Ανίχνευση Επικαλυπτόμενων Ακουστικών Γεγονότων
Overlapping Acoustic Event Detection

Διπλωματική Εργασία

Λύκα Θωμαή

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Πανεπιστήμιο Θεσσαλίας

Επιβλέποντες Καθηγητές

Γεράσιμος Ποταμιάνος, Αναπληρωτής Καθηγητής

Αντώνιος Αργυρίου, Επίκουρος Καθηγητής

Οκτώβριος 2016

Ανίχνευση Επικαλυπτόμενων Ακουστικών Γεγονότων

ΠΕΡΙΛΗΨΗ

Η ανάλυση της ακουστικής σκηνής βασίζεται στην αναγνώριση των γεγονότων, αλλά και στον εντοπισμό τους μέσα στο χώρο. Στις μέρες μας, παρόλο που τα συστήματα για αναγνώριση μεμονωμένων γεγονότων έχουν αναπτυχθεί αρκετά ώστε να ταξινομούν τα γεγονότα με ακρίβεια, στην περίπτωση των επικαλυπτόμενων γεγονότων η απόδοση των συστημάτων δεν είναι επαρκής. Στη παρούσα διπλωματική αναπτύχθηκε ένα σύστημα αναγνώρισης επικαλυπτόμενων ακουστικών γεγονότων, που βασίζεται στην ανίχνευση και τον εντοπισμό έως και δύο ακουστικών πηγών σε ένα έξυπνο δωμάτιο, δηλαδή σε έναν ειδικά εξοπλισμένο χώρο με πολλαπλές συστοιχίες μικροφώνων και κάμερες. Η διαδικασία της ανίχνευσης βασίζεται στη δημιουργία Κρυφών Μαρκοβιανών Μοντέλων, ενώ η διαδικασία του εντοπισμού βασίζεται στη πολυκαναλική επεξεργασία των σημάτων εισόδου με τη μέθοδο της Κατευθυντήριας Δύναμης Απόκρισης. Τα ποσοστά επιτυχίας αναγνώρισης και εντοπισμού τόσο των επικαλυπτόμενων όσο και των μεμονωμένων γεγονότων ύστερα από τη πειραματική ανάλυση των δεδομένων της βάσης UPC-TALP προσέγγισαν το 60% και 56% αντίστοιχα, ενώ με το συνδυασμό των αποτελεσμάτων των δύο μεθόδων αναπτύχθηκε ένα βελτιστοποιημένο σύστημα αναγνώρισης, όπου το ποσοστό επιτυχίας αυξήθηκε στο 81%. Τα ενθαρρυντικά αποτελέσματα του προτεινόμενου συνδυασμού μεθόδων αποτελεί το εφαλτήριο βήμα για την επέκταση της μεθόδου με στόχο τη δημιουργία ενός συστήματος για αναγνώριση πολλαπλών πηγών.

Overlapping Acoustic Event Detection

ABSTRACT

Acoustic source localization and acoustic event detection are the main tasks in acoustic scene analysis. Nowadays, the majority of research is focused on isolated acoustic event detection, and, as a result, there are many systems developed with high performance. On the other hand, research on overlapped events is still immature, and the existing systems don't yield good results. Motivated by these facts, the thesis is focused on implementing both an overlapped Acoustic Event Detection (AED) system and an overlapped Acoustic Source Localization (ASL) system on data collected inside the UPC's smart room, which detects up to two sources. The overlapped AED system is based on HMMs that operate using MFCC features, and the analysis is performed on a frame basis using the Viterbi decoder, while the ASL system is based on the Steered Response Power using the signals captured by the set of distant microphones available in the smart room. The experimental results from the AED and ASL systems showed a high success rate close to 60% and 56% respectively. By combining the results of the two methods, an optimized recognition system was developed with a success rate close to 81%. The encouraging results of the proposed system is the initial step to extend the method for creating a system that recognizes multiple acoustic sources.

ΕΥΧΑΡΙΣΤΙΕΣ

Με το τέλος της διπλωματικής μου εργασίας, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή, κ. Γεράσιμο Ποταμιάνο, αναπληρωτή καθηγητή του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών στο Βόλο, για την καθοδήγηση και τη βοήθεια που μου προσέφερε. Επιπρόσθετα, θα ήθελα να ευχαριστήσω την οικογένειά μου για την αγάπη και την υποστήριξη σε όλη τη διάρκεια των προπτυχιακών μου σπουδών. Ένα μεγάλο ευχαριστώ οφείλω στην αδερφή μου Ερασμία, που σαν μεγάλη αδερφή, παρά την απόσταση, ήταν πάντα δίπλα μου για να με υποστηρίζει και να με συμβουλεύει. Τέλος, ευχαριστώ όλους τους φίλους μου για την αλληλοϋποστήριξη, τη συνεργασία, αλλά κυρίως για τις όμορφες στιγμές που περάσαμε αυτά τα χρόνια.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΕΦΑΛΑΙΟ 1–ΕΙΣΑΓΩΓΗ

1.1 Αναγνώριση Ακουστικών Γεγονότων

Ο ήχος είναι ο δεύτερος πιο σημαντικός τρόπος με τον οποίο οι άνθρωποι αισθάνονται και κατανοούν το κόσμο. Πρόκειται για μία πηγή πλούσια σε πληροφορίες και για αυτό το λόγο χρησιμοποιείται σε πολλές εφαρμογές, βοηθώντας για παράδειγμα στην αναγνώριση και τη κατηγοριοποίηση όλων των ήχων ενός δωματίου σε μία συγκεκριμένη χρονική στιγμή. Η αναγνώριση ακουστικών γεγονότων αποτελεί κεντρικό θέμα στο κομμάτι της Υπολογιστής Ανάλυσης Ακουστικής Σκηνής (Computational Auditory Scene Analysis-CASA) που εντάσσεται στο κλάδο της Αναγνώρισης Προτύπων. Οι εφαρμογές στις οποίες είναι απαραίτητη η χρήση της αναγνώρισης ακουστικών γεγονότων είναι πολυάριθμες, με πιο γνωστές να αποτελούν τα συστήματα παρακολούθησης και φύλαξης χώρων, τα έξυπνα σπίτια, και η ρομποτική. Η ανθρώπινη δραστηριότητα που λαμβάνει χώρα είτε σε εσωτερικό είτε σε εξωτερικό χώρο, αντικατοπτρίζεται σε μία πληθώρα ακουστικών γεγονότων, τα οποία είτε παράγονται από το ανθρώπινο σώμα (ομιλία, βήχας) είτε παράγονται από τη χρήση κάποιων αντικειμένων (τηλέφωνο, χτύπημα πόρτας). Με αυτό τον τρόπο, τα συστήματα αναγνώρισης ακουστικών γεγονότων αποσκοπούν στον προσδιορισμό της ταυτότητας των ήχων, της χρονικής τους θέσης στο ακουστικό σήμα, αλλά και στον εντοπισμό της θέσης τους στο χώρο.

1.2 Το Πρόβλημα των Επικαλυπτόμενων Ακουστικών Γεγονότων

Στις μέρες μας, η τεχνολογία για την αναγνώριση και την ταξινόμηση μεμονωμένων ακουστικών γεγονότων έχει αναπτυχθεί αρκετά ώστε να αναγνωρίζει τα γεγονότα με πολύ μεγάλη ακρίβεια. Το πρόβλημα δημιουργείται στην αναγνώριση επικαλυπτόμενων ακουστικών γεγονότων, όπου η ακρίβεια των ταξινομητών μειώνεται πάρα πολύ. Στα πειράματα που έγιναν κατά τη διάρκεια του προγράμματος CLEAR 2007 σχετικά με την αναγνώριση ακουστικών γεγονότων σε χώρους σεμιναρίων, παρατηρήθηκε ότι το μεγαλύτερο ποσοστό λάθους αναγνώρισης προερχόταν από την λανθασμένη αναγνώριση των επικαλυπτόμενων ακουστικών γεγονότων [1]. Με τον όρο επικαλυπτόμενα ακουστικά γεγονότα εννοούμε όταν ένα ακουστικό γεγονός έχει χρονική επικάλυψη με άλλα ακουστικά γεγονότα, δηλαδή κατά τη διάρκεια ενός ακουστικού γεγονότος, υπάρχουν ταυτόχρονα και άλλα ηχητικά γεγονότα που λαμβάνουν χώρα. Η επικάλυψη ακουστικών γεγονότων έχει παρατηρηθεί ως χαρακτηριστικό σε πολλά γεγονότα της πραγματικής ζωής και συχνά χαρακτηρίζεται ως “cocktail party” πρόβλημα [2]. Σε αυτό το πρόβλημα γίνεται προσπάθεια για να διαχωριστεί

- Μια πηγή ομιλίας από μία άλλη πηγή
- Μια πηγή ομιλίας από ένα ακουστικό γεγονός
- Ένα ακουστικό γεγονός από ένα άλλο ακουστικό γεγονός

1.3 Παρόμοιες Εργασίες

Για την περίπτωση της αναγνώρισης και της ταξινόμησης μεμονωμένων ακουστικών γεγονότων, βασισμένων σε Κρυφά Μαρκοβιανά Μοντέλα, έχει γίνει μεγάλη έρευνα και έχουν επιτευχθεί πολύ ικανοποιητικές αποδόσεις. Αντίστοιχες μέθοδοι που αντιμετωπίζουν το πρόβλημα των επικαλυπτόμενων ακουστικών γεγονότων περιλαμβάνονται στην διατριβή του A. Temko [1] με τίτλο "Acoustic event detection and classification". Επιπλέον, οι [3] χρησιμοποίησαν φασματογράφημα και το γενικευμένο μετασχηματισμό Hough για να αναγνωρίσουν τα επικαλυπτόμενα ακουστικά γεγονότα. Σημαντική ήταν και η προσπάθεια των [4] που βασίστηκε σε Κρυφά Μαρκοβιανά Μοντέλα με πολλαπλά μονοπάτια αποκωδικοποίησης. Οι [5] πρότειναν τη χρήση μη-αρνητικών πινάκων παραγοντοποίησης (non-negative matrix factorization - NMF). Σημαντική ήταν η συμβολή του [6] με την ανίχνευση επικαλυπτόμενων ακουστικών γεγονότων βάση της οπτικής συσχέτισής τους. Τέλος, ως μέλος της πρόκλησης DCASE, οι [7] πρότειναν ένα σύστημα αναγνώρισης βασισμένο στο πλαίσιο sparse-CNMF. Για την περίπτωση του εντοπισμού των ακουστικών πηγών, δημιουργήθηκε ένα σύστημα από τους [8] που βασίζεται στην κατευθυντήρια δύναμη απόκρισης χρησιμοποιώντας το μετασχηματισμό φάσης, ενώ οι M. Omologo και P. Svaizer [9] ανέπτυξαν ένα σύστημα εντοπισμού βασισμένο στο φάσμα ισχύος κάθε σημείου του χώρου.

1.4 Σκοπός της Διπλωματικής

Η ανάλυση της ακουστικής σκηνής με σκοπό την ανίχνευση και τον εντοπισμό των ακουστικών γεγονότων είναι ένα από τα βασικά προβλήματα που βρίσκονται υπό έρευνα στο κλάδο της Αναγνώρισης Ακουστικών Γεγονότων. Σκοπός αυτής της διπλωματικής είναι να μελετηθεί το θέμα της Αναγνώρισης και του Εντοπισμού των Ακουστικών Γεγονότων σε έξυπνα δωμάτια. Βασισμένοι στο γεγονός ότι το πρόβλημα της επικάλυψης ακουστικών γεγονότων παραμένει, η κύρια προσπάθεια στη διπλωματική έγκειται στη ανάπτυξη ενός συστήματος για την αναγνώριση αυτών των γεγονότων. Για αυτό το λόγο δημιουργήθηκε ένα σύστημα αναγνώρισης βασισμένο στα κρυφά μοντέλα Markov όπου η ακουστική ανάλυση γίνεται σε επίπεδο χρονικών πλαισίων, η εκπαίδευση γίνεται με Γκαουσιανά μοντέλα μίξης, και η αναγνώριση πραγματοποιείται με τον αλγόριθμο Viterbi. Επιπλέον, αναπτύχθηκε ένα σύστημα που ανιχνεύει τις ηχητικές πηγές που υπάρχουν στο δωμάτιο. Το έργο για τον εντοπισμό των πηγών δυσκολεύει λόγω του θορύβου και της αντήχησης, αλλά κυρίως λόγω των πολλαπλών πηγών. Για αυτό το λόγο, το σύστημά μας βασίστηκε στην τεχνική του αλγορίθμου της Κατευθυντήριας Δύναμης Απόκρισης (Steered Response Power ή SRP), καθώς το βασικό χαρακτηριστικό του είναι η ανοχή σε αντήχηση και η ανεξαρτησία του από τον προσανατολισμό των ηχείων. Τα πειράματα πραγματοποιήθηκαν πάνω στη πολυκαναλική βάση δεδομένων UPC-TALP με βάση τα σήματα που λαμβάνονται από τα μικρόφωνα και με την υπόθεση ότι υπάρχουν έως και δύο πιθανές ακουστικές πηγές σε κάθε έξυπνο δωμάτιο. Με λίγα λόγια, η συνεισφορά αυτής της διπλωματικής μπορεί να συνοψιστεί στα εξής σημεία:

- Περαιτέρω έρευνα και μελέτη στον κλάδο της Αναγνώρισης Ακουστικών Γεγονότων
- Εκμάθηση συγκεκριμένων εργαλείων που χρησιμοποιούνται σε αυτό το κλάδο

- Δημιουργία συστήματος που να προσαρμόζεται εύκολα ώστε να λειτουργεί αποδοτικά και σε άλλα δεδομένα
- Πειράματα με διαφορετικές παραμέτρους στα εργαλεία HTK και Matlab
- Χρησιμοποίηση διαφορετικών μετρικών λάθους για την εύρεση του αποδοτικότερου μοντέλου

1.5 Οργάνωση της Διπλωματικής

Η διπλωματική χωρίζεται σε 6 Κεφάλαια σχετικά με το θέμα της αναγνώρισης και του εντοπισμού των επικαλυπτόμενων ακουστικών γεγονότων, όπου το κάθε κεφάλαιο επικεντρώνεται σε ένα συγκεκριμένο θέμα. Ακολουθεί μία σύντομη περιγραφή για το τι καλύπτει κάθε κεφάλαιο.

- Το ***Κεφάλαιο 2*** περιέχει όλη τη θεωρία για τη μοντελοποίηση του προβλήματος και το μαθηματικό υπόβαθρο. Περιγράφεται η εξαγωγή χαρακτηριστικών σύμφωνα με τους φασματικούς συντελεστές συχνότητας Mel, η θεωρία πίσω από τα Κρυφά Μαρκοβιανά μοντέλα, καθώς και τα Γκαουσιανά μοντέλα μίξης που χρησιμοποιήθηκαν για την εκπαίδευση των δεδομένων. Παρουσιάζεται ο αλγόριθμος Κατευθυντήριας Δύναμης Απόκρισης με το μετασχηματισμό φάσης (Steered Response Power PHAse Transform ή SRP-PHAT) που χρησιμοποιήθηκε για την εκτίμηση των κατευθύνσεων άφιξης και ο αλγόριθμος Στοχαστικής Συστολής Περιοχής (Stochastic Region Contraction ή SRC) που χρησιμοποιήθηκε για την βελτίωση των αποτελεσμάτων σχετικά με τον εντοπισμό των ακουστικών πηγών.
- Το ***Κεφάλαιο 3*** περιέχει τις βασικές πληροφορίες σχετικά με την βάση δεδομένων που χρησιμοποιήθηκε για τα πειράματά μας. Πρόκειται για την πολυκαναλική βάση δεδομένων UPC-TALP, όπου παρουσιάζονται όλες οι τεχνικές της πληροφορίες σχετικά με το πως δημιουργήθηκε και οργανώθηκε καθώς και η γεωμετρία του χώρου.
- Το ***Κεφάλαιο 4*** περιέχει αναλυτικά όλη τη διαδικασία για την εκτέλεση των πειραμάτων. Περιγράφει τα εργαλεία που χρησιμοποιήθηκαν για τη δημιουργία του κάθε συστήματος, καθώς και τη διαδικασία για την επεξεργασία των δεδομένων σε κάθε περίπτωση.
- Το ***Κεφάλαιο 5*** περιέχει τις μετρικές αξιολόγησης που ακολουθήθηκαν κατά τη διάρκεια των πειραμάτων, καθώς και τα αποτελέσματα που εξήχθησαν από τα διάφορα πειράματα. Τα αποτελέσματα αναλύονται τόσο γραπτά όσο και με γραφήματα για την καλύτερη κατανόησή τους.
- Το ***Κεφάλαιο 6*** κλείνει τη διπλωματική με συμπεράσματα καθώς και κάποιες μελλοντικές ερευνητικές κατευθύνσεις για την συγκεκριμένη περιοχή έρευνας με σκοπό τη βελτίωση των αποτελεσμάτων.

ΚΕΦΑΛΑΙΟ 2-ΘΕΩΡΗΤΙΚΟ ΠΛΑΙΣΙΟ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

2.1 Φασματικοί Συντελεστές Συχνότητας Mel

Το πρώτο βήμα στα συστήματα για Αυτόματη Αναγνώριση Ομιλίας (ASR) είναι η εξαγωγή χαρακτηριστικών, ο προσδιορισμός δηλαδή των συστατικών του ηχητικού σήματος που είναι καλά για ανάλυση και η απόρριψη των κομματιών με θόρυβο. Τα περισσότερα, λοιπόν, συστήματα βασίζουν την εξαγωγή χαρακτηριστικών στους Φασματικούς Συντελεστές Συχνότητας Mel ή MFCC, οι οποίοι έχουν αποδειχθεί εξαιρετικά αποτελεσματικοί και αξιόπιστοι [10]. Αυτή η μέθοδος εξαγωγής χαρακτηριστικών αναφέρθηκε για πρώτη φορά από τους Bridle και Brown το 1974 και αναπτύχθηκε περαιτέρω από τον Mermelstein το 1976, ενώ τα πειράματα βασίστηκαν πάνω στον ανθρώπινο λόγο [10] . Μιας και οι άνθρωποι αντιλαμβάνονται πιο εύκολα τις χαμηλές συχνότητες απ' ότι τις υψηλές , δημιουργήθηκε η κλίμακα Mel , όπου σχετίζει έναν ήχο ή μία αντιληπτή συχνότητα με την πραγματική της συχνότητα και τα διανύσματα χαρακτηριστικών αντιπροσωπεύουν αυτή την ιδιότητα. Η μετατροπή της συχνότητας από την κλίμακα Hertz στην κλίμακα Mel γίνεται με τον εξής τύπο :

$$\boldsymbol{M(f) = 1125 \ln(1 + f/700)}$$

ενώ η αντίστροφη διαδικασία γίνεται με τον τύπο :

$$\boldsymbol{\mathrm{M}^{-1}(m) = 700(e^{m/1125} - 1)}$$

όπου $\boldsymbol{f}$ είναι η συχνότητα σε Hertz και $\boldsymbol{m}$ η συχνότητα σε Mel. Τα βήματα για τον υπολογισμό των Φασματικών Συντελεστών Mel παρουσιάζονται στο Σχήμα 1 και περιγράφονται παρακάτω:

1. Διαίρεση σήματος σε επικαλυπτόμενα πλαίσια.

Στόχος της εξαγωγής χαρακτηριστικών είναι να παρέχει τα φασματικά χαρακτηριστικά από ένα μικρό παράθυρο ομιλίας, στο οποίο μπορούμε να κάνουμε την παραδοχή ότι το σήμα παραμένει στάσιμο. Αυτό το πετυχαίνουμε χρησιμοποιώντας ένα παράθυρο, το οποίο είναι μη μηδενικό σε μία συγκεκριμένη εσωτερική περιοχή και μηδενικό οπουδήποτε αλλού. Ο πιο συνηθισμένος τύπος παραθύρου που χρησιμοποιείται είναι ο ορθογώνιος, αλλά κόβει απότομα το σήμα στα όριά του και προκαλεί προβλήματα. Για αυτό το λόγο, χρησιμοποιούμε το παράθυρο πλαισίου Hamming με το μέγεθός του να είναι στα 25ms και μετατοπισμένο κατά 10ms , το οποίο στα όρια του παραθύρου συρρικνώνει τις τιμές του σήματος κοντά στο μηδέν ώστε να μην υπάρχουν ασυνέχειες [11]. Η εξίσωση για το παράθυρο Hamming έχει ως εξής :

$$w[n] = \begin{cases} 0.54 - 0.46\cos\dfrac{2\pi n}{L}, & 0 \le n \le L-1 \\ 0, & otherwise \end{cases}$$

2. Διακριτός Μετασχηματισμός Fourier (Discrete Fourier Transform ή DFT)

Το επόμενο βήμα είναι να γνωρίζουμε πόση ενέργεια περιέχει το σήμα σε διαφορετικές ζώνες συχνοτήτων. Το εργαλείο για την εξαγωγή αυτής της πληροφορίας είναι ο διακριτός μετασχηματισμός Fourier, ο οποίος ορίζεται ως εξής :

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\frac{\pi}{N}kn}$$

3. Εφαρμογή του φίλτρου Mel

Στη συνέχεια εφαρμόζεται το φίλτρο Mel στο φάσμα ισχύος και αθροίζεται η ενέργεια σε κάθε φίλτρο.

4. Υπολογισμός Λογαρίθμου όλων των Filterbank ενεργειών

Στο τέταρτο βήμα υπολογίζουμε το λογάριθμο των ενεργειών, γιατί πειράματα έχουν δείξει ότι οι άνθρωποι αντιλαμβάνονται την ένταση σε λογαριθμική κλίμακα, διότι η πράξη του λογαρίθμου επιτρέπει το διαχωρισμό διέγερσης/φίλτρου στο μοντέλο παραγωγής φωνής.

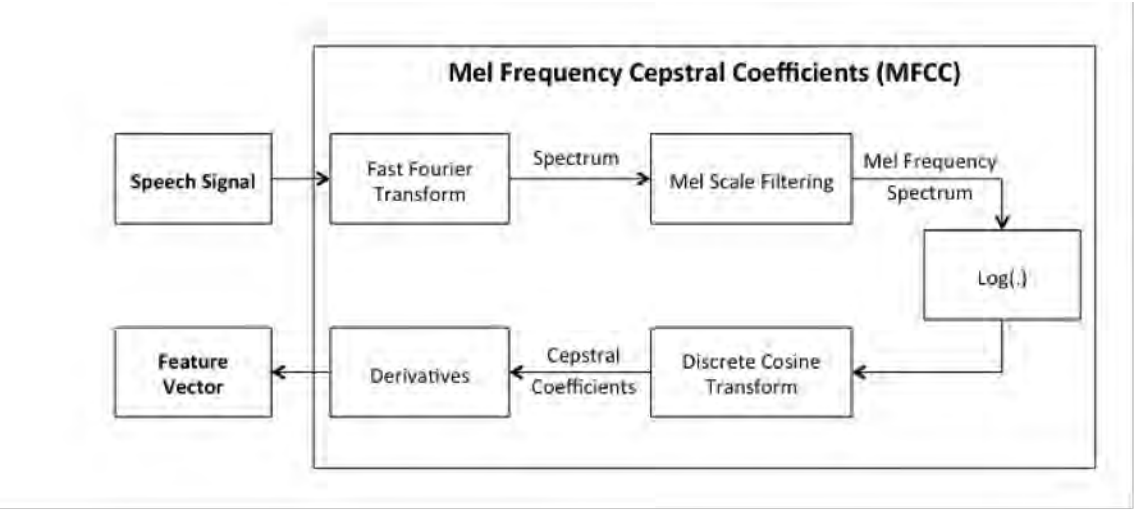
5. Υπολογισμός του Διακριτού Μετασχηματισμού Συνημιτόνου (Discrete Cosine Transform ή DCT)

Τελευταίο βήμα αποτελεί ο υπολογισμός του διακριτού μετασχηματισμού συνημιτόνου. Ο DCT χρησιμοποιείται σαν μέσο συμπίεσης της πληροφορίας του σήματος στους συντελεστές Mel και αποσυσχετίζει τις ενέργειες, επιτρέποντας την κατασκευή διαγώνιων πινάκων συνδιασποράς [12].

6. Το διάνυσμα χαρακτηριστικών που προκύπτει αποτελείται από τους συντελεστές 1 έως 13.
7. **Προαιρετικό βήμα** αποτελεί ο υπολογισμός των φασματικών συντελεστών της 1ης και 2ης παραγώγου ή αλλιώς Differential and Acceleration coefficients (Delta and Delta-Deltas), επεκτείνοντας το διάνυσμα. Για τον υπολογισμό των συντελεστών της 1ης παραγώγου (Delta) χρησιμοποιείται ο εξής τύπος:

$$d_t = \frac{\sum_{n=1}^{2} n(c_{t-n} - c_{t+n})}{2\sum_{n=1}^{2} n^2}$$

όπου d_t είναι ο συντελεστής 1ης παραγώγου για το πλαίσιο t που υπολογίζεται με βάση τους συντελεστές c_t που δημιουργήθηκαν στον βήμα 6. Οι συντελεστές 2η; παραγώγου υπολογίζονται με τον ίδιο τρόπο, χρησιμοποιώντας όμως τους συντελεστές 1ης παραγώγου.



Σχήμα 1 Διάγραμμα για τη δημιουργία των MFCC συντελεστών [13]

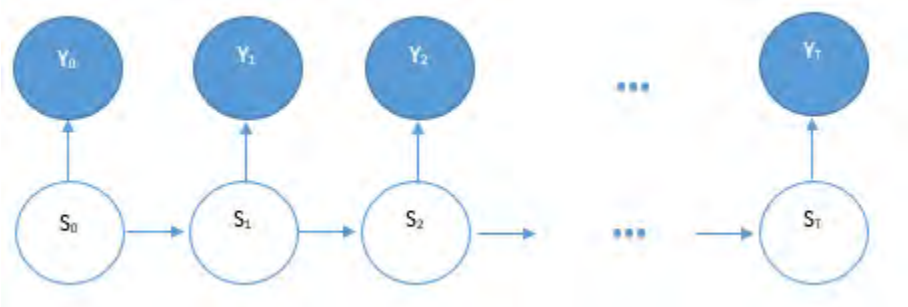
2.2 Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models-HMMs)

Τα κρυφά μοντέλα Markov αποτελούν το πιο διαδεδομένο εργαλείο για την μοντελοποίηση δεδομένων που αλλάζουν με τη πάροδο του χρόνου. Η θεωρία πίσω από τα HMMs αναπτύχθηκε στα τέλη του 1960 από τους Baum, Eagon, Petrie, Soules και Weiss [14], ενώ χρησιμοποιήθηκε για πρώτη φορά στην αναγνώριση ομιλίας στις αρχές του 1970 από τον Jim Baker στο Πανεπιστήμιο Carnegie-Mellon. Πλέον χρησιμοποιούνται σχεδόν σε όλα τα συστήματα αναγνώρισης ομιλίας, αλλά και σε εφαρμογές όπως η μοντελοποίηση ακολουθιών εικόνων και εντοπισμού αντικειμένων.

Ένα κρυφό μοντέλο Markov αντιπροσωπεύει την κατανομή πιθανότητας με βάση μία ακολουθία παρατηρήσεων. Πρόκειται για ένα πιθανοτικό μοντέλο, το οποίο μοντελοποιείται σύμφωνα με το απλό μοντέλο Markov αλλά με κρυφές καταστάσεις. Στα απλά μοντέλα Markov, η κάθε κατάσταση είναι απευθείας ορατή στον παρατηρητή, με αποτέλεσμα οι πιθανότητες μετάβασης να είναι η μοναδική παράμετρος. Στα κρυφά Μαρκοβιανά μοντέλα, η έξοδος, που εξαρτάται από την κάθε κατάσταση, είναι απευθείας ορατή, ενώ η κατάσταση δεν είναι απευθείας ορατή. Θεωρώντας ότι μία παρατήρηση συμβολίζεται με **Y** και η κατάσταση της με **S**, τότε την χρονική στιγμή t η παρατήρηση και η κατάσταση μπορούν να συμβολιστούν με $\mathbf{Y_t}$ και $\mathbf{S_t}$ αντίστοιχα.

Τα κρυφά μοντέλα Markov μπορούν να συνοψιστούν σε δύο κύρια σημεία. Πρώτον, μία παρατήρηση τη χρονική στιγμή t δημιουργήθηκε από μία διαδικασία στη κατάσταση $\mathbf{S_t}$, η οποία είναι κρυφή στον παρατηρητή. Δεύτερον, κάθε κατάσταση **S** του μοντέλου ακολουθεί τις ιδιότητες Markov, δηλαδή η τωρινή κατάσταση S_t εξαρτάται μόνο από την προηγούμενή της κατάσταση $\mathbf{S_{t-1}}$ και είναι ανεξάρτητη από τις προηγούμενες t-2 καταστάσεις. Αυτό σημαίνει ότι η κατάσταση σε μία χρονική στιγμή διαθέτει όλες τις πληροφορίες που χρειαζόμαστε για το ιστορικό της διαδικασίας, ώστε να μπορέσουμε να προβλέψουμε το μέλλον της. Το μοντέλο πιθανότητας ενός HMM φαίνεται στο Σχήμα 2, ενώ ο τύπος για την από κοινού κατανομή πιθανότητας της ακολουθίας καταστάσεων είναι ο εξής [15]:

$$\mathbf{P(S_{1:T}, Y_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t)}$$



Σχήμα 2 Μοντέλο πιθανότητας HMM

Με την δημιουργία ενός Κρυφού Μαρκοβιανού Μοντέλου, τρία προβλήματα πρέπει να λυθούν:

1. **Το πρόβλημα της Αξιολόγησης (Evaluation Problem):** έχοντας ως δεδομένο τα μοντέλα HMM και μία ακολουθία παρατηρήσεων, θέλουμε να υπολογίσουμε το πιο πιθανό μοντέλο που δημιούργησε αυτή την ακολουθία.
2. **Το πρόβλημα της Αποκωδικοποίησης (Decoding Problem):** η εύρεση της πιο πιθανής ακολουθίας κρυφών καταστάσεων, δεδομένης της ακολουθίας παρατηρήσεων. Για το σκοπό αυτό, χρησιμοποιείται ο αλγόριθμος Viterbi.
3. **Το πρόβλημα της Μάθησης (Learning Problem):** η εκτίμηση των παραμέτρων του HMM από μία ακολουθία παρατηρήσεων, δηλαδή τα δεδομένα εκπαίδευσης χρησιμοποιούνται για να υπολογιστούν οι πιθανότητες μετάβασης μεταξύ των καταστάσεων (states) και των πιθανοτήτων εκπομπής των παρατηρήσεων. Ο αλγόριθμος Forward-Backward χρησιμοποιείται για το σκοπό αυτό.

2.3 Γκαουσιανά Μοντέλα Μίξης

Τα Γκαουσιανά Μοντέλα Μίξης (Gaussian Mixture Model-GMM) είναι από τις πιο διαδεδομένες μεθόδους ομαδοποίησης και χρησιμοποιούνται συχνά ως μοντέλα για την κατανομή πιθανοτήτων. Πρόκειται για μία παραμετρική συνάρτηση πυκνότητας-πιθανότητας που υποθέτει ότι όλα τα δεδομένα παράγονται από μία μίξη ενός πεπερασμένου αριθμού Γκαουσιανών κατανομών. Μπορεί, επίσης, να θεωρηθεί ως γενίκευση του αλγορίθμου των k-Μέσων ενσωματώνοντας πληροφορίες σχετικά με την συνδιασπορά των δεδομένων [16]. Ένα GMM ορίζεται ως ένας κυρτός συνδυασμός Γκαουσιανών πυκνοτήτων. Μία Γκαουσιανή πυκνότητα σε ένα ***d***-διάστατο χώρο που χαρακτηρίζεται από το διάνυσμα μέσης τιμής $\boldsymbol{\mu}$ και έναν ***d x d*** πίνακα συνδιασποράς ***C,*** ορίζεται ως εξής:

$$\varphi(\boldsymbol{x};\boldsymbol{\theta}) = 2\pi^{-d/2} det(\boldsymbol{C})^{-1/2} exp(\frac{-(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2})$$

όπου το $\boldsymbol{\vartheta}$ υποδηλώνει τα $\boldsymbol{\mu}$ και $\boldsymbol{C}$.

Οι παράμετροι του GMM υπολογίζονται από τα δεδομένα εκπαίδευσης χρησιμοποιώντας τον επαναληπτικό αλγόριθμο Expectation-Maximization (EM). Ο αλγόριθμος πήρε το όνομά του από τα δύο στάδια που εκτελεί: το στάδιο E (αναμενόμενης τιμής), όπου δημιουργεί μία συνάρτηση για την εκτίμηση της λογαριθμικής πιθανότητας, που υπολογίζεται χρησιμοποιώντας την τρέχουσα τιμή των παραμέτρων, και το στάδιο M (μεγιστοποίηση), όπου προσπαθεί να εκτιμήσει τη μέγιστη πιθανοφάνεια της λογαριθμικής πιθανότητας. Ο αλγόριθμος ξεκινά από κάποια αρχική εκτίμηση του $\boldsymbol{\vartheta}$ και στη συνέχεια ενημερώνει επαναληπτικά το $\boldsymbol{\vartheta}$ μέχρι να βρεθεί σύγκλιση.

2.4 Συστοιχίες Μικροφώνων

Η επεξεργασία μιας συστοιχίας σημάτων περιλαμβάνει τη χρήση πολλαπλών αισθητήρων προκειμένου να ληφθεί ένα σήμα που διαδίδεται μέσω κυμάτων. Μια συστοιχία μικροφώνων αποτελείται από ένα σύνολο μικροφώνων, τα οποία είναι τοποθετημένα σε διαφορετικές θέσεις στο χώρο προκειμένου να λάβουν το σήμα από διαφορετικές κατευθύνσεις. Βασισμένοι στις αρχές διάδοσης

του ήχου, οι πολλαπλές είσοδοι μπορούν να υποστούν επεξεργασία για να ενισχύσουν ή για να εξασθενήσουν τα σήματα που έρχονται από συγκεκριμένες κατευθύνσεις. Η διπλωματική επικεντρώθηκε στη χρήση συστοιχιών μικροφώνων που λαμβάνουν ακουστικά σήματα ή πιο συγκεκριμένα σήματα ομιλίας και ακουστικών γεγονότων που μπορούν να λάβουν χώρα σε ένα δωμάτιο. Με τη βοήθεια κατάλληλων αλγορίθμων, τα σήματα από κάθε μικρόφωνο, επεξεργάστηκαν και συνδυάστηκαν για τη δημιουργία του σχηματισμού δέσμης (beamforming). Η διάταξη μιας συστοιχίας αποτελείται, στη συγκεκριμένη περίπτωση που εξετάζεται, από τέσσερα μικρόφωνα, όπου κάθε ένα από τα μικρόφωνα θα λάβει το σήμα σε διαφορετικό χρόνο από τα υπόλοιπα και είτε θα προηγείται χρονικά είτε θα είναι καθυστερημένο.

2.5 Σχηματισμός Δέσμης (Beamforming)

Για τον εντοπισμό των ηχητικών πηγών, οι περισσότερες εφαρμογές χρησιμοποιούν συστοιχίες μικροφώνων. Σε τέτοιες εφαρμογές, η ακριβής γνώση της θέσης των μικροφώνων είναι απαραίτητη για την αποτελεσματικότητα των συστημάτων. Οι υπάρχοντες διαδικασίες για τον εντοπισμό πηγών χωρίζονται σε τρεις κατηγορίες. Η πρώτη κατηγορία βασίζεται στη μεγιστοποίηση της ενέργειας από έναν σχηματισμό δέσμης, η δεύτερη κατηγορία βασίζεται στην εκτίμηση από την υψηλή ανάλυση του φάσματος και η τρίτη κατηγορία βασίζεται στη διαφορά άφιξης των σημάτων σε κάθε μικρόφωνο. Η παρούσα διπλωματική βασίζεται στη πρώτη κατηγορία του σχηματισμού δέσμης. Η διαδικασία του σχηματισμού δέσμης ή beamforming προσπαθεί να υπολογίσει τη θέση της ηχητικής πηγής μεγιστοποιώντας μία μετρική. Ο πιο απλός σχηματισμός δέσμης είναι με τη διαδικασία της καθυστέρησης και άθροισης (delay-and-sum), ενώ ένας πιο πολύπλοκος σχηματισμός δέσμης είναι με τη διαδικασία του φιλτραρίσματος και άθροισης (filter-and-sum) όπου εφαρμόζονται προσαρμοστικά φίλτρα στα σήματα πριν την άθροισή τους. Καθυστερώντας τις εξόδους των μικροφώνων και προσθέτοντας όλα τα σήματα των μικροφώνων, ενισχύεται το σήμα σε σχέση με το θόρυβο αλλά και με τα σήματα που έρχονται από διαφορετικές κατευθύνσεις [17]. Ένας beamformer χρησιμοποιείται για να κατευθύνει πάνω σε μία συγκεκριμένη περιοχή τα σήματα, θεωρώντας την περιοχή ως πιθανή πηγή. Το σημείο στο χώρο όπου προκύπτει η μέγιστη ισχύς από τον σχηματισμό δέσμης αποτελεί τη πιθανή προέλευση του ήχου.

Θεωρώντας ότι ο χώρος που εξετάζεται στη διπλωματική καλύπτεται από M μικρόφωνα, τότε υπάρχουν M διαφορετικές κατευθύνσεις άφιξης, όπου κάθε μία αποτελεί το άμεσο μονοπάτι από την ηχητική πηγή στο μικρόφωνο. Το σήμα $x_m(t)$ που καταλήγει σε κάθε μικρόφωνο m δίνεται από τον τύπο

$$x_m(n) = a_m s\big(n - t - f_m(\tau)\big) + w_m(n)$$

όπου a_m είναι οι παράγοντες εξασθένησης, εξαιτίας των φαινομένων διάδοσης στο μέσο, t είναι ο χρόνος διάδοσης από την άγνωστη πηγή s(n) στο μικρόφωνο 0, $w_m(n)$ είναι το σήμα προσθετικού θορύβου στο m-οστό μικρόφωνο, τ η σχετική χρονική καθυστέρηση μεταξύ των μικροφώνων 0 και 1 και $f_m(\tau)$ είναι η σχετική καθυστέρηση μεταξύ των μικροφώνων 0 και m. Σε μία συστοιχία M μικροφώνων, με τις κατάλληλες κατευθυντήριες καθυστερήσεις (steering delays) δ_m , η μέθοδος καθυστέρησης και άθροισης μπορεί να ευθυγραμμίσει στο χρόνο, και στη συνέχεια να αθροίσει όλα τα σήματα μαζί. Η διαδικασία μπορεί να οριστεί μαθηματικώς ως εξής:

$$y(t, \delta_1, \delta_2, \dots, \delta_M) = \sum_{m=1}^{m=M} x_m(t - \delta_m)$$

όπου οι κατευθυντήριες καθυστερήσεις μπορούν να ορισθούν ως

$$\delta_m = \tau_m - \tau_0$$

όπου τ_m είναι η χρονική καθυστέρηση από την πηγή στο μικρόφωνο m και τ_0 είναι η ελάχιστη χρονική καθυστέρηση από όλα τα ζευγάρια μικροφώνων $\tau_m, m = [1,2, \dots, M]$. Η χρονική καθυστέρηση (time traveling) του σήματος από την πηγή στο μικρόφωνο m υπολογίζεται σύμφωνα με τον τύπο $\tau_m = \frac{r_m}{c}$, όπου r_m είναι η απόσταση από το μικρόφωνο m στη πηγή και c η ταχύτητα του ήχου. Επομένως, η έξοδος του σχηματισμού δέσμης μπορεί να ορισθεί ως εξής:

$$y(t, \delta_1, \delta_2, \dots, \delta_M) = s(t) * \sum_{m=1}^{m=M} \alpha_m + \sum_{m=1}^{m=M} w_m(t - \tau_m - \tau_0)$$

Προσθέτοντας ένα προσαρμοστικό φίλτρο στο σχηματισμό δέσμης καθυστέρησης και άθροισης προκύπτει ο σχηματισμός δέσμης με φιλτράρισμα και άθροιση. Για την έξοδο με τη δεύτερη μέθοδο του σχηματισμού δέσμης χρησιμοποιείται ο τύπος :

$$Y(\omega, \delta_1, \delta_2, \dots, \delta_M) = \sum_{m=1}^{m=M} G_m(\omega) X_m(\omega) e^{-j\omega\delta_m} \quad (1)$$

όπου $X_m(\omega)$ είναι ο μετασχηματισμός Fourier του σήματος του μικροφώνου $x_m(t)$ και $G_m(\omega)$ είναι ο μετασχηματισμός Fourier του φίλτρου.

2.6 Η Μέθοδος Κατευθυντήριας Δύναμης Απόκρισης

Η μέθοδος για την εύρεση του σημείου με τη μέγιστη ισχύ εξόδου είναι γνωστή ως Κατευθυντήρια Δύναμη Απόκρισης (Steered Response Power- SRP). Το πιο σημαντικό χαρακτηριστικό της μεθόδου είναι η ικανότητα να συνδυάζει πολλαπλά σήματα ταυτόχρονα, αλλά και η δυνατότητα να επεξεργάζεται αποτελεσματικά μικρά πλαίσια δεδομένων. Επιπλέον, μπορεί να εντοπίζει πολλαπλές πηγές που συμβαίνουν ταυτόχρονα, καθώς ο σχηματισμός δέσμης θα μεγιστοποιεί την ισχύ σε πολλαπλά σημεία, τα οποία θα αντιστοιχούν στις θέσεις των πολλαπλών ομιλητών. Η κατευθυντήρια δύναμη απόκρισης αποτελεί την έξοδο του σχηματισμού δέσμης με φιλτράρισμα και άθροιση όταν το κατευθύνουμε ως προς όλα τα σημεία σε μία προκαθορισμένη περιοχή. Για κάθε σημείο υπάρχει η συνάρτηση των κατευθυντήριων καθυστερήσεων, που στο πεδίο της συχνότητας ορίζεται ως εξής:

$$P(\delta_1, \dots, \delta_M) = \int_{-\infty}^{+\infty} Y(\omega, \delta_1, \dots, \delta_M) Y^*(\omega, \delta_1, \dots, \delta_M) d\omega \quad (2)$$

Αντικαθιστώντας τις εξισώσεις (1) και (2) έχουμε:

$$P(\delta_1, \dots, \delta_M) = \int_{-\infty}^{+\infty} \left(\sum_{k=1}^{k=M} G_k(\omega) X_k(\omega) e^{-j\omega\delta_k}\right) \left(\sum_{l=1}^{l=M} G_l^*(\omega) X_l^*(\omega) e^{-j\omega\delta_l}\right) d\omega$$

Ενώ μεταθέτοντας τους όρους και γνωρίζοντας ότι $\delta_l - \delta_k = \tau_l - \tau_k$ προκύπτει η εξίσωση

$$P(\delta_1, \dots, \delta_M) = \int_{-\infty}^{+\infty} \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} (G_k(\omega) G_l^*(\omega))(X_k(\omega) X_l^*(\omega)) e^{-j\omega(\tau_l - \tau_k)} d\omega$$

Το παραπάνω ολοκλήρωμα συγκλίνει, διότι στον πραγματικό κόσμο τα σήματα των μικροφώνων καθώς και τα φίλτρα έχουν πεπερασμένη ενέργεια, επομένως τα αθροίσματα μπορούν να εναλλαχθούν με το ολοκλήρωμα ως εξής:

$$P(\delta_1, \dots, \delta_M) = \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} \int_{-\infty}^{+\infty} (G_k(\omega) G_l^*(\omega))(X_k(\omega) X_l^*(\omega)) e^{-j\omega(\tau_l - \tau_k)} d\omega$$

Ορίζοντας τη συνδυασμένη συνάρτηση βαρών $\Psi_{kl} = G_k(\omega) G_l^*(\omega)$, και δεδομένου ότι $\tau_l - \tau_k = \tau_{lk}$, αντικαθιστούμε στη παραπάνω εξίσωση και παίρνουμε την έκφραση για την κατευθυντήρια δύναμη απόκρισης:

$$P(\delta_1, \dots, \delta_M) = \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} \int_{-\infty}^{+\infty} \Psi_{kl}(\omega) X_k(\omega) X_l^*(\omega) e^{-j\omega\tau_{lk}} d\omega$$

Εφαρμόζοντας στη κατευθυντήρια δύναμη απόκρισης τη συνάρτηση βαρών του μετασχηματισμού φάσης (Phase transform), όπου $\Psi_{kl}(\omega) = \frac{1}{|X_k(\omega) X_l^*(\omega)|}$, παίρνουμε την κατευθυντήρια δύναμη απόκρισης με μετασχηματισμό φάσης (SRP-PHAT). Έτσι, για κάθε σημείο στο χώρο έχουμε :

$$P(\delta_1, \dots, \delta_M) = \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} \int_{=\infty}^{+\infty} \frac{1}{|X_k(\omega) X_l^*(\omega)|} X_k(\omega) X_l^*(\omega) e^{-j\omega(\tau_{lk})} d\omega \quad (3)$$

Επομένως, κατευθύνοντας το σχηματισμό δέσμης σε όλα τα πιθανά σημεία του χώρου μπορούμε να βρούμε τη θέση της πηγής. Τα σημεία που επιστρέφουν τη μεγαλύτερη τιμή της κατευθυντήριας δύναμης για το μετασχηματισμό φάσης είναι οι πιθανές θέσεις της ηχητικής πηγής. Έχοντας όλες τις πιθανές συντεταγμένες των σημείων στο χώρο, υπολογίστηκε η κατευθυντήρια δύναμη απόκρισης με μετασχηματισμό φάσης για κάθε σημείο. Το σημείο με τη μεγαλύτερη τιμή, δηλαδή για το σημείο που ισχύει

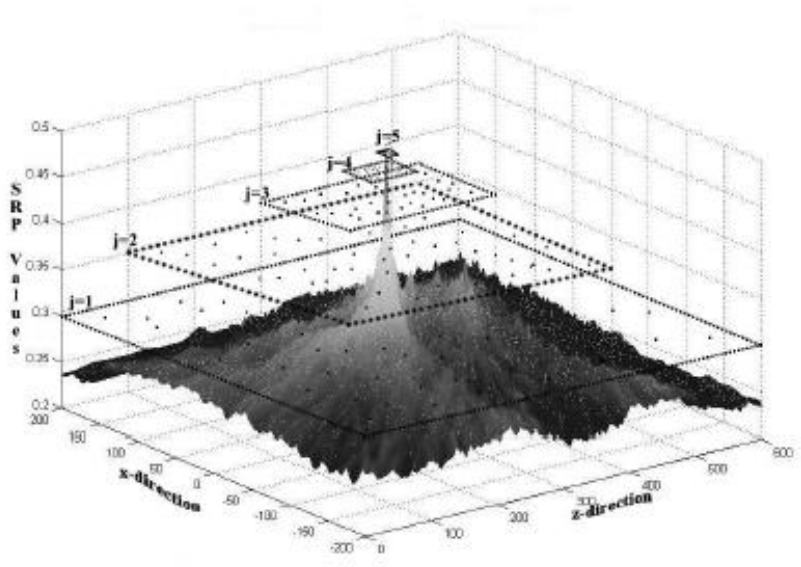
$$\underset{\vec{x}}{\operatorname{argmax}}\ P(\delta_1, \dots, \delta_M)$$

προκύπτει ότι αποτελεί και τη θέση της πηγής.

2.7 Αλγόριθμος Στοχαστικής Συστολής Περιοχής (Stochastic Region Contraction-SRC)

Η διαδικασία της βελτιστοποίησης μιας λειτουργίας χωρίς τη γνώση των μεταβλητών είναι ένα δύσκολο πρόβλημα. Οι λειτουργίες μπορεί να έχουν πολλές μεταβλητές και ταυτόχρονα πολλά τοπικά μέγιστα και ελάχιστα. Οι περισσότερες τεχνικές για βελτιστοποίηση είναι ακριβές ως προς τον αριθμό των πράξεων. Ο αλγόριθμος της Στοχαστικής Συστολής περιοχής δημιουργήθηκε για να λύσει έναν αριθμό από αυτά τα προβλήματα. Αναπτύχθηκε από τους Berger και Silverman [18] ως μία διαδικασία βελτιστοποίησης των σημάτων από τις συστοιχίες μικροφώνων. Για αυτό το λόγο αποτελεί ιδανική λύση στη βελτιστοποίηση της αναζήτησης της καλύτερης τιμής του SRP-PHAT, καθιστώντας τον αλγόριθμο πιο αποτελεσματικό σε πραγματικό χρόνο.

Η μέθοδος SRC βασίζεται στη σταδιακή μείωση του εύρους αναζήτησης κάθε μεταβλητής. Δίνοντας έναν αρχικό ορθογώνιο χώρο αναζήτησης που περιέχει τις επιθυμητές βέλτιστες τιμές, ξεκινάει μία επαναληπτική διαδικασία η οποία διαιρεί τον αρχικό χώρο σε μικρότερους, οι οποίοι περιέχουν τη μέγιστη τιμή. Σε κάθε επανάληψη επιλέγονται τυχαία σημεία στο δοθέν χώρο και από αυτά επιλέγονται μόνο όσα έχουν τιμή καλύτερη από το μέσο όρο της προηγούμενης επανάληψης. Μόλις βρεθεί ένας επαρκής αριθμός νέων λύσεων, τότε το μέγεθος του χώρου ενημερώνεται. Υπάρχει περίπτωση, κατά τη διάρκεια του αλγορίθμου η περιοχή αναζήτησης να επεκταθεί περιστασιακά αν βρεθεί μία καλή υποψήφια λύση, αλλά σε καμία περίπτωση δεν επιτρέπεται να επεκταθεί περισσότερο από τα αρχικά όρια. Η επαναληπτική διαδικασία επαναλαμβάνεται μέχρι είτε να πραγματοποιηθεί ο μέγιστος αριθμός επαναλήψεων είτε να δημιουργηθεί ο ελάχιστος χώρος προς αναζήτηση. Στο Σχήμα 3 φαίνεται η επαναληπτική διαδικασία του αλγορίθμου για τους διάφορους χώρους αναζήτησης που προκύπτουν μέχρι να βρεθεί η μέγιστη τιμή για το SRP-PHAT.



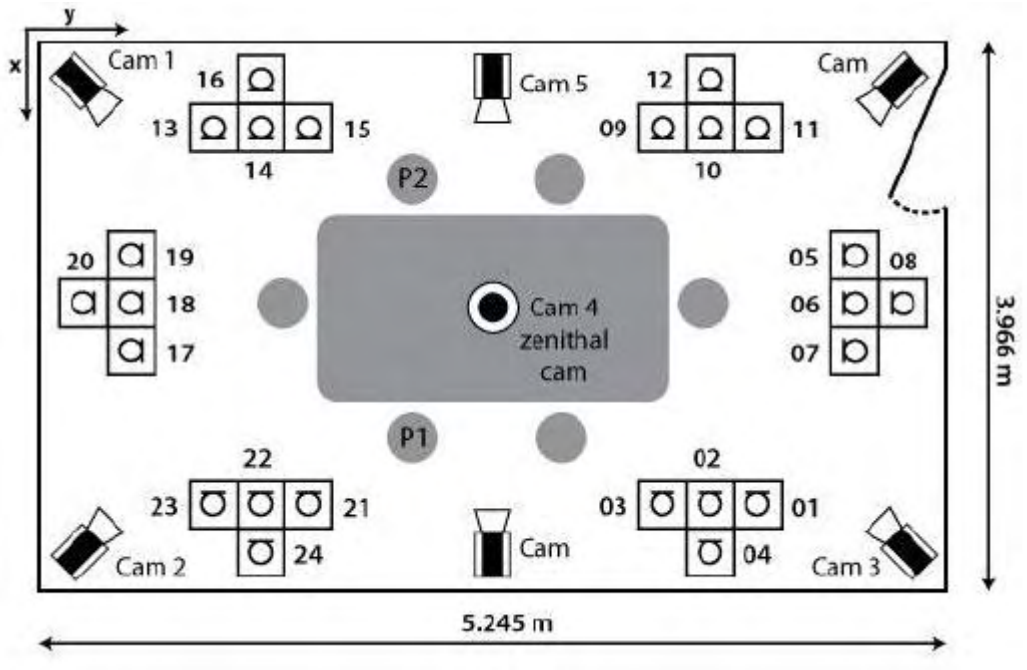
Σχήμα 3 Παράδειγμα αλγορίθμου SRC [18]

ΚΕΦΑΛΑΙΟ 3-ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ UPC-TALP

3.1 Βασικές Πληροφορίες

Η βάση δεδομένων που χρησιμοποιήθηκε στην διπλωματική δημιουργήθηκε κατά τη διάρκεια του CHIL Project (Computers in the Human Interaction Loop) στο πλαίσιο ενός ολοκληρωμένου προγράμματος κάτω από την εποπτεία της European Commission's Sixth Framework και περιέχει μια σειρά από ακουστικά γεγονότα που μπορούν να πάρουν μέρος σε ένα χώρο συνεδριάσεων [19]. Στο Σχήμα 4 παρουσιάζεται η κάτοψη του δωματίου καθώς και η κατανομή των μικροφώνων, όπου υπάρχουν 6 T-σχήματος ομάδες μικροφώνων και 6 κάμερες. Επιπλέον επισημαίνονται οι θέσεις P1 και P2, όπου βρίσκονται οι δύο συμμετέχοντες, ενώ τα υπόλοιπα ακουστικά γεγονότα (AEs) μπορούν να παραχθούν οπουδήποτε αλλού μέσα στο χώρο. Τα ηχητικά αρχεία που δημιουργήθηκαν ηχογραφήθηκαν από τα 24 μικρόφωνα που ήταν πλήρως συγχρονισμένα μεταξύ τους. Το σενάριο που εφαρμόστηκε στη διπλωματική αποτελείται από 3 διαφορετικά μοντέλα:

- Δεν υπάρχει καθόλου ακουστική δραστηριότητα.
- Υπάρχει μόνο μία ακουστική πηγή στο δωμάτιο.
- Υπάρχουν δύο ταυτόχρονες ακουστικές πηγές.



Σχήμα 4 Κάτοψη του UPC δωματίου [13]

3.2 Τεχνικές Πληροφορίες

Η βάση δεδομένων UPC-TALP περιέχει τόσο μεμονωμένα όσο και επικαλυπτόμενα ακουστικά γεγονότα στο χώρο συνεδριάσεων. Υπάρχουν 8 συνεδρίες με μεμονωμένα γεγονότα (S01-S08) και 9 συνεδρίες με επικαλυπτόμενα γεγονότα (T01-T09). Η δομή της βάσης έχει ως εξής:

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

Στην παρούσα διπλωματική χρησιμοποιηθήκαν μόνο τα ακουστικά αρχεία T01-T07 που παρήχθησαν από 6 διαφορετικά άτομα. Το όνομα κάθε ακουστικού αρχείου έχει τη μορφή «NAME_N.wv», όπου «NAME» είναι το όνομα της συνεδρίας και N ο αριθμός του μικροφώνου. Κάθε συνεδρία περιέχει ακουστικές ακολουθίες διάρκειας 5-6 λεπτών με 44100 Hz συχνότητα δειγματοληψίας και 24 bits ακρίβειας που παρατηρήθηκαν από 24 μικρόφωνα, ενώ αποθηκεύτηκαν σε μορφή *.wv . Καθώς χρησιμοποιούμε μόνο τα επικαλυπτόμενα ακουστικά αρχεία, στη βάση μας κατηγοριοποιήσαμε 38 κλάσεις ηχητικών γεγονότων, μεμονωμένων και επικαλυπτόμενων, όπως περιγράφεται στο Κεφάλαιο 4. Τέλος, εκτός από τα ακουστικά αρχεία, υπάρχει και ένα αρχείο επισημείωσης με την εξής μορφή «session_name.csv», όπου εμπεριέχονται όλα τα σύμβολα των ηχητικών γεγονότων μαζί με την ακριβή στιγμή έναρξης και λήξης τους.

ΚΕΦΑΛΑΙΟ 4-ΠΕΙΡΑΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

4.1 Ανίχνευση Επικαλυπτόμενων Ακουστικών Γεγονότων

4.1.1 SoX

Η ονομασία του προέρχεται από τα αρχικά του Sound Exchange. Πρόκειται για ένα λογισμικό κονσόλας που λειτουργεί μέσω γραμμών εντολών, δέχεται ακουστικά αρχεία διαφόρων τύπων και τα επεξεργάζεται. Προσφέρει στους χρήστες πολλές δυνατότητες, όπως να εφαρμόσει διάφορα εφέ στα αρχεία, να αλλάξει τον αριθμό των καναλιών, να παρέχει στατιστικές πληροφορίες του κομματιού, να αλλάξει το ρυθμό δειγματοληψίας, να συνενώσει πολλά ακουστικά αρχεία αλλά και να τμηματοποιήσει συγκεκριμένα σημεία ενός κομματιού.

Ως αναφορά την διπλωματική χρησιμοποιηθήκαν οι παρακάτω 4 δυνατότητες του SoX:

- Αλλαγή της μορφής αρχείου σε wav με την εντολή :
  sox input.wv output.wav

- Αλλαγή του ρυθμού δειγματοληψίας με την εντολή :
  sox input.wav -r 16000 output.wav

- Αναζήτηση πληροφορίες σχετικά με το ακουστικό αρχείο με την εντολή :
  sox filename.wav -n stat

- Διάσπαση ακουστικού αρχείου σε επιμέρους με την εντολή :
  sox input.wav output.wav trim start_point duration

Εφόσον οι εντολές έπρεπε να εκτελεστούν για κάθε αρχείο της βάσης δεδομένων δημιουργήθηκε script αρχείο που να εκτελεί τις παραπάνω εντολές επαναληπτικά για κάθε αρχείο εισόδου.

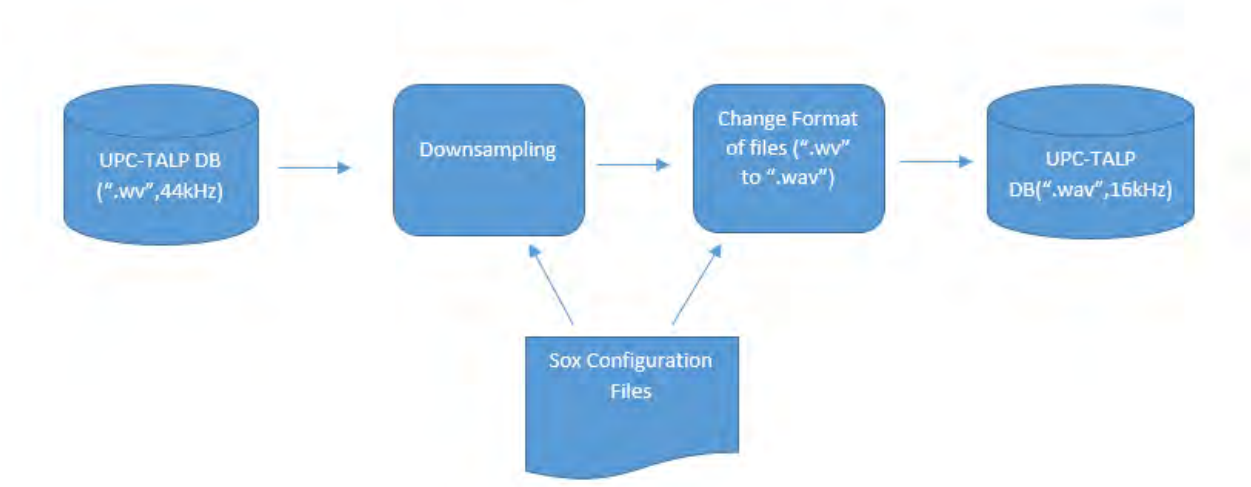
4.1.2 HTK

Η ονομασία του HTK προέρχεται από το Hidden Markov Model Toolkit. Πρόκειται για μία εργαλειοθήκη η οποία χρησιμοποιείται κυρίως για αναγνώριση ομιλίας. Αναπτύχθηκε στο εργαστήριο Μηχανικής Μάθησης του Πανεπιστημίου Cambridge από το τμήμα Μηχανικών [20]. Αποτελείται από μια μεγάλη πληθώρα βιβλιοθηκών υλοποιημένες στη γλώσσα C. Δημιουργεί και διαχειρίζεται κρυφά Μαρκοβιανά μοντέλα, επεξεργάζεται τις ιδιότητες τους, τα εκπαιδεύει ανάλογα με χαρακτηριστικά που δέχεται, και τέλος χρησιμοποιεί τον αλγόριθμο Viterbi για να αναγνωρίσει τα διάφορα ηχητικά δεδομένα.

4.1.3 Ροή Εργασίας

4.1.3.1 Προετοιμασία δεδομένων

Προτού τα ακουστικά αρχεία της βάσης UPC-TALP οδηγηθούν για μελέτη στο σύστημα αναγνώρισης ακουστικών γεγονότων, έπρεπε πρώτα να υποστούν επεξεργασία. Όπως αναφέραμε και προηγουμένως τα αρχεία προσφέρονται με μορφοποίηση “wv” στα 44.1 kHz και με ακρίβεια των 24-bits, όπου κάθε ακουστικό αρχείο αποτελεί την παρατήρηση ενός συγκεκριμένου μικροφώνου. Για την αρχική επεξεργασία των δεδομένων χρησιμοποιήθηκε το εργαλείο SoX, όπου δημιουργήθηκε μια ρουτίνα κονσόλας (bash script), για την εύκολη και άμεση προσπέλαση όλων των αρχείων της βάσης, η οποία άλλαζε τη συχνότητα δειγματοληψίας από 44.1 kHz σε 16 kHz και την ακρίβεια από 24-bits σε 16-bits. Τέλος, απαραίτητη ήταν η αλλαγή της μορφής (format) των ακουστικών αρχείων από “.wv” σε “.wav”. Μετά το τέλος των παραπάνω βημάτων έχουμε ουσιαστικά δημιουργήσει ένα αντίγραφο της αρχικής βάσης με διαφορετική μορφή και συχνότητα δειγματοληψίας. Η διαδικασία φαίνεται και στο Σχήμα 5.



Σχήμα 5 Διάγραμμα για την προετοιμασία των δεδομένων

Μετά την επεξεργασία της βάσης, ώστε να έρθει στην επιθυμητή μορφή σύμφωνα με το HTK, έγινε μία μελέτη πάνω στα αρχεία επισημείωσης. Αρχικά, δημιουργήθηκε κώδικας σε Python ώστε να βρεθούν όλα τα ακουστικά γεγονότα που χρησιμοποιούνται, καθώς και η διάρκεια τους. Επιπλέον, πραγματοποιήθηκε ξεχωριστός έλεγχος για επικαλυπτόμενα ακουστικά γεγονότα δύο μόνο πηγών, αφού στο πρόβλημα μας δεν λάβαμε υπόψιν επικαλύψεις που αποτελούνταν από 3 ξεχωριστές πηγές και πάνω. Τέλος, πριν τη δημιουργία των τελικών κλάσεων, υπολογίσαμε τις συνολικές φορές που εμφανίζεται κάθε ακουστικό γεγονός σε όλες τις συνεδρίες της βάσης, καθώς και τη διάρκεια τους, ώστε να είμαστε σίγουροι ότι η εκπαίδευση τους θα γίνει σωστά. Το αποτέλεσμα της επεξεργασίας μας οδήγησε στη εύρεση 13 μεμονωμένων ακουστικών γεγονότων και 25 επικαλυπτόμενων ακουστικών γεγονότων, όπως αυτά παρουσιάζονται στον Πίνακα 1. Οι επικαλυπτόμενες ακουστικές κλάσεις που δημιουργήθηκαν αποτελούνται από 2 πηγές, από τις οποίες τις περισσότερες φορές η μία είναι πάντα η

ομιλία. Παρόλα αυτά, υπάρχουν και ακουστικές κλάσεις όπου δύο διαφορετικά γεγονότα λαμβάνουν χώρα ταυτόχρονα, όπως για παράδειγμα χτύπημα πόρτας (ds) με βήματα (st).

Label	Event Type	Label	Event Type
ap	applause	**pw-cl**	paperwork - spoon
cl	spoon	**pw-co**	paperwork- cough
cm	chairmoving	**pw-kn**	paperwork - knock
co	cough	**pw-kt**	paperwork - keyboard
ds	doorslam	**pr-kt**	phone - keyboard
kt	keyboard	**pr-sp**	phone - speech
kj	keyjingle	**sp-ap**	speech - applause
kn	knock	**sp-cl**	speech - spoon
pw	paperwork	**sp-cm**	speech - chairmoving
pr	phone	**sp-co**	speech - cough
sp	speech	**sp-ds**	speech - doorslam
si	silence	**sp-kj**	speech - keyjingle
st	steps	**sp-kt**	speech - keyboard
cl-cm	spoon - chairmoving	**sp-pw**	speech - paperwork
cm-ds	chairmoving - doorslam	**sp-st**	speech - steps
cm-pw	chairmoving - paperwork	**st-cl**	steps - spoon
ds-st	doorslam- steps	**st-cm**	steps - chairmoving
kt-kn	keyboard- knock	**st-kj**	steps - keyjingle
kj-cm	keyjingle - chairmoving	**st-kn**	steps - knock

Πίνακας 1 Κατανομή των κλάσεων της βάσης UPC-TALP

Σε αυτό το στάδιο της εργασίας, τα ακουστικά αρχεία είναι έτοιμα για το τελικό στάδιο της προ-επεξεργασίας. Έτσι συντάχθηκε κώδικας σε Python που εκτελεί τα παρακάτω βήματα για κάθε ένα αρχείο σχολιασμού :

1. Διατρέχει το αρχείο επισημείωσης και υπολογίζει την χρονική στιγμή έναρξης καθώς και τη διάρκεια του κάθε γεγονότος.
2. Τμηματοποιεί το κάθε ακουστικό γεγονός σε ξεχωριστό ακουστικό αρχείο με τη βοήθεια του εργαλείου SoX.
3. Εξετάζει αν μεταξύ δύο ακουστικών γεγονότων υπάρχει ησυχία με τον εξής τρόπο. Αν υπάρχει διαφορά μεταξύ της χρονικής στιγμής που τελείωσε το $1^{ο}$ γεγονός και της στιγμής που ξεκίνησε το $2^{ο}$ γεγονός, τότε δημιουργεί το ακουστικό αρχείο της κλάσης «ησυχία».

Τελικά, τα ακουστικά αρχεία της βάσης δεδομένων έχουν επεξεργαστεί και τμηματοποιηθεί ως ξεχωριστά ακουστικά γεγονότα σε ξεχωριστούς φακέλους ανάλογα με την συνεδρία στην οποία ανήκουν.

4.1.3.2 Δημιουργία Γραμματικής-Λεξικού

Το πρώτο βήμα για την αναγνώριση ακουστικών γεγονότων είναι ο ορισμός της γραμματικής που παράγει τις λέξεις. Για τη δημιουργία της γραμματικής στο HTK χρειάζονται δύο αρχεία:

- το αρχείο "**.grammar**" το οποίο περιέχει τους κανόνες που διέπουν την γραμματική
- και το αρχείο "**.voca**" το οποίο περιέχει όλες τις πιθανές λέξεις προς αναγνώριση, καθώς και την προφορά τους.

Τα πειράματα έγιναν για αναγνώριση τόσο μεμονωμένων γεγονότων όσο και για ακολουθίες συμβάντων. Γι' αυτό το λόγο κατασκευάστηκαν δύο γραμματικές, μία για κάθε περίπτωση, οι οποίες παρουσιάζονται στο Σχήμα 2 και Σχήμα 7 αντίστοιχα, ενώ οι κανόνες που ακολουθούνται για τη σύνταξη του αρχείου ".grammar" είναι της μορφής BNF.

```
S:EVENTS
```

*Σχήμα 6 Ορισμός γραμματικής για αναγνώριση μεμονωμένων γεγονότων, αρχείο **.grammar***

```
S:SENT
SENT:EVENTS
```

*Σχήμα 7 Ορισμός γραμματικής για αναγνώριση ακολουθίας γεγονότων, αρχείο **.grammar***

Τα S και SENT αποτελούν μη τερματικά σύμβολα, ενώ το EVENTS αποτελεί τερματικό σύμβολο και αντιπροσωπεύει μία σταθερή τιμή.

Το αρχείο "**.voca**" περιέχει όλους τους ορισμούς για κάθε κατηγορία που υπάρχει στο "**.grammar**" και περιγράφει με ποια φωνήματα μπορούν να αντικατασταθούν τα γεγονότα. Στην δικιά μας περίπτωση έχουμε μόνο την κατηγορία EVENTS και έτσι το αρχείο θα έχει τη μορφή που φαίνεται στο Σχήμα 8.

```
% EVENTS
% EVENTS
ap ap
cl cl
cl-cm cl-cm
cm cm
cm-ds cm-ds
cm-pw cm-pw
co co
ds ds
ds-st ds-st
kt kt
kt-kn kt-kn
kj kj
kj-cm kj-cm
kn kn
pw pw
pw-cl pw-cl
pw-co pw-co
pw-kn pw-kn
pw-kt pw-kt
pr pr
etc
```

*Σχήμα 8 Περιεχόμενο του λεξιλογίου στο αρχείο **.voca***

Όπως φαίνεται από τα παραπάνω αρχεία προσπαθούμε να αναγνωρίσουμε κάποια συγκεκριμένα γεγονότα, είτε μεμονωμένα είτε επικαλυπτόμενα, ή κάποια ακολουθία αυτών. Έτσι η γραμματική μας αναφέρεται στην ακολουθία που περιμένουμε να πάρουμε κατά τη διάρκεια της αναγνώρισης. Η παραγωγή, λοιπόν, των πιθανών προτάσεων γίνεται στο αρχείο **gram** , όπως φαίνεται στο Σχήμα 9.

```
$events= ap|cl|cl-cm|cm|cm-ds|cm-pw|co|ds|ds-st|kt|kt-kn|kj|kj-
cm|kn|pw|pw-cl|pw-co|pw-kn|pw-kt|pr|pr-kt|pr-sp|sp|sp-ap|sp-cl|sp-cm|sp-
co|sp-ds|sp-kj|sp-kt|sp-pw|sp-st|si|st|st-cl|st-cm|st-kj|st-kn;

($events)
```

*Σχήμα 9 Περιεχόμενο του αρχείου **gram***

Αφού δημιουργήθηκαν τα απαραίτητα αρχεία, κλήθηκε η ακόλουθη εντολή του HTK δημιουργώντας το αρχείο **wdnet** :

HParse gram wdnet

Μετά τη δημιουργία των απαραίτητων αρχείων για τη γραμματική, σειρά έχει η δημιουργία του Λεξικού. Πρόκειται για μία ταξινομημένη λίστα με όλες τις λέξεις που δέχεται η γραμματική, καθώς και συνδυασμό τους. Το HTK διαθέτει το λεξικό Voxforge, όπου υπάρχουν καταγεγραμμένες το μεγαλύτερο μέρος των αγγλικών λέξεων με μια πληθώρα προφορών. Στη συγκεκριμένη περίπτωση όμως, το πρόβλημα που μελετάτε δεν ασχολείται με αναγνώριση λέξεων, αλλά ακουστικών γεγονότων. Συνεπώς, το αρχείο **lexicon** που δημιουργήθηκε είναι ίδιο με το αρχείο "**.voca**" .

4.1.3.3 Εξαγωγή Διανυσμάτων MFCCs

Το πρώτο στάδιο για την εκπαίδευση των μοντέλων είναι η μετατροπή των ακουστικών γεγονότων σε ακολουθίες διανυσμάτων [20]. Για αυτό το λόγο, εξήχθησαν τα διανύσματα MFCC για κάθε ακουστικό αρχείο και τοποθετήθηκαν αυτόματα στο κατάλληλο φάκελο. Η δημιουργία των διανυσμάτων MFCC εκτελέστηκε χρησιμοποιώντας το εργαλείο **HCopy**, το οποίο παίρνει ως παραμέτρους το αρχείο ρυθμίσεων **config,** καθώς και ένα αρχείο script, με όνομα **codetrain.scp** σύμφωνα με το HTK, που καθορίζει το όνομα και τη θέση των ακουστικών αρχείων του συνόλου εκπαίδευσης, μαζί με το όνομα και τη θέση των διανυσμάτων που θα προκύψουν. Το περιεχόμενο του αρχείου **config** παρουσιάζεται στο Σχήμα 10.

```
SOURCEFORMAT=WAV
TARGETKIND = MFCC_0_D_A
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
```

*Σχήμα 10 Αρχείο ρυθμίσεων **config***

Με ένα τέτοιο αρχείο ρυθμίσεων θα εκτελεστεί ανάλυση MFCCs (Mel Frequency Cepstral Coefficient) όπου οι πιο σημαντικές ιδιότητες αναλύονται παρακάτω:

- Με την επιλογή **SOURCEFORMAT** δηλώνουμε τον τύπο αρχείου από τα δεδομένα εισόδου, που στην περίπτωσή μας είναι της μορφής .wav .
- Με την επιλογή **TARGETKIND** δηλώνουμε τη μορφή των διανυσμάτων που θα εξαχθούν. Έχοντας ως παράμετρο την επιλογή MFCC_0_D_A για κάθε πλαίσιο του ηχητικού σήματος θα εξαχθούν 39 συντελεστές. Έτσι θα εξαχθούν οι 13 πρώτοι MFCC συντελεστές, αντικαθιστώντας τον τελευταίο με τον μηδενικό συντελεστή που αντιστοιχεί στην ενέργεια του κάθε πλαισίου (λόγω της επιλογής _0), έπειτα θα εξαχθούν οι 13 Delta συντελεστές που αποτελούν την παράγωγο $1^{ης}$ τάξης των 13 MFCC συντελεστών (λόγω της επιλογής _D) και τέλος θα εξαχθούν οι 13 Acceleration συντελεστές που αποτελούν την παράγωγο της $2^{ης}$ τάξης (λόγω της επιλογής _A).
- Με την επιλογή **WINDOWSIZE** ορίζουμε το μέγεθος του πλαισίου με το οποίο σαρώνεται το ηχητικό σήμα. Η τιμή του είναι προκαθορισμένα στα 25ms.
- Με την επιλογή **TARGETRATE** ορίζουμε κατά πόσα δευτερόλεπτα μετακινούμαστε πάνω στο σήμα. Στο πρόβλημα μας θέσαμε αυτή την τιμή στα 10ms.

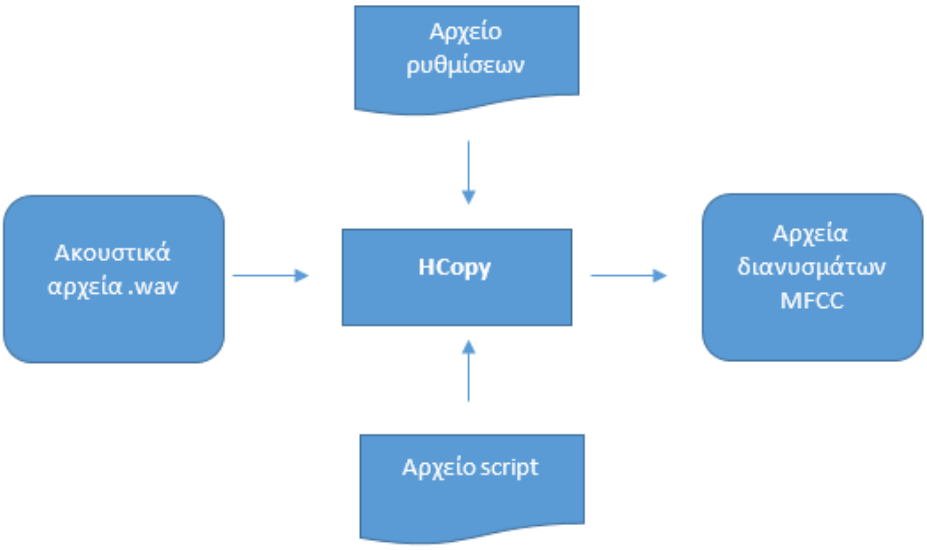
Μετά τη δημιουργία των δύο απαραίτητων αρχείων κλήθηκε η ακόλουθη εντολή :

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

όπου

- Η επιλογή -A έχει ως αποτέλεσμα την εκτύπωση των ορισμάτων στην κονσόλα
- Η επιλογή -C δηλώνει την ύπαρξη αρχείου ρυθμίσεων
- Η επιλογή -S δηλώνει την ύπαρξη script αρχείου

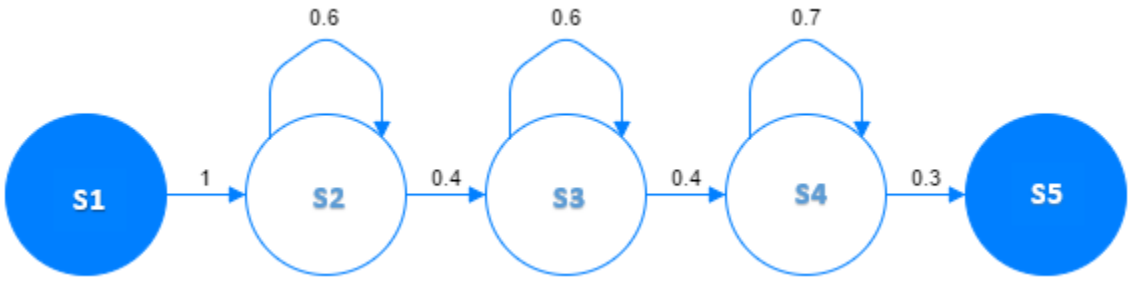
Η εκτέλεση της παραπάνω εντολής έχει ως αποτέλεσμα τη δημιουργία των MFCC διανυσμάτων. Για κάθε ακουστικό αρχείο δημιουργήθηκε ένα αρχείο με το ίδιο όνομα και κατάληξη **.mfc**. Πρέπει να σημειωθεί ότι μετά την εξαγωγή των χαρακτηριστικών, το HTK επιβάλει να διαγραφεί η παράμετρος **SOURCEFORMAT** από το αρχείο ρυθμίσεων προτού προχωρήσουμε στην εκπαίδευση των HMM μοντέλων [20]. Η διαδικασία απεικονίζεται γραφικά στο Σχήμα 11.



Σχήμα 11 Διαδικασία μετατροπής ακουστικών δεδομένων

4.1.3.4 Ορισμός Πρωτότυπου Μοντέλου HMM

Το επόμενο βήμα για την εκπαίδευση των Μαρκοβιανών μοντέλων είναι ο ορισμός ενός πρωτότυπου μοντέλου για τη δημιουργία ενός καλά εκπαιδευμένου συνόλου. Το σημαντικό στοιχείο του πρωτότυπου μοντέλου δεν είναι οι παράμετροί του, αλλά η τοπολογία του. Για μοντέλα αναγνώρισης ακουστικών γεγονότων ένα πρωτότυπο με πέντε καταστάσεις αποτελεί την καλύτερη τοπολογία [20]. Η κάθε κλάση αποτελείται από πέντε HMM καταστάσεις εκ των οποίων οι τρεις (S2, S3, S4) αποτελούν ενεργές καταστάσεις και εκπέμπουν πιθανότητες, ενώ η πρώτη (S1) και η τελευταία (S5) είναι τερματικές και δεν εκπέμπουν πιθανότητες, όπως παρουσιάζεται και στο Σχήμα 12. Έτσι, δημιουργήθηκε το αρχείο **prototype**, όπου κάθε κλάση έχει την ίδια μέση τιμή και διακύμανση.



Σχήμα 12 Πρωτότυπο μοντέλο HMM

Κάθε ενεργή κατάσταση του μοντέλου περιγράφεται από ένα Γκαουσιανό μοντέλο μίξης. Το Γκαουσιανό μοντέλο περιγράφεται πλήρως από ένα διάνυσμα μέσης τιμής και ένα διάνυσμα διακύμανσης, ενώ οι καταστάσεις S1 και S5 δεν περιγράφονται αφού δεν εκπέμπουν καμία πιθανότητα. Έπειτα, αφού δημιουργήθηκε το αρχείο **trainHMM.scp**, που περιέχει τη θέση των MFCC χαρακτηριστικών, καλέσαμε την παρακάτω εντολή:

HCompV -C config -f 0.01 -m -S trainscript.scp -M hmm0 prototype

Το εργαλείο HCompV δημιουργεί μία νέα έκδοση του αρχείου **prototype** στο φάκελο **hmm0** με βάση το σύνολο εκπαίδευσης, στο οποίο οι μηδενικές μέσες τιμές και οι μοναδιαίες διακυμάνσεις έχουν αντικατασταθεί. Επιπλέον, λόγω της επιλογής **-f** δημιουργήθηκε το αρχείο **vFloors** (variance floor macro) που περιέχει ένα διάνυσμα, οι τιμές του οποίου θα χρησιμοποιηθούν για να ορίσουν ένα κάτω όριο στις διακυμάνσεις που θα υπολογιστούν στα επόμενα βήματα [20].

4.1.3.5 Αρχεία Ετικετών

Το εργαλείο HTK για κάθε ακουστικό γεγονός χρειάζεται να ξέρει τι αντιπροσωπεύει. Έτσι δημιουργήσαμε Αρχεία Ετικετών ή αλλιώς Label Files, δηλαδή αρχεία που αποθηκεύουν πληροφορίες για τον αριθμό και την ακολουθία των γεγονότων κάθε ακουστικού αρχείου. Η τυπική μορφή ενός Αρχείου Ετικετών είναι ένα **.lab** αρχείο, με όνομα ίδιο με το αρχείο που περιέχει το διάνυσμα χαρακτηριστικών, και περιέχει την ακολουθία γεγονότων. Όλες οι πληροφορίες για τα Αρχεία Ετικετών αποθηκεύονται τελικά σε ένα Master Label αρχείο με κατάληξη **.mlf**.

4.1.3.6 Εκπαίδευση Κρυφών Μαρκοβιανών Μοντέλων (HMM)

Μοναδικό Μοντέλο Μίξης

Μετά την δημιουργία όλων των απαραίτητων αρχείων τα δεδομένα μας είναι έτοιμα για να εκπαιδευτούν με χρήση του αλγορίθμου Baum-Welch. Εκτελώντας την εντολή

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

το αρχικό μοντέλο **hmm0** εκπαιδεύεται και εξάγει το νέο σύνολο μοντέλων του δεύτερου σταδίου **hmm1**. Η διαδικασία αυτή επαναλαμβάνεται μέχρι το ένατο στάδιο, δηλαδή μέχρι να δημιουργηθεί το μοντέλο σταδίου **hmm9**. Η διαδικασία της εκπαίδευσης περιγράφεται στο Σχήμα 13.

Σχήμα 13 Διαδικασία εκπαίδευσης με μοναδικό μοντέλο μίξης[21]

Πολλαπλά Μοντέλα Μίξης

Στην προηγούμενη ενότητα, εκπαιδεύσαμε τα δεδομένα μας χρησιμοποιώντας ένα μοναδικό Γκαουσιανό μοντέλο. Σε αυτή τη ενότητα, περιγράφεται η διαδικασία εκπαίδευσης χρησιμοποιώντας πολλαπλά μοντέλα μίξης. Η εκπαίδευση ξεκίνησε με ένα μοναδικό μοντέλο και διπλασίαζε τον αριθμό των μοντέλων σε κάθε επανάληψη, μέχρι να φτάσει τις 16 Γκαουσιανές [22]. Ο κύριος λόγος που χρησιμοποιήθηκαν πολλαπλά μοντέλα μίξης για εκπαίδευση είναι για να διαπιστωθεί με ποιο μοντέλο μπορούν να υπάρξουν τα καλύτερα αποτελέσματα στην αναγνώριση. Σε όλες τις περιπτώσεις, όμως, υπάρχει ένα άνω όριο στον αριθμό των φορών που μπορεί να εκπαιδευτεί ένα μοντέλο, καθώς στα προβλήματα ταξινόμησης υπάρχει πάντα το πρόβλημα της υπερεκπαίδευσης (overtraining). Για την γρήγορη εκπαίδευση των δεδομένων δημιουργήθηκε script αρχείο που εκτελεί όλη τη διαδικασία της εκπαίδευσης αυξάνοντας τα μοντέλα μίξης. Πιο συγκεκριμένα για κάθε μοντέλο μίξης (2,4,8,16) εκτελούνται τα παρακάτω βήματα:

- Παίρνει ως αρχικό μοντέλο το μοντέλο **hmm9** που δημιουργήθηκε από το μοναδικό Γκαουσιανό μοντέλο μίξης.
- Εκτελεί την παραπάνω εντολή HERest για να εκπαιδεύσει το σύστημα, δημιουργώντας την κατάσταση **hmm9** για το αντίστοιχο μοντέλο μίξης.

Τα μοντέλα μίξης καθορίζονται σε ένα αρχείο **split.hed** που προσδιορίζει τον αριθμό των μοντέλων που χρησιμοποιεί, ενώ τα τελικά μοντέλα που χρησιμοποιούμε για αναγνώριση είναι τα εξής:

- **hmm9**, εκπαίδευσε 9 φορές το μοντέλο **hmm0** χρησιμοποιώντας 1 GMM
- **hmm9-GMM 2**, εκπαίδευσε 9 φορές το μοντέλο **hmm9** χρησιμοποιώντας 2 GMM
- **hmm9-GMM 4**, εκπαίδευσε 9 φορές το μοντέλο **hmm9-GMM 2** χρησιμοποιώντας 4 GMMs
- **hmm9-GMM 8**, εκπαίδευσε 9 φορές το μοντέλο **hmm9-GMM 4** χρησιμοποιώντας 8 GMMs
- **hmm9-GMM 16**, εκπαίδευσε 9 φορές το μοντέλο **hmm9-GMM 8** χρησιμοποιώντας 16 GMMs

4.1.3.7 Αναγνώριση

Η διαδικασία της αναγνώρισης χωρίζεται σε δύο μέρη. Και στις δύο περιπτώσεις το σύνολο εκπαίδευσης, όπως είδαμε, αποτελούνταν από μεμονωμένα γεγονότα, ενώ το σύνολο δοκιμής στη μία περίπτωση αποτελούνταν από μεμονωμένα γεγονότα και στην άλλη περίπτωση αποτελούνταν από ολόκληρη τη συνεδρία 7. Η πρώτη περίπτωση ονομάζεται isolated training & testing, ενώ η δεύτερη περίπτωση isolated training & embedded testing.

Αρχικά, έχουμε ως είσοδο το ακουστικό αρχείο μορφής ".wav" που θα χρησιμοποιηθεί σαν δεδομένο δοκιμής. Μέσω του εργαλείου **HCopy** θα μετατραπεί σε μία ακολουθία MFCC διανυσμάτων, ακολουθώντας την ίδια διαδικασία που έγινε κατά τη διάρκεια της εκπαίδευσης. Έτσι, θα δημιουργηθεί το αρχείο με τα διανύσματα, το οποίο θα δοθεί ως είσοδο στον αλγόριθμο Viterbi, όπου και θα ταξινομηθούν τα γεγονότα. Η εντολή που χρησιμοποιεί το εργαλείο HVite είναι η παρακάτω:

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

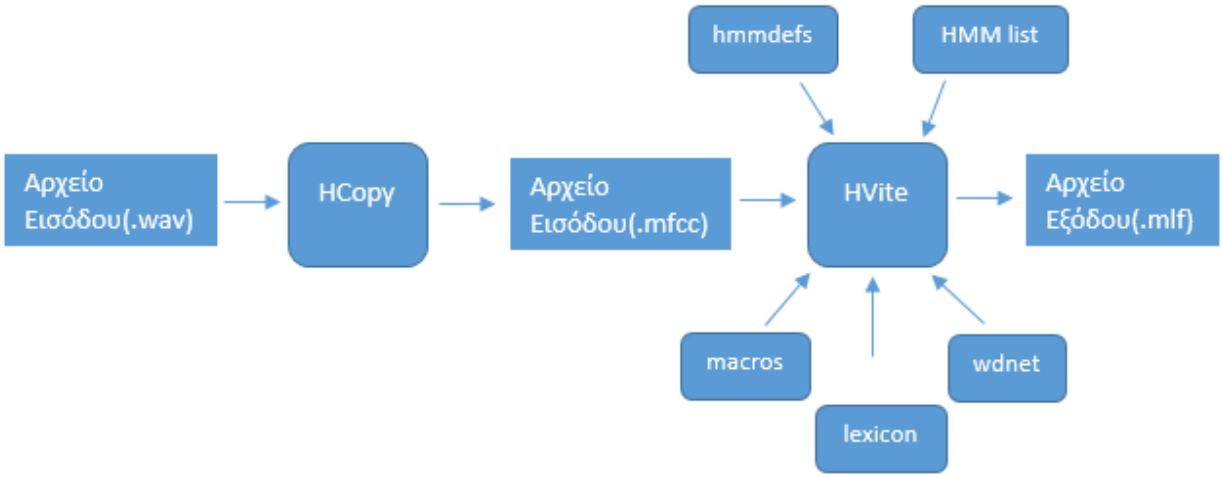
το οποίο δίνει σαν έξοδο το αρχείο **recout.mlf**, όπου φαίνονται τα δεδομένα που αναγνωρίστηκαν. Το περιεχόμενο του αρχείου φαίνεται στο Σχήμα 14, ενώ η διαδικασία φαίνεται σχηματικά στο Σχήμα 15. Τέλος για την λήψη των αποτελεσμάτων χρησιμοποιήθηκε το εργαλείο **HResults** του HTK με την εντολή

HResults -I test.mlf tiedlist recout.mlf

το οποίο εκτυπώνει ένα πίνακα με το ποσοστό αναγνώρισης με βάση δύο μετρικές αξιολόγησης (βλέπε Κεφάλαιο 5).

```
#!MLF!#
"*/T07.rec"
0 10200000 si -3950.679688
10200000 27600000 si -6431.375488
27600000 28500000 stkn -386.689148
28500000 37100000 si -3206.938721
37100000 38300000 cl -669.760986
38300000 39200000 stkn -503.703949
39200000 43300000 si -1557.547974
43300000 51800000 ds -3038.382812
51800000 54900000 si -1078.237183
54900000 56100000 st -505.288910
56100000 60100000 ds -1529.178101
60100000 69800000 st -3397.456055
69800000 81900000 ds -4723.501953
81900000 85900000 si -1377.684326
85900000 91000000 st -1938.895752
91000000 93400000 st -937.998352
93400000 94600000 stcm -412.708832
94600000 96400000 cm -679.244202
96400000 106700000 st -3804.529785
106700000 117100000 st -3913.436523
117100000 118800000 spst -679.684448
118800000 122200000 st -1235.037720
122200000 127500000 st -1901.644409
```

*Σχήμα 14 Περιεχόμενο αρχείου εξόδου **recout.mlf***



Σχήμα 15 Διαδικασία αναγνώρισης

4.2 Εντοπισμός Επικαλυπτόμενων Ακουστικών Γεγονότων

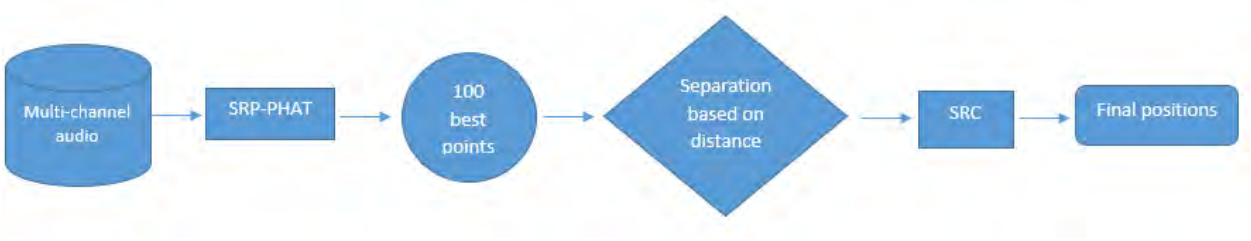
4.2.1 Matlab

Η ονομασία του προέρχεται από το Matrix Laboratory. Πρόκειται για μία πλατφόρμα που δημιουργήθηκε για να επιλύει μαθηματικά, μηχανικά και επιστημονικά προβλήματα. Η γλώσσα του Matlab βασίζεται στις πράξεις πινάκων, ενώ με τη βοήθεια των ενσωματωμένων γραφημάτων καθίσταται πιο εύκολη η απεικόνιση των δεδομένων και των αποτελεσμάτων. Η πιο πρόσφατη έκδοση του Matlab είναι η R2016b, η οποία κυκλοφόρησε το Σεπτέμβριο του 2016.

4.2.2 Ροή Εργασίας

Η ανάλυση της ακουστικής σκηνής και ο εντοπισμός των ακουστικών γεγονότων είναι ένα από τα βασικά προβλήματα που βρίσκονται υπό έρευνα στο κλάδο της Εντοπισμού Ακουστικών Γεγονότων. Στόχος αυτής της μεθόδου είναι η δημιουργία ενός συστήματος που θα ανιχνεύει τις ηχητικές πηγές που υπάρχουν σε ένα έξυπνο δωμάτιο, είτε παρουσία μίας μεμονωμένης πηγής είτε παρουσία δύο πηγών. Το έργο όμως για τον εντοπισμό της πηγής δυσκολεύει λόγω του θορύβου και της αντήχησης, αλλά κυρίως λόγω των πολλαπλών πηγών. Για αυτό το λόγο, το σύστημα βασίστηκε στην τεχνική του αλγορίθμου SRP-PHAT, καθώς το βασικό χαρακτηριστικό του είναι η ανοχή σε αντήχηση και η ανεξαρτησία του από τον προσανατολισμό των ηχείων. Η ροή εργασίας που ακολουθήθηκε για τη

δημιουργία του συστήματος εντοπισμού μέσω του εργαλείου Matlab παρουσιάζεται γραφικά στο Σχήμα 16.



Σχήμα 16 Ροή εργασίας για το σχεδιασμό του συστήματος εντοπισμού ακουστικών πηγών

Η μελέτη και ο σχεδιασμός του συστήματος βασίζεται στην πολυκαναλική επεξεργασία του δωματίου, χρησιμοποιώντας τις συστοιχίες μικροφώνων όπως αυτά έχουν κατανεμηθεί στο δωμάτιο. Θεωρώντας το δωμάτιο ως ένα τρισδιάστατο χώρο, πραγματοποιείται η προσπάθεια εκτίμησης ενός συνόλου συντεταγμένων βασισμένων στις παρατηρήσεις των συστοιχιών μικροφώνων, που θα αντικατοπτρίζουν τις θέσεις είτε του ομιλητή είτε των ακουστικών γεγονότων που συμβαίνουν.

Πιο συγκεκριμένα έχοντας 24 μικρόφωνα και γνωρίζοντας τις συντεταγμένες τους, τμηματοποιήσαμε το χώρο μας σε διαστάσεις των 10x1) cm^2 για ένα μέσο ύψος του δωματίου γύρω στο 1.5 m. Η επιλογή της τιμής των 10 cm για τη διαίρεση του χώρου αποτελεί μία μέση περίπτωση, καθώς δίνοντας μεγαλύτερες ή μικρότερες τιμές δεν πραγματοποιούνταν σωστός εντοπισμός των πηγών στο συγκεκριμένο χώρο που μελετάμε. Έπειτα, συλλέξαμε τα ηχητικά δεδομένα απ' όλα τα μικρόφωνα και τα μετατρέψαμε σε ένα πίνακα διανυσμάτων. Για τη διαδικασία του εντοπισμού, χρησιμοποιήθηκαν πλαίσια παραθύρων των 250ms, τα οποία κάθε φορά μετατοπίζονταν κατά 10ms. Για κάθε πλαίσιο ο αλγόριθμος υπολογίζει τις χρονικές καθυστερήσεις των ζευγαριών μικροφώνων (time difference of arrival - TDOA) και έπειτα το μετασχηματισμό φάσης κάθε ζεύγους. Αφού υπολογίστηκαν οι τιμές της κατευθυντήριας δύναμης απόκρισης με τον μετασχηματισμό φάσης (SRP-PHAT), πήραμε ένα χάρτη του χώρου για το ηχητικό αρχείο που δώσαμε με βάση τις τιμές SRP που προέκυψαν. Καθώς στο πρόβλημα που μελετάμε ο αριθμός των ακουστικών πηγών δεν είναι μοναδιαίος, κρίθηκε αναγκαία η επεξεργασία αυτής της ρουτίνας, ώστε να εντοπίζει παραπάνω από μία ακουστικές πηγές στο χώρο. Για αυτό το λόγο, αφού δημιουργήθηκαν οι τιμές SRP-PHAT και ταξινομήθηκαν με φθίνουσα σειρά, έγινε διαχωρισμός σύμφωνα με τις αποστάσεις τους. Λαμβάνοντας ως δεδομένο ότι το σημείο με το μεγαλύτερο SRP-PHAT θα αποτελεί και την κύρια πηγή πληροφορίας ($1^η$ ακουστική πηγή), θέσαμε ένα όριο 40 cm , έτσι ώστε οποιοδήποτε άλλο σημείο απέχει λιγότερο από αυτό το όριο να θεωρείται ότι αποτελεί το ίδιο ακουστικό γεγονός. Η τιμή αυτή επιλέχθηκε εμπειρικά και στηρίζεται στο γεγονός ότι όταν δύο άνθρωποι στον ίδιο χώρο συνομιλούν απέχουν περίπου τόση απόσταση. Δοκιμές πραγματοποιήθηκαν τόσο για μεγαλύτερες όσο και μικρότερες τιμές. Αυξάνοντας την τιμή, το ποσοστό των σωστά εντοπισμένων ακουστικών γεγονότων μειωνόταν ραγδαία, καθώς δυσκολευόταν να εντοπίσει δύο ταυτόχρονες ακουστικές πηγές, ενώ μειώνοντας την τιμή, αναγνώριζε λανθασμένα πρόσθετες ακουστικές πηγές με αποτέλεσμα το ποσοστό να μειώνεται και σε αυτή τη περίπτωση. Στη συνέχεια, και με δεδομένο ότι ο αλγόριθμος αναγνωρίζει έως και δύο ακουστικές πηγές, το πρώτο σημείο που δεν θα ανήκει στην $1^η$ ακουστική πηγή θα σημαίνει την ύπαρξη $2^ης$ πηγής. Έχοντας εντοπίσει μία ή δύο ακουστικές πηγές, κλήθηκε ο αλγόριθμος βελτιστοποίησης SRC με σκοπό να εντοπίσει καλύτερα μέσα στο χώρο από που προέρχονται τα ακουστικά γεγονότα. Για τον υπολογισμό των τιμών του SRP-PHAT, καθώς και για τον αλγόριθμο βελτιστοποίησης SRC, χρησιμοποιήθηκε η ρουτίνα που αναπτύχθηκε από τον H. Do [17], [23]. Τέλος,

αναπτύχθηκε μία ρουτίνα για την τελική επιβεβαίωση του αριθμού των πηγών, καθώς και την αναγνώριση του χώρου στον οποίο ανήκουν. Η ρουτίνα σχεδιάζει τους χώρους του δωματίου, όπου μπορούν να εντοπιστούν ηχητικές πηγές. Από τα δεδομένα της βάσης γνωρίζουμε ότι υπάρχουν 2 γενικοί χώροι όπου παράγονται γεγονότα:

- κοντά στην πόρτα
- στους χώρους γύρω και πάνω από το τραπέζι

Με βάση αυτές τις συνθήκες, αν το σύστημα εντόπιζε μία ακουστική πηγή σε κάποιο άλλο χώρο του δωματίου, την αγνοούσε θεωρώντας ότι δεν υπάρχει κάτι. Έχοντας ως δεδομένο αυτές τις συνθήκες και γνωρίζοντας τις γεωμετρικές συντεταγμένες του χώρου, δημιουργήσαμε εμπειρικά τις γεωμετρικές συντεταγμένες αυτών των χώρων. Αυτή η ρουτίνα μας βοήθησε κατά τη διάρκεια των πειραμάτων, ώστε να ελέγχουμε τα αποτελέσματα μας με βάση το χώρο στον οποίο ανήκουν.

Επιπλέον, έγιναν δοκιμές με την προσθήκη χρονικής ομαλοποίησης (temporal smoothing). Πιο συγκεκριμένα, θεωρήσαμε ένα τυχαίο βήμα των 10 frames. Για κάθε frame, υπολογίστηκαν οι αρχικές τιμές του SRP-PHAT σε κάθε σημείο του χώρου. Μετά και τον υπολογισμό του SRP-PHAT του $10^{ου}$ συνεχόμενου frame, υπολογίστηκε για κάθε σημείο το μέσο SRP-PHAT βάση όλων των προηγούμενων τιμών. Έπειτα, με βάση αυτές τις τιμές ανιχνεύτηκε το μέγιστο SRP-PHAT σε όλο το χώρο για τη συγκεκριμένη χρονική περίοδο και εντοπίστηκαν οι αντίστοιχες πηγές. Για παράδειγμα, στη $1^{η}$ επανάληψη για να βρούμε το μέσο όρο για κάθε σημείο πήραμε τα frames που αντιστοιχούσαν στη χρονική περίοδο από 0.01sec μέχρι 0.35sec, στη $2^{η}$ επανάληψη πήραμε για τη χρονική περίοδο από 0.02sec μέχρι 0.36 κ.ο.κ. Η μέθοδος αυτή χρησιμοποιήθηκε με σκοπό την ελαχιστοποίηση του θορύβου και την ελαχιστοποίηση των λάθος εντοπισμένων πηγών, με την υπόθεση ότι ο θόρυβος εμφανίζεται τυχαία στα διάφορα σημεία του χώρου. Παίρνοντας, λοιπόν, το μέσο όρο των τιμών ανά 10 frames η παρουσία του θορύβου θα έπρεπε να ελαχιστοποιηθεί.

Και στις δύο περιπτώσεις το σύστημα, για κάθε πλαίσιο του σήματος που αναλύει, δίνει ως έξοδο τρία πράγματα:

- τον αριθμό των εντοπισμένων ακουστικών πηγών
- τις τιμές του SRP για κάθε πηγή που αναγνώρισε
- το χώρο του δωματίου στον οποίο ανήκει, καθώς και τις συντεταγμένες των πηγών

4.3 Συνδυασμός των Μεθόδων

Για τη δημιουργία ενός σωστού συστήματος αναγνώρισης ακουστικών γεγονότων είναι συχνά προτιμότερο να συνδυάζονται μέθοδοι με σκοπό την καλύτερη απόδοση. Σε αυτό το κομμάτι της διπλωματικής, συνδυάζονται οι δύο μέθοδοι που αναπτύχθηκαν με σκοπό τη βελτιστοποίηση των αποτελεσμάτων. Για το σκοπό αυτό χρησιμοποιήθηκε ο αλγόριθμος N-best, που παρέχεται σαν εργαλείο στο HTK, όπου δίνει σαν αποτέλεσμα τις N πιο πιθανές επιλογές σύμφωνα με το αρχείο εισόδου. Για να δημιουργηθεί η N-best λίστα κατά τη διαδικασία της αναγνώρισης χρησιμοποιείται η επιλογή **-n** στην εντολή του **HVite**, όπου ορίζουμε τον αριθμό των N-best tokens που θα χρησιμοποιηθούν σε κάθε κατάσταση του Μαρκοβιανού μοντέλου καθώς και τις N-best υποθέσεις που θα δημιουργηθούν. Με

αυτό τον τρόπο, για κάθε μία είσοδο των T πλαισίων, δημιουργούνται μονοπάτια από τον αρχικό κόμβο μέχρι τον κόμβο εξόδου που διαπερνούν ακριβώς T καταστάσεις. Κάθε ένα τέτοιο μονοπάτι αποτελεί και μία υπόθεση προς αναγνώριση. Η δουλειά του αποκωδικοποιητή είναι να βρει τα N καλύτερα μονοπάτια, δηλαδή τα μονοπάτια με την μεγαλύτερη λογαριθμική πιθανότητα. Όταν υπάρχουν πολλές πιθανές έξοδοι από μία κατάσταση, το token αντιγράφεται έτσι ώστε όλα τα μονοπάτια να δημιουργηθούν παράλληλα. Ορίζοντας τον αριθμό N, το σύστημα κρατάει μέχρι τόσα μονοπάτια. Πιο συγκεκριμένα, για κάθε κατάσταση του HMM:

- Κρατάει σε ξεχωριστή μεταβλητή κάθε υπόθεση που έχει διαφορετικό μονοπάτι λέξεων
- Για κάθε υπόθεση αθροίζει τις πιθανότητες των καταστάσεων
- Κρατάει τις N καλύτερες υποθέσεις των οποίων οι πιθανότητες είναι οι μεγαλύτερες.

Δημιουργώντας μέσω του HTK τα N-best μονοπάτια και χρησιμοποιώντας τα δεδομένα που προέκυψαν από το σύστημα εντοπισμού των ηχητικών πηγών σχετικά με τον αριθμό των πηγών που υπάρχουν σε κάθε χρονικό πλαίσιο, είχαμε σαν αποτέλεσμα μία βελτιστοποιημένη αναγνώριση του αρχείου εισόδου. Η διαδικασία για τη δημιουργία του συστήματος βασίστηκε στον αριθμό των πηγών που εντοπίστηκαν σε κάθε χρονικό πλαίσιο και χωρίστηκε σε τρεις περιπτώσεις:

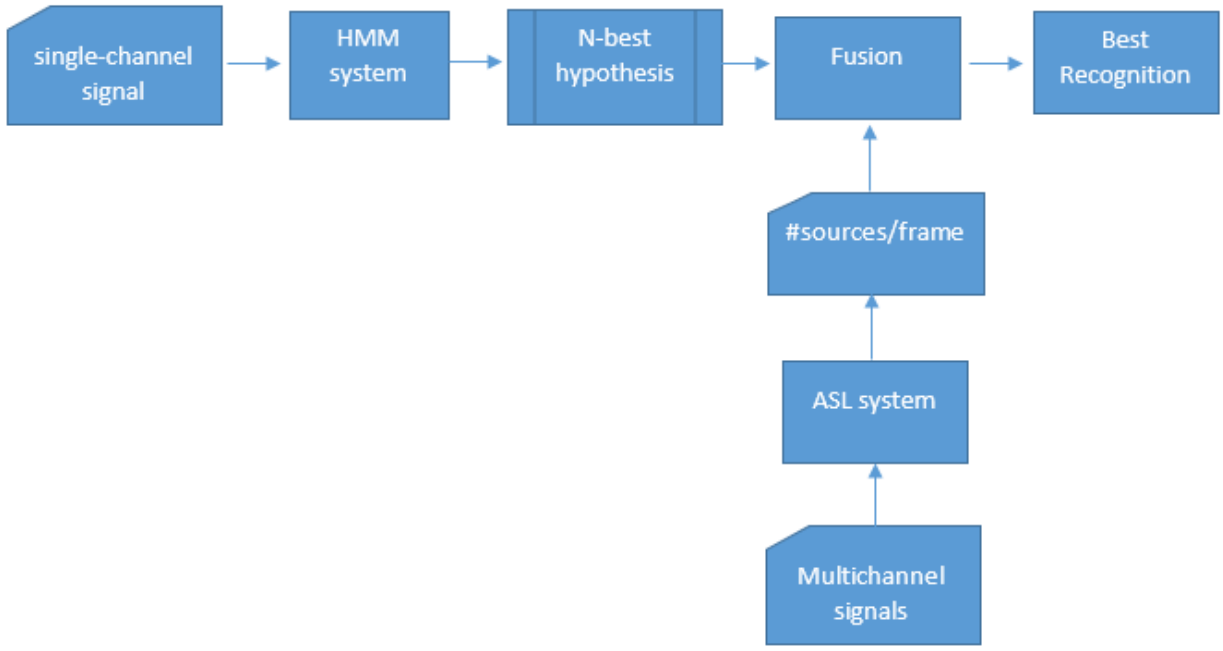
1. Όταν το σύστημα δεν εντόπισε καμία πηγή
2. Όταν το σύστημα εντόπισε μία πηγή
3. Όταν το σύστημα εντόπισε δύο πηγές

Στην πρώτη περίπτωση, αν κάποια από τις N καλύτερες υποθέσεις έχει αναγνωρίσει ησυχία τότε το σύστημα μας δέχεται αυτή σαν πιθανή αναγνώριση, ενώ σε αντίθετη περίπτωση δέχεται το ακουστικό γεγονός με την μεγαλύτερη πιθανότητα. Στην δεύτερη περίπτωση, το σύστημα δέχεται σαν πιθανή αναγνώριση το γεγονός με την μεγαλύτερη πιθανότητα από τις N υποθέσεις. Στην τρίτη περίπτωση, το σύστημα ψάχνει τις N υποθέσεις για να εντοπίσει ένα επικαλυπτόμενο γεγονός. Αν δεν καταφέρει να εντοπίσει δύο ταυτόχρονα ακουστικά γεγονότα, κρατάει τα δύο γεγονότα με τις μεγαλύτερες πιθανότητες που αναγνωρίστηκαν. Με βάση αυτή τη διαδικασία δημιουργείται η τελική πιθανή ακολουθία των γεγονότων. Στη συνέχεια, η ακολουθία ελέγχθηκε με τον ίδιο τρόπο που ελέγχθηκαν και τα αντίστοιχα αρχεία στη μέθοδο ανίχνευσης για τον υπολογισμό της απόδοσης. Πιο συγκεκριμένα, δημιουργήθηκε κώδικας που εκτελεί τα παρακάτω βήματα:

- Διαβάζει τον αριθμό των πηγών που υπάρχουν σε κάθε χρονικό πλαίσιο των 25ms, όπου κάθε πλαίσιο μετατοπίζεται κατά 10ms
- Σύμφωνα με τον αριθμό των πηγών, κρατάει τον αντίστοιχο αριθμό αποτελεσμάτων από τα N-best για εκείνο το πλαίσιο
- Δημιουργεί την τελική καλύτερη λίστα αποτελεσμάτων
- Ανοίγει το αρχείο αναφοράς και διασπά τα δεδομένα σε ίδια χρονικά πλαίσια
- Συγκρίνει ένα προς ένα τα πλαίσια με το αρχείο αναφοράς υπολογίζοντας πόσα γεγονότα αναγνώρισε σωστά

- Υπολογίζει τους δύο λόγους, οι οποίο αποτελούν το ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου

Η διαδικασία που ακολουθήθηκε φαίνεται στο Σχήμα 17.



Σχήμα 17 Διαδικασία δημιουργίας βελτιστοποιημένης αναγνώρισης με συνδυασμό των μεθόδων

ΚΕΦΑΛΑΙΟ 5-ΠΕΙΡΑΜΑΤΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

5.1 Ανίχνευση Επικαλυπτόμενων Ακουστικών Γεγονότων

5.1.1 Μετρικές Αξιολόγησης

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκαν δύο βασικές μετρικές. Όπως ειπώθηκε, η διαδικασία της αναγνώρισης χωρίζεται σε δύο περιπτώσεις. Η πρώτη περίπτωση ονομάζεται isolated training & testing, ενώ η δεύτερη περίπτωση isolated training & embedded testing. Στην πρώτη περίπτωση που τα δεδομένα δοκιμής ήταν μεμονωμένα χρησιμοποιήθηκε η μετρική AEER (acoustic event error rate) ή ποσοστό λάθους αναγνώρισης ακουστικών γεγονότων. Η μετρική αυτή συγκρίνει την αντιστοίχιση των γεγονότων που προκύπτουν από τον ταξινομητή σε σχέση με τα γεγονότα που γνωρίζουμε από το αρχείο αναφοράς, ενώ δε λαμβάνει υπόψη τη χρονική διάρκεια των γεγονότων. Πιο συγκεκριμένα, η μετρική AEER υπολογίζεται ως εξής:

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

ή

$$AEER = \frac{S + D + I}{S + D + H}$$

όπου

- N είναι ο συνολικός αριθμός των παρατηρήσεων
- I είναι ο αριθμός των λαθών εισαγωγής
- D είναι ο αριθμός των λαθών διαγραφής
- S είναι ο αριθμός των λαθών αντικατάστασης
- H είναι ο αριθμός των σωστά ταξινομημένων παρατηρήσεων

Τα σύμβολα I,D και S δηλώνουν τρείς διαφορετικούς τύπους λαθών που χρησιμοποιούνται ευρέως στην αναγνώριση ακουστικών γεγονότων. Πιο συγκεκριμένα έχουμε τις εξής μετρικές λάθους:

- **Λάθος αντικατάστασης (substitution error).** Τα λάθη αντικατάστασης αναφέρονται στα γεγονότα που ταξινομήθηκαν λάθος.
- **Λάθος εισαγωγής (insertion error).** Τα λάθη εισαγωγής προκύπτουν όταν ένα γεγονός αναγνωριστεί και ταξινομηθεί, ενώ στο αρχείο αναφοράς δεν υπάρχει στην ακολουθία των δεδομένων. Αυτό σημαίνει ότι ο ταξινομητής αναγνωρίζει περισσότερα γεγονότα από όσα θα έπρεπε κανονικά να αναγνωρίζει.
- **Λάθος διαγραφής (deletion error).** Τα λάθη διαγραφής αποτελούν την αντίθετη περίπτωση από τα λάθη εισαγωγής, δηλαδή ο ταξινομητής κατά τη διάρκεια της αναγνώρισης χάνει γεγονότα με αποτέλεσμα να αναγνωρίζει λιγότερα γεγονότα.

Για τον υπολογισμό της μετρικής γράφτηκε κώδικας σε Python που δέχεται το αρχείο εξόδου **recout.mlf** καθώς και το αρχείο ετικετών και συγκρίνει τα αποτελέσματα για κάθε ακουστικό γεγονός. Πιο συγκεκριμένα, με βάση το αρχείο ετικετών, ελέγχει αν κάθε ακουστικό γεγονός είναι μεμονωμένο ή επικαλυπτόμενο, ώστε να υπολογίσει αν έχει 1 ή 2 γεγονότα στο συγκεκριμένο αρχείο. Έπειτα ελέγχει το αρχείο εξόδου για να υπολογίσει πόσα από τα γεγονότα αναγνώρισε σωστά, υπολογίζοντας το λόγο της μετρικής AEER.

Στη δεύτερη περίπτωση που τα δεδομένα δοκιμής ήταν ολόκληρη η συνεδρία, χρησιμοποιήθηκε η μετρική FMR (frame misclassification rate) ή ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου που συγκρίνει τα πλαίσια 1-1 και υπολογίζει τρία πράγματα:

- Τα συνολικά λάθη που έκανε ο ταξινομητής
- Το λόγο $\frac{\#\text{frame}-\text{misrecognized}-\text{AEs}}{\#frames}$ (α)
- Το λόγο $\frac{\#frame-misrecognized-AEs}{\#\text{frame}-\text{groundtruth}-\text{AEs}}$(β)

Για του υπολογισμό της μετρικής FMR γράφτηκε κώδικας σε python που δέχεται το αρχείο εξόδου **recout.mlf** και το αρχείο σχολιασμού των δεδομένων δοκιμής και συγκρίνει τα αποτελέσματα σε επίπεδο πλαισίου. Πιο συγκεκριμένα ο κώδικας εκτελεί τα παρακάτω βήματα:

- Ανοίγει το αρχείο σχολιασμού των δεδομένων δοκιμής και ελέγχει τη συνολική διάρκεια του ακουστικού αρχείου. Υπολογίζει πόσα πλαίσια των 10ms υπάρχουν συνολικά στο αρχείο και αναλύει το κάθε γεγονός σε πλαίσια.
- Ανοίγει το αρχείο εξόδου **recout.mlf** και εκτελεί την ίδια διαδικασία
- Συγκρίνει τα πλαίσια 1-1 από τα δύο αρχεία υπολογίζοντας πόσα από τα γεγονότα αναγνώρισε σωστά
- Υπολογίζει τους δύο λόγους, οι οποίοι αποτελούν το ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου.

Πιο συγκεκριμένα στη διαδικασία σύγκρισης των πλαισίων εκτελούνται οι εξής ενέργειες. Για κάθε γεγονός που υπάρχει στο κάθε πλαίσιο ελέγχουμε αν αποτελεί μεμονωμένο γεγονός ή υπάρχει επικάλυψη με κάποιο άλλο. Έτσι δημιουργούνται οι εξής 4 περιπτώσεις σύγκρισης ακουστικών γεγονότων ώστε να υπολογίσουμε την απόδοση του ταξινομητή:

	1η περίπτωση	*2η περίπτωση*	*3η περίπτωση*	*4η περίπτωση*
Αναγνώριση του ταξινομητή	Μεμονωμένο	Μεμονωμένο	Επικαλυπτόμενο	Επικαλυπτόμενο
Επισημείωση	Μεμονωμένο	Επικαλυπτόμενο	Μεμονωμένο	Επικαλυπτόμενο

Η μετρική FMR υπολογίζει τα λάθη του ταξινομητή ανάλογα με την περίπτωση της σύγκρισης ως εξής:

1. 1 λάθος αν αναγνωρίσει λάθος το μεμονωμένο γεγονός, αλλιώς κανένα λάθος
2. 1 λάθος καθώς απέτυχε να αναγνωρίσει δύο γεγονότα, και άλλο 1 λάθος αν το γεγονός που αναγνώρισε ήταν λάθος
3. 1 λάθος καθώς αναγνώρισε δύο γεγονότα ενώ έπρεπε να αναγνωρίσει ένα, και άλλο 1 λάθος αν κανένα από τα δύο γεγονότα που αναγνώρισε δεν ήταν το σωστό
4. 2 λάθη αν δεν αναγνωρίσει σωστά κανένα γεγονός, 1 λάθος αν αναγνωρίσει σωστά ένα γεγονός, και κανένα λάθος αν κάνει σωστή αναγνώριση

5.1.2 Αποτελέσματα Πειραμάτων

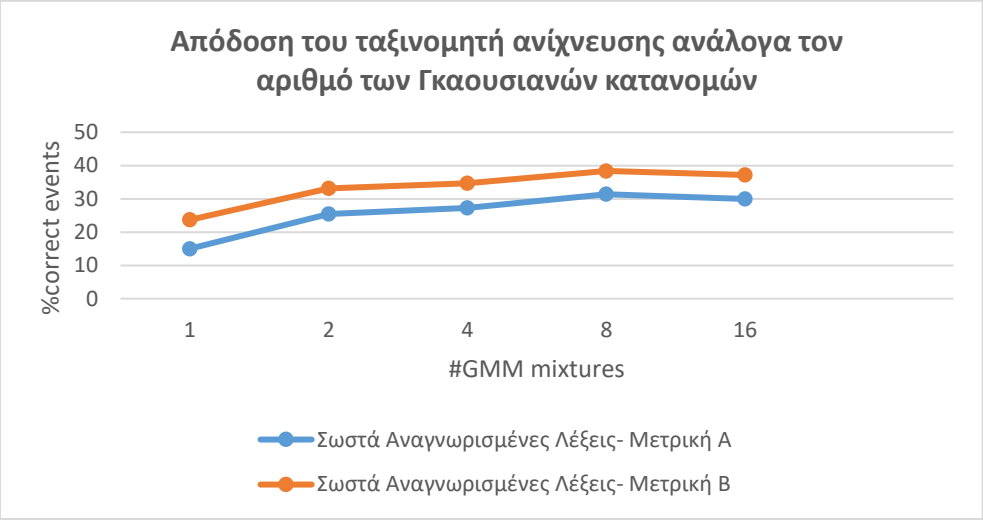
Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν πάνω στη βάση δεδομένων. Αρχικά δοκιμάσαμε την περίπτωση isolated training & testing, δηλαδή εκπαιδεύσαμε το σύστημα με μεμονωμένα κομμάτια ήχου, και το ελέγξαμε με τον ίδιο τρόπο, δηλαδή τα δεδομένα δοκιμής τμηματοποιήθηκαν με τον ίδιο τρόπο. Ως δεδομένα εκπαίδευσης χρησιμοποιήθηκαν τα δεδομένα από τις συνεδρίες 1-6, ενώ ως δεδομένα δοκιμής χρησιμοποιήθηκαν τα δεδομένα τις συνεδρίας 7. Εφόσον το HTK μας δίνει τη δυνατότητα να επεξεργαστούμε τη γραμματική μας σε πολύ χαμηλό επίπεδο, δημιουργήσαμε μία γραμματική η οποία να παράγει μόνο προτάσεις μίας λέξης. Με αυτό το τρόπο, τα αποτελέσματα των μετρήσεων δεν εξαρτώνται από την μεταβλητή του penalty ή της λάθος εισαγωγής λέξης. Επιπλέον το HTK εξ' ορισμού χρησιμοποιεί μία Γκαουσιανή κατανομή για την εκπαίδευση των γεγονότων. Όπως αναφέραμε και προηγουμένως, τρέξαμε τα πειράματά μας επαναληπτικά για 2, 4, 8, και 16 Γκαουσιανές.

Το HTK με το εργαλείο **HResults** υπολογίζει δύο μετρικές. Η πρώτη μετρική με το όνομα SENT αποτελεί την ακρίβεια του συστήματος να αναγνωρίσει μία πρόταση ή αλληλουχία γεγονότων. Έχοντας όμως ορίσει τη γραμματική μας ώστε να παράγει μόνο μεμονωμένα γεγονότα αυτή η μετρική δεν μας δίνει κάποιο χρήσιμο αποτέλεσμα. Η δεύτερη μετρική με το όνομα WORD είναι ανάλογη του AEER και αναφέρεται στην επιτυχία του ταξινομητή να αναγνωρίσει τα μεμονωμένα γεγονότα. Οι παράμετροι μέσα στις αγκύλες αποτελούν τις ίδιες παραμέτρους με αυτές που ορίσαμε προηγουμένως και μας δίνουν τις ίδιες πληροφορίες. Καθώς, όμως, ο ταξινομητής μας υπολογίζει και επικαλυπτόμενα γεγονότα, αυτή η μετρική αδυνατεί να ξεχωρίσει σωστά τα γεγονότα, με αποτέλεσμα να πρέπει να επεξεργαστούμε τα αποτελέσματα που παίρνουμε από το HTK ώστε να υπολογίζουμε σωστά τις λάθος αναγνωρίσεις του συστήματός μας, και πιο συγκεκριμένα να μετράμε τα λάθη εισαγωγής και διαγραφής. Αυτό το πετύχαμε με τον ίδιο τρόπο που υπολογίσαμε και τα λάθη στο FMR.

Στον Πίνακα 2 παρουσιάζονται τα αποτελέσματα του ταξινομητή για τις διάφορες Γκαουσιανές κατανομές και στο Σχήμα 18 παρουσιάζεται γραφικά η ακρίβεια του ταξινομητή ανάλογα με τον αριθμό των Γκαουσιανών κατανομών που χρησιμοποιήθηκαν με βάση τις δύο μετρικές.

Πλήθος GMM κατανομών	Σωστά Αναγνωρισμένες Λέξεις (%) Μετρική Α	Σωστά Αναγνωρισμένες Λέξεις (%) Μετρική Β
1	15.00	23.68
2	25.46	33.10
4	27.28	34.70
8	31.37	38.37
16	30.00	37.15
Βέλτιστο Ποσοστό Λάθους(%)	68.32%	**61.63%**

Πίνακας 2 Αποτελέσματα του ταξινομητή ανίχνευσης για isolated testing



Σχήμα 18 Διάγραμμα της απόδοσης του ταξινομητή ανίχνευσης για isolated testing ανάλογα τον αριθμό των Γκαουσιανών κατανομών

Όπως παρατηρούμε, η απόδοση του ταξινομητή αυξάνεται με τη χρήση περισσότερων Γκαουσιανών κατανομών. Για πλήθος Γκαουσιανών κατανομών ίσο με 8 πετυχαίνουμε το καλύτερο αποτέλεσμα στις αναγνωρισμένες λέξεις, ενώ αν αυξήσουμε τις κατανομές σε 16 το σύστημά μας μειώνει την απόδοση του και γίνεται υπερεκπαιδευμένο (overtrained) με αποτέλεσμα την αύξηση του AEER.

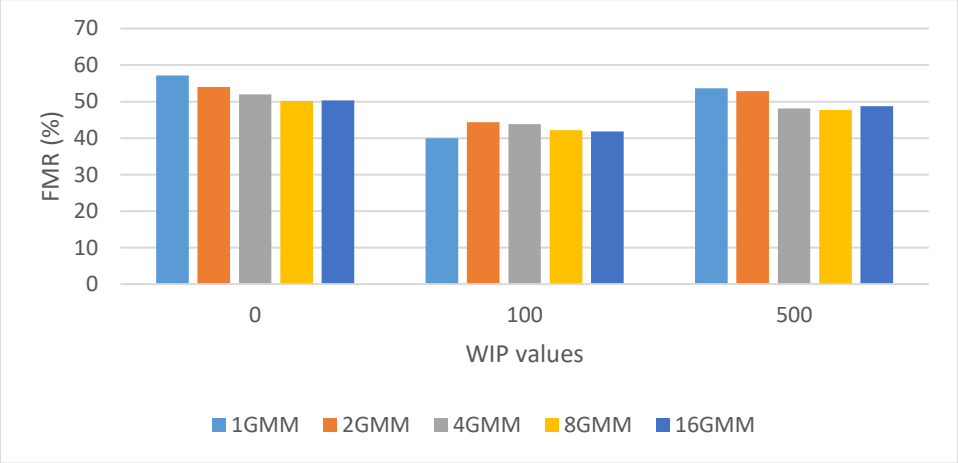
Σε δεύτερη φάση δοκιμάσαμε τον ταξινομητή για τη περίπτωση του **isolated training & embedded testing**. Το σύνολο εκπαίδευσης παρέμεινε το ίδιο, ενώ το σύνολο δοκιμής και η γραμματική άλλαξαν. Στην περίπτωση αυτή ορίσαμε τη γραμματική μας έτσι ώστε να παράγει μία ακολουθία γεγονότων. Επιπλέον για την αξιολόγηση του ταξινομητή χρησιμοποιήσαμε ως σύνολο δοκιμής ολόκληρη τη συνεδρία 7. Το αρχείο **recout.mlf** που παίρνουμε από την αποκωδικοποίηση Viterbi έχει την ίδια σχεδόν μορφή με το αρχείο αναφορά. Όμως σε αυτή τη περίπτωση που έχουμε αλληλουχία γεγονότων οι μετρικές που μας δίνει το εργαλείο **HResults** δεν μας βοηθάνε να βγάλουμε κάποιο

συμπέρασμα για την επιτυχία του ταξινομητή με αποτέλεσμα να χρησιμοποιήσουμε τις *μετρικές ποσοστού λάθους ταξινόμησης σε επίπεδο πλαισίου (FMR)*.

Και στις τέσσερις περιπτώσεις ακολουθείται μία ρουτίνα που υπολογίζει τα λάθη του ταξινομητή. Επιπλέον, σε αυτή την περίπτωση η παράμετρος της λάθος εισαγωγής λέξης (word insertion penalty ή WIP) παίζει ρόλο στα αποτελέσματα. Εκτελώντας την εντολή του **HVite** για διαφορετικές τιμές της παραμέτρου του WIP παίρνουμε τα αποτελέσματα που φαίνονται στον Πίνακα 3 με βάση τη Μετρική Α και στον Πίνακα 4 με βάση τη Μετρική Β. Στο Σχήμα 19 και Σχήμα 20 βλέπουμε γραφικά την απόδοση του ταξινομητή.

	Λανθασμένες Προβλέψεις-Μετρική Α (%)				
WIP	***1GMM***	***2GMM***	**4GMM**	**8GMM**	**16GMM**
0	57.15	53.96	51.95	50.12	50.34
100	39.97	44.33	43.78	42.17	41.81
500	53.62	52.82	48.08	47.68	48.77
Ποσοστό λάθους FMR			**14152/35381=39.97**		

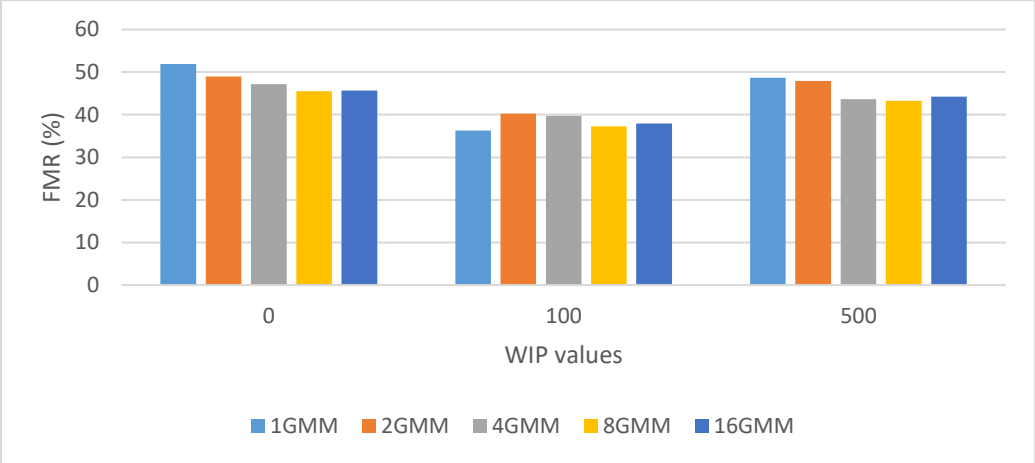
Πίνακας 3 Αποτελέσματα του ταξινομητή ανίχνευσης για embedded testing και ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου με βάση τη Μετρική Α



Σχήμα 19 Γραφική παράσταση της απόδοσης του ταξινομητή ανίχνευσης με βάση τη Μετρική Α

	Λανθασμένες Προβλέψεις-Μετρική Β (%)				
WIP	***1GMM***	***2GMM***	**4GMM**	**8GMM**	**16GMM**
0	51.88	48.99	47.16	45.50	45.70
100	36.28	40.25	39.74	37.28	37.96
500	48.68	47.95	43.65	43.29	44.27
Ποσοστό λάθους FMR				**36.28%**	

Πίνακας 4 Αποτελέσματα του ταξινομητή ανίχνευσης για embedded testing και ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου με βάση τη Μετρική Β



Σχήμα 20 Γραφική παράσταση της απόδοσης του ταξινομητή ανίχνευσης με βάση τη Μετρική Β

Σύμφωνα με τους παραπάνω πίνακες παρατηρούμε ότι το βέλτιστο ποσοστό λάθους αγγίζει το 36%, ή διαφορετικά ότι οι επιτυχείς προβλέψεις του ταξινομητή έφτασαν το 64%, που είναι ικανοποιητικό αν αναλογιστούμε ότι υπάρχουν επικαλύψεις μεταξύ των γεγονότων και ότι καθώς τα γεγονότα συμβαίνουν υπάρχουν παύσεις. Επίσης παρατηρούμε ότι η χρήση περισσότερων Γκαουσιανών κατανομών βελτίωσε την απόδοση του ταξινομητή μας. Για πλήθος Γκαουσιανών κατανομών ίσο με 8 πετυχαίνουμε το 2ο καλύτερο αποτέλεσμα στις αναγνωρισμένες λέξεις, ενώ αν αυξήσουμε τις κατανομές σε 16 ή παραπάνω το σύστημά μας μειώνει την απόδοση του και γίνεται υπερεκπαιδευμένο (overtrained) με αποτέλεσμα να αυξάνονται οι λάθος αναγνωρίσεις.

5.2 Εντοπισμός Επικαλυπτόμενων Ακουστικών Πηγών

5.2.1 Μετρική Αξιολόγησης

Η μετρική που χρησιμοποιήθηκε για να ελέγξουμε τη απόδοση του συστήματός μας υπολογίζει τα λάθη του συστήματος ως εξής:

- 1 λάθος, αν εντοπίσει λανθασμένα ένα μεμονωμένο γεγονός
- 1 λάθος, αν εντοπίσει σωστά μόνο ένα γεγονός στην περίπτωση επικαλυπτόμενων γεγονότων σε διαφορετικές θέσεις
- 1 λάθος, αν στην περίπτωση επικαλυπτόμενων γεγονότων στην ίδια θέση εντοπίσει ένα σωστά
- 2 λάθη, αν στην περίπτωση επικαλυπτόμενων γεγονότων δεν εντοπίσει σωστά κανένα

Με βάση τις παραπάνω περιπτώσεις λαθών, η απόδοση του συστήματος υπολογίζεται σύμφωνα με τις δύο μετρικές :

1. $$\frac{\#\ frame\_correct\_localized\_AEs}{\#frames}$$

2. $$\frac{\#\ frame\_correct\_localized\_AEs}{\sum ground\_truth\_AEs}$$

5.2.2 Αποτελέσματα Πειραμάτων

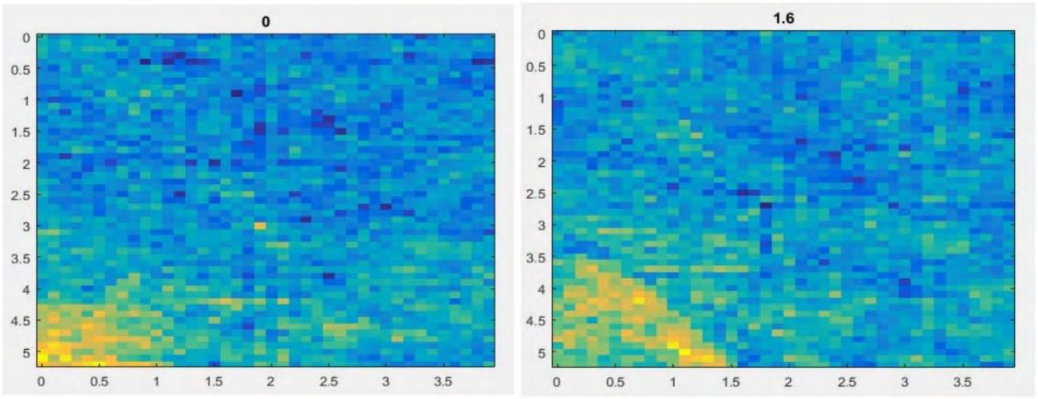
Τα πειράματα πραγματοποιηθήκαν πάνω στα ηχητικά δεδομένα από τις συνεδρίες T01-T03. Το σύστημα αναγνωρίζει τρεις περιπτώσεις δεδομένων:

- 1 ακουστική πηγή
- 2 ακουστικές πηγές σε διαφορετικές θέσεις
- 2 ακουστικές πηγές στην ίδια θέση

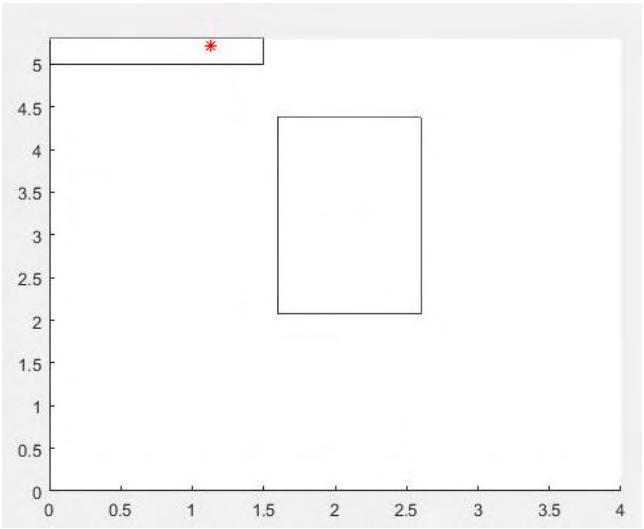
Καθώς δεν υπήρχε αρχείο αναφοράς σχετικά τις θέσεις των πηγών, ο έλεγχος των αποτελεσμάτων έγινε συνδυάζοντας τα αρχεία βίντεο καθώς και τα δεδομένα από το Σχήμα 1, που μας δείχνουν τις γενικές περιοχές που συμβαίνει κάθε γεγονός. Για αυτό το λόγο, δημιουργήθηκε κώδικας όπου ορίζονται οι 4 διαφορετικοί χώροι του δωματίου όπου μπορούν να συμβούν ηχητικά γεγονότα. Έπειτα, τα τελικά σημεία όπου εντοπίστηκαν οι πηγές, ελέγχονται για το αν ανήκουν σε έναν από τους χώρους.

Η διαδικασία της αξιολόγησης, όπως και στην αξιολόγηση για τη μέθοδο της Ανίχνευσης Επικαλυπτόμενων Ακουστικών Γεγονότων, χωρίζεται σε δύο κατηγορίες. Την κατηγορία isolated testing και την κατηγορία embedded testing.

Στην πρώτη κατηγορία δώσαμε ως είσοδο στο σύστημά μας ηχητικά κομμάτια που αποτελούνταν από ένα γεγονός, είτε μεμονωμένο είτε επικαλυπτόμενο. Δοκιμάσαμε το σύστημα για όλες τις περιπτώσεις γεγονότων και τα αποτελέσματα φαίνονται παρακάτω. Στο Σχήμα 21 φαίνονται οι τιμές των SRP-PHAT για κάθε σημείο του χώρου στην περίπτωση που υπήρχε μόνο *μία ακουστική πηγή*, ενώ στο Σχήμα 22 φαίνεται το τελικό σημείο του χώρου όπου εντοπίστηκε το γεγονός. Οι τιμές 0 και 1.6 αποτελούν την τιμή για την z-συντεταγμένη του χώρου, ενώ ο οριζόντιος άξονας αποτελεί την x-συντεταγμένη και ο κάθετος άξονας την y-συντεταγμένη. Παρατηρούμε ότι οι μέγιστες τιμές του SRP-PHAT υπολογίζονται σωστά κοντά στο κομμάτι του χώρου όπου βρίσκεται η πόρτα. Στο Σχήμα 22 φαίνεται το έξυπνο δωμάτιο που εξετάζουμε με τους χώρους της πόρτας και του τραπεζιού να είναι ζωγραφισμένοι. Με την κουκίδα επισημαίνεται το τελικό σημείο του χώρου στο οποίο εντοπίστηκε η ακουστική πηγή από το σύστημα.



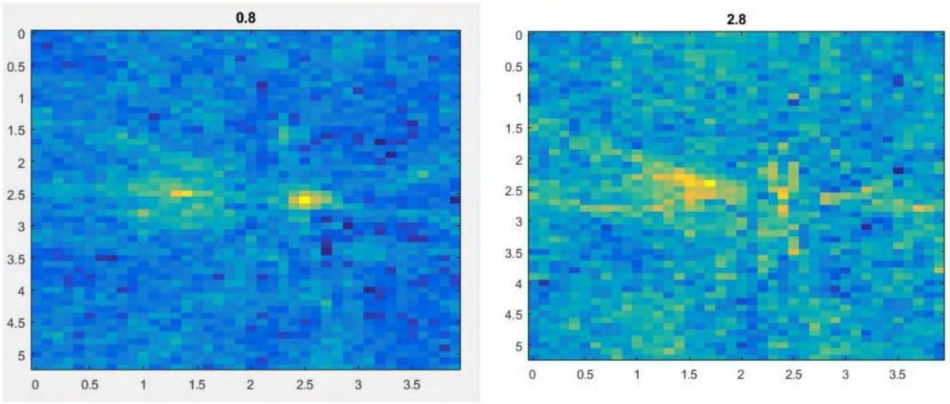
Σχήμα 21 Τιμές SRP-PHAT για τον εντοπισμό του ακουστικού γεγονότος ds



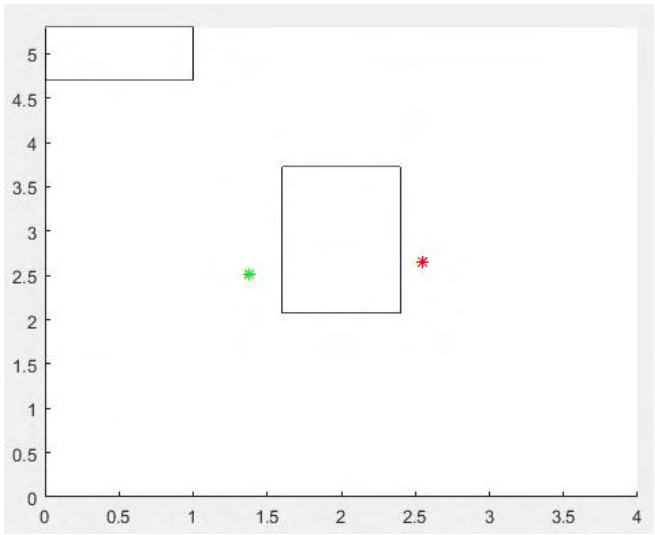
Σχήμα 22 Εντοπισμός ακουστικού γεγονότος ds

Στο Σχήμα 23 φαίνονται οι τιμές των SRP-PHAT για κάθε σημείο του χώρου στην περίπτωση που υπήρχαν *δύο ακουστικές πηγές σε διαφορετικές θέσεις* και στο Σχήμα 24 φαίνονται τα τελικά σημεία του χώρου όπου εντοπίστηκαν τα δύο γεγονότα. Παρατηρούμε ότι το σύστημα αναγνωρίζει σωστά τις

συντεταγμένες των δύο ακουστικών πηγών, οι οποίες προέρχονται από τις δύο πλευρές του τραπεζιού. Επίσης, παρατηρούμε ότι για ύψος 0.8 m το σύστημα αναγνωρίζει με μεγαλύτερη ακρίβεια ότι υπάρχουν δύο πηγές στο χώρο, σε αντίθεση με την περίπτωση που το ύψος είναι στα 2.8 m. Το συγκεκριμένο γεγονός που προσπαθεί το σύστημα να αναγνωρίσει είναι η ομιλία του ενός ανθρώπου και το χτύπημα των χεριών από τον άλλο ομιλητή, οι οποίοι κάθονται απέναντι ο ένας από τον άλλον στο γραφείο του δωματίου. Επομένως, το σύστημα αντιλαμβάνεται καλύτερα τον ήχο σε ένα σημείο που βρίσκεται στο ίδιο περίπου ύψος με το πραγματικό γεγονός.

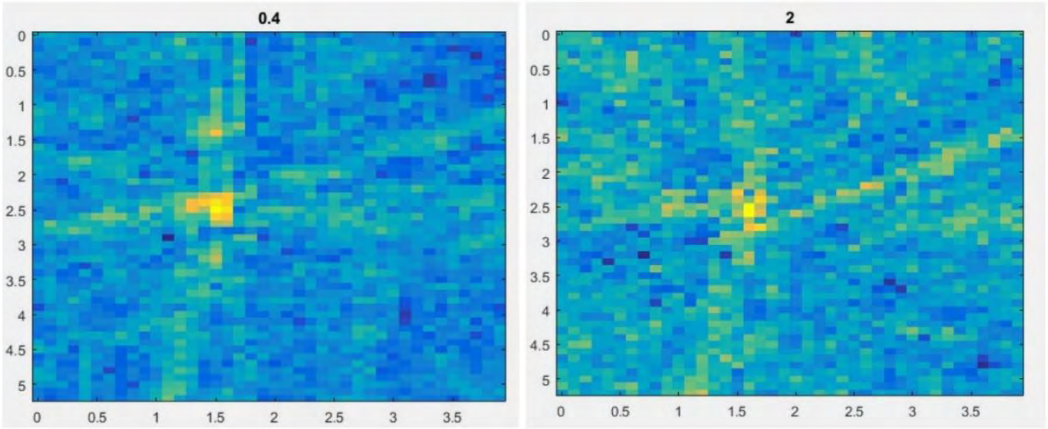


Σχήμα 23 Τιμές SRP-PHAT για τον εντοπισμό των ακουστικών γεγονότων sp-cl



Σχήμα 24 Εντοπισμός ακουστικών γεγονότων sp-cl

Στο Σχήμα 25 φαίνονται οι τιμές των SRP-PHAT για κάθε σημείο του χώρου στην περίπτωση που υπήρχαν *δύο επικαλυπτόμενα ακουστικά γεγονότα στην ίδια θέση* μέσα στο χώρο, όπως για παράδειγμα όταν ο ομιλητής μιλάει και ταυτόχρονα κουνάει τα κλειδιά του. Σε αυτή τη περίπτωση το σύστημα βγάζει μία τοποθεσία για τα γεγονότα. Το αποτέλεσμα θεωρείται σωστό καθώς η ένταση των δύο γεγονότων δε διαφέρει αρκετά και η τοποθεσία από όπου προέρχονται είναι η ίδια.



Σχήμα 25 Τιμές SRP-PHAT για τον εντοπισμό των ακουστικών γεγονότων sp-kj

Στον Πίνακα 5 φαίνονται τα αποτελέσματα του συστήματος για τον εντοπισμό των ακουστικών πηγών στις συνεδρίες T01-T03.

	T01	T02	T03
$\#\ correctly\ localized\ AEs$	181	189	164
$\#\ erroneously\ localized\ AEs$	38	54	42
$\frac{\#\ frame\_wrong\_localized\_AEs}{\sum ground\_truth\_AEs} 100\%$	17.4%	22.3%	20.4%

Πίνακας 5 Αποτελέσματα συστήματος εντοπισμού ακουστικών πηγών για μεμονωμένα γεγονότα εισόδου

Στη δεύτερη κατηγορία δώσαμε ως είσοδο στο σύστημα ολόκληρα αρχεία ήχου, που αποτελούνταν από πολλά διαδοχικά ακουστικά γεγονότα. Για κάθε frame της διαδικασίας που περιγράψαμε προηγουμένως, πήραμε τα αποτελέσματα σχετικά με το πόσες πηγές αναγνωρίστηκαν και σε ποιο χώρο του δωματίου. Για τον έλεγχο των δεδομένων, δημιουργήθηκε script αρχείο που χωρίζει τα δεδομένα του αρχείου αναφοράς σε αντίστοιχα frames της ίδιας διάρκειας, ενώ οι συντεταγμένες τους και το κομμάτι του χώρου που τοποθετήθηκαν εκτιμήθηκαν κατά προσέγγιση παρακολουθώντας το αντίστοιχο βίντεο. Η συνεδρία T07 δοκιμάστηκε για τις δύο περιπτώσεις του Αλγορίθμου (χωρίς temporal smoothing και με temporal smoothing). Λόγω των αυξημένων υπολογισμών και επαναλήψεων του αλγορίθμου, η συνεδρία διασπάστηκε σε μικρότερα κομμάτια και στα οποία έγινε ο έλεγχος. Τα αποτελέσματα από τα δύο πειράματα φαίνονται στον Πίνακα 6.

	T07 (όχι temporal smoothing)	**T07** (με temporal smoothing-part 1)	**T07** (με temporal smoothing-part 2)
$\#\ \boldsymbol{correctly\ localized\ AEs}$	15432	2306	2002
$\#\ \boldsymbol{erroneously\ localized\ AEs}$	19257	1095	1651
$\frac{\#\ \boldsymbol{frame\_wrong\_localized\_AEs}}{\sum \boldsymbol{ground\_truth\_AEs}}\ \mathbf{100}\%$	54%	48%	56%
$\frac{\#\ \boldsymbol{frame\_wrong\_localized\_AEs}}{\#\boldsymbol{frames}}\ \mathbf{100}\%$	60%	44%	54%

Πίνακας 6 Αποτελέσματα συστήματος εντοπισμού ακουστικών πηγών για embedded testing

Σύμφωνα με τα πειράματα, παρατηρήθηκε ότι τα περισσότερα λάθη πραγματοποιήθηκαν σε 3 περιπτώσεις:

1. Όταν δεν υπήρχε κανένα ακουστικό γεγονός, αλλά λόγω θορύβου και κάποιας αντήχησης το σύστημα αναγνώριζε ένα ακουστικό γεγονός.
2. Όταν κάποια ακουστική πηγή άλλαζε θέση στο χώρο κατά τη διάρκεια ενός γεγονότος.
3. Λόγω λανθασμένης επισημείωσης στο αρχείο αναφοράς.

Αξίζει να σημειωθεί πως η εύρεση της τέλειας λύσης για το χώρο που μελετάμε δεν είναι εφικτή εξαιτίας των παρακάτω τριών παραγόντων:

- Ανακρίβειες στη γεωμετρία των πηγών αλλά και των μικροφώνων
- Έλλειψη ακριβούς αρχείου με τις συντεταγμένες του κάθε ακουστικού αρχείου
- Παρουσία θορύβου και αντήχησης

5.3 Συνδυασμός των Μεθόδων

5.3.1 Μετρική Αξιολόγησης

Για την αξιολόγηση των αποτελεσμάτων, χρησιμοποιήθηκε η μετρική FMR (frame misclassification rate) ή ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου που συγκρίνει τα πλαίσια 1-1 και υπολογίζει τρία πράγματα:

- Τα συνολικά λάθη που έκανε ο ταξινομητής

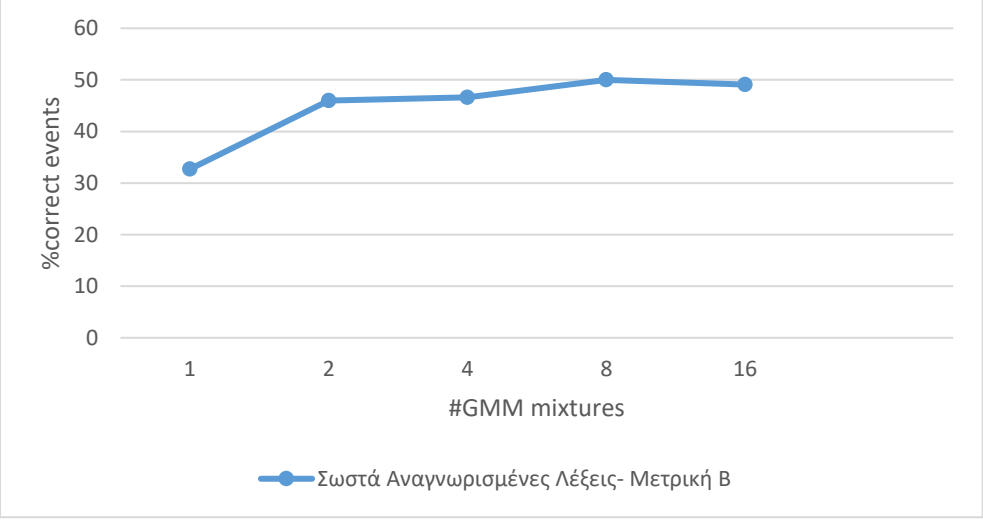
- Το λόγο $\frac{\#\text{frame}-\text{misrecognized}-\text{AEs}}{\#frames}$ (α)
- Το λόγο $\frac{\#frame-misrecognized-AEs}{\#\text{frame}-\text{groundtruth}-\text{AEs}}$(β)

5.3.2 Αποτελέσματα Πειραμάτων

Τα πειράματα και σε αυτή την περίπτωση χωρίστηκαν στις δύο κατηγορίες, isolated testing και embedded testing. Στη πρώτη περίπτωση ελέγξαμε τα δεδομένα της συνεδρίας T07 ως μεμονωμένα γεγονότα και στον Πίνακα 7 και Πίνακα 8 φαίνονται τα αποτελέσματα που προέκυψαν.

Πλήθος GMM κατανομών	Λάθος Αναγνωρισμένες Λέξεις (%) Μετρική B
1	67.30
2	54.00
4	53.40
8	50.00
16	50.90
Βέλτιστο Ποσοστό Λάθους(%)	**50.00%**

Πίνακας 7 Αποτελέσματα του ταξινομητή για το συνδυασμό μεθόδων για μεμονωμένα δεδομένα δοκιμής

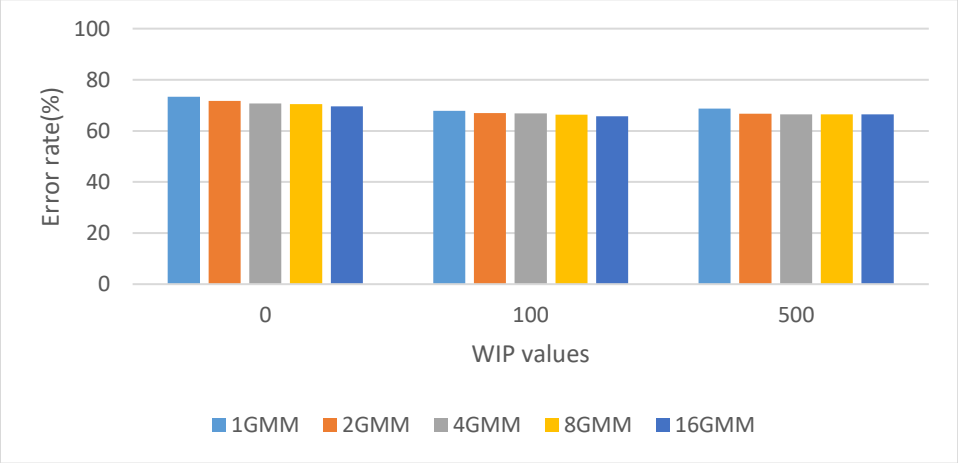


Πίνακας 8 Διάγραμμα της απόδοσης του ταξινομητή για το συνδυασμό μεθόδων για μεμονωμένα δεδομένα δοκιμής ανάλογα με τον αριθμό των Γκαουσιανών κατανομών

Σε δεύτερη φάση δοκιμάστηκε το σύστημα έχοντας ως σύνολο δοκιμής ολόκληρη τη συνεδρία T07. Ακολουθώντας την ίδια διαδικασία και χρησιμοποιώντας τις ίδιες μετρικές, προέκυψαν τα αποτελέσματα που φαίνονται στον Πίνακα 9 και στον Πίνακα 10 με βάση τη μετρική Α και στον Πίνακα 11 και Πίνακα 12 με βάση τη μετρική Β.

	Λανθασμένες Προβλέψεις-Μετρική Α (%)				
WIP	***1GMM***	***2GMM***	**4GMM**	**8GMM**	**16GMM**
0	73.30	71.67	70.67	70.42	69.61
100	67.80	66.90	66.80	66.35	65.65
500	68.70	66.68	66.44	66.41	66.41
Ποσοστό λάθους FMR			**66.35%**		

Πίνακας 9 Αποτελέσματα του ταξινομητή για το συνδυασμό μεθόδων για embedded testing και ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου με βάση τη Μετρική Α



Πίνακας 10 Γραφική παράσταση της απόδοσης του ταξινομητή για το συνδυασμό μεθόδων σε embedded testing με βάση τη Μετρική Α

	Λανθασμένες Προβλέψεις-Μετρική Β (%)				
WIP	***1GMM***	***2GMM***	**4GMM**	**8GMM**	**16GMM**
0	36.60	35.83	35.33	35.21	34.8
100	33.90	33.40	33.40	33.10	32.82
500	35.09	33.80	33.22	33.20	33.20
Ποσοστό λάθους FMR			**33.00%**		

Πίνακας 11 Αποτελέσματα του ταξινομητή συνδυασμού μεθόδων για embedded testing και ποσοστό λάθους ταξινόμησης σε επίπεδο πλαισίου με βάση τη Μετρική Β

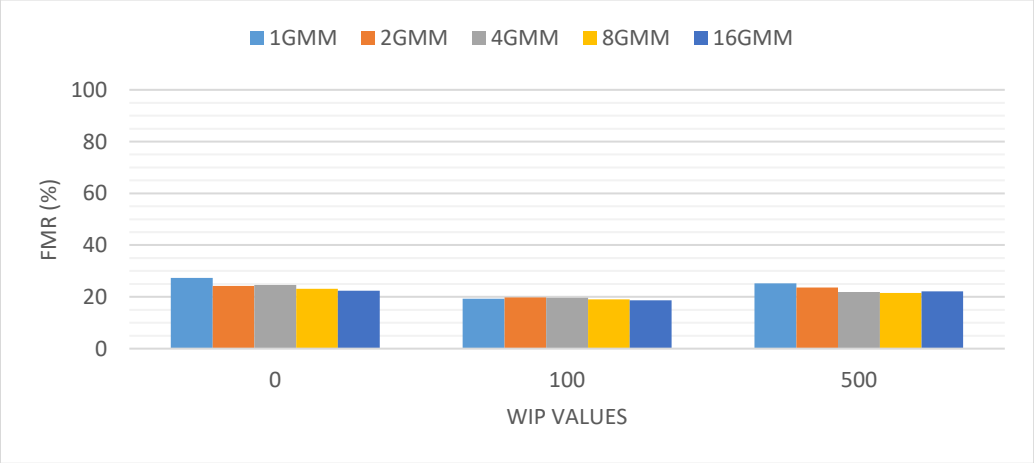
Πίνακας 12 Γραφική παράσταση της απόδοσης του ταξινομητή για το συνδυασμό μεθόδων σε embedded testing με βάση τη Μετρική Β

Επιπλέον έγιναν δοκιμές με την προσθήκη χρονικής ομαλοποίησης (temporal smoothing). Πιο συγκεκριμένα, θεωρήθηκε ένα τυχαίο βήμα των 100 frames. Για κάθε frame ακολουθήθηκε η διαδικασία που περιεγράφηκε παραπάνω και στη συνέχεια για κάθε 100 frames υπολογίσθηκε το ακουστικό γεγονός που ανιχνεύτηκε τις περισσότερες φορές. Το γεγονός με τις περισσότερες ψήφους σε κάθε 100 frames θεωρήθηκε ότι είναι το πιο πιθανό προς αναγνώριση, δημιουργώντας με αυτό τον τρόπο μία καινούργια υπόθεση με τα πιο πιθανά ακουστικά γεγονότα. Η καινούργια υπόθεση που δημιουργήθηκε ελέγχθηκε για την απόδοσή της με τον ίδιο τρόπο. Τα αποτελέσματα που προέκυψαν παρουσιάζονται στον Πίνακα 13, και αναλύονται γραφικά στον Πίνακα 14 για τη μετρική Α και στον Πίνακα 15 για τη μετρική Β.

	Λανθασμένες Προβλέψεις-Μετρική Α (%)					Λανθασμένες Προβλέψεις-Μετρική Β (%)				
WIP	***1GMM***	***2GMM***	**4GMM**	**8GMM**	**16GMM**	**1GMM**	**2GMM**	**4GMM**	**8GMM**	**16GMM**
0	54.58	48.47	49.23	46.18	44.65	27.29	24.23	24.61	23.09	22.32
100	38.50	39.60	39.31	38.16	37.40	19.27	19.80	19.60	19.08	18.70
500	51.52	47.32	43.89	43.12	44.27	25.27	23.66	21.94	21.56	22.13
Ποσοστό λάθους FMR			**38.16%**					**19.08%**		

Πίνακας 13 Αποτελέσματα του ταξινομητή συνδυασμού μεθόδων με temporal smoothing για embedded testing με βάση τις μετρικές Α και Β

Πίνακας 14 Γραφική παράσταση της απόδοσης του ταξινομητή συνδυασμού μεθόδων για embedded testing με temporal smoothing με βάση τη Μετρική Α



Πίνακας 15 Γραφική παράσταση της απόδοσης του ταξινομητή συνδυασμού μεθόδων για embedded testing με temporal smoothing με βάση τη Μετρική Β

ΚΕΦΑΛΑΙΟ 6-ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1 Ανασκόπηση της Διπλωματικής

Η παρούσα διπλωματική παρουσιάζει τη συστηματική μελέτη στον τομέα της ανίχνευσης και του εντοπισμού επικαλυπτόμενων ακουστικών γεγονότων, χρησιμοποιώντας τα σήματα που λαμβάνουν οι συστοιχίες μικροφώνων σε ένα έξυπνο δωμάτιο. Η εργασία ξεκίνησε με την περιγραφή του προβλήματος, καθώς και συναφείς εργασίες. Ακολούθησε η μοντελοποίηση του προβλήματος, το μαθηματικό υπόβαθρο των μεθοδολογιών, καθώς και η περιγραφή της βάσης. Έπειτα, αναλύθηκε ο τρόπος διεξαγωγής των πειραμάτων και παρουσιάστηκαν τα αποτελέσματά τους. Η διαδικασία των πειραμάτων μπορεί να χωριστεί σε τρεις κατηγορίες:

1. Ανίχνευση επικαλυπτόμενων ακουστικών γεγονότων
2. Εντοπισμός επικαλυπτόμενων ακουστικών γεγονότων
3. Συνδυασμός των δύο μεθόδων

Με βάση τα πειράματα, παρατηρήθηκε ότι το σύστημα ανίχνευσης αναγνώρισε σωστά το 38.5% των ακουστικών γεγονότων για την περίπτωση του isolated testing και το 63% των ακουστικών γεγονότων για την περίπτωση του embedded testing. Στο σύστημα εντοπισμού ακουστικών πηγών για isolated testing η απόδοση του συστήματος ήταν στο 80%, ενώ για embedded testing ήταν στο 45%. Έπειτα, ο συνδυασμός των δύο μεθόδων για isolated testing ανέβασε το ποσοστό επιτυχίας στο 50%, ενώ για embedded testing το ποσοστό επιτυχίας έφτασε στο 67%. Τέλος, στο σύστημα με το συνδυασμό των μεθόδων για embedded testing έγινε προσθήκη χρονικής ομαλοποίησης, με αποτέλεσμα το σύστημα να αναγνωρίζει ακόμα πιο αποτελεσματικά τα ακουστικά γεγονότα, πετυχαίνοντας ποσοστό επιτυχίας 81%.

Εξετάζοντας τα αποτελέσματα είναι φανερό, πως με το συνδυασμό των δύο μεθόδων και τη χρονική ομαλοποίηση, η αποτελεσματικότητα του συστήματος αυξάνεται ραγδαία. Συμπεραίνουμε, λοιπόν, πως μία λύση στο πρόβλημα της ανίχνευσης των επικαλυπτόμενων ακουστικών γεγονότων είναι ο εντοπισμός των ακουστικών πηγών και ο αριθμός τους πριν τη διαδικασία της ανίχνευσης.

6.2 Πιθανές Μελλοντικές Κατευθύνσεις

Παρόλο που με τη διεκπεραίωση της παρούσας διπλωματικής εργασίας η μελέτη και τα αποτελέσματα των συστημάτων ανίχνευσης και εντοπισμού ήταν θετικά, υπάρχουν ακόμα περιθώρια βελτίωσης και επέκτασης αυτών των συστημάτων. Ορισμένες μελλοντικές κατευθύνσεις ορίζονται παρακάτω:

- Επέκταση του συστήματος ανίχνευσης με συνδυασμό των καναλιών στο στάδιο της εκπαίδευσης, καθώς κάποια γεγονότα μπορεί να έχουν ηχογραφηθεί καλύτερα από κάποιο άλλο μικρόφωνο
- Εφαρμογή διαφορετικών ακουστικών χαρακτηριστικών, όπως τα μη-φασματικά χαρακτηριστικά με θετικό συνελικτικό πίνακα παραγοντοποίησης (convolutive non-negative matrix factorization)

- Επέκταση του συστήματος εντοπισμού ώστε να ανιχνεύει περισσότερες από δύο επικαλυπτόμενες ακουστικές πηγές
- Έλεγχος των συστημάτων με δεδομένα άλλων βάσεων για επανεξέταση της απόδοσής τους

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, and M. Omologo, “Acoustic Event Detection and Classification,” in *Computers in the Human Interaction Loop*, A. Waibel and R. Stiefelhagen, Eds. Springer London, 2009, pp. 61–73.

[2] J. R. Hershey, P. A. Olsen, S. J. Rennie, A. Aaron, “Audio Alchemy: Getting Computers to Understand Overlapping Speech,” *Scientific American*. [Online]. Available: http://www.scientificamerican.com/article/speech-getting-computers-understand-overlapping/. [Accessed: 05-Oct-2016].

[3] J. Dennis, H. D. Tran, and E. S. Chng, “Overlapping sound event recognition using local spectrogram features and the generalised Hough transform,” *Pattern Recognit. Lett.*, vol. 34, no. 9, pp. 1085–1093, Jul. 2013.

[4] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Context-dependent sound event detection,” *EURASIP J. Audio Speech Music Process.*, vol. 2013, no. 1, pp. 1-4, Jan. 2013.

[5] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. V. hamme, “An exemplar-based NMF approach to audio event detection,” in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[6] “Practical Cryptography.” [Online]. Available: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/. [Accessed: 05-Oct-2016].

[7] P. Giannoulis, G. Potamianos, P. Maragos, and A. Katsamanis, “Improved dictionary selection and detection schemes in sparse CNMF-based overlapping acoustic event detection,” *Detect. Classif. Acoust. Scenes Events 2016*, Sep. 2016.

[8] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, “An integrated framework for multi-channel multi-source localization and voice activity detection,” in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011, pp. 92–97.

[9] M. Omologo and P. Svaizer, “Acoustic event localization using a crosspower-spectrum phase based technique,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94*, 1994, vol. 2, pp. 273-276.

[10] P. Mermelstein, “Distance Measures for speech recognition- psychological and instrumental,” *Int. J. Pattern Recognit. Artif. Intell.*, pp. 374–388, 1976.

[11] B. Kotnik, D. Vlaj, Z. Kacic, and B. Horvat, “Robust MFCC feature extraction algorithm using efficient additive and convolutional noise reduction procedures,” Proceedings of *ICSLP*, pp. 445–448, 2002.

[12] M. Lutter, “Mel-Frequency Cepstral Coefficients,” *SR Wiki*, 25-Nov-2014. .

[13] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, “Two-source acoustic event detection and localization: Online implementation in a smart-room,” in *EUSIPCO*, 2011, pp. 1317–1321.

[14] D. B. Paul, “Speech recognition using hidden Markov models,” *Linc. Lab. J.*, vol. iii, no. 1, 1990.

[15] Z. Ghahramani, “An introduction to hidden Markov models and Bayesian networks,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 15, no. 1, pp. 9–42, Feb. 2001.

[16] D. Reynolds, “Gaussian Mixture Models,” in *Encyclopedia of Biometrics*, S. Z. Li and A. K. Jain, Eds. Springer US, 2015, pp. 827–832.

[17] H. Do, H. F. Silverman, and Y. Yu, “A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array,” in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, vol. 1, pp. 121-124.

[18] M. F. Berger and H. F. Silverman, “Microphone array optimization by stochastic region contraction,” *IEEE Trans. Signal Process.*, vol. 39, no. 11, pp. 2377–2386, Nov. 1991.

[19] A. Temko, R. Malkin, D. Macho, C. Nadeu, and M. Omologo, “UPC-TALP database of isolated meeting-room acoustic events,” *Springer-Verl. Berl. Heidelb.*, pp. 311 – 322, 2007.
[20] S. Young *et al.*, *The HTK Book*, vol. 3. Entropic Cambridge Research Laboratory Cambridge, 1997.
[21] N. Moreau, “HTK: Basic Tutorial.” 02-Feb-2002.
[22] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, “Overlapped speech detection for improved speaker diarization in multiparty meetings,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4353–4356.
[23] H. Do, *SRP-PHAT*. 2009.