Πανεπιστημιο Θεσσαλιας

Διπλωματικη εργασια

Αναγνώριση ακουστικών γεγονότων με βαθιά νευρωνικά δίκτυα

Συγγραφέας:
Κωνσταντίνος Θεμελής

Επιβλέποντες:
Γεράσιμος Ποταμιάνος
Αντώνιος Αργυρίου

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

14 Σεπτεμβρίου 2015

Πανεπιστήμιο Θεσσαλίας

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Περίληψη

Η αναγνώριση ακουστικών γεγονότων σε έξυπνα περιβάλλοντα ξεκινάει να έχει όλο και περισσότερο εφαρμογή σε αρκετά προβλήματα στη σημερινή εποχή. Συγκεκριμένα, αυτό το τμήμα της επιστήμης της Αναγνώρισης Προτύπων ασχολείται με τον εντοπισμό ενός ή περισσοτέρων ηχητικών γεγονότων που λαμβάνουν χώρα σε ένα περιβάλλον όπου τα άτομα παράγουν διαφορετικούς ήχους. Υπάρχουν πολλοί ταξινομητές όπως το πολυεπίπεδο perceptron (MLP multilayer perceptron-MLP), SOM (structure adaptive self-organizing map), LDA, Bayes κτλπ, αλλά στο πλαίσιο αυτή της διπλωματικής θα πειραματιστούμε με τα κρυφά μοντέλα Markov (hidden Markov models ή HMM) και κυρίως τα βαθιά νευρωνικά δίκτυα (deep neural networks ή DNN).

Κατά τη διάρκεια του σχεδίου εργασίας, χρησιμοποιήσαμε την πολυκαναλική βάση δεδομένων του ερευνητικού κέντρου UPC-TALP που ηχογραφήθηκε από 24 μικρόφωνα τοποθετημένα σε ένα έξυπνο περιβάλλον. Σαν πρώτο βήμα επεξεργάστηκαν τα ακουστικά αρχεία ώστε να έχουν μία κοινή μορφή (16kHz) και τμηματοποιήθηκαν ανάλογα με τα αρχεία ετικετών. Ύστερα χρησιμοποιήσαμε δύο εργαλεία, το HTK και το Kaldi για να εκπαιδεύσουμε και να εξετάσουμε τη βάση δεδομένων. Σε αυτή τη διπλωματική, μοντελοποιήσαμε τα δεδομένα σαν κρυφά μοντέλα Markov και έπειτα σαν νευρωνικά δίκτυα. Στο τέλος, συγκρίναμε τα αποτελέσματα και διακρίναμε ποιά μέθοδος ανταποκρίνεται καλύτερα στο πρόβλημά μας.

Ευχαριστίες

Κατ΄ αρχήν θα ήθελα να ευχαριστήσω τον επικεφαλή επόπτη αυτής της διπλωματικής κ.Γεράσιμο Ποταμιάνο, αναπληρωτή καθηγητή του τμήματος Ηλετρολόγων Μηχανικών & Μηχανικών Υπολογιστών στον Βόλο, για την παραπάνω από χρήσιμη βοήθειά του. Η υποστήριξή και καθοδήγησή του μέσα από κάθε στάδιο, ήταν εξέχουσας σημασίας καθώς κατάφερα να προβληματιστώ με το επιστημονικό πεδίο του ενδιαφέροντός μου και να αναπτύξω τις ικανότητές μου πάνω σε αυτό. Επιπρόσθετα, θα ήθελα να ευχαριστήσω από καρδιάς την οικογένειά μου για την αγάπη και την υποστήριξή τους που με βοήθησε πνευματικά. Για την παροχή ενός εργασιακού χώρου, θα ήθελα να ευχαριστήσω τον αναπληρωτή καθηγητή κ.Νικόλαο Μπέλλα, που μου παρείχε είσοδο στα εργαστήρια Β2 του τμήματος. Τέλος, δίνω την ευγνωμοσύνη μου στους συναδέλφους και φίλους μου Χρήστο Ιωαννίδη, Ιωάννη Κωνσταντέλια και Ευάγγελο Νόνα για την βοήθειά τους σε προβλήματα του κώδικα. Χωρίς τα προαναφερθέντα άτομα αυτή η διπλωματική δεν θα υπήρχε.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Στην οικογένεια μου και στους φίλους μου...

Κεφάλαιο 1

Εισαγωγή

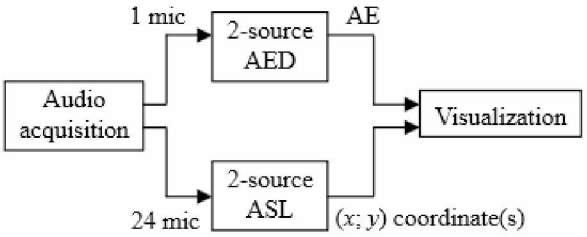
1.1 Αναγνώριση ακουστικών γεγονότων

Η αναγνώριση ακουστικών γεγονότων στοχεύει στον καθορισμό της ταυτότητας των ήχων και της πηγής με τη χρήση ενός ή περισσότερων μικροφώνων. Μπορεί έτσι να παρέχει υποστήριξη όταν εφαρμόζεται ανάλυση της ακουστικής σκηνής. Επιπρόσθετα, αυτός ο κλάδος μπορεί να λειτουργήσει σαν βελτιωτικός παράγοντας σε ταξινομητές που χρησιμοποιούνται για αναγνώριση φωνής [5]. Παρόλο που η ομιλία αποτελεί το πιο πληροφοριακό ακουστικό γεγονός, τα διαφορετικά γεγονότα που λαμβάνουν μέρος σε ένα χώρο μπορεί να φέρουν εξίσου σημαντική πληροφορία. Και αυτό γιατί μέσα από αυτά τα συμβάντα αντανακλάται η ανθρώπινη δραστηριότητα είτε από άμμεσες παρεμβάσεις στο γύρω περιβάλλον (ήχος βημάτων) είτε μέσα από τη χρήση αντικειμένων (ήχος πληκτρολογίου). Συνεπώς, με τον εντοπισμό και την κατηγοριοποίηση τέτοιων γεγονότων, μπορεί να χαρτογραφηθεί η ανθρώπινη και κοινωνική δραστηριότητα σε ένα δωμάτιο [6]. Η αναγνώριση ακουστικών γεγονότων αποτελεί παρακλάδι της επιστήμης της Αναγνώρισης Προτύπων και συγκεκριμένα της υπολογιστικής ανάλυσης ακουστικής σκηνής όπου τα ηχητικά σήματα επεξεργάζονται ώστε να εξαχθούν χαρακτηριστικά που μπορούν να τα περιγράψουν. Σύμφωνα με αυτά τα χαρακτηριστικά, εκπαιδεύονται ταξινομητές ώστε να καταστούν ικανοί να εντοπίσουν και να αναγνωρίσουν μεμονωμένα ακουστικά γεγονότα (μεμονωμένα πειράματα) ή ακολουθίες (ενσωματωμένα πειράματα).

1.2 Σχετική εργασία

Η βάση δεδομένων που χρησιμοποιήθηκε σε αυτή τη διπλωματική έχει χρησιμοποιηθεί σαν υλικό εκπαίδευσης και αξιολόγησης στο σχέδιο εργασίας "Multi-microphone fusion for detection of speech and acoustic events in smart spaces" [7]. Πειράματα πάνω στην ίδια

ακριβώς βάση πραγματοποίησαν οι Taras Butko, Fran González Pla κ.α που προσπάθησαν να αναγνωρίσουν τα ακουστικά γεγονότα μέσω HMM ταξινομητή και να εντοπίσουν την ηχητική πηγή. Η ροή εργασίας που ακολούθησαν, δημιουργώντας ένα τετοιο σύστημα φαίνεται στο Σχήμα 1.1 [1]. Το αποτέλεσμα το οποίο επέτυχαν με την εξαγωγή ποσοστού επιτυχίας σε επίπεδο πλαισίου άγγιξε το 91.5%.



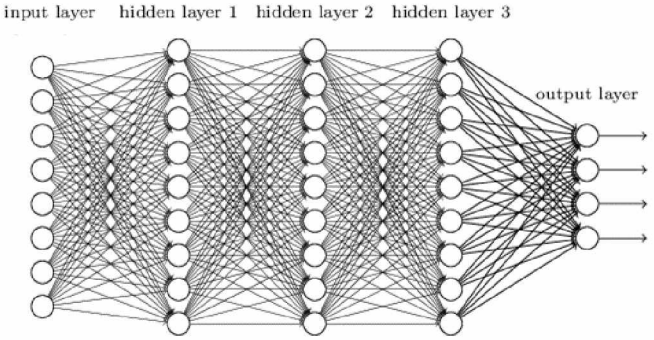
Σχημα 1.1: Αναπαράσταση του προτεινόμενου συστήματος. Λήφθηκε από το [1]

Πειράματα πάνω στα ίδια δεδομένα έγιναν επίσης και από Xiaodan Zhuang, Xi Zhou κ.α, αφού προσπάθησαν να προσεγγίσουν το πρόβλημα με συνδυασμό *μηχανών διανυσματικής στήριξης* (support vector machines ή SVM) και κρυφών μοντέλων Markov [8]. Με βάση τα MFCC χαρακτηριστικά, που θα χρησιμοποιηθούν και για αυτή την εργασία, το ποσοστό ακρίβειας του ταξινομητή έφτασε το 41.2%.

1.3 Βαθιά μάθηση και νευρωνικά δίκτυα

1.3.1 Θεωρητικό υπόβαθρο

Τα νευρωνικά δίκτυα είναι εμπνευσμένα από το βιολογικό μοντέλο που προτάθηκε από τον H. Hubel και Torsten Wiesel το 1959 [9]. Από βιολογική άποψη ο εγκέφαλος αποτελείται από νευρικά κύτταρα που ονομάζονται νευρώνες και είναι συνδεδεμένοι μεταξύ τους με δεσμούς. Αυτός ο μηχανισμός επιτρέπει τον ηλεκτρισμό να κυλλά διαμέσου του εγκεφάλου σε όλο το σώμα και έτσι καθιστά ικανή την επικοινωνία μεταξύ των κυττάρων. Έχοντας, ο εγκέφαλος αυτόν τον μηχανισμό, λαμβάνει αποφάσεις και επιτρέπει σωματικές δράσεις. Τα τεχνητά νευρωνικά δίκτυα προσπαθούν να προσομοιώσουν αυτή τη συμπεριφορά δημιουργώντας ένα σύστημα κόμβων διαιρεμένο σε συνδεδεμένα επίπεδα. Θεωρητικά μιλώντας, μπορεί να ειπωθεί ότι αν το τέλειο νευρωνικό δίκτυο κατασκευαζόταν, θα προσέγγιζε την λειτουργικότητα του ανθρώπινου εγκεφάλου ή αλλιώς όπως ονομάζεται, τεχνητή νοημοσύνη. Στο Σχήμα 1.1 παρουσιάζεται παράδειγμα ενός νευρωνικού δικτύου όπου φαίνεται η οργάνωση των επιπέδων και οι συνδέσεις μεταξύ τους.



Σχημα 1.2: Οπτικό παράδειγμα νευρωνικού δικτύου. Το μέγεθος της εισόδου στο νευρωνικό δίκτυο συμπίπτει με τη διάσταση των χαρακτηριστικών.[2]

1.4 Η διπλωματική σε μία ματιά

1.4.1 Σκοπός της διπλωματικής

Η παρούσα διπλωματική ασχολείται με το πρόβλημα του εντοπισμού και της αναγνώρισης ακουστικών γεγονότων που λαμβάνουν χώρο σε ένα έξυπνο περιβάλλον, όπως ένα γραφείο. Τα γεγονότα ηχογραφήθηκαν ταυτόχρονα από 24 μικρόφωνα. Ο σκοπός αυτού του σχεδίου εργασίας είναι να μοντελοποιήσει το πρόβλημα με δύο πιθανοτικά μοντέλα, αυτό των κρυφών μοντέλων Markov και των βαθιών νευρωνικών δικτύων, να μελετήσει την επίδοσή τους μέσα από μία σειρά πειραμάτων, να εξαγάγει τα αποτελέσματα, να εφαρμόσει πιθανές βελτιστοποιήσεις και τέλος να συγκρίνει τα αποτελέσματα.

1.4.2 Συνεισφορά της διπλωματικής

Η κύρια συνεισφορά αυτής της διπλωματικής έγκειται στην εφαρμογή της μεθόδου των βαθιών νευρωνικών δικτύων στα δεδομένα της πολυκαναλικής βάσης δεδομένων UPC-TALP με σκοπό να εκτιμηθεί κατά πόσο ανιχνεύουν τα ακουστικά γεγονότα σε έξυπνα περιβάλλοντα. Πιο ειδικά, η επιστημονική συνεισφορά της εργασίας μπορεί να συνοψιστεί στους ακόλουθους τομείς:

- Πειράματα με διαφορετικές παραμετρικοποιήσεις μοντέλων και των εργαλείων HTK και Kaldi.
- Τη χρήση διαφορετικών μετρικών λάθους με σκοπό την καλύτερη και ορθότερη αξιολόγηση των ταξινομητών.
- Εύρεση του αποδοτικότερου μοντέλου βαθιού νευρωνικού δικτύου για την μοντελοποίηση του προβλήματος.

- Εξαγωγή τελικών συμπερασμάτων και αξιολόγηση της εφαρμογής νευρωνικών δικτύων στην ηχογραφημμένη βάση δεδομένων.

1.4.3 Περιεχόμενο κεφαλαίων

Εκτός της εισαγωγής, αυτή η διπλωματική αποτελείται από πέντε ακόμα κεφάλαια που συμπεριλαμβάνουν το ακόλουθο περιεχόμενο:

- ***Κεφάλαιο 2***: περιέχει όλη την πληροφορία σχετικά με τον τρόπο επίλυσης του προβλήματος, την θεωρία πίσω από τα μοντέλα Markov και τα νευρωνικά δίκτυα, την εξαγωγή των χαρακτηριστικών MFCC που χρησιμοποιήθηκαν για την αναγνώριση ακουστικών γεγονότων και τις ιδιότητές τους. Επιπλέον, παραθέτονται μαθηματικοί τύποι με σκοπό την πλήρη κατανόηση της διαδικασίας εξαγωγής των συχνοτικών-φασματικών συντελεστών Mel από ένα ηχητικό σήμα.

- ***Κεφάλαιο 3***: σε αυτό το τμήμα παρουσιάζεται η πολυκαναλική βάση δεδομένων UPC-TALP, τεχνικές πληροφορίες σχετικά με το υπόβαθρο στο οποίο ηχογραφήθηκε και οργανώθηκε. Ακόμη, αναρτείται και πίνακας όπου τα δεδομένα της βάσης έχουν κατηγοριοποιηθεί στις 8 συνεδρίες ανάλογα με το γεγονός που περιγράφουν. Σε αυτό το σημείο πρέπει να σημειωθεί ότι το οπτικό τμήμα της βάσης δεν περιγράφεται καθώς δεν χρησιμοποιήθηκε στο πλαίσιο της εργασίας.

- ***Κεφάλαιο 4***: το πιο εκτενές και αναλυτικό κεφάλαιο της διπλωαμτικής αφού περιγράφονται τα εργαλεία που χρησιμοποιήθηκαν για την διεκπεραίωση της διπλωματικής. Το Kaldi και το HTK λοιπόν αποτελούν δύο παρόμοια εργαλεία, εξειδικευμένα στην ηχητική και φωνητική αναγνώριση. Η χρήση και των δύο εργαλείων βασίζεται σε script αρχεία που διευκολύνουν τη διαδικασία της εκπαίδευσης και κατηγοριοποίησης της βάσης δεδομένων. Έπειτα, παραθέτονται αναλυτικές πληροφορίες για το εργαλείο SoX με το οποίο έγινε η επεξεργασία των ακουστικών αρχείων σε μία κοινή μορφή. Στο τέλος του κεφαλαίου, περιγράφεται η διαδικασίας της εξαγωγής χαρακτηριστικών, της εκπαίδευσης και αποκωδικοποίησης των δεδομένων με χρήση HMM και βαθιών νευρωνικών δικτύων.

- ***Κεφάλαιο 5***: εδώ παρουσιάζονται σε πίνακες τα αποτελέσματα που εξήχθησαν από το Kaldi και το HTK. Υπάρχει πλήρης σχολιασμός των αποτελεσμάτων καθώς και γραφήματα για την ευκολότερη κατανόησή τους.

- ***Κεφάλαιο 6***: στο τελευταίο κεφάλαιο πραγματοποιείται συζήτηση σχετικά με τα αποτελέσματα του Κεφαλαίου 5 και προτείνονται μελλοντικοί οδοί βελτίωσης των αποτελεσμάτων μέσα από την εξαγωγή διαφορετικών χαρακτηριστικών ή συνδυασμό καναλιών.

Κεφάλαιο 2

Εξαγωγή Χαρακτηριστικών - Ταξινομητές

Η επιλογή των κατάλληλων χαρακτηριστικών που αντιπροσωπεύουν τα δεδομένα μας, μπορεί να χαρακτηριστεί ως μία δύσκολη αλλά σημαντική διαδικασία για την κατασκευή ενός αποδοτικού ταξινομητή στην αναγνώριση ακουστικών γεγονότων. Καθώς τα δεδομένα μας έχουν την μορφή ακουστικών αρχείων, τα χαρακτηριστικά που χρησιμοποιήθηκαν για την εκπαίδευση HMM και DNN ταξινομητών είναι οι φασματικοί συντελεστές της κλίμακας Mel ή αλλιώς MFCC για συντομογραφία. Μετά από την τμηματοποίηση των αρχείων της βάσης, ακολουθήθηκε μία αυστηρή διαδικασία για την εξαγωγή των διανυσμάτων χαρακτηριστικών. Θα αποτελούσε σοβαρή παράλειψη να μην αναφερθεί ότι τα χαρακτηριστικά MFCC εξάγονται σταδιακά από ένα ηχητικό σήμα με τη χρήση ενός κυλιόμενου παραθύρου. Περισσότερες πληροφορίες σχετικά με αυτή τη διαδικασία, θα παρουσιαστούν στα ακόλουθα υποκεφάλαια. Ύστερα από τη μελέτη του τύπου χαρακτηριστικών, ακολουθεί εκτενής ανάλυση των δύο τύπων ταξινομητών που χρησιμοποιήθηκαν στο πλαίσιο της διπλωματικής, των κρυφών μοντέλων Markov και των βαθιών νευρωνικών δικτύων καθώς και την προσαρμογή του προβλήματός μας σε αυτούς.

2.1 Φασματικοί συντελεστές κλίμακας Mel

Το ανθρώπινο αυτί αντιλαμβάνεται πιο εύκολα τις αλλαγές σε έναν ήχο στις χαμηλές συχνότητες παρά στις υψηλές. Η κλίμακα Mel σχεδιάστηκε ώστε τα διανύσματα χαρακτηριστικών να αντιπροσωπεύουν αυτή τη δυνατότητα. Στην αναγνώριση ακουστικών γεγονότων το επιθυμητό αποτέλεσμα είναι ο εντοπισμός και η αναγνώριση των γεγονότων που μπορεί να ακούσει ο άνθρωπος, αποκλείοντας άλλες συχνότητες. Η μετατροπή της συχνότητας από την κλίμακα Hertz στην κλίμακα Mel πραγματοποιείται με τον ακόλουθο μαθηματικό τύπο:

$$M(f) = 1125 \ln(1 + \frac{f}{700}) \tag{2.1}$$

Η αντίστροφη μετατροπή γίνεται ως εξής:

$$M^{-1}(m) = 700(\exp \frac{m}{1125} - 1) \tag{2.2}$$

όπου f και m είναι η συχνότητα στην κλίμακα των Hertz και Mel αντίστοιχα.

Οι συντελεστές MFCC προτάθηκαν τη δεκαετία του '80 από τους S. Davis και P. Mermelstein [10] και έκτοτε χρησιμοποιούνται σαν χαρακτηριστικά που περιγράφουν ηχητικά σήματα στην ηχητική και φωνητική αναγνώριση. Τα βήματα που πρέπει να ακολουθηθούν για τον υπολογισμό των συντελεστών Mel παρουσιάζονται παρακάτω και στο Σχήμα 2.1:

1. Ορίζεται ένα παράθυρο-πλαισίου (προκαθορισμένη τιμή τα 25ms). Το ηχητικό σήμα είναι ορισμένο στον χρόνο και μεταβάλλεται σε συγκεκριμένες χρονικές στιγμές. Σε πληθώρα ηχητικών σημάτων, τέτοιες αλλαγές συμβαίνουν κάθε ms. Έχοντας αυτό κατά νου, καθίσταται πλήρως κατανοητό πως τα διανύσματα χαρακτηριστικών θα εξαχθούν σε αυτά τα χρονικά διαστήματα (τις χρονικές περιόδους που το σήμα δεν αλλάζει). Έτσι λοιπόν το σήμα τμηματοποιείται σε 25ms πλαίσια. Το πλαίσιο που ορίζουμε πρέπει να έχει χρονική διάστημα 20-40ms, αφού μικρότερες τιμές από 20 θα μας παρέχουν ανεπαρκή πληροφορία, ενώ τιμές μεγαλύτερες του 40 θα αντιπροσωπεύουν πολλαπλές αλλαγές στο σήμα.

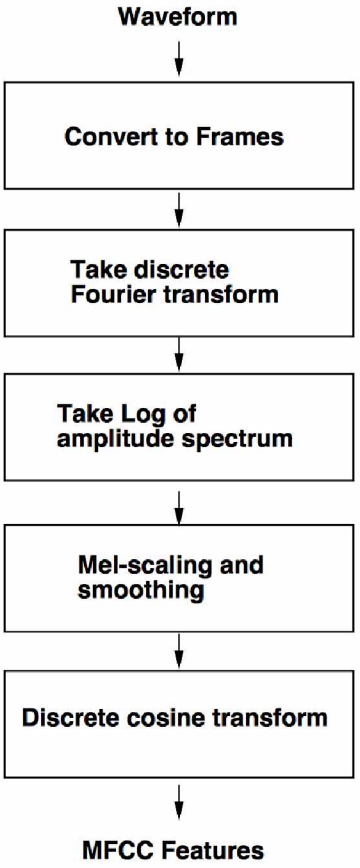
2. Για κάθε πλαίσιο υπολογίζεται ο διακριτός μετασχηματισμός Fourier (discrete Fourier transform ή DFT).

3. Εφαρμόζεται το φίλτρο Mel στο φάσμα ισχύος και αθροίζεται η ενέργεια σε κάθε πλαίσιο. Συγκεκριμένα, το πρώτο φίλτρο υποδεικνύει την ενέργεια κοντά σε μηδενικές συχνότητες. Η κλίμακα Mel καθορίζει πόσο ευρύ θα είναι κάθε φίλτρο για να υπολογιστεί η ενέργεια.

4. Υπολογίζεται ο λογάριθμος όλων των ενεργειών με σκοπό τα χαρακτηριστικά που θα εξαχθούν να προσεγγίζουν τι ακούει το ανθρώπινο αυτί.

5. Εφαρμόζεται ο διακριτός μετασχηματισμός συνημιτόνου (discrete cosine transform ή DCT) των προηγούμενων αποτελεσμάτων. Αυτή η ενέργεια αποσκοπεί στο να εξαλήψει της επικαλύψεις-συσχετίσεις των φίλτρων. Ο DCT χρησιμοποιείται σαν μέσο συμπίεσης της πληροφορίας του σήματος στους συντελεστές Mel και αποσυσχετίζει τις ενέργειες, πράγμα που επιτρέπει την κατασκευή διαγώνιων πινάκων συνδιασποράς (χρήσιμο στην εκπαίδευση HMM) [11].

6. Σαν διάνυσμα χαρακτηριστικών κρατάμε τους συντελεστές 1 έως 13 και απορρίπτονται οι υπόλοιποι. Αυτό συμβαίνει γιατί υψηλότεροι συντελεστές DCT συνεπάγονται γρήγορες αλλαγές στην ενέργεια του σήματος, άρα και του ίδιου ηχητικού σήματος, πράγμα που υποβαθμίζει την απόδοση συστημάτων φωνητικής και ηχητικής αναγνώρισης.



Σχημα 2.1: Ροή εργασίας για την εξαγωγή των χαρακτηριστικών. Λήφθηκε από το [3].

2.2 Τεχνικές μάθησης

Οι τρεις βασικές κατηγορίες εκμάθησης στον κλάδο της αναγνώρισης προτύπων είναι οι ακόλουθες:

- Η **μάθηση με επίβλεψη** αποτελεί την πιο απλή τεχνική μάθησης μηχανής. Στην ουσία, οι ταξινομητές σχεδιάζονται αξιοποιώντας μία εκ των προτέρων πληροφορία που μπορεί να είναι οι ετικέτες στα δεδομένα (labelled data).
- Στην άλλη περίπτωση προβλήματος αναγνώρισης προτύπων, οι κλάσεις δεν είναι γνωστές εκ των προτέρων, αλλά παρέχονται ομοιότητες μεταξύ των δεδομένων. Στόχος

είναι να ομαδοποιηθούν κατάλληλα τα δεδομένα με την ανάδειξη αυτών των ομοιοτήτων [12]. Αυτή η διαδικασία ονομάζεται **μάθηση χωρίς επίβλεψη**.

Το δικό μας πρόβλημα κατηγοριοποιείται στη μέθοδο της μάθησης με επίβλεψη, καθώς η βάση δεδομένων εμπεριέχει αρχεία ετικετών που βοηθούν τον ταξινομητή στην εκπαίδευσή του. Στη συνέχεια ακολουθεί περιγραφή της θεωρίας πίσω από τα μοντέλα που χρησιμοποιήθηκαν για τους ταξινομητές.

2.3 Γκαουσιανά μοντέλα μίξης

Τα μοντέλα αυτά αποτελούν μία κυρίαρχη προσέγγιση για την αναγνώριση ομιλητών όταν υπάρχει ανεξαρτησία από τον γραπτό λόγο [13]. Γι' αυτό, η χρήση τους είναι ευρύτερη σε προβλήματα ηχητικής αναγνώρισης σαν πιθανοτικά μοντέλα πολυμεταβλητών κατανομών που προσπαθούν να περιγράψουν άλλες αυθαίρετες. Μία μίξη Γκαουσιανών ορίζεται σαν ένας κυρτός συνδυασμός κατανομών πυκνότητας πιθανανότητας. Πιο ειδικά, κατανομή τέτοιου είδους στον $d-$διάστατο χώρο που χαρακτηρίζεται από τη μέση τιμή του χώρου $m \in \mathbb{R}^{\mathrm{d}}$ και τον $d \times d$ πίνακα συνδιασποράς C, ορίζεται ως εξής [13]:

$$\varphi(x;\theta) = (2\pi)^{-\frac{d}{2}} det(C)^{-\frac{1}{2}} exp(\frac{-(x-m)^T C^{-1}(x-m)}{2}) \tag{2.3}$$

όπου το θ υποδηλώνει τις παραμέτρους m και C.

Τα παραπάνω μοντέλα εκπαιδεύονται με τη χρήση του αλγορίθμου Αναμενόμενης Τιμής-Μεγιστοποίησης (Expectation Maximization-EM). Ο συγκεκριμένος αλγόριθμος χρησιμοποιείται κυρίως όταν τα δεδομένα είναι ελλιπή. Ακολουθεί περιγραφή του αλγορίθμου.

Ας συμβολίσουμε με y το σύνολο των παρατηρήσεων και την αντίστοιχη συνάρτηση πυκνότητας πιθανότητας (σ.π.π) $p(y|\theta)$, όπου θ είναι το σύνολο των παραμέτρων. Επίσης, ορίζουμε ένα σύνολο τυχαίων συνεχών μεταβλητών X με σ.π.π $p(x|\theta)$. Στην πραγματικότητα δεν μπορούμε να παρατηρήσουμε απευθείας τις μεταβλητές X, αλλά έχουμε μόνο την αντίληψη y της $Y = T(X)$ όπου T είναι μία συνάρτηση αντιστοίχησης του συνόλου X στη μέση τιμή του. Γνωρίζοντας λοιπόν μόνο τις παρατηρήσεις y, προσπαθούμε να εκτιμήσουμε τη μέγιστη πιθανοφάνεια των παραμέτρων (MLE) θ [14]:

$$\hat{\theta}_{MLE} = arg \max_{\theta \in \Theta} p(y|\theta). \tag{2.4}$$

Στην πράξη υπολογίζεται πιο εύκολα το σύνολο θ που μεγιστοποιεί τη λογαριθμική πιθανότητα του y,

$$\hat{\theta}_{MLE} = arg\max_{\theta\in\Theta} \log p(y|\theta). \tag{2.5}$$

Προφανώς, οι (2.5) και (2.6) δίνουν την ίδια λύση. Παρόλα αυτά, σε πολλά προβλήματα καθίσταται δύσκολος ο υπολογισμός των παραπάνω παραστάσεων. Έτσι ακολουθούνται τα παρακάτω βήματα του EM [15]:

1. Οι παράμετροι του συνόλου θ αρχικοποιούνται.
2. Με βάση αυτή την υπόθεση και τις παρατηρήσεις y, υπολογίζεται η υπο-συνθήκη πιθανότητα, $p(x|y)$. Δηλαδή, ολόκληρα τα δεδομένα να είναι x.
3. Απορρίπτεται η υπόθεση του βήματος 1, αλλά διατηρείται η πιθανότητα του βήματος 2.
4. Πρέπει να μεγιστοποιηθεί η πιθανότητα $\log p(x|\theta)$. Όμως το μόνο γνωστό είναι η πιθανότητα $p(x|y)$. Έτσι υπολογίζεται η αναμενόμενη πιθανότητα:

$$E[\log p(x|\theta)] = \int_{X(y)} \log p(x|\theta) p(x|y) dx \tag{2.6}$$

5. Επανυπολογίζεται το σύνολο θ ώστε να μεγιστοποιεί την πιθανότητα του βήματος 4.
6. Ο αλγόριθμος επιστρέφει στο βήμα 2.

2.4 Κρυφά Μοντέλα Markov

Τα κρυφά μοντέλα Markov αποτελούν ένα εργαλείο για την μοντελοποίηση δεδομένων που έχουν υπόσταση στον χρόνο. Τα HMM χρησιμοποιούνται κυρίως στην ηχητική και φωνητική αναγνώριση και καθιστούν την αναπαράσταση της κατανομής πιθανότητας σε μία ακολουθία παρατηρήσεων. Θεωρώντας ότι μία παρατήρηση συμβολίζεται με Y και η κατάσταση με S, τότε η παρατήρηση και η κατάσταση τη χρονική στιγμή t μπορεί να συμβολιστεί Y_t και S_t αντίστοιχα. Τα κρυφά μοντέλα Markov διατηρούν τις ιδιότητες των μοντέλων Markov, αλλά η λέξη-κλειδί *κρυφά* αναφέρεται σε δύο βασικές ιδιότητες:

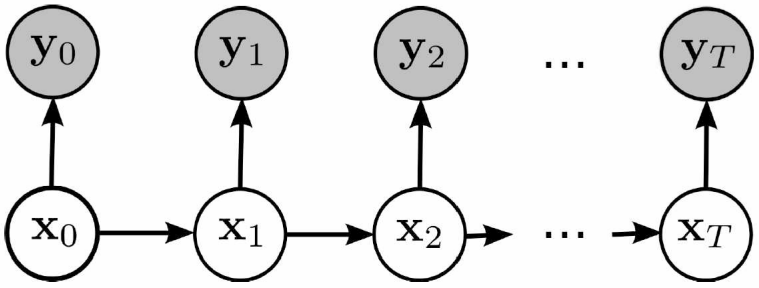
1. Μία παρατήρηση τη χρονική στιγμή t παρήχθη στην κατάσταση S_t που είναι κρυφή από τον παρατηρητή.

2. Η κάθε κατάσταση ακολουθεί τις ιδιότητες Markov, το οποίο σημαίνει πως οι τιμές των μεταβλητών στην κατάσταση S_t εξαρτώνται μόνο από την κατάσταση S_{t-1} και όχι από τις $t-2$ προηγούμενες.

Ο τύπος για την από κοινού κατανομή πιθανότητας της ακολουθίας των καταστάσεων είναι ο εξής:

$$P(S_{1:T}, Y_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_T|S_T) \tag{2.7}$$

Το μοντέλο πιθανότητας HMM παρουσιάζεται στο Σχήμα 2.2



Σχημα 2.2: Παράδειγμα κρυφού μοντέλου Markov. Λήφθηκε από το [4].

2.5 Perceptron

Το perceptron θεμελιώθηκε τη δεκαετία του '60 από τον Frank Rosenblatt και συνιστά την πιο απλή μορφή ενός νευρωνικού δικτύου [16]. Έχοντας κατά νου το Σχήμα 1.2. το perceptron έχει σαν είσοδο ένα σύνολο από δυαδικές τιμές (κάθε κόμβος εισαγωγής του Σχήματος 1.1 αποτελεί μία δυαδική τιμή). Καθώς λοιπόν η είσοδος είναι σε δυαδική μορφή, αναμένεται να συμβαίνει το ίδιο και με την έξοδο. Ας υποθέσουμε ότι η είσοδος αποτελείται από τις μεταβλητές $x_1, x_2, ..., x_n$, που είναι συνδεδεμένες κατ' ευθείαν με τον νευρώνα εξόδου (μηδενικά κρυφά επίπεδα). Οι συνδέσεις αυτές χαρακτηρίζονται από συγκεκριμένα βάρη $w_1, w_2, ..., w_n$. Πρακτικά, αυτό συμβαίνει επειδή κάποια είσοδος μπορεί να επηρεάζει πιο πολύ την έξοδο σε σχέση με τις άλλες. Η έξοδος δηλαδή θα δίνεται από τον ακόλουθο τύπο:

$$f(x) = \begin{cases} 0, & \sum_j w_j x_j \leq 0 \\ 1, & \sum_j w_j x_j > 0 \end{cases} \tag{2.8}$$

Επιπρόσθετα, κάθε κρυφό επίπεδο περιγράφεται από μία ειδική τιμή που ονομάζεται bias. Για την περίπτωση του παραδείγματος που διατυπώθηκε παραπάνω υπάρχει μία και μοναδική τιμή

bias ή b για συντομογραφία. Αν αυτή η τιμή είναι πολύ αρνητική, είναι δύκολο να υπάρχει έξοδος 1. Το αντίστροφο αν ήταν πολύ θετική. Γενικά, ως bias ορίζεται η μετρική ευκολίας ενεργοποίησης ενός νευρώνα. Έχοντας ως δεδομένα τα παραπάνω, η έξοδος διαμορφώνεται ως εξής:

$$f(x) = \begin{cases} 0, & \sum_j (w_j x_j) + b \leq 0 \\ 1, & \sum_j (w_j x_j) + b > 0 \end{cases} \tag{2.9}$$

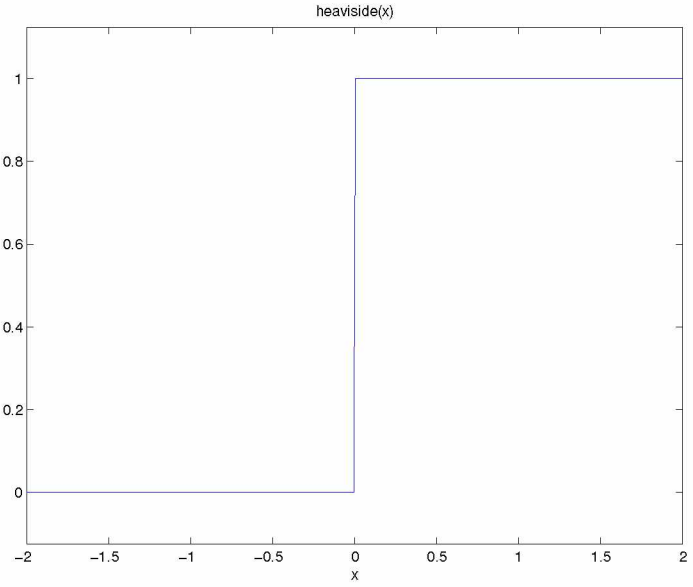
2.6 Συνάρτηση ενεργοποίησης

Στις παραπάνω περιπτώσεις τα κρυφά επίπεδα δεν ήταν παρόντα. Αν σε ένα νευρωνικό δίκτυο υπάρχουν κρυφά επίπεδα μεταξύ εισόδου και εξόδου, κάθε νευρώνας του $i-$στου κρυφού επιπέδου θα είναι 0 ή 1, το οποίο σημαίνει ότι κάλλιστα η συνάρτηση βήματος μπορεί να χρησιμοποιηθεί για αυτή την εναλλαγή τιμών. Η συμπεριφορά της συνάρτησης βήματος παρουσιάζεται στο Σχήμα 2.3. Για να αποφευχθεί όμως αυτή η απότομη μετάβαση από το 0 στο 1, χρησιμοποιήθηκε η σιγμοειδής συνάρτηση που παρουσιάζεται στο Σχήμα 2.4. Οι μαθηματικοί τύποι που περιγράφουν τις τρεις βασικές συναρτήσεις ενεργοποίησης είναι:

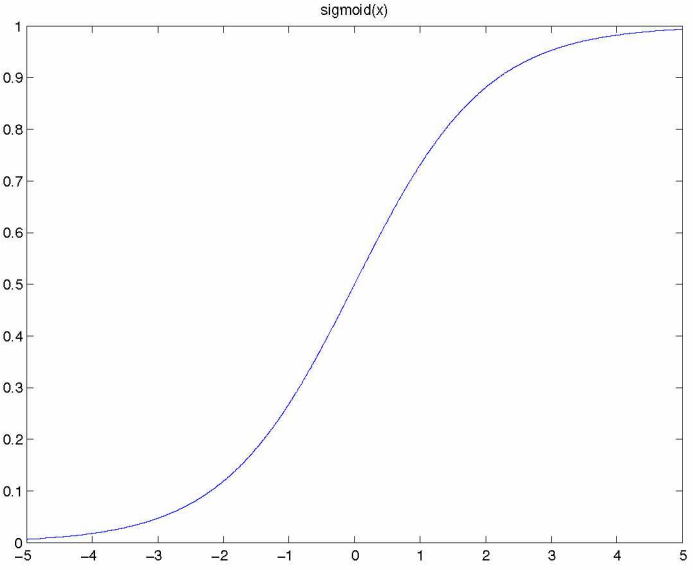
$$step(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases} \tag{2.10}$$

$$sigm(x) = \frac{1}{1 + e^{-\alpha x}} \tag{2.11}$$

$$tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{2.12}$$

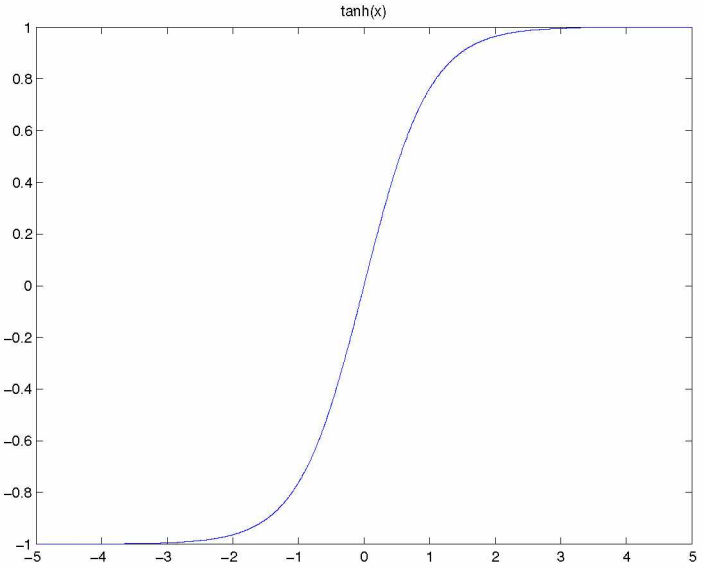


Σχημα 2.3: Συνάρτηση βήματος



Σχημα 2.4: Σιγμοειδής συνάρτηση

Σε ένα δηλαδή νευρωνικό δίκτυο με δύο κρυφά επίπεδα, η είσοδος στους νευρώνες του 2ου επιπέδου θα έχουν τιμές που ανήκουν στο διάστημα $[0, 1]$. Το Kaldi χρησιμοποιεί σαν συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη (hyperbolic tangent function ή tanh) που μετατοπίζει την έξοδο στο $[-1, 1]$.



ΣΧΗΜΑ 2.5: Συνάρτηση υπερβολικής εφαπτομένης

2.7 Αρχιτεκτονικές βαθιάς μάθησης

Κατά την υλοποίηση αυτής τη διπλωματικής, πειραματιστήκαμε με την αρχιτεκτονική της βαθιάς μάθησης. Ένα βαθύ νευρωνικό δίκτυο αποτελεί ένα τεχνητό νευρωνικό δίκτυο με πολλαπλά κρυφά επίπεδα μεταξύ της εισόδου και της εξόδου. Τα βαθιά νευρωνικά δίκτυα ή DNN εκπαιδεύονται με τη χρήση του αλγορίθμου οπισθοδιάδοσης. Ύστερα από κάθε επανάληψη, τα βάρη ανανεώνονται μέσα από τον ακόλουθο τύπο:

$$w(t+1) = w(t) + \eta \Delta w \tag{2.13}$$

όπου η είναι ο ρυθμός εκπαίδευσης, $w(t)$ η τρέχουσα εκτίμηση των βαρών και Δw η αντίστοιχη διόρθωση ώστε να προκύψει η επόμενη εκτίμηση $w(t+1)$. Η επιλογή της συνάρτησης κόστους έγκειται αποκλειστικά στον τύπο των δεδομένων (με επίβλεψη, χωρίς επίβλεψη) και στην συνάρτηση ενεργοποίησης (βηματική, σιγμοειδής). Για παράδειγμα, στο Kaldi χρησιμοποιήθηκε η συνάρτηση διεντροπίας (cross-entropy) σαν συνάρτηση κόστους:

$$J = -\sum_{i=1}^{N}\sum_{k=1}^{k_L} y_k(i) \ln \frac{\hat{y}_k(i)}{y_k(i)} \tag{2.14}$$

και σαν συνάρτηση εξόδου του DNN η softmax:

$$\hat{y}_k = \frac{\exp(v_k^L)}{\sum_k \exp(v_{k'}^L)} \tag{2.15}$$

όπου N είναι το πλήθος των διανυσμάτων εκπαίδευσης, L το πλήθος των κρυφών επιπέδων και v η συνάρτηση ενεργοποίησης. Τέλος, οι μεταβλητές y_k και $\hat{y}_k$ δηλώνουν την επιθυμητή και πραγματική αντίστοιχα εξόδο του νευρωνικού δικτύου.

2.8 Προβλήματα με τα βαθιά νευρωνικά δίκτυα

Τα βαθιά νευρωνικά δίκτυα χαρακτηρίζονται από δύο βασικά προβλήματα. Το πρώτο είναι αυτό της υπερ-εκπαίδευσης και το δεύτερο είναι ο χρόνος υπολογισμού τους. Κατά τη διάρκεια εκπαίδευσης ενός DNN με το Kaldi, ακόμα και με τετραπύρηνο επεξεργαστή, ο χρόνος ολοκλήρωσης ήταν 15-25 λεπτά. Μεγάλες βελτιώσεις έχουν γίνει όμως με τη χρήση της τεχνολογίας CUDA στις κάρτες γραφικών. Παρόλο που και τα δύο εργαλεία έχουν υλοποιημένες τις συναρτήσεις τους σε τέτοια τεχνολογία, για αυτή την εργασία δεν χρησιμοποιήθηκε. Σε αυτό το σημείο πρέπει να αναφερθεί πως η καθυστέρηση της διάτρεξης των νευρωνικών δικτύων δεν οφείλεται στο μέγεθος της βάσης δεδομένων, αλλά στις επιπλέον επαναλήψεις που λαμβάνουν χώρα κατά τη διαδικασία της εκπαίδευσης, δημιουργώντας έτσι ανεπιθύμητες εξαρτήσεις.

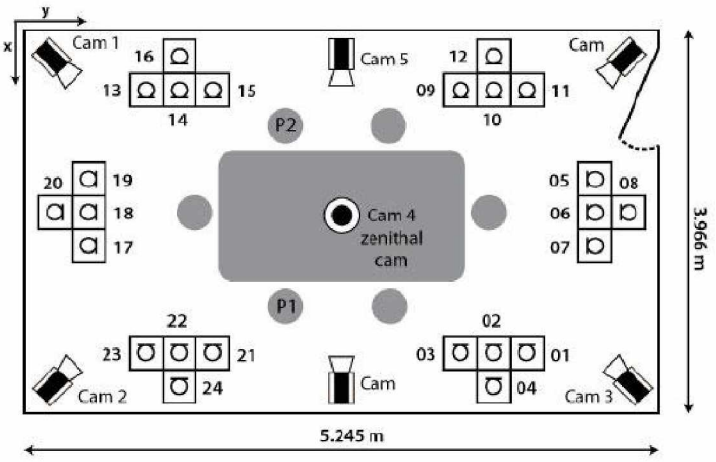
Ο αλγόριθμος οπισθοδιάδοσης σε συνδυασμό με αυτόν της απότομης κατάδυσης [12] (gradient descent) αποτελεί τον προτεινόμενο τρόπο εκπαίδευσης των νευρωνικών δικτύων, αν και το κόστος όπως αναφέρθηκε είναι ένα μειονέκτημα. Ωστόσο, τα βαθιά νευρωνικά δίκτυα τείνουν να είναι πιο αποδοτικά χάρη στην λιγότερη ανάγκη για παραμέτρους και υπολογιστικά στοιχεία, χωρίς όμως αυτό να σημαίνει ότι αποτελούν πάντοτε καλύτερη λύση από τις ρηχές μεθόδους. Η εκπαίδευση ενός βαθιού νευρωνικού δικτύου είναι γνωστό ότι αποτελεί μία δύσκολη διαδικασία. Κατά τον προκαθορισμένο τρόπο εκπαίδευσης τα βάρη του δικτύου αρχικοποιούνται με τυχαίες τιμές και εφαρμόζεται ο αλγόριθμος απότομης κατάδυσης, πράγμα που δίνει φτωχές λύσεις για νευρωνικά δίκτυα με 3 ή περισσότερα κρυφά επίπεδα. Για αυτό τον λόγο, τα νευρωνικά δίκτυα περιορίζονται στο ένα ή δύο κρυφά επίπεδα [17]. Η δήλωση αυτή θα γίνει πιο κατανοητή κατά την διαδικασία αξιολόγησης του ταξινομητή στο Κεφάλαιο 5 όπου παρουσιάζονται αναλυτικά τα αποτελέσματα.

Κεφάλαιο 3

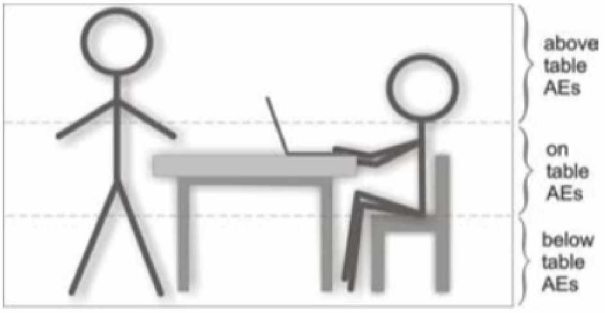
Η Πολυ-καναλική Βάση Δεδομένων UPC-TALP

3.1 Βασικές Πληροφορίες

Η συγκεκριμένη βάση δεδομένων καταγράφηκε κατά τη διάρκεια του CHIL Project (Computers in the Human Interaction Loop) στο πλαίσιο του ολοκληρωμένου προγράμματος εργασίας European Commission's Sixth Framework και στην ουσία περιέχει μία σειρά από ακουστικά αρχεία με γεγονότα που μπορούν να συμβούν σε ένα δωμάτιο συνεδριάσεων [18]. Τα ηχητικά αρχεία της βάσης ηχογραφήθηκαν από 24 συνολικά μικρόφωνα, τα οποία ήταν συγχρονισμένα και οργανωμένα σε ομάδες. Μέσα στην αίθουσα συνεδριάσεων υπήρχαν 6 τέτοιες T-σχήματος ομάδες μικροφώνων, όπως φαίνεται στο Σχήμα 3.1 και 3.2 όπου παρουσιάζεται η κάτοψη του δωματίου που έγιναν οι ηχογραφήσεις και ο τύπος των δεδομένων που προσπαθήσαμε να ταξινομήσουμε αντίστοιχα. Η επιπρόσθετη πληροφορία πάνω στο σχήμα είναι η θέση των δύο συμμετόχοντων P1 και P2. Μεταξύ των διαδοχικών γεγονότων υπήρχε παύση μερικών δευτερολέπτων, ώστε ο ταξινομητής που θα εκπαιδευτεί να αναγνωρίζει και την απουσία κάποιου ηχητικού γεγονότος (ησυχία). Εκτός της ηχητικής κάλυψης, το δωμάτιο ήταν εξοπλισμένο με έξι κάμερες που κατέγραφαν την όλη διαδικασία. Η συγκεκριμένη βάση δεδομένων μπορεί να χρησιμοποιηθεί σαν υλικό εκπαίδευσης για τεχνολογίες αναγνώρισης ακουστικών γεγονότων, όπως και για αλγορίθμους αποκωδικοποίησης, οι οποίοι δοκιμάζονται σε ήσυχα περιβάλλοντα χωρίς μόνιμες επικαλύψεις μεταξύ των ήχων.



Σχημα 3.1: Κάτοψη του UPC-δωματίου. Παρουσιάζονται οι θέσεις και η κατανομή των 24 μικροφώνων. Η εικόνα λήφθηκε από το [5].



Σχημα 3.2: Διαχωρισμός των δεδομένων ανάλογα με τον χώρο του δωματίου. Η εικόνα λήφθηκε από το [5].

3.2 Τεχνικές Πληροφορίες

Όπως προαναφέρθηκε, η βάση δεδομένων UPC-TALP περιέχει μεμονωμένα και αυθόρμητα ηχητικά συμβάντα σε ένα έξυπνο περιβάλλον. Η δομή της αποθηκευμένης βάσης παρουσιάζεται παρακάτω:

- 1^{st} DVD: S01, S02
- 2^{nd} DVD: S03, S04
- 3^{rd} DVD: S05, S06
- 4^{th} DVD: S07, T02, T03, T04
- 5^{th} DVD: S08, T01
- 6^{th} DVD: T05, T06, T07, T08, T09

Τα ακουστικά γεγονότα σε κάθε συνεδρία (S01-S08) παρήχθηκαν από έξι διαφορετικά άτομα. Σε κάθε προσπάθειά τους, στόχος ήταν να αναπαράγουν μία συγκεκριμένη ακολουθία από ήχους που θα καταγράφονταν από τα 24 μικρόφωνα. Πιο ειδικά, η συχνότητα στην οποία πραγματοποιήθηκαν οι ηχογραφήσεις ήταν τα 44.1kHz και τα αρχεία αποθηκεύτηκαν σε μορφή *.wv. Στη βάση δεδομένων έχουν κατηγοριοποιηθεί 12 κλάσεις ηχητικών γεγονότων που μπορούν να συμβούν ρεαλιστικά σε ένα δωμάτιο συνεδριάσεων. Από αυτά, το χειροκρότημα και το γέλιο συνέβησαν όταν στο δωμάτιο ήταν παραπάνω από ένας συμμετέχοντες. Το πλήθος των γεγονότων όπως και η κατανομή τους κατά το πέρας της καταγραφής της βάσης δεδομένων παρουσιάζεται στον Πίνακα 3.1.

Ηχητικό Γεγονός	S01	S02	S03	S04	S05	S06	S07	S08
Χτύπημα Πόρτας	9	8	10	10	10	8	11	13
Κλείσιμο Πόρτας	17	15	19	20	40	37	56	52
Βήματα	10	10	8	23	43	34	28	50
Μετακίνηση Καρέκλας	19	37	32	22	23	38	34	40
Κουτάλι/Κουδούνισμα φλιτζανιού	10	11	13	11	10	15	11	15
Τύλιγμα Χαρτιού Εργασίας	9	11	10	8	17	12	12	12
Ήχος Κλειδιών	11	11	11	8	0	13	10	18
Ήχος Πληκτρολογίου	10	10	13	12	10	13	10	11
Ήχος Τηλεφώνου/Μουσική	11	18	11	14	8	11	13	15
Χειροκρότημα	9	5	9	11	12	9	14	14
Βήχας	10	10	12	13	9	13	11	12
Ομιλία	0	0	0	0	8	20	12	34

ΠΙΝΑΚΑΣ 3.1: Κατανομή των κλάσεων της βάσης δεδομένων UPC-TALP ανά συνεδρία.

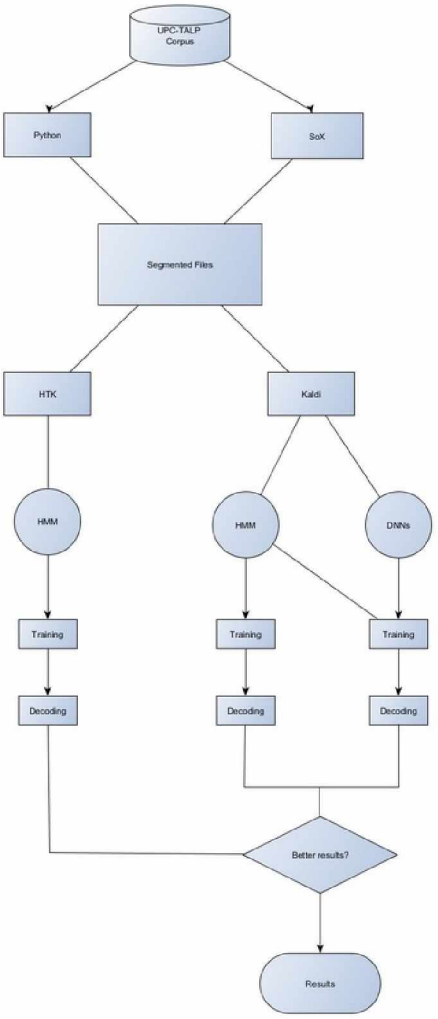
Τα σύμβολα των κλάσεων (ηχητικών γεγονότων) μαζί με τη διάρκεια τους καταγράφονται ανά συνεδρία σε ένα *.csv αρχείο.

Κεφάλαιο 4

Πειραματικά Εργαλεία

4.1 Ροή εργασίας

Η ροή εργασίας που ακολουθήθηκε παρουσιάζεται γραφικά παρακάτω:



Σχημα 4.1: Ροή εργασίας

4.2 HTK

Το HTK (εργαλειοθήκη κρυφών μοντέλων Markov) αναπτύχθηκε στο εργαστήριο Μηχανικής Μάθησης του Τμήματος Μηχανικών στο Πανεπιστήμιο του Cambridge με σκοπό τη δημιουργία και τη διαχείριση των μοντέλων Markov [19]. Το σύνολο του εργαλείου έχει υλοποιηθεί σε γλώσσα C, άρα και τα εκτελέσιμα αρχεία και οι βιβλιοθήκες του. Τα προαναφερθέντα είναι χωρισμένα σε φακέλους για την ευκολότερη εύρεση από τον χρήστη. Σχετικά με την εγκατάσταση, αρκεί η μεταγλώττιση και παραγωγή του εκτελέσιμου κώδικα C (περισσότερες πληροφορίες εδώ http://htk.eng.cam.ac.uk). Παρόλο που το HTK εξειδικεύεται σε έρευνες σχετικά με την φωνητική αναγνώριση, έχει πολυάριθμες εφαρμογές συμπεριλαμβανομένων της φωνητικής σύνθεσης και αναγνώρισης γραμματικών χαρακτήρων. Μέσα από το HTK, ο χρήστης δημιουργεί HMM μοντέλα, επεξεργάζεται τις ιδιότητες των Γκαουσιανών κατανομών, τα εκπαιδεύει με δικές του παραμετρικοποιήσεις και τέλος χρησιμοποιεί τον αλγόριθμο Viterbi για να αναγνωρίσει τα ηχητικά γεγονότα.

4.2.1 Προετοιμασία δεδομένων - Εξαγωγή MFCCs

Σε αυτό το στάδιο του σχεδίου εργασίας τα ακουστικά αρχεία της βάσης δεδομένων έχουν επεξεργαστεί και χωριστεί σε ξεχωριστούς φακέλους με τη βοήθεια του SoX ανάλογα με τη συνεδρία στην οποία ανήκουν. Έτσι δημιουργήθηκε ο φάκελος **train** με δύο υπο-φακέλους **mfcc** και **wav** που περιέχουν τα διανύσματα χαρακτηριστικών και τα ακουστικά αρχεία αντίστοιχα. Αρχικά, τα MFCC χαρακτηριστικά εξήχθησαν από κάθε ακουστικό αρχείο και τοποθετήθηκαν (αυτόματα πάντα) στο κατάλληλο φάκελο. Γι' αυτό τον λόγο δημιουργήθηκε ένα αρχείο ρυθμίσεων με το όνομα **config**. Το περιεχόμενο αυτού παρουσιάζεται παρακάτω στο Σχήμα 4.2.

```
SOURCEFORMAT = WAV
TARGETKIND = MFCC_0
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
```

Σχημα 4.2: Αρχείο ρυθμίσεων

Το παραπάνω περιεχόμενο ουσιαστικά περιγράφει τις ιδιότητες των χαρακτηριστικών με τις οποίες θα εξαχθούν τα χαρακτηριστικά MFCC. Κρατώντας τις πιο σημαντικές από αυτές, έχουμε:

- **SOURCEFORMAT**, δηλώνει τον τύπο του ακουστικού αρχείου εισόδου, WAV στην περίπτωσή μας.
- **TARGETKIND MFCC_0**, εξάγει τα διανύσματα MFCC διάστασης 13 αντικαθιστώντας τον τελευταίο με τον μηδενικό συντελεστή που αντιστοιχεί στην ενέργεια του κάθε παραθύρου. Η παράμετρος αυτή απαιτεί μία σειμβολοσειρά που προσδιορίζει αυτή την ιδιότητα. Με σκοπό τη βελτίωση του συνόλου εκπαίδευσης χρησιμοποιήθηκαν και οι 39 συντελεστές MFCC. Έτσι η παράμετρος TARGETKIND άλλαξε σε MFCC_0_D_A, το οποίο πρόσθεσε τους δελτα-συντελεστές και δέλτα-δέλτα-συντελεστές.
- **WINDOWSIZE**, το μέγεθος του πλαισίου με το οποίο σαρώνουμε το ηχητικό σήμα. Η τιμή του είναι προκαθορισμένη στα 25ms και απαιτεί μία θετική πραγματική τιμή.
- **TARGETRATE**, δηλώνει κατά πόσο κινούμαστε πάνω στο σήμα. Για παράδειγμα, το αρχικό παράθυρο διαρκεί από τη χρονική στιγμή 0 μέχρι 25ms και το επόμενο από 10 έως 35ms (σχετικά με τις επικαλύψεις υπάρχει αναφορά στο υποκεφάλαιο 2.2).
- **NUMCHANS**, απαιτεί έναν ακέραιο αριθμό που δηλώνει τον αριθμό των filter-bank channels που χρησιμοποιούνται στην ανάλυση.
- **NUMCEPS**, αποτελεί τη μεταβλητή που υποχρεώνει την διατήρηση των 13 πρώτων συντελεστών για τον λόγο που υπόθηκε στο υποκεφάλαιο 2.2.

Το επόμενο βήμα ήταν η δημιουργία ενός script αρχείου που περιείχε τη θέση του συνόλου εκπαίδευσης και τον φάκελο που θα αποθηκεύονταν τα αρχεία με τα διανύσματα χαρακτηριστικών. Το αρχείο ονομάστηκε, σύμφωνα με το HTK, **codetrain.scp**.

```
../train/wav/s01/S01-eventid_0.wav ../train/mfcc/s01/S01-eventid_0.mfc
../train/wav/s01/S01-eventid_1.wav ../train/mfcc/s01/S01eventid_1.mfc
../train/wav/s01/S01-eventid_2.wav ../train/mfcc/s01/S01-eventid_2.mfc
../train/wav/s01/S01-eventid_3.wav ../train/mfcc/s01/S01-eventid_3.mfc
../train/wav/s01/S01-eventid_4.wav ../train/mfcc/s01/S01-eventid_4.mfc
etc
```

Σχημα 4.3: Παράδειγμα περιεχομένου του αρχείου codetrain.scp

Μετά τη δημιουργία των δύο αυτών βασικών αρχείων κλήθηκε η εντολή:

HCopy -A -D -T 1 -C config -S codetrain.scp

1. Η επιλογή -A έχει ως αποτέλεσμα την εκτύπωση των ορισμάτων της εντολής στο τερματικό. Καθώς τρέχουν τα πειράματα μέσω script αρχείων, η παράμετρος αυτή αποτελεί προτεινόμενη λύση, ώστε να ξέρει ο χρήστης να διατρέξει στο ημερολόγιο του εργαλείου και να ανακαλέσει ποιές εντολές έχει τρέξει.

2. Η επιλογή -C χρησιμοποιείται για να δηλώσει την ύπαρξη αρχείου ρυθμίσεων.

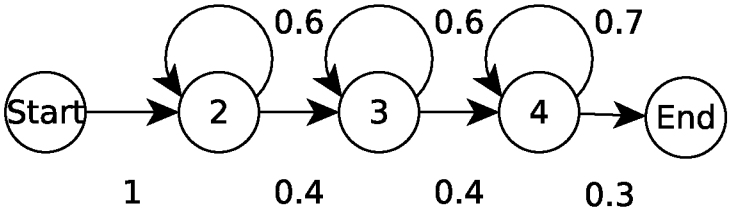
3. Η επιλογή -S χρησιμοποείται για να συγκεκριμενοποιήσει το όνομα του script αρχείου του Σχήματος 4.3.

4. Το -D απλά εκτυπώνει στο τερματικό τις επιλογές μας στο αρχείο ρυθμίσεων.

Μετά την εκτέλεση της παραπάνω εντολής δημιουργήθηκαν τα αρχεία **.mfc**, που περιέχουν τα διανύσματα χαρακτηριστικών. Πρέπει να επισημανθεί ότι σε κάθε ακουστικό αρχείο αντιστοιχεί ένα αρχείο .mfc με το ίδιο όνομα. Ύστερα από την εξαγωγή των χαρακτηριστικών, η παράμετρος SOURCEFORMAT διαγράφηκε, όπως επιτάζει το HTK, ώστε να προχωρήσουμε στην εκπαίδευση των HMM μοντέλων.

4.2.2 Ορισμός πρωτοτύπου

Το πρώτο βήμα στην δημιουργία των μοντέλων είναι να ορισθεί ένα πρωτότυπο με πέντε καταστάσεις. Ουσιαστικά η κάθε κλάση θα αποτελείται από πέντε HMM καταστάσεις εκ των οποίων οι 2, 3 και 4 θα εκπέμπουν πιθανότητες, ενώ οι 1 και 5 θεωρούνται τερματικές. Έτσι δημιουργήθηκε το αρχείο **prototype** που αποτέλεσε βάση ολόκληρου του HMM μοντέλου, το περιεχόμενο του οποίου φαίνεται στο Σχήμα 4.5. Επίσης, παραθέτουμε γραφική αναπαράσταση του αρχείου στο Σχήμα 4.4 για να γίνει πιο κατανοητός ο ορισμός του μοντέλου.



Σχημα 4.4: Πρωτότυπο μοντέλο HMM.

```
 o <VecSize> 13 <MFCC_0>
h "prototype"
<BeginHMM>
<NumStates> 5
<State> 2
<Mean> 13
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 13
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<State> 3
<Mean> 13
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 13
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<State> 4
<Mean> 13
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 13
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<TransP> 5
0.0 1.0 0.0 0.0 0.0
0.0 0.6 0.4 0.0 0.0
0.0 0.0 0.6 0.4 0.0
0.0 0.0 0.0 0.7 0.3
0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

Σχημα 4.5: Περιεχόμενο αρχείου prototype.

Εύκολα γίνεται αντιληπτό πως κάθε κατάσταση του Markov μοντέλου περιγράφεται από ένα Γκαουσιανό μοντέλο μίξης. Παρακάτω σε αυτή τη διπλωματική θα αποδειχθεί ότι η χρήση περισσότερων από ένα Γκαουσιανό μοντέλο μίξης δημιουργεί πιο αποδοτικούς ταξινομητές. Πριν ολοκληρωθεί η δημιουργία του μοντέλου πρώτου σταδίου, ορίστηκε ένα ακόμη αρχείο, το **trainHMM.scp** . που περιλαμβάνει τη θέση των χαρακτηριστικών στον σκληρό δίσκο.

```
../train/mfcc/s01/S01-eventid_0.mfc
../train/mfcc/s01/S01-eventid_1.mfc
../train/mfcc/s01/S01-eventid_2.mfc
../train/mfcc/s01/S01-eventid_3.mfc
../train/mfcc/s01/S01-eventid_4.mfc
../train/mfcc/s01/S01-eventid_5.mfc
etc
```

Σχημα 4.6: Περιεχόμενο αρχείου trainHMM.scp

Έπειτα, η παρακάτω εντολή χρησιμοποιήθηκε

HCompV -A -D -T 1 -C config -f 0.01 -m -S trainHMM.scp -M hmm0 prototype

δίνοντας στην έξοδο τα αρχεία:

- ***prototype***, το νέο πρωτότυπο μοντέλο, δημιουργημένο με βάση το σύνολο εκπαίδευσης
- ***vfloors***, που διατηρεί γενικές ρυθμίσεις για το μοντέλο Markov

Επιπρόσθετα, δημιουργήθηκε το αρχείο **hmmdefs** που περιέχει τον ορισμό του μοντέλου Markov για κάθε κλάση (το τελικό μοντέλο μας αποτελείται από 13 HMM, ένα για κάθε κλάση). Κάθε γεγονός κατέχει 3 καταστάσεις που εκπέμπουν και περιγράφεται, όπως προαναφέραμε, από ένα Γκαουσιανό μοντέλο μίξης. Γι' αυτό ακριβώς τον λόγο, υπάρχει ένα διάνυσμα μέσης τιμής όπως και ένα διάνυσμα των διαγωνίων στοιχείων του διαγώνιου πίνακα διασποράς. Όσον αφορά τον πίνακα μεταβάσεων, η πρώτη γραμμή έχει πιθανότητα 1.0 στο κελί (1,2) και 0.0 στη τελευταία. Αυτό συμβαίνει επειδή η κατάσταση 1 δεν δέχεται συνδέσεις και η 5 δεν δείχνει σε κάποια άλλη.

Παράδειγμα του αρχείου **hmmdefs** για το γεγονός *knock* παρουσιάζεται στο Σχήμα 4.7.

```
  h "kn"

<BEGINHMM>

<NUMSTATES> 5

<STATE> 2

<MEAN> 25

2.285674e-07 5.184949e-08 1.189473e-08 -2.638324e-08 3.905087e-08 -3.028702e-08 -6.018104e-08 -1.403002e-08 6.340363e-08

1.347327e-08 -8.192691e-08 -9.772542e-08 -7.355682e-03 -3.789886e-03 1.184963e-02 -7.089332e-03 -3.232978e-04 4.847161e-03

-5.158526e-03 6.903737e-04 9.217438e-03 -6.614167e-03 -1.175443e-03 1.114908e-02 4.575837e-03

<VARIANCE> 25

6.580434e+01 3.732679e+01 3.525515e+01 4.770429e+01 4.332327e+01 4.544641e+01 5.620689e+01 2.553866e+01 4.001572e+01 3.416671e+01

2.128212e+01 2.660224e+01 1.668585e+00 1.700366e+00 1.616409e+00 1.768895e+00 1.718035e+00 2.098122e+00 2.326025e+00 1.677738e+00

2.010739e+00 1.595870e+00 1.417548e+00 1.510511e+00 1.447709e+00

<GCONST> 9.662931e+01

<STATE> 3

<MEAN> 25

2.285674e-07 5.184949e-08 1.189473e-08 -2.638324e-08 3.905087e-08 -3.028702e-08 -6.018104e-08 -1.403002e-08 6.340363e-08

1.347327e-08 -8.192691e-08 -9.772542e-08 -7.355682e-03 -3.789886e-03 1.184963e-02 -7.089332e-03 -3.232978e-04 4.847161e-03

-5.158526e-03 6.903737e-04 9.217438e-03 -6.614167e-03 -1.175443e-03 1.114908e-02 4.575837e-03

<VARIANCE> 25

6.580434e+01 3.732679e+01 3.525515e+01 4.770429e+01 4.332327e+01 4.544641e+01 5.620689e+01 2.553866e+01 4.001572e+01 3.416671e+01

2.128212e+01 2.660224e+01 1.668585e+00 1.700366e+00 1.616409e+00 1.768895e+00 1.718035e+00 2.098122e+00 2.326025e+00 1.677738e+00

2.010739e+00 1.595870e+00 1.417548e+00 1.510511e+00 1.447709e+00

<GCONST> 9.662931e+01

<STATE> 4

<MEAN> 25

2.285674e-07 5.184949e-08 1.189473e-08 -2.638324e-08 3.905087e-08 -3.028702e-08 -6.018104e-08 -1.403002e-08 6.340363e-08

1.347327e-08 -8.192691e-08 -9.772542e-08 -7.355682e-03 -3.789886e-03 1.184963e-02 -7.089332e-03 -3.232978e-04 4.847161e-03

-5.158526e-03 6.903737e-04 9.217438e-03 -6.614167e-03 -1.175443e-03 1.114908e-02 4.575837e-03

<VARIANCE> 25

6.580434e+01 3.732679e+01 3.525515e+01 4.770429e+01 4.332327e+01 4.544641e+01 5.620689e+01 2.553866e+01 4.001572e+01 3.416671e+01

2.128212e+01 2.660224e+01 1.668585e+00 1.700366e+00 1.616409e+00 1.768895e+00 1.718035e+00 2.098122e+00 2.326025e+00 1.677738e+00

2.010739e+00 1.595870e+00 1.417548e+00 1.510511e+00 1.447709e+00

<GCONST> 9.662931e+01

<TRANSP> 5

0.000000e+00 1.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 6.000000e-01 4.000000e-01 0.000000e+00 0.000000e+00

0.000000e+00 0.000000e+00 6.000000e-01 4.000000e-01 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 7.000000e-01 3.000000e-01

0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00

<ENDHMM>
```

Σχημα 4.7: Τμήμα περιεχομένου αρχείου hmmdefs

4.2.3 Flat start monophones

Σε αυτό το τμήμα της προετοιμασίας των δεδομένων δημιουργήθηκε ένα αρχείο που περιέχει τις λέξεις της γραμματικής μας, που ισοδυναμούν με τα γεγονότα. Στο πρόβλημά μας δεν έχουμε λέξεις, αλλά με στόχο τη χρήση του HTK έπρεπε να μεταφραστούν με κάποιο τρόπο τα δεδομένα στη μορφή που ορίζεται στο API. Τα φωνήματα (flat start monophones) λοιπόν ορίζονται στο αρχείο **monophones0** με αλφαβητική σειρά. Κάθε γεγονός γράφτηκε με συντομογραφία για να αποφευχθούν περίπλοκες περιγραφές. Ιδιαίτερη προσοχή πρέπει να δοθεί στο γεγονός ότι η ***ησυχία*** αποτελεί ένα ξεχωριστό συμβάν. Στο σχήμα 4.8 παρουσιάζονται οι δηλώσεις των συμβάντων που λαμβάνουν μέρος στο πρόβλημά μας.

```
ap #applause
cm #chair moving
co #cough
ds #door slam
kt #keyboard typing
kj #keyjingle
kn #knock
pw #paperwork
pr #phone ringing
spe #speech
cl #spoon
si #silence
st #steps
```

Σχημα 4.8: Αρχείο monophones0

4.2.4 Αρχεία ετικετών

Πριν την εκπαίδευση των κρυφών μοντέλων Markov τοποθετήθηκαν αρχεία ετικετών στα ακουστικά αρχεία. Μέσα από τα αρχεία ετικετών δηλώνεται σε ποιο συμβάν αντιστοιχεί το αρχείου ήχου ή χαρακτηριστικών. Παραδείγματος χάριν, το τμήμα ενός αρχείου ετικετών έχει τη μορφή του Σχήματος 4.9.

```
#!MLF!# "*/S01-eventid_0.lab"
kn
.  "*/S01-eventid_1.lab"
si
.
"*/S01-eventid_2.lab"
ds
.
"*/S01-eventid_3.lab"
si
.
"*/S01-eventid_4.lab"
ds
.
"*/S01-eventid_5.lab"
si
```

Σχημα 4.9: Το αρχείο αποτελεί ένα Κύριο Αρχείο Ετικετών Master Label File (MLF).

4.2.5 Εκπαίδευση κρυφών μοντέλων Markov

4.2.5.1 Μοναδικό μοντέλο μίξης ανά κατάσταση

Τα μοντέλα Markov εκπαιδεύονται κατά κόρον με τη χρήση του αλγορίθμου EM (Baum-Welch). Ακολουθεί περιγραφή του αλγορίθου:

Σαν είσοδος του αλγορίθμου δίνεται ο πίνακας μεταβάσεων T, η αρχική κατανομή πιθανότητας των καταστάσεων π_i, η πιθανότητα μίας συγκεκριμένης παρατήρησης τη χρονική στιγμή t για την κατάσταση j, $b_j(y_t) = P(Y_t = y_t|X_t = j)$ και το άθροισμα των πιθανοτήτων αυτών στον πίνακα B.

Initialization:

Set $\theta =$ A,B,π with random values

For iter=0 until maxIters:

Forward Step:

Find the probability of noticing observation $y_1, y_2, ..., y_t$ it state i at time t.

This is done recursively by using the following formulas

1) $a_i(1) = \pi_i b_i(y_1)$

2) $a_j(t+1) = b_j(y_{t+1}) \sum_{i=1}^{N} a_i(t) a_{ij}$

Backward Step:

Find the probability of ending sequence $y_{t+1}, ..., y_T$ in state i at time t, $\beta_i(t)$.

1) $\beta_i(T) = 1$

2) $\beta_i(t) = \sum_{j=1}^{N} \beta_j(t+1) a_{ij} b_j(y_{t+1})$

Update:

Firstly set $\gamma_i(t) = \frac{a_i t \beta_i(t)}{\sum_{j=1}^{N} a_j(t)\beta_j(t)}$

Calculate the probability of being in state i at time t given the observed sequence Y and the parameters A,B,π.

$$P(X_t = i, X_{t+1} = j | Y, \theta) = \frac{a_i t a_{ij} \beta_j(t+1) b_j(y_t+1)}{\sum_{k=1}^{N} a_k(t)\beta_k(t)}$$

Now the parameters are updated:

$\pi_i^* = \gamma_i(1)$

$a_{ij}^* = \frac{\sum_{t=1}^{T-1} P(X_t=i, X_{t+1}=j|Y,\theta)}{\sum_{t=1}^{T-1} \gamma_i(t)}$

$b_i^*(u_k) = \frac{\sum_{t=1}^{T} k\gamma_i(t)}{\sum_{t=1}^{T} \gamma_i(t)}$, where $k = 1$ if $y_t = u_t$ and 0 otherwise.

Μετά τη δημιουργία όλων των απαραίτητων αρχείων καλείται η εντολή του HTK:

HERest -A -D -T 1 -C config -I labs.mlf -S trainHMM.scp -H hmm0/macros -H hmm0/hmmdefs -M hmm1 monophones0

που ανάλογα με τον ορισμό των HMM που έχουμε περιγράψει παραπάνω, εκπαιδεύει το αρχικό στάδιο της αρχιτεκτονικής μας, hmm0, και εξάγει το νέο σύνολο μοντέλων δεύτερου σταδίου hmm1. Αυτή η διαδικασία επαναλήφθηκε μέχρι το hmm9, έτσι ώστε η τελευταία εντολή να έχει τη μορφή:

HERest -A -D -T 1 -C config -I labs.mlf -S trainHMM.scp -H hmm8/macros -H hmm8/hmmdefs -M hmm9 monophones0

4.2.5.2 Πολλαπλά μοντέλα μίξης ανά κατάσταση

Σε αυτό το τμήμα του κεφαλαίου χρησιμοποιήθηκαν οι ίδιες εντολές όπως προηγουμένως, με τη διαφορά ότι εφαρμόσαμε αύξηση των μοντέλων μίξης σε 2, 4 και 8. Πιο ειδικά, ακολουθήσαμε τα εξής βήματα:

1. Σαν σημείο έναρξης χρησιμοποιήθηκε το σύνολο μοντέλων hmm9 που αποτελούταν από ένα Γκαουσιανό μοντέλο μίξης.
2. Κάθε κατάσταση του γεγονότος-i ορίσθηκε ώστε να έχει 2 (4, 8) μοντέλα μίξης.
3. Το σύστημα εκπαιδεύτηκε με τη χρήση της εντολής HERest.
4. Εξήχθησε η κατάσταση hmm9 του συστήματος με 2 (4, 8) μοντέλα μίξης ανά κατάσταση
5. Επανάληψη των βημάτων 1-4.

Στοχεύοντας την αύξηση των μοντέλων μίξης, δημιουργήθηκε το αρχείο **split.hed** με το περιεχόμενο του Σχήματος 4.10.

```
MU 2 *.state[2-4].mix
```

Σχημα 4.10: Περιεχόμενο αρχείου split.hed

Το παραπάνω αρχείο λειτουργεί σαν ρυθμιστής, καθώς ορίζει το πλήθος των GMM. Η τιμή 2 είναι ενδεικτική και μπορεί να αλλάξει ανάλογα με τους στόχους του χρήστη. Τέλος, δημιουργήσαμε ένα ακριβές αντίγραφο του αρχείου monophones0 με το όνομα tiedlist.
Η εντολή του HTK που χρησιμοποιήθηκε για να εφαρμοστεί ο διαχωρισμός είναι:

HHed -H hmm0/hmmdefs -M hmm1 split.hed tiedlist

όπου hmm1 αποτελεί την έξοδο της εντολής.

4.2.6 Αποκωδικοποίηση φωνημάτων

Πρωτού συνεχίσουμε, πρέπει να αναφερθεί ότι θα παρουσιαστεί μόνο η διαδικασία που μας οδήγησε στα αποτελέσματα, καθώς αυτά αναλύονται στο Κεφάλαιο 5. Μετά το τέλος της εκπαίδευσης των μοντέλων μας, ορίσθηκε η γραμματική που θα παράγει τις λέξεις-γεγονότα. Το βήμα αυτό είναι υποχρεωτικό καθώς το HTK έχει διαμορφωθεί ώστε να χρησιμοποιείται σε φωνητική αναγνώριση όπου λέξεις είναι το αντικείμενο που πρέπει να εντοπισθεί και να κατηγοριοποιηθεί. Σε αυτό τον κλάδο της αναγνώρισης προτύπων τα μοντέλα Markov αποτελούν την πιο αποδοτική λύση. Στη δική μας περίπτωση, δεν έχουμε λέξεις αλλά γεγονότα. Συνεπώς, η γραμματική που ορίστηκε παρήγαγε τέτοια γεγονότα,

είτε μεμονωμένα, είτε ακολουθίες.

Η γραμματική που παράγει μεμονωμένα γεγονότα ορίζεται πολύ εύκολα όπως στο Σχήμα 4.11.

```
S: EVENTS
```

Σχημα 4.11: Ορισμός της γραμματικής στο αρχείο smartplaces.grammar

και το λεξιλόγιο περιγράφει με ποια φωνήματα μπορούν να αντικατασταθούν τα γεγονότα.

```
% EVENTS
KNOCK kn
DOORSLAM ds
STEPS st
CHAIRMOVING cm
SPOON cl
PAPERWORK pw
KEYJINGLE kj
KEYBOARD kt
PHONE pr
APPLAUSE ap
COUGH co
SPEECH spe
SILENCE si
```

Σχημα 4.12: Ορισμός του λεξιλογίου στο αρχείο smartplaces.voca

Η παραγωγή προτάσεων ορίζεται στο αρχείο **gram**:

```
$events= KNOCK | DOORSLAM | STEPS | CHAIRMOVING | SPOON | PAPERWORK |
KEYJINGLE | KEYBOARD | PHONE | APPLAUSE | COUGH | SPEECH | SILENCE;
($events)
```

Σχημα 4.13: Περιεχόμενο αρχείου gram

Με τη δημιουργία των παραπάνω αρχείων, κλήθηκε η εντολή του HTK:

HParse gram wdnet

δημιουργώντας έτσι το αρχείο δικτύου λέξεων **wdnet** που είναι χρήσιμο στη διαδικασία της αποκωδικοποίησης.

Το τελευταίο αρχείο που χρειαζόταν ήταν το λεξικό (**lexicon**). Το λεξικό αποτελείται από όλες τις λέξεις ή συνδυασμό αυτών που υφίστανται σε μία γλώσσα. Το HTK εξ' ορισμού περιέχει το λεξικό του voxforge όπου είναι καταγεγραμμένες η πλεινότητα των αγγλικών λέξεων. Στο δικό μας πρόβλημα όμως, που ασχολείται με ακουστικά γεγονότα, τα φωνητικά δεδομένα δεν υφίστανται. Γίνεται λοιπόν αντιληπτό, πως το αρχείο lexicon θα έχει το ίδιο περιεχόμενο με το smartplaces.voca.

Αρχικά, δοκιμάσαμε τον ταξινομητή HMM για μεμονωμένα γεγονότα. Η εντολή του HTK που κλήθηκε είναι:

HVite -A -D -T 1 -H macros -H hmm9/hmmdefs -C config -S test.scp -l '*' -i test/mfcc/recout.mlf -w wdnet -p 0.0 -s 5.0 voxforge_lexicon tiedlist

το οποίο δίνει σαν έξοδο το αρχείο recout.mlf. Η χρήση του αναλύεται καλύτερα στο Κεφάλαιο 5. Παράδειγμα του περιεχομένου του παραθέτεται στο Σχήμα 4.14. Τα νούμερα αντιπροσωπεύουν τη χρονική στιγμή εντοπισμού, τη διάρκεια του αναγνωρισμένου γεγονότος και τη λογαριθμική πιθανότητά του.

```
#!MLF!#
"*/S08-eventid-0.rec"
0 2200000 ds -803.827332
.
"*/S08-eventid-1.rec"
0 3600000 ds -1316.290039
.
"*/S08-eventid-2.rec"
0 4400000 ds -1754.076294
.
"*/S08-eventid-3.rec"
0 2200000 ds -809.735901
.
"*/S08-eventid-4.rec"
0 4100000 ds -1460.559570
.
"*/S08-eventid-5.rec"
0 10000000 kn -3700.865234
.
"*/S08-eventid-6.rec"
0 3300000 ds -1223.540161
...and so on...
```

Σχημα 4.14: Περιεχόμενο αρχείου recout.mlf

Ακολουθήθηκαν δύο οδοί κατά την διαδικασία της κατηγοριοποίησης:

- Το σύνολο εκπαίδευσης και αξιολόγησης αποτελούταν από μεμονωμένα γεγονότα-ακουστικά αρχεία (isolated training-testing).
- Το σύνολο εκπαίδευσης αποτελούταν από μεμονωμένα γεγονότα, ενώ σαν δεδομένα δοκιμών χρησιμοποιήθηκε ολόκληρο το αρχείο της συνεδρίας 8 (embedded testing).

Για τη λήψη των αποτελεσμάτων, χωρίς όμως να παρουσιάζονται εδώ, κλήθηκε η εντολή του HTK:

HResults -I testref.mlf tiedlist recout.mlf

Σε αυτό το σημείο τελειώνει η περιγραφή της πορείας που ακολουθήσαμε στο HTK. Στο επόμενο υποκεφάλαιο αναλύουμε τη διαδικασία εκπαίδευσης και δοκιμών σε HMM και βαθιά νευρωνικά δίκτυα επί του Kaldi.

4.3 Kaldi

Το δεύτερο και πιο βασικό εργαλείο που χρησιμοποιήθηκε είναι το Kaldi που αρχικά υλοποιήθηκε στο πανεπιστήμιο Johns Hopkins (2009) και πρόσφατα στο ερευνητικό τμήμα της Microsoft (2012) και είναι προγραμματισμένο σε C++. Χρησιμοποιείται κυρίως για αναγνώριση φωνής όπως και το HTK [20]. Κάποια σημαντικά χαρακτηριστικά αυτού του εργαλείου είναι:

1. η χρήση μηχανών πεπερασμένων καταστάσεων (FSM).
2. η υποστήριξη γραμμικής άλγεβρας με τη χρήση βιβλιοθηκών πινάκων μέσα από τις ρουτίνες των εργαλειών BLAS και LAPACK.
3. η ποικιλία των αλγορίθμων που μπορεί να βρει ο χρήστης στη γενική τους μορφή, ώστε να τους αλλάξει μετέπειτα κατά το πρόβλημά του (π.χ αποκωδικοποίηση HMM και νευρωνικών δικτύων με τον ίδιο αλγόριθμο).

Στα ακόλουθα υποκεφάλαια, παρουσιάζουμε την εκτέλεση του Kaldi. Η περιγραφή είναι πιο σύντομη απ΄ ότι στο HTK, καθώς σαν εργαλείο είναι περισσότερο αυτοματοποιημένο και με τη διάτρεξη του βασικού αρχείου script, **run.sh** ο χρήστης επιτυγχάνει:

- Δημιουργία γραμματικής, λεξιλογίου και FSM.
- Διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης και αξιολόγησης.
- Εξαγωγή διανυσμάτων χαρακτηριστικών MFCC.
- Εκπαίδευση κρυφών μοντέλων Markov.
- Εξαγωγή ποσοστού λάθους ταξινομητή HMM.
- Εκπαίδευση βαθιών νευρωνικών δικτύων.
- Εξαγωγή ποσοστού λάθους ταξινομητή DNN.

Όπως στο HTK, έτσι και εδώ τα πειράματα έτρεξαν για δύο περιπτώσεις. Όταν το σύνολο αξιολόγησης αποτελούταν από μεμονωμένα γεγονότα και δεύτερον από αλληλουχίες γεγονότων.

4.3.1 Προετοιμασία δεδομένων

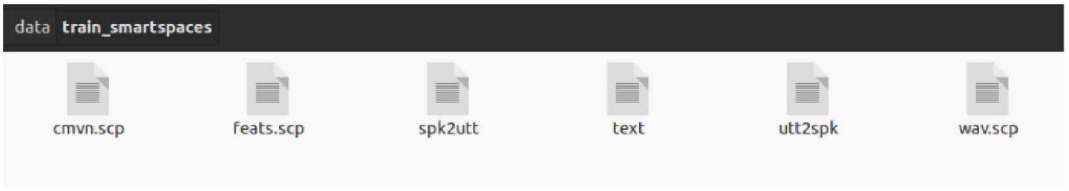
Κατά την εκτέλεση του πρώτου βήματος του πλάνου, όλα τα ακουστικά αρχεία της βάσης μεταφέρθηκαν σε φάκελο με όνομα **waves_smartspaces** όπου κάθε αρχείο είχε την ακόλουθη δομή, S0X-eventid-00Y.wav:

- X αντιπροσωπεύει το νούμερο της συνεδρίας.

- Υ αποτελεί σειριακό αριθμό του γεγονότος (π.χ eventid-001, eventid-202). Χρησιμοποιείται τριψήφια αναπαράσταση ώστε η λίστα με τα αρχεία να είναι ταξινομημένη όπως επιτάζει το Kaldi.

Όταν δοκικάστηκαν ονόματα με *eventid* που είχε κανονική αναπαράσταση ψηφίων το kaldi σταματούσε τη διάτρεξη του run.sh.

Σαν επόμενο βήμα πραγματοποιήθηκε η διάτρεξη του αρχείου **prepare_data.sh**, που οδήγησε στο διαχωρισμό των αρχείων σε σύνολο εκπαίδευσης και αξιολόγησης. Πιο ειδικά, μέσα από αυτό καλείται ένα δευτερεύον perl αρχείο που πραγματοποιεί τη μετακίνηση των αρχείων με εντολές συστήματος. Σε αυτό επιβάλαμε όσα αρχεία **δεν** έχουν όνομα S08 να μετακινηθούν στα δεδομένα εκπαίδευσης, αλλιώς στα δεδομένα για την εξέταση του ταξινομητή.Μετά την εξέταση δημιουργήθηκαν δύο υποφακέλοι στο φάκελο του project, οι train_smartplaces και test_smartplaces που περιέχουν τα ακουστικά αρχεία. Επίσης, δημιουργείται το αρχείο **text** το οποίο αποτελεί ένα αρχέιο ετικετών και αντιστοιχεί τα ακουστικά αρχεία σε γεγονότα. Παραθέτουμε στο Σχήμα 4.15 την έξοδο της εντολής *local/prepape_data.sh waves_smartplaces*.



Σχημα 4.15: Περιεχόμενο καταλόγου train_smartplaces

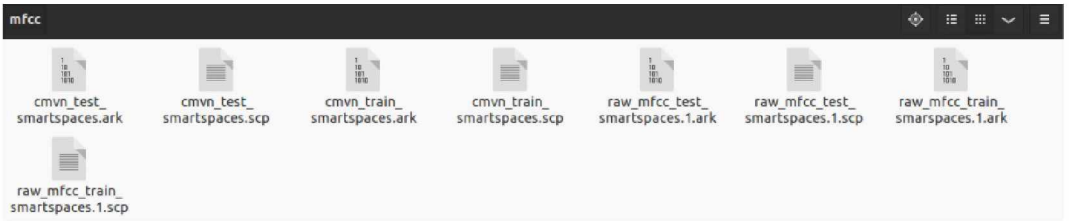
Μετά τον διαχωρισμό των αρχείων ακολούθησε η προετοιμασία της γραμματικής μέσα από την χρήση των εντολών:

- local/prepare_dict.sh
- local/prepare_lang.sh – position-dependent-phones false data/local/dict "<SIL>" data/local data/lang
- local/prepare_lm.sh

Στη γενική τους περίπτωση, για να εκτελεστούν αυτές οι εντολές έπρεπε να οριστεί ο τύπος της γραμματικής. Όταν ορίστηκε το πρωτότυπο της γραμματικής, το αρχείο prepare_dict.sh ξεχώρισε τις λέξεις από τα φωνήματα και το prepare_lang.sh δημιούργησε το σχετικό FSM. Γι΄ αυτό τον λόγο, πριν την εκτέλεση των παραπάνω εντολών δημιουργήσαμε ένα φάκελο με το όνομα **input**.

Όσον αφορά τον σωστό ορισμό της γραμαμτικής για το πρόβλημά μας, όπως και στο HTK, η γραμματική σχεδιάστηκε ώστε κάθε πρόταση να εμπεριέχει μία λέξη, δηλαδή ένα γεγονός. Επίσης, εφόσον ασχολούμαστε με αναγνώριση ακουστικών γεγονότων το κάθε γεγονός αποτελείται από ένα φώνημα. Δυστυχώς όμως, εξ' ορισμού δεν δίνεται η δυνατότητα στον χρήστη να επιλέξει τέτοιου είδους γραμματική. Έτσι έπρεπε να δημιουργήσουμε το δικό μας FSM και να το μεταγλωττίσουμε με το script prepare_lm.sh. Παρόλο που το Kaldi δίνει τη δυνατότητα ορισμού του μοντέλου ***ησυχίας***, το παραλείψαμε και ορίσαμε το δικό μας (δες input/G.txt).

Το επόμενο βήμα ήταν η εξαγωγή των χαρακτηριστικών MFCC από τα ακουστικά αρχεία. Στο κεντρικό αρχείο διάτρεξης αναγράφεται η εντολή steps/make_mfcc.sh που παράγει τα διανύσματα χαρακτηριστικών αυτοματοποιημένα. Στο σημείο αυτό να επισημανθεί πως το Kaldi εξάγει και τους 39 συντελεστές χαρακτηριστικών. Παραθέτουμε ένα απόσπασμα του φακέλου στο Σχήμα 4.16 που περιέχει τα αρχεία χαρακτηριστικών.



Σχημα 4.16: Περιεχόμενο καταλόγου mfcc

4.3.2 Εκπαίδευση & αξιολόγηση κρυφών μοντέλων Markov

Έπειτα από την εξαγωγή χαρακτηριστικών ακολούθησε η εκπαίδευση των HMM ώστε κάθε γεγονός να περιγράφεται από 5 καταστάσεις Markov. Το ίδιο ίσχυσε και για το γεγονός ***ησυχία***. Στο κεντρικό αρχείο διάτρεξης ο χρήστης μπορεί να αλλάξει τις παραμέτρους εκπαίδευσης αλλά εμείς κρατήσαμε τις προτεινόμενες. Συγκεκριμένα, η εκπαίδευση ολοκληρώνεται μετά από 40 επαναλήψεις, παράγοντας έτσι ένα δέντρο καταστάσεων σε μορφή αρχείου fst (δες εργαλείο OpenFST)[21].

Κατά τη διαδικασία της αποκωδικοποίησης γίνεται διάτρεξη αυτού του δέντρου με σκοπό την εύρεση του καλύτερου μονοπατιού (Viterbi decoder) [22]. Ο χρήστης μπορεί να επιβάλλει αναδιάτρεξη του δέντρου με σκοπό τη μείωση του ποσοστού λάθους. Παρόλα αυτά μέσα από δικές μας δοκιμές, η βελτίωση αγγίζει το 2%, γι' αυτό παραλείψαμε αυτή την ενέργεια. Ολα τα παραπάνω επιτυγχάνονται στο κεντρικό αρχείο διάτρεξης με τη χρήση του αρχείου steps/train_mono.sh και steps/decode.sh. Στο τέλος, εμφανίζεται στο τερματικό το

καλύτερο ποσοστό λάθους. Όλα τα αποτελέσματα της αξιολόγησης του ταξινομητή μπορούν να βρεθούν στον κατάλογο exp/mono0a/decode_test_smartplaces που δημιουργήθηκε αυτόματα κατά τη διαδικασία της αποκωδικοποίησης Viterbi.

4.3.3 Εκπαίδευση & αξιολόγηση βαθιών νευρωνικών δικτύων

Σχετικά με τη διαδικασία που ακολουθήσαμε στα βαθιά νευρωνικά δίκτυα, δεν χρειάζεται να ειπωθούν πολλά καθώς τα βήματα ταυτίζονται με αυτά των HMM. Αυτό που πρέπει να επισημανθεί είναι ότι στην εκπαίδευσής του τα βαθιά νευρωνικά δίκτυα χρησιμοποιούν το δέντρο FST και δημιουργούν ένα καινούργιο για την αξιολόγησή τους. Όπως παρουσιάζεται και στο επόμενο κεφάλαιο, δοκιμάστηκαν πολλές διαφορετικοί παράμετροι που αφορούν τα DNN με στόχο τη μελέτη τους και τη λήψη του καλύτερου αποτελέσματος. Τέλος, πρέπει να αναφερθεί ότι στη διαδικασία της κατηγοριοποίησης το Kaldi δεν εξάγει τις χρονικές στιγμές αναγνώρισης ενός γεγονότος (ειδικά στο embedded testing που το χρειαστήκαμε) όπως το HTK, αλλά βρέθηκε λύση μέσω μίας ειδικής εντολής που περιγράφεται και αυτή στο Κεφάλαιο 5.

4.4 SoX

Το SoX αποτελεί μία πλατφόρμα μέσω γραμμής εντολών που επεξεργάζεται αρχεία ήχου ή τα μετατρέπει σε διαφορετικά φορμα. Επιπλέον, μπορεί να εφαρμόσει διάφορες τροποποιήσεις όπως να πραγματοποιήσει ηχογράφηση, να αλλάξει το ρυθμό δειγματοληψίας ή να τμηματοποιήσει και να συγχωνέψει τα ακουστικά αρχεία. Όσον αφορά τη διπλωματική εργασία, το SoX χρησιμοποιήθηκε στην τμηματοποίηση των ακουστικών αρχείων ώστε κάθε ένα να αποτελεί ένα συμβάν. Επίσης, έγινε αλλαγή του ρυθμού της δειγματοληψίας στα 16kHz. Τέλος από .wv τα αρχεία μετατράπηκαν σε φορμα .wav. Χρειάζεται επίσης να σημεωθεί πως οι εντολές του SoX συσσωρεύτηκαν σε script αρχεία Python με σκοπό την εφαρμογή τους σε ολόκληρη τη βάση δεδομένων.

Οι εντολές που χρησιμοποιήθηκαν ήταν οι εξής:

- ανάκληση πληροφορίας σχετικά με το αρχείο, sox *filename* -n stat.
- αλλαγή του ρυθμού δειγματοληψίας, sox *input.wav* -r 16000 *output.wav* .
- εξαγωγή τμήματος από ακουστικό αρχείο, sox *input output* trim *start duration*.

4.5 Σημειώσεις

Κάθε χρήστης του Kaldi και του HTK μπορεί να ανακαλύψει πολλές ομοιότητες μεταξύ των δύο εργαλείων καθώς χρησιμοποιούν shell script αρχεία για να λειτουργήσουν.

Γίνεται επομένως εύκολα αντιληπτό, πως ο χρήστης δεν χρειάζεται να προβληματιστεί με τον κώδικα υλοποίησης, αλλά να χρησιμοποιήσει σωστά τις κλήσεις των συναρτήσεων όπως ορίζει το API. Παρόλο που το Kaldi αποτελεί ένα πιο καινοτόμο εργαλείο πρέπει να ειπωθεί ότι το API του δεν είναι ευανάγνωστο, λείπουν λεπτομέρειες που ο κοινός προγραμματιστής μπορεί να μην γνωρίζει. Παρόλα αυτά, το forum στο sourceforge.net είναι χρήσιμο έως απαραίτητο στον χρήστη που κάνει τα πρώτα του βήματα σε script γλώσσα.

Η βασική μορφή εντολών του Kaldi όπως και του HTK είναι: <εντολή> <ρυθμίσεις> <είσοδος> <έξοδος>. Στη συνέχεια θα παρουσιάσουμε ποιές εντολές χρησιμοποιήσαμε για τη δική μας υλοποίηση.

Κεφάλαιο 5

Πειράματα & Αποτελέσματα

5.1 Πειραματικό πλαίσιο

Για την πραγματοποίηση των πειραμάτων σχετικά με το πρόβλημα της κατηγοριοποίησης γεγονότων χρησιμοποίηθηκαν τα εργαλεία Kaldi και HTK σε συνδυασμό με τη βάση δεδομένων UPC-TALP Multimodal που περιγράψαμε στο Κεφάλαιο 3. Συγκεκριμένα, με το πρώτο εργαλείο εκπαιδεύσαμε 13 μοντέλα Markov και έπειτα τα αξιολογήσαμε ως προς τη δυνατότητα αναγνώρισης των γεγονότων με τον αποκωδικοποιητή Viterbi. Όλα τα πειράματα διεξήχθησαν στα εργαστήρια Β2 του Πανεπιστημίου Θεσσαλίας στον Βόλο. Κατά την διαδικασία της εκπαίδευσης χρησιμοποιήθηκαν μεμονωμένα αρχεία ήχου που το καθένα περιέγραφε ένα γεγονός. Σε αντίθετη περίπτωση, στην αποκωδικοποίηση χρησιμοποιήθηκαν τόσο μεμονωμένα αρχεία ήχου, όσο και ολόκληρο το ακουστικό αρχείο της συνεδρίασης 8. Τα χαρακτηριστικά που χρησιμοποιήθηκαν για να ομαδοποιήσουν τα γεγονότα είναι οι συντελεστές MFCC, DELTA-MFCC και DELTA-DELTA-MFCC, ενώ οι ταξινομητές αυτοί του Κεφαλαίου 2, δηλαδή τα κρυφά μοντέλα Markov και τα DNN. Επίσης, πρέπει να επισημανθεί ότι από τις 8 συνολικά συνεδρίες, οι πρώτες 7 αποτέλεσαν υλικό εκπαίδευσης των ταξινομητών, ενώ η συνεδρία 8 χρησιμοποιήθηκε για την κατηγοριοποίηση των γεγονότων. Από τα 24 συνολικά κανάλια χρησιμοποιήθηκε μόνο το κανάλι/μικρόφωνο 1.

Κατά τη διάρκεια των πειραμάτων χρησιμοποιήθηκαν δύο υπολογιστικά συστήματα. Ένας φορητός υπολογιστής Lenovo Y550p Ideapad και ένας Η/Υ τύπου πύργου. Η ολοκληρωμένη διαδικασία της εκπαίδευσης και του ελέγχου αποτελεσμάτων ταξινόμησης σε HMM και βαθιά νευρωνικά δίκτυα διήρκησε στο σύνολό της περίπου 30-45 λεπτά ανάλογα με τις παραμέτρους που είχαν δοκιμαστεί. Πρέπει να σημειωθεί ότι παρόλο που τα εργαλεία επιτρέπουν τη χρήση τεχνολογίας CUDA, στα συγκεκριμένα πειράματα χρησιμοποιήθηκε μόνο επεξεργαστική δύναμη.

Τα χαρακτηριστικά των δύο μηχανημάτων είναι:

Lenovo Ideapad Y550p

- Ubuntu 14.04 LTS 64bit
- 4GB DDR3 RAM
- Intel Core i7 Q720@1.60GHz
- GeForce GT 240M
- 50GB HDD (partition)

Desktop

- Ubuntu 14.04 LTS 64bit
- 12GB DDR3 RAM @1666MHz
- Intel Core i7 870@2.93GHz
- GeForce GTX 750Ti
- 54GB HDD (partition)

Τέλος, Ο πίνακας 5.1 παρουσιάζει τις 13 κλάσεις στις οποίες βασίστηκαν τα πειράματα, με 13η να είναι αυτή της ησυχίας.

Ηχητικό Γεγονός	Σύμβολο
Χτύπημα Πόρτας	kn
Κλείσιμο Πόρτας	ds
Βήματα	st
Μετακίνηση Καρέκλας	cm
Κουτάλι/Κουδούνισμα φλιτζανιού	cl
Τύλιγμα Χαρτιού Εργασίας	pw
Ήχος Κλειδιών	kj
Ήχος Πληκτρολογίου	kt
Ήχος Τηλεφώνου/Μουσική	pr
Χειροκρότημα	ap
Βήχας	co
Ομιλία	sp
Ησυχία	si

Πινακας 5.1: Σύνολο κλάσεων της UPC-TALP Multimodal Database

5.2 Μετρικές αξιολόγησης

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκαν δύο βασικές μετρικές. Η πρώτη μετρική είναι ανάλογη του WER στον κλάδο της Αυτόματης Αναγνώρισης Ομιλίας, όπου οι λέξεις της γραμματικής αναφέρονται στα ακουστικά γεγονότα. Όμως, αυτή η μετρική δεν λαμβάνει υπόψιν τη χρονική διάρκεια του γεγονότος, αλλά την ακριβή χρονική αντιστοίχιση της πρόβλεψης του ταξινομητή σε σχέση με το ground truth. Συγκεκριμένα, απευθυνόμαστε σε αυτή ως *ποσοστό λάθους αναγνώρισης ακουστικών γεγονότων* acoustic event error rate ή AEER) και χρησιμοποιήθηκε κατά τη διάρκεια της εργασίας όταν τα δεδομένα ήταν μεμονωμένα χωρίς επικαλύψεις. Στη περίπτωση που η ταξινόμηση πραγματοποιήθηκε με αλληλουχία γεγονότων χρησιμοποιήθηκε η μετρική *ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου* (frame misclassification rate ή FMR). Η ακρίβεια ενός ταξινομητή [23] σε σχέση με το λόγο σφάλματος λέξης δίνεται από τον ακόλουθο τύπο.

$$Acc = 1 - AEER \tag{5.1}$$

Πιο συγκεκριμένα, το AEER ορίζεται ως ακολούθως

$$AEER = \frac{S + D + I}{N} \tag{5.2}$$

ή

$$AEER = \frac{S + D + I}{S + D + C} \tag{5.3}$$

όπου

- S είναι ο αριθμός των αντικαταστάσεων
- D είναι ο αριθμός των διαγραφών
- I είναι ο αριθμός των εισαγωγών
- C είναι ο αριθμός των σωστών ταξινομημένων παρατηρήσεων
- N είναι το πλήθος των παρατηρήσεων

Τα παραπάνω σύμβολα συνεισφέρουν στον υπολογισμό κάποιων επιμέρους μετρικών λάθους που χρησιμοποιούνται ευρέως στην αναγνώριση ήχων και γεγονότων. Πιο αναλυτικά ορίζουμε ως:

- **Λάθος εισαγωγής (insertion error)**. Το λάθος που υπολογίζεται όταν ο ταξινομητής αναγνωρίζει περισσότερα γεγονότα από όσα έπρεπε.
- **Λάθος διαγραφής (deletion error)**. Αποτελεί ουσιαστικά την αντίθετη περίπτωση από το λάθος εισαγωγής.

- **Λάθος αντικατάστασης (substitution error)**. Αποτελεί την πιο απλή μορφή λάθους καθώς αναφέρεται στο λάθος του ταξινομητή να κατηγοριοποιήσει το γεγονός σωστά. Για παράδειγμα, ενώ έπρεπε να αναγνωριστεί το γεγονός βήματα, τελικά αναγνωρίστηκε χειροκρότημα.

5.3 Αποτελέσματα μετρήσεων

Στο παρών τμήμα του κεφαλαίου θα παρουσιαστούν τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν πάνω στη βάση δεδομένων. Αρχικά, γίνεται αναφορά στα αποτελέσματα του εργαλείου HTK τόσο με τη χρήση μεμονομένων γεγονότων, όσο και με ακολουθίες συμβάντων. Αυτή τη διαδικασία ακολουθεί και το υποκεφάλαιο 5.3.2 με τα αποτελέσματα του εργαλείου Kaldi. Η διαδικασία που οδήγησε στη λήψη των αποτελεσμάτων περιγράφεται αναλυτικά στο Κεφάλαιο 4, ενώ μπορούν να βρεθούν και σύνδεσμοι στο τμήμα *Παραρτήματα* σε βασικά τμήματα κώδικα. Εκτός λοιπόν από την παρουσιάση των αποτελεσμάτων, πραγματοποιείται αναφορά σε τεχνικές βελτίωσης των δεδομένων και της τεχνικής εξαγωγής χαρακτηριστικών με σκοπό τη μείωση λάθους τόσο στη μέθοδο των κρυφών μοντέλων Markov όσο και στα βαθιά νευρωνικά δίκτυα.

Τα χαρακτηριστικά που χρησιμοποιήθηκαν, όπως περιγράφεται και στο Κεφάλαιο 2, ήταν οι συντελεστές MFCC είτε στην απλή τους μορφή είτε σε συνδυασμό τους (DELTA-MFCC, DD-MFCC). Μέσα από αυτά τα αποτελέσματα γίνεται μία εκτίμηση της απόδοσης των επιλεγμένων χαρακτηριστικών αλλά και των εργαλείων όσον αφορά την αναγνώριση ακουστικών γεγονότων.

5.3.1 Αποτελέσματα HTK

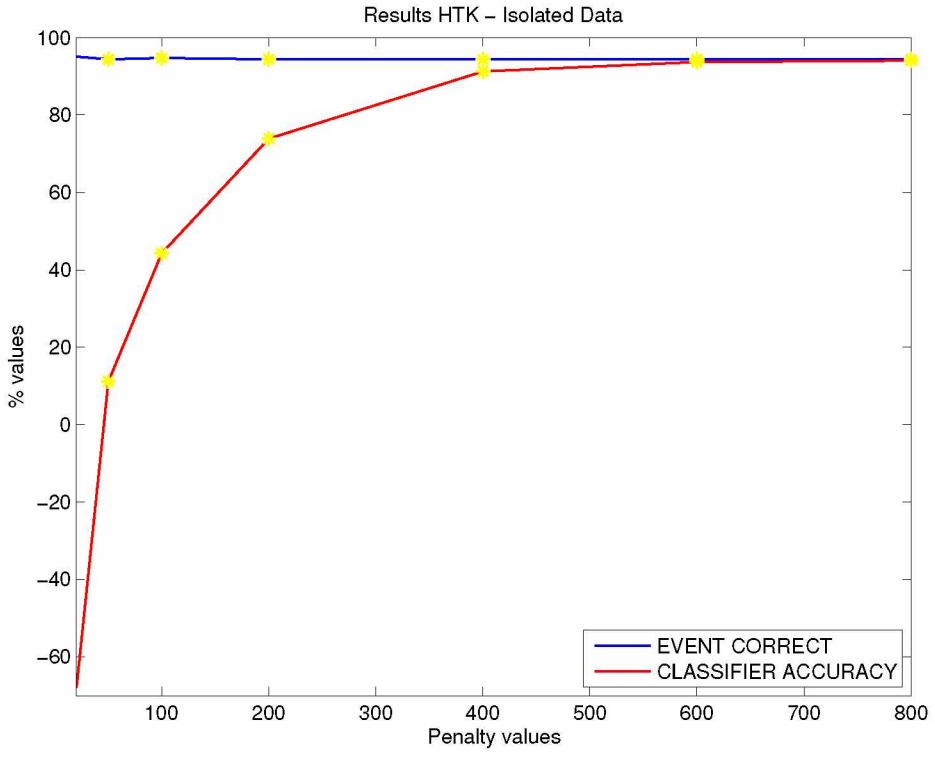
Στην υποενότητα αυτή θα γίνει παρουσίαση των εκτιμήσεων του ταξινομητή HMM χρησιμοποιώντας το εργαλείο HTK, καθώς τα νευρωνικά δίκτυα εμπεριέχονται μόνο στο Kaldi. Όπως προαναφέρθηκε, τα ηχητικά κομμάτια της συνεδρίας 8 αποτελούν τα στοιχεία αναφοράς για τον ταξινομητή.

Αρχικά, δοκιμάστηκαν τα μεμονωμένα κομμάτια ήχου (κάθε κομμάτι ήχου αποτελεί ένα συμβάν) και ως validation data η συνεδρία 8 τμηματοποιήθηκε με την ίδια μέθοδο. Δοκιμάστηκαν διάφορες τιμές για την παράμετρο *λάθος εισαγωγής λέξης* (word insertion penalty ή WIP). Καθώς αυξάνεται η μεταβλητή αυτή, η γραμματική που παράγει τις λέξεις-συμβάντα υποχρεώνεται να έχει σαν έξοδο μεμονωμένες λέξεις. Αυτή η παραμετρικοποίηση βοηθάει στην κατακόρυφη μείωση του λαθους εισαγωγής και διαγραφής με αποτέλεσμα την καλύτερη αξιολόγηση του ταξινομητή σε μεμονωμένα γεγονότα ή γνωστό ως isolated

validation. Στον Πίνακα 5.2 παρουσιάζονται τα αποτελέσματα αυτά και στο Σχήμα 5.1 η εξέλιξη την ακρίβειας του ταξινομητή με διάφορες τιμές της προαναφερθείσας παραμέτρου.

WIP	Σωστές Αναγνωρισμένες Λέξεις (%)	Ακρίβεια (%)
20	95.12	-67.94
50	94.43	11.15
100	94.77	44.25
200	94.43	73.87
400	94.43	91.29
600	94.43	93.73
800	94.43	94.08
ΒΕΛΤΙΣΤΟΣ ΛΟΓΟΣ		16/287
ΛΑΘΟΥΣ AEER(%)		5.57

Πινακας 5.2: Αποτελέσματα HMM ταξινομητή με μεμονωμένα δεδομένα και χρήση μοναδικής Γκαουσιανής κατανομής



Σχημα 5.1: Πρόοδος της σωστής πρόβλεψης λέξης και της ακρίβειας του ταξινομητή με χρήση μοναδικής Γκαουσιανής κατανομής

Όπως παρουσιάζεται στο Σχήμα 5.1, αυξάνοντας την τιμή της παραμέτρου WIP, μπορεί από τη μία μεριά ο ταξινομητής να κάνει λιγότερες σωστές προβλέψεις, αλλά από την άλλη

κερδίζει σε ακρίβεια. Μέσα λοιπόν από πολλαπλές διατρέξεις επιτεύχθηκε το ποσοστό λάθους αναγνώρισης ακουστικών γεγονότων να μειωθεί στο 5.57% ή διαφορετικά, οι επιτυχείς προβλέψεις από τον ταξινομητή έφτασαν το 94.43%.

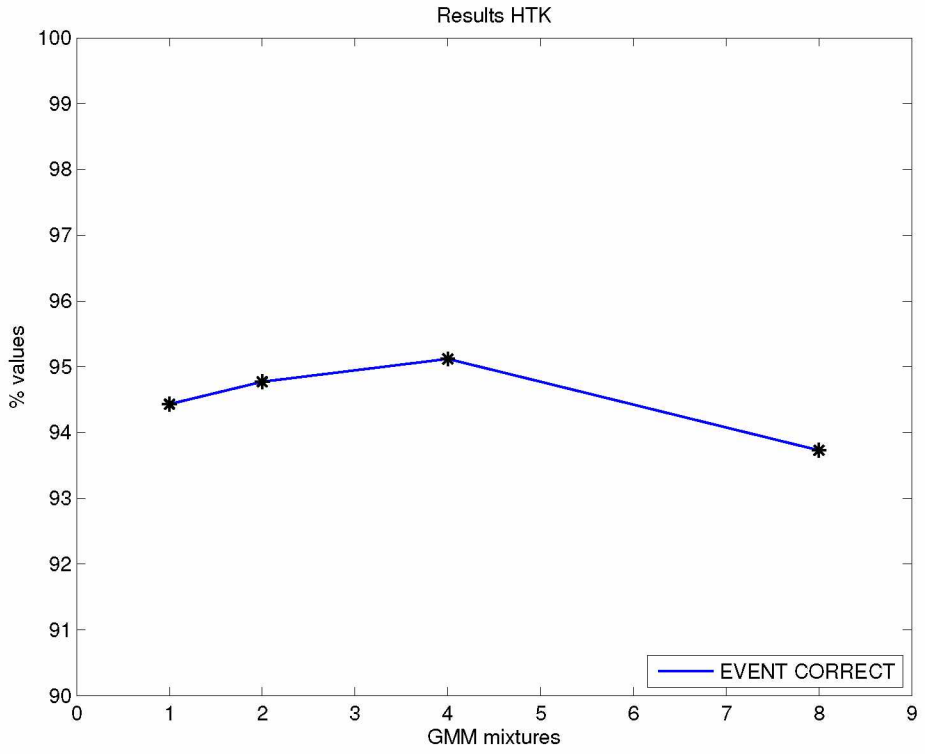
Επιπρόσθετα, τα δεδομένα δοκιμάστηκαν μέσω μίας διαφορετικής πειραματικής οδού. Σε πρώτη φάση, να αναφερθεί ότι το HTK δίνει τη δυνατότητα τροποποίησης της γραμματικής σε αρκετά χαμηλό επίπεδο, αφού μπορεί να ορισθεί άμμεσα από έναν γράφο. Στο συγκεκριμένο στάδιο, δεν χρειάστηκε να ορισθεί από εμάς τέτοιος γράφος, αλλα ήταν αναγκαίο να αλλαχθούν οι μεταβλητές της γραμματικής ώστε να παραχθούν προτάσεις μίας λέξεις, όπως διατυπώσαμε στο Κεφάλαιο 4. Με αυτή την πράξη, απελευθερώσαμε τις μετρήσεις από την τιμή του WIP.

Το HTK εξ' ορισμού χρησιμοποιεί μία Γκαουσιανή κατανομή για την μοντελοποίηση κάθε γεγονότος. Για αυτό τον λόγο, τα ίδια πειράματα έτρεξαν επαναλλελημενα για 2, 4 και 8 Γκαουσιανές κατανομές. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 5.3 και στο Σχήμα 5.2.

Πλήθος GMM κατανομών	Σωστές Αναγνωρισμένες Λέξεις (%)	Ακρίβεια (%)
1	94.43	94.43
2	94.77	94.77
4	95.12	95.12
8	93.73	93.73
ΒΕΛΤΙΣΤΟ ΠΟΣΟΣΤΟ		14/287
ΛΑΘΟΥΣ AEER(%)		4.88

Πινακας 5.3: Αποτελέσματα HMM ταξινομητή με μεμονωμένα δεδομένα επικύρωσης

Όπως παρατηρούμε από το Σχήμα 5.2 η απόδοση του ταξινομητή αυξάνεται με τη χρήση περισσότερων Γκαουσιανών κατανομών, αλλά για πλήθος Γκαουσιανών ίσο με 8, οι σωστές αναγνωρισμένες λέξεις μειώνονται. Αυτό συμβαίνει λόγω της συνεχής αύξησης των GMM κατανομών, πράγμα που κάνει το σύστημά μας υπερεκπαιδευμένο (overtrained) πράγμα που οδηγεί στην αυξηση του λόγου AEER.



Σχημα 5.2: Πρόοδος της σωστής πρόβλεψης λέξης και της ακρίβειας του ταξινομητή με χρήση πολλαπλών Γκαουσιανών κατανομών

Τα παραπάνω αποτελέσματα εμφανίζονταν στο τερματικό του Ubuntu μετά από τη διάτρεξη των εντολών που παρουσιάστηκαν στο Κεφάλαιο 4. Συγκεκριμένα η αυτούσια έξοδος του HTK προσαρτάται στο Σχήμα 5.3:

```
SENT: %Correct=94.76 [H=271, S=15, N=286]
WORD: %Corr=94.77, Acc=94.43 [H=272, D=1, S=14, I=1, N=287]
```

Σχημα 5.3: Έξοδος HTK στο τερματικό

Αναλύοντας το αποτέλεσμα, το SENT αναφέρεται στο ποσοστό επιτυχίας του ταξινομητή στη διαδικασία αναγνωρίσης αλληλουχίας γεγονότων. Όμως, έχοντας μεμονωμένα δεδομένα, αυτή η μετρική δεν έχει υπόσταση και όπως περιγράψαμε στην εισαγωγή του κεφαλαίου, η αναγνώριση ακουστικών αλληλουχιών αποτελεί ένα ξεχωριστό πρόβλημα που αξιολογείται με τη δική του μετρική.

Πιο σημαντικά, το WORD είναι ανάλογο του AEER που ορίσαμε στην υποενότητα των μετρικών. Πρακτικά, αναφέρεται στην επιτυχία του ταξινομητή να αναγνωρίσει μεμονωμένα γεγονότα. Μέσα στις αγκύλες παρατηρούνται κάποιες παράμετροι που δίνουν σημαντικές πληροφορίες:

- H είναι το σύνολο των σωστά αναγνωρισμένων ακουστικών γεγονότων
- D αναφέρεται στο λάθος διαγραφής
- S αποτελεί το λάθος αντικατάστασης
- I είναι το λαθος εισαγωγής
- N αναφέρεται στο σύνολο των μεμονωμένων γεγονότων που χρησιμοποιήθηκαν σαν σημείο αναφοράς

Κλείνοντας αυτή την ενότητα πρέπει να γίνει ξεκάθαρος ο τρόπος με τον οποίο εξήχθησαν τα πραπάνω αποτελέσματα από το HTK. Κατά την εκτέλεση του αποκωδικοποιητή Viterbi, το HTK αποθηκεύει δυναμικά στο αρχείο recout.mlf το γεγονός που αναγνώρισε ο ταξινομητής σε κάθε ηχητικό απόσπασμα της συνεδρίας 8. Παράδειγμα αυτού του αρχείου μπορεί να βρεθεί στο τμήμα *Παραρτήματα* στο τέλος της διπλωματικής εργασίας. Έχοντας λοιπόν δημιουργήσει αυτό το αρχείο, το HTK συγκρίνει το ground truth που του έχει ορισθεί στα αρχεία ετικετών με το περιεχόμενο του recout.mlf. Επειδή όμως τα δεδομένα μας αποτελούνται από μεμονωμένα ακουστικά συμβάντα, το λάθος εισαγωγής και διαγραφής πρέπει να τείνει στο 0, όπως και συμβαίνει, αφού με την διάτρεξη διάφορων πειραμάτων οι τιμές τους παρέμειναν στο 1.

Σε δεύτερη φάση δοκιμάσαμε το σύστημα του HTK σε ενσωματωμένες δοκιμές (embedded testing). Το σύνολο των δεδομένων εκπαίδευσης παρέμεινε το ίδιο αλλά έγινε εξαγωγή των acceleration MFCC coefficients ή αλλιώς delta-delta MFCC coefficients που ο ορισμός των οποίων γίνεται στο Κεφάλαιο 2. Επίσης, για να ενισχύσουμε το μοντέλο της *ησυχίας* (si) εξαγάγαμε περισσότερα κομμάτια ήχου από τα ήδη υπάρχοντα. Αυτό επιτεύχθηκε αφού στις πρώτες 7 συνεδρίες εμφανίζονται χρονικά κενά μεταξύ του τέλους ενός γεγονότος και της αρχής του επόμενου. Υπολογίζοντας έτσι αυτή τη χρονική διαφορά με τη χρήση του εργαλείου SoX, σε κάθε συνεδρία αντιστοιχούσαν τουλάχιστον 150 κομμάτια μικρού μήκους που δεν ανήκαν σε κάποια από τις 12 κλάσεις. Χρειάζεται επίσης να σημειωθεί ότι έχοντας (embedded testing), ως σημείο αναφοράς για την αξιολόγηση του ταξινομητή χρησιμοποιήθηκε ολόκληρο το ακουστικό αρχείο της συνεδρίας 8. Μολαταύτα, αυτό το ακουστικό αρχείο θα μπορούσε να τμηματοποποιηθεί σε 4 επιμέρους κομμάτια ξανά με τη χρήση του εργαλείου SoX. Κάτι τέτοιο όμως δεν πραγματοποιήθηκε αφού το κόψιμο των ακουστικών αρχείων θα επέφερε απώλεια σε ακουστικά παράθυρα. Όλα τα προαναφερθέντα στοιχεία διατυπώθηκαν σαν εισαγωγή για την περιγραφή των αποτελεσμάτων αλλά και της μετρικής FMR που αναφέρθηκε στο υποκεφάλαιο 5.2.

Πριν προχωρήσουμε στην περιγραφή εξαγωγής των αποτελεσμάτων πρέπει να αναφερθεί ότι το αρχείο recout.mlf μετά τη διαδικασία του decoding είχε σχεδόν την ίδια μορφή με αυτή

των μεμονωμένων πειραμάτων, περιέχοντας όμως ένα αρχείο, της συνεδρίας 8, διατρέχοντάς το χρονικά και αναγνωρίζοντας τα γεγονότα σε επίπεδο ms. Με σκοπό λοιπόν τη χρήση αυτής της μετρικής, γράφτηκε κώδικας σε python που πραγματοποιούσε τις εξείς ενέργειες:

1. Ανοίγει το αρχείο S08.csv και αναλύει τα γεγονότα σε πλαίσια. Για παράδειγμα, αν ένα γεγονός διαρκεί από 0 έως 2,7 δευτερόλεπτα, τότε υπάρχουν 270 παράθυρα των 10ms που περιγράφουν αυτό το γεγονός. Στο σύνολο της διάρκειας του ακουστικού αρχείου της συνεδρίας 8, υπολογίζονται 149991 παράθυρα. Συνεπώς, από το αρχείο εξόδου του HTK εξαγάγαμε το ανάλογο πλήθος παραθύρων. Τα χρονικά κενά προφανώς συμπληρώθηκαν με το γεγονός της σιωπής (si).

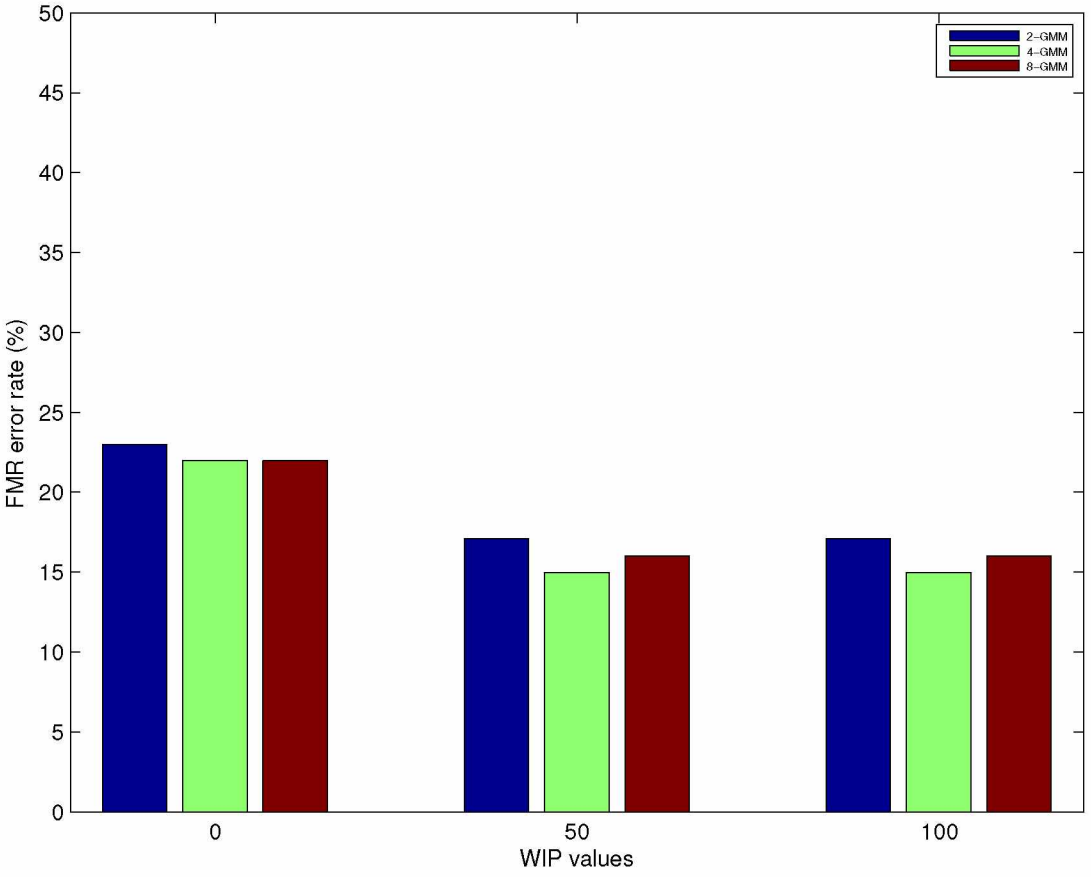
2. Ανοίγει το αρχείο εξόδου recout.mlf και επαναλαμβάνει την ίδια διαδικασία όπως στο βήμα 1.

3. Συγκρίνει τα παράθυρα 1-1 και υπολογίζει τον λόγο $\frac{\#matched\_frames}{\#total\_frames}$. Συμπερασματικά, έχει καταστεί σαφές ότι αυτός ο λόγος περιγράφει το ποσοστό λάθους ταξινόμισης σε επίπεδο πλαισίου.

Έχοντας συνθέσει αυτόν τον μικρό αλγόριθμο καλούμε τις εντολές του HTK που αναφέρθηκαν στο Κεφάλαιο 4 με διαφορετικές παραμέτρους για το WIP. Στο πλαίσιο αυτό κατανοούμε ότι δεν μπορούμε να χρησιμοποιήσουμε μία γραμματική που παράγει μία λέξη, αφού το ακουστικό αρχείο για την αξιολόγηση του ταξινομητή αποτελείται από μία πρόταση με την ακολουθία όλων των γεγονότων που συνέβησαν στην συνεδρία 8. Συνεπώς, μέσα από τις δοκιμές διαφόρων τιμών, όπως παρουσιάζεται στον Πίνακα 5.4, καταλήγουμε στο βέλτιστο αποτέλεσμα που μπορεί να δώσει το HTK.

	Λανθασμένες προβλέψεις		
WIP	2GMM	4GMM	8GMM
0	22.99%	21.98%	21.98%
50-80	17.11%	14.98%	16.02%
100	17.11%	14.98%	16.02%
ΠΟΣΟΣΤΟ	22498/149991		
ΛΑΘΟΥΣ FMR(%)	14.98		

Πινακας 5.4: Αποτελέσματα HMM ταξινομητή και ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου με ενσωματωμένα δεδομένα της συνεδρίας 8



Σχημα 5.4: Πρόοδος της απόδοσης του ταξινομητή HMM σε επίπεδο FMR.

Με αναφορά τον παραπάνω πίνακα παρατηρείται ότι το βέλτιστο ποσοστό λάθους αγγίζει το 14.98%, που είναι αρκετά ικανοποιητικό αν αναλογιστούμε ότι καθώς τα γεγονότα συμβαίνουν στον χρόνο, υπάρχουν παύσεις που μπορεί να λειτουργούν ως εμπόδια στην ακριβή οριοθέτηση του γεγονότος μέσα σε αυτόν. Το αποτέλεσμα αυτό έγινε εφικτό με την καλή εκπαίδευση του μοντέλου σιωπής. Χωρίς την σωστή εκπαίδευση της κλάσης si, το ποσοστό επιτυχίας δεν μπορούσε να ξεπεράσει το 60%. Γίνεται, επομένως, εύκολα αντιληπτό ότι η ενέργειά μας πάνω στα δεδομένα εκπαίδευσης βελτίωσε σε μεγάλο βαθμό την απόδοση του HMM ταξινομητή. Αναντίρρητα λοιπόν, η ίδια λογική ακολουθήθηκε και στο εργαλείο Kaldi στα βαθιά νευρωνικά δίκτυα.

5.3.2 Αποτελέσματα Kaldi

Στην υποενότητα αυτή θα παρουσιαστούν τα αποτελέσματα του Kaldi τόσο στη χρήση HMM ταξινομητή όσο και βαθιών νευρωνικών δικτύων, που είναι και ο κύριος στόχος της διπλωματικής. Δηλαδή θα αξιολογηθεί η συμπεριφορά τους στο πρόβλημα της αναγνώρισης ακουστικών γεγονότων. Με αφετηρία τη διαδικασία και το σκεπτικό που ακολουθήσαμε για το Kaldi στο Κεφάλαιο 4 εξαγάγαμε τα αποτελέσματα που θα παρουσιαστούν παρακάτω.

Στο σημείο αυτό πρέπει να αναφερθεί ότι το Kaldi είναι αρκετά πιο παραμετρικοποιήσιμο από το HTK, αλλά στις περισσότερες περιπτώσεις χρησιμοποιήθηκαν οι default τιμές στα αρχεία ρύθμισης. Ένα τυπικό παράδειγμα αποτελεί η χρήση των acceleration συντελεστών ως χαρακτηριστικά εκπαίδευσης και αξιολόγησης. Μέσα από τη διαδικασία χρήσης και αξιοποίησης των πόρων του Kaldi, ο προγραμματιστής μπορεί να αλλάξει τον τύπο χαρακτηριστικών, τον αριθμό HMM καταστάσεων στο μοντέλο κάθε κλάσης, την γραμματική και το λεξικό, τις παραμέτρους στα νευρωνικά δίκτυα, και να επιλέξει μεθόδους εκπαίδευσης και αξιολόγησης ανάλογα με το πρόβλημα του.

Αρχικά, όπως και στο HTK, δοκιμάστηκαν τα μεμονωμένα γεγονότα. Η προετοιμασία των δεδομένων και η εκπαίδευση τους περιγράφεται αναλυτικά στο Κεφάλαιο 4. Επιπλέον, στα μοντέλα Markov χρησιμοποιήθηκαν 5 καταστάσεις για την μοντελοποίηση μία κλάσης, σε αντίθεση με τον αρχικό σχεδιασμό του προβλήματος που προέβλεπε την ύπαρξη 3 καταστάσεων σε κάθε κλάση. Κάθε HMM κατάσταση ουσιαστικά διαχωρίζει το διάνυσμα χαρακτηριστικών σε μικρότερα τμήματα. Άρα ουσιαστικά διαχωρίζει το παράθυρο του σήματος σε μικρότερα χρονικά κομμάτια. Όπως λοιπόν γίνεται αντιληπτό, η αύξηση των καταστάσεων ενός HMM μοντέλου είναι πεπερασμένη, αφού και ο χρόνος αποτελεί ένα πεπερασμένο μέγεθος. Είναι χρήσιμο να τονιστεί επίσης, ότι παρόλο που η μεταβλητή WIP υπάρχει ορισμένη, δεν επήλθε τροποίησή της. Στον Πίνακα 5.5 παρουσιάζονται τα αποτελέσματα με διάφορες παραμέτρους για τα νευρωνικά δίκτυα και τις προκαθορισμένες για τα μοντέλα Markov.

HMM	DNN			
	Επαναλήψεις	Πρόσθετες Επαναλήψεις	Κρυφά Επίπεδα	AEER(%)
	15	5	3	16.08
	15	5	5	15.73
	15	2	3	15.73
	10	5	3	13.64
	2	2	1	13.29
	10	5	1	12.24
	5	5	1	11.89
	3	2	1	11.89
7.34				11.89

ΠΙΝΑΚΑΣ 5.5: Ποσοστό λάθους αναγνώρισης ακουστικών γεγονότων, με χρήση βαθιών νευρωνικών δικτύων και κρυφών μοντέλων Markov.

Όπως παρατηρούμε τα μοντέλα Markov συμπεριφέρονται καλύτερα από τα βαθιά νευρωνικά δίκτυα, παρόλο που η διαφορά στο ποσοστό λάθους είναι μικρή. Όσον αφορά τα νευρωνικά δίκτυα γίνεται αντιληπτό ότι στο πρόβλημα μας η χρήση μικρών παραμέτρων επιφέρει τα καλύτερα αποτελέσματα, πράγμα που επικυρώνει την τοποθέτησή μας στην θεωρία των νευρωνικών δικτύων στο Κεφάλαιο 1. Αυτό μπορεί να οφείλεται στο γεγονός ότι ο όγκος των δεδομένων δεν είναι μεγάλος και τα χαρακτηριστικά δεν έχουν μεγάλη διάσταση (μέχρι 39 συντελεστές MFCC). Δηλαδή, η περαιτέρω ανάλυση σε πολλαπλά κρυφά επίπεδα (περισσότερα των 5) δημιουργεί φαινόμενα υπερ-εκπαίδευσης, πράγμα που μειώνει την απόδοση του ταξινομητή.

Έπειτα από την επιτυχημένη εξαξωγή των αποτελεσμάτων για μεμονωμένα συμβάντα, δοκιμάστηκαν αλληλουχίες γεγονότων. Ακολουθώντας λοιπόν τη διαδικασία του HTK, στο σύνολο δεδομένων εκπαίδευσης τοποθετήθηκαν τα τμηματοποιημένα ακουστικά αρχεία και σαν σημείο αξιολόγησης ολόκληρο το ακουστικό αρχείο της συνεδρίας 8. Στο σημείο αυτό πρέπει να επισημανθεί ότι εξ΄ ορισμού το Kaldi δεν εξάγει τις χρονικές περιόδους στις οποίες έχει αναγνωρίσει ένα γεγονός (αρχείο recout.mlf του HTK). Γι΄ αυτό τον λόγο, χρησιμοποιήθηκε μία επιπλέον εντολή του Kaldi, που περιγράφεται στον προηγούμενο κεφάλαιο, και ουσιαστικά αποσκοπεί στην ανάκτηση αυτής της πληροφορίας. Όπως και προηγουμένως, μέσω του κώδικα python, κατέστη εφικτός ο υπολογισμός του ποσοστού λάθους ταξινόμησης σε επίπεδο πλαισίου. Στον Πίνακα 5.6 παρουσιάζονται τα αποτελέσματα των μετρήσεων με τη χρήση κρυφών μοντέλων Markov.

Πλήθος HMM καταστάσεων	Λανθασμένες προβλέψεις
3	22.99%
5	20.58%
10	20.93%
ΠΟΣΟΣΤΟ ΛΑΘΟΥΣ	30868/149991
FMR(%)	20.58

ΠΙΝΑΚΑΣ 5.6: Ποσοστό λάθους ταξινόμησης ακολουθίας γεγονότων σε επίπεδο πλαισίου, με χρήση μοντέλων Markov.

Όπως παρατηρείται το λάθος του HMM ταξινομητή στο Kaldi αγγίζει το 20%. Αυτό όμως που μας ενδιαφέρει είναι να αξιολογήσουμε την απόδοση των βαθιών νευρωνικών δικτύων σε σχέση με τον HMM ταξινομητή. Έτσι λοιπόν ως τελευταία πειραματική πράξη και έχοντας επιλέξει τα καλύτερα μοντέλα νευρωνικών δικτύων από τον Πίνακα 5.5, δοκιμάστηκαν στο να προβλέψουν την αλληλουχία των γεγονότων της συνεδρίας 8.

	Λανθασμένες προβλέψεις		
επαναλήψεις, κρυφά επίπεδα	3 HMM καταστάσεις	5 HMM καταστάσεις	10 HMM καταστάσεις
5-5-1	22.99%	22.99%	23.94%
2-2-1	24.95%	20.58%	23.94%
3-2-1	22.99%	20.93%	22.99%
ΠΟΣΟΣΤΟ ΛΑΘΟΥΣ		31498/149991	
FMR(%)		20.58	

Πινακας 5.7: Ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου, με χρήση βαθιών νευρωνικών δικτύων.

Τα παραπάνω αποτελέσματα μας οδηγούν στο πειραματικό συμπέρασμα ότι για το πρόβλημά μας τα βαθιά νευρωνικά δίκτυα και τα κρυφά μοντέλα Markov συμπεριφέρονται το ίδιο καθώς αγγίζουν το ίδιο ποσοστό λάθους FMR, δηλαδή 20.58%.

Κεφάλαιο 6

Συμπεράσματα

6.1 Συνεισφορά της διπλωματικής εργασίας

Μέσα από την εκπόνηση αυτής της διπλωματικής εργασίας πραγματοποιήθηκε συστηματική μελέτη της απόδοσης των βαθιών νευρωνικών δικτύων στο πρόβλημα της αναγνώρισης και κατηγοριοποίησης ακουστικών γεγονότων που λαμβάνουν χώρα σε ένα δωμάτιο συνεδρίασης. Επίσης, έγινε σύγκριση μεταξύ της απόδοσης του DNN ταξινομητή και των κρυφών μοντέλων Markov, αλλά και των δύο εργαλίων, HTK και Kaldi. Μέσα από τα πειραματικά αποτελέσματα που παρουσιάστηκαν στο προηγούμενο κεφάλαιο, η επιστημονική συμβολή της διπλωματικής καλύπτει τους εξείς άξονες:

- Τη μελέτη της πολυκαναλικής βάσης δεδομένων UPC-TALP που δημιουργήθηκε από την ηχογράφηση 12 διαφορετικών γεγονότων σε ένα έξυπνο περιβάλλον. Επίσης, έγινε επεξεργασία των ακουστικών δεδομένων με τη χρήση του λογισμικού SoX.

- Την εξαγωγή των κατάλληλων χαρακτηριστικών που θα μπορούσαν να περιγράψουν τα δεδομένα μας καλύτερα. Στην αρχή των πειραμάτων, όπως αναφέρουμε και στα προηγούμενα κεφάλαια, χρησιμοποιήθηκαν μόνο οι 13 συντελεστές MFCC, ενώ στη συνέχεια δοκιμάστηκαν και οι 39 (MFCC, D-MFCC, DD-MFCC). Ο συνδυασμός αυτός των χαρακτηριστικών ενίσχυσε κατά μεγάλο βαθμό τα μοντέλα εκπαίδευσής (είτε σε HMM είτε σε DNN).

- Πραγματοποιήθηκε διεξοδική μελέτη και εξέταση των δύο βασικών εργαλείων. Πιο ειδικά, και τα δύο εργαλεία έρχονται με κάποιες βασικές εντολές που αποτελούν γενικές περιπτώσεις προβλημάτων. Όπως παρουσιάζεται και στο Κεφάλαιο 4, έπρεπε να χρησιμοποιηθούν πολλές παραλλαγές εντολών ή και επιπλέον για να προσαρμοστούν οι υλοποιήσεις τους στο πρόβλημά μας.

- Τη χρήση διαφορετικών μετρικών εκτός των προκαθορισμένων, καθώς με τη μελέτη ακουστικών γεγονότων και όχι λέξεων, το word error rate δεν είχε υπόσταση. Έτσι, έγινε χρήση της μετρικής *ποσοστό λάθους ταξινόμησης ακουστικών γεγονότων* (για

μεμονομένα συμβάντα) και *ποσοστό λάθους ακολουθίας γεγονότων σε επίπεδο πλαισίου*. Με τον ορισμό αυτών των μετρικών, μπορέσαμε να αξιολογήσουμε καλύτερα τους ταξινομητές σε διαφορετικές φάσεις του προβλήματος. Η μετρική FMR είναι αυτή που παρουσίασε ιδιαίτερο ενδιαφέρον αφού εξέταζε τη συνεδρία 8 σε λεπτομερές επίπεδο χρόνου. Σχετικά με την τελευταία μετρική, το Kaldi μας έδωσε λόγο λάθους 20% με τη χρήση HMM, ενώ 21% με βαθιά νευρωνικά δίκτυα. Εδώ πρέπει να σχολιαστεί ότι το HTK έδωσε ποσοστό λάθους FMR ίσο με 15%. Παρόλο που και τα δύο εργαλεία χρησιμοποιούν τους ίδιους αλγορίθμους εκπαίδευσης και αποκωδικοποίησης, αυτή η διαφοροποίηση στο αποτέλεσμα πιθανόν οφείλεται στο διαφορετικό σχεδιασμό των εργαλείων παρά στην υλοποίηση των αλγορίθμων.

- Τέλος, η πιο σημαντική συνεισφορά της διπλωματικής αυτής ήταν τα πειράματα με διαφορετικές ρυθμίσεις για τα νευρωνικά δίκτυα ώστε να μελετηθεί καλύτερα η συμπεριφορά τους. Όπως αναφέρουν και στην εργασίας τους οι Hugo Larochelle, Yoshua Bengio κ.α [17], η στρατηγική εκμάθησης του αλγορίθμου οπισθοδιάδοσης δίνει μηικανοποιητικές λύσεις για νευρωνικά δίκτυα με 3 ή παραπάνω κρυφά επίπεδα. Αυτή τους τη δήλωση την επικυρώσαμε καθώς στο Kaldi τα καλύτερα αποτελέσματα στα νευρωνικά δίκτυα επήλθαν με τη χρήση ενός κρυφού επιπέδου (δες τέλος Κεφαλαίου 5).

Σχολιάζοντας τα αποτελέσματα αυτά, φαίνεται ότι τα μοντέλα Markov συμπεριφέρονται λίγο καλύτερα στο πρόβλημά μας, χωρίς όμως η διαφορά αυτή να είναι ουσιαστική.

6.2 Μελλοντικές ερευνητικές κατευθύνσεις

Παρόλο που με τη διεκπεραίωση της παρούσας διπλωματικής εργασίας η μελέτη και τα αποτελέσματα των βαθιών νευρωνικών δικτύων σαν ταξινομητές ήταν ικανοποιητικά, μπορούν να ακολουθηθούν επιπρόσθετοι βελτιωτικοί οδοί σε μελλοντικές έρευνες. Μερικές από αυτές παρουσιάζονται παρακάτω:

- Με σκοπό τη βελτίωση των ακουστικών χαρακτηριστικών θα μπορούσε να γίνει περαιτέρω έρευνα πάνω στη χρήση και εφαρμογή των μη-φασματικών χαρακτηριστικών με θετικό συνελικτικό πίνακα παραγοντοποίησης (convolutive non-negative matrix factorization ή NMF). Με βάση την εργασία των Courtenay V. Cotton και Daniel P. W. Ellis, ένα σύστημα εντοπισμού ακουστικών γεγονότων αποδίδει πολύ καλύτερα όταν χρησιμοποιηθεί συνδυασμός των NMF και MFCC χαρακτηριστικών παρά των MFCC μόνο [24]. Επιπλέον, με σκοπό την αποσυσχέτιση της λογαριθμικής ενέργειας στις συστοιχίες φίλτρων (logarithmic filter bank energies ή logFBE), έχουν μελετηθεί τα χαρακτηριστικά φιλτραρίσματος συχνότητας που παρουσιάζουν καλά αποτελέσματα σε μοντέλα Markov [25].

- Ένα άλλο πεδίο μελλοντικής έρευνας αποτελεί ο επαναπροσδιορισμός των αρχείων ετικετών της βάσης δεδομένων UPC-TALP, με σκοπό την επαναξέταση της απόδοσης του συστήματος βαθιών νευρωνικών δικτύων σε αυτή. Συγκεκριμένα, τα γεγονότα που συμβαίνουν στη συνεδρία 8 είναι περισσότερα από τα καταγεγραμμένα και λειτουργούν σαν θόρυβος παρασκηνίου.

- Ακόμη, ένα πεδίο που πρέπει να επικεντρωθούμε είναι ο συνδυασμός των καναλιών, είτε στο στάδιο της εκπαίδευσης του ταξινομητή, είτε στη φάση της κατηγοριοποίησης. Δηλαδή, κάποια γεγονότα μπορεί να έχουν ηχογραφηθεί καλύτερα από κάποιο άλλο μικρόφωνο. Αυτό θα ενίσχυε αρκετά το κάθε μοντέλο-κλάση κατά τη διαδικασία της εκμάθησης του ταξινομητή. Με αυτόν τον τρόπο θα αποφεύγαμε την υπερ-εκπαίδευση, καθώς δεν θα χρησιμοποιούνταν ολόκληρες οι ηχογραφήσεις και απο τα 24 μικρόφωνα.

- Τέλος, κάθώς η βάση δεδομένων εμπεριέχει και οπτικό υλικό θα μπορούσε να πραγματοποιηθεί συνδυασμός της οπτικοακουστικής πληροφορίας. Δηλαδή να επιτευχθεί συνδυασμός τροπικοτήτων (modality fusion) με σκοπό την αύξηση της ακρίβειας των βαθιών νευρωνικών δικτύων σαν ταξινομητές. Παραδείγματα τέτοιων πειραμάτων εμφανίζονται στα πειράματα των Samira Ebrahimi Kahou1, Christopher Pal κ.α όπου συνδύασαν βαθιά συνελικτικά νευρωνικά δίκτυα (deep convolutional neural networks), DNN και SVM με σκοπό την αναγνώριση κίνησης σε βίντεο [26].

Bibliography

[1] Taras Butko, Fran González Pla, Carlos Segura, Climent Nadeu, and Javier Hernando. Two-source acoustic event detection and localization: Online implementation in a smart-room. In *Signal Processing Conference, 2011 19th European*, pages 1317–1321. IEEE, 2011.

[2] Deep neural network figure. RSIP Vision, Global Leaders in Image Processing and Computer Vision. URL http://www.rsipvision.com/exploring-deep-learning. [Online; accessed 28-August-2015].

[3] Beth Logan et al. Mel Frequency Cepstral Coefficients for Music Modeling. In *ISMIR*, 2000.

[4] Hidden Markov Model figure. Factor graphs: HMM. URL http://www.igi.tugraz.at/lehre/MLA/WS13/MLA_Exercises_2013/node6.html. [Online; accessed 29-August-2015].

[5] Taras Butko, Climent Nadeu Camprubí, et al. Detection of overlapped acoustic events using fusion of audio and video modalities. 2010.

[6] Andrey Temko, Robert Malkin, Christian Zieger, Dusan Macho, Climent Nadeu, and Maurizio Omologo. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. *Cough*, 65(48):5, 2006.

[7] Panagiotis Giannoulis, Gerasimos Potamianos, Athanasios Katsamanis, and Petros Maragos. Multi-microphone fusion for detection of speech and acoustic events in smart spaces. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pages 2375–2379. IEEE, 2014.

[8] Xiaodan Zhuang, Xi Zhou, Mark A Hasegawa-Johnson, and Thomas S Huang. Real-world acoustic event detection. *Pattern Recognition Letters*, 31(12):1543–1551, 2010.

[9] David H Hubel MD John Franklin et al. *Brain and Visual Perception: The Story of a 25-Year Collaboration*. Oxford University Press, 2004.

[10] Steven B Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

[11] Bertrand Denis, Jean Côté, and René Laprise. Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (dct). *Monthly Weather Review*, 130(7):1812–1829, 2002.

[12] Sergios Theodoridis, Aggelos Pikrakis, Konstantinos Koutroumbas, and Dionisis Cavouras. *Introduction to Pattern Recognition: A Matlab Approach*. Academic Press, 2010.

[13] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

[14] Yihua Chen and Maya R Gupta. EM demystified: An expectation-maximization tutorial. *University of Washington*, 2010.

[15] Sean Borman. The expectation maximization algorithm - a short tutorial. *Submitted for publication*, pages 1–9, 2004.

[16] Yoav Freund and Robert E Schapire. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296, 1999.

[17] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. Exploring strategies for training deep neural networks. *The Journal of Machine Learning Research*, 10:1–40, 2009.

[18] A Temko, D Macho, C Nadeu, and C Segura. UPC-TALP database of isolated acoustic events. *Internal UPC report*, 85, 2005.

[19] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. *The HTK book*, volume 2. Entropic Cambridge Research Laboratory Cambridge, 1997.

[20] Arnab Ghoshal and Daniel Povey. The Kaldi Speech Recognition Toolkit.

[21] Kyle Gorman. OpenFst. URL http://www.openfst.org/twiki/bin/view/FST/WebHome. [Online; accessed 29-August-2015].

[22] KS Arunlal and Dr SA Hariprasad. An efficient viterbi decoder. *International Journal of Advanced Information Technology (IJAIT) Vol*, 2, 2012.

[23] Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C Gates. An empirical analysis of word error rate and keyword error rate. In *INTERSPEECH*, pages 2070–2073, 2008.

[24] Courtenay V Cotton and Daniel PW Ellis. Spectral vs. spectro-temporal features for acoustic event detection. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 69–72. IEEE, 2011.

[25] Pere Pujol Marsal, Susagna Pol, Astrid Hagen, Hervé Bourlard, and Climent Nadeu. Comparison and combination of RASTA-PLP and FF features in a hybrid HMM/MLP speech recognition system. In *INTERSPEECH*, 2002.

[26] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM, 2013.