την αναγνώριση, να δημιουργήσει το ίδιο e-mail σε περίπου δύο λεπτά. Σε αυτό τον χρόνο δεν υπολογίζεται τυχόν εξοικονόμηση που υπάρχει χρησιμοποιώντας φωνητικές εντολές για να ανοίξει το λογισμικό για την αποστολή e-mail, να αναζητήσει την διεύθυνση ηλεκτρονικού ταχυδρομείου του παραλήπτη, και να στείλει το e-mail. Ένας μέσος άνθρωπος, μιλάει 5-7 φορές γρηγορότερα από το να γράφει. Είναι φανερό ότι η τεχνολογία αναγνώρισης ομιλίας μπορεί να γίνει πολύ χρήσιμη. Παρακάτω παρουσιάζονται διάφοροι τομείς που μπορεί να χρησιμοποιηθεί.

3.1. Χρήσεις σε συστήματα Interactive Voice Response

Τα Interactive Voice Response (IVR) είναι συστήματα που επιτρέπουν σε ένα υπολογιστή να δεχθεί φωνητικές εντολές και εντολές από τα πλήκτρα μιας συσκευής, όπως π.χ. ένα τηλέφωνο. Χρησιμοποιούνται κυρίως για να αυτοματοποιήσουν την αλληλεπίδραση ενός τηλεφωνικού κέντρου και ενός καλούντα. Τα συστήματα αυτά χρησιμοποιούν την τεχνολογία αναγνώρισης ομιλίας, για να κάνουν πιο προσιτή την αλληλεπίδραση με ανθρώπους. Οι εφαρμογές της αναγνώρισης ομιλίας στα IVR συστήματα περιγράφονται παρακάτω.

3.1.1. Χρήση σε συστήματα πληροφόρησης

Σε κάποιες περιπτώσεις οι χρήστες δεν χρειάζονται να μιλήσουν σε κάποιο άτομο σε ένα τηλεφωνικό κέντρο για να μάθουν κάποιες πληροφορίες. Για παράδειγμα, όταν δεν έχουν πολύ χρόνο να διαθέσουν ή χρειάζονται απλές πληροφορίες όπως ενημέρωση για δρομολόγια αεροπλάνων, η αναγνώριση ομιλίας μπορεί να χρησιμοποιηθεί για να μειώσει τον χρόνο αναμονής και να δώσει στους ενδιαφερόμενους τις πληροφορίες που χρειάζονται.

Ένα ακόμα πλεονέκτημα της χρήσης συστημάτων αναγνώρισης ομιλίας είναι ότι μπορούν να βοηθήσουν αποτελεσματικά σε στιγμές όπου υπάρχει μεγάλος φόρτος, και οι πελάτες χρειάζονται γρήγορη εξυπηρέτηση. Ένα παράδειγμα είναι οι εταιρείες στοιχημάτων. Σε μία κανονική ημέρα, υπάρχουν αγώνες κάθε 10 λεπτά, με το 80% των στοιχημάτων μέσω τηλεφώνου να γίνεται τα τελευταία λεπτά. Η εταιρεία Ladbrokes, που δραστηριοποιείται στον τομέα των τυχερών παιχνιδιών, χρησιμοποιεί σύστημα αναγνώρισης ομιλίας, το οποίο χειρίζεται τις περιπτώσεις όπου ο παίκτης θέλει να κάνει απλές ενέργειες, που είναι και οι πιο συνηθισμένες, όπως να ποντάρει σε κάποιον αγώνα, ή να ρωτήσει για τις αποδόσεις. Αυτό έχει ως αποτέλεσμα την διαχείριση μεγάλου φόρτου, χωρίς επιπλέον αύξηση του προσωπικού. Για πιο σύνθετα ζητήματα όπως ποντάρισμα με κάποιο σύστημα, ο παίκτης μπορεί να μιλήσει με κάποιο υπάλληλο.

Το αεροδρόμιο του Δουβλίνου, χρησιμοποιεί ένα τέτοιο σύστημα και κατάφερε να διαχειριστεί 30% αύξηση των επιβατών χωρίς να χρειαστεί να γίνει καμία προσθήκη προσωπικού. Οι πελάτες που επιθυμούν να ενημερωθούν σχετικά με τις αφίξεις ή τις αναχωρήσεις μπορούν να πουν “arrivals” ή “departures” αντίστοιχα, και το σύστημα επιστρέφει τις σχετικές πληροφορίες. Περίπου το 80% των χρηστών λαμβάνουν τις πληροφορίες που χρειάζονται από αυτό το σύστημα. Με την χρήση αυτού του συστήματος ο μέσος χρόνος αναμονής είναι 53 δευτερόλεπτα, και οι εργαζόμενοι στις πληροφορίες μπορούν να ασχοληθούν μόνο για πελάτες που χρειάζονται πιο σύνθετες πληροφορίες.

3.1.2. Προώθηση των χρηστών στο κατάλληλο τμήμα

Οι εντολές μέσω πλήκτρων που απαιτούνται από ένα IVR σύστημα, όπως στα τηλεφωνικά κέντρα των εταιριών κινητής τηλεφωνίας, μπορούν πολλές φορές να γίνουν πολύπλοκες, και να μπερδέψουν τον χρήστη, με αποτέλεσμα να καταλήξει σε διαφορετικό τμήμα από αυτό που θα μπορούσε να τον εξυπηρετήσει. Ακόμα, δαπανάται περισσότερος χρόνος καθώς ο χρήστης ίσως χρειαστεί να ακούσει και τις επιλογές που δεν του χρειάζονται, μέχρι να ακούσει αυτή που τον ενδιαφέρει. Η αναμονή σε μία σειρά προτεραιότητας έως ότου κάποιος υπάλληλος προωθήσει σε κάποιο τμήμα τον πελάτη, και κυρίως η κατάληξη τελικά σε λάθος τμήμα, προκαλούν εκνευρισμό και μη ικανοποίηση. Τα συστήματα που χρησιμοποιούν αναγνώριση ομιλίας είναι κατά 40% ταχύτερα από αυτά που περιμένουν τις εντολές του χρήστη μέσω πλήκτρων (The Wall Street Journal,2003). Έτσι, είναι προφανές ότι η τεχνολογία αναγνώρισης ομιλίας μπορεί να δώσει την λύση σε αυτό το ζήτημα. Με την τεχνολογία Intelligent Call Steering (ICS), δεν απαιτείται από τον χρήστη να χρησιμοποιήσει πλήκτρα για να δώσει εντολές. Το σύστημα ρωτάει τον χρήστη τι χρειάζεται, και αυτός απλά αποκρίνεται λέγοντας το αίτημα του. Στη συνέχεια το σύστημα καταλαβαίνει τι είπε ο καλών, και τον προωθεί στην κατάλληλη υπηρεσία. Το σύστημα προωθεί τον χρήστη στο προσωπικό που πρέπει σε 20 με 30 δευτερόλεπτα, ενώ οι κλήσεις που δεν προωθούνται σωστά είναι το 3-5% της συνολικής κίνησης.

Η εταιρεία Standard Life που δραστηριοποιείται σε όλο τον κόσμο χρησιμοποιεί σύστημα αυτόματης αναγνώρισης ομιλίας για τις ασφαλιστικές υπηρεσίες που παρέχει. Το σύστημα πρώτα εξακριβώνει για ποιο λόγο κάλεσε ο χρήστης, εάν χρειάζεται υποβάλλει τον χρήστη σε ελέγχους ασφαλείας και τέλος, προωθεί τον χρήστη στο κατάλληλο τμήμα. Η Standard Life έτσι αύξησε τον αριθμό των κλήσεων που μπορεί να εξυπηρετήσει κατά 25%, και μείωσε τον αριθμό των κλήσεων που κατέληγαν σε λάθος τμήμα κατά 66%.

Η εταιρεία Suncorp, που παρέχει ασφαλιστικές υπηρεσίες, αντικατέστησε το προηγούμενο IVR σύστημα που απαιτούσε από τον χρήστη να χρησιμοποιήσει πλήκτρα, με ένα IVR σύστημα που λειτουργεί με αναγνώριση φυσικού λόγου. Το σύστημα αυτό διαθέτει ένα λεξικό με περισσότερες από 100.000 εκφράσεις, ώστε να πετυχαίνει μεγάλη ακρίβεια.

3.1.3. Αυτόματη αναγνώριση χρήστη

Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, η αναγνώριση ομιλίας μπορεί να χρησιμοποιηθεί για αναγνώριση ατόμων. Οι απάτες που έχουν σχέση με αναγνώριση χρηστών στο Ηνωμένο Βασίλειο, σύμφωνα με την υπηρεσία κατά της απάτης του Ηνωμένου Βασιλείου (CIFAS), εκτιμάται ότι κοστίζει 1,5 δισεκατομμύριο ευρώ ετησίως. Είναι πολύ χρήσιμη τεχνική, γιατί μπορεί να γίνει η αναγνώριση κάποιου ατόμου, χωρίς να χρειάζονται δεδομένα όπως κωδικοί, που από ένα μη ασφαλές μέσον όπως το τηλέφωνο, μπορούν να υποκλαπούν. Ακόμα, είναι πιο εύκολο για άτομα που δεν είναι εξοικειωμένα με τη χρήση πληροφοριακών συστημάτων να τα χρησιμοποιήσουν. Επίσης, χρειάζονται λιγότερα από δύο λεπτά για να δημιουργηθεί ένα φωνητικό αποτύπωμα. Αυτό αποθηκεύεται σε μία βάση δεδομένων, και κάθε φορά που θα χρειαστεί να γίνει αναγνώριση, η διαδικασία θα διαρκέσει λιγότερο από 30 δευτερόλεπτα.

Η εταιρεία Ahm Health Management που δραστηριοποιείται στην Αυστραλία, παρέχοντας ασφαλιστικές υπηρεσίες διαθέτει ένα τέτοιο σύστημα που επιτρέπει στους πελάτες της να επικοινωνούν με τους εκπροσώπους της, γρήγορα και με ασφάλεια. Οι εγγεγραμμένοι πελάτες απλά λένε τον κωδικό που τους έχει δοθεί, χωρίς να χρειάζεται να δώσουν κωδικούς, όνομα, ή άλλες πληροφορίες που πιθανώς να κουράζουν τον χρήστη. Για τον ίδιο λόγο, το τηλεφωνικό κέντρο έχει λιγότερο φόρτο. Τέλος, εξοικονομείται χρόνος και από τους εκπροσώπους της εταιρείας, γιατί δεν χρειάζεται να κάνουν ερωτήσεις για να μάθουν ποιος είναι ο πελάτης. Τους ενημερώνει το σύστημα.

3.2. Χρήση σε μαθητές με αναπηρίες

Μαθητές οι οποίοι αντιμετωπίζουν προβλήματα μάθησης και αδυναμία συγκέντρωσης, σωματική αναπηρία, διανοητικά προβλήματα ή προβλήματα στην ομιλία μπορούν να επωφεληθούν από την τεχνολογία της αναγνώρισης ομιλίας (Speech recognition for people with disabilities, www.customtyping.com) για να διδαχθούν, να επικοινωνήσουν και να δημιουργήσουν. Παρακάτω παρουσιάζονται διάφοροι τομείς της μάθησης που η αναγνώριση ομιλίας μπορεί να βοηθήσει.

3.2.1. Χρήση σε τομείς που απαιτούν οπτική επαφή

Η τεχνολογία της αναγνώρισης ομιλίας επειδή δεν απαιτεί την ανάγκη γνώσης του χειρισμού του ποντικιού και του πληκτρολογίου, ελαχιστοποιεί την ανάγκη χρήσης τους. Μπορεί να χρησιμοποιηθεί για να βοηθήσει μαθητές με προβλήματα όρασης.

3.2.2. Λύση σε εργονομικά προβλήματα

Το γεγονός ότι δεν υπάρχει η ανάγκη χρήσης του ποντικιού και του πληκτρολογίου δίνει την δυνατότητα σε μαθητές που δεν μπορούν να πάρουν την ενδεδειγμένη στάση χειρισμού ενός υπολογιστή, σε αυτούς που δεν μπορούν να μείνουν σταθεροί, αλλά και σε μαθητές με προβλήματα αναπηρίας στα χέρια, να μπορούν να δουλεύουν και να γράφουν στον υπολογιστή,

3.2.3. Ευκολότερος έλεγχος λογισμικού

Για τους μαθητές που αντιμετωπίζουν προβλήματα μάθησης είναι δύσκολο να μάθουν να χειρίζονται αποτελεσματικά το ποντίκι, να ανοίγουν και να κλείνουν ένα πρόγραμμα και να μετακινούνται στο σύστημα αρχείων του υπολογιστή. Αυτοί οι μαθητές, μπορεί να βρουν πιο ελκυστική την χρήση αναγνώρισης ομιλίας για τον έλεγχο του λογισμικού.

3.2.4. Βοήθημα για τους μαθητές που κουράζονται εύκολα

Η αναγνώριση ομιλίας μπορεί να βοηθήσει τους μαθητές που κουράζονται εύκολα όταν εργάζονται στον υπολογιστή, αφού χρειάζονται λιγότερη ενέργεια για τον χειριστούν και να δουλέψουν με αυτόν. Επίσης, η δυνατότητα να κάθονται σε στάση που τους βολεύει, με καλύτερη στήριξη για το σώμα έχει ως αποτέλεσμα μεγαλύτερη αντοχή και μπορούν να δώσουν σημασία στην εργασία που έχουν, χωρίς να αποσπώνται από τον έλεγχο της στάσης του σώματος και από τις κινήσεις των χεριών που χρειάζονται για να μπορεί κάποιος να γράφει με το πληκτρολόγιο.

3.2.5. Αποτελέσματα χρήσης προγραμμάτων αναγνώρισης ομιλίας σε μαθητές με αναπηρίες

Η ικανότητα της χρησιμοποίησης ενός προγράμματος αναγνώρισης ομιλίας, καθώς και ο χειρισμός του είναι ζητήματα που απαιτούν προσήλωση και συγκέντρωση. Οι μαθητές που αντιμετωπίζουν προβλήματα συγκέντρωσης, μπορεί να βρουν δύσκολη την αρχική εκπαίδευση, και μπορεί ακόμη να αντιμετωπίσουν δυσκολίες στην επισήμανση λανθασμένων αναγνωρίσεων. Όμως, θετικά αποτελέσματα των μαθητών με προβλήματα συγκέντρωσης στον τομέα της ποιότητας της εργασίας που τους έχει ανατεθεί, συνεπακολουθούμενη αύξηση της βαθμολογίας τους, αλλά και βελτίωση της ποιότητας και του μεγέθους των κειμένων που γράφουν, έχουν ως αποτέλεσμα μεγαλύτερη επιθυμία για να επιτύχουν, το οποίο επιφέρει αύξηση της συγκέντρωσης που επιδεικνύουν.

Η συγγραφή κειμένων με χρήση αναγνώρισης ομιλίας μπορεί να είναι ένα κίνητρο για βελτίωση ενός μαθητή ο οποίος είναι απογοητευμένος μετά από προσπάθεια χρόνων για την συγγραφή κειμένου χειρωνακτικά. Παρόλο που η αναγνώριση ομιλίας απαιτεί να θυμούνται οι μαθητές τις φωνητικές εντολές και να κατανοούν που και πως πρέπει να κάνουν διορθώσεις στο αποτέλεσμα της αναγνώρισης, το επίπεδο της προσοχής που πρέπει να έχει κάποιος όταν χρησιμοποιεί το πληκτρολόγιο είναι μεγαλύτερο. Αυτό συμβαίνει γιατί η εργασία με πληκτρολόγιο, δεν απαιτεί μόνο προσοχή στην οθόνη και στα αποτελέσματα που προκύπτουν, αλλά και στην θέση των πλήκτρων επάνω στο πληκτρολόγιο. Η αναγνώριση ομιλίας επιτρέπει στους μαθητές να επικεντρωθούν μόνο στο τι λένε, και στην διόρθωση πιθανών λαθών.

3.2.5.1. Βελτίωση της ικανότητας αναγνώρισης εκφράσεων που προφέρονται λανθασμένα

Η χρήση της αναγνώρισης ομιλίας μαζί με την ανάγνωση κειμένου (text to speech) μπορεί να βοηθήσει τους μαθητές που έχουν προβλήματα στην ομιλία. Το πρόγραμμα αναγνώρισης ομιλίας Dragon NaturallySpeaking έχει την δυνατότητα ηχογράφησης της ομιλίας του χρήστη, και την παραγωγή κειμένου από αυτή. Κατά την αναπαραγωγή της φωνής το πρόγραμμα τονίζει την λέξη που ακούγεται

εκείνη τη στιγμή. Αυτό επιτρέπει στον χρήστη σε περίπτωση λάθους να εξακριβώσει εάν το λάθος συνέβη από σφάλμα κατά την ομιλία, π.χ. λόγω κακής προφοράς μιας λέξης, ή κατά την αναγνώριση.

3.2.5.2 Βελτίωση ικανότητας γραφής

Η υπαγόρευση (dictation) σε έναν υπολογιστή μπορεί να βελτιώσει την ικανότητα γραφής ατόμων με προβλήματα μάθησης. Αυτό γίνεται γιατί αίρεται το πρόβλημα της γνώσης του συλλαβισμού ώστε να γραφεί μία λέξη, της γνώσης της στίξης, της τοποθέτησης κεφαλαίων γραμμάτων και της ικανότητας κάποιος να μπορεί να γράψει με το χέρι. Για μαθητές που δεν είναι καλοί στη συγγραφή κειμένου, η αναγνώριση ομιλίας βοηθά σε καλύτερα αποτελέσματα.

3.2.5.3. Βελτίωση ανάγνωσης

Η αναγνώριση ομιλίας μπορεί να βοηθήσει στην εκμάθηση κάποιων αναγνωστικών ικανοτήτων. Έρευνες έχουν δείξει ότι η αναγνώριση ομιλίας μπορεί να είναι αποτελεσματική στην θεραπεία των προβλημάτων ανάγνωσης και συλλαβισμού, των παιδιών με προβλήματα μάθησης. Οι ερευνητές Raskind και Higgins (1999) βρήκαν ότι μαθητές με προβλήματα μάθησης σε ηλικίες από 9-18 ετών, που χρησιμοποιούσαν την αναγνώριση ομιλίας για να γράψουν κείμενα 50 λεπτά την εβδομάδα, για 16 εβδομάδες, έδειξαν σημαντική βελτίωση στους τομείς της αναγνώρισης κάποιας λέξης, συλλαβισμού, της φωνολογικής τους έκφρασης και κατανόησης της διαδικασίας της ανάγνωσης.

Υπάρχουν προγράμματα, όπως της εταιρείας Soliloquy Learning, που ειδικεύονται στην βελτίωση των ικανοτήτων της ανάγνωσης. Καθώς οι μαθητές προφέρουν κάποια λέξη, το πρόγραμμα ανιχνεύει για ποια λέξη πρόκειται, και βοηθά τους μαθητές προφέροντας την λέξη.

3.2.6. Προγράμματα αναγνώρισης ομιλίας ως βοήθημα στην εκπαίδευση

Προγράμματα αναγνώρισης ομιλίας χρησιμοποιούνται από καθηγητές ως βοήθημα για μαθητές που έχουν προβλήματα ακοής ή για κάποιο λόγο δεν μπορούν να κρατήσουν σημειώσεις. Καθώς ο καθηγητής ομιλεί, το περιεχόμενο της ομιλίας του μετατρέπεται σε κείμενο από ένα πρόγραμμα αναγνώρισης ομιλίας. Το κείμενο αυτό είναι διαθέσιμο στους μαθητές του. Διαθέσιμα προγράμματα για αυτό το σκοπό είναι το *TypeWell*, και το *Liberated Learning*. Τέλος, το πρόγραμμα *ICommunicator*, μετατρέπει την ομιλία του καθηγητή σε κείμενο αλλά και στην νοηματική γλώσσα σε πραγματικό χρόνο.

3.3. Έλεγχος λογισμικού μέσω φωνητικών εντολών

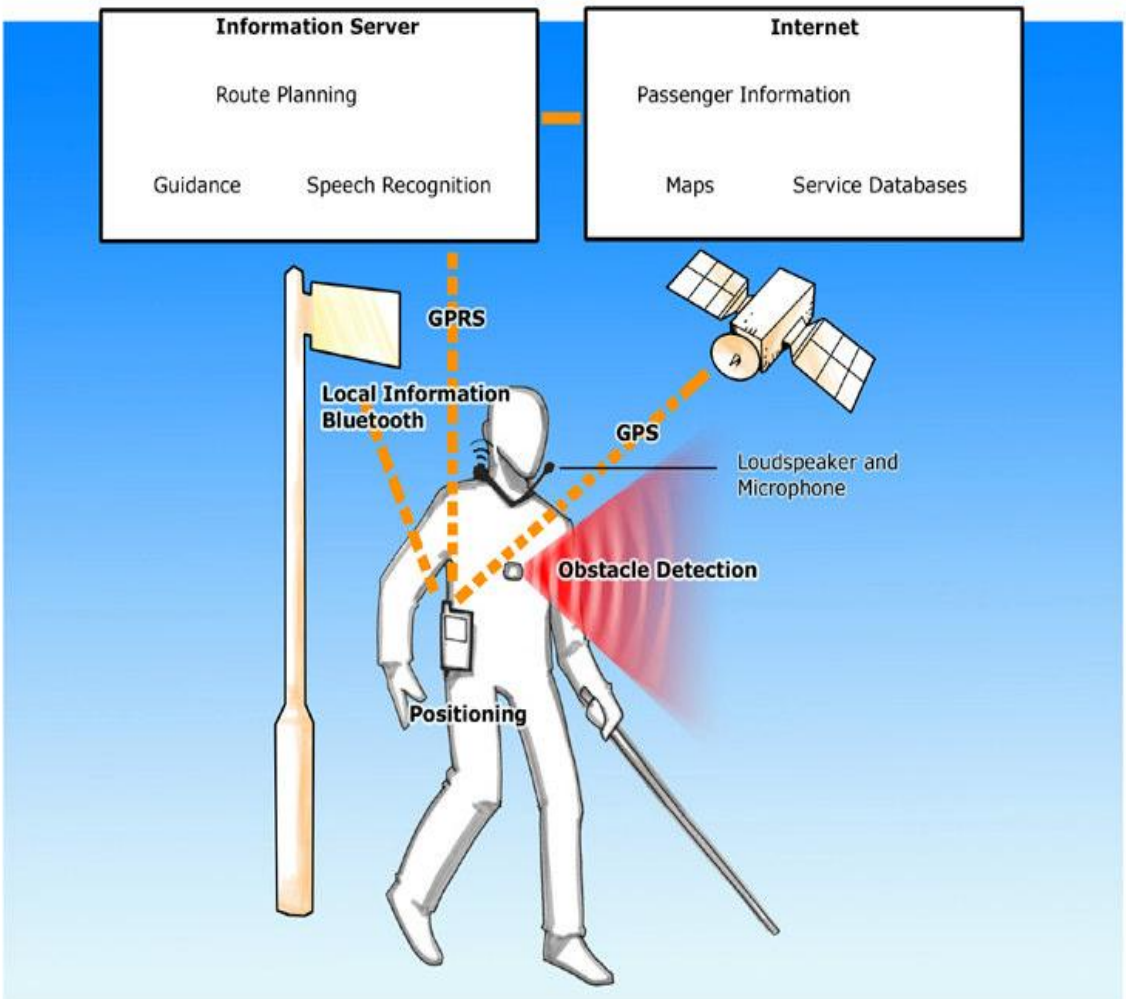
Η τεχνολογία της αναγνώρισης ομιλίας επιτρέπει στον χρήστη ενός ηλεκτρονικού υπολογιστή να τον ελέγξει μέσω φωνητικών εντολών. Οποιαδήποτε συντόμευση της επιφάνειας εργασίας και οποιοδήποτε μενού επιλογών, μπορεί να επιλεγεί με χρήση φωνητικών εντολών. Ακόμα, ο χρήστης μπορεί να ανοίξει και να τροποποιήσει αρχεία κειμένου, να πλοηγηθεί μέσα στο σύστημα αρχείων και να βρει το αρχείο ή τον κατάλογο που χρειάζεται και να τον ανοίξει. Με την χρήση φωνητικών εντολών, είναι πιθανό χρήστες οι οποίοι σήμερα δεν χρησιμοποιούν υπολογιστή, διότι τους φαίνεται δύσκολο να μάθουν να τον χειρίζονται, να χρησιμοποιήσουν τελικά, καθώς ο έλεγχός του έτσι είναι πιο φυσικός. Επίσης, με την χρήση αναγνώρισης ομιλίας, μονότονες εργασίες όπως η συμπλήρωση μίας ηλεκτρονικής φόρμας, ή η εισαγωγή δεδομένων, μπορούν να γίνουν γρηγορότερα.

3.3.1. Χρήση αναγνώρισης ομιλίας από άτομα με ειδικές ανάγκες

Όπως αναφέρθηκε και παραπάνω, με την χρήση φωνητικών εντολών γίνεται πολύ πιο εύκολο σε τυφλούς χρήστες ή χρήστες με αναπηρίες, να χειριστούν τον υπολογιστή.

Ένα πρόγραμμα, που δημιουργήθηκε στην Φινλανδία από κρατικούς φορείς με το όνομα Noppa, επιτρέπει σε τυφλούς να περιηγηθούν σε μία πόλη, και να

χρησιμοποιήσουν τα κατάλληλα μέσα μαζικής μεταφοράς για να πάνε στον προορισμό τους. Το μόνο που χρειάζεται είναι ένα κινητό τηλέφωνο που να περιέχει λογισμικό αναγνώρισης ομιλίας, δέκτη GPS (Global Positioning System) που είναι ένα παγκόσμιο σύστημα εντοπισμού θέσης, και να είναι ικανό να συνδεθεί στο Internet με την βοήθεια GPRS. Η συσκευή του κινητού τηλεφώνου είναι συνδεδεμένη με ένα server όπου γίνεται η επιλογή της κατεύθυνσης που θα δοθεί στον χρήστη. Αυτός ο server συνδέεται και με το Internet και παρέχει πληροφορίες σχετικές με τον καιρό ή την κίνηση. Το λογισμικό αναγνώρισης ομιλίας επιτρέπει στον χρήστη να πει τον προορισμό του, και στην συνέχεια με χρήση λογισμικού για σύνθεση ομιλίας (speech synthesis), υποδεικνύεται στον χρήστη πως θα φτάσει στην σωστή στάση λεωφορείου ή στο σωστό σταθμό τρένου, και ποιο είναι το κατάλληλο όχημα για να επιβιβαστεί.



Εικόνα 4. Αρχιτεκτονική του Noppa

3.3.2. Χρήση της αναγνώρισης ομιλίας για υπαγόρευση κειμένου

Ένα παράδειγμα του πως η αναγνώριση ομιλίας μπορεί να χρησιμοποιηθεί για υπαγόρευση κειμένου είναι οι καταγραφές EMR (electronic medical record). Οι EMR είναι πληροφορίες αποθηκευμένες σε ηλεκτρονική μορφή, δημιουργημένες από ένα νοσοκομείο ή από ένα γιατρό, και παρέχουν το ιστορικό της κατάστασης της υγείας ενός ασθενή. Η δημιουργία μίας καταγραφής, είναι χρονοβόρα υπόθεση. Αυτό συμβαίνει γιατί ο γιατρός πρέπει να ηχογραφήσει σε μία συσκευή την διάγνωση της κατάστασης της υγείας του ασθενή. Στη συνέχεια, το περιεχόμενο της ηχογράφησης μετατρέπεται σε κείμενο από κάποια εταιρεία που δραστηριοποιείται σε αυτό το τομέα.

Η χρήση της τεχνολογίας της αναγνώρισης ομιλίας κάνει αποτελεσματικότερη και ευκολότερη την σύνταξη τους, και έχει και οικονομικά οφέλη. Η διαδικασία γίνεται ταχύτερη αφού το βήμα της μετατροπής της ομιλίας του γιατρού σε κείμενο αφαιρείται. Όμως, ακόμα και αν ο συντάκτης της ιατρικής αναφοράς είναι ο ίδιος ο γιατρός, εξοικονομείται χρόνος αφού κάθε γιατρός χρειάζεται περίπου 15 ώρες την εβδομάδα για αυτό το σκοπό. Με την χρήση της αναγνώρισης ομιλίας ο χρόνος αυτός μειώνεται στο μισό. Ακόμα, οι αναζητήσεις πληροφοριών για τον ασθενή, η συμπλήρωση ηλεκτρονικών φορμών, η συνταγογράφηση και οι οδηγίες προς τους νοσηλευτές γίνονται ταχύτερα με την ομιλία από ότι με το πληκτρολόγιο. Τέλος, γιατροί μεγάλης ηλικίας, που δεν είναι ακόμα εξοικειωμένοι με την χρήση υπολογιστή, μπορούν να συντάσσουν τις αναφορές τους μέσω της υπαγόρευσης.

3.3.3. Χρήση σε περιπτώσεις όπου τα μάτια και τα χέρια του χρήστη είναι δεσμευμένα

Η αναγνώριση ομιλίας μπορεί να είναι πολύ χρήσιμη σε περιπτώσεις όπου τα μάτια και τα χέρια του χρήστη είναι δεσμευμένα, όπως για παράδειγμα στην οδήγηση. Πολλές από τις συσκευές GPS (Global Positioning System) που κυκλοφορούν στην αγορά, έχουν αυτή την δυνατότητα, όπως το Pioneer AVIC-F500BT, το Garmin 855 and 885T κ.ά. Υπάρχουν συσκευές που ο έλεγχός τους γίνεται από συγκεκριμένες φωνητικές εντολές, και άλλες που οι φωνητικές εντολές που δέχονται μπορούν να μοιάζουν στον φυσικό λόγο. Η χρήση φωνητικών

εντολών μπορεί να κάνει ασφαλέστερη την οδήγηση, αφού ο οδηγός δεν χρειάζεται να χειρίζεται χειρωνακτικά την συσκευή, αλλά κάνει και ευκολότερη την χρήση της. Σύμφωνα με στοιχεία της εταιρείας Nuance Communications, που δραστηριοποιείται στον χώρο των συστημάτων αναγνώρισης ομιλίας, 8 στους 9 χρήστες χρησιμοποιούν φωνητικές εντολές στις συσκευές που το επιτρέπουν.

Όμως, σύμφωνα με την έρευνα Cell Phones and Driving (McKnight, 1991), του Αμερικανικού Αυτοκινητιστικού Συλλόγου, (American Automobile Association-AAA), η χρήση φωνητικά ελεγχόμενων συσκευών μπορεί να είναι εξίσου επικίνδυνη με την χρήση συσκευών που ελέγχονται χειρωνακτικά. Αυτό οφείλεται, υποστηρίζουν, στο γεγονός ότι ο νοητικός περισπασμός είναι περίπου το ίδιο επικίνδυνος με τον χειρωνακτικό. Μία άλλη έρευνα, (Maciej, J. and Vollrath, M., 2009) για λογαριασμό της εταιρείας Nuance Communications, έδειξε ότι η χρήση φωνητικών εντολών μειώνει σημαντικά τον περισπασμό του οδηγού, συγκριτικά με τον χειρωνακτικό τρόπο. Εξετάζοντας τις κινήσεις των ματιών των οδηγών, ανακάλυψαν ότι με την χρήση φωνητικών εντολών στα GPS, τα μάτια τους κοίταζαν τον δρόμο 200 με 300% περισσότερο από τα απλά GPS.

3.3.4. Χρήση σε ηλεκτρονικά παιχνίδια

Η τεχνολογία της αναγνώρισης ομιλίας χρησιμοποιείται και στα ηλεκτρονικά παιχνίδια. Τα πλεονεκτήματα της χρήσης της είναι αρκετά. Με την χρήση φωνητικών εντολών, μειώνεται το πρόβλημα της αναζήτησης του κατάλληλου κουμπιού ή μενού, ενώ παίζεται το παιχνίδι. Ο παίκτης απλά λέει την εντολή και το παιχνίδι την εκτελεί, χωρίς να χρειάζεται να απασχοληθούν τα μάτια του στο πληκτρολόγιο. Η χρήση φωνητικών εντολών δεν κάνει καλύτερο μόνο τον τρόπο παιξίματος, αλλά επιτρέπει και σε παίκτες που δεν είναι εξοικειωμένοι με τον χειρισμό υπολογιστή ή ειδικότερα κάποιου παιχνιδιού, να παίξουν με σχετική ευκολία ένα παιχνίδι. Επίσης, το να δίνει ένας παίκτης φωνητικές εντολές στο παιχνίδι, αυξάνει τον ρεαλισμό και την ευχαρίστηση που προκύπτει από το παίξιμο του. Στον πίνακα 1 αναφέρονται κάποια παιχνίδια που μπορούν να χρησιμοποιήσουν φωνητικές εντολές. Ακόμα, παρουσιάζεται και σε ποιες κονσόλες μπορούν να παιχτούν αυτά τα παιχνίδια. Φαίνεται ότι η πλειοψηφία των δημοφιλών παιχνιδομηχανών υποστηρίζει αναγνώριση ομιλίας.

Πίνακας 1. Παιχνίδια που χρησιμοποιούν φωνητικές εντολές

Παιχνίδι	*Κονσόλα*
Tom Clancy's EndWar	Sony Playstation 3, Microsoft Xbox, Sony PSP, PC, Nintendo DS
Mario Party 7	Nintendo GameCube
NASCAR 06: Total Team Control	Sony Playstation 2, Microsoft Xbox
Battlefield 1942: World War II Anthology	PC
SOCOM: U.S. Navy SEALS Combined Assault	Sony Playstation 2
SingStar Pop Edition	Sony Playstation 3
Phoenix Wright: Ace Attorney - Trials and Tribulations	Nintendo Wii, Nintendo DS, PC, Game Boy Advance

3.4. Χρήση για στρατιωτικούς σκοπούς

3.4.1. Χρήση σε μαχητικά αεροσκάφη

Τα πιλοτήρια των μαχητικών αεροσκαφών γίνονται όλο και πιο σύνθετα, όσο προχωρά η εξέλιξη τους, περιέχοντας ένα πλήθος από συσκευές ελέγχου. Αυτό το πολύπλοκο και αγχωτικό περιβάλλον απορροφά πολλή από την προσοχή του πιλότου. Η χρησιμοποίηση αναγνώρισης ομιλίας ως εργαλείο ελέγχου, μπορεί να ελευθερώσει τα μάτια και τα χέρια του πιλότου με αποτέλεσμα την μείωση του φόρτου του, και την συγκέντρωση του στην αποστολή του. Ακόμα, με την χρησιμοποίηση φωνητικών εντολών αντί για πλήκτρα ή μοχλούς ελέγχου, εξοικονομείται χώρος, που σε ένα πιλοτήριο μαχητικού αεροσκάφους είναι πολύτιμος. Αυτό το γεγονός μπορεί να ωφελήσει και από οικονομικής πλευράς.

Τα τελευταία χρόνια έχουν γίνει διάφορες απόπειρες ώστε να χρησιμοποιηθεί η τεχνολογία της αναγνώρισης ομιλίας στα πιλοτήρια των μαχητικών αεροσκαφών. Τα προγράμματα που υπάρχουν για αυτό το σκοπό είναι το αμερικανικό Advanced

Fighter Technology Integration (AFTI), για τα αεροσκάφη τύπου F-16, το γαλλικό πρόγραμμα για την εισαγωγή τεχνολογίας αναγνώρισης ομιλίας στα αεροσκάφη τύπου Mirage και αγγλικά προγράμματα για διάφορους τύπους αεροσκαφών.

Οι εργασίες που τα συστήματα αναγνώρισης ομιλίας καλούνται να διεκπεραιώσουν είναι: ορισμός συχνότητας του ασυρμάτου, φωνητικές εντολές σε σύστημα αυτόματου πιλότου, ορισμός συντεταγμένων στόχου, ορισμός δεδομένων για τον χειρισμό του οπλισμού και χειρισμός δεδομένων πτήσης. Γενικά, χρησιμοποιούνται μικρά σε μέγεθος λεξικά, ώστε να επιτυγχάνεται μεγαλύτερη ακρίβεια. Η αύξηση όμως της επιτάχυνσης της βαρύτητας (G), που προκαλείται όταν τα αεροσκάφη κάνουν απότομους ελιγμούς, προκαλεί μείωση της ακρίβειας.

Η RAF (Royal Air Force) , που είναι η πολεμική αεροπορία της Μεγάλης Βρετανίας, χρησιμοποιεί στο αεροσκάφος Eurofighter Typhoon σύστημα αναγνώρισης ομιλίας το οποίο εξαρτάται από τον χρήστη. Έτσι απαιτείται από κάθε πιλότο να εκπαιδεύσει το σύστημα προτού το χρησιμοποιήσει. Το σύστημα προς το παρών δεν χρησιμοποιείται σε κρίσιμα ζητήματα, όπως η προσγείωση του αεροσκάφους ή η ρίψη των βλημάτων.

Για να χρησιμοποιηθεί η τεχνολογία της αναγνώρισης ομιλίας στα πολεμικά αεροσκάφη θα πρέπει η ακρίβεια που επιτυγχάνεται να είναι από 95% και πάνω. Γενικά, η αναγνώριση ομιλίας δεν είναι ακόμη τόσο ακριβής ώστε να χρησιμοποιείται σε κρίσιμα ζητήματα, όπου ακόμη και ένα μικρό λάθος μπορεί να έχει δραματικές επιπτώσεις. Για παράδειγμα, αν αναγνωρισθεί η λέξη “eject” (εκτόξευση) αντί της “reject” (απόρριψη).

3.4.2. Χρήση σε πολεμικά ελικόπτερα

Εφαρμογές της χρήσης αναγνώρισης ομιλίας για τους ίδιους λόγους με των αεροσκαφών, υπάρχουν και στα πολεμικά ελικόπτερα. Η διαφορά όμως των μαχητικών αεροσκαφών με τα ελικόπτερα, είναι ότι η καμπίνα των ελικοπτέρων είναι περισσότερο θορυβώδης, λόγω θορύβων που παράγονται από τα μηχανικά μέρη του ελικοπτέρου, με αποτέλεσμα μικρότερη ακρίβεια στην αναγνώριση. Ένας ακόμη λόγος είναι ότι συνήθως οι πιλότοι των ελικοπτέρων δεν φορούν μάσκα, η οποία μειώνει τους εξωτερικούς θορύβους.

Προγράμματα που διερευνούν αυτή την δυνατότητα είναι το πρόγραμμα Army Avionics Research and Development Activity (AVRADA) των ΗΠΑ, και το Royal Aerospace Establishment (RAE) της Μεγάλης Βρετανίας. Τα γαλλικά ελικόπτερα Puma έχουν ήδη αυτή την δυνατότητα. Τα αποτελέσματα είναι ενθαρρυντικά, και οι εφαρμογές τους είναι: έλεγχος και ρύθμιση των ασύρματων συστημάτων επικοινωνίας, των συστημάτων πλοήγησης και των συστημάτων στόχευσης.

Η χρήση αναγνώρισης ομιλίας στα πολεμικά ελικόπτερα γίνεται για να αυξηθεί η αποτελεσματικότητα του πιλότου. Τα αποτελέσματα από τα προγράμματα που αναφέρθηκαν είναι ενθαρρυντικά, αλλά ακόμη δεν έχουν χρησιμοποιηθεί σε πραγματικές καταστάσεις. Πρέπει να υπάρξουν εξελίξεις στην τεχνολογία της αναγνώρισης ομιλίας, προκειμένου να μπορούν να χρησιμοποιηθούν και στην πράξη.

3.5. Χρήση σε πολυμέσα

Με την χρήση αυτόματης αναγνώρισης ομιλίας, είναι πλέον εφικτό να εξάγεται το περιεχόμενο των διαλόγων ενός αρχείου video, για παράδειγμα μίας ταινίας, σε μορφή κειμένου. Αυτό που γίνεται είναι σε κάθε video οι ομιλίες να μετατρέπονται σε κείμενα με την χρήση λογισμικού αναγνώρισης ομιλίας και στη συνέχεια να γίνεται ευρετηριοποιήση με βάση το περιεχόμενο των διαλόγων. Εφαρμογές αυτής της τεχνικής είναι η αναζήτηση διαλόγου μέσα από ένα video χρησιμοποιώντας την φωνή του χρήστη και η αυτόματη δημιουργία υποτίτλων.

3.5.1. Αναζήτηση διαλόγων ενός αρχείου video

Η αναγνώριση ομιλίας μπορεί να χρησιμοποιηθεί για την αναζήτηση του περιεχομένου των διαλόγων ενός video. Ο χρήστης μπορεί όχι μόνο να κάνει αναζήτηση με βάση τον τίτλο ή την περιγραφή μίας ταινίας, αλλά και να ψάξει τους διαλόγους μέσα στην ταινία. Ακόμα, μπορεί κάποιος να πλοηγηθεί μέσα στην ταινία, αφού αυτό το σύστημα ξέρει σε ποιο σημείο της ταινίας βρίσκεται ο διάλογος που αναζητείται. Ο χρήστης κάνει αναζήτηση με την χρήση της ομιλίας του, η ομιλία του μετατρέπεται σε μορφή κειμένου, και το σύστημα ψάχνει αν

υπάρχει κάποιος διάλογος αποθηκευμένος στην βάση δεδομένων του, που να καλύπτει τα κριτήρια αναζήτησης.

Μία από τις μηχανές αναζήτησης διαλόγων video με χρήση αναγνώρισης ομιλίας την παρέχει το Google. Αυτό γίνεται με την εισαγωγή ενός gadget, που είναι μία εφαρμογή δημιουργημένη με HTML και JavaScript, που μπορεί να προστεθεί σε μία σελίδα, ως περιεχόμενο της. Το gadget αυτό παρουσιάστηκε και χρησιμοποιήθηκε πρώτη φορά στις αμερικανικές εκλογές του 2008. Άλλες μηχανές αναζήτησης είναι το Blinkx και το Everyzing (από το Νοέμβριο του 2009 μετονομάστηκε σε RAMP).

3.5.2. Αυτόματη δημιουργία υποτίτλων

Η τεχνολογία αναγνώρισης ομιλίας μπορεί να χρησιμοποιηθεί και για την αυτόματη δημιουργία υποτίτλων σε ένα αρχείο video. Η αυτόματη δημιουργία υποτίτλων σε video που ο δημιουργός τους δεν προτίθεται να φτιάξει, βοηθά τους κωφούς χρήστες να δουν video, αλλά και τους ανθρώπους που μιλούν διαφορετική γλώσσα από αυτή του video.

Στον δικτυακό τόπο YouTube, που περιέχει αρχεία video, ανεβαίνουν από τους χρήστες κάθε λεπτό 20 ώρες video. Στην πλειοψηφία των αρχείων, δεν βρίσκονται ενσωματωμένοι υπότιτλοι. Το YouTube εφαρμόζει πλέον την τεχνολογία από το Google Voice, που είναι ένα σύστημα αναγνώρισης ομιλίας, για την δημιουργία υποτίτλων. Τα αποτελέσματα δεν είναι τέλεια, όπως δηλώνουν οι υπεύθυνοι της Google, όμως ακόμα και έτσι μπορεί να είναι χρήσιμα ώστε να κατανοηθεί το περιεχόμενο των διαλόγων. Άλλα προγράμματα όπως τα Dragon NaturallySpeaking και το IBM ViaVoice δίνουν την επιλογή εξαγωγής των διαλόγων.

Το SDK που χρησιμοποιούν τα Microsoft Windows XP, και το SPHINX του πανεπιστημίου Carnegie Mellon των Ηνωμένων Πολιτειών, που είναι συστήματα αναγνώρισης ομιλίας, δοκιμάστηκαν στην αναγνώριση ομιλίας από την τηλεόραση. Επιλέχθηκαν τα δελτία ειδήσεων, όπου οι εκφωνητές ομιλούν με καθαρή φωνή και σωστή άρθρωση. Το SDK της Microsoft, έχει ρυθμό εμφάνισης λανθασμένων αναγνωρίσεων (WER) περίπου 60% στις ειδήσεις του τηλεοπτικού σταθμού CNN, για 30 λεπτά χωρίς διαφημίσεις. Το SPHINX, δοκιμάστηκε στην αναγνώριση 30

λεπτών του δελτίου ειδήσεων, και ο ρυθμός εμφάνισης λανθασμένων αναγνωρίσεων ήταν 35%. Η ακρίβεια του SDK της Microsoft θα μπορούσε να είναι καλύτερη, αφού το σύστημα δεν είχε εκπαιδευτεί χρησιμοποιώντας τέτοιου είδους κείμενα. Ακόμα, η προεπιλεγμένη ρύθμιση της μηχανής αναγνώρισης ομιλίας των XP, είναι για προσεκτική υπαγόρευση, δηλαδή πιο αργή και με μεγαλύτερη προσοχή στην άρθρωση των λέξεων, συγκριτικά με τον φυσικό λόγο των παρουσιαστών του CNN.

3.6. Οι χρήσεις της αναγνώρισης ομιλίας στο μέλλον

Εκτιμάται ότι μεταξύ 2010 με 2015 η τεχνολογία της αναγνώρισης ομιλίας από ηλεκτρονικό υπολογιστή, θα φτάσει το ανθρώπινο επίπεδο αναγνώρισης. Οι επιπτώσεις από αυτή την εξέλιξη θα είναι σημαντικές, και θα αλλάξουν τον τρόπο που χειριζόμαστε τον υπολογιστή μας. Παρακάτω παρουσιάζονται εκτιμήσεις σχετικά με την χρήση της στο μέλλον.

Οι φραγμοί της αλληλεπίδρασης, για τους ανθρώπους που δεν έχουν οικειότητα χρήσης ηλεκτρονικού υπολογιστή θα μειωθούν, με αποτέλεσμα συχνότερη παραγωγή, αναζήτηση και χρήση πληροφοριών. Η αύξηση της παραγωγής θα οδηγήσει στην δημιουργία τεράστιων ποσοτήτων πληροφορίας, όπως σκέψεις, ιδέες, αναμνήσεις. Η πληροφορία θα μπορεί να παρέχεται στο διαδίκτυο και σε προφορική μορφή.

3.6.1. Επικοινωνία

Μία εφαρμογή αποστολής στιγμιαίων γραπτών μηνυμάτων από υπολογιστή σε υπολογιστή (instant messenger), θα δέχεται την ομιλία του χρήστη από ένα μικρόφωνο. Όταν ο χρήστης πει «Γιάννη, τι κάνεις;», το σύστημα θα καταλαβαίνει ότι ο χρήστης θέλει να μιλήσει με τον χρήστη «Γιάννη», θα συνδέεται με αυτόν, και θα του στέλνει το κείμενο «τι κάνεις;». Και ο χρήστης «Γιάννης» θα απαντά πάλι με ομιλία, η οποία θα μετατρέπεται σε κείμενο και θα αποστέλλεται. Με την χρήση αναγνώρισης ομιλίας, καθίσταται ευκολότερη και πιο γρήγορη η σύνταξη μηνυμάτων.

3.6.2. Αναζήτηση συζητήσεων

Σήμερα είναι εφικτό κάποιος να κάνει αναζήτηση και να βρει χρήσιμες πληροφορίες, σε συζητήσεις που έχουν γίνει ήδη σε φόρουμ. Με την χρήση της αναγνώρισης ομιλίας είναι δυνατό να γίνονται αναζητήσεις σε όλες τις ανοικτές συνομιλίες μέσω φωνής σε πραγματικό χρόνο.

Για παράδειγμα, ας υποθέσουμε ότι κάποιος μελετά βιολογία, τον απασχολεί η θεωρία της εξέλιξης, και η συζήτηση με φωνή μέσω διαδικτύου με ένα άλλο φίλο του είναι δημόσια. Ένα πρόγραμμα καταγράφει σε κείμενο την συζήτηση σε πραγματικό χρόνο, ενώ ένα άλλο την ευρετηριοποιεί, ώστε να διευκολύνεται κάποια μελλοντική αναζήτηση. Εν τω μεταξύ, κάποιος άλλος ενδιαφέρεται για το ίδιο θέμα. Κάνει μία αναζήτηση, και ανακαλύπτει αυτή τη συζήτηση που τον ενδιαφέρει. Ίσως και η ίδια η εφαρμογή, τον ενημερώνει ότι κάποιοι συζητούν το ζήτημα που τον ενδιαφέρει. Θα μπορεί να παρακολουθήσει την συζήτηση, και αν οι αρχικοί συνομιλητές το επιθυμούν, να πάρει μέρος σε αυτή.

Επίλογος

Σε αυτό το κεφάλαιο έγινε αναφορά στις πολλές χρήσεις που η τεχνολογία αναγνώρισης ομιλίας έχει. Συγκεκριμένα, η αναγνώριση ομιλίας χρησιμοποιείται στα συστήματα IVR (Interactive Voice Response), συστήματα που επιτρέπουν σε ένα χρήστη να κάνει επιλογές πλοήγησης μέσα σε ένα μενού επιλογών. Τα IVR συστήματα χρησιμοποιούνται από εταιρείες κινητής τηλεφωνίας, από εταιρίες για κρατήσεις θέσεων σε αεροπλάνα και αλλού. Τα πλεονεκτήματα της χρήσης φωνητικών εντολών από αυτά τα συστήματα είναι γρηγορότερη εξυπηρέτηση, αφού ο χρήστης δεν χρειάζεται να περιμένει να τον εξυπηρετήσει κάποιος υπάλληλος, και καλύτερη εξυπηρέτηση, γιατί ο χρήστης μπορεί να πει τι χρειάζεται και το σύστημα μπορεί να τον προωθήσει στο κατάλληλο τμήμα.

Ακόμα μία χρήση είναι η αναγνώριση χρήστη, η οποία χρησιμοποιείται για ταυτοποίηση από απόσταση από τράπεζες, ασφαλιστικές εταιρείες, και είναι πολύ αποτελεσματική καθώς χρήστες μη εξοικειωμένοι με πληροφοριακά συστήματα

μπορούν να τα χρησιμοποιήσουν. Επίσης, δεν υπάρχει κίνδυνος υποκλοπής κωδικών.

Η τεχνολογία αναγνώρισης ομιλίας είναι πολύ χρήσιμη σε άτομα με αναπηρίες. Επιτρέπει σε άτομα που δεν μπορούν να χειριστούν το πληκτρολόγιο και το ποντίκι, λόγω αναπηρίας στα χέρια τους, λόγω προβλημάτων όρασης ή άλλων λόγων να χρησιμοποιήσουν ηλεκτρονικό υπολογιστή. Έχει βρεθεί, ότι τα άτομα με ειδικές ανάγκες που χρησιμοποιούν αναγνώριση ομιλίας εμφανίζουν βελτίωση στους τομείς της συγκέντρωσης, της ανάγνωσης και της γραφής.

Η αναγνώριση ομιλίας χρησιμοποιείται όμως και στο στρατιωτικό τομέα, στα μαχητικά αεροσκάφη και ελικόπτερα για έλεγχο των συστημάτων τους, και έτσι ο πιλότος μπορεί να επικεντρώνεται σε πιο σημαντικές εργασίες κατά την πτήση.

Στις εμπορικές εφαρμογές, που είναι και οι πιο δημοφιλείς, η αναγνώριση ομιλίας χρησιμοποιείται για τον έλεγχο λογισμικού. Για αυτό το λόγο χρησιμοποιείται σε ηλεκτρονικά παιχνίδια, σε GPS για να μην αφαιρείται ο οδηγός από τον έλεγχο της συσκευής, για υπαγορεύσεις κειμένων, και κάνει την διαδικασία σύνταξης ενός κειμένου γρηγορότερη. Η αναγνώριση ομιλίας χρησιμοποιείται στον τομέα των πολυμέσων για την αυτόματη δημιουργία υποτίτλων σε ένα αρχείο video. Αυτή είναι μία πολύ χρήσιμη εφαρμογή, καθώς λόγω του μεγάλου όγκου τέτοιων αρχείων στο διαδίκτυο, χρειάζεται μία αυτοματοποιημένη διαδικασία που να μπορεί να δημιουργήσει υπότιτλους. Μία άλλη εφαρμογή σε αυτό το τομέα είναι η αναζήτηση διαλόγων που υπάρχουν σε ένα αρχείο πολυμέσων με την χρήση ομιλίας.

Στο επόμενο κεφάλαιο παρουσιάζονται μερικές βασικές έννοιες των συστημάτων αναγνώρισης ομιλίας, που θα βοηθήσουν να κατανοηθεί ο τρόπος λειτουργίας τους.

Κεφάλαιο 4. Χαρακτηριστικά των συστημάτων αναγνώρισης ομιλίας

Εισαγωγή

Ένα σύστημα αναγνώρισης ομιλίας εκτός από την μηχανή αναγνώρισης ομιλίας περιέχει και άλλα τμήματα. Ένα από αυτά είναι το λεξικό. Το λεξικό είναι πολύ σημαντικό τμήμα, καθώς περιέχει τις λέξεις που μπορούν να αναγνωριστούν, και επηρεάζει την ακρίβεια του συστήματος. Αυτά τα ζητήματα θίγονται σε αυτό το κεφάλαιο.

4.1. Λεξικά

Τα λεξικά είναι κατάλογοι λέξεων ή εκφράσεων που μπορούν να αναγνωριστούν από το σύστημα αναγνώρισης ομιλίας. Ένα λεξικό περιέχει τον τρόπο προφοράς της έκφρασης σε ηχητική μορφή, και την έκφραση σε μορφή κειμένου.

Για να επιλεχτεί το κατάλληλο λεξικό, χρησιμοποιούνται συλλογές κειμένων (speech corpus) σχετικές με το περιβάλλον που θα χρησιμοποιηθεί το σύστημα αναγνώρισης ομιλίας σε συνδυασμό με το λεξιλόγιο της γλώσσας. Για να αναλυθούν οι συλλογές κειμένων χρειάζεται ένας tokenizer (ένα σύστημα που χρησιμοποιείται για να κατακερματίζει ένα κείμενο σε λέξεις). Οι συλλογές κειμένων χρησιμοποιούνται για να δημιουργηθεί το ακουστικό μοντέλο. Παραδείγματα τέτοιων συλλογών είναι το Isis Switchboard και το TIMIT.

Γενικά, είναι ευκολότερο ένα σύστημα να αναγνωρίσει φωνήματα από μικρότερα λεξικά. Αντίθετα από τα κανονικά λεξικά, κάθε λήμμα δεν είναι απαραίτητο να είναι μία ενιαία λέξη. Μπορούν να είναι μια πρόταση ή δύο. Τα μικρότερα λεξικά μπορούν να έχουν μόνο 1 ή 2 αναγνωρισμένες εκφράσεις, π.χ. «πάρε στο σπίτι», ενώ τα πολύ μεγάλα λεξικά μπορούν να έχουν αρκετές χιλιάδες.

Ένας από τους παράγοντες που επηρεάζουν τον αριθμό των λανθασμένων αναγνωρίσεων ενός συστήματος αναγνώρισης ομιλίας είναι ο αριθμός των

εκφράσεων που δεν βρίσκονται στο λεξικό του. Για αυτό το λόγο, ένα σημαντικό ζήτημα όταν δημιουργείται ένα γλωσσικό μοντέλο (βλέπε κεφ.5) είναι να επιλεχτεί ένα λεξικό που θα έχει την μεγαλύτερη δυνατή κάλυψη σε ένα κείμενο που δεν έχει τεθεί ξανά στην μηχανή αναγνώρισης.

Ο πίνακας 2 (Roukos, 1997) παρουσιάζει το ποσοστό ενός αγνώστου κειμένου στο οποίο ένα σύστημα αναγνώρισης ομιλίας κάνει σωστή αναγνώριση χρησιμοποιώντας ένα λεξικό του οποίου το μέγεθος είναι σταθερό, δηλαδή δεν προστίθενται σε αυτό νέες εκφράσεις με την διαδικασία της εκπαίδευσης, σε αναλογία με το μέγεθος του λεξικού. Η γλώσσα του κειμένου είναι η αγγλική.

Πίνακας 2. Ποσοστό κειμένου που γίνεται σωστή αναγνώριση σε σχέση με το μέγεθος του λεξικού

Μέγεθος του λεξικού	Ποσοστό κειμένου
20.000	94.1 %
64.000	98.7 %
100.000	99.3 %
200.000	99.4 %

Για ένα χρήστη ενός συστήματος αναγνώρισης ομιλίας, τα αποτελέσματα μπορούν να είναι καλύτερα, εάν το λεξικό είναι προσαρμοσμένο πάνω του, με βάση τα λεγόμενα του. Ο πίνακας 3 (Roukos, 1997) παρουσιάζει το ποσοστό ενός αγνώστου κειμένου στο οποίο ένα σύστημα αναγνώρισης ομιλίας κάνει σωστή αναγνώριση χρησιμοποιώντας ένα λεξικό με αρχικό μέγεθος 20.000 εκφράσεις, του οποίου το μέγεθος αυξάνεται με την προσθήκη νέων εκφράσεων που προκύπτουν κατά την εκπαίδευση. Ακόμα, ο πίνακας παρουσιάζει το μέγεθος του κειμένου που πρέπει να αναγνωρισθεί για να προστεθούν αυτές οι εκφράσεις στο λεξικό.

Πίνακας 3. Ποσοστό κειμένου που γίνεται σωστή αναγνώριση σε σχέση με ένα στατικό και ένα δυναμικό λεξικό

Αριθμός εκφράσεων που προστέθηκαν	Μέγεθος κειμένου που χρειάστηκε	Ποσοστό αναγνώρισης κειμένου με στατικό λεξικό	Ποσοστό αναγνώρισης κειμένου με δυναμικό λεξικό
100	1.800	93.4 %	94.5 %
400	12.800	94.8 %	97.5 %
3.100	81.600	94.8 %	98.1 %
6.400	211.000	94.4 %	98.9 %

Για παράδειγμα, στο λεξικό της μηχανής αναγνώρισης των Microsoft Windows, ο χρήστης μπορεί να εισάγει με απλό τρόπο νέες εκφράσεις, όπως ονόματα ή μη συχνά χρησιμοποιούμενες λέξεις. Η διαδικασία είναι η εξής:

1. Στη γραμμή γλώσσας επιλέγουμε **Εργαλεία** →**Προσθήκη/Διαγραφή** λέξεων.
2. Στο παράθυρο που ανοίγει, στο πεδίο **Word** γράφουμε την λέξη που θέλουμε να προστεθεί.
3. Στη συνέχεια πατάμε το πλήκτρο **Record** και προφέρουμε την λέξη.
4. Η λέξη έχει προστεθεί. Πιέζουμε **Close** και κλείνουμε το παράθυρο.

4.2. Ακρίβεια

Η ακρίβεια ενός συστήματος αναγνώρισης ομιλίας είναι ένα από τα βασικότερα χαρακτηριστικά του. Η ακρίβεια μετράται με τον ρυθμό εμφάνισης λάθος λέξεων που στα αγγλικά ορίζεται ως Word Error Rate (WER). Ο ρυθμός εμφάνισης λάθος λέξεων υπολογίζεται με βάση τον τύπο: $$WER = \frac{S + D + I}{N} \quad \textbf{(4.1)}$$

όπου,

- S: είναι ο αριθμός των αντικαταστάσεων,
- D: είναι ο αριθμός των διαγραφών,
- I: είναι ο αριθμός των εισαγωγών,
- N: είναι ο αριθμός των λέξεων που προφέρθηκαν.

Για παράδειγμα, αν η έκφραση που έπρεπε να αναγνωρισθεί ήταν η: «Οι περισσότερες περιοχές της Ελλάδας έχουν λιακάδα», και το αποτέλεσμα της αναγνώρισης ήταν η έκφραση: «Οι περισσότερες αρχές κοντά στην Ελλάδα έχουν λιακάδα», ο ρυθμός εμφάνισης λάθος λέξεων θα ήταν περίπου 57%. Αυτό το αποτέλεσμα προκύπτει καθώς υπάρχουν 3 αντικαταστάσεις, οι λέξεις «αρχές», «την», «Ελλάδα», και μία εισαγωγή, η λέξη «κοντά». Σύμφωνα με τον τύπο, 4/7 κάνει περίπου 0,57. Εκφρασμένο σε ποσοστό, αυτό το νούμερο είναι περίπου 57%. Πολλές φορές, η μέτρηση της ακρίβειας μετράται με τον ρυθμό αναγνώρισης λέξεων που στα αγγλικά μεταφράζεται ως Word Recognition Rate (WRR) και ορίζεται από τον τύπο:

$$WRR = 1 - WER = \frac{H - I}{N}$$, όπου το H είναι το N-(S+D). **(4.2)**

Ένα θεωρητικό πρόβλημα όταν υπολογίζεται η ακρίβεια ενός συστήματος είναι εάν το λανθασμένο αποτέλεσμα της αναγνώρισης προέκυψε από λάθος της μηχανής αναγνώρισης ομιλίας ή σε λάθος κατά την προφορά της λέξης από τον χρήστη, π.χ. επειδή η γλώσσα που ομιλείται δεν είναι η μητρική γλώσσα του χρήστη, ή επειδή έχει βαριά προφορά.

Οι δυνατότητες ενός συστήματος αναγνώρισης ομιλίας μπορούν να αξιολογηθούν με τη μέτρηση της ακρίβειας του στην αναγνώριση εκφράσεων. Αυτό περιλαμβάνει όχι μόνο να αναγνωρίσει σωστά μια έκφραση αλλά και να προσδιορίσει εάν η έκφραση είναι στο λεξικό του. Στις αρχές της δεκαετίας του '90, τα καλύτερα συστήματα είχαν ακρίβεια 85% σε έναν σχετικά μικρό λεξικό των 20.000 λέξεων. Σήμερα, τα υψηλής ποιότητας συστήματα έχουν ακρίβεια 98% ή περισσότερο, αν και αυτό μπορεί να ποικίλει πολύ μεταξύ των ομιλητών. Η αποδεκτή ακρίβεια ενός συστήματος εξαρτάται από την εφαρμογή.

4.2.1. Παράγοντες που επηρεάζουν την ακρίβεια

Κανένα σύστημα αναγνώρισης ομιλίας δεν είναι 100% ακριβές. Διάφοροι παράγοντες μπορούν να μειώσουν την ακρίβεια. (Speech recognition-Weaknesses and Flaws, www.howstuffworks.com) Μερικοί από αυτούς τους παράγοντες είναι ζητήματα που συνεχίζουν να βελτιώνονται καθώς η τεχνολογία βελτιώνεται. Άλλοι μπορούν να ελαττωθούν από τον ίδιο τον χρήστη.

4.2.1.1. Χαμηλή αναλογία σήματος προς θόρυβο(SNR)

Η ακρίβεια επηρεάζεται από τον λόγο του ηχητικού σήματος προς τον θόρυβο. Συνεπώς, εάν το ηχητικό σήμα που θέλουμε να αναγνωρισθεί έχει παρόμοια ισχύ με τον θόρυβο, το αποτέλεσμα της αναγνώρισης δεν θα είναι καλό. Τα φωνήματα πρέπει να εκφέρονται ευδιάκριτα, και οποιοσδήποτε πρόσθετος θόρυβος που εισάγεται στον ήχο δημιουργεί προβλήματα στην αναγνώριση. Ο θόρυβος μπορεί να προέλθει από διάφορες πηγές, συμπεριλαμβανομένου του δυνατού παρασιτικού θορύβου σε ένα περιβάλλον γραφείου. Οι χρήστες πρέπει να χρησιμοποιήσουν το σύστημα αναγνώρισης ομιλίας σε ένα δωμάτιο με ησυχία, με ένα μικρόφωνο υψηλής ποιότητας που πρέπει να τοποθετείται όσο γίνεται πιο κοντά στο στόμα του ομιλητή. Χαμηλής ποιότητας κάρτες ήχου, που παρέχουν την θύρα για το μικρόφωνο ώστε να σταλεί το σήμα στον υπολογιστή, συχνά δεν έχουν αρκετό προστατευτικό κάλυμμα για να προστατευθεί το σήμα από άλλα ηλεκτρικά σήματα που παράγονται από άλλα τμήματα του υπολογιστή. Έτσι μπορεί να εισαχθεί βόμβος ή συριγμός (σφύριγμα) στο σήμα.

4.2.1.2. Ισχύς του υπολογιστικού συστήματος

Η εκτέλεση των στατιστικών μοντέλων που απαιτούνται για τη αναγνώριση ομιλίας προσθέτει στο επεξεργαστή του υπολογιστή μεγάλο φόρτο. Οι γρηγορότεροι προσωπικοί υπολογιστές σήμερα μπορεί να έχουν μεγάλο χρόνο απόκρισης με τις περίπλοκες εντολές ή φράσεις ενώ τα λεξικά που απαιτούνται από τα προγράμματα απορροφούν ένα μεγάλο μέρος του σκληρού δίσκου.

4.2.1.3. Ομώνυμα

Τα ομώνυμα είναι δύο λέξεις που συλλαβίζονται διαφορετικά και έχουν διαφορετικές έννοιες αλλά εκφέρονται με τον ίδιο τρόπο. Π.χ. «πιάνο» και «πιάνω». Δεν υπάρχει κανένας τρόπος για ένα σύστημα αυτόματης αναγνώρισης ώστε να αναγνωριστεί η διαφορά μεταξύ αυτών των λέξεων βασιζόμενο στον ήχο μόνο. Εντούτοις, η εκπαίδευση των συστημάτων και τα στατιστικά πρότυπα που

λαμβάνουν υπόψη τα συμφραζόμενα των λέξεων έχουν βελτιώσει πολύ την απόδοση τους.

4.3. Εκπαίδευση

Μερικά συστήματα αναγνώρισης ομιλίας έχουν τη δυνατότητα να προσαρμοστούν σε έναν ομιλητή. Το σύστημα εκπαιδεύεται με την επανάληψη από τη πλευρά του χρήστη φράσεων που χρησιμοποιούνται συχνά. Ακόμα, το σύστημα αναλύει και προσαρμόζεται στον τρόπο ομιλίας του χρήστη. Με την εκπαίδευση το σύστημα αναγνώρισης βελτιώνει την ακρίβεια του. Για παράδειγμα, αν το σύστημα αναγνώρισης ομιλίας πρόκειται να χρησιμοποιηθεί για ιατρικούς σκοπούς, θα πρέπει κατά την εκπαίδευση του συστήματος να διαβαστούν κείμενα σχετικά με ιατρικά θέματα. Έτσι το σύστημα θα είναι πιθανότερο να επιστρέψει το σωστό αποτέλεσμα αν ο γιατρός έχει πει μία ιατρική έκφραση, σε σχέση με τον αν λεγόταν η ίδια έκφραση σε ένα σύστημα που το έχει εκπαιδεύσει ένας δημοσιογράφος που ασχολείται με το πολιτικό ρεπορτάζ.

Χάρη στην εκπαίδευση, ένα σύστημα μπορεί να χρησιμοποιηθεί από ομιλητές που έχουν δυσκολίες στην ομιλία. Με δεδομένο ότι ο ομιλητής επαναλαμβάνει μια έκφραση με παρόμοιο τρόπο, το σύστημα είναι σε θέση να προσαρμοσθεί πάνω σε αυτόν τον ομιλητή και να αναγνωρίσει τις ιδιότυπα εκφρασμένες λέξεις ή φράσεις. Περισσότερα για την εκπαίδευση αναφέρονται στο κεφάλαιο 5.

Επίλογος

Σε αυτό το κεφάλαιο αναφέρθηκαν διάφορες βασικές έννοιες των συστημάτων αναγνώρισης ομιλίας. Ένα από τα πιο σημαντικά συστατικά ενός συστήματος αναγνώρισης ομιλίας είναι τα λεξικά. Τα λεξικά είναι κατάλογοι λέξεων ή εκφράσεων που μπορούν να αναγνωριστούν από το σύστημα αναγνώρισης ομιλίας. Ένα λεξικό περιέχει τον τρόπο προφοράς της έκφρασης σε ηχητική μορφή, και την έκφραση σε μορφή κειμένου. Η ακρίβεια ενός συστήματος αναγνώρισης ομιλίας, είναι ένα μέγεθος που απεικονίζει το ποσοστό του αριθμού των λανθασμένων αναγνωρίσεων. Προφανώς, αν η ακρίβεια δεν είναι καλή, ένα σύστημα αναγνώρισης ομιλίας δεν μπορεί να χρησιμοποιηθεί και να είναι

αποδοτικό για την χρήστη. Τα σημερινά συστήματα σε ιδανικές συνθήκες επιτυγχάνουν ακρίβεια μεγαλύτερη του 98%.

Παρουσιάζονται επίσης, διάφοροι παράγοντες που μπορούν να μειώσουν την ακρίβεια του συστήματος, εκτός από τον τρόπο ομιλίας του χρήστη. Οι παράγοντες που επηρεάζουν είναι τα υψηλά επίπεδα θορύβου στο περιβάλλον που γίνεται η αναγνώριση, η υπολογιστική ισχύς, καθώς πρόκειται για διαδικασία που απαιτεί πόρους, και τέλος τα ομώνυμα, που είναι διάφορες λέξεις που έχουν το ίδιο άκουσμα, όμως έχουν εντελώς διαφορετική σημασία, όπως οι λέξεις «πιάνω» και «πιάνο». Όμως αυτά τα ζητήματα λύνονται κατά ένα μεγάλο ποσοστό από την εκπαίδευση. Αφού παρουσιάστηκαν όλα αυτά τα ζητήματα σχετικά με τα συστήματα αναγνώρισης ομιλίας, στο επόμενο κεφάλαιο παρουσιάζεται ο τρόπος λειτουργίας αυτών των συστημάτων.

Κεφάλαιο 5. Πως λειτουργεί η αυτόματη αναγνώριση ομιλίας

Εισαγωγή

Όπως είδαμε και παραπάνω, η προσπάθεια του ανθρώπου να δημιουργήσει software ικανό να αναγνωρίζει την ανθρώπινη ομιλία ξεκινά από τα δεκαετία του ’60. Μισό αιώνα αργότερα, η τεχνολογία αναγνώρισης ομιλίας έχει φτάσει σε ένα σημείο ώστε να μπορεί να αποτελέσει ένα σημαντικό εργαλείο για τους ανθρώπους. Σε αυτό το κεφάλαιο, παρουσιάζεται ο τρόπος λειτουργίας αυτής της τεχνολογίας.

5.1. Φωνήματα

Τα φωνήματα (phonemes) αποτελούν τα ελάχιστα στοιχεία/μονάδες μίας γλώσσας, που παρέχουν διαφοροποιητική λειτουργία στο φωνητικό επίπεδο, για το νόημα του γλωσσικού ήχου. Τα φωνήματα είναι τμηματικές και ασυνεχείς μονάδες, περιορισμένου αριθμού σε κάθε γλώσσα. (φώνημα, www.greek-language.gr) Κάθε φώνημα έχει μοναδική σημασία για το σύστημα αναγνώρισης ομιλίας. Αν π.χ. στη λέξη *θέμα* [thema] αντικαταστήσουμε το αρχικό [θ] με το [δ] θα προκύψει μια λέξη με διαφορετική σημασία, *δέμα* [Dhema]. Σε αυτή την περίπτωση, που η αντικατάσταση ενός φθόγγου με έναν άλλο οδηγεί σε σημασιολογική διαφοροποίηση, λέμε ότι οι δύο αυτοί φθόγγοι έχουν *διακριτική λειτουργία* και αντιπροσωπεύουν δυο διαφορετικά φωνήματα.

Γενικά, ένας φθόγγος αναγνωρίζεται ως ξεχωριστό φώνημα όταν με τη χρήση του στη θέση ενός άλλου φθόγγου δημιουργείται μια διαφορετική λέξη. Αν, αντίθετα, δυο φθόγγοι δεν μπορούν να αντικαταστήσουν ο ένας τον άλλο και να προκύψει διαφοροποίηση σημασίας, αλλά ο καθένας εμφανίζεται σε φωνητικά περιβάλλοντα όπου δεν μπορεί να εμφανιστεί ο άλλος, όπως π.χ. οι φθόγγοι [c] και [k] της ελληνικής γλώσσας στις λέξεις *κιλό* [ci-lo], *καλά* [ka-la], λέμε ότι οι δυο αυτοί φθόγγοι δεν αντιπροσωπεύουν διαφορετικά φωνήματα, αλλά αποτελούν *αλλόφωνα* ενός και του αυτού φωνήματος, δηλαδή διαφορετικές προφορές του

ίδιου φωνήματος που οφείλονται στην επίδραση του διαφορετικού φωνητικού περιβάλλοντος στο οποίο εμφανίζονται. Τα φωνήματα τοποθετούνται σε πλάγιες καθέτους (//), ενώ οι παραλλαγές ή αλλόφωνα τοποθετούνται σε ορθές αγκύλες ([ ]). Τα αλλόφωνα είναι ένα ζήτημα που τα συστήματα αναγνώρισης ομιλίας λαμβάνουν υπόψη.

Τα φωνήματα είναι αφηρημένες και ασυνεχείς μονάδες που είναι εγγεγραμμένες στη συνείδηση των ομιλητών. Οι αφηρημένες αυτές μονάδες πραγματώνονται κατά την ομιλία με τη μορφή των φθόγγων, που είναι φυσικές οντότητες αντιληπτές με τις αισθήσεις. Παρόλο που τα φωνήματα είναι αφηρημένες μονάδες, μη αντιληπτές με τις αισθήσεις, οι ομιλητές, όταν χρησιμοποιούν τη γλώσσα, έχουν συνείδηση των φωνημάτων και όχι των πραγματώσεών τους που είναι οι φθόγγοι. Έτσι εξηγείται πώς οι έλληνες ομιλητές, ενώ αντιλαμβάνονται ότι τα [θ] και [δ] στις λέξεις *θέμα* και *δέμα* είναι διαφορετικά, δεν αντιλαμβάνονται καμιά διαφορά ανάμεσα στα [c] και [k] που είναι οι αρχικοί φθόγγοι στις λέξεις *κιλό* και *καλά* αντίστοιχα. Στη συνείδησή τους και οι δυο λέξεις αρχίζουν με το φώνημα /k/ και αυτό αντιλαμβάνονται. Αυτό οφείλεται στο γεγονός ότι οι μονάδες που συμβάλλουν στη διαφοροποίηση των μηνυμάτων είναι τα φωνήματα και όχι οι πραγματώσεις τους που είναι οι φθόγγοι.

Τα φωνήματα διαφοροποιούνται μεταξύ τους στη βάση των διακριτικών τους χαρακτηριστικών. Κάθε φώνημα απαρτίζεται από έναν αριθμό διακριτικών χαρακτηριστικών με τα οποία διαφοροποιείται από τα άλλα φωνήματα της γλώσσας στην οποία ανήκει. Τα διακριτικά χαρακτηριστικά που συγκροτούν ένα φώνημα δεν είναι τμηματικές μονάδες, δηλαδή δεν πραγματώνονται διαδοχικά το ένα μετά το άλλο, αλλά εμφανίζονται όλα μαζί ταυτόχρονα. Κάθε φώνημα διαφέρει από τα άλλα φωνήματα με τα οποία ανήκει στο ίδιο φωνολογικό σύστημα (π.χ. τα φωνήματα της ελληνικής γλώσσας) ως προς ένα ή περισσότερα διακριτικά χαρακτηριστικά. Το /p/ π.χ., που είναι χειλικό, κλειστό και άηχο, διαφέρει από το /b/, που είναι χειλικό, κλειστό και ηχηρό, ως προς ένα διακριτικό χαρακτηριστικό που είναι η ηχηρότητα.

Η Διεθνής Φωνητική Ενωση (International Phonetic Association) έχει δημιουργήσει το Διεθνές Φωνητικό Αλφάβητο (International Phonetic Alphabet) το οποίο περιλαμβάνει όλα τα φωνήματα όλων των γλωσσών του πλανήτη. Η Ελληνική γλώσσα έχει περίπου 30 φωνήματα, ενώ η αγγλική περίπου 50.

Πίνακας 4. Τα ελληνικά φωνήματα

Φωνήματα	Γραφήματα	Παράδειγμα
/a/	α	κ**α**λός
/e/	ε, αι	**έ**χω, π**αί**ζω
/i/	ι, η, υ, ει, οι, υι	πάλ**ι**, ζάλ**η**, μ**ύ**λος, **εί**μαι, **οί**κος, **υι**ός
/o/	ο, ω	**ό**χι, **ώ**ρα
/u/	ου	**ού**τε
/j/	ι, η, ει, οι	π**ι**α, κελά**η**δημα, αλήθ**ει**α, π**οι**ος
/p/	π, ππ, ψ	κό**π**ος, ί**ππ**ος, **ψ**άρι
/b/	μπ	**μπ**ορώ
/f/	φ, (α)υ, (ε)υ, (ε)υφ	α**φ**ήνω, α**υ**τός, ε**υ**τυχία, ε**ύφ**ορος
/v/	β, ββ, (α)υ, (ε)υ, (ε)υβ	λα**β**ή, Σά**ββ**ατο, α**ύ**ριο, ε**ύ**λογος, ε**υβ**οϊκός
/t/	τ, ττ	κά**τ**ι, α**ττ**ικός
/d/	ντ	**ντ**ύνω
/θ/	θ	α**θ**ώος
/δ/	δ, δδ	παι**δ**ί, Σα**δδ**ουκαίοι
/k/	κ, κκ, ξ	**κ**άπως, έ**κκ**ληση, **ξ**ύλο
/g/	γκ, γγ	**γκ**αρίζω, ε**γγ**ύηση
/x/	χ	**χ**άρη
/γ/	γ	α**γ**όρι
/m/	μ, μμ	αί**μ**α, ά**μμ**ος
/n/	ν, νν	έ**ν**ας, ε**νν**έα
/l/	λ, λλ	πά**λ**ι, ά**λλ**ος
/r/	ρ, ρρ	αέ**ρ**ας, ά**ρρ**ωστος

Φωνήματα	Γραφήματα	Παράδειγμα
/s/	σ, σσ, ς, ξ, ψ	ε**σ**ύ, γλώ**σσ**α, θε**ς**, **ξ**ύλο, **ψ**άρι
/z/	ζ, σ	**ζ**ωή, κό**σ**μος

Πίνακας 5. Συνδυασμοί των ελληνικών φωνημάτων

Συνδυασμοί Φωνημάτων	Γραφήματα	Παράδειγμα
/ks/	ξ, κσ, κς	έ**ξ**ι, ε**κσ**τρατεία, ταν**κς**
/ps/	ψ, πσ, πς	**ψ**άρι, κλι**πσ**άκι, κλι**πς**
/mb/	μπ	κά**μπ**ος
/nd/	ντ	έ**ντ**ιμος
/ng/	γκ, γγ	έ**γκ**υος, ε**γγ**ονός

5.2. Hidden Markov Models

Πριν γίνει αναφορά στον τρόπο που λειτουργεί η αναγνώριση ομιλίας, είναι σκόπιμο να αναφερθεί ένα χρήσιμο μαθηματικό εργαλείο που χρησιμοποιείται για την αναγνώριση των φωνημάτων και των λέξεων. Το εργαλείο αυτό είναι τα Hidden Markov Models (HMM).

Τα Hidden Markov Models είναι ένα ισχυρό στατιστικό εργαλείο, για την μοντελοποίηση ακολουθιών, στα οποία μία διαδικασία δημιουργεί μία καταφανή ακολουθία (Phil Blunsom, 2004).

Τα HMM αποτελούνται από καταστάσεις (states). Ας ορίσουμε ένα σύνολο καταστάσεων $S=\{S_0, S_1, ..., S_t, ...\}$, που παίρνουν τιμές που ανήκουν στο σύνολο των φυσικών αριθμών, δηλαδή από 1 έως N. Για κάθε μία κατάσταση (εκτός από την αρχική) ισχύει:

$$P(S_i \mid S_{i-1}) \qquad \textbf{(5.1)}$$

Το οποίο σημαίνει ότι η δεσμευμένη πιθανότητα κάποιας κατάστασης εξαρτάται μόνο από την προηγούμενη. Έτσι, συμβολίζουμε με $S_t=j$ την κατάσταση j που η διαδικασία βρίσκεται την χρονική στιγμή t. Επειδή η διαδικασία μετακινείται από μία κατάσταση σε μία άλλη οι δεσμευμένες πιθανότητες:

$$a_{ij} = P(S_t = j \mid S_{t-1} = i) \qquad \text{για κάθε } i,j \in S. \qquad \textbf{(5.2)}$$

ονομάζονται πιθανότητες μετάβασης (transition probabilities). Οι πιθανότητες αυτές μπορούν να αναπαρασταθούν σε ένα πίνακα,

Πίνακας 6. Πιθανότητες μετάβασης *a*

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix}$$

όπου N είναι ο αριθμός των καταστάσεων. Σε κάθε κατάσταση, το μοντέλο επιστρέφει μία πιθανότητα. Οι πιθανότητες αυτές λέγονται παρατηρήσεις (observations), και συμβολίζονται με το γράμμα O. Οι πιθανότητες αυτές μπορούν να αναπαρασταθούν σε ένα πίνακα,

Πίνακας 7. Επιστρεφόμενες πιθανότητες *b*

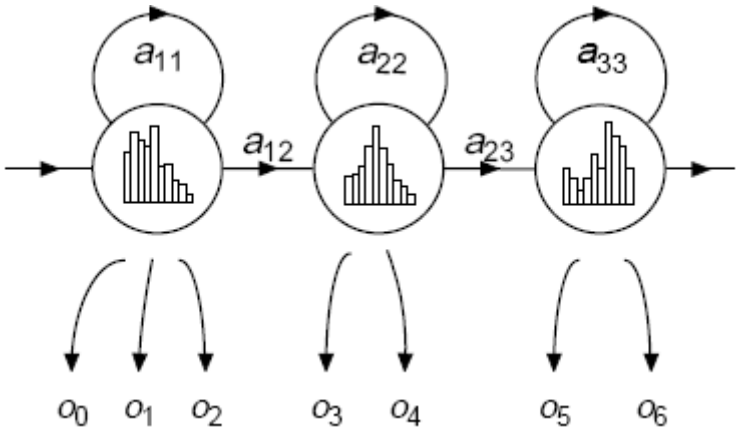
$$\mathbf{B} = \begin{pmatrix} b_1(1) & b_1(2) & \cdots & b_1(M) \\ b_2(1) & b_2(2) & \cdots & b_2(M) \\ \vdots & \vdots & \ddots & \vdots \\ b_N(1) & b_N(2) & \cdots & b_N(M) \end{pmatrix}$$

όπου $b_i(k)$, είναι η πιθανότητα της τιμής k να προέρχεται από μία κατάσταση S. Ισχύει δηλαδή:

$$b_i(k) = P(O_t = k \mid S_t = i) \qquad \text{για κάθε } i \in S,\ k \in \text{O και } t \geq 0. \qquad \textbf{(5.3)}$$

το οποίο σημαίνει ότι κάθε παρατήρηση *O,* εξαρτάται από την κατάσταση *S* που προέρχεται αυτή η παρατήρηση σε κάποια χρονική στιγμή *t*.

Σε διακριτές χρονικές στιγμές, το μοντέλο αλλάζει κατάσταση, σύμφωνα με τις πιθανότητες μετάβασης *A*. Σε κάθε κατάσταση παράγεται μια παρατήρηση $o \in O$. Ένας παρατηρητής μπορεί να δει μόνο τις παραγόμενες παρατηρήσεις, και όχι τις καταστάσεις από τις οποίες προήλθαν. Για αυτό και ο χαρακτηρισμός Hidden (κρυφά). Στην εικόνα 5 παρουσιάζεται ένα Hidden Markov Model.



Εικόνα 5. Ένα Hidden Markov Model

Τα Hidden Markov Models είναι χρήσιμα στην αναγνώριση προτύπων, και για αυτό το λόγο χρησιμοποιούνται στην αναγνώριση ομιλίας για αναγνώριση φωνημάτων και λέξεων. Αξίζει να αναφερθεί ότι και νευρωνικά δίκτυα χρησιμοποιούνται από τα συστήματα αναγνώρισης ομιλίας για τον ίδιο σκοπό.

5.3. Αρχιτεκτονική ενός συστήματος αναγνώρισης ομιλίας

Η αρχιτεκτονική των πρώτων συστημάτων αναγνώρισης ομιλίας ήταν απλή. Χρησιμοποιούσαν μία μέθοδο γνωστή ως ταίριασμα προτύπων (template matching). Το σύστημα είχε μία βάση δεδομένων με πρότυπα λέξεων. Έτσι, το σήμα της φωνής συγκρινόταν με τα πρότυπα και η εγγραφή που ήταν πιο κοντά στο σήμα ήταν το αποτέλεσμα της αναγνώρισης. Αυτή η αρχιτεκτονική απέδιδε σχετικά καλά για μικρά λεξικά. Για μεγάλα λεξικά όμως, ήταν αδύνατο να λειτουργήσει, γιατί δεν γίνεται να δημιουργηθούν τόσο πολλά πρότυπα και το

υπολογιστικό κόστος για την εύρεση του καλύτερου ταιριάσματος είναι πολύ μεγάλο.

Σήμερα, η αρχιτεκτονική ενός συστήματος αναγνώρισης ομιλίας είναι διαφορετική. Η βασική αρχή είναι ότι οι λέξεις παράγουν σήματα. Ο σκοπός της αναγνώρισης ομιλίας είναι να αναγνωρισθούν οι λέξεις που παρήγαγαν αυτό το σήμα. Αυτός ο σκοπός δεν είναι εύκολα επιτεύξιμος, γιατί η σχέση μεταξύ λέξεων και σημάτων δεν είναι ντετερμινιστική. Η χρήση μαθηματικών, βοηθά στην ανάλυση και κατανόηση του προβλήματος.

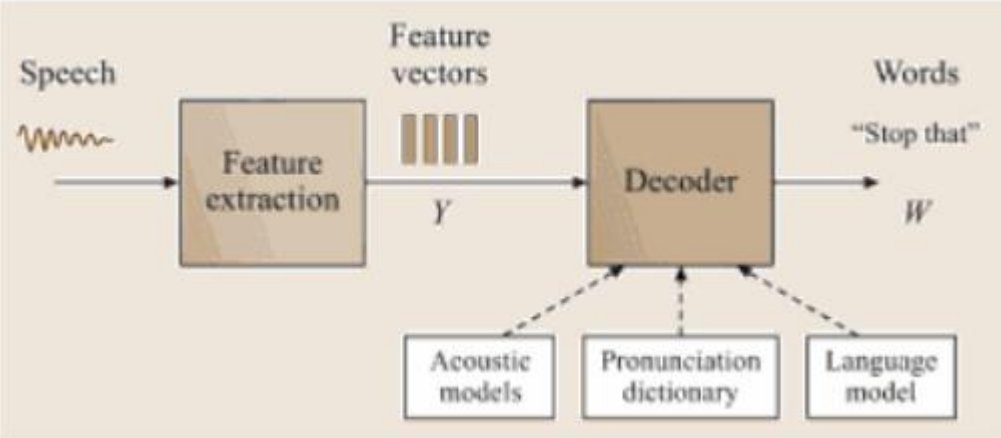
Τα βασικά μέρη ενός συστήματος αναγνώρισης ομιλίας (Young, 2008) παρουσιάζονται στην εικόνα 6. Το αναλογικό σήμα της ανθρώπινης ομιλίας μέσω ενός μικροφώνου εισάγεται στο σύστημα, όπου τεμαχίζεται σε πλαίσια ήχου (speech frames), $Y= y_1,\ldots, y_r$. Στη συνέχεια, ο αποκωδικοποιητής (decoder), προσπαθεί να βρει την ακολουθία των λέξεων $W= w_{1,\ldots,} w_r$ που είναι πιο πιθανό να έχουν δημιουργηθεί από τα *Y*. Δηλαδή, μαθηματικά, ο αποκωδικοποιητής προσπαθεί να βρει:

$$\hat{W} = \arg\max_{w}[p(W \mid Y)] \qquad \textbf{(5.4)}$$

Όμως, επειδή η δεσμευμένη πιθανότητα *p(W|Y)*, είναι δύσκολο να μοντελοποιηθεί, διότι υπάρχουν πολλά ζεύγη που ικανοποιούν το *p(W,Y),* το θεώρημα του Bayes χρησιμοποιείται ώστε να γίνει η μετατροπή του κανόνα 5.4 σε μορφή ευκολότερη για μοντελοποίηση.

$$\hat{W} = \arg\max_{W}[p(Y \mid W)\,p(W)] \qquad \textbf{(5.5)}$$

Η πιθανότητα *p(Y|W)*, που εκφράζει την πιθανότητα ότι όταν μία ακολουθία λέξεων *W* έχει ειπωθεί, τα ηχητικά πλαίσια *Y* θα έχουν παρατηρηθεί, υπολογίζεται από το ακουστικό μοντέλο (acoustical model). Η πιθανότητα *p(W),* που εκφράζει την πιθανότητα κάποια ακολουθία λέξεων να έχει ειπωθεί, υπολογίζεται από το γλωσσικό μοντέλο (language model). Παρακάτω, αναλύεται αυτή η διαδικασία.

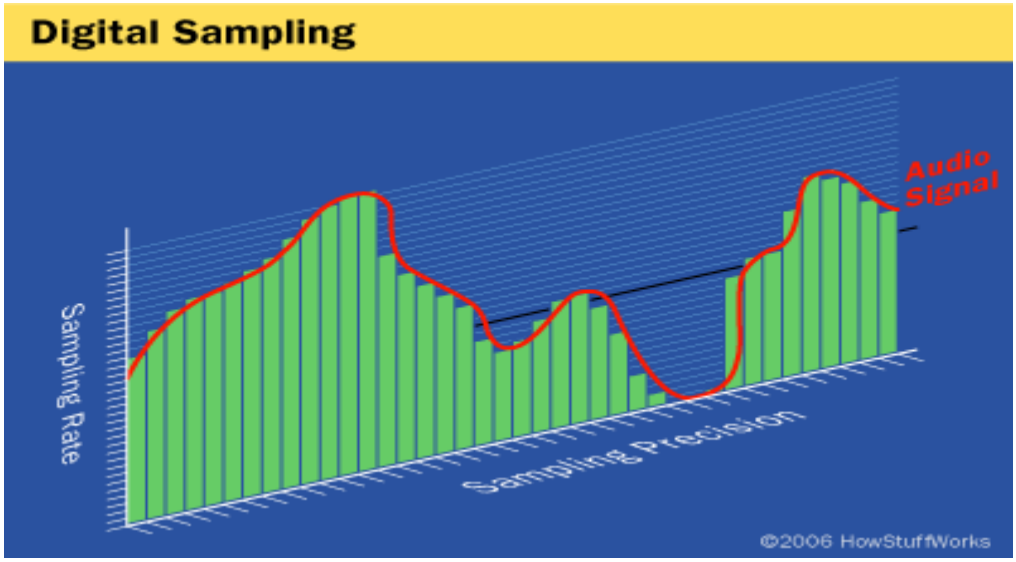


Εικόνα 6. Αρχιτεκτονική ενός συστήματος αναγνώρισης ομιλίας

5.4. Επεξεργασία του σήματος

5.4.1. Μετατροπή του αναλογικού σήματος της φωνής μας σε ψηφιακό

Κατά την ομιλία, δημιουργούνται δονήσεις στον αέρα. Αυτές οι δονήσεις είναι ο ήχος που ακούμε. Από την γραφική αναπαράσταση της συχνότητας της φωνής σε σχέση με τον χρόνο, προκύπτει ένα αναλογικό σήμα. Το πρώτο στάδιο στην διαδικασία αυτόματης αναγνώρισης ομιλίας είναι να μετατραπεί το αναλογικό σήμα της ομιλίας μας σε ψηφιακή μορφή, αναγνωρίσιμη από τον υπολογιστή. Αυτή η μετατροπή γίνεται από την κάρτα ήχου του υπολογιστή μας. Η πιο συνηθισμένη συχνότητα δειγματοληψίας είναι τα 16.000 Hz. Συνεπώς, το ψηφιακό σήμα είναι ένα σύνολο από πλάτη τα οποία παράγονται με ρυθμό 16.000 περίπου το δευτερόλεπτο. Το σήμα της φωνής είναι ένα διάνυσμα πολύπλοκο και πλούσιο επειδή παρέχει πληροφορίες σχετικές με το αποδιδόμενο μήνυμα αλλά και πληροφορίες σχετικές με τον ομιλητή (χροιά φωνής, συναισθηματική κατάσταση ομιλητή κ.α.) Η Εικόνα 7 παρουσιάζει τον τρόπο μετατροπής του σήματος από αναλογικό σε ψηφιακό.



Εικόνα 7. Μετατροπή του αναλογικού σήματος ομιλίας σε ψηφιακό

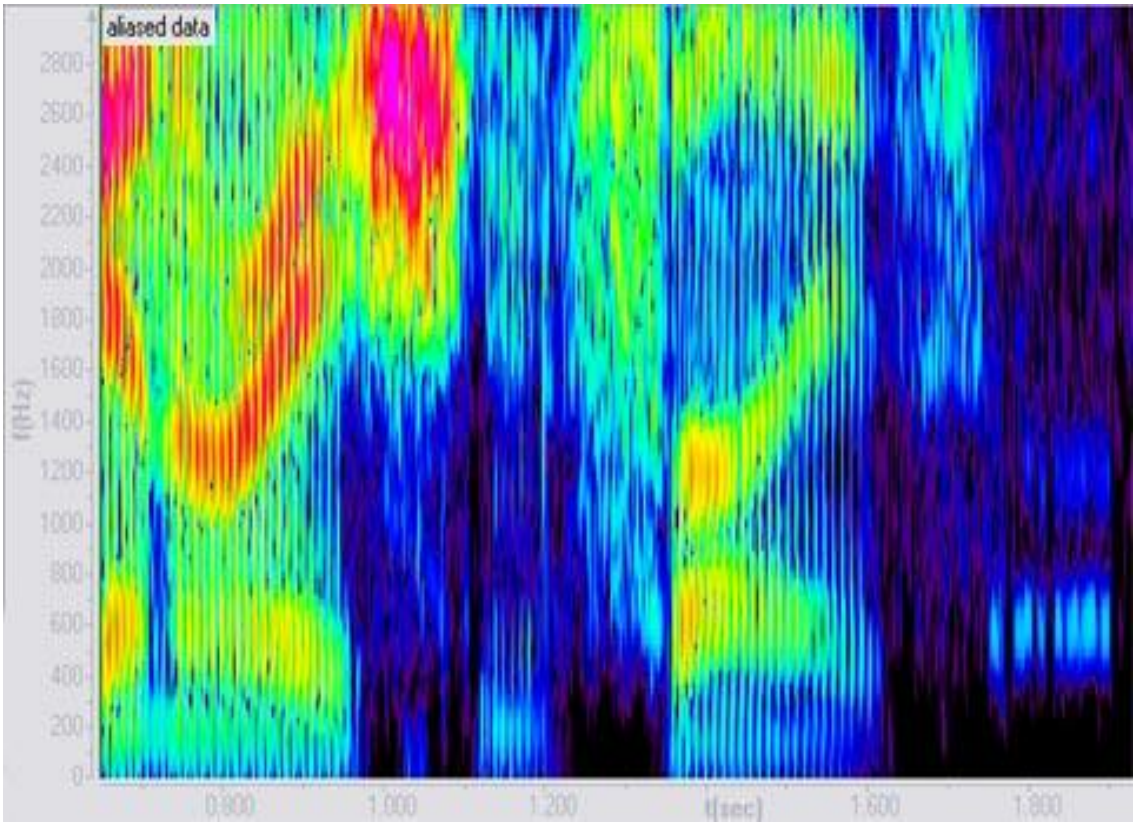
5.4.2. Κωδικοποίηση του σήματος

Το ψηφιακό σήμα είναι η είσοδος μίας μηχανής αναγνώρισης ομιλίας. Το σήμα όμως, χρειάζεται να υποστεί περαιτέρω επεξεργασία γιατί περιέχει πληροφορίες που δεν χρειάζονται, και μπορούν να προκαλέσουν προβλήματα κατά την αναγνώριση, αλλά και για να συμπιεστεί. Το πιο σημαντικό χαρακτηριστικό που πρέπει να εξαχθεί από ένα σήμα ομιλίας είναι ο προσδιορισμός της ενέργειας (δηλαδή της πλάτους), που έχει το σήμα σε κάποια συχνότητα σε μία χρονική στιγμή. Πρώτα, το σύστημα φιλτράρει τον ήχο, ώστε να απομακρυνθεί ο θόρυβος. Στη συνέχεια, περνώντας την ψηφιακή κυματομορφή μέσα από ένα γρήγορο μετασχηματισμό Fourier (Fast Fourier Transform-FFT) αναλύεται η κυματομορφή στο πεδίο της συχνότητας. Αυτό γίνεται με τον τύπο:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi k \frac{n}{N}} \qquad k = 0, \ldots, N-1. \tag{5.6}$$

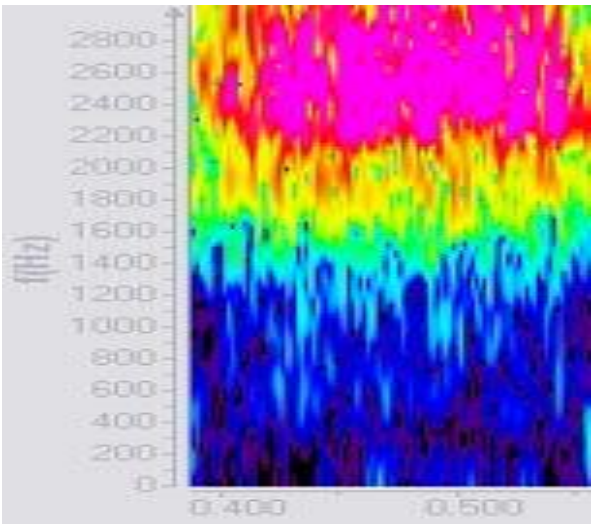
Με την εφαρμογή του γρήγορου μετασχηματισμού Fourier αναλύεται κάθε 1/100 του δευτερολέπτου από το ψηφιακό σήμα παράγοντας ένα φάσμα συχνοτήτων, το οποίο περιγράφει τον ήχο που ακούστηκε για κάθε 1/100 του δευτερολέπτου. Για να γίνουν κατανοητά όλα αυτά, είναι καλύτερο να δούμε ένα φασματογράφημα ομιλίας. Ένας φασματογράφημα είναι μια τρισδιάστατη γραφική

παράσταση της συχνότητας και του πλάτους (δηλαδή της έντασης) ενός κύματος έναντι του χρόνου. Η συχνότητα απεικονίζεται στο κάθετο άξονα, ο χρόνος στον οριζόντιο και το πλάτος απεικονίζεται με διαφορετικό χρώμα, και ισχύει ότι για μεγαλύτερο πλάτος το χρώμα είναι πιο έντονο. Είναι χρήσιμος ώστε να κατανοηθεί καλύτερα η διαδικασία μετατροπής του σήματος και το αποτέλεσμα της. Στην Εικόνα 8 απεικονίζεται ένα φασματογράφημα ομιλίας των αγγλικών λέξεων "Generation5".



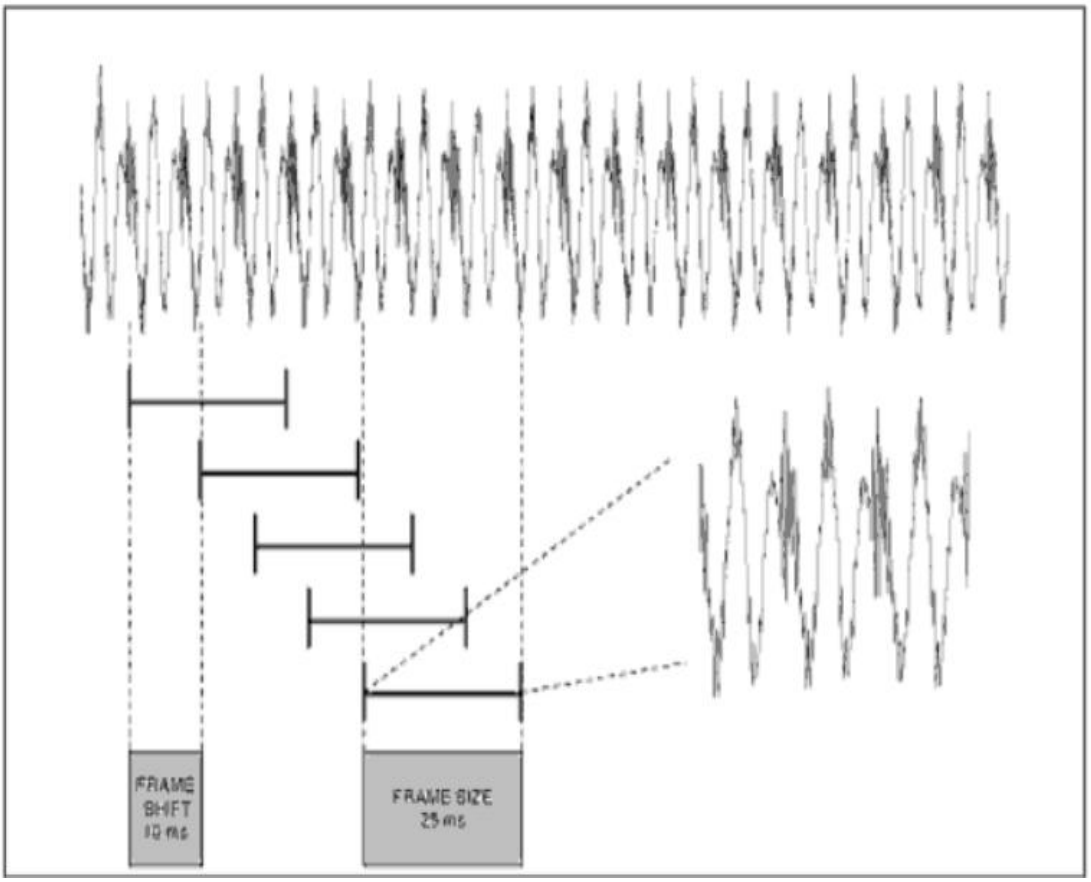
Εικόνα 8. Φασματογράφημα των αγγλικών λέξεων "Generation 5"

Στην Εικόνα 9 παρουσιάζεται το φασματογράφημα του αγγλικού φωνήματος “ss”, που περιέχεται στη λέξη “a**ss**ure” (διαβεβαιώνω).

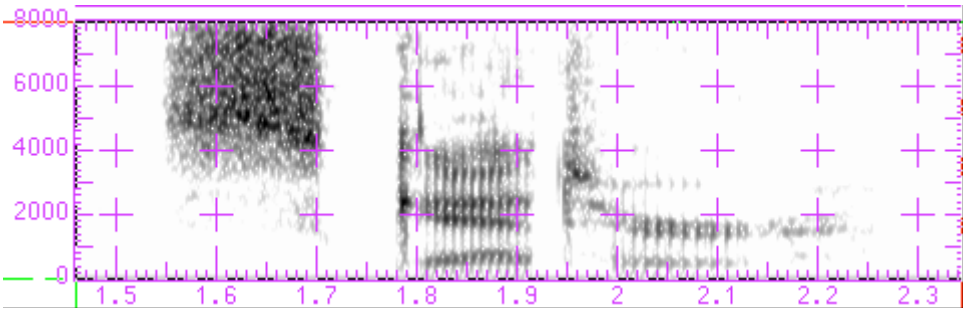


Εικόνα 9. Φασματογράφημα του αγγλικού φωνήματος "ss"

Ο στόχος αυτής της επεξεργασίας είναι να παραχθούν τμήματα με φασματικά χαρακτηριστικά που μπορούν να συνθέσουν φωνήματα. Το πρόβλημα με το ηχητικό σήμα είναι ότι τα στατιστικά χαρακτηριστικά των ιδιοτήτων του δεν παραμένουν σταθερά στο χρόνο. Αυτό σημαίνει ότι ακόμα και ο ίδιος ομιλητής, που λέει ακριβώς την ίδια έκφραση, δεν μπορεί να παράξει ακριβώς το ίδιο σήμα. Άρα, θέλουμε να εξάγουμε φασματικά χαρακτηριστικά από ένα μικρό τμήμα της ομιλίας που μπορεί να χαρακτηρίσει ένα φώνημα, (Jurafsky and Martin, 2009) και για το οποίο κάνουμε την παραδοχή ότι τα φασματικά του χαρακτηριστικά για αυτό το μικρό χρονικό διάστημα παραμένουν σταθερά. Για αυτό το λόγο, εφαρμόζεται πάνω στο σήμα ομιλίας μία συνάρτηση παραθύρου (window process), από την οποία το σήμα τεμαχίζεται παράγοντας πλαίσια ήχου (speech frames). Τα χαρακτηριστικά μίας συνάρτησης παραθύρου είναι το πλάτος του παραθύρου (σε milliseconds), η μετατόπιση για την εκκίνηση δημιουργίας νέου παραθύρου (offset) και το σχήμα του παραθύρου. Το πλάτος του παραθύρου εκφράζει το μέγεθος των πλαισίων ήχου. Η εικόνα 10 παρουσιάζει αυτή την διαδικασία.



Εικόνα 10. Εφαρμογή της συνάρτησης παραθύρου, με μέγεθος παραθύρου 25ms και την μετατόπιση του παραθύρου κάθε 10ms



Εικόνα 11. Φασματογράφημα της αγγλικής λέξης "sad"

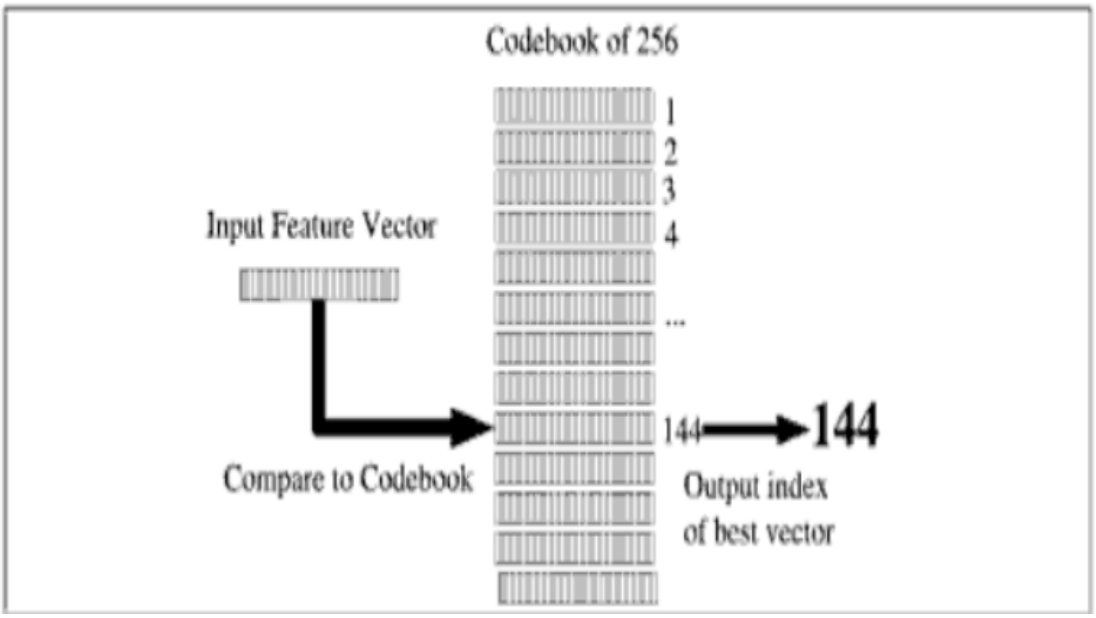
Αναφέρθηκε παραπάνω ότι η ανθρώπινη ομιλία αναλύεται καλύτερα εξετάζοντας την ένταση (δηλαδή το πλάτος) της συχνότητας σε σχέση με τον χρόνο. Ένα παράδειγμα είναι το φασματογράφημα της λέξης "sad" στην εικόνα 11. Η λέξη ξεκινά σε χρόνο 1.55 με υψηλές συχνότητες, χωρίς κάποιες σημαντικές χαμηλές συχνότητες, χωρίς αντήχηση (το φαινόμενο κατά το οποίο ο ήχος προσκρούει σε εμπόδιο κοντινό στην πηγή που τον παράγει, ανακλάται και επιστρέφει δυνατότερος, χωρίς όμως να ακούγονται καθαρά οι λεπτομέρειές του)

και στο χρόνο 1.7 υπάρχει μία περίοδος ησυχίας. Αυτά τα χαρακτηριστικά υποδηλώνουν ένα φώνημα /s/ ή /sh/. Μεταξύ 1.8 και 1.9 υπάρχουν έντονες χαμηλές συχνότητες, με εμφανείς «ράβδους» αντήχησης, χαρακτηριστικά ενός φωνήεντος. Από τον χρόνο 1.9 και μετά, οι χαμηλές συχνότητες και αντηχήσεις δείχνουν ένα φώνημα που αντιστοιχεί σε τελικό σύμφωνο, όπως τα /t/, /d/, /p/.

Το σύστημα αναγνώρισης ομιλίας διαθέτει μια βάση δεδομένων με φασματογραφήματα συχνοτήτων και κωδικούς αριθμούς (feature numbers) που αντιστοιχούν σε αυτά, που απεικονίζουν διαφορετικούς ήχους της ανθρώπινης ομιλίας. Αυτή η βάση δεδομένων ονομάζεται *codebook*. Το μέγεθος του κυμαίνεται από 256 έως 1024. Η δημιουργία του codebook γίνεται μέσω της εκπαίδευσης. Η παράμετρος που διαχωρίζει τους ήχους που βρίσκονται στο codebook είναι το πλάτος της συχνότητας. Το κάθε πλαίσιο ήχου αναγνωρίζεται συγκρίνοντας το με την εγγραφή του codebook που είναι πιο κοντά σε αυτό και του αποδίδεται ο κωδικός αριθμός της εγγραφής. Η σύγκριση αυτή γίνεται με την χρήση του πυθαγορείου θεωρήματος. Το πυθαγόρειο θεώρημα για δύο σημεία a=(x1, y1) και b=(x2,y2) είναι:

$$d(a,b) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad \textbf{(5.7)}$$

Όπου οι συντεταγμένες *x, y* των σημείων *a* και *b* εκφράζουν συχνότητα και πλάτος αντίστοιχα.



Εικόνα 12. Ταυτοποίηση ενός πλαισίου ήχου με μία εγγραφή του codebook

5.5. Ακουστικό μοντέλο

5.5.1. Αναγνώριση φωνημάτων

Το ακουστικό μοντέλο περιέχει για κάθε φώνημα της γλώσσας ένα Hidden Markov μοντέλο. Κάθε ένα από αυτά δέχεται ως είσοδο τα πλαίσια ήχου και επιστρέφει την πιθανότητα να εκπροσωπεί η ακολουθία των πλαισίων ήχου που εισήχθησαν το φώνημα που μοντελοποιεί το συγκεκριμένο HMM. Θεωρητικά, θα μπορούσε να συγκριθεί κάθε πλαίσιο ήχου με τα φάσματα συχνοτήτων που είναι αποθηκευμένα στο codebook. Εάν ένα πλαίσιο ήταν ίδιο με κάποιο φασματογράφημα με κωδικό αριθμό #52, θα μπορούσε να σημαίνει ότι ο χρήστης παρήγαγε το φώνημα /x/. Ο κωδικός #53 θα μπορούσε να είναι ένα φώνημα /f/, κ.τ.λ. Εάν αυτό ίσχυε, θα ήταν εύκολο να υπολογιστεί ποια φωνήματα είπε ο χρήστης. Δυστυχώς, αυτό δεν ισχύει λόγω διαφόρων λόγων:

- Κάθε φορά που ο χρήστης προφέρει μια λέξη αυτή ακούγεται διαφορετικά. Ποτέ δεν παράγεται ακριβώς ο ίδιος ήχος για το ίδιο φώνημα. Για αυτό το λόγο ποτέ κάποιο φασματογράφημα του συστήματος αναγνώρισης ομιλίας δεν είναι ακριβώς ίδιο με κάποιο ήχο που έχει ακουστεί.

- Ο θόρυβος από το μικρόφωνο και από το περιβάλλον στο οποίο βρίσκεται ο χρήστης κάποιες φορές αλλοιώνει τον ήχο και προκαλεί λανθασμένες εκτιμήσεις από το σύστημα αναγνώρισης ομιλίας.
- Η εκφορά κάθε φωνήματος εξαρτάται από το σημείο που βρίσκεται μέσα σε μία λέξη και από τα φωνήματα που βρίσκονται γύρω του. Π.χ. στη λέξη «σίγουρος» το αρχικό «σ» ακούγεται διαφορετικά από ότι το τελικό.

Για αυτούς τους λόγους δεν γίνεται το σύστημα αναγνώρισης ομιλίας να είναι σίγουρο σχετικά με το ποιο φώνημα έχει ειπωθεί. Αυτά τα προβλήματα της αναγνώρισης φωνημάτων λύνονται με την χρήση στατιστικών προτύπων. Το μαθηματικό εργαλείο για την εφαρμογή στατιστικών προτύπων είναι τα Hidden Markov Models. Κάθε εγγραφή του codebook αντιστοιχίζεται σε περισσότερα από ένα φωνήματα με διαφορετικές πιθανότητες για κάθε ένα. Για να είναι σε θέση το Hidden Markov μοντέλο να επιστρέφει αυτή την πιθανότητα στην εκμάθηση του χρησιμοποιούνται πολλά και διαφορετικά δείγματα του φωνήματος. Η λειτουργία αναγνώρισης φωνημάτων με χρήση Hidden Markov μοντέλων είναι η εξής:

Για να αναγνωρίσει ένα σύστημα αναγνώρισης ομιλίας πως ηχεί ένα φώνημα, εκπαιδεύεται αναλύοντας πολλές εκφορές του. Αναλύει κάθε πλαίσιο ήχου αυτών των εκφορών και παράγει έναν κωδικό αριθμό. Επειδή ένα φώνημα διαρκεί για έναν σχετικά μεγάλο χρόνο, 50 έως 100 πλαισίων ήχου είναι πιθανό μέσα σε αυτό το διάστημα να έχουν ακουστεί και άλλοι ήχοι, που δημιουργούν προβλήματα στην αναγνώριση (θόρυβος). Το σύστημα μαθαίνει στατιστικά πόσο πιθανό είναι για κάποιον ήχο να εμφανιστεί μέσα σε ένα φώνημα. Έτσι, μπορεί να αναγνωρίσει αν πρόκειται για θόρυβο ή όχι. Για παράδειγμα, για το φώνημα /x/, υπάρχει μια πιθανότητα 80% του ήχου με κωδικό αριθμό #52 να εμφανιστεί σε οποιοδήποτε πλαίσιο ήχου που το αποτελεί, πιθανότητα 30% του ήχου με κωδικό αριθμό #189, και πιθανότητα 15% του ήχου με κωδικό αριθμό #53. Για κάθε πλαίσιο ήχου ενός φωνήματος /f/ μπορεί να υπάρξει μια πιθανότητα 10% του ήχου με κωδικό αριθμό #52 να εμφανιστεί, 10% του ήχου με κωδικό αριθμό #189 και 80% του ήχου με κωδικό αριθμό # 53. Ας υποθέσουμε ότι τα 6 πλαίσια ήχου που αναγνωρίσθηκαν έχουν τους ακόλουθους κωδικούς αριθμούς:

52, 52 ,189 , 53 , 52 , 52

Το σύστημα αναγνώρισης υπολογίζει την πιθανότητα του ήχου που ακούστηκε να είναι ένα /x/ και την πιθανότητα να είναι οποιοδήποτε άλλο φώνημα, όπως το /f/. (How speech recognition works, http://project.uet.itgo.com/speech.htm)

Η πιθανότητα του /x/ είναι:

80% * 80% * 30% * 15% * 80% * 80% = 1,84%

Η πιθανότητα του ήχου να είναι /f/ είναι:

10% * 10% * 10% * 80% * 10% * 10% = 0,0008%

Από αυτά τα αποτελέσματα φαίνεται ότι το /x/ είναι πιθανότερο να είναι το φώνημα που ειπώθηκε. (Στην πραγματικότητα, το αποτέλεσμα των πράξεων αποτελεί μη κανονικοποιημένη πιθανότητα διότι είναι μεγαλύτερο της μονάδας.) Γενικά, η διαδικασία αναγνώρισης έχει ως εξής:

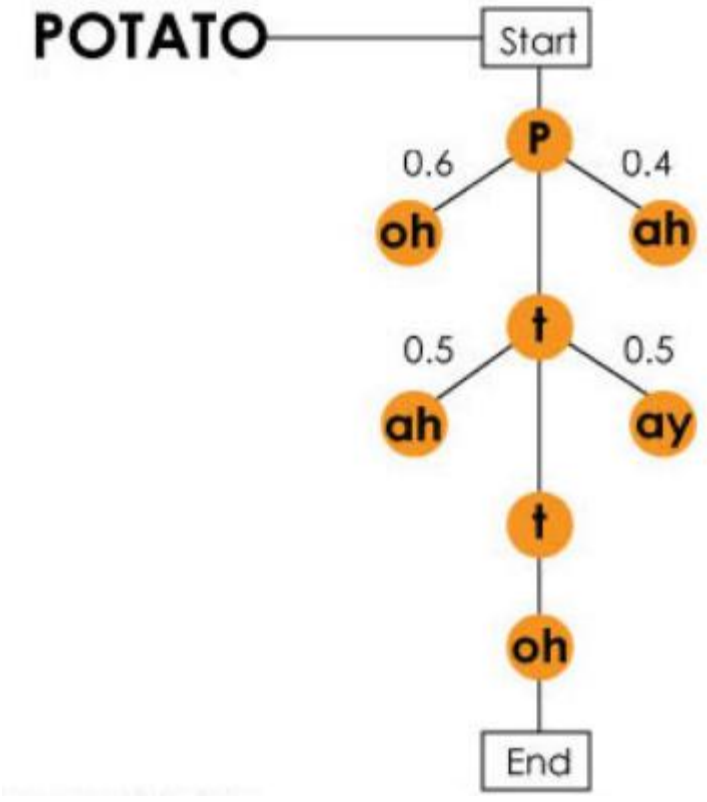
Όταν το σύστημα αναγνώρισης αρχίζει να εκτελείται, δημιουργεί μία υποθετική κατάσταση. Υποθέτει ότι ο χρήστης δεν έχει αρχίσει να ομιλεί, άρα αυτό που «ακούει» είναι το φώνημα «σιωπής». Κάθε 1/100 του δευτερολέπτου, όσο διαρκεί δηλαδή ένα πλαίσιο ήχου, και ενώ το σύστημα έχει αρχίσει να «ακούει» την ομιλία του χρήστη, προσθέτει μία καινούργια υποθετική κατάσταση για κάθε φώνημα, δηλαδή στην αγγλική γλώσσα που περιέχει 50 φωνήματα, δημιουργεί 50 υποθετικές καταστάσεις και σε κάθε μία από αυτές αποδίδει μία πιθανότητα, σύμφωνα με το ποιο πλαίσιο ήχου έχει αναγνωριστεί. Μετά το πρώτο 1/100 του δευτερολέπτου το σύστημα έχει δημιουργήσει 51 υποθετικές καταστάσεις. Στο 2/100 του δευτερολέπτου από την έναρξη, ένα νέο πλαίσιο ήχου εισάγεται. Οι πιθανότητες των 51 υποθετικών καταστάσεων επαναϋπολογίζονται μαζί με το νέο κωδικό αριθμό. Τότε, σε κάθε υποθετική κατάσταση που υπάρχει από το πρώτο 1/100, προστίθενται άλλες 50 υποθετικές καταστάσεις, σαν παιδιά ενός κόμβου ενός δέντρου. Έτσι, δημιουργούνται 51*50=2550 καινούργιες υποθετικές καταστάσεις. Τώρα, η πιθανότητα κάθε κατάστασης είναι η πιθανότητα που είχε στο πρώτο 1/100 επί την πιθανότητα στο δεύτερο 1/100 του δευτερολέπτου. Η ίδια διαδικασία ακολουθείται για κάθε 1/100 του δευτερολέπτου. Το κλαδί του δέντρου με την μεγαλύτερη πιθανότητα, υπολογισμένη σύμφωνα με την διαδικασία που περιγράφηκε παραπάνω, είναι το τελικό αποτέλεσμα.

Φυσικά, λόγω του μεγάλου αριθμού των υποθετικών καταστάσεων γίνονται βελτιστοποιήσεις στην διαδικασία αυτή. Αν μία πιθανότητα είναι πολύ μικρή σε σχέση με την μεγαλύτερη πιθανότητα, τότε αυτή αφαιρείται. Αυτή η διαδικασία

ονομάζεται περικοπή (pruning). Μία άλλη τεχνική βελτιστοποίησης είναι να μην δημιουργούνται καινούργιες καταστάσεις κάθε 1/100 του δευτερολέπτου. Για να το κάνει όμως αυτό ένα σύστημα αναγνώρισης ομιλίας, θα πρέπει να ξέρει ποια φωνήματα έπονται μετά από κάθε φώνημα.

5.5.2. Αναγνώριση λέξεων

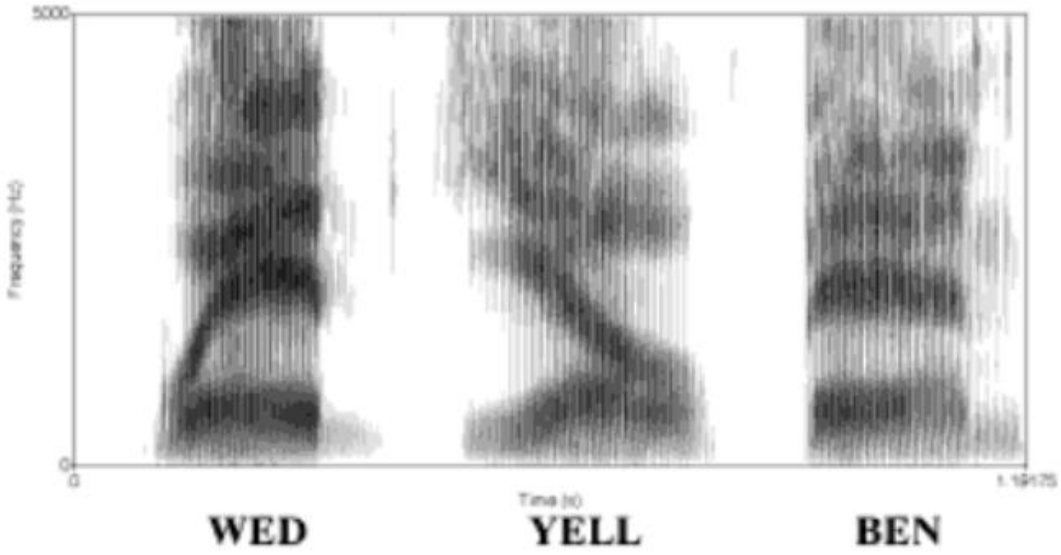
Τα ηχητικά τμήματα που αποτελούν τις λέξεις είναι τα φωνήματα. Είδαμε παραπάνω τον τρόπο αναγνώρισης τους. Αυτά τα τμήματα ενώνονται στη συνέχεια σχηματίζοντας μία ακολουθία που δημιουργεί τις λέξεις η οποία μοντελοποιείται από ένα Hidden Markov μοντέλο. Η ακολουθία των φωνημάτων που έχει αναγνωρισθεί συγκρίνεται με τις εγγραφές του λεξικού (βλέπε κεφ.4) με σκοπό να βρεθεί μία ίδια εγγραφή. Μία τέτοια ακολουθία παρουσιάζεται στην εικόνα 13.



Εικόνα 13. Hmm για την αγγλική λέξη "potato"

Όμως, υπάρχει ένα πρόβλημα στην χρήση φωνημάτων για την σύνθεση λέξεων. Το πρόβλημα είναι ότι ο ήχος των φωνημάτων αλλάζει αρκετά ανάλογα με ποιο φώνημα προηγείται ή έπεται ενός φωνήματος (συνάρθρωση-coarticulation).

Π.χ. στις λέξεις «άοπλος» και «αεροπλάνο» το φώνημα /a/ ακούγεται διαφορετικά. Το /a/ δεν εκφέρεται ακριβώς ως /a/ αλλά ενσωματώνει και το /o/ ή το /e/ στον τελικό ήχο. Στην εικόνα 14 παρουσιάζονται τα φασματογραφήματα τριών λέξεων που περιέχουν το αγγλικό φωνήεν [e], και πως τα γειτονικά φωνήματα μπορούν να επηρεάσουν το άκουσμά του. Παρατηρούμε ότι το μέσο περίπου κάθε φασματογραφήματος που απεικονίζει το [e], δεν είναι ίδιο με τα υπόλοιπα.



Εικόνα 14. Φασματογραφήματα των αγγλικών λέξεων "Wed", "Yell" και "Ben". Το φώνημα /e/ στο μέσον κάθε λέξης αποτυπώνεται διαφορετικά

Οι μηχανές αναγνώρισης ομιλίας λύνουν το πρόβλημα με τη λύση των τριφωνημάτων (triphones), όπου κάθε φώνημα έχει ένα διαφορετικό Hidden Markov Model για κάθε μοναδικό ζεύγος από δεξιούς και αριστερούς γείτονες. Δηλαδή, για κάθε φώνημα λαμβάνονται υπόψη και τα φωνήματα που το περιβάλλουν. Η αναπαράσταση των τριφωνημάτων γίνεται μέσα σε αγκύλες ως εξής: [xyz], το οποίο σημαίνει ότι το φώνημα /y/ έπεται του /x/, και ακολουθείται από το /z/. Άλλη αναπαράσταση είναι η [x-y+z], που έχει ακριβώς το ίδιο νόημα. Με τα τριφωνήματα αναγνωρίζονται επιτυχώς διάφορες παραλλαγές στον ήχο, και είναι σημαντικό μέλος των σύγχρονων συστημάτων αναγνώρισης ομιλίας. Για μία γλώσσα με περίπου 30 φωνήματα όπως η ελληνική, θα πρέπει να δημιουργηθούν 30*30*30= 27.000 τριφωνήματα. Αυτό το μέγεθος δημιουργεί προβλήματα στην εκπαίδευση, αφού όσο πιο πολύπλοκα είναι τα μοντέλα που χρησιμοποιούνται, τόσο λιγότερο δυνατό είναι να βρεθούν αρκετές εμφανίσεις για κάθε τύπο φωνήματος στα δεδομένα εκπαίδευσης ώστε να γίνει αποτελεσματική εκπαίδευση.

Στην πράξη, δεν χρειάζονται τόσα πολλά τριφωνήματα, αφού στα ελληνικά από την ακολουθία φωνημάτων [xzg] δεν προκύπτει κάποια λέξη. Να αναφερθεί ότι υπάρχουν και διφωνήματα (diphones), αλλά τα πιο δημοφιλή μοντέλα για την αναπαράσταση της επιρροής άλλων φωνημάτων είναι τα τριφωνήματα.

Για παράδειγμα, η λέξη «δέκα», μπορεί να αναπαρασταθεί από την ακολουθία [sil δ e k a sil]. Χρησιμοποιώντας τριφωνήματα, η ακολουθία θα ήταν η εξής:

$$\text{sil}_{\text{sil}}\delta_{\text{e}}\ _{\delta}\text{e}_{\text{k}}\ _{\text{e}}\text{k}_{\text{a}}\ _{\text{k}}\text{a}_{\text{sil}}\ \text{sil}.$$

Κάθε ένα τριφώνημα μοντελοποιείται από ένα Hidden Markov μοντέλο. Όπως και στην αναγνώριση φωνημάτων, με βάση την ακολουθία των πλαισίων ήχου, το κάθε HMM επιστρέφει την πιθανότητα να είναι αυτό που μοντελοποιεί το τριφώνημα που ειπώθηκε.

5.6. Γλωσσικό μοντέλο

Σε μία γλώσσα δεν επιτρέπεται οποιαδήποτε σειρά λέξεων (συντακτικά, γραμματικά και σημασιολογικά) π.χ. «δεν θα πάω» είναι δεκτή πρόταση συντακτικά ενώ η πρόταση «θα πάω δεν» δεν είναι. Το γλωσσικό μοντέλο (language model) είναι ένας μηχανισμός για τον υπολογισμό της πιθανότητας να υπάρχει μία λέξη W σε μία εκφώνηση όπου προηγούνται N λέξεις. Στο προηγούμενο παράδειγμα, ένα σωστά εκπαιδευμένο σύστημα αναγνώρισης ομιλίας θα επιστρέψει μεγαλύτερη πιθανότητα στην πρόταση «Η αρκούδα κοιμήθηκε» από ότι στην πρόταση «Η θάλασσα κοιμήθηκε». (Γλωσσικό μοντέλο, www.logografos.gr). Με την χρήση του γλωσσικού μοντέλου βελτιστοποιείται η απόδοση του συστήματος και αυξάνεται η ακρίβεια του.

Ο ρόλος του γλωσσικού μοντέλου είναι να αποδίδει κάποια πιθανότητα σε κάθε μία ακολουθία λέξεων. Μαθηματικά, πρέπει να υπολογίσει την πιθανότητα ***P(W),*** όπου το *W* εκφράζει μία ακολουθία λέξεων, και η πιθανότητα *P(W),* εκφράζει την πιθανότητα ο χρήστης να πει την ακολουθία W. Η πιθανότητα P(W), χρησιμοποιώντας τον κανόνα της δεσμευμένης πιθανότητας, μπορεί να οριστεί ως:

$$\mathrm{P(W)} = \mathrm{P(w1)P(w2|\ w1)P(w3|\ w2, w1)\ldots P(wn\ |\ w1, w2, \ldots., wn\ -1).} \quad \textbf{(5.8)}$$

Το γλωσσικό μοντέλο υπολογίζει στατιστικά αυτή τη πιθανότητα, αναλύοντας κείμενα κατά την εκπαίδευση του. Το πλεονέκτημα της χρήσης του κανόνα 5.8 είναι ότι η πιθανότητα W_i, εξαρτάται μόνο από τις προηγούμενες λέξεις που έχουν ήδη ειπωθεί, και όχι από κάποια που θα ειπωθεί στο μέλλον. Αυτό που το γλωσσικό μοντέλο πρέπει να κάνει είναι να υπολογίσει τις δεσμευμένες πιθανότητες. Όμως, η πιθανότητα $P(w_i|w_1, w_2,\ldots,w_{i-1})$, είναι αδύνατο να εκτιμηθεί, ακόμα και για μέτριες τιμές του *i*. Αυτό συμβαίνει, γιατί δεν μπορούν να βρεθούν κείμενα που να περιέχουν κάθε ακολουθία, και ειδικά μεγάλες ακολουθίες.

Στην πράξη, ο αριθμός του *n* στον κανόνα 5.8 είναι συνήθως 3. Για αυτή την τιμή του n, ο κανόνας ισοδυναμεί με τον ακόλουθο:

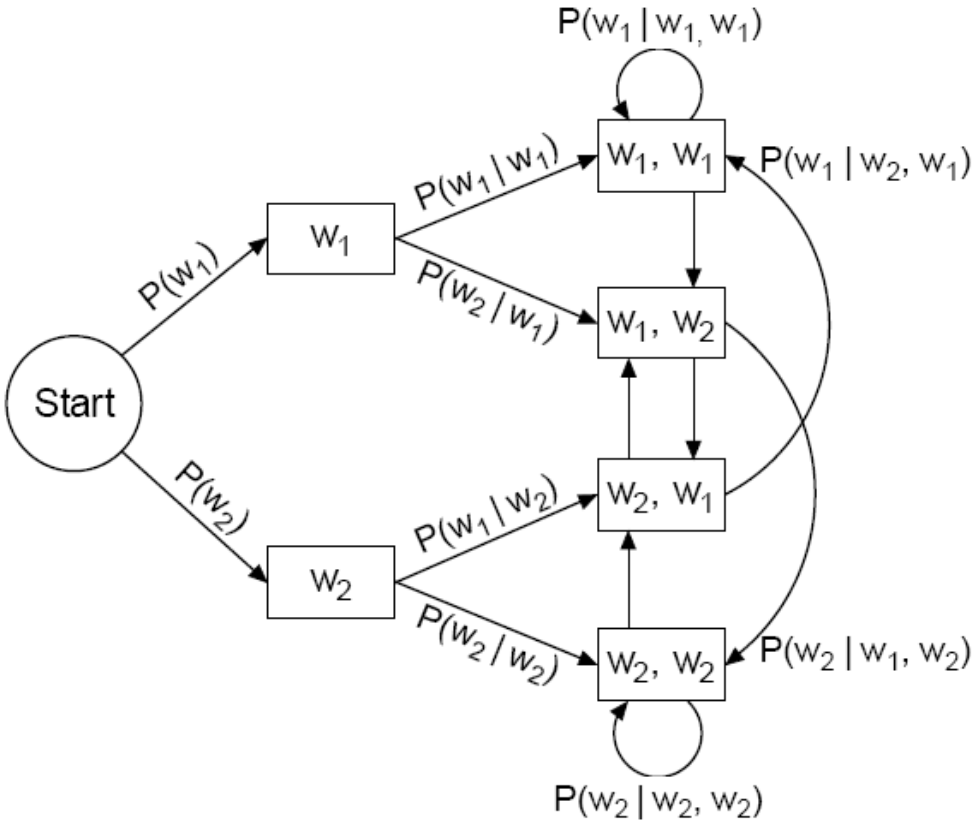
$$P(W) = \prod_{i=1}^{n} P(w_i \mid w_{i-2}, w_{i-1}) \qquad \textbf{(5.9)}$$

Οι ακολουθίες τριών λέξεων λέγονται τριγράμματα (trigrams). Τα περισσότερα δεν απαντώνται σε κάποιο κείμενο εκπαίδευσης, όμως το να τα προφέρει ο χρήστης δεν είναι απίθανο.

Εκτός από την χρήση στατιστικών πληροφοριών για τον υπολογισμό της πιθανότητας, μοντελοποιείται και το συντακτικό και η γραμματική μίας γλώσσας. Έτσι, σε κάθε λέξη σε μία πρόταση, εκχωρούνται ετικέτες που προσδιορίζουν τι μέρος του λόγου είναι κάθε λέξη. Για παράδειγμα, για την έκφραση «μέσα στο...», είναι προφανές ότι δεν μπορεί να ακολουθεί ρήμα, ενώ είναι πολύ πιθανό να ακολουθεί ουσιαστικό. Όμως, υπάρχουν κάποια προβλήματα σε αυτή τη διαδικασία. Το πρώτο είναι ότι τα μοντέλα αυτά αποτελούνται από τρεις λέξεις, ενώ για να γίνει αποτελεσματικότερος ο προσδιορισμός του τι μέρος του λόγου είναι κάθε λέξη, απαιτούνται μεγαλύτερες προτάσεις. Το δεύτερο πρόβλημα είναι ότι στον προφορικό λόγο γίνονται από τους ανθρώπους πολλά γραμματικά και συντακτικά λάθη.

Τα τριγράμματα, και γενικά τα *n* μεγέθους γλωσσικά μοντέλα, μπορούν να αναπαρασταθούν από Hidden Markov Models. Χρειάζεται μία αρχική κατάσταση, και στη συνέχεια πιθανότητες μετάβασης καθορίζουν τις μεταβάσεις από την αρχική κατάσταση στις επόμενες. Για n=3, υπάρχει μία μετάβαση από την

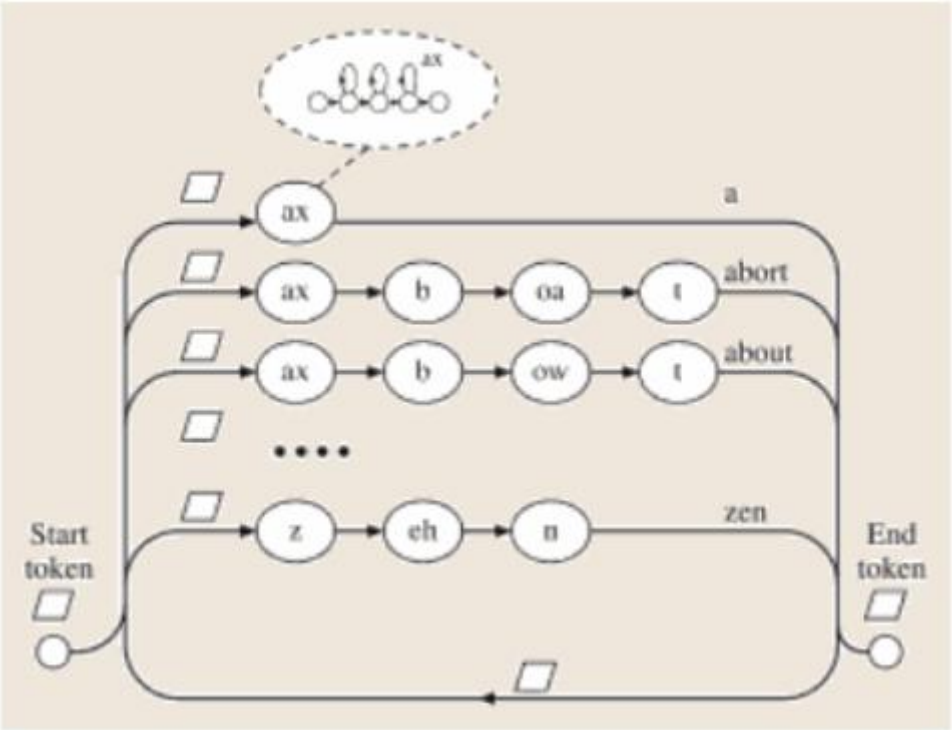
κατάσταση (w_i, w_j) σε μία κατάσταση (w_k, w_l), αν και μόνο αν $w_j = w_k$. Σε αυτή την περίπτωση, η πιθανότητα μετάβασης θα είναι $P(w_l | w_i, w_j)$.



Εικόνα 15. Ένα γλωσσικό μοντέλο τριών καταστάσεων για ένα λεξικό δύο λέξεων. Κάποιες από τις μεταβάσεις δεν αναγράφονται για λόγους ευκρίνειας.

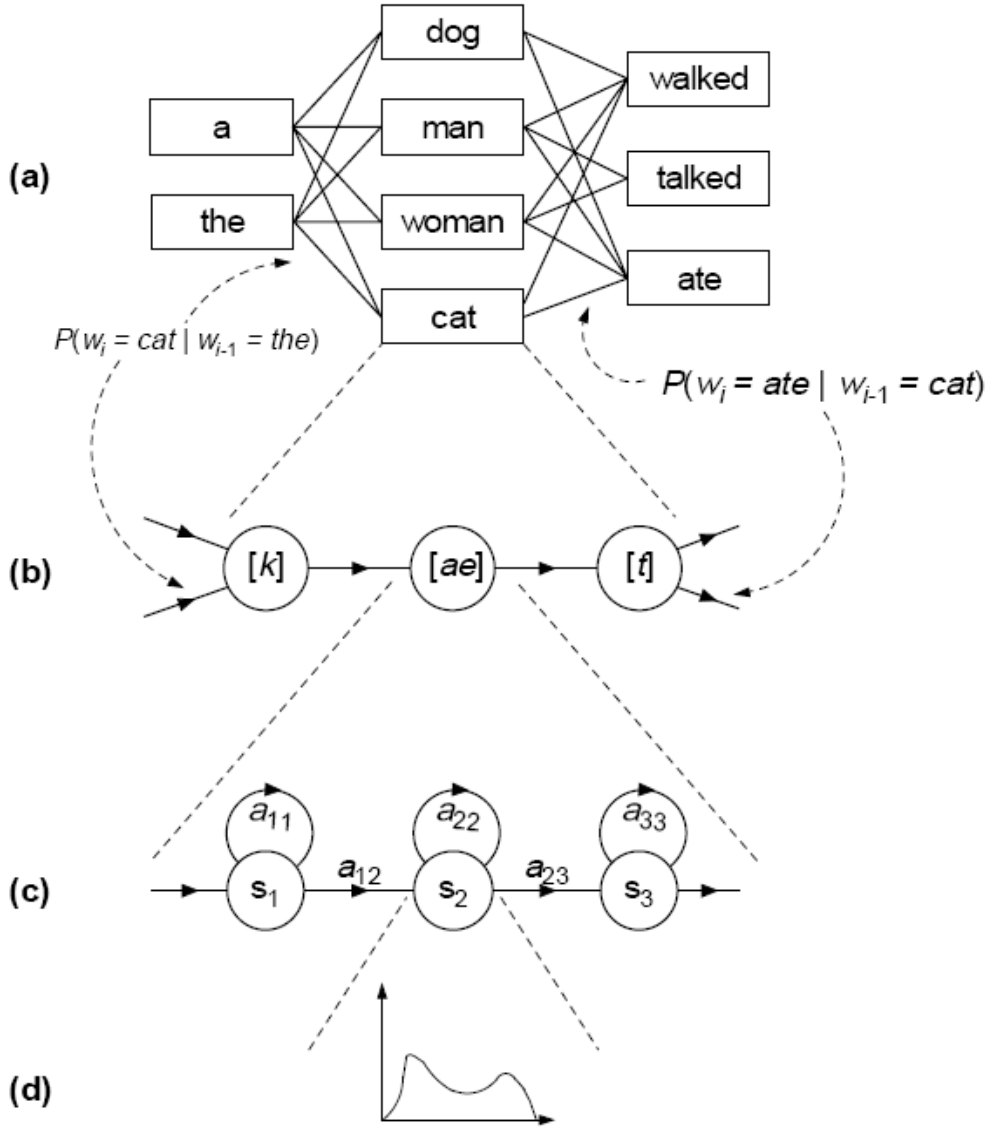
5.7. Εύρεση του πιθανότερου αποτελέσματος

Η εισαγωγή Hidden Markov μοντέλων μέσα σε άλλα, δημιουργεί νέα Hidden Markov Models. Έτσι δημιουργείται ένα HMM που ενσωματώνει το ακουστικό και το γλωσσικό μοντέλο. Αυτό το HMM ονομάζεται δίκτυο αναγνώρισης (recognition network).



Εικόνα 16. Ένα δίκτυο αναγνώρισης

Το δίκτυο αναγνώρισης συντίθεται ιεραρχικά, (Aarnio,1998) από την αναγνώριση των φωνημάτων, στην συνέχεια την αναγνώριση των λέξεων, και τέλος των προτάσεων που έχουν ειπωθεί. Η εικόνα 17 παρουσιάζει αυτή την διαδικασία για μονοφωνικά ακουστικά μοντέλα και ένα απλό λεξικό.



Εικόνα 17. Δημιουργία ενός δικτύου αναγνώρισης με ιεραρχική συνένωση Hidden Markov μοντέλων. a) Ένα μικρό λεξικό με τις πιθανότητες μετάβασης μεταξύ των λέξεων να ορίζονται από ένα Hidden Markov μοντέλο. b) Η προφορά μίας λέξης αναπαρίσταται από ένα Hidden Markov μοντέλο. c) Κάθε φώνημα μοντελοποιείται από ένα Hidden Markov μοντέλο. d) Κάθε φώνημα αναγνωρίζεται τεμαχίζοντας το σε πλαίσια ήχου και αναγνωρίζοντας κάθε ένα από αυτά.

Τα αποτελέσματα της αναγνώρισης μέσα από αυτό το δίκτυο ανακτώνται χρησιμοποιώντας τον αλγόριθμο *Viterbi*. Ο αλγόριθμος Viterbi βρίσκει το μονοπάτι μέσα σε ένα Hidden Markov μοντέλο που επιστρέφει την μεγαλύτερη πιθανότητα. Η συνθήκη τερματισμού επιτυγχάνεται όταν ο αλγόριθμος φτάσει στον χρόνο *T* όπου η τελευταία παρατήρηση *O* επεξεργάζεται.

$$\delta_j(t) = \max_{i \in S} \delta_i(t-1) a_{ij} b_j o(t) \qquad \text{για κάθε } j \in S \text{ και } t \geq 1 \qquad \textbf{(5.10)}$$

$$\delta(T) = \max_{j \in S} \delta_j(T) \qquad \text{όπου } t=T \qquad \textbf{(5.11)}$$

Η τιμή $\delta_j(t)$ εκφράζει την πιθανότητα κάθε μονοπατιού. Η πιθανότητα $\delta(T)$ είναι η πιθανότητα του καλύτερου μονοπατιού μέσα σε ένα Hidden Markov μοντέλο.

Μέσα σε ένα δίκτυο αναγνώρισης, σε οποιοδήποτε χρόνο *t,* μία υπόθεση αποτελείται από ένα μονοπάτι μέσα στο δίκτυο, που εκφράζει μία ακολουθία από καταστάσεις, αρχίζει από την αρχική κατάσταση και τελειώνει στην τελική κατάσταση *j*, και έχει μία πιθανότητα $\delta_j(t)$. Αυτή η διαδικασία μπορεί να γίνει περισσότερο κατανοητή από την εισαγωγή ενός *token*, που αποτελείται από τις τιμές $\delta_j(t)$ και ένα δείκτη *link* σε μία εγγραφή που περιέχει πληροφορίες για το δίκτυο πριν το χρόνο *t*. Η αναγνώριση προκύπτει καθώς αυτά τα token, περνούν από κάθε κατάσταση μέσα από το δίκτυο και καταλήγουν στην τελική κατάσταση. Το token είναι της μορφής ($\delta_j(t)$, *link*).

Κάθε φορά που το token περνά από το τέλος μίας λέξης στην αρχή της επόμενης, η πιθανότητα $\delta_j(t)$ επαναϋπολογίζεται από το γλωσσικό μοντέλο. Την ίδια στιγμή, αποθηκεύεται μία εγγραφή *R* που περιέχει ένα αντίγραφο του token, τη χρονική στιγμή που αυτό συνέβη, και την λέξη που προηγήθηκε. Ο δείκτης link του token αλλάζει και πλέον δείχνει την τελευταία εγγραφή *R* που έχει προκύψει. Καθώς κάθε token περνά μέσα από το δίκτυο, αποθηκεύει όλες τις αλλαγές στην εγγραφή *R*. Το token που συγκεντρώνει την μεγαλύτερη πιθανότητα στον χρόνο *T* που είναι ο τελικός χρόνος επεξεργάζεται ώστε να εξακριβωθεί ποια είναι η ακολουθία των λέξεων που πέρασε.

Το πρόβλημα που υπάρχει είναι ότι με αυτή την διαδικασία, εισάγεται πολύ μεγάλο υπολογιστικό κόστος. Ενδεικτικά, για ένα λεξικό που αποτελείται από 10.000 λέξεις και γλωσσικό μοντέλο μεγέθους 3, δημιουργούνται 10^8 διαφορετικές καταστάσεις, και 10^{12} μεταβάσεις. Εάν το μέσο μέγεθος μίας λέξης σε φωνήματα είναι 5, και κάθε φώνημα μοντελοποιείται από 4 Hidden Markov Models, ο αριθμός των διαφορετικών καταστάσεων σε ένα δίκτυο είναι: $4*5*10^{8=}2*10^9$. Συνεπώς, θα πρέπει να υπάρξει μία διαδικασία βελτιστοποίησης.

Μόνο τα token που έχουν μία σημαντική πιθανότητα να βρίσκονται στο σωστό μονοπάτι μέσα στο δίκτυο συνεχίζουν. Για να γίνει αυτό αποθηκεύεται η μεγαλύτερη πιθανότητα που έχει βρεθεί μέχρι εκείνη τη στιγμή, και όσων token η πιθανότητα είναι πολύ κάτω από αυτή σταματούν. Αυτή η τεχνική ονομάζεται

Beam search. Το μειονέκτημα της χρήσης της είναι ότι μπορεί να αφαιρεθεί ένα μονοπάτι που στη συνέχεια μπορούσε να αποκτήσει μεγαλύτερη πιθανότητα. Ως αποτέλεσμα της *Beam search* 90% του υπολογιστικού κόστους δαπανάται στα πρώτα δύο φωνήματα, καθώς στη συνέχεια προκύπτουν οι βελτιστοποιήσεις.

5.8. Εκπαίδευση

Ο στόχος της εκπαίδευσης είναι να αναγνωριστεί ομιλία που δεν έχει αναγνωριστεί στο παρελθόν, και η αναγνώριση να προκύψει με βάση τα δεδομένα που έχουν πάρει από την διαδικασία της εκπαίδευσης.

Κατά την εκπαίδευση, εκτιμούνται παράμετροι των Hidden Markov μοντέλων, όπως οι τιμές των πιθανοτήτων μετάβασης, ή η αρχική κατάσταση των μοντέλων. Υπάρχουν πολλά HMM μέσα σε ένα σύστημα αναγνώρισης ομιλίας, και κάθε ένα από αυτά είναι εκπαιδευμένο να αναπαριστά μία διαφορετική πηγή ομιλίας. Στη φάση της αναγνώρισης, κάθε μοντέλο συναγωνίζεται με τα υπόλοιπα για το ποιο εκφράζει καλύτερα τα δεδομένα ομιλίας που δέχθηκαν ως είσοδο.

Ας υποθέσουμε ότι έχουμε πολλά διαφορετικά HMM και μία ακολουθία από παρατηρήσεις *O* που παράγονται. Το θέμα είναι να βρεθεί ποιο μοντέλο *λ* αναπαριστά καλύτερα την ομιλία που δέχθηκε ως είσοδο, με βάση τις παρατηρήσεις που εξάγονται, δηλαδή, μαθηματικά αυτό που μεγιστοποιεί την σχέση:

$$\hat{\lambda}_{MAP} = \arg\max_{\lambda} P(\lambda \mid O) \qquad \textbf{(5.12)}$$

Η σχέση 5.12 με την εφαρμογή του κανόνα του Baye's, μετατρέπεται στην σχέση:

$$\hat{\lambda}_{MAP} = \arg\max_{\lambda} \frac{P(O \mid \lambda)P(\lambda)}{P(O)} = \arg\max_{\lambda} P(O \mid \lambda)P(\lambda) \qquad \textbf{(5.13)}$$

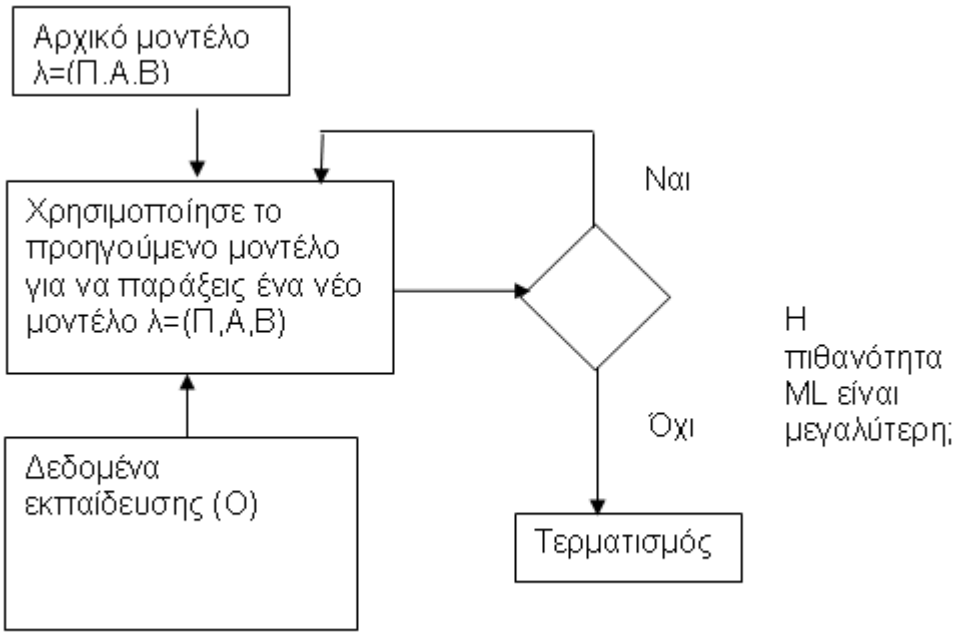
Αυτή η πιθανότητα ονομάζεται *maximum a posterior* (MAP). Ο παρανομαστής του κλάσματος P(O), αφαιρείται καθώς παραμένει σταθερός, και συνεπώς δεν επηρεάζει την σχέση. Επίσης, επειδή τιμές για την πιθανότητα P(λ) δεν μπορούν να βρεθούν, υποθέτουμε ότι ακολουθεί ενιαία κατανομή, που σημαίνει ότι όλες οι τιμές του *λ* έχουν ίδια πιθανότητα. Συνεπώς, η 5.13 μετατρέπεται στην:

$$\hat{\lambda}_{ML} = \arg\max_{\lambda} P(O \mid \lambda) \qquad \textbf{(5.14)}$$

Η πιθανότητα *λ* ονομάζεται *Maximum likelihood* (ML). Οπότε, με βάση αυτή τη σχέση το καλύτερο μοντέλο είναι αυτό που έχει την μεγαλύτερη πιθανότητα να δημιουργήσει την δοθείσα ακολουθία παρατηρήσεων.

Για την εκπαίδευση, ο σκοπός είναι να βελτιστοποιηθεί ένα μοντέλο ή ένα σετ από μοντέλα, με βάση κάποιο κριτήριο βελτιστοποίησης. Εάν αυτό το κριτήριο είναι να βρεθεί η μεγαλύτερη πιθανότητα που δίνει η 5.14 μπορεί να εφαρμοστεί ο αλγόριθμος *Baum-Welch*.

Ο αλγόριθμος *Baum-Welch* είναι μία επαναληπτική διαδικασία, που επαναπροσδιορίζει τις παραμέτρους των πιθανοτήτων σύνδεσης *A*, τις πιθανότητες που παράγονται *B*, και την αρχική κατάσταση Π έως ότου το μέγιστο της Maximum likelihood συνάρτησης βρεθεί. Η λειτουργία του αλγορίθμου Baum-Welch παρουσιάζεται στην εικόνα 18.



Εικόνα 18. Απεικόνιση της λειτουργίας του αλγόριθμου Baum-Welch

Σε κάθε βήμα, το μοντέλο θα πρέπει να δίνει μεγαλύτερη πιθανότητα, αλλιώς ο αλγόριθμος τερματίζεται. Το μειονέκτημα αυτού του αλγορίθμου είναι ότι μετά από κάποιο τοπικό μέγιστο, ο αλγόριθμος τερματίζεται, αν και στη συνέχεια θα μπορούσε να υπάρξει κάποια μεγαλύτερη πιθανότητα από αυτή που τερμάτισε τον αλγόριθμο.

Επίλογος

Σε αυτό το κεφάλαιο παρουσιάστηκε ο τρόπος λειτουργίας των συστημάτων αναγνώρισης ομιλίας. Επειδή το δομικό συστατικό της προφορικής ομιλίας είναι τα φωνήματα, είναι επόμενο ο τρόπος λειτουργίας των συστημάτων αναγνώρισης ομιλίας να έχει ως βασικό στοιχείο την αναγνώριση φωνημάτων. Τα φωνήματα είναι τα μικρότερα στοιχεία της προφορικής ομιλίας, που παρέχουν διαφοροποιητική λειτουργία. Είναι το ανάλογο των φθόγγων του γραπτού κειμένου. Τα συστήματα αναγνώρισης ομιλίας, προσπαθούν να αναγνωρίσουν πρώτα τα φωνήματα από μία ομιλία και μετά να τα συνθέσουν για να δημιουργηθούν λέξεις και προτάσεις.

Πυρήνας της λειτουργίας των συστημάτων αναγνώρισης ομιλίας είναι τα Hidden Markov Models (HMM). Τα HMM είναι μαθηματικά στατιστικά μοντέλα, που μοντελοποιούν τις μεταβάσεις μεταξύ ακολουθιών. Για να μπορέσει να γίνει η αναγνώριση, το σήμα πρέπει πρώτα να επεξεργαστεί. Το αναλογικό σήμα της ανθρώπινης ομιλίας μετατρέπεται σε ψηφιακό. Στη συνέχεια, το ψηφιακό σήμα κατακερματίζεται σε τμήματα ήχου, που ονομάζονται πλαίσια ήχου (speech frames), και έχουν διάρκεια 1/100 του δευτερολέπτου. Έπειτα, το σύστημα προσπαθεί να ταυτοποιήσει τα σήματα αυτά, συγκρίνοντας τα με μία βάση δεδομένων που περιέχει παρόμοιους ήχους. Αυτή η βάση δεδομένων ονομάζεται *codebook*.

Σε κάθε πλαίσιο ήχου, που ταυτοποιήται με την εγγραφή π.χ. 32 του codebook, του αποδίδεται ο κωδικός αριθμός 32 της εγγραφής (feature number). Τώρα είναι η σειρά του ακουστικού μοντέλου (acoustical model) να εφαρμοστεί κάνοντας τα ακόλουθα. Το σύστημα προσπαθεί να αναγνωρίσει φωνήματα, που όπως ειπώθηκε στο προηγούμενο κεφάλαιο, είναι οι δομικές μονάδες της προφορικής ομιλίας. Με την χρήση HMM, υπολογίζει με βάση την είσοδο από πλαίσια ήχου που δέχεται, ποιο φώνημα είναι πιθανότερο τα συγκεκριμένα πλαίσια ήχου να μοντελοποιούν. Στη συνέχεια, διαθέτοντας μία ακολουθία από φωνήματα, εφαρμόζει ένα ακόμα HMM, όπου γνωρίζοντας τις πιθανότητες μετάβασης μεταξύ των φωνημάτων, και έχοντας και ένα λεξικό που περιέχει όλες τις πιθανές λέξεις, αναγνωρίζει ποιες λέξεις έχουν ειπωθεί.

Τέλος, για να είναι σε θέση να αναγνωρίζει συνεχόμενο διάλογο, το σύστημα εφαρμόζει πάνω στην ακολουθία των λέξεων ένα ακόμα HMM μοντέλο, το γλωσσικό (language model). Το γλωσσικό μοντέλο είναι ένας μηχανισμός για τον υπολογισμό της πιθανότητας να υπάρχει μία λέξη W σε μία εκφώνηση όπου προηγούνται N λέξεις, και για το πετύχει αυτό μοντελοποιεί το συντακτικό και την γραμματική μίας γλώσσας. Για παράδειγμα, ένα γλωσσικό μοντέλο θα δώσει μικρότερη πιθανότητα να έχει ειπωθεί η φράση «θα πάω δεν» από ότι στη φράση «δεν θα πάω». Έτσι, πλέον συντίθενται οι πιο πιθανές ακολουθίες λέξεων *n* μεγέθους, όπου το n συνήθως είναι 3.

Όλα αυτά τα HMM που έχουν δημιουργηθεί, δημιουργούν ένα δίκτυο από HMM. Με την χρήση του αλγορίθμου *viterbi*, βρίσκεται το καλύτερο μονοπάτι, που αποτελεί και το αποτέλεσμα της αναγνώρισης.

Τέλος, τα HMM με την διαδικασία της εκπαίδευσης μπορούν να προσαρμοστούν, και να αλλάξουν τα χαρακτηριστικά τους με αποτέλεσμα να παράγουν διαφορετικές εξόδους. Ο πιο γνωστός αλγόριθμος για την εκπαίδευση είναι ο αλγόριθμος *Baum-Welch.* Με αυτό το τρόπο, ένα σύστημα αναγνώρισης ομιλίας μπορεί να προσαρμοστεί πάνω σε κάποιο ομιλητή.

Κεφάλαιο 6. Επεξήγηση του κώδικα της εφαρμογής

Εισαγωγή

Σκοπός της εφαρμογής είναι η αναγνώριση φωνητικών εντολών, και η εκτέλεση των λειτουργιών που είναι συνδεδεμένες με την κάθε φωνητική εντολή. Για αυτό το λόγο, μπορεί να χρησιμοποιηθεί από χρήστες με προβλήματα όρασης. Η εφαρμογή είναι γραμμένη σε γλώσσα προγραμματισμού JavaScript, ώστε να μπορεί να εκτελεστεί σε έναν browser, και για να αναγνωρίζει την ομιλία χρησιμοποιεί λειτουργίες που παρέχονται από το πακέτο SAPI (Speech Application Programming Interface) της εταιρείας Microsoft. Οι λειτουργίες που υποστηρίζει είναι απομόνωση των συνδέσμων, των παραγράφων, των περιγραφών των εικόνων, των επικεφαλίδων που υπάρχουν σε μία HTML σελίδα , και η εμφάνιση τους σε μία λίστα. Επίσης, γράφει σε μία περιοχή κειμένου (TextArea) ότι του υπαγορεύει ο χρήστης. Ο κώδικας που ακολουθεί δημιουργεί τα αντικείμενα που χρειάζεται.

Σημείωση: Για να «τρέξει» η εφαρμογή, θα πρέπει να έχουμε εγκατεστημένο σε λειτουργικό σύστημα Windows, το Speech SDK (SAPI). Μπορεί να κατεβαστεί από εδώ: http://www.microsoft.com/downloads/details.aspx?FamilyID=5e86ec97-40a7-453f-b0ee-6583171b4530&displaylang=en.

Στην ίδια σελίδα αναφέρονται όλες οι απαιτήσεις για να «τρέξει» το SAPI.

6.1. Δηλώσεις

```
<script type="text/javascript">


var Recog= new ActiveXObject ("Sapi.SpSharedRecognizer");

var Ctxt = Recog.CreateRecoContext ( );

Ctxt.EventInterests = 16;

var myGrammar = Ctxt.CreateGrammar (0);
```

```
myGrammar.DictationLoad ( );

myGrammar.DictationSetState (1);

</script>
```

Η δήλωση: var Recog= new ActiveXObject("Sapi.SpSharedRecognizer");
δημιουργεί μία μηχανή αναγνώρισης ομιλίας (*speech recognition engine*), την οποία εκχωρεί στην μεταβλητή **Recog** και είναι τύπου ActiveX, ώστε να μπορεί να χρησιμοποιηθεί σε web περιβάλλον.

Η δήλωση: var Ctxt = Recog.CreateRecoContext ( );
δημιουργεί ένα πλαίσιο (*Context*), το εκχωρεί στην μεταβλητή **Ctxt**, μέσα στο οποίο οργανώνεται οτιδήποτε έχει σχέση με την αναγνώριση ομιλίας, όπως διάφορα γεγονότα και λεξικά που χρησιμοποιούνται κατά την διάρκεια της αναγνώρισης ομίλιας. Είναι το αντικείμενο που πραγματοποιεί την επικοινωνία μεταξύ του SAPI και της εφαρμογής.

Η δήλωση: Ctxt.EventInterests = 16;
δηλώνει ότι το γεγονός το οποίο θα γίνεται δεκτό από την μηχανή αναγνώρισης ομιλίας προς την εφαρμογή είναι το *SRERecognition*, το οποίο μας επιστρέφει την καλύτερη υπόθεση από κάποιο ηχητικό ερέθισμα. Δηλαδή, αυτό που γίνεται είναι ένα φιλτράρισμα, όπου αποκλείονται τα γεγονότα που δεν μας ενδιαφέρουν. Από προεπιλογή, το SAPI επιτρέπει όλα τα γεγονότα. Τα γεγονότα αυτά είναι τύπου enumerate και η τιμή που έχει εκχωρηθεί στο *SRERecognition* είναι 16. Εναλλακτικά, θα μπορούσε να έχει γίνει η δήλωση:
Ctxt.EventInterests = SRERecognition.

Η δήλωση: var myGrammar = Ctxt.CreateGrammar (0);
δημιουργεί ένα λεξικό για την εφαρμογή μας και το καταχωρεί στην μεταβλητή myGrammar. Το στοιχείο Ctxt περιλαμβάνει την μέθοδο CreateGrammar ( ), η οποία δημιουργεί το λεξικό. Το λεξικό αυτό είναι πλέον συνδεδεμένο με το Ctxt. Το όρισμα που έχει η μέθοδος CreateGrammar ( ), δηλώνει το id του λεξικού και είναι προαιρετικό. Εδώ χρησιμοποιείται ο αριθμός 0, που είναι το default όρισμα. Το λεξικό μέχρι τώρα δεν περιέχει όμως καμία λέξη ή έκφραση.

Η δήλωση: myGrammar.DictationLoad ( );
με την χρήση της μέθοδου DictationLoad ( ), κάνει το λεξικό της εφαρμογής ικανό να δεχθεί υπαγόρευση ομιλίας (dictation). Φορτώνονται δηλαδή στο λεξικό λέξεις

και εκφράσεις της αγγλικής γλώσσας. Επίσης, προσδιορίζεται ότι το λεξικό θα είναι στατικό, το οποίο σημαίνει ότι οι εκφράσεις που περιέχει δεν μπορούν να αλλάξουν κατά την διάρκεια της εκτέλεσης.

Η δήλωση: myGrammar.DictationSetState (1);

θέτει την κατάσταση του λεξικού σε ενεργή. Αυτό σημαίνει ότι το λέξικό μπορεί να χρησιμοποιηθεί. Η DictationSetState ( ) δέχεται ως όρισμα ένα αντικείμενο τύπου *SpeechRuleState*, το οποίο ορίζει την κατάσταση μίας έκφρασης (ενεργή, ανενεργή, σε παύση). Η τιμή 1 που δίνεται εδώ, ισοδυναμεί με την τιμή *SGDSActive*, που θέτει μία έκφραση ενεργή.

6.2. Χειρισμός γεγονότων

```
<script type="text/javascript">

  //event handling

   function Ctxt::Recognition(StreamNum,StreamPos,RecogType,Result)
  {

   var newItem;

   switch(Result.PhraseInfo.GetText())
   {

case "links":
    GetLinks();
    break;

    case "paragraphs":
    GetParagraphs();
    break;

    case "images":
```

```
        GetImages();
        break;

        case "headers":
        GetHeaders();
        break;

        case "all":
        GetAll();
        break;

    default:
    createAlternates();


     }
```

Για λόγους ευκρίνειας του κώδικα δημιουργώ ξεχωριστό script για τον χειρισμό γεγονότων. Στο παραπάνω απόσπασμα κώδικα, καλείται το γεγονός *Recognition* του αντικειμένου *Ctxt*, το οποίο προκύπτει όταν η μηχανή αναγνώρισης αναγνωρίζει μία έκφραση. Για να προκύψει αυτό το γεγονός, θα πρέπει η έκφραση που ειπώθηκε να βρίσκεται στο λεξικό, και η ποιότητα του ήχου να ξεπερνά ένα συγκεκριμένο όριο (*confidence score*). Αυτό συμβαίνει για όσο το δυνατό μεγαλύτερη ακρίβεια στην αναγνώριση. Το όριο αυτό ορίζεται από το SAPI. Εάν αυτά τα κριτήρια δεν ικανοποιηθούν, τότε δεν παράγεται αυτό το γεγονός, και συνεπώς δεν υπάρχει καμία αναγνωρισμένη έκφραση. Είναι το σημαντικότερο γεγονός που υπάρχει στο SAPI, καθώς μας επιστρέφει το αποτέλεσμα μιας επιτυχούς αναγνώρισης ομιλίας. Το γεγονός Recognition δηλώνεται με την εξής γραμμή κώδικα:

```
function Ctxt::Recognition(StreamNum,StreamPos,RecogType,Result).
```

Η τιμή **StreamNum** είναι μία μεταβλητή Long τύπου και προσδιορίζει τον αριθμό του Stream (Stream Number) που έχει η συγκεκριμένη αναγνώριση. Κάθε φορά είναι διαφορετικό.

Η τιμή **StreamPos** προσδιορίζει την θέση της έκφρασης μέσα στο Stream και είναι τύπου Variant (προσαρμοζόμενου τύπου).

Η τιμή **RecogType** προσδιορίζει το είδος της αναγνώρισης από το οποίο προήλθε το αποτέλεσμα. Το είδος της αναγνώρισης είναι *SRTStandard*, δηλαδή κανονική αναγνώριση. Το RecogType είναι τύπου *SpeechRecognitionType*, ο οποίος είναι enumerate. Η τιμή *SRTStandard* ισούται με 0. Σε αυτή την εφαρμογή δεν χρησιμοποιείται κανένας άλλος τύπος αναγνώρισης, εκτός από την κανονική.

Η τιμή **Result** είναι ένα αντικείμενο τύπου *IspeechRecoResult*, το οποίο και περιέχει το αποτέλεσμα της αναγνώρισης ομιλίας, καθώς και άλλες πληροφορίες, όπως εναλλακτικά αποτελέσματα της έκφρασης που αναγνωρίσθηκε.

Στη συνέχεια, με την δήλωση: switch(Result.PhraseInfo.GetText αν το αποτέλεσμα της αναγνώρισης είναι μία απο τις λέξεις*: links, paragraphs, images, headers* ή *all*, καλούνται και εκτελούνται αντιστοίχως οι μέθοδοι: *GetLinks( ), GetParagraphs( ), GetImages( ), GetHeaders( )* και *GetAll( )*. Εάν το αποτέλεσμα της αναγνώρισης δεν είναι κάποια από αυτές τις λέξεις, τότε καλείται η default επιλογή, που είναι η εκτέλεση της *createAltenates* ( ). Ακόμα, κάθε λέξη που υπαγορεύεται γράφεται στην TextArea με τον κώδικα:

```
dictation_filling.innerHTML=dictation_filling.innerHTML+
" "+Result.PhraseInfo.GetText();
```

6.3. Ανάλυση της λειτουργίας των συναρτήσεων

6.3.1. GetLinks

Η συνάρτηση **GetLinks( )** χρησιμοποιείται για να βρίσκει και να εμφανίζει σε μία λίστα όλους τους συνδέσμους ενός HTML αρχείου. Ο κώδικας της συνάρτησης είναι ο εξής:

```
function GetLinks()
{
 for(i=0;i<(document.links.length);i++)
   {
    newItem=  document.createElement("OPTION");
```

```
        newItem.innerText=document.links(i);
        resultList.add(new Option(newItem.innerText),null);
       }
     }
```

Κάθε σύνδεσμος που προκύπτει, τοποθετείται στο drop-down μενού *resultList*, δημιουργούνται αντικείμενα τύπου <OPTION>, με την χρήση της μεθόδου document.createElement("OPTION"); και εκχωρούνται στην μεταβλητή *newItem.*

Το κάθε μέλος του μενού, που πλέον περιέχει ένα σύνδεσμο, προστίθεται με την μέθοδο *add( )* στην *resultList* με την έκφραση:
resultList.add(new Option(newItem.innerText),null);

6.3.2. GetParagraphs

Η συνάρτηση **GetParagraphs( )** συγκεντρώνει και εμφανίζει στο μενού *resultList* όλες τις παραγράφους ενός HTML κειμένου, που έχουν σημανθεί με την σήμανση <P>. Ο κώδικας της συνάρτησης GetParagraphs( ), είναι ο εξής:

```
function GetParagraphs()
{
   for(i=0;(i<document.getElementsByTagName("P").length);i++)
    {
   P_table=document.getElementsByTagName("P");
   newItem=document.createElement("OPTION");
   newItem.innerText=P_table[i].innerText;
   resultList.add(new Option(newItem.innerText),null);
 }
  }
```

Αποθηκεύει σε ένα πίνακα όλες τις <P> σημάνσεις, χρησιμοποιώντας την μέθοδο *getElementsByTagName( )*:
P_table=document.getElementsByTagName("P");

Η λειτουργία της έκφρασης document.createElement("OPTION"); έχει αναλυθεί στην *GetLinks( )*;

Στη συνέχεια, το κείμενο της κάθε παραγράφου αποθηκεύεται στην *newItem* ως εξής: newItem.innerText=P_table[i].innerText;

Επίσης, η λειτουργία της έκφρασης resultList.add(new Option(newItem.innerText),null); έχει αναλυθεί στην *GetLinks( )*;.

6.3.3. GetImages

Η συνάρτηση **GetImages( )** συγκεντρώνει και εμφανίζει στο *resultList* όλα τα alts των στοιχείων IMG που έχουν δηλωθεί σε ένα HTML κείμενο. Ο κώδικας της είναι ο εξής:

```
function GetImages()
 {
 for(i=0;i<document.images.length;i++)
          {
 newItem=document.createElement("OPTION");
 img_collection=document.images;
 newItem.innerText=img_collection[i].alt;
 resultList.add(new Option(newItem.innerText),null);
           }
}
```

Ανακτάται μία συλλογή με όλα τα images του HTML κειμένου και αποθηκεύεται στην *img_collection* με την έκφραση:
img_collection=document.images;

Στην επόμενη γραμμή κώδικα, από κάθε στοιχείο IMG ανακτάται το χαρακτηριστικό *alt* και αποθηκεύεται στην *newItem* ως εξής:
newItem.innerText=img_collection[i].alt;

Τέλος, χρησιμοποιείται η έκφραση: resultList.add(new Option(newItem.innerText),null); της οποίας η λειτουργία έχει αναλυθεί.

6.3.4. GetHeaders

Η συνάρτηση **GetHeaders( )** συγκεντρώνει και εμφανίζει σε μία λίστα όλες τις επικεφαλίδες που δηλώνονται σε ένα HTML κείμενο με την σήμανση <H>. Ο κώδικας της είναι ο εξής:

```
function GetHeaders()
   {
   for(i=1;i<7;i++)
    {
   heds=document.getElementsByTagName("h"+i);
   hd_len=heds.length;

   for(j=0;j<hd_len;j++)
            {
   newItem=document.createElement("OPTION");
   newItem.innerText=heds.item(j).firstChild.data;
   resultList.add(new Option(newItem.innerText),null);
                  }
                 }
   }
```

Η GetHeaders( ) χρησιμοποιεί δύο βρόχους επανάληψης. Ο πρώτος, ο οποίος εκτελείται 6 φορές, όσο δηλαδή και το πλήθος των διαφορετικού τύπου επικεφαλίδων, εκτελεί την λειτουργία της αποθήκευσης σε ένα πίνακα των επικεφαλίδων ανά τύπο και του υπολογισμού του μεγέθους του πίνακα που προκύπτει κάθε φορά. Αυτό γίνεται με τις εκφράσεις:

```
heds=document.getElementsByTagName("h"+i);
hd_len=heds.length;
```

Ο δεύτερος βρόχος επανάληψης εκτελείται τόσες φορές, όσο και το πλήθος των επικεφαλίδων ανά τύπο. Το πλήθος αυτό είναι αποθηκευμένο στη μεταβλητή *hd_len*. Στη συνέχεια, δημιουργείται ένα στοιχείο μενού με την εντολή:

newItem=document.createElement("OPTION");

Στην επόμενη γραμμή κώδικα ανακτάται το κείμενο κάθε καταχώρησης του πίνακα *heds* και αποθηκεύεται στην *newItem* ως εξής:

newItem.innerText=heds.item(j).firstChild.data;

Τέλος, χρησιμοποιείται η έκφραση: resultList.add(new Option(newItem.innerText),null); για να προσθέσει κάθε νέα καταχώρηση στο *resultList*.

Για παράδειγμα, ας υποθέσουμε ότι το i έχει την τιμή 1. Ο πίνακας heds θα αποθηκεύσει όλες τις επικεφαλίδες τύπου h1 αφού στο όρισμα της getElementsByTagName( ) γίνεται συνένωση συμβολοσειρών, και προκύπτει έτσι η συμβολοσειρά ‘’h1’’. Στην μεταβλητή hd_len θα αποθηκευθεί το πλήθος των h1 επικεφαλίδων του HTML κειμένου. Ας υποθέσουμε ότι είναι 2. Αυτό σημαίνει ότι ο δεύτερος βρόχος θα εκτελεστεί 2 φορές, και έτσι το κείμενο των δυο επικεφαλίδων θα καταχωρηθεί στο r*esultList*.

6.3.5. GetAll

Η συνάρτηση **GetAll( )** συγκεντρώνει και εμφανίζει στο r*esultList* όλα τα στοιχεία που υπάρχουν σε ένα HTML κείμενο. Ο κώδικας της είναι ο εξής:

```
function GetAll()
 {
 for(i=0;i<(document.all.length);i++)
        {
 newItem=document.createElement("OPTION");
 newItem.innerText=document.all(i);
 resultList.add(new Option(newItem.innerText),null);
        }
}
```

Κάθε στοιχείο αποθηκεύεται στην *newItem* ως εξής: newItem.innerText=document.all(i);

Τέλος, κατά τα γνωστά, προστίθεται στο *result_list* η *newItem*.

6.3.6. createAlternates

Η **createAlternates( )** είναι η συνάρτηση που εκτελείται όταν δεν επιλέγεται καμία από τις προηγούμενες επιλογές. Είναι δηλαδή η default επιλογή. Το αποτέλεσμα της εκτέλεσής της είναι η εμφάνιση σε μία λίστα 10 εναλλακτικών εκφράσεων, που η πιθανότητα που συγκέντρωσαν είναι μικρότερη από αυτή του αποτελέσματος, και πιθανώς κάποια από αυτές τις εναλλακτικές εκφράσεις να είναι αυτό που ο χρήστης είπε. Ο κώδικας της συνάρτησης είναι ο ακόλουθος:

```
function createAlternates()
  {

        var alter=Result.Alternates(10);
        var alterObject;

        for(i=0;i<(alter.Count);i++)
                {
        newItem=document.createElement("OPTION");
        alterObject=alter.Item(i);
        newItem.innerText=alterObject.PhraseInfo.GetText();
         document.getElementById("alter_selections").options[i]=new
        Option(newItem.innerText);
                 }
       }
```

Στην πρώτη γραμμή της μεθόδου δημιουργούνται οι δέκα εκφράσεις που έχουν την μεγαλύτερη πιθανότητα να έχει εκφέρει ο χρήστης. Οι εναλλακτικές αυτές εκφράσεις ανακτώνται με την μέθοδο *Alternates* του αντικειμένου *Result* και αποθηκεύονται στη συλλογή αντικειμένων τύπου *ISpeechPhraseAlternate* ***alter***. Η *alter* είναι τύπου *ISpeechPhraseAlternates,* που είναι ο τύπος της συλλογής αντικειμένων *ISpeechPhraseAlternate.* Ο αριθμός 10 ως όρισμα της *Alternates* ορίζει τον αριθμό των εναλλάκτικων εκφράσεων που θα ανακτηθούν με το πρώτο αποτέλεσμα να έχει την μεγαλύτερη πιθανότητα να είναι αυτό που ο χρήστης έχει

πει, ενώ το δέκατο έχει την μικρότερη. Η έκφραση που δημιουργεί τις εναλλακτικές εκφράσεις είναι η εξής:

```
var alter=Result.Alternates(10);
```

Τα αντικείμενα τύπου *ISpeechPhraseAlternates* έχουν δύο ιδιότητες. Την **Count**, η οποία επιστρέφει τον αριθμό των *ISpeechPhraseAlternate* αντικειμένων της συλλογής, και την ιδιότητα **Item** η οποία επιστρέφει ένα στοιχείο της συλλογής με βάση την θέση του. Αυτές οι δύο ιδιότητες χρησιμοποιούνται στη συνέχεια.

Η *Count* χρησιμοποιείται για να υπολογιστεί ο αριθμός των επαναλήψεων του βρόχου.

Με την χρήση της ιδιότητας *Item*, ανάλογα με την τιμή του *i*, η *alter* επιστρέφει στην *alterObject* ένα αντικείμενο τύπου *ISpeechPhraseAlternate*. Η έκφραση είναι η ακόλουθη: alterObject=alter.Item(i);

Μία από τις ιδιότητες ενός αντικειμένου *ISpeechRecoResult* (όπως είναι το *Result*) είναι η *PhraseInfo*, η οποία επιστρέφει ένα αντικείμενο τύπου *ISpeechPhraseInfo*. Στο αντικείμενο αυτό περιέχονται πληροφορίες σχετικά με την έκφραση που ειπώθηκε, όπως το μέγεθος της σε bytes, το id του λεξικού που χρησιμοποιήθηκε για να γίνει η αναγνώριση και άλλες. Επίσης περιέχει και την μέθοδο *GetText*, που επιστρέφει το αποτέλεσμα της αναγνώρισης σε μορφή *String*. Έτσι ανακτάται η έκφραση που ειπώθηκε σε μορφή *String.*

Στη συνέχεια, χρησιμοποιείται η ιδιότητα *PhraseInfo* για να ανακτηθεί η έκφραση σε μορφή *String*. Αυτό το String αποθηκεύεται στην μεταβλητή *newItem*. Η γραμμή κώδικα είναι:

```
newItem.innerText=alterObject.PhraseInfo.GetText();
```

Τέλος, κάθε εναλλακτική έκφραση που έχει ανακτηθεί προστίθεται στο drop-down μενού *alter_selections*. Αυτό πραγματοποιείται με την έκφραση:

```
document.getElementById("alter_selections").options[i]=new
Option(newItem.innerText);
```

Συμπεράσματα

Το κυριότερο συμπέρασμα που εξήχθη από αυτή την εργασία είναι η συνειδητοποίηση της χρησιμότητας των συστημάτων αναγνώρισης ομιλίας. Διαπιστώθηκε ότι ήδη χρησιμοποιούνται σε πολλούς και διαφορετικούς τομείς, όπως από άτομα με ειδικές ανάγκες ως βοήθημα στον έλεγχο του υπολογιστή ή ως βοήθημα στην εκπαίδευση, σε στρατιωτικές εφαρμογές, στον χειρισμό των συσκευών πλοήγησης και του ηχοσυστήματος των αυτοκινήτων, σε IVR συστήματα και σε άλλα πεδία.

Η χρήση από την δεκαετία 2000-2010 από αεροδρόμια, εταιρείες στοιχημάτων με μεγάλο αριθμό πελατών κ.ά. δείχνει ότι υπάρχει μία τάση χρήσης της και ανοίγει τον δρόμο για την χρησιμοποίηση της σε μεγαλύτερο βαθμό. Ακόμα, μετά και την πρόσφατη σχετικά ενασχόληση της Google με σχετικές τεχνολογίες, ο ανταγωνισμός των εταιρειών που προσφέρουν ανάλογο λογισμικό μεγαλώνει. Να σημειωθεί ότι εταιρείες κολοσσοί του χώρου της πληροφορικής ασχολούνται με την αναγνώριση ομιλίας, όπως η Microsoft, πακέτο της οποίας χρησιμοποιήθηκε για την δημιουργία της εφαρμογής, και η IBM, πράγμα που λογικά θα επιφέρει γρήγορη εξέλιξη των υπαρχόντων συστημάτων.

Η ακρίβεια των σημερινών συστημάτων αναγνώρισης ομιλίας φτάνει το 98%. Αυτό σημαίνει ότι στις 100 λέξεις που ένας χρήστης προφέρει, οι 98 αναγνωρίζονται επιτυχώς. Όμως αυτές οι επιδόσεις επιτυγχάνονται μόνο σε ιδανικές συνθήκες, δηλαδή σε χώρο όπου δεν υπάρχει θόρυβος από το περιβάλλον, και ο ομιλητής έχει καλή άρθρωση. Αν η έρευνα που γίνεται σχετικά με την αναγνώριση σε μη ιδανικό περιβάλλον, δηλαδή στο καθημερινό περιβάλλον που οι άνθρωποι καλούνται να επικοινωνήσουν και να αναγνωρίσουν τα λεγόμενα κάποιου άλλου επιτύχει, η αναγνώριση ομιλίας θα γνωρίσει μεγάλη άνθηση. Τότε, ίσως να αρχίσουμε να χρησιμοποιούμε εφαρμογές αναγνώρισης ομιλίας σε τομείς που σήμερα ούτε καν φανταζόμαστε.

Ακόμα, για την εφαρμογή που δημιουργήθηκε, παρατηρήθηκε ότι: η δημιουργία της δεν είναι ιδιαιτέρως πολύπλοκη υπόθεση. Το μόνο που χρειάζεται είναι να ανατρέξει κάποιος στο εγχειρίδιο του SAPI. Το μειονέκτημα της εφαρμογής είναι ότι λόγω της δημιουργίας και χρήσης ActiveXObject για την ενσωμάτωση λειτουργιών αναγνώρισης ομιλίας, μπορεί να εκτελεστεί μόνο από

browser Internet Explorer. Ακόμα, η αναγνώριση των φωνητικών εντολών, παρουσιάζει μικρά ποσοστά ακρίβειας, παρότι το σύστημα εκπαιδεύτηκε με τη φωνή μου. Από τις δοκιμές που έγιναν διαπιστώθηκε ότι είναι προτιμότερο να εκφέρουμε προτάσεις, παρά μεμονωμένες λέξεις. Αυτό συμβαίνει λόγω της ύπαρξης του γλωσσικού μοντέλου. Το SAPI έχει επιλογή για χρήση ενός λεξικού ορισμένο από τον χρήστη, με το οποίο θα επιτυγχάνονταν μεγαλύτερα ποσοστά ακριβείας για μεμονωμένες λέξεις, όμως χρειάζεται να δημιουργηθεί ένα ξεχωριστό αρχείο που να περιέχει τις επιτρεπόμενες λέξεις, πράγμα που είναι αδύνατο για την JavaScript να το χειριστεί. Γενικά, η αναγνώριση προτάσεων είναι σε αποδεκτά επίπεδα.

Βιβλιογραφία

Βοσνίδης, Χ. (2002). *Υλοποίηση του συστήματος αναγνώρισης ομιλίας*. Retrieved from Technical University of Crete: http://www.telecom.tuc.gr/courses/speech/speech_lectures/docs/11-Recognition_System.pdf

Γλώσσα και ορθογραφία. Retrieved from Ηλεκτρονικός κόμβος για την υποστήριξη των διδασκόντων την Ελληνική γλώσσα: http://www.komvos.edu.gr/glwssa/Odigos/thema_d10/main.htm

Έγκλιση. (n.d.) In *Πύλη για την ελληνική γλώσσα*. Retrieved from http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/search.html?lq=%CE%B5%CE%B3%CE%BA%CE%BB%CE%B9%CF%83%CE%B7&dq=

Κουρεμένος, Κ. (2009, 25 March). *Γραμματική της Ελληνικής Γλώσσας(Holton, Mackridge, Φιλιππάκη-Warburton), 01*. Message posted on http://enaskitis.blogspot.com/

Μιχαλέτου, Ε. (2008). *Παραμετροποίηση σήματος ομιλίας για αναγνώριση συναισθήματος ομιλητή*. Retrieved from Νημερτής: http://nemertes.lis.upatras.gr/dspace/bitstream/123456789/1191/3/Nimertis_Michaletou.pdf

Τεχνολογία αναγνώρισης φωνής. (2004). Retrieved from Ηλεκτρονικός Λογογράφος: http://www.logografos.gr/technologynews.aspx?lID=EL&technologyNewID=2&si=1

Φώνημα (n.d.) In *Πύλη για την ελληνική γλώσσα*. Retrieved from: http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/glossology/show.html?id=72

Aarnio, T. (1999), *Speech Recognition with Hidden Markov Models in Visual Communication*. Retrieved from Nokia Research Center:

Automated Speech Recognition (n.d.) In *SearchMobileComputing Definitions*. Retrieved from: http://searchmobilecomputing.techtarget.com/sDefinition/0,,sid40_gci786138,00.html

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C, and Morgan N. (2007), *Historical Development and Future Directions in Speech Recognition and Understanding.* Retrieved from Retrieval Group: http://www-nlpir.nist.gov/MINDS/FINAL/speech.web.pdf

Cook, S. (2002). *Speech Recognition HOWTO*. Retrieved from faqs.org: http://www.faqs.org/docs/Linux-HOWTO/Speech-Recognition-HOWTO.html

Dynamic Programming (n.d.). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Dynamic_programming

Electronic Medical record (n.d). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Electronic_medical_record

Englund, C, (2004). *Speech Recognition in the JAS 39 Gripen aircraft-adaptation to speech at different G-loads*. Retrieved from: http://www.speech.kth.se/prod/publications/files/1664.pdf

Feldon, B. (2008). *The top five uses of speech recognition technology.* Retrieved from Call Centre Helper: http://www.callcentrehelper.com/the-top-five-uses-of-speech-recognition-technology-1536.htm

Filter Bank(n.d.) Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Filter_bank

Finite State Machine (n.d.). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Finite_state_machine

Furui, S. (1997). Speaker Recognition. Retrieved from Center for Spoken Language Understanding: http://cslu.cse.ogi.edu/HLTsurvey/ch1node9.html

Getting started with MS Speech Recognition. Retrieved from Tufts University: http://training.uit.tufts.edu/pdftips/speechtips.pdf

Global Positioning System (n.d.) Retrieved from Wikipedia: http://el.wikipedia.org/wiki/Global_Positioning_System

How Speech Recognition Works. (2001). Retrieved from: http://project.uet.itgo.com/speech.htm

How Speech Recognition Works. Retrieved from: http://electronics.howstuffworks.com/speech-recognition.htm

Huang, C. Hsu, W., Chang, SF. (2003). *Automatic Closed Caption Alignment Based on Speech Recognition Transcripts*. Retrieved from Center For Telecommunications Research, Columbia University: http://www.ctr.columbia.edu/papers_advent/03/align03huang.pdf

Interactive Voice response (n.d.) Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Interactive_voice_response

Juang, B.H. Rabiner, L. (2004). *Automatic Speech Recognition-A Brief History of the Technology Development*. Retrieved from University of California, Santa Barbara: http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf

Jurafsky, D. Martin,J.(2009). *An Introduction to Natural Language Procecing, Computational Linguistics, and Speech Recognition*. Retrieved from: http://books.google.com/books?id=fZmj5UNK8AQC&pg=PA346#v=onepage&q&f=false

Kotadia, M. (2007). *Voice biometrics replaces ID check at AHM*. Retrieved from Zdnet: http://www.zdnet.com.au/news/software/soa/Voice-biometrics-replaces-ID-check-at-AHM/0,130061733,339274060,00.htm

Matthews, J. (2002). *How does speech recognition works?* Retrieved from Generation5: http://www.generation5.org/content/2002/howsrworks.asp

McKnight, J. and McKnight, S. (1991), *The Effect of Cellular Phone Use Upon Driver Attention*, Retrieved from AAA Foundation for Traffic Safety: http://www.aaafoundation.org/resources/index.cfm?button=cellphone

Mike Faraone (2009). Voice Recognition Transcription Software: The Advantages. *Articlesbase*. http://www.articlesbase.com/software-articles/voice-recognition-transcription-software-the-advantages-1527014.html

Moores, M. and Minchin, S. (2 October 2007). *Dublin Airport enhances customer service with a new "speech" based Flight Information Service*. Retrieved from Sourcewire.com: http://www.sourcewire.com/releases/rel_display.php?relid=34124

Muthusamy, Y., and Spitz, L., (1997). *Automatic Language Identification*. Retrieved from Center for Spoken Language Understanding: http://cslu.cse.ogi.edu/HLTsurvey/ch8node9.html

Noppa Navigation and Guidance for the Blind. (2004). Retrieved from European Local Transport Information Service: http://www.eltis.org/docs/studies/noppa_vtt.pdf

Pattern Recognition (n.d.). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Pattern_recognition

Revis, M. (2005). The Case for Speech Recognition. *For The Record, volume 17*. Retrieved from ForTheRecord: http://www.fortherecordmag.com/archives/ftr_042505p20.shtml

Roukos, S. (1997). *Language Representation*. Retrieved from Center for Spoken Language Understanding: http://cslu.cse.ogi.edu/HLTsurvey/ch1node8.html#SECTION163

Sahuguet, A. and Bezman, A. (2008, 14 June). “In their own words” Political Videos meet Google speech-to-text technology. Message posted to: http://googleblog.blogspot.com/2008/07/in-their-own-words-political-videos.html

Speaker Recognition (n.d.) Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Speaker_recognition

Speech and Language in Military Application. Retrieved from NATO: http://ftp.rta.nato.int/public//PubFullText/RTO/TR/RTO-TR-IST-037///TR-IST-037-02.pdf

Speech corpus (n.d.) Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Speech_corpus

Speech Recognition (n.d). Retrieved from the Wiki Wikia Science: http://future.wikia.com/wiki/Speech_Recognition

Speech Recognition (n.d.) In *Babylon online dictionary*. Retrieved from: http://dictionary.babylon.com/

Speech Recognition (n.d.) Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Speech_recognition

Speech Recognition Accuracy in 80db of Noise. (2001). Retrieved from Sensory: http://www.sensoryinc.com/support/docs/80-0176-B.pdf

Speech recognition for students with disabilities. (2007). Retrieved from Custom Typing Training: http://www.customtyping.com/tutorials/sr/students_disabilities.htm

Speech Recognition GPS Navigation. (2009). Retrieved from GPS Technology Review: http://gpstekreviews.com/2009/03/27/speech-recognition-gps-navigation/

Speech Recognition: Accelerating the Adoption of Electronic Medical reports. (2008). Retrieved from Nuance Healthcare Solutions: http://www.nuance.com/healthcare/pdf/wp_healthcare_MDEMRadopt.pdf

StateWorks. [Graph Illustration of a finite state machine]. *StateWorks: Finite State Machine*. Retrieved from: http://www.stateworks.com/technology/finite_state_machine/

Types of Speech Recognition. Retrieved from Lumenvox: http://www.lumenvox.com/resources/tips/types-of-speech-recognition.aspx

Voice Commands. Retrieved from Giant Bombs: http://www.giantbomb.com/voice-commands/92-295/

What is IVR? Retrieved from Voxeo: http://www.voxeo.com/library/ivr.jsp

Word error rate (n.d.) Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Word_error_rate

Xbox 360 Video Game to Feature Voice Recognition from Fonix. (2006). Retrieved from Eweek: http://www.eweek.com/c/a/VOIP-and-Telephony/Xbox-360-Video-Game-to-Feature-Voice-Recognition-Technology-from-Fonix/

Young, S. (2008), Hmm and Related Speech Recognition Technologies. *In Springer handbook of speech processing (*Speech Recognition). Retrieved from: http://books.google.com/books?id=Slg10ekZBkAC&pg=PA519&dq=springer+speech+recognition&hl=el&cd=1#v=onepage&q=springer%20speech%20recognition&f=false

Zue,V., Cole, R., Ward, W. (1995). *Speech Recognition.* Retrieved from Center for Spoken Language Understanding: http://cslu.cse.ogi.edu/HLTsurvey/ch1node4.html#seczuecoleward

Παράρτημα Α - Ο κώδικας της εφαρμογής

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"

"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">


<html xmlns="http://www.w3.org/1999/xhtml">
<head>
   <title>Speech Recognition Tool</title>
   <style type="text/css">
      .style1
      {
         font-size: xx-large;
         text-align: center;
      }

      #alter_selections
      {
         width: 176px;
         margin-left: 8px;
         height: 143px;
         margin-top: 0px;
      }

      #resultList
      {
         width: 214px;
         height: 27px;
         margin-right: 0px;
         margin-top: 0px;
      }

      #dictation_filling
      {
         width: 496px;
         height: 229px;
         margin-left: 205px;
         margin-top: 21px;
```

```
        margin-right: 0px;
      }
  </style>
</head>
  <body>

  <script type="text/javascript">

  var Recog = new ActiveXObject("Sapi.SpSharedRecognizer");
  var Ctxt = Recog.CreateRecoContext();
  Ctxt.EventInterests = 16;
  var myGrammar = Ctxt.CreateGrammar(0);
  myGrammar.DictationLoad();
  myGrammar.DictationSetState(1);


</script>

<script type="text/javascript">

 //event handling

  function Ctxt::Recognition(StreamNum,StreamPos,RecogType,Result)

  var newItem;

  switch(Result.PhraseInfo.GetText())
  {

  case "links":
  GetLinks();
 break;

 case "paragraphs":
 GetParagraphs();
 break;

 case "images":
 GetImages();
 break;
```

```
case "headers":
GetHeaders();
break;

case "all":
 GetAll();
 break;

  default:
  createAlternates();

}

dictation_filling.innerHTML=dictation_filling.innerHTML+
  " "+Result.PhraseInfo.GetText();

  function GetLinks()
  {
  for(i=0;i<(document.links.length);i++)
         {
  newItem=document.createElement("OPTION");
  newItem.innerText=document.links(i);
  resultList.add(new Option(newItem.innerText),null);
         }
}

  function GetParagraphs()
  {
  for(i=0;(i<document.getElementsByTagName("P").length);i++)
           {
  P_table=document.getElementsByTagName("P");
  newItem=document.createElement("OPTION");
  newItem.innerText=P_table[i].innerText;
  resultList.add(new Option(newItem.innerText),null);
          }
}

  function GetImages()
  {
```

```
 for(i=0;i<document.images.length;i++)
        {
newItem=document.createElement("OPTION");
img_collection=document.images;
newItem.innerText=img_collection[i].alt;
resultList.add(new Option(newItem.innerText),null);
                    }
}

 function GetHeaders()
 {
 for(i=1;i<7;i++)
 {
 heds=document.getElementsByTagName("h"+i);
 hd_len=heds.length;

 for(j=0;j<hd_len;j++)
 {
 newItem=document.createElement("OPTION");
 newItem.innerText=heds.item(j).firstChild.data;
 resultList.add(new Option(newItem.innerText),null);

                  }

        }

 }

 function GetAll()
 {
 for(i=0;i<(document.all.length);i++)
 {
 newItem=document.createElement("OPTION");
 newItem.innerText=document.all(i);
 resultList.add(new Option(newItem.innerText),null);

          }
 }

function createAlternates()
```

```
{

        var alter=Result.Alternates(10);
        var alterObject;


        for(i=0;i<(alter.Count);i++)
        {
        newItem=document.createElement("OPTION");
        alterObject=alter.Item(i);
        newItem.innerText=alterObject.PhraseInfo.GetText();
         document.getElementById("alter_selections").options[i]=
        new Option(newItem.innerText);

                }

        }

 }

</script>

    <p class="style1">
       Speech recognition Tool</p>

    <p class="style1">

  <textarea id="dictation_filling" name="dictation_filling">
  </textarea>

  <select id="resultList" name="resultList" size="2">
   </select></p>

  <select id="alter_selections" name="alter_selections" size="10">
  </select></p>

</body>
 </html>
```