Ανίχνευση Ακουστικών Γεγονότων και Ανάλυση Ακουστικού Περιβάλλοντος

του Πεχλιβάνη Πλάτων

9 Οκτωβρίου 2012

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον καθηγητή κ. Γεράσιμο Ποταμιάνο για την επίβλεψη αυτής της διπλωματικής εργασίας, την καθοδήγησή του και για την εξαιρετική συνεργασία που είχαμε. Ακόμη θα ήθελα να ευχαριστήσω τον Παναγιώτη Γιαννούλη, διπλωματούχο μηχανικό της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, για την προετοιμασία και προεπεξεργασία της βάσης δεδομένων που χρησιμοποιήθηκε στα πειράματα της διπλωματικής εργασίας. Επίσης ευχαριστώ τον φίλο μου Κώστα Χατζηιωάννου για τη βοήθεια του σχετικά με την εγγραφή της διπλωματικής σε LaTex. Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου και τους φίλους μου για την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Περίληψη

Η συγκεκριμένη διπλωματική έχει σαν αντικείμενο τη μελέτη και την επίλυση του προβλήματος της κατηγοριοποίησης των ακουστικών γεγονότων, τα οποία λαμβάνουν χώρα κατά τη διάρκεια συνεδριάσεων σε χώρο εξοπλισμένο με πολλαπλά μικρόφωνα. Για τους σκοπούς των πειραμάτων που πραγματοποιούνται χρησιμοποιείται μια βάση δεδομένων η οποία διαθέτει 24 κανάλια καταγραφής των γεγονότων. Βασικός σκοπός της διπλωματικής αποτελεί η βελτίωση των αποτελεσμάτων του προβλήματος της κατηγοριοποίησης με το συνδυασμό των καναλιών αυτών. Αρχικά, γίνεται μια συστηματική και συνοπτική επισκόπηση των βασικών ειδών των ακουστικών χαρακτηριστικών και ταξινομητών της βιβλιογραφίας. Στη συνέχεια, επιλέγονται τα MFCC χαρακτηριστικά και ο GMM ταξινομητής για τη δημιουργία μοντέλων που συνδυάζουν τα ηχητικά σήματα των καναλιών στο επίπεδο της απόφασης, αλλά και στο επίπεδο του ηχητικού σήματος με την χρησιμοποίηση του BeamformIt, ενός λογισμικού που λαμβάνει ως είσοδο εναν αριθμό σημάτων και παράγει ένα ενισχυμένο/βελτιωμένο σήμα. Ακόμη, χρησιμοποιούνται κάποιες μετρικές εκτίμησης της καναλικής πληροφορίας, όπως ο σηματοθορυβικός λόγος (SNR), η διαφορά της λογαριθμικής πιθανοφάνειας και η διασπορά της λογαριθμικής πιθανοφάνειας, για την επιλογή και το συνδυασμό μόνο των καλύτερων καναλιών. Τέλος, γίνεται προσπάθεια βελτίωσης των αποτελεσμάτων με το συνδυασμό των MFCC χαρακτηριστικών και των χρονικών καθυστερήσεων των σημάτων (TDOA features) κατά τη φάση της κατηγοριοποίησης. Το μοντέλο που χρησιμοποιεί για κάθε συνεδρία το καλύτερο βάσει SNR κανάλι και συνδυάζει MFCC και TDOA χαρακτηριστικά κατά την κατηγοριοποίηση των ακουστικών γεγονότων επιτυγχάνει ποσοστό αναγνώρισης 96,7%. Ίδιο ποσοστό επιτυγχάνει και το μοντέλο που χρησιμοποιεί για κάθε συνεδρία το καλύτερο κανάλι με βάση την μετρική της διαφοράς της λογαριθμικής πιθανοφάνειας. Οι επιδόσεις αυτές αποτελούν τις καλύτερες που λαμβάνουμε και είναι κατά 1,4% (απόλυτα) βελτιωμένες από την επίδοση του μοντέλου που κάνει χρήση μόνο του πρώτου καναλιού. Η βελτίωση αυτή αντιστοιχεί σε 30% σχετική μείωση του σφάλματος.

Abstract

The aim of this thesis is to study classification of acoustic events that may occur in a meeting room environment, equipped with multiple far-field microphones. For the experiments a database is used that has 24 audio channels available. Our main target in this thesis is to improve the classification results with the fusion of these channels. Initially, a systematic and compact review of the bibliography main acoustic features and classifiers is made. Based on our results, we choose to use MFCC features and GMM classifier to create models of acoustic events. For channel combination, both desicion-level fusion and signal-level fusion are employed, the latter using BeamformIt, which is a software that takes as input a number of signals and produces an enhanced one. Furthermore, several metrics are used, such as SNR, log-likelihood difference and log-likelihood dispersion, to select the best channels to participate in the fusion process. Finally, an attempt to improve classification results is made by fusing MFCC features with the time delays of the signals (TDOA features) during the classification phase. The model that selects the best SNR channel for each session and performs MFCC and TDOA fusion during the classification achieves 96.7% accuracy. The same classification rate is achieved by the model that selects the best channel for each session according to the log-likelihood difference metric. These two results are the best that we achieved in experiments and they are 1.4% improved compared to the classification rate of the model that uses only the first channel. This improvement corresponds to a 30% relative reduction in classification error.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Ανίχνευση και Κατηγοριοποίηση Ακουστικών Γεγονότων

Παρ' όλο που η ομιλία είναι κατά γενική ομολογία το ακουστικό γεγονός που περιέχει τις περισσότερες και σημαντικότερες πληροφορίες, υπάρχουν και άλλοι ήχοι που εμπεριέχουν χρήσιμες πληροφορίες, όπως για παράδειγμα οι ήχοι που διαδραματίζονται σε μία αίθουσα συνεδριάσεων. Σε τέτοιου είδους περιβάλλοντα η ανθρώπινη δραστηριότητα παράγει μία μεγάλη γκάμα ακουστικών γεγονότων, τα οποία προκαλούνται είτε από τον ίδιο τον άνθρωπο είτε από τα αντικείμενα που αυτός χρησιμοποιεί. Παραδείγματα τέτοιων ήχων αποτελούν ο λόγος κάποιου ομιλητή, το κλείσιμο/άνοιγμα της πόρτας, ο ήχος του πληκτρολογίου, το κούνημα μιας καρέκλας, βηματισμοί και πολλοί άλλοι.

Η Ανίχνευση και Κατηγοριοποίηση Ακουστικών Γεγονότων (Acoustic Event Detection/Classification) [1] αποτελούν μέρος της Υπολογιστικής Ακουστικής Ανάλυσης Σκηνής [2] η οποία ασχολείται με την επεξεργασία ακουστικών σημάτων και την μετατροπή τους σε συμβολικές περιγραφές. Ενώ η κατηγοριοποίηση ακουστικών γεγονότων ασχολείται με γεγονότα τα οποία έχουν απομονωθεί από το χρονικό τους πλαίσιο, η ανίχνευση ακουστικών γεγονότων έγκειται στην αναγνώριση και τον εντοπισμό τους μέσα από μία συνεχή ηχητική ροή. Οι ήχοι αυτοί, σε αντίθεση με τους Ηλεκτρονικούς Υπολογιστές, αναγνωρίζονται σχετικά εύκολα από το ανθρώπινο αυτί, το οποίο επιδεικνύει εξαιρετική ανεκτικότητα σε μια σειρά από ανωμαλίες που μπορεί να έχει υποστεί το ακουστικό σήμα, οπως π.χ. ο θόρυβος.

Σκοπός των τεχνολογιών αυτών αποτελεί η πλήρης κατανόηση και ερμηνεία της σκηνής και της ανθρώπινης δραστηριότητας που εξελίσσεται μέσα σε αυτήν καθώς και η βελτίωση της αποτελεσματικότητας

των τεχνολογιών ομιλίας, όπως π.χ. τα συστήματα αυτόματης αναγνώρισης φωνής (Automatic Speech Recognition Systems). Βασικά σημεία στην ανίχνευση αλλά και στην κατηγοριοποίηση των ακουστικών γεγονότων αποτελούν η εξαγωγή των ιδιαίτερων χαρακτηριστικών (Feature Extraction) κάθε ηχητικού σήματος και η εκπαίδευση ενός ταξινομητή (Classifier Training) για την δημιουργία του μοντέλου που θα χρησιμοποιηθεί αργότερα κατά την φάση της κατηγοριοποίησης.

1.2 Περιεχόμενο και δομή της παρούσας εργασίας

1.2.1 Σκοπός της διπλωματικής

Η παρούσα διπλωματική ασχολείται με το πρόβλημα της κατηγοριοποίησης διαφόρων ακουστικών γεγονότων που λαμβάνουν χώρα σε ένα "έξυπνο" περιβάλλον εξοπλισμένο με πολλαπλά μικρόφωνα, όπως π.χ. μια αίθουσα συνεδριάσεων κατά τη διάρκεια κάποιου σεμιναρίου ή συνεδρίασης. Οι στόχοι της είναι η εφαρμογή και αξιολόγηση των κύριων μεθόδων κατηγοριοποίησης, η μελέτη και αξιολόγηση των ιδιαίτερων χαρακτηρηστικών που εξάγονται από τους ήχους καθώς και των ταξινομητών που χρησιμοποιούνται, και τέλος η βελτίωση των αποτελεσμάτων με τη δημιουργία μοντέλων που συνδυάζουν τις ηχητικές πληροφορίες των πολλαπλών καναλιών που καταγράφουν την συνεδρίαση.

1.2.2 Συνεισφορά της διπλωματικής

Η κύρια συνεισφορά της συγκεκριμένης διπλωματικής έγκειται στη συστηματική μελέτη του προβλήματος της κατηγοριοποίησης ακουστικών γεγονότων τα οποία λαμβάνουν χώρα κατά την διάρκεια μιας συνεδρίασης με την χρησιμοποίηση πληροφορίας από πολλαπλά ηχητικά κανάλια και η αξιολόγηση των διαφορετικών μοντέλων και προσεγγίσεων στο πρόβλημα αυτό. Πιο συγκεκριμένα οι επιστημονικές της συνεισφορές συνοψίζονται στους εξής άξονες:

- Στην μελέτη μιας σειράς από τρόπους συνδυασμού των καναλιών σε περιβάλλοντα στα οποία χρησιμοποιούνται πολλά μικρόφωνα για την καταγραφή των ήχων.
- Στην χρησιμοποίηση μετρικών εκτίμησης της ποιότητας της καναλικής πληροφορίας για την επιλογή και το συνδυασμό των καλύτερων καναλιών.

- Στο συνδυασμό των χρόνων καθυστέρησης των ηχητικών σημάτων κάθε καναλιού (TDOA χαρακτηριστικά) με τα κλασικά MFCC χαρακτηριστικά για την βελτίωση της επίδοσης των μοντέλων κατηγοριοποίησης.

1.2.3 Οργάνωση του περιοχομένου της διπλωματικής

Το περιεχόμενο της διπλωματικής είναι οργανωμένο σε πέντε ακόμη κεφάλαια, πέραν του εισαγωγικού, ως εξής:

- Στο **κεφάλαιο 2** περιγράφονται συνοπτικά οι κύριες κατηγορίες χαρακτηριστικών που χρησιμοποιούνται στα προβλήματα της ανίχνευσης και κατηγοριοποίησης ακουστικών γεγονότων. Συγκεκριμένα γίνεται αναφορά στα RASTA-PLP, MFCC και AM-FM ηχητικά χαρακτηριστικά.

- Στο **κεφάλαιο 3** περιγράφονται συνοπτικά οι βασικές κατηγορίες ταξινομητών. Συγκεκριμένα αναφέρονται ο ταξινομητής των k κοντινότερων γειτόνων (k nearest neighbors - kNN), το μοντέλο μείγματος Γκαουσιανών (Gaussian Mixture Model - GMM) και οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM).

- Στο **κεφάλαιο 4** μελετάται ο συνδυασμός των ηχητικών σημάτων από πολλαπλά κανάλια με σκοπό την βελτίωση των αποτελεσμάτων της κατηγοριοποίησης. Αρχικά, υλοποιούνται δύο μοντέλα με συνδυασμό καναλιών στο επίπεδο της απόφασης (decision level fusion). Στη συνέχεια αναλύονται μοντέλα που προέρχονται από την συνένωση πολλαπλών ηχητικών σημάτων (signal level fusion) και περιγράφεται η βελτίωση της επίδοσής τους με την επιλογή των καλύτερων, σύμφωνα με συγκεκριμένα μέτρα, καναλιών. Τέλος, γίνεται αναφορά και στο συνδυασμό των κλασικών MFCC χαρακτηριστικών με τα χαρακτηριστικά που υπολογίζονται από τις καθυστερήσεις στους χρόνους άφιξης των σημάτων (Time Delay Of Arrival - TDOA).

- Στο **κεφάλαιο 5** περιγράφεται συνοπτικά η βάση δεδομένων η οποία χρησιμοποιήθηκε για τον έλεγχο των μοντέλων στο πλαίσιο της παρούσας εργασίας και παρουσιάζονται συγκεντρωτικά όλα τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν και περιγράφηκαν στα προηγούμενα κεφάλαια.

- Τέλος, στο **κεφάλαιο 6** παρουσιάζονται τα συμπεράσματα που προκύπτουν από το σύνολο της διπλωματικής και συνοψίζονται

οι επιστημονικές συνεισφορές της. Επίσης, αναφέρονται και κάποιες μελλοντικές κατευθύνσεις και προεκτάσεις της.

1.3 Επισκόπηση της βιβλιογραφίας

Οι ροές ακουστικών γεγονότων όπως η μετάδοση ειδήσεων, ηχογραφήσεις συνεδριάσεων και προσωπικά βίντεο περιέχουν ήχους από μια ποικιλία πηγών. Τέτοια παραδείγματα περιλαμβάνουν ακουστικά γεγονότα που σχετίζονται με την ανθρώπινη παρουσία, όπως η ομιλία, γέλια ή βήχας, ή και ήχους ζώων και αντικειμένων. Η ανίχνευση αυτών των γεγονότων είναι χρήσιμη, π.χ. για την αυτόματη ανάλυση ή ταξινόμησή τους σε ένα συγκεκριμένο πλαίσιο. Ένα πλαίσιο ακουστικών γεγονότων χαρακτηρίζεται από την παρουσία των επιμέρους συμβάντων, όπως η περιγραφή ενός ακουστικού αρχείου και επίσης τον εντοπισμό των κατηγοριών ακουστικών γεγονότων [3].

Είναι γεγονός ότι το ερευνητικό πεδίο της επεξεργασίας ακουστικών σημάτων και ειδικότερα των σημάτων εκείνων που δεν ανήκουν στην κατηγορία της ανθρώπινης ομιλίας έχει λάβει λιγότερη προσοχή απ' ότι το πεδίο της τεχνολογίας ομιλίας [4], [5]. Επισημαίνεται στην υπάρχουσα βιβλιογραφία ότι κατά τα τελευταία χρόνια ο μεγαλύτερος όγκος της έρευνας έχει προσανατολιστεί προς την ανάλυση του λόγου. Αυτό οφείλεται στον αδιαμφισβήτητο ρόλο που έχει η ομιλία ως τον πιο φυσικό και διαδεδομένο τρόπο επικοινωνίας μεταξύ των ανθρώπων αλλά και στο μεγάλο αριθμό εφαρμογών που απαιτούν την αλληλεπίδραση μεταξύ ανθρώπου και μηχανής.

Ωστόσο, τα τελευταία χρόνια πραγματοποιτοποιείται συστηματική έρευνα πάνω στα προβλήματα της ανίχνευσης και της κατηγοριοποίησης πολλών ακουστικών γεγονότων, που παρουσιάζονται μέσα στο περιβάλλον στο οποίο συναναστρέφονται οι άνθρωποι, πέραν της ομιλίας. Αρχικά, η έρευνα στο πεδίο της κατηγοριοποίησης ήχων περιοριζόταν στη χρήση ένος μικρού αριθμού κλάσεων όπως π.χ. ομιλία/μουσική ή μουσική/ομιλία/άλλο, όπου στο "άλλο" περιλαμβάνονταν όλοι οι υπόλοιποι περιβαλλοντικοί ήχοι. Πλέον χρησιμοποιούνται βάσεις δεδομένων οι οποίες περιέχουν μια μεγάλη γκάμα ακουστικών γεγονότων οι οποίοι σχετίζονται είτε με ένα συγκεκριμένο περιβάλλον είτε με κάποια συγκεκριμένη δραστηριότητα. Χαρακτηρηστικό παράδειγμα αποτελούν οι βάσεις δεδομένων που καταγράφουν τους ήχους κατά τη διάρκεια συνεδριάσεων.

Η ανίχνευση ακουστικών γεγονότων μέσα στα πλαίσια των ροών ήχου και τη μετέπειτα κατάταξη είναι ένας σχετικά νέος τομέας έρευνας και τα υφιστάμενα συστήματα βασίζονται στην χρήση τεχνικών ταξινόμησης [6]. Τυπικά παραδείγματα είναι η χρήση των κρυφών μο-

ντέλων Markov (HMM) ή οι Μηχανές Υποστήριξης Διανυσμάτων (SVM). Οι προσεγγίσεις αυτές απαιτούν την ταξινόμηση των δεδομένων για τη βελτιστοποίηση των παραμέτρων των συστημάτων.

Οι ανιχνεύσεις ακουστικών συμβάντων έχουν ευρεία εφαρμογή. Πιο συγκεκριμένα η συλλογή πληροφοριών που σχετίζονται με ακουστικά γεγονότα οδηγεί στην αποκάλυψη ανθρώπινων κοινωνικών δραστηριοτήτων. Τέτοια παραδείγματα περιλαμβάνουν την μετακίνηση μιας καρέκλας, το ανοιγοκλείσιμο μιας πόρτας κατά την διάρκεια μιας σύσκεψης, τις επευφημίες ενός ακροατηρίου κατά την διάρκεια ενός αθλητικού γεγονότος, έναν πυροβολισμό στον δρόμο ή τα βιαστικά βήματα σε ένα σπίτι. Η συλλογή, ανάλυση και ταξινόμηση των ακουστικών γεγονότων είναι χρήσιμη σε εφαρμογές επιτήρησης για λόγους ασφάλειας, πολυμέσα ανάκτησης πληροφοριών και ευφυών αιθουσών. Μερικά από τα ακουστικά γεγονότα είναι συγκριτικά σημαντικότερα, όπως οι επευφημίες, ενώ άλλα είναι πιο διακριτά, όπως τα βήματα στον διάδρομο μιας αίθουσας συσκέψεων [7].

Βασικός στόχος των ερευνητών είναι η βελτίωση της αποτελεσματικότητας και της επίδοσης των μοντέλων που χρησιμοποιούνται στην ανίχνευση και κατηγοριοποίηση ακουστικών γεγονότων. Γι' αυτόν το λόγο γίνεται συστηματική προσπάθεια για συνένωση των ήδη χρησιμοποιούμενων στην αναγνώριση φωνής ηχητικών χαρακτηριστικών ή ακόμη και δημιουργία νέων, ικανών να αντιπροσωπεύσουν μια μεγάλη γκάμα ήχων πέραν της ομιλίας. Ακόμη, για τον ίδιο σκοπό δοκιμάζονται και μελετώνται τρόποι συνδυασμού αλλά και τροποποιήσεις των κύριων ταξινομητών που χρησιμοποιούνται κατά την φάση της κατηγοριοποίησης. Τέλος, ο συνδυασμός της πληροφορίας των πολλαπλών καναλιών που καταγράφουν τα ηχητικά γεγονότα είναι ένας τομέας για τον οποίο δεν έχει γίνει συστηματική έρευνα και ο οποίος μπορεί να επιφέρει αισθητή βελτίωση των αποτελεσμάτων της κατηγοριοποίησης. Ο συνδυασμός των καναλιών μπορεί να πραγματοποιηθεί είτε στο επίπεδο της απόφασης είτε στο επίπεδο των σημάτων. Ακόμη, υπάρχει στη βιβλιογραφία ένας μικρός αριθμός εργασιών που ασχολούνται με το συνδυασμό των ηχητικών και οπτικών πληροφοριών που καταγράφονται από τα μικρόφωνα και τις κάμερες του δωματίου της συνεδρίασης [8].

Οι συγκεκριμένες τεχνολογίες έχουν ακόμα να αντιμετωπίσουν διαφόρων ειδών ανοιχτά ζητήματα. Η πλειονότητα της ερευνητικής δραστηριότητας έχει επικεντρωθεί σε βάσεις "καθαρές" από κάθε είδους παραμόρφωση, όπως π.χ. ο θόρυβος. Εκτός από τις βάσεις αυτές υπάρχουν και οι βάσεις δεδομένων οι οποίες περιέχουν ακουστικά συμβάντα επικαλυπτόμενα με ομιλία, καθιστώντας την ανίχνευση και κατηγοριοποίηση των γεγονότων ιδιαίτερα προκλητική [9]. Ακόμη, είναι δυνατόν τα ακουστικά γεγονότα να μην είναι απομονωμένα και

να υπάρχει επικάλυψη μεταξύ τους. Γίνεται λοιπόν κατανοητό πως τη μεγαλύτερη πρόκληση αποτελεί ο πειραματισμός πάνω σε βάσεις δεδομένων πραγματικού κόσμου (real-world). Στη βιβλιογραφία υπάρχει μικρός αριθμός τέτοιων εργασιών.

Κεφάλαιο 2

Εξαγωγή Χαρακτηριστικών

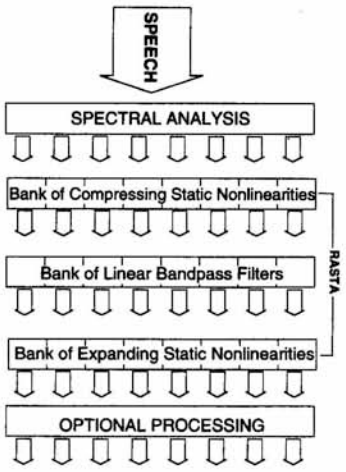
2.1 Εισαγωγή

Η φάση της εξαγωγής των ακουστικών χαρακτηριστικών είναι ίσως το σημαντικότερο ζήτημα στην διαδικασία της κατηγοριοποίησης και αποτελείται από τη μετατροπή του ακουστικού σήματος σε μια σειρά διανυσμάτων μικρών διαστάσεων, όπου κάθε ένα συνοψίζει ένα τμήμα του σήματος. Η διαδικασία εξαγωγής χαρακτηριστικών αποτελείται από έναν αριθμό διαδοχικών επιπέδων. Αρχικά, μετά την ηχογράφηση και προεπεξεργασία των σημάτων, πραγματοποιείται ο χωρισμός και ανάλυση μικρών χρονικών τμημάτων (short-time analysis). Στη συνέχεια, υπολογίζονται τα χαρακτηριστικά στο πεδίο της συχνότητας και τέλος εφαρμόζεται ο μετασχηματισμός των ακουστικών παραμέτρων. Στη συνέχεια του κεφαλαίου θα περιγραφούν συνοπτικά τα τρία είδη των ιδιαίτερων ακουστικών χαρακτηριστικών που χρησιμοποιήθηκαν στα πειράματα: τα RASTA-PLP, τα MFCCs και τα AM-FM.

2.2 Χαρακτηριστικά RASTA-PLP

Τα PLP χαρακτηριστικά (Perceptually weighted Linear Prediction) βασίζονται στην εξαγωγή ενός λείου φάσματος βασισμένο στην γραμμική πρόβλεψη και στάθμιση με βάρη που προκύπτουν από ακουστικές καμπύλες [10]. Μια εξέλιξη των PLP αποτελούν τα RASTA-PLP, τα οποία έχουν καλύτερη συμπεριφορά ως προς τον θόρυβο, και προκύπτουν φιλτράροντας τις αργές φασματικές μεταβολές [11]. Τα βήματα για τον υπολογισμό των RASTA-PLP χαρακτηριστικών φαίνονται στο Σχήμα 2.2 και είναι τα εξής:

- Υπολογισμός του κριτικού-εύρους φάσματος ισχύος



Σχήμα 2.1: Βασικά στάδια για τον υπολογισμό των RASTA-PLP συντελεστών. Έχει ληφθεί από το [11]

- Μετασχηματισμός του φασματικού πλάτους μέσω ενός συμπιεσμένου, στατικού, μη-γραμμικού μετασχηματισμού
- Φιλτράρισμα της τροχιάς χρόνου καθεμιάς από τις μετασχηματισμένες φασματικές συνιστώσες
- Μετασχηματισμός της φιλτραρισμένης ηχητικής αναπαράστασης μέσω ενός επεκταμένου, στατικού, μη-γραμμικού μετασχηματισμού
- Πολλαπλασιασμός με την καμπύλη ίσης έντασης για την προσομείωση της ακουστικής ανάλυσης ισχύος
- Υπολογισμός ενός μοντέλου ολικών πόλων από το προκύπτων φάσμα και χρήση του μοντέλου αυτού για την εύρεση των συντελεστών

2.3 Χαρακτηριστικά MFCC

Τα πιο διαδεδομένα ακουστικά χαρακτηριστικά είναι το διάνυσμα MFCC (Mel Frequency Cepstral Coefficients) [12], [13]. Αν και προέρχονται από το πεδίο της αυτόματης αναγνώρισης ομιλίας, χρησιμοποιούνται ευρέως σε πολλές εφαρμογές κατηγοριοποίησης ακουστικών σημάτων. Τα MFCC αποτελούν μια προσέγγιση του ανθρώπινου ακουστικού συστήματος, καθώς λαμβάνουν υπόψη τη μη-γραμμική φύση

της αντίληψης των θεμελιωδών συχνοτήτων, αλλά και τη μη-γραμμική σχέση μεταξύ έντασης και ηχηρότητας. Αυτές οι ιδιότητες κάνουν τα MFCC αρκετά επαρκή γνωρίσματα για την αναγνώριση ομιλίας. Η επιτυχία των MFCC, συνδυασμένη με τον προτυποποιημένο και υπολογιστικά αποδοτικό υπολογισμό τους, τα μετέτρεψε σε τυποποιημένη επιλογή και σε άλλα πεδία όπως την αναγνώριση ομιλητών, την αναγνώριση γλώσσας, την αναγνώριση συναισθήματος καθώς και σε άλλες εφαρμογές της τεχνολογίας ομιλίας.

Οι συντελεστές MFCC υπολογίζονται με βάση το λογαριθμικό φάσμα ισχύος (power spectrum), το οποίο προκύπτει από το μέτρο του μετασχηματισμού Fourier (FFT) και τη χρήση ενός τριγωνικού filterbank στο πεδίο της συχνότητας, το οποίο ουσιαστικά κάνει δειγματοληψία της ενέργειας του σήματος σε διαφορετικές ζώνες. Στη συνέχεια, ένας μικρός αριθμός συνιστωσών υπολογίζεται με την εφαρμογή του DCT πάνω στις λογαριθμικά συμπιεσμένες εξόδους των φίλτρων.

Τα MFCC χαρακτηριστικά έχουν σχετικά καλή απόδοση αναγνώρισης σε χαμηλά επίπεδα θορύβου. Ωστόσο, η απόδοσή τους φθίνει πολύ γρήγορα καθώς αυξάνεται το επίπεδο του θορύβου, αφού βασίζονται στην φασματική περιβάλλουσα του σήματος.

2.4 Χαρακτηριστικά AM-FM

Στις αρχές της δεκαετίας του '90 προτάθηκε το μοντέλο AM-FM ως ένα εναλλακτικό μη-γραμμικό μοντέλο εμπνευσμένο από τα μη-γραμμικά φαινόμενα κατά την φώνηση. Σύμφωνα με το μοντέλο αυτό το κάθε ηχητικό σήμα αναλύεται σε μια σειρά από ημιτονοειδή σήματα στιγμιαίας συχνότητας και πλάτους [14], [15]. Τα σήματα αυτά θεωρούνται ως χρονο-συχνοτικές κατανομές και περιέχουν σημαντικές ακουστικές πληροφορίες που δεν μπορούν να καταγραφούν από το γραμμικό μοντέλο. Το μοντέλο AM-FM λαμβάνει τα μη-γραμμικά φαινόμενα ως διαμορφώσεις πλάτους (AM) και συχνότητας (FM) στο ακουστικό σήμα.

Τα χαρακτηριστικά AM-FM υπολογίζονται από τα σήματα στιγμιαίας συχνότητας και πλάτους. Τα βασικά βήματα για την εξαγωγή τους είναι τα εξής:

- Εξαγωγή των σημάτων συντονισμού κατά το μήκος των τροχιών των formants με τη χρήση φίλτρων Gabor
- Αποδιαμόρφωση των σημάτων συντονισμού σε στιγμιαίο πλάτος και στιγμιαία συχνότητα

- Υπολογισμός των δεικτών διαμόρφωσης πλάτους και συχνότητας

Κεφάλαιο 3

Ταξινομητές

3.1 Εισαγωγή

Ο ρόλος του ταξινομητή είναι να αντιστοιχίσει τα δεδομένα προς κατηγοριοποίηση, και πιο συγκεκριμένα τα ηχητικά χαρακτηριστικά που εξάγονται από αυτά, σε έναν αριθμό διαφορετικών κατηγοριών. Στη συνέχεια του κεφαλαίου, αφού πρώτα γίνει μια σύντομη αναφορά στις βασικές κατηγορίες των ταξινομητών, θα περιγραφούν συνοπτικά τα τρία βασικά είδη που χρησιμοποιήθηκαν στα πειράματα: ο kNN, ο GMM και ο SVM ταξινομητής.

Κατηγορίες Ταξινομητών

Οι ταξινομητές μπορούν να διακριθούν σε δύο βασικές κατηγορίες: στους διαχωριστικούς (discriminative) και στους μη-διαχωριστικούς (non-discriminative).

Οι διαχωριστικοί ταξινομητές εκπαιδεύονται για να ελαχιστοποιήσουν το λάθος ταξινόμησης σε ένα σύνολο από δεδομένα εκπαίδευσης. Συνεπώς, πρέπει μόνο να διαμορφώσουν το όριο μεταξύ των κατηγοριών και είναι ανεκτικοί σε οποιεσδήποτε παραλλαγές μέσα στα όρια αυτά. Οι σημαντικότεροι απο τους διαχωριστικούς ταξινομητές είναι ο multilayer perceptron και οι διανυσματικές μηχανές υποστήριξης (SVM - support vector machines).

Οι μη-διαχωριστικές προσεγγίσεις δεν στοχεύουν άμεσα στην ελαχιστοποίηση του λάθους ταξινόμησης. Μια σημαντική ομάδα μη-διαχωριστικών προσεγγίσεων καλείται παραγωγική (generative). Οι ταξινομητές αυτοί προσπαθούν να χτίσουν την κατανομή στηριζόμενοι απόλυτα στα δεδομένα εκπαίδευσης. Ακόμη, επεξεργάζονται τα δείγματα κάθε κατηγορίας ανεξάρτητα από αυτά των άλλων κατηγοριών. Οι σημαντικότεροι ταξινομητές αυτής της κατηγορίας είναι το κρυμμένο μοντέλο Markov (HMM - hidden Markov models) και το μοντέ-

λο μείγματος Γκαουσσιανών κατανομών (GMM - Gaussian mixture models). Εκτός από τους παραγωγικούς ταξινομητές, η ομάδα των μη-διαχωριστικών ταξινομητών περιλαμβάνει και ταξινομητές που δεν μπορουν να χαρακτηριστούν παραγωγικοί επειδή δεν δημιουργούν τις συναρτήσεις κατανομών των δεδομένων. Χαρακτηριστικό παράδειγμα της κατηγορίας αυτής είναι οι k κοντινότεροι γείτονες (kNN - k-nearest neighbor).

Οι διαχωριστικές, όπως και οι παραγωγικές προσεγγίσεις έχουν τους περιορισμούς τους και κανένας από αυτούς δεν παρέχει μια τέλεια λύση σε πρακτικές εφαρμογές. Επομένως, χρησιμοποιούνται οι υβριδικές προσεγγίσεις που συνδυάζουν ιδιότητες και των δύο κατηγοριών.

3.2 Ταξινομητής kNN

Μια πολύ γνωστή και ευρεία χρησιμοποιούμενη τεχνική κατηγοριοποίησης που βασίζεται στη χρήση μέτρων βασισμένων στην απόσταση είναι αυτή των k κοντινότερων γειτόνων (k nearest neighbors - kNN). Κατά το στάδιο της εκπαίδευσης, ο kNN αποθηκεύει τα διανύσματα όλου του συνόλου εκπαίδευσης (training set) καθώς και τις κλάσεις στις οποίες αυτά ανήκουν. Ουσιαστικά, δηλαδη, αποθηκεύει τα σημεία του πολυδιάστατου χώρου που αντιστοιχούν στο σύνολο εκπαίδευσης μαζί με τις κατηγορίες τους. Αυτό έχει σαν αποτέλεσμα τα δεδομένα εκπαίδευσης να αποτελούν το μοντέλο κατηγοριοποίησης.

Όταν πρόκειται να γίνει μια κατηγοριοποίηση για ένα νέο στοιχείο πρέπει να καθοριστεί η απόστασή του από κάθε στοιχείο του συνόλου εκπαίδευσης. Για τον υπολογισμό της απόστασης υπάρχουν αρκετές μετρικές, μια από τις οποίες είναι η Ευκλείδεια απόσταση η οποία χρησιμοποιήθηκε στα πειράματα και ο υπολογισμός της φαίνεται στην (3.1) :

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \tag{3.1}$$

όπου $\mathbf{x}, \mathbf{y}$ δύο διανύσματα του n-διάστατου διανυσματικού χώρου. Στη συνέχεια κάθε νέο στοιχείο κατατάσσεται στην κατηγορία που πλειοψηφεί μεταξύ των k κοντινότερων διανυσμάτων του συνόλου εκπαίδευσης. Η μεταβλητή k αναφέρεται στον αριθμό των κοντινότερων γειτόνων που έχουν λόγο στην κατηγοριοποίηση. Ο αριθμός αυτός εξαρτάται από το σύνολο εκπαίδευσης που χρησιμοποιείται και συνήθως ρυθμίζεται με τρόπο ώστε η κατηγοριοποίηση να πραγματοποιείται με την υψηλότερη ακρίβεια. Στα πειράματα δόθηκε η τιμή 1 στο k,

που σημαίνει πως κάθε νέο στοιχείο κατατάσσονταν στην κατηγορία του κοντινότερού του γείτονα.

Ο αλγόριθμος kNN απαιτεί περισσότερους υπολογισμούς κατά την κατηγοριοποίηση νέων στοιχείων όσο αυξάνει το πλήθος των διανυσμάτων του συνόλου εκπαίδευσης. Αυτό αποτελεί σημαντικό μειονέκτημα του αλγορίθμου, καθώς παρατηρούνται μεγάλες καθυστερήσεις κατά τη φάση της κατηγοριοποίησης όταν το σύνολο εκπαίδευσης είναι μεγάλο. Τα πλεονεκτήματα του αλγορίθμου είναι η απλότητά του και η ταχύτητα κατά την φάση της εκπαίδευσης.

Υπάρχει επίσης και μια παραλλαγή του αλγορίθμου των k κοντινότερων γειτόνων, η οποία είναι γνωστή με το όνομα k κοντινότερος γείτονας σταθμισμένης απόστασης. Σύμφωνα με την παραλλαγή αυτή, το πόσο συνεισφέρει κάθε γείτονας στην κατηγοριοποίηση υπολογίζεται βάσει ενός βάρους. Οι κοντινότεροι γείτονες έχουν μεγαλύτερη συνεισφορά ενώ οι μακρινότεροι μικρότερη. Χρησιμοποιώντας την μέθοδο αυτή, μπορούμε να λάβουμε υπόψη όλα τα διανύσματα του συνόλου των δεδομένων εκπαίδευσης και όχι μόνο τις κοντινότερες.

3.3 Ταξινομητής GMM

Το μοντέλο μείγματος Γκαουσσιανών κατανομών (Gaussian Mixture Model - GMM) αποτελεί υποπερίπτωση μιας άλλης μεγάλης κατηγορίας ταξινομητών, του κρυμμένου μοντέλου Markov (Hidden Markov Model - HMM). Συγκεκριμένα πρόκειται για ταξινομητές HMM μίας κατάστασης. Οι ταξινομητές που βασίζονται σε GMM παρέχουν μια καλή περιγραφή της κατανομής δεδομένων και αποτελούν μια λογική επιλογή στα προβλήματα κατηγοριοποίησης, δεδομένου ότι αποδίδουν ικανοποιητικά σε πολλές εφαρμογές ταξινόμησης ακουστικών σημάτων.

Το GMM μοντέλο αποτελεί μια παραμετρική συνάρτηση πυκνότητας πιθανότητας η οποία αναπαρίσταται ως ένας γραμμικός συνδυασμός (μείγμα) Γκαουσσιανών κατανομών που χαρακτηρίζονται από διαφορετικές παραμέτρους [16]. Οι παράμετροι του μοντέλου είναι το βάρος κάθε στοιχείου του μείγματος, το διάνυσμα μέσης τιμής (mean vector) και ο πίνακας συνδιακύμανσης (covariance matrix), ο οποίος μπορεί να είναι πλήρης (full) ή διαγώνιος (diagonal). Ο αλγόριθμος k-means χρησιμοποιείται για να προσδώσουμε αρχικές τιμές στις παραμέτρους οι οποίες επανυπολογίζονται από τον επαναληπτικό αλγόριθμο Μεγιστοποίησης Προσδοκίας (Expectation Maximization - EM). Άλλοι λιγότερο "δημοφιλείς" αλγόριθμοι για τον υπολογισμό των παραμέτρων αυτών είναι ο αλγόριθμος Figueiredo-Jain (FJ) και ο άπληστος EM αλγόριθμος (Greedy EM).

Ένα GMM που αποτελείται από M Γκαουσσιανές δίνεται από τον ακόλουθο τύπο:

$$p(\mathbf{x}_t) = \sum_{m=1}^{M} \pi_m N(\mathbf{x}_t, \mu_m, \mathbf{\Sigma}_m) \tag{3.2}$$

όπου το $\mathbf{x}_t$ αποτελεί τους συντελεστές των χαρακτηριστικών στο χρόνο t, το N είναι μία Γκαουσσιανή συνάρτηση με μέσο μ_m και πίνακα συνδιακύμανσης $\mathbf{\Sigma}_m$ ενώ π_m είναι η εκ των προτέρων πιθανότητα της συγκεκριμένης κατάστασης.

Άλλη μια σημαντική παράμετρος κάθε GMM μοντέλου, η οποία δεν αναφέρθηκε παραπάνω είναι ο αριθμός των μειγμάτων (mixtures/components), δηλαδή η μεταβλητή M στον τύπο (3.2). Ο αριθμός των μειγμάτων αντιπροσωπεύει την ευκαμψία του GMM, μπορεί να διαφέρει μεταξύ των κλάσεων του προβλήματος της κατηγοριοποίησης και έχει άμεση σχέση με την απόδοση του μοντέλου. Στα πειράματα που πραγματοποιήθηκαν χρησιμοποιήθηκε GMM μοντέλο 2 μειγμάτων με διαγώνιο πίνακα συνδιακύμανσης.

3.4 Ταξινομητής SVM

Οι Μηχανές Υποστήριξης Διανυσμάτων (Support Vector Machines - SVM) [17] αποτελούν τους πιο γνωστούς και διαδεδομένους διαχωριστικούς ταξινομητές. Πρόκειται για μια μέθοδο μηχανικής μάθησης η οποία χρησιμοποιείται κυρίως για την επίλυση δυαδικών προβλημάτων κατηγοριοποίησης. Ωστόσο, μπορεί εύκολα να επεκταθεί και για προβλήματα περισσότερων κλάσεων. Οι Μηχανές Υποστήριξης Διανυσμάτων δεν προσπαθούν να διαμορφώσουν την εσωτερική κατανομή των δεδομένων εκπαίδευσης αλλά αναζητούν το βέλτιστο χωρισμό μεταξύ των κατηγοριών.

Συγκεκριμένα ο ταξινομητής SVM προσπαθεί να βρει ένα υπερεπίπεδο απόφασης το οποίο να διαχωρίζει το σύνολο των παραδειγμάτων εκπαίδευσης με τέτοιο τρόπο ώστε τα παραδείγματα που ανήκουν στην ίδια κατηγορία να είναι στην ίδια πλευρά του υπερεπιπέδου. Μεταξύ όλων των πιθανών υπερεπιπέδων αναζητά εκείνο για το οποίο η απόσταση από το κοντινότερο παράδειγμα είναι μέγιστη, δηλαδή αναζητά το υπερεπίπεδο μεγίστου περιθωρίου (maximal margin hyperplane) [18]. Ουσιαστικά, ο ταξινομητής SVM μεγιστοποιεί την απόσταση μεταξύ των διανυσμάτων υποστήριξης (support vectors) και ενός υπερεπιπέδου απόφασης. Τα διανύσματα υποστήριξης είναι τα παραδείγματα εκπαίδευσης που βρίσκονται πιο κοντά στο υπερεπίπεδο και καθορίζουν το περιθώριό του (margin).

Στα περισσότερα προβλήματα ταξινόμησης τα παραδείγματα εκπαίδευσης δεν είναι γραμμικά διαχωρίσιμα. Σ' αυτή την περίπτωση, τα SVM απεικονίζουν το αρχικό σύνολο χαρακτηριστικών σε ένα σύνολο μεγαλύτερης διάστασης χρησιμοποιώντας μια συνάρτηση $\Phi(\mathbf{x})$. Το εσωτερικό γινόμενο στον χώρο $\Phi(\mathbf{x})$ καλείται συνάρτηση πυρήνα (kernel function). Η επιτυχία της μεθόδου εξαρτάται από τη σωστή επιλογή της συνάρτησης πυρήνα. Οι πιο γνωστοί πυρήνες είναι:

- Γραμμικός (linear) : $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \bullet \mathbf{x}_j$
- Πολυωνυμικός (polynomial) : $k(\mathbf{x}_i, \mathbf{x}_j) = (a\mathbf{x}_i \bullet \mathbf{x}_j + r)^p, a > 0$
- Ακτινικής Συνάρτησης Βάσης (radial basis function) : $k(\mathbf{x}_i, \mathbf{x}_j) = e^{-g\|\mathbf{x}_i - \mathbf{x}_j\|^2}$, όπου $g = 1/2\sigma^2$ και σ το εύρος του πυρήνα
- Πυρήνας τετραγώνου : $k(\mathbf{x}_i, \mathbf{x}_j) = \sum_m \frac{(x_{im} - x_{jm})^2}{x_{im} + x_{jm}}$

Για προβλήματα ταξινόμησης πολλών κατηγοριών υπάρχουν δύο προσεγγίσεις: η ένας εναντίον όλων προσέγγιση (one-against-all) και η ένας εναντίον ένα (one-against-one). Στην πρώτη προσέγγιση για ένα σύνολο K κατηγοριών απαιτείται η εκπαίδευση K δυαδικών SVM ταξινομητών. Κάθε SVM υπολογίζει ένα υπερεπίπεδο απόφασης το οποίο διαχωρίζει τα παραδείγματα της κατηγορίας του από τα παραδείγματα των υπόλοιπων $K-1$ κατηγοριών. Στη δεύτερη προσέγγιση για ένα σύνολο K κατηγοριών απαιτείται η εκπαίδευση $K(K-1)/2$ δυαδικών SVM. Κάθε τέτοιος ταξινομητής υπολογίζει ένα υπερεπίπεδο απόφασης το οποίο διαχωρίζει τα παραδείγματα της κατηγορίας του από τα παραδείγματα κάθε μιας από τις υπόλοιπες κατηγορίες.

Κεφάλαιο 4

Συνδυασμός Καναλιών

4.1 Εισαγωγή

Όπως θα δούμε στο κεφάλαιο 5, στο οποίο παρουσιάζεται η βάση δεδομένων που χρησιμοποιήθηκε στα πειράματα, 24 συνολικά μικρόφωνα χρησιμοποιήθηκαν για την καταγραφή των ακουστικών γεγονότων κατά την διάρκεια των συνεδριάσεων. Στο παρών κεφάλαιο περιγράφονται τα μοντέλα που δημιουργήθηκαν με το συνδυασμό των καναλιών αυτών με σκοπό την βελτίωση των αποτελεσμάτων της κατηγοριοποίησης. Αρχικά, αναλύονται τα μοντέλα που προήλθαν από το συνδυασμό των καναλιών στο επίπεδο της απόφασης. Στη συνέχεια, περιγράφονται τα μοντέλα που δημιουργήθηκαν από το συνδυασμό όλων ή ενός συγκεκριμένου αριθμού καναλιών στο επίπεδο του ηχητικού σήματος και τέλος γίνεται αναφορά στον συνδιασμό ηχητικών χαρακτηριστικών, και πιο συγκεκριμένα των MFCCs με τα χαρακτηριστικά που προέρχονται από τις καθυστερήσεις των σημάτων (TDOA features). Να τονιστεί ότι για την δημιουργία των μοντέλων που θα περιγραφούν σε αυτό το κεφάλαιο χρησιμοποιήθηκαν τα MFCC χαρακτηριστικά και ο ταξινομητής GMM, ενώ τα αντίστοιχα αποτελέσματα παρουσιάζονται στο κεφάλαιο 5.

Μέθοδος Κατηγοριοποίησης

Στο σημείο αυτό θα γίνει μια σύντομη περιγραφή του τρόπου με τον οποίο γίνεται η κατηγοριοποίηση των ακουστικών γεγονότων και η εξαγωγή των αποτελεσμάτων στα πειράματα. Όπως θα δούμε στο κεφάλαιο 5.1, στο οποίο παρουσιάζεται η βάση δεδομένων, συνολικά στα πειράματα της κατηγοριοποίησης χρησιμοποιούνται 8 καταγεγραμμένες συνεδριάσεις (S01-S08). Σέ κάθε πείραμα μια από αυτές τις συνεδρίες επιλέγεται ως σύνολο δοκιμής (test set), ενώ οι υπόλοιπες αποτελούν το σύνολο εκπαίδευσης (training set) με το οποίο

εκπαιδεύεται ο ταξινομητής. Αν π.χ η συνεδρία 1 (S01) αποτελεί το σύνολο δοκιμής, οι υπόλοιπες 7 συνεδρίες (S02-S08) αποτελούν το σύνολο εκπαίδευσης. Στη συνέχεια, το μοντέλο που κατασκευάζεται χρησιμοποιεί την προς δοκιμήν συνεδρία και προβλέπει τις κατηγορίες των ακουστικών γεγονότων που λαμβάνουν χώρα σε αυτή. Η παραπάνω διαδικασία επαναλαμβάνεται 8 φορές έτσι ώστε κάθε φορά καθεμιά από τις 8 συνεδρίες να αποτελεί το σύνολο δοκιμής. Η μέθοδος αυτή (8-cross fold validation method) βοηθάει στην εξαγωγή "δικαιότερων" αποτελεσμάτων, καθώς δεν γίνεται τυχαία η επιλογή των συνεδριών για τα σύνολα δοκιμής και εκπαίδευσης.

Αν c_i είναι ο αριθμός των σωστών προβλέψεων του μοντέλου για τα ακουστικά γεγονότα της συνεδρίας i, όπου $i = 1, \ldots, 8$ και t_i ο συνολικός αριθμός των γεγονότων της συνεδρίας i, ο αριθμός σωστών προβλέψεων C του μοντέλου σύμφωνα με την παραπάνω μέθοδο δίνεται από τον τύπο:

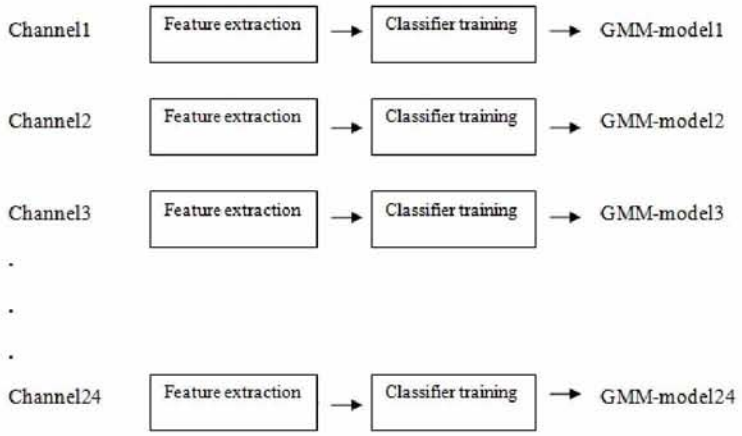
$$C = \frac{c_1 + c_2 + \ldots + c_8}{t_1 + t_2 + \ldots + t_8} \tag{4.1}$$

4.2 Συνδυασμός καναλιών στο επίπεδο της απόφασης

Στην ενότητα αυτή γίνεται λόγος για δύο μοντέλα που συνδυάζουν και τα 24 κανάλια στο επίπεδο της απόφασης, δηλαδή το Μοντέλο Μονού Καναλιού (Single Channel Model) και το Μοντέλο Πολλαπλών Καναλιών (Multi Channel Model) [8]. Το κοινό των δύο αυτών μοντέλων είναι ότι προβλέπουν τις κατηγορίες για κάθε ακουστικό γεγονός ξεχωριστά και στα 24 κανάλια και στη συνέχεια αποφασίζουν την κατηγορία που πλειοψηφεί μεταξύ των προβλέψεων.

4.2.1 Μοντέλο Μονού Καναλιού

Στο Μοντέλο Μονού Καναλιού η όλη διαδικασία της κατηγοριοποίησης πραγματοποιείται σε καθένα από τα 24 κανάλια ξεχωριστά όπως ακριβώς θα πραγματοποιούνταν σε ένα μόνο κανάλι. Έτσι, για κάθε συνεδρία που αποτελεί το σύνολο δοκιμής (test set), χρησιμοποιούνται οι υπόλοιπες για την εξαγωγή των ηχητικών χαρακτηριστικών, την εκπαίδευση ενός ταξινομητή και τη δημιουργία ενός Μοντέλου Γκαουσσιανών Κατανομών (GMM). Η παραπάνω διαδικασία πραγματοποιείται και στα 24 κανάλια με αποτέλεσμα για κάθε μια συνεδρία να δημιουργούνται 24 μοντέλα GMM. Κάθε τέτοιο σύνολο από 24 μοντέλα κατηγοριοποίησης αντιστοιχεί σε μια συνεδρία, η οποία αποτελεί



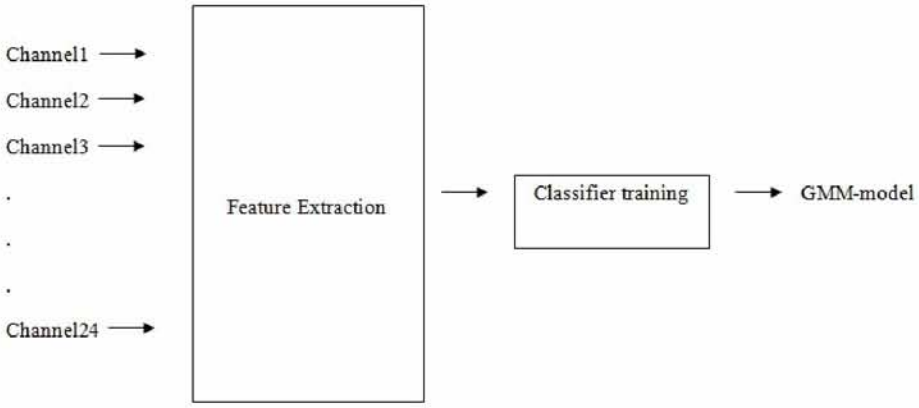
Σχήμα 4.1: Μοντέλο Απλού Καναλιού. Δημιουργία των μοντέλων GMM στο πλαίσιο μιας συνεδρίας

το σύνολο δοκιμής και κατά συνέπεια είναι η μόνη από τις 8 συνεδρίες που απουσιάζει από το σύνολο εκπαίδευσης. Η διαδικασία δημιουργίας των μοντέλων GMM για μια συνεδρία i, όπου $i = 1, \ldots, 8$ φαίνεται στο σχήμα 4.1.

Μετά την δημιουργία των μοντέλων ακολουθεί η φάση της ταξινόμησης. Για κάθε συνεδρία εξάγονται τα χαρακτηριστικά των γεγονότων του συνόλου δοκιμής σε καθένα από τα 24 κανάλια και χρησιμοποιείται ο ταξινομητής GMM που αντιστοιχεί στο κάθε κανάλι για την πρόβλεψη της κλάσης του κάθε ακουστικού γεγονότος. Κάθε GMM ταξινομητής, δηλαδή, λειτουργεί ανεξάρτητα από τους άλλους και δίνει τη δική του "γνώμη" για το άγνωστο ακουστικό γεγονός. Έτσι για κάθε γεγονός υπάρχουν 24 προβλέψεις, μία για κάθε κανάλι, και από αυτές επιλέγεται ως τελική πρόβλεψη του μοντέλου αυτή που πλειοψηφεί (majority voting).

4.2.2 Μοντέλο Πολλαπλών Καναλιών

Το Μοντέλο Πολλαπλών Καναλιών παρουσιάζει αρκετές ομοιότητες με το Μοντέλο Μονού Καναλιού, έχοντας ωστόσο μια σημαντική διαφορά στον τρόπο δημιουργίας κάθε GMM μοντέλου. Για κάθε μια συνεδρία που περιλαμβάνεται στο σύνολο εκπαίδευσης εξάγονται τα ιδιαίτερα χαρακτηριστικά των γεγονότων της σε καθε ένα από τα 24 κανάλια. Τα χαρακτηριστικά αυτά του κάθε καναλιού δεν χρησιμοποιούνται για την δημιουργία ενός ξεχωριστού GMM μοντέλου, όπως στο Μοντέλο Μονού Καναλιού, αλλά συγκεντρώνονται μαζί με



Σχήμα 4.2: Μοντέλο Πολλαπλών Καναλιών. Δημιουργία του μοντέλου GMM στο πλαίσιο μιας συνεδρίας

τα χαρακτηριστικά και των υπολοίπων καναλιών σε ένα μεγάλο πίνακα και χρησιμοποιούνται όλα μαζί για την εκπαίδευση ενός μόνο ταξινομητή GMM. Ουσιαστικά, για καθε συνεδρία που αποτελεί το σύνολο δοκιμής δημιουργείται ένα μοντέλο GMM, το οποίο έχει εκπαιδευτεί με τα ακουστικά χαρακτηριστικά των γεγονότων των υπόλοιπων 7 συνεδριών και στα 24 κανάλια. Η διαδικασία αυτή φαίνεται στο σχήμα 4.2.

Αν X_s είναι ο πίνακας χαρακτηριστικών των γεγονότων μιας συνεδρίας στο κανάλι s, τότε ο πίνακας χαρακτηριστικών Z που χρησιμοποιείται για την εκπαίδευση του GMM ταξινομητή δημιουργείται από την συνένωση των πινάκων X_s καθενός από τα 24 κανάλια:

$$Z = X_1 \cup X_2 \cup X_3 \cup \ldots \cup X_{24} \tag{4.2}$$

Κατά την φάση της ταξινόμησης, το κοινό GMM μοντέλο που αντιστοιχεί στην προς δοκιμήν συνεδρία (test session) χρησιμοποιείται για την κατηγοριοποίηση των άγνωστων ακουστικών γεγονότων σε καθένα από τα 24 κανάλια. Στη συνέχεια, και εφόσον υπάρχουν 24 προβλέψεις, μια για κάθε ένα από τα 24 κανάλια, επιλέγεται ως κλάση του άγνωστου ακουστικού γεγονότος αυτή που πλειοψηφεί.

4.3 Συνδυασμός καναλιών στο επίπεδο του σήματος

Στο παρών κεφάλαιο γίνεται αναφορά στα μοντέλα που δημιουργήθηκαν με συνδυασμό όλων ή ενός συγκεκριμένου αριθμού καναλιών στο επίπεδο του ηχητικού σήματος. Με τη χρήση του κατάλληλου λογισμικού γίνεται ο συνδυασμός των σημάτων των επιλεγμένων καναλιών με αποτέλεσμα τη δημιουργία ενός μόνο βελτιωμένου (enhanced) ηχητικού σήματος. Στη συνέχεια, ακολουθεί η διαδικασία της κατηγοριοποίησης ακριβώς όπως θα πραγματοποιούνταν εαν επιλέγαμε ένα μόνο από τα 24 κανάλια.

Πιο συγκεκριμένα τα πειράματα που αντιστοιχούν στην ενότητα αυτή, και τα αποτελέσματά των οποίων θα περιγραφούν στο κεφάλαιο 5, χωρίζονται σε 2 κύριες κατηγορίες. Στην πρώτη χρησιμοποιούμε το λογισμικό BeamformIt, το οποίο περιγράφεται στην υποενότητα 4.3.1, στο οποίο δίνουμε ως είσοδο τα σήματα και των 24 καναλιών μιας συνεδρίασης, και μετά την επεξεργασία και τον συνδιασμό τους, λαμβάνουμε στην έξοδο ένα ενισχυμένο/βελτιωμένο σήμα. Με τον τρόπο αυτόν, σε κάθε συνεδρίαση αντιστοιχεί ένα μοναδικό σήμα. Ακολουθούν οι φάσεις της εκπαίδευσης ενός ταξινομητή και της κατηγοριοποίησης των ακουστικών γεγονότων με την μέθοδο 8-cross fold validation, όπως ακριβώς θα συνέβαινε αν εφαρμόζαμε την παραπάνω διαδικασία στα σήματα ενός τυχαίου καναλιού από τα 24 που είναι διαθέσιμα.

Τα πειράματα της δεύτερης κατηγορίας είναι παρόμοια με αυτά της πρώτης με μια σημαντική διαφοροποίηση. Αντί να εισάγουμε τα σήματα και των 24 καναλιών στο BeamformIt, επιλέγουμε τα σήματα των N "καλύτερων" καναλιών. Η επιλογή αυτή βασίζεται στις μετρικές εκτίμησης της ποιότητας των καναλιών, οι οποίες παρουσιάζονται αναλυτικά στην υποενότητα 4.3.2. Με τον τρόπο αυτό συνδυάζουμε μόνο "ποιοτικά" κανάλια και αποφεύγουμε τη χρήση καναλιών χαμηλής ποιότητας τα οποία μπορεί να επηρεάσουν αρνητικά στην δημιουργία του εξαγόμενου σήματος. Παράδειγμα τέτοιων καναλιών αποτελούν αυτά που περιέχουν υψηλά επίπεδα περιβαλλοντικού θορύβου. Στα πειράματα που πραγματοποιήθηκαν δώσαμε στο N διάφορες τιμές. Ακόμη, στα πειράματα και των δύο αυτών κατηγοριών αλλάξαμε και δώσαμε διάφορες τιμές στις παραμέτρους του BeamformIt καταγράφοντας τις αλλαγές στα αποτελέσματα της κατηγοριοποίησης.

4.3.1 Λογισμικό BeamformIt

Στο σημείο αυτό θα γίνει η περιγραφή του λογισμικού που χρησιμοποιήθηκε για το συνδυασμό των καναλιών και την δημιουργία ενός μό-

νο σήματος με χρήση κάποιας τεχνικής διαμόρφωσης ακτίνας/δέσμης (beamforming technique). Χρησιμοποιήθηκε το BeamformIt [19], [20] το οποίο αποτελεί ένα αποτελεσματικό εργαλείο που δέχεται ως είσοδο εναν αριθμό καναλιών και υπολογίζει ως έξοδο ένα μοναδικό ακουστικό σήμα μέσω μιας τεχνικής φιλτραρίσματος και αθροίσματος (filter&sum beamforming technique).

Το BeamformIt αρχικά δημιουργήθηκε από τον Xavier Anguera για την συμμετοχή στην NIST RT05s αξιολόγηση με σκοπό να ασχοληθεί με τα διαφορετικά μικρόφωνα που ήταν διαθέσιμα σε μια αίθουσα συνεδριάσεων. Στη συνέχεια το λογισμικό βελτιώθηκε για την RT06s αξιολόγηση και τελικά κυκλοφόρησε για δημόσια χρήση. Το σήμα εξόδου που παράγεται από την επεξεργασία του BeamformIt αποθηκεύεται σε ένα *.SPH αρχείο με ρυθμό δειγματοληψίας (sample rate) 16 kHz. Για την βέλτιστη απόδοση του BeamformIt και τα σήματα εισόδου πρέπει να έχουν ρυθμό δειγματοληψίας 16 kHz. Αν έχουν μεγαλύτερο ρυθμό δειγματοληψίας, όπως στην δική μας περίπτωση, αφαρμόζεται η τεχνική της υποδειγματοληψίας (downsampling).

Τεχνικές Beamforming

Οι τεχνικές beamforming εκμεταλεύονται το γεγονός ότι το ίδιο ακουστικό σήμα καταφθάνει σε καθένα από τα μικρόφωνα σε μια ελάχιστα διαφορετική χρονική στιγμή εξαιτίας της καθυστέρησης στη διάδοση τους σήματος μέσω του αέρα. Συνδυάζοντας τα σήματα όλων των μικροφώνων με διάφορους τρόπους γίνεται δυνατή η προσομοίωση ενός μικροφώνου του οποίου η ακουστική δέσμη επικεντρώνεται στον ομιλητή ή στο ακουστικό γεγονός που λαμβάνει χώρα κάθε στιγμή στο πλαίσιο μιας συνεδρίασης. Υπάρχουν πολλές ακουστικές τεχνικές beamforming οι οποίες απαιτούν διαφορετικό επίπεδο γνώσης των χαρακτηριστικών των μικροφώνων και της τοποθεσίας των ομιλητών.

Υπάρχουν δύο κύριες κατηγορίες τεχνικών beamforming. Η ανεξάρτητη δεδομένων (data-independent) και η εξαρτημένη δεδομένων (data-dependent). Οι τεχνικές οι οποίες είναι ανεξάρτητες των δεδομένων ορίζουν τις παραμέτρους τους και τις διατηρούν καθόλη την διάρκεια επεξεργασίας των σημάτων εισόδου. Από την άλλη μεριά οι τεχνικές που εξαρτώνται από τα δεδομένα ενημερώνουν συνεχώς τις παραμέτρους τους ώστε να ταιριάζουν καλύτερα στα σήματα εισόδου. Ακόμη, υπάρχουν πολλές τεχνικές επεξεργασίας οι οποίες εφαρμόζονται μετά το beamforming.

Η απλούστερη τεχνική beamforming ειναι αυτή της καθυστέρησης και αθροίσματος (Delay&Sum technique). Το εξαγώμενο σήμα $y(n)$ δίνεται από τον τύπο:

$$y[n] = \frac{1}{N} \sum_{m=1}^{N} x_m[n - \tau_m] \tag{4.3}$$

δεδομένου ενός συνόλου N μικροφώνων, όπου κάθε μικρόφωνο έχει μια σχετική με τα άλλα μικρόφωνα καθυστέρηση τ_m. Σε αυτή την τεχνική όλα τα κανάλια είναι ισοσταθμισμένα στην έξοδο. Η τεχνική καθυστέρησης και αθροίσματος αποτελεί μια υποπερίπτωση μιας πιο γενικής τεχνικής beamforming, της τεχνικής φιλτραρίσματος και αθροίσματος (Filter&Sum technique). Το εξαγώμενο σήμα $y(n)$ της τεχνικής αυτής, η οποία χρησιμοποιείται στο BeamformIt, δίνεται από τον τύπο:

$$y[n] = \sum_{m=1}^{N} w_m(n) x_m[n - \tau_m] \tag{4.4}$$

Σε αυτή την τεχνική τα σήματα από τα διαφορετικά μικρόφωνα περνούν μέσω ενός ανεξάρτητου για κάθε μικρόφωνο φίλτρου w_m. Το εξαγώμενο σήμα είναι το άθροισμα όλων των φιλτραρισμένων σημάτων.

Παράμετροι του BeamformIt

Όπως έχουμε αναφέρει το BeamformIt αποτελεί ένα λογισμικό που δέχεται ως είσοδο έναν αριθμό σημάτων καταγεγραμμένα από διαφορετικά μικρόφωνα σε μια αίθουσα συνεδριάσεων, στη συνέχεια ακολουθεί μια επεξεργασία των σημάτων αυτών και μέσω κάποιας τεχνικής beamforming δημιουργεί ένα ενισχυμένο σήμα με καλύτερη ποιότητα από κάθε ένα από τα επιμέρους αρχικά σήματα. Το νέο σήμα είναι και το πιο "καθαρό" από άποψης θορύβου και αυτό με το μεγαλύτερο Λόγο Σήματος προς Θόρυβο (Signal-to-Noise Ratio - SNR).

Η διαδικασία της επεξεργασίας των σημάτων που ακολουθείται κάθε φορά διαφέρει ανάλογα με τις παραμέτρους που εμείς επιλέγουμε. Παρακάτω περιγράφονται οι σημαντικότερες παράμετροι του λογισμικού με τις οποίες πειραματιστήκαμε και καταγράψαμε τις αλλαγές που επιφέρουν στα αποτελέσματα του προβλήματος της κατηγοριοποίησης, τα οποία παρουσιάζονται στο κεφάλαιο 5.

Υπολογισμός του καναλιού αναφοράς

Σε ένα τυπικό συστημα beamforming βασισμένο στις χρονικές καθυστερήσεις είναι απαραίτητη η επιλογή ενός καναλιού ως κανάλι αναφοράς (reference channel). Το κανάλι αυτό συγκρίνεται με όλα τα

υπόλοιπα και οι χρονικές καθυστερήσεις άφιξης των σημάτων (time delay of arrival - TDOA) υπολογίζονται για κάθε ζευγάρι. Είναι πολύ σημαντικό το κανάλι αναφοράς να είναι ο καλύτερος εκπρόσωπος των καναλιών μιας συνεδρίασης καθώς ο σωστός υπολογισμός των καθυστερήσεων εξαρτάται από την επιλογή του καναλιού αυτού.

Το BeamformIt δίνει την δυνατότητα στον χρήστη είτε να επιλέξει το κανάλι αναφοράς είτε αυτό να επιλεγεί μέσω ενός αλγορίθμου βασισμένου στην ετεροσυσχέτιση (cross-correlation algorithm). Στα πειράματα που πραγματοποιήθηκαν επιλέγουμε εμείς ως κανάλι αναφοράς το καλύτερο από τα 24 κανάλια, βάσει κάποιων μετρικών εκτίμησης της ποιότητας της καναλικής πληροφορίας, οι οποίες θα περιγραφούν αναλυτικά στην υποενότητα 4.3.2. Ένα τέτοιο αξιόπιστο μέτρο σύγκρισης μεταξύ των καναλιών ειναι το SNR.

Ωστόσο, σε κάποια πειράματα το κανάλι αναφοράς υπολογίζεται από το ίδιο το λογισμικό μέσω του αλγορίθμου ετεροσυσχέτισης. Ο αλγόριθμος αυτός υπολογίζει τη ετεροσυσχέτιση (GCC-PHAT) για όλους τους πιθανούς συνδυασμούς καναλιών για ένα μπλοκ διάρκειας $1s$. Αυτό επαναλαμβάνεται για M μπλοκς. Για κάθε κανάλι i η μέση ετεροσυσχέτιση υπολογίζεται από τον τύπο:

$$cross\_correlation_i = \frac{1}{MN} \sum_{m=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{N} xcorr(i,j) \tag{4.5}$$

όπου N είναι ο αριθμός των καναλιών και το M υποδεικνύει τον αριθμό των μπλοκς που λαμβάνονται υπόψη στην εύρεση του μέσου. Το κανάλι με την μεγαλύτερη μέση ετεροσυσχέτιση επιλέγεται ως το κανάλι αναφοράς.

Κατώφλι θορύβου

Άλλη μια σημαντική παράμετρος με την οποία πειραματιστήκαμε είναι αυτή του κατωφλίου θορύβου (noise threshold). Το BeamformIt δίνει την δυνατότητα στον χρήστη να επιλέξει την εφαρμογή του αλγόριθμου του κατωφλίου θορύβου (noise thresholding algorithm) μετά τον υπολογισμό των TDOA τιμών. Ο αλγόριθμος αυτός ανιχνεύει τις TDOA τιμές που πιθανώς προέρχονται από κάποιο κομμάτι του σήματος κατά το οποίο δεν λαμβάνει χώρα κάποιο ακουστικό γεγονός και τις αντικαθιστά με προηγούμενες πιο σταθερές τιμές καθυστέρησης. Ουσιαστικά, δηλαδή, ο αλγόριθμος ανιχνεύει και αφαιρεί τις TDOA τιμές που δεν μπορούν να θεωρηθούν αξιόπιστες. Κάθε φορά εφαρμόζεται ένα απλό φίλτρο στις TDOA τιμές βασισμένο στις τιμές ετεροσυσχέτισής τους (GCC-PHAT values) χρησιμοποιώντας ένα κατώφλι θορύβου Thr_{noise}.

Υπόλοιπες παράμετροι

Εκτός από τις παραπάνω παραμέτρους υπάρχουν και αυτές που παρέμειναν σταθερές κατά τη διάρκεια των πειραμάτων, καθώς διατηρήθηκαν οι προεπιλεγμένες τιμές τους. Αυτές είναι η χρησιμοποίηση ενός αλγορίθμου για τον υπολογισμό του σχετικού βάρους του κάθε καναλιού καθώς και η εξάλειψη των πλαισίων (frames) από κάποια χαμηλής ποιότητας κανάλια. Ο υπολογισμός του σχετικού βάρους του κάθε καναλιού είναι πολύ σημαντικός καθώς με αυτόν τον τρόπο μειώνεται η επίδραση των καναλιών χαμηλής ποιότητας και αυξάνεται η επίδραση των "ποιοτικών" καναλιών. Όσον αφορά την εξάλειψη πλαισίων, χρησιμοποιείται η πληροφορία διασυσχέτισης για να αποφασιστεί αν κάποιο πλαίσιο ενός καναλιού είναι χαμηλής ποιότητας και είναι καλύτερο να μην λάβει μέρος στον υπολογισμό του σήματος εξόδου.

4.3.2 Μετρικές εκτίμησης της ποιότητας των καναλιών

Σε αυτή την υποενότητα θα προσπαθήσουμε να εξηγήσουμε τι σημαίνει όταν λέμε ότι ένα κανάλι είναι "ποιοτικό" ή ότι ένα κανάλι είναι χαμηλής ποιότητας. Για τον σκοπό αυτό θα γίνει παρουσίαση και ανάλυση κάποιων μετρικών εκτίμησης της ποιότητας της καναλικής πληροφορίας, και πιο συγκεκριμένα του Λόγου Σήματος προς Θόρυβο (Signal-to-Noise Ratio - SNR), της μέσης διαφοράς της λογαριθμικής πιθανοφάνειας (average log-likelihood difference) και της διασποράς της λογαριθμικής πιθανοφάνειας (log-likelihood dispersion). Στα πειράματα που πραγματοποιήθηκαν επιλέξαμε αρκετές φορές έναν αριθμό καλύτερων καναλιών ή το καλύτερο από τα 24 κανάλια σύμφωνα με αυτές τις μετρικές για την δημιουργία των μοντέλων και την κατηγοριοποίηση των ακουστικών γεγονότων.

Σηματοθορυβικός Λόγος (SNR)

Κάθε ηχητικό σήμα περιέχει και κάποιο ποσοστό περιβαλλοντικού θορύβου, δηλαδή ήχων του περιβάλλοντος οι οποίοι παράγονται από ανεπιθύμητες δραστηριότητες. Αυτό συμβαίνει και κατά την καταγραφή των ακουστικών γεγονότων των συνεδριάσεων από τα μικρόφωνα του δωματίου, με το επίπεδο του περιβαλλοντικού θορύβου να διαφέρει σε καθένα από αυτά. Το SNR [21] είναι μια μετρική, η οποία χρησιμοποιείται κυρίως σε εφαρμογές επεξεργασίας σημάτων, και η οποία συγκρίνει το επίπεδο του επιθυμητού "χρήσιμου" σήματος με το επίπεδο του θορύβου του περιβάλλοντος. Ορίζεται ως ο λόγος της

μέσης ισχύος της χρήσιμης πληροφορίας του σήματος ως προς την μέση ισχύ του θορύβου, όπως φαίνεται στην (4.6). Όσο μεγαλύτερος είναι ο λόγος αυτός τόσο περισσότερο το "χρήσιμο" σήμα υπερισχύει του θορύβου. Μονάδα μέτρησης του SNR είναι τα decibels (dB).

$$SNR_{dB} = 10log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \tag{4.6}$$

Στα πειράματα υπολογίσαμε το SNR για κάθε μια από τις 8 συνεδρίες και στα 24 κανάλια καταγραφής των ηχητικών σημάτων. Το SNR παρουσιάζει σημαντικές διαφοροποιήσεις μεταξύ των καναλιών κάτι που δηλώνει πως υπάρχουν κανάλια που το ποσοστό της χρήσιμης πληροφορίας που περιέχουν είναι μεγαλύτερο από το ποσοστό κάποιων άλλων καναλιών. Με βάση αυτή την μετρική τα κανάλια μπορούν να διαχωριστούν ως "ποιοτικά" ή ως λιγότερο "ποιοτικά". Για τον υπολογισμό του SNR στα αρχεία ήχου των συνεδριάσεων της βάσης δεδομένων χρησιμοποιήσαμε τον τύπο :

$$SNR_{dB} = 10log_{10}\left(\frac{\frac{1}{M}\sum_{i=0}^{M-1} s^2[i]}{\frac{1}{K}\sum_{i=0}^{K-1} n^2[i]}\right) \tag{4.7}$$

όπου $s[i]$ είναι τα δείματα (samples) που αντιστοιχούν στο χρήσιμο σήμα και $n[i]$ τα δείγματα που αντιστοιχούν στο θόρυβο. Ακόμη M και K είναι ο συνολικός αριθμός των δειγμάτων που περιέχουν χρήσιμο σήμα και θόρυβο αντίστοιχα. Τα δείγματα που περιέχουν θόρυβο ουσιαστικά αντιστοιχούν στα τμήματα των αρχείων ήχου κατά τα οποία δεν λαμβάνει χώρα κάποιο ακουστικό γεγονός και χαρακτηρίζονται με το σύμβολο "si" στα *.csv αρχεία της βάσης δεδομένων.

Average N-best log-likelihood difference

Το συγκεκριμένο μέτρο, καθώς και το επόμενο που θα αναλυθεί στην επόμενη παράγραφο, προσπαθούν να εκτιμήσουν την ικανότητα διαχωρισμού (discriminating power) των ταξινομητών σε κάθε κανάλι. Με τον όρο ικανότητα διαχωρισμού εννοούμε κατά πόσο ένας ταξινομητής μπορεί να διαχωρίσει τις κλάσεις του προβλήματος της κατηγοριοποίησης. Στην δική μας περίπτωση, το μέτρο της μέσης διαφοράς της λογαριθμικής πιθανοφάνειας [22] δείχνει την ικανότητα του GMM ταξινομητή να διαχωρίσει τις κλάσεις των ακουστικών γεγονότων που συμβαίνουν κατά την διάρκεια μιας συνεδρίασης σε καθένα από τα 24 κανάλια. Όσο μεγαλύτερη είναι η τιμή του σε ένα κανάλι, τόσο μεγαλύτερη είναι και η πιθανότητα το GMM μοντέλο του καναλιού αυτού να επιτύχει στις προβλέψεις των κλάσεων των άγνωστων γεγονότων.

Έστω ότι $c_{s,n}^{(t)}$, $n = 1, \ldots, N$ είναι οι N κλάσεις με την μεγαλύτερη πιθανότητα να αποτελούν την κλάση μιας δεδομένης παρατήρησης $O_s^{(t)}$ στο κανάλι s, όπου $s = 1, \ldots, 24$. Γίνεται κατανοητό πως η κλάση με την μεγαλύτερη πιθανότητα, δηλαδή η $c_{s,1}^{(t)}$ είναι αυτή που τελικά προβλέπεται από τον ταξινομητή. Ακόμη, έστω ότι $R_{s,n}^{(t)}$ είναι η λογαριθμική πιθανότητα της n-στής πιθανότερης κλάσης στο κανάλι s, δηλαδή $R_{s,n}^{(t)} = logPr(O_s^{(t)}|c_{s,n}^{(t)})$. Ο τύπος με τον οποίο υπολογίζουμε το μέτρο της μέσης διαφοράς της λογαριθμικής πιθανοφάνειας για ένα συγκεκριμένο ακουστικό γεγονός στο κανάλι s είναι ο εξής :

$$I_{s,L}^{(t)} = \frac{1}{N-1} \sum_{n=2}^{N} (R_{s,1}^{(t)} - R_{s,n}^{(t)}) \tag{4.8}$$

Μετά τον υπολογισμό του $I_{s,L}^{(t)}$ για κάθε παρατήρηση, προσθέτουμε τις τιμές αυτές που αφορούν ακουστικά γεγονότα του ίδιου καναλιού ώστε να έχουμε μια γενική τιμή για το κάθε κανάλι. Αν I_s είναι η τιμή του μέτρου σε ένα κανάλι s τότε :

$$I_s = I_{s,1} + I_{s,2} + \ldots + I_{s,T} \tag{4.9}$$

όπου T είναι ο συνολικός αριθμός των ακουστικών γεγονότων μιας συνεδρίασης στο κανάλι s. Όσο μεγαλύτερη είναι η τιμή του I_s τόσο "καλύτερο" θεωρείται το κανάλι s καθώς ο ταξινομητής GMM σε αυτό έχει την ικανότητα να διαχωρίσει ευκολότερα τις κλάσεις των ακουστικών γεγονότων.

N-best log-likelihood dispersion

Το μέτρο της διασποράς της λογαριθμικής πιθανοφάνειας [22] είναι παρόμοιο με το μέτρο που μόλις περιγράψαμε. Υπολογίζει και αυτό την ικανότητα ενός ταξινομητή στο διαχωρισμό των κλάσεων των ακουστικών γεγονότων μιας συνεδρίασης σε ένα συγκεκριμένο κανάλι. Υπολογίζεται από τον τύπο :

$$I_{s,D}^{(t)} = \frac{2}{N(N-1)} \sum_{n=1}^{N} \sum_{n'=n+1}^{N} (R_{s,n}^{(t)} - R_{s,n'}^{(t)}) \tag{4.10}$$

Στη συνέχεια ακολουθείται η διαδικασία πρόσθεσης των επιμέρους τιμών των γεγονότων ενός συγκεκριμένου καναλιού s, όπως ακριβώς στην (4.9), ώστε να έχουμε μια γενική εικόνα της αξίας του κάθε καναλιού.

4.4 Συνδυασμός MFCC και TDOA χαρακτηριστικών

Στο παρών κεφάλαιο γίνεται περιγραφή των χαρακτηριστικών TDOA (time delay of arrival), τα οποία προέρχονται από τον υπολογισμό των καθυστερήσεων της άφιξης των σημάτων στα μικρόφωνα που καταγράφουν τις συνεδριάσεις. Στα πειράματα που πραγματοποιήσαμε έγινε προσπάθεια για το συνδυασμό των χαρακτηριστικών αυτών με τα κλασικά ακουστικά χαρακτηριστικά MFCC σε ήδη υπάρχοντα μοντέλα που έχουν περιγραφεί σε προηγούμενες ενότητες του κεφαλαίου αυτού. Τα αποτελέσματα των πειραμάτων αυτών, τα οποία παρουσιάζονται στο κεφάλαιο 5, δείχνουν κατά πόσο ο συνδυασμός αυτών των διαφορετικών χαρακτηριστικών μπορεί να επιφέρει σημαντική βελτίωση στα αποτελέσματα του προβλήματος της κατηγοριοποίησης των ακουστικών γεγονότων μιας συνεδρίασης.

4.4.1 Χαρακτηριστικά TDOA

Κατά τη διάρκεια των συνεδριάσεων της βάσης δεδομένων που χρησιμοποιούμε, τα ακουστικά γεγονότα που παράγονται καταγράφονται από 24 μικρόφωνα που έχουν διαφορετική θέση μέσα στο δωμάτιο (Multiple Distant Microphones - MDM). Σε τέτοιου είδους πολυκαναλικά περιβάλλοντα το ίδιο ακουστικό σήμα καταφθάνει σε καθένα από τα μικρόφωνα σε μια ελάχιστα διαφορετική χρονική στιγμή εξαιτίας της καθυστέρησης της διάδοσης του σήματος μέσω του αέρα. Οι χρονικές αυτές καθυστερήσεις της άφιξης των σημάτων αποτελούν ένα νέο είδος χαρακτηριστικών, τα TDOA χαρακτηριστικά [23].

Τα TDOA χαρακτηριστικά υπολογίζονται με τη χρήση ενός γενικευμένου μετασχηματισμού ετεροσυσχέτισης φάσεως (generalized cross correlation phase transform - GCC-PHAT). Όλες οι χρονικές καθυστερήσεις υπολογίζονται βάσει ενός καναλιού αναφοράς. Το κανάλι αναφοράς επιλέγεται βάσει του SNR ή βάσει της τιμής της μέσης ετεροσυσχέτισης του καναλιού με τα υπόλοιπα κανάλια. Οι χρονικές καθυστερήσεις της άφιξης των σημάτων υπολογίζονται κάθε 250 ms (segment size) ενώ γίνεται χρήση και ενός παραθύρου ανάλυσης (analysis window) 500 ms. Η TDOA τιμή μεταξύ του καναλιού αναφοράς και οποιουδήποτε άλλου από τα υπόλοιπα κανάλια ορίζεται ως η καθυστέρηση που αντιστοιχεί στη μέγιστη ετεροσυσχέτιση μεταξύ των δύο καναλιών.

Δεδομένου δύο σημάτων $s_i(n)$ και $s_j(n)$ η GCC-PHAT τιμή ορίζεται ως :

$$G_{PHAT}(f) = \frac{X_i(f)X_j^*(f)}{|X_i(f)||X_j(f)|} \tag{4.11}$$

όπου $X_i(f)$ και $X_j(f)$ είναι οι μετασχηματισμοί Fourier των δύο σημάτων. Η τιμή TDOA για τα κανάλια s_i και s_j υπολογίζεται ως εξής:

$$d_{PHAT}(i,j) = \arg\max_d R_{PHAT}(d) \tag{4.12}$$

όπου $R_{PHAT}(d)$ είναι ο αντίστροφος μετασχηματισμός Fourier του $G_{PHAT}(f)$.

4.4.2 Διαδικασία συνδυασμού MFCC και TDOA χαρακτηριστικών

Σε αυτή την υποενότητα θα περιγραφεί η διαδικασία συνδυασμού των χαρακτηριστικών MFCC και TDOA [23], [24]. Αρχικά, γίνεται ο υπολογισμός των χρονικών καθυστερήσεων στην άφιξη των σημάτων στα μικρόφωνα της αίθουσας συνεδριάσεων. Ο υπολογισμός αυτός γίνεται μέσω του BeamformIt, το οποίο παράγει στην έξοδό του ένα αρχείο που περιέχει τις τιμές αυτές υπολογισμένες κάθε 250 ms. Οι τιμές του καναλιού αναφοράς παραλείπονται καθώς έχουν την τιμή 0, ενώ οι υπόλοιπες $N-1$ τιμές αποτελούν τα TDOA χαρακτηριστικά, όπου N είναι ο αριθμός των καναλιών που δέχεται ως είσοδο το BeamformIt. Στην περίπτωση των πειραμάτων μας, όταν εισάγουμε και τα 24 κανάλια στο BeamformIt, οι 23 τιμές καθυστέρησης των υπόλοιπων καναλιών χρησιμοποιούνται ως TDOA χαρακτηριστικά.

Στη συνέχεια, κατά την φάση της εκπαίδευσης, εκπαιδεύουμε δύο ταξινομητές GMM με αποτέλεσμα την δημιουργία 2 μοντέλων, ανεξάρτητων μεταξύ τους. Το ένα είναι το μοντέλο M_{MFCC} το οποίο παράγεται με τη χρήση των MFCC χαρακτηριστικών κατά την φάση της εκπαίδευσης του ταξινομητή, όπως ακριβώς γίνεται και σε όλα τα μοντέλα που έχουν ήδη περιγραφεί, και το άλλο είναι το μοντέλο M_{TDOA} το οποίο παράγεται με τη χρήση των TDOA χαρακτηριστικών κατά την φάση της εκπαίδευσης. Τα δύο αυτά μοντέλα είναι τελείως διαφορετικά μεταξύ τους, δεν αλληλεπιδρούν και λειτουργούν ανεξάρτητα το ένα με το άλλο.

Τέλος, κατά την φάση της κατηγοριοποίησης και αφού εξάγουμε τα MFCC και τα TDOA χαρακτηριστικά ενός άγνωστου ακουστικού γεγονότος O, υπολογίζουμε τις PDF τιμές $p(O_i|c_k)$ ξεχωριστά για τα δύο μοντέλα M_{MFCC} και M_{TDOA}, όπου O_i είναι το πλαίσιο (frame) i

του γεγονότος O και c_k είναι μια συγκεκριμένη κλάση από τις 12 συνολικά που διαθέτει η βάση δεδομένων, δηλαδή $k = 1, \ldots, 12$. Μετά την διαδικασία αυτή, υπολογίζουμε το άθροισμα των λογαρίθμων των PDF τιμών για κάθε μία κλάση c_k ως εξής :

$$\mathcal{L}(O|c_k) = \sum_{i=1}^{N} \log[p(O_i|c_k)] \tag{4.13}$$

όπου N είναι ο αριθμός των πλαισίων του άγνωστου ακουστικού γεγονότος O. Η συγκεκριμένη διαδικασία πραγματοποιείται ξεχωριστά για τα δύο μοντέλα. Αφού γίνει και αυτό το βήμα, ακολουθεί ο συνδυασμός των πιθανοτήτων των δύο μοντέλων με τον τύπο :

$$\mathcal{L}(O|c_k) = \mathcal{L}(O|c_k, M_{MFCC})W_1 + \mathcal{L}(O|c_k, M_{TDOA})W_2 \tag{4.14}$$

όπου W_1 και W_2 είναι τα βάρη με τα οποία συμμετέχουν στο συνδυασμό τα μοντέλα M_{MFCC} και M_{TDOA} αντίστοιχα, και για τα οποία ισχύει $W_1 + W_2 = 1$. Τα βάρη αυτά [25] παίζουν σημαντικό ρόλο στην βελτίωση των αποτελεσμάτων της κατηγοριοποίησης και οι τιμές τους εξαρτώνται από την βάση δεδομένων που χρησιμοποιείται. Με τη χρήση μόνο των TDOA χαρακτηριστικών λαμβάνουμε πολύ χαμηλά ποσοστά σωστών προβλέψεων, γι' αυτό και τις περισσότερες φορές το W_2 έχει μικρότερη τιμή από το W_1, δηλαδή το βάρος συμμετοχής των MFCC χαρακτηριστικών. Στα πειράματα δώσαμε διάφορες τιμές στα βάρη αυτά, λαμβάνοντας τα καλύτερα αποτελέσματα για τα ζεύγη $W_1 = 0.7, W_2 = 0.3$ και $W_1 = 0.8, W_2 = 0.2$.

Τελικά, μετά και τον υπολογισμό των πιθανοτήτων της (4.14), με τους οποίους επιτυγχάνεται ο συνδυασμός των MFCC και TDOA χαρακτηριστικών, επιλέγουμε ως κλάση του άγνωστου ακουστικού γεγονότος O την κλάση c_k για την οποία ισχύει :

$$c_k = \arg\max_k \mathcal{L}(O|c_k) \tag{4.15}$$

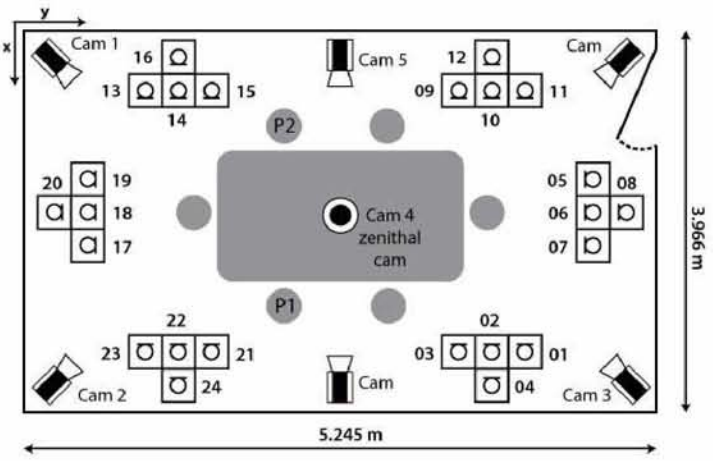
Κεφάλαιο 5

Πειράματα και Αποτελέσματα

5.1 Βάση δεδομένων

Για την πραγματοποίηση των πειραμάτων σχετικά με το πρόβλημα της κατηγοριοποίησης απαραίτητη είναι η χρήση κάποιας βάσης δεδομένων. Για τα πειράματα που πραγματοποιήθηκαν στο πλαίσιο της παρούσας εργασίας χρησιμοποιήθηκε η UPC-TALP Multimodal Database [26], η οποία περιέχει καταγεγραμμένα ακουστικά γεγονότα που μπορούν να συμβούν κατά τη διάρκεια μιας συνεδρίασης. Η παραγωγή της πραγματοποιήθηκε στο Τμήμα Θεωρίας Σημάτων και Τηλεπικοινωνιών και στο Ερευνητικό Κέντρο TALP του Πανεπιστημίου της Καταλονίας. Οι συνεδριάσεις (sessions) οι οποίες χρησιμοποιήθηκαν στα πειράματα κατά τις φάσεις της εκπαίδευσης και της κατηγοριοποίησης περιείχαν μόνο απομονωμένα (isolated) ακουστικά γεγονότα, δηλαδή δεν υπήρχαν επικαλύψεις μεταξύ των ήχων. Μεταξύ των γεγονότων παρεμβάλλονταν παύσεις μερικών δευτερολέπτων. Στη βάση δεδομένων υπάρχουν και συνεδριάσεις με ήχους που επικαλύπτονται, οι οποίες ωστόσο δεν χρησιμοποιήθηκαν στα πειράματα.

Τα ηχητικά σήματα της βάσης καταγράφηκαν από 24 συνολικά μικρόφωνα, τα οποία ήταν οργανωμένα σε ομάδες. Μέσα στην αίθουσα συνεδριάσεων υπήρχαν 6 τέτοιες T-σχήματος ομάδες μικροφώνων, όπως φαίνεται στο Σχήμα 5.1 το οποίο παρουσιάζει την κάτοψη του UPC-δωματίου στο οποίο έγιναν οι ηχογραφήσεις. Ακόμη, στο Σχήμα 5.1 εκτός από τις θέσεις των μικροφώνων και των καμερών φαίνεται και η θέση P1 από την οποία καθε φορά ένας από τους συμμετέχοντες έπρεπε να παράγει μια σειρά από ήχους. Τα ηχητικά δεδομένα καταγράφηκαν με ρυθμό δειγματοληψίας (sample rate) 44.1 kHz και αποθηκεύτηκαν σε *.wv αρχεία. Όλα τα κανάλια ήταν συγχρονισμένα.



Σχήμα 5.1: Κάτοψη του UPC-δωματίου στο οποίο έγιναν οι ηχογραφήσεις των συνεδριάσεων. Έχει ληφθεί από το [8]

Από τις καταγεγραμμένες συνεδριάσεις οι 8 (συνεδρίες S01-S08) χρησιμοποιήθηκαν στα πειράματα αφού πρώτα μετατράπηκαν σε *.wav αρχεία. Σε κάθε πείραμα κάποιες από αυτές επιλέγονταν για την δημιουργία του μοντέλου κατά την φάση της εκπαίδευσης ενώ οι υπόλοιπες για την κατηγοριοποίηση των γεγονότων κατά την φάση της δοκιμής. Η UPC-TALP Multimodal Database περιέχει 12 κλάσεις ηχητικών γεγονότων που μπορούν να λάβουν χώρα σε μια συνεδρίαση, οι οποίες φαίνονται στον Πίνακα 5.1 μαζί με τον συνολικό αριθμό των γεγονότων κάθε κλάσης στις συνεδρίες S01-S08. Εκτός από τις 12 αυτές κλάσεις με το σύμβολο "si" καταγράφεται η μη ύπαρξη κάποιου γεγονότος. Τα σύμβολα των κλάσεων, δηλαδή τα ηχητικά γεγονότα, και οι χρονικές στιγμές που αυτά ξεκινούν και ολοκληρώνονται σε κάθε συνεδρίαση καταγράφονται σε ένα *.csv αρχείο. Σε κάθε συνεδρία αντιστοιχεί και ένα τέτοιο *.csv αρχείο.

5.2 Αποτελέσματα πειραμάτων

Στο παρών κεφάλαιο θα παρουσιαστούν αναλυτικά τα αποτελέσματα όλων των πειραμάτων που πραγματοποιήσαμε στο πλαίσιο αυτής της εργασίας. Αρχικά, γίνεται αναφορά στα αποτελέσματα των πειραμάτων στα οποία χρησιμοποιήθηκαν τα βασικά είδη ακουστικών χαρακτηριστικών και ταξινομητών. Γίνεται συνδυασμός των χαρακτηριστικών που περιγράφονται στο κεφάλαιο 2, δηλαδή των MFCC, των RASTA-PLP και των AM-FM χαρακτηριστικών, με τους ταξινομητές του κεφαλαίου 3, δηλαδή τους ταξινομητές GMM, kNN

Πίνακας 5.1: Κλάσεις της UPC-TALP Multimodal Database

Ηχητικό Γεγονός	Σύμβολο	Αριθμός Γεγονότων
Χτύπημα Πόρτας/Τραπεζιού	kn	79
Δυνατό Κλείσιμο Πόρτας	ds	256
Βήματα	st	206
Μετακίνηση Καρέκλας	cm	245
Κουτάλι/Κουδούνισμα Φλιτζανιού	cl	96
Ήχος/Τύλιγμα Χαρτιού Εργασίας	pw	91
Ήχος Κλειδιών	kj	82
Ήχος Πληκτρολογίου	kt	89
Ήχος Τηλεφώνου/Μουσική	pr	101
Χειροκροτήματα	ap	83
Βήχας	co	90
Ομιλία	sp	74

και SVM. Έτσι γίνεται μια πρώτη προσπάθεια εκτίμησης της συμβολής των διάφορων χαρακτηριστικών και ταξινομητών στο πρόβλημα της κατηγοριοποίησης των ακουστικών γεγονότων στο πλαίσιο μιας συνεδρίασης. Στη συνέχεια του κεφαλαίου, παρουσιάζονται τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν με τη χρήση των μοντέλων του κεφαλαίου 4. Ουσιαστικά θα εξεταστεί κατά πόσο ο συνδυασμός της πληροφορίας των ηχητικών σημάτων των 24 καναλιών που κατέγραφαν τις συνεδριάσεις βοηθάει στην βελτίωση των αποτελεσμάτων της κατηγοριοποίησης.

5.2.1 Αποτελέσματα πειραμάτων με το συνδυασμό των βασικών ειδών χαρακτηριστικών και ταξινομητών

Στην υποενότητα αυτή θα γίνει μια σύντομη παρουσίαση των αποτελεσμάτων τως αρχικών πειραμάτων, κατά τα οποία χρησιμοποιούνται τα MFCC, RASTA-PLP και AM-FM χαρακτηριστικά κατά την φάση της εξαγωγής χαρακτηριστικών και οι GMM, kNN και SVM ταξινομητές κατά την φάση της εκπαίδευσης. Τα πειράματα της υποενότητας αυτής έχουν ως στόχο μια πρώτη εκτίμηση της επίδοσης των βασικών ακουστικών χαρακτηριστικών και ταξινομητών στο πλαίσιο του προβλήματος της κατηγοριοποίησης.

Σε ότι αφορά την οργάνωση των πειραμάτων χρησιμοποιούνται οι 6 πρώτες συνεδρίες της βάσης δεδομένων (S01-S06) ως σύνολο εκπαίδευσης για την εκπαίδευση του εκάστοτε ταξινομητή, και οι υπόλοιπες δύο (S07-S08) ως σύνολο δοκιμής για την εξαγωγή των

σωστών προβλέψεων του κάθε μοντέλου. Η χρήση αυτών των σταθερών συνόλων εκπαίδευσης και δοκιμής δεν επιτρέπει την εξαγωγή απόλυτα αντιπροσωπευτικών αποτελεσμάτων. Για την εξαγωγή πιο "δίκαιων" αποτελεσμάτων το κάθε πείραμα θα έπρεπε να επαναλαμβάνεται 8 φορές, ώστε κάθε φορά μια διαφορετική από τις 8 συνεδρίες να αποτελεί το σύνολο δοκιμής. Κατά τις φάσεις της εξαγωγής χαρακτηριστικών αλλά και της κατηγοριοποίησης των άγνωστων ακουστικών γεγονότων χρησιμοποιήθηκε μόνο το πρώτο από τα 24 συνολικά κανάλια. Ακόμη, να τονιστεί ότι στα αρχεία των ηχητικών σημάτων εφαρμόζεται η τεχνική της υποδειγματοληψίας (downsampling) ώστε να μειωθεί ο ρυθμός δειγματοληψίας από τα 44.1 kHz στα 16 kHz.

Κατά την φάση της εκπαίδευσης, οι πίνακες των χαρακτηριστικών, που παράγονται κατά την φάση της εξαγωγής χαρακτηριστικών, μπαίνουν αυτούσιοι στα δεδομένα του συνόλου εκπαίδευσης. Στη συνέχεια, κατά την φάση της κατηγοριοποίησης, εξάγουμε τον πίνακα χαρακτηριστικών του κάθε άγνωστου ακουστικού γεγονότος του συνόλου δοκιμής και το μοντέλο μας προβλέπει την κλάση καθενός από τα διανύσματα του πίνακα αυτού ξεχωριστά. Τέλος, επιλέγεται ως κλάση του άγνωστου ακουστικού γεγονότος αυτή που πλειοψηφεί μεταξύ των προβέψεων (majority voting). Κατά την εξαγωγή των αποτελεσμάτων, αν c_i είναι ο αριθμός των σωστών προβλέψεων του μοντέλου για τα ακουστικά γεγονότα της συνεδρίας i, και t_i ο συνολικός αριθμός των γεγονότων της συνεδρίας i, ο αριθμός των σωστών προβλέψεων C σε κάθε πείραμα είναι :

$$C = \frac{c_7 + c_8}{t_7 + t_8} \tag{5.1}$$

Τα αποτελέσματα αυτών των πειραμάτων φαίνονται στον Πίνακα 5.2. Αρχικά, παρατηρούμε ότι ο συνδυασμός του GMM ταξινομητή και των RASTA-PLP χαρακτηριστικών πετυχαίνουν το καλύτερο

Πίνακας 5.2: Αποτελέσματα πειραμάτων με συνδυασμό των βασικών χαρακτηριστικών - ταξινομητών

Ταξινομητής	Χαρακτηριστικά	Ποσοστό Επιτυχίας (%)
SVM	RASTA-PLP	88,1
kNN	RASTA-PLP	74,7
kNN	MFCC	92,4
kNN	AM-FM	82,4
GMM	RASTA-PLP	97,1
GMM	MFCC	96,3
GMM	AM-FM	81,8

αποτέλεσμα. Ωστόσο, τα αποτελέσματα των RASTA-PLP με τους άλλους δύο ταξινομητές δεν είναι το ίδιο καλά. Αντιθέτως, τα MFCC διατηρούν τα υψηλά ποσοστά τους σε συνδυασμό και με τον GMM και με τον kNN ταξινομητή. Παρατηρούμε ακόμη πως λαμβάνουμε τα χειρότερα αποτελέσματα με τη χρήση των AM-FM χαρακτηριστικών. Από τους ταξινομητές αυτός που έχει τα υψηλότερα ποσοστά επιτυχίας είναι ο GMM.

Τέλος, να αναφερθεί ότι έγιναν κάποιες προσπάθειες βελτίωσης των αποτελεσμάτων αυτών, με την εφαρμογή της Κανονικοποίησης Μέσης Τιμής (Mean Normalization) ή με την προσθήκη της πρώτης παραγώγου και της ενέργειας των σημάτων στους πίνακες των χαρακτηριστικών, χωρίς ωστόσο να υπάρξει αποτέλεσμα.

5.2.2 Αποτελέσματα των πειραμάτων με το συνδυασμό των καναλιών

Σε αυτή την υποενότητα θα γίνει αναλυτική παρουσίαση των αποτελεσμάτων από τα πειράματα που πραγματοποιήθηκαν στο πλαίσιο του κεφαλαίου 4. Στο κεφάλαιο αυτό δημιουργήθηκαν μοντέλα που συνδύαζαν την πληροφορία των 24 καναλιών που κατέγραφαν τις συνεδριάσεις με σκοπό την βελτίωση των αποτελεσμάτων της κατηγοριοποίησης.

Πριν παρουσιάσουμε τα αποτελέσματα αυτά, θα γίνει αναφορά στο πλαίσιο κάτω από το οποίο πραγματοποιούνται αυτά τα πειράματα. Όλα τα πειράματα διεξάγονται κάτω από ένα κοινό πλαίσιο, έστι ώστε να είναι δυνατή η σύγκριση των αποτελεσμάτων τους. Αρχικά, για την δημιουργία όλων των μοντέλων που θα ακολουθήσουν επιλέγεται η χρήση των MFCC ακουστικών χαρακτηριστικών και του GMM ταξινομητή. Στα ηχητικά αρχεία των συνεδριάσεων της βάσης δεδομένων εφαρμόζεται η τεχνική της υποδειγματοληψίας (downsampling) ώστε να μειωθεί ο ρυθμός δειγματοληψίας από τα 44.1 kHz στα 16 kHz.

Σε ότι αφορά την οργάνωση των πειραμάτων εφαρμόζεται η μέθοδος 8-cross fold validation, η οποία περιγράφεται στη εισαγωγή του κεφαλαίου 4. Η χρήση της μεθόδου αυτής έχει σαν αποτέλεσμα την εξαγωγή "δίκαιων" αποτελεσμάτων, αφού το κάθε πείραμα επαναλαμβάνεται 8 φορές, ώστε κάθε φορά μια διαφορετική από τις 8 συνεδρίες να αποτελεί το σύνολο δοκιμής. Η χρήση αυτής της μεθόδου διαφοροποιεί αρκετά τα πειράματα της υποενότητας αυτής από τα πειράματα της 5.2.1. Άλλη μια σημαντική διαφορά παρατηρείται στον τρόπο κατηγοριοποίησης των άγνωστων ακουστικών γεγονότων. Σε αυτά τα πειράματα, αφού εξάγουμε τα MFCC χαρακτηριστικά ενός άγνωστου ακουστικού γεγονότος O, υπολογίζουμε τις PDF τιμές με τη χρήση του μοντέλου GMM που έχει δημιουργηθεί. Έτσι παράγε-

Πίνακας 5.3: Αποτελέσματα μοντέλου με τη χρήση των ηχητικών σημάτων μόνο του πρώτου καναλιού

Συνεδρία Δοκιμής	Σωστές Προβλέψεις/Σύνολο Γεγονότων
1	124/125
2	142/146
3	140/144
4	141/152
5	173/191
6	209/223
7	216/222
8	274/286
ΣΥΝΟΛΟ	1419/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)	95,3

ται ένας πίνακας μεγέθους 12 x N ο οποίος περιέχει τις PDF τιμές $p(O_i|c_k)$ και των 12 κλάσεων c_k σε καθένα από τα N πλαίσια του άγνωστου ακουστικού γεγονότος. Στη συνέχεια, υπολογίζουμε το άθροισμα των λογαρίθμων των PDF τιμών για κάθε μία κλάση c_k, $k = 1, \ldots, 12$, ως εξής :

$$\mathcal{L}(O|c_k) = \sum_{i=1}^{N} \log[p(O_i|c_k)] \tag{5.2}$$

Τελικά, μετά και τον υπολογισμό των πιθανοτήτων της 5.2, επιλέγεται ως κλάση του άγνωστου γεγονότος O η κλάση c_k για την οποία ισχύει:

$$c_k = \arg\max_k \mathcal{L}(O|c_k) \tag{5.3}$$

Αρχικά, το πρώτο μοντέλο που δημιουργείται, χρησιμοποιεί τα MFCC κατά την φάση της εξαγωγής χαρακτηριστικών και τον GMM ως ταξινομητή. Χρησιμοποιούνται τα ηχητικά σήματα μόνο του πρώτου από τα 24 κανάλια της βάσης δεδομένων. Τα αποτελέσματα του μοντέλου, με την διεξαγωγή των πειραμάτων σύμφωνα με το πλαίσιο που περιγράφηκε παραπάνω, φαίνονται στον Πίνακα 5.3, όπου κάθε φορά μια διαφορετική συνεδρία αποτελεί το σύνολο δοκιμής. Ακόμη, φαίνονται οι σωστές προβλέψεις του μοντέλου ξεχωριστά στα άγνωστα ακουστικά γεγονότα κάθε συνεδρίας. Το ποσοστό επιτυχίας του μοντέλου είναι 95,3%, και είναι κατά μία μονάδα μικρότερο από αυτό του αντίστοιχου μοντέλου της προηγούμενης υποενότητας (Πίνακας 5.2).

Πίνακας 5.4: Αποτελέσματα μοντέλων μονού καναλιού και πολλαπλών καναλιών

Συνεδρία Δοκιμής	Μοντέλο Μονού Καναλιού	Μοντέλο Πολλαπλών Καναλιών
1	118/125	120/125
2	142/146	141/146
3	140/144	139/144
4	138/152	141/152
5	184/191	184/191
6	219/223	219/223
7	217/222	217/222
8	275/286	276/286
ΣΥΝΟΛΟ	1433/1489	1437/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)	96,2	96,5

Το ποσοστό αυτό είναι ιδιαίτερα σημαντικό καθώς θα αποτελέσει μέτρο σύγκρισης για την επιτυχία των μοντέλων που θα παρουσιαστούν παρακάτω και τα οποία δεν χρησιμοποιούν μόνο το πρώτο κανάλι αλλά συνδυάζουν με διαφορετικό τρόπο και τα 24 κανάλια.

Στη συνέχεια, παρουσιάζονται τα αποτελέσματα των δύο μοντέλων που δημιουργούνται με το συνδυασμό και των 24 καναλιών στο επίπεδο της απόφασης, και τα οποία περιγράφονται αναλυτικά στην ενότητα 4.2. Πρόκειται για το Μοντέλο Μονού Καναλιού και το Μοντέλο Πολλαπλών Καναλιών. Το πρώτο επιτυγχάνει ποσοστό επιτυχίας 96,2% ενώ το δεύτερο 96,5%. Η αναλυτική εξαγωγή των αποτελεσμάτων των δύο αυτών μοντέλων παρουσιάζεται στον Πίνακα 5.4.

Από τα αποτελέσματα του Πίνακα 5.4, παρατηρούμε πως με το συνδυασμό της ηχητικής πληροφορίας και των 24 καναλιών στο επίπεδο της απόφασης επιτυγχάνεται μια σημαντική βελτίωση, σε σχέση με το μοντέλο που έκανε χρήση μόνο του πρώτου καναλιού (Πίνακας 5.3). Η βελτίωση αυτή πλησιάζει τη μία ποσοστιαία μονάδα στην περίπτωση του Μοντέλου Μονού Καναλιού και γίνεται ακόμα μεγαλύτερη στο Μοντέλο Πολλαπλών Καναλιών. Ακόμη, παρατηρούμε πως τα δύο αυτά μοντέλα διατηρούν την πολύ καλή τους απόδοση σε κάθε μία από τις 8 συνεδρίες. Αντίθετα με τη χρήση μόνο ενός τυχαίου καναλιού, που στην περίπτωσή μας είναι το πρώτο από τα 24, παρατηρούνται εξαιρετικά αποτελέσματα σε ορισμένες από τις συνεδρίες και εξαιρετικά χαμηλά σε κάποιες άλλες. Το μοντέλο, δηλαδή, δεν διατηρεί την σταθερότητά του, καθώς το τυχαίο κανάλι που επιλέγεται μπορεί να είναι το "καλύτερο" σε κάποιες συνεδρίες αλλά και το "χειρότερο" σε κάποιες άλλες.

Πίνακας 5.5: Αποτελέσματα μοντέλων μονού καναλιού και πολλαπλών καναλιών με τη χρήση του καλύτερου βάσει SNR καναλιού

Συνεδρία Δοκιμής	Καλύτερο Κανάλι	Μοντέλο Μονού Καναλιού	Μοντέλο Πολλαπλών Καναλιών
1	4	119/125	123/125
2	12	142/146	138/146
3	13	140/144	139/144
4	4	141/152	142/152
5	13	184/191	184/191
6	13	217/223	219/223
7	10	218/222	216/222
8	13	275/286	277/286
ΣΥΝΟΛΟ		1436/1489	1438/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)		96,5	96,6

Στη συνέχεια, θα παρουσιάσουμε δύο μοντέλα παρόμοια με τα μοντέλα Μονού και Πολλαπλών Καναλιών που περιγράφηκαν παραπάνω, τα οποία ωστόσο έχουν μια σημαντική διαφοροποίηση. Κατά την φάση της κατηγοριοποίησης δεν επιλέγεται πλέον η πλειοψηφούσα κλάση μεταξύ των 24 προβλέψεων των ταξινομητών, που αντιστοιχούν στα 24 κανάλια. Για κάθε συνεδρίαση επιλέγονται ως κλάσεις των άγνωστων ακουστικών γεγονότων οι προβλέψεις του ταξινομητή που αντιστοιχεί στο καλύτερο από τα 24 κανάλια, σύμφωνα με την μετρική του SNR. Για αυτόν το λόγο, για κάθε συνεδρίαση υπολογίζονται οι τιμές του SNR σε καθένα από τα 24 κανάλια καταγραφής των ακουστικών γεγονότων. Τα αποτελέσματα των δύο μοντέλων με την μικρή αυτή διαφοροποίηση, αλλά και το καλύτερο βάσει του SNR κανάλι της κάθε συνεδρίας παρουσιάζονται στον Πίνακα 5.5.

Στον Πίνακα 5.5 παρατηρούμε ότι το Μοντέλο Μονού Καναλιού με τη χρήση του καλύτερου βάσει SNR καναλιού κατά τη φάση της κατηγοριοποίησης έχει ποσοστό επιτυχίας 96,5%, δηλαδή παρουσιάζει μια βελτίωση της τάξης του 0,3% από το μοντέλο που κάνει χρήση και των 24 καναλιών (Πίνακας 5.4). Γενικά επειδή τα ποσοστά των μοντέλων στο πρόβλημα της κατηγοριοποίησης είναι αρκετά υψηλά, βελτιώσεις αυτής της τάξης θεωρούνται ιδιαίτερα σημαντικές. Στο αντίστοιχο Μοντέλο Πολλαπλών Καναλιών η βελτίωση που επιτυγχάνεται είναι αρκετά μικρή και αγγίζει το 0,1%.

Εκτός από το SNR, υπάρχουν άλλες δύο μετρικές εκτίμησης της καναλικής πληροφορίας οι οποίες περιγράφονται στην υποενότητα 4.3.2. Αυτές είναι η Average N-best log-likelihood difference και η N-best

Πίνακας 5.6: Αποτελέσματα μοντέλου μονού καναλιού με τη χρήση του καλύτερου καναλιού βάσει της μετρικής Average 4-best log-likelihood difference

Συνεδρία Δοκιμής	Καλύτερο Κανάλι	Σωστές Προβλέψεις/ Σύνολο Γεγονότων
1	10	120/125
2	6	144/146
3	12	140/144
4	2	139/152
5	17	183/191
6	21	219/223
7	3	218/222
8	24	276/286
ΣΥΝΟΛΟ		1439/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)		96,7

log-likelihood dispersion. Τα αποτελέσματα των μοντέλων, τα οποία χρησιμοποιούν τις δύο παραπάνω μετρικές αντί του SNR για την επιλογή του καλύτερου καναλιού κατά την φάση της κατηγοριοποίησης, παρουσιάζονται στους Πίνακες 5.6 και 5.7.

Το μοντέλο που χρησιμοποιεί την μέση διαφορά της λογαριθμικής πιθανοφάνειας επιτυγχάνει ποσοστό 96,7% και το μοντέλο που χρησιμοποιεί την διασπορά της λογαριθμικής πιθανοφάνειας 96,6%. Παρατηρούμε πως υπάρχει περεταίρω βελτίωση των αποτελεσμάτων σε σχέση με το μοντέλο κατά το οποίο γίνεται η επιλογή του καλύτερου καναλιού βάσει του SNR (Πίνακας 5.5). Μπορεί η βελτίωση αυτή να φαίνεται μικρή, ωστόσο αν αναλογιστούμε την επίδοση του αρχικού Μοντέλου Μονού Καναλιού (Πίνακας 5.4) κατανοούμε πως είναι ιδιαίτερα σημαντική καθώς φτάνει το 0,5%.

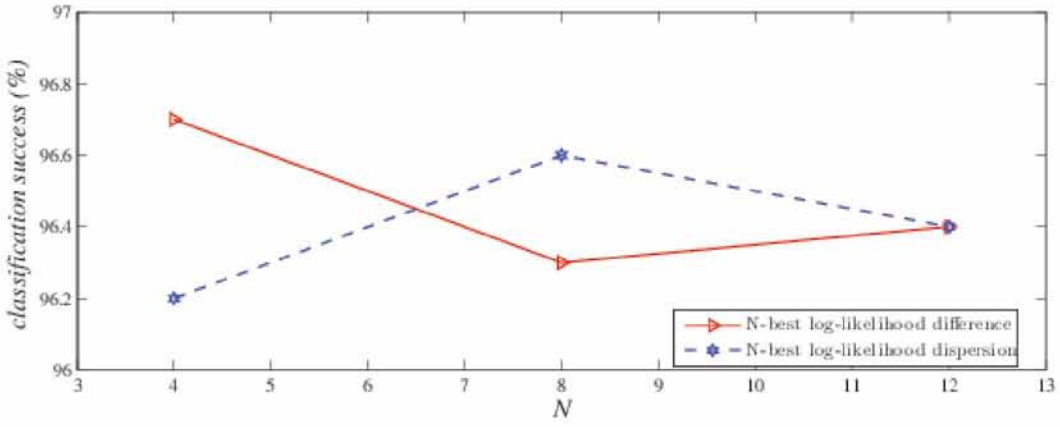
Μια σημαντική παράμετρος των δύο αυτών μέτρων, όπως φαίνεται και από την ονομασία τους, είναι το N. Η μεταβλητή αυτή ορίζει τον αριθμό των πιθανότερων κλάσεων κατά την κατηγοριοποίηση, οι οποίες λαμβάνουν μέρος στον υπολογισμό των μέτρων, και επομένως $N = 1, \ldots, 12$. Η αλλαγή της μεταβλητής αυτής επηρεάζει το τελικό αποτέλεσμα των μέτρων και επομένως και το κανάλι που επιλέγεται κατά την φάση της κατηγοριοποίησης. Στα πειράματα δίνουμε στην μετάβλητη αυτή τις τιμές 4, 8, 12, και λαμβάνουμε τα καλύτερα αποτελέσματα για $N = 4$ στο μέτρο της μέσης διαφοράς της λογαριθμικής πιθανοφάνειας και για $N = 8$ στο μέτρο της διασποράς της λογαριθμικής πιθανοφάνειας (Σχήμα 5.2). Τα καλύτερα αποτελέσματα των δύο μέτρων είναι και αυτά που παρουσιάζονται στους Πίνακες 5.6 και 5.7. Η επίδραση της αλλαγής της μεταβλητής N στα αποτελέσματα των

Πίνακας 5.7: Αποτελέσματα μοντέλου μονού καναλιού με τη χρήση του καλύτερου καναλιού βάσει της μετρικής 8-best log-likelihood dispersion

Συνεδρία Δοκιμής	Καλύτερο Κανάλι	Σωστές Προβλέψεις/ Σύνολο Γεγονότων
1	1	124/125
2	12	142/146
3	7	138/144
4	6	138/152
5	17	183/191
6	21	219/223
7	3	218/222
8	24	276/286
ΣΥΝΟΛΟ		1438/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)		96,6

μοντέλων που κάνουν χρήση των δύο αυτών μέτρων για την επιλογή του καλύτερου καναλιού κατά την φάση της κατηγοριοποίησης φαίνεται στο Σχήμα 5.2.

Στο διάγραμμα του σχήματος αυτού παρατηρούμε πως με τη χρήση των 4 καλύτερων κλάσεων στον υπολογισμό των μέτρων, το μοντέλο που χρησιμοποιεί την διαφορά της λογαριθμικής πιθανοφάνειας επιτυγχάνει την καλύτερη επίδοσή του ενώ στη συνέχεια φθίνει, καθώς ο αριθμός των κλάσεων αυξάνεται στο 8. Αντίθετη ακριβώς πορεία σε αυτό το διάστημα έχει το άλλο μοντέλο, το οποίο χρησιμοποιεί την



Σχήμα 5.2: Επίδραση της μεταβλητης N στα αποτελέσματα της κατηγοριοποίησης

Πίνακας 5.8: Αποτελέσματα μοντέλου με το συνδυασμό και των 24 καναλιών στο επίπεδο του ηχητικού σήματος με ή χωρίς τη χρήση κατωφλίου θορύβου

Συνεδρία Δοκιμής	Κατώφλι Θορύβου=1	Κατώφλι Θορύβου=0
1	119/125	120/125
2	138/146	140/146
3	133/144	132/144
4	139/152	141/152
5	182/191	183/191
6	218/223	217/223
7	219/222	217/222
8	270/286	272/286
ΣΥΝΟΛΟ	1418/1489	1422/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)	95,2	95,5

διασπορά της λογαριθμικής πιθανοφάνειας για την επιλογή καναλιού. Τα δύο μοντέλα καταλήγουν να έχουν την ίδια απόδοση όταν και οι 12 κλάσεις του προβλήματος λαμβάνονται υπόψη στον υπολογισμό των μέτρων.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα των μοντέλων που χρησιμοποιούν το λογισμικό BeamformIt για το συνδυασμό των καναλιών στο επίπεδο του ηχητικού σήματος, όπως περιγράφεται στην ενότητα 4.3. Στα μοντέλα αυτής την ενότητας εισάγουμε έναν αριθμό ηχητικών σημάτων στην είσοδο του BeamformIt και λαμβάνουμε στην έξοδο ένα ενισχυμένο σήμα, το οποίο στη συνέχεια χρησιμοποιείται στις φάσεις του προβλήματος της κατηγοριοποίησης. Πολύ σημαντικό ρόλο στα αποτελέσματα διαδραματίζουν οι παράμετροι που δίνουμε στο λογισμικό, οι οποίες περιγράφονται αναλυτικά στην υποενότητα 4.3.1.

Στην πρώτη μας προσπάθεια, για κάθε συνεδρία εισάγουμε τα ηχητικά σήματα και των 24 καναλιών στο BeamformIt. Γνωρίζουμε πως η χρήση και των 24 καναλιών, από τα οποία μερικά είναι χαμηλής ποιότητας, ίσως να οδηγήσει σε μικρή μείωση των αποτελεσμάτων. Γι' αυτό και στη συνέχεια θα προσπαθήσουμε να μειώσουμε τον αριθμό των σημάτων που εισάγονται στο λογισμικό. Ακόμη, δεν επιλέγουμε εμείς το κανάλι αναφοράς το οποίο υπολογίζεται με την μέθοδο της ετεροσυσχέτισης (cross-correlation). Τέλος, στο ένα πείραμα γίνεται χρήση του αλγορίθμου του κατωφλίου θορύβου (noise thresholding algorithm) για την απαλοιφή, κατά τον υπολογισμό των χρονικών καθυστερήσεων, των κομματιών των ηχητικών σημάτων που θεω-

Πίνακας 5.9: Αποτελέσματα μοντέλων με την επιλογή του καναλιού αναφοράς βάσει του SNR και της log-likelihood difference

Συνεδρία Δοκιμής	SNR	log-likelihood difference
1	118/125	123/125
2	138/146	136/146
3	138/144	133/144
4	142/152	141/152
5	184/191	183/191
6	215/223	218/223
7	218/222	216/222
8	272/286	272/286
ΣΥΝΟΛΟ	1425/1489	1422/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)	95,7	95,5

ρούνται ως θόρυβος απο το BeamformIt, ενώ στο άλλο πείραμα δεν γίνεται χρήση του αλγορίθμου. Τα αποτελέσματα του μοντέλου αυτού, με ή χωρίς τη χρήση του κατωφλίου θορύβου παρουσιάζονται στον Πίνακα 5.8. Το μοντέλο που χρησιμοποιεί το κατώφλι θορύβου επιτυγχάνει ποσοστό 95,2% ενώ αυτό που δεν το χρησιμοποιεί 95,5%.

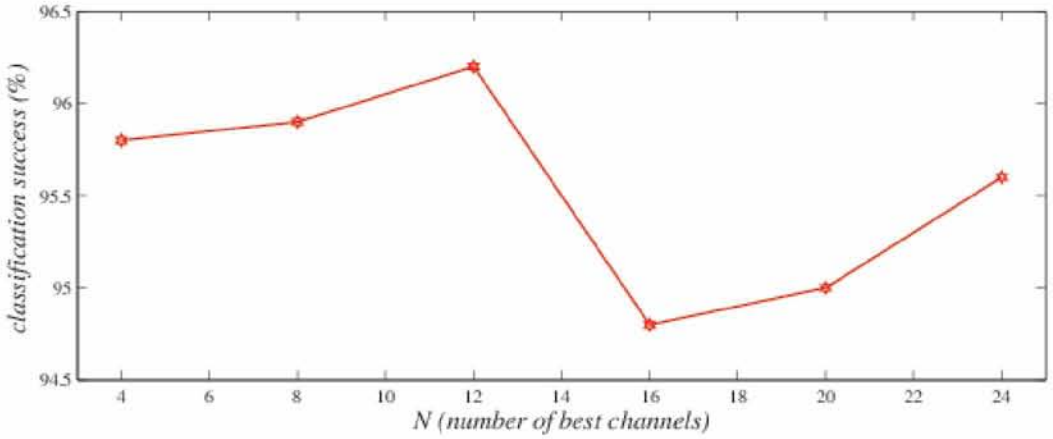
Παρατηρούμε πως υπάρχει μια μικρή μείωση της επίδοσης του μοντέλου που κάνει χρήση του BeamformIt, σε σχέση με τα μοντέλα που έχουν περιγραφεί μέχρι στιγμής. Αυτό, όπως έχουμε αναφέρει, ίσως να οφείλεται στο συνδυασμό και των 24 καναλιών, μερικά από τα οποία να μην είναι της απαιτούμενης ποιότητας και έτσι να επηρεάζουν αρνητικά το παραγόμενο σήμα. Ακόμη, προκαλεί εντύπωση πως με τη μη χρήση του κατωφλίου θορύβου το μοντέλο παρουσιάζει βελτίωση της επίδοσής του, η οποία είναι της τάξης του 0,3%. Μια πιθανή εξήγηση είναι ότι το λογισμικό με τη χρήση του κατωφλίου θορύβου λαμβάνει λανθασμένα ως θόρυβο χρήσιμα τμήματα των σημάτων και δεν τα συμπεριλαμβάνει στον υπολογισμό των χρονικών καθυστερήσεων, όπως θα έπρεπε να κάνει.

Στο σημείο αυτό, θα γίνει προσπάθεια βελτίωσης αυτών των αποτελεσμάτων με την επιλογή του καναλιού αναφοράς σύμφωνα με τις μετρικές του SNR και της διαφοράς της λογαριθμικής πιθανοφάνειας (log-likelihood difference). Αυτό έχει ως αποτέλεσμα τη μη διεξαγωγή της μεθόδου της ετεροσυσχέτισης κατά την επεξεργασία των ηχητικών σημάτων από το BeamformIt. Επιλέγουμε εμείς το καλύτερο κανάλι, σύμφωνα με τις παραπάνω δύο μετρικές, ως κανάλι αναφοράς καθώς ο σωστός υπολογισμός των καθυστερήσεων εξαρτάται από την επιλογή του καναλιού αυτού. Ακόμη, από εδώ και

στο εξής επιλέγουμε τη μη διεξαγωγή του αλγορίθμου του κατωφλίου θορύβου ο οποίος οδήγησε στο καλύτερο αποτέλεσμα (Πίνακας 5.8). Τα αποτελέσματα των δύο αυτών μοντέλων, με επιλογή του καναλιού αναφοράς βάσει του SNR και της διαφοράς της λογαριθμικής πιθανοφάνειας αντίστοιχα, παρουσιάζονται στον Πίνακα 5.9. Παρατηρούμε πως με την επιλογή του καναλιού αναφοράς βάσει του SNR το μοντέλο επιτυγχάνει ποσοστό 95,7%, δηλαδή παρουσιάζει βελτίωση 0,2%. Αντίθετα, με τη χρήση του μέτρου της διαφοράς της λογαριθμικής πιθανοφάνειας στην επιλογή του καναλιού αναφοράς το αποτέλεσμα του μοντέλου παραμένει στα ίδια ακριβώς επίπεδα (95,5%) και δεν παρουσιάζει κάποια βελτίωση.

Στα μοντέλα που ακολουθούν πραγματοποιούμε μια σημαντική αλλαγή, με την οποία αποβλέπουμε στη βελτίωση των αποτελεσμάτων των μοντέλων που έχουν ήδη περιγραφεί και συνδυάζουν τα ηχητικά σήματα των 24 καναλιών με τη χρήση του BeamformIt. Πλέον, με βάση τη μετρική του SNR επιλέγουμε ως είσοδο στο BeamformIt μόνο τα N καλύτερα κανάλια από τα 24 συνολικά που καταγράφουν τα ακουστικά γεγονότα των συνεδριάσεων. Με τη βασική αυτη μετατροπή στα μοντέλα κατηγοριοποίησης αποφεύγεται η συμμετοχή των μη "ποιοτικών" καναλιών στην δημιουργία του ενισχυμένου ακουστικού σήματος. Για την παραγωγή του σήματος αυτού της κάθε συνεδρίας δεν συνδυάζονται πλέον και τα 24 κανάλια παρά μόνο τα N καλύτερα.

Η επιλογή του N, όπου $N = 1, \ldots, 24$, είναι ιδιαίτερα σημαντική και παίζει σημαντικό ρόλο στο ποσοστό επιτυχίας του μοντέλου. Στο διάγραμμα του Σχήματος 5.3 παρουσιάζεται το ποσοστό επιτυχίας του



Σχήμα 5.3: Ποσοστά επιτυχίας του μοντέλου με το συνδυασμό των N καλύτερων καναλιών στο επίπεδο του ηχητικού σήματος με χρήση της μετρικής του SNR

Πίνακας 5.10: Αποτελέσματα του μοντέλου με το συνδυασμό των 12 καλύτερων καναλιών στο επίπεδο του ηχητικού σήματος με χρήση της μετρικής του SNR

Συνεδρία Δοκιμής	Σωστές Προβλέψεις/Σύνολο Γεγονότων
1	123/125
2	142/146
3	139/144
4	140/152
5	183/191
6	219/223
7	217/222
8	269/286
ΣΥΝΟΛΟ	1432/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)	96,2

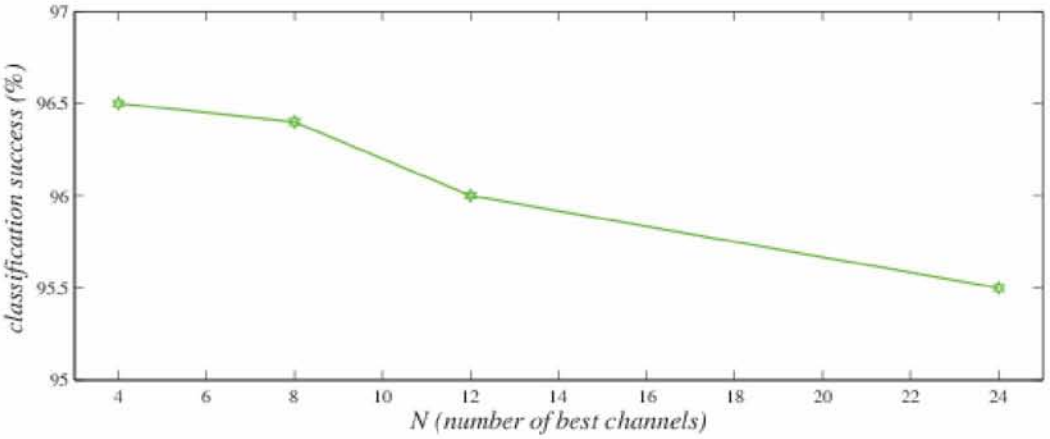
μοντέλου, αρχίζοντας με το συνδυασμό των 4 καλύτερων βάσει SNR καναλιών και καταλήγοντας με το συνδυασμό και των 24 καναλιών. Παρατηρούμε πως το ποσοστό επιτυχίας του μοντέλου παρουσιάζει μια μικρή αύξηση από $N = 4$ έως και $N = 12$, όπου και λαμβάνει την μέγιστη τιμή του (96,2%). Η εξαγωγή του αποτελέσματος αυτού, κατά το οποίο το μοντέλο συνδυάζει τα 12 καλύτερα βάσει SNR κανάλια στο επίπεδο του ηχητικού σήματος, παρουσιάζεται αναλυτικά στον Πίνακα 5.10. Η επίδοση αυτή είναι ιδιαίτερα σημαντική, καθώς βελτιώνει κατά 0,5% την μέχρι στιγμής καλύτερη επίδοση των μοντέλων που κάνουν χρήση του BeamformIt (Πίνακας 5.9). Για $N > 12$ τα ποσοστά επιτυχίας του μοντέλου βρίσκονται σε αρκετά πιο χαμηλά επίπεδα από την μέγιστη τιμή του, καθώς τα μη "ποιοτικά" κανάλια που προστίθενται επηρεάζουν αρνητικά τα παραγώμενα σήματα.

Εκτός από το μέτρο του SNR χρησιμοποιήσαμε και τη μετρική της διαφοράς της λογαριθμικής πιθανοφάνειας (log-likelihood difference) για την επιλογή των N καλύτερων καναλιών που συνδυάζονται στο επίπεδο της απόφασης με τη χρήση του BeamformIt. Η χρήση της μετρικής αυτής για την επιλογή του καναλιού αναφοράς δεν βοήθησε ιδιαίτερα στη βελτίωση των αποτελεσμάτων των μοντέλων μας (Πίνακας 5.9). Ωστόσο, η χρησιμοποίησή της στην επιλογή των πιο ποιοτικών καναλιών, τα οποία συμμετέχουν στη δημιουργία ενός ενισχυμένου σήματος για την κάθε συνεδρία, βοηθάει σημαντικά στην βελτίωση της απόδοσης του μοντέλου κατηγοριοποίησης. Και σε αυτή την περίπτωση η επιλογή της τιμής του N, δηλαδή του αριθμού των καλύτερων καναλιών που επιλέγονται ως είσοδος στο BeamformIt, είναι πολύ σημαντική και καθορίζει το ποσοστό επιτυχίας του

Πίνακας 5.11: Αποτελέσματα του μοντέλου με το συνδυασμό των 4 καλύτερων καναλιών στο επίπεδο του ηχητικού σήματος με χρήση της μετρικής της διαφοράς της λογαριθμικής πιθανοφάνειας

Συνεδρία Δοκιμής	Σωστές Προβλέψεις/Σύνολο Γεγονότων
1	123/125
2	141/146
3	137/144
4	141/152
5	187/191
6	216/223
7	217/222
8	274/286
ΣΥΝΟΛΟ	1436/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)	96,5

μοντέλου. Στο διάγραμμα του Σχήματος 5.4 παρουσιάζεται η σχέση μεταξύ της τιμής του N και του ποσοστού επιτυχίας του μοντέλου. Παρατηρούμε πως το μοντέλο επιτυγχάνει την καλύτερη επίδοση (96,5%) με την επιλογή των 4 καλύτερων καναλιών βάσει της μετρικής της διαφοράς της λογαριθμικής πιθανοφάνειας. Η εξαγωγή του αποτελέσματος αυτού παρουσιάζεται αναλυτικά στον Πίνακα 5.11. Το συγκεκριμένο ποσοστό επιτυχίας είναι εξαιρετικά σημαντικό καθώς αποτελεί την καλύτερη από τις επιδόσεις των μοντέλων που κάνουν



Σχήμα 5.4: Ποσοστά επιτυχίας του μοντέλου με το συνδυασμό των N καλύτερων καναλιών στο επίπεδο του ηχητικού σήματος με χρήση της μετρικής της διαφοράς της λογαριθμικής πιθανοφάνειας

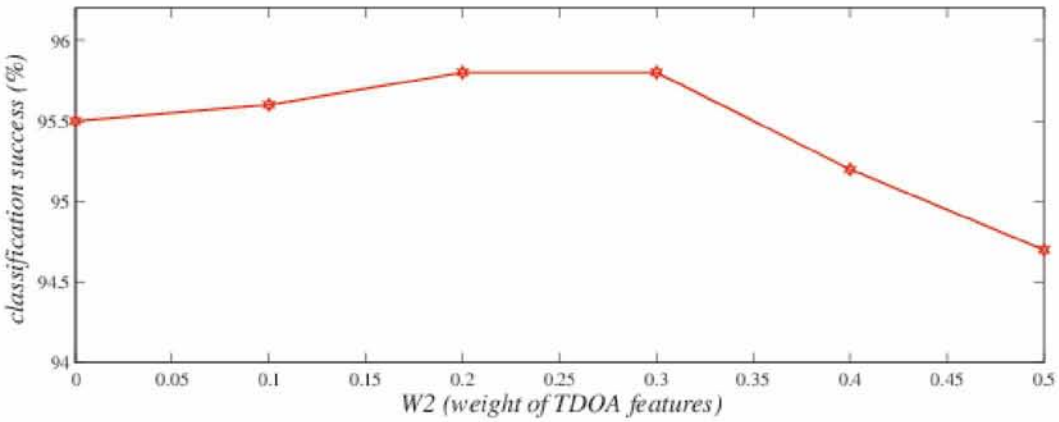
Πίνακας 5.12: Αποτελέσματα του συνδυασμού των MFCC και TDOA χαρακτηριστικών στο μοντέλο που συνδυάζει και τα 24 κανάλια στο επίπεδο του ηχητικού σήματος

Συνεδρία Δοκιμής	MFCC	MFCC+TDOA
1	120/125	120/125
2	140/146	140/146
3	132/144	132/144
4	141/152	142/152
5	183/191	184/191
6	217/223	217/223
7	217/222	218/222
8	272/286	273/286
ΣΥΝΟΛΟ	1422/1489	1426/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)	95,5	95,8

χρήση του BeamformIt. Το μοντέλο διατηρεί την υψηλή του απόδοση και για $N = 8$, όπου επιτυγχάνει ποσοστό 96,4%. Ωστόσο, στη συνέχεια η γραφική παράσταση φθίνει έως ότου λάβει την χαμηλότερη τιμή της με τη χρήση και των 24 καναλιών στη δημιουργία των παραγόμενων σημάτων.

Στο τελευταίο μέρος αυτής της υποενότητας προσπαθούμε να βελτιώσουμε τα αποτελέσματα κάποιων μοντέλων που έχουν ήδη αναφερθεί παραπάνω, με το συνδυασμό των MFCC και των TDOA χαρακτηριστικών, με τον τρόπο που περιγράφεται στην ενότητα 4.4. Το πρώτο μοντέλο συνδυάζει τα ηχητικά σήματα και των 24 κανάλιων μέσω του BeamformIt. Χρησιμοποιεί τα MFCC χαρακτηριστικά για την εκπαίδευση του GMM ταξινομητή. Το κανάλι αναφοράς υπολογίζεται μέσω της μεθόδου της ετεροσυσχέτισης ενώ δεν χρησιμοποιείται ο αλγόριθμος του κατωφλίου θορύβου. Το ποσοστό επιτυχίας του μοντέλου αυτού είναι 95,5% (Πίνακας 5.8). Τώρα δημιουργούμε ένα ίδιο μοντέλο με τη διαφορά ότι εκπαιδεύουμε τον ταξινομητή GMM με τη χρήση των TDOA χαρακτηριστικών. Τα δύο αυτά ανεξάρτητα μοντέλα συνδυάζονται κατά τη φάση της κατηγοριοποίησης σύμφωνα με τον τύπο (4.14). Τα συγκριτικά αποτελέσματα των μοντέλων παρουσιάζονται αναλυτικά στον Πίνακα 5.12. Παρατηρούμε πως με το συνδυασμό των MFCC και TDOA χαρακτηριστικών το ποσοστό επιτυχίας φτάνει το 95,8% και παρουσιάζει μια σημαντική βελτίωση της τάξης του 0,3%. Αυτό συμβαίνει καθώς οι σωστές προβλέψεις του μοντέλου με τη χρήση και των TDOA χαρακτηριστικών αυξάνονται ή στην χειρότερη περίπτωση παραμένουν στα ίδια επίπεδα σε κάθε μία από τις 8 συνεδρίες.

Δύο σημαντικές παράμετροι που καθορίζουν την απόδοση των



Σχήμα 5.5: Σχέση του ποσοστού επιτυχίας του μοντέλου και του W_2, δηλαδή του βάρους συμμετοχής των TDOA χαρακτηριστικών κατά την φάση της κατηγοριοποίησης

μοντέλων που συνδυάζουν τα MFCC και TDOA χαρακτηριστικά είναι τα W_1 και W_2 του τύπου 4.14. Το W_1 είναι το βάρος με το οποίο το μοντέλο των MFCC χαρακτηριστικών συμμετέχει στο συνδυασμό, ενω το W_2 είναι το βάρος με το οποίο συμμετέχει το μοντέλο των TDOA χαρακτηριστικών. Για τα δύο αυτά βάρη ισχύει ότι $W_1 + W_2 = 1$. Η σχέση της τιμής του W_2 με το ποσοστό επιτυχίας του παρόντος μοντέλου φαίνεται στο Σχήμα 5.5.

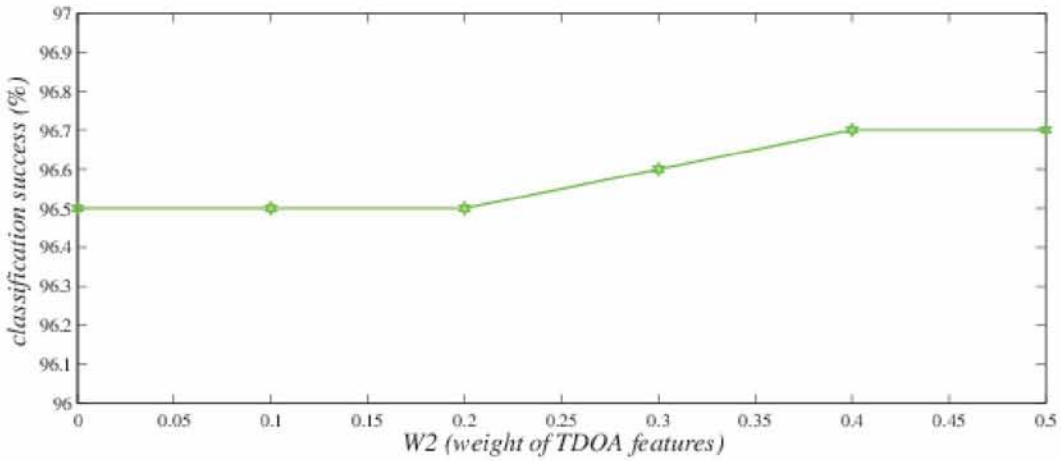
Στο σχήμα αυτό παρατηρούμε πως για $W_2 = 0$, δηλαδή όταν συμμετέχουν μόνο τα MFCC χαρακτηριστικά στην κατηγοριοποίηση το μοντέλο επιτυγχάνει ποσοστό 95,5%. Στη συνέχεια, με την αύξηση της συμμετοχής των TDOA χαρακτηριστικών το ποσοστό αυτό παρουσιάζει αύξηση και λαμβάνει την μέγιστη τιμή του (95,8%) για $W_2 = 0,2$ και $W_2 = 0,3$. Η εξαγωγή του μέγιστου αυτού ποσοστού παρουσιάζεται στον Πίνακα 5.12. Για $W_2 > 0,3$ και όσο το βάρος των TDOA χαρακτηριστικών αυξάνεται η γραφική παράσταση φθίνει. Το μοντέλο επιτυγχάνει ποσοστό 94,7% για $W_2 = 0,5$ όπου τα MFCC και τα TDOA χαρακτηριστικά συμμετέχουν ισομερώς στην κατηγοριοποίηση. Με τη συμμετοχή μόνο των TDOA χαρακτηριστικών, δηλαδή για $W_2 = 1$, το ποσοστό επιτυχίας του μοντέλου είναι ιδιαίτερα χαμηλό (40,4%).

Η επόμενή μας προσπάθεια αφορά το μοντέλο που συνδυάζει και τα 24 κανάλια στο επίπεδο του ηχητικού σήματος. Το μοντέλο αυτό έχει τις ίδιες παραμέτρους με το προηγούμενο με τη διαφορά ότι το κανάλι αναφοράς επιλέγεται με τη χρήση της μετρικής του SNR. Με τη χρήση μόνο των MFCC χαρακτηριστικών το μοντέλο επιτυγχάνει

Πίνακας 5.13: Αποτελέσματα του συνδυασμού των MFCC και TDOA χαρακτηριστικών στο μοντέλο που συνδυάζει και τα 24 κανάλια στο επίπεδο του ηχητικού σήματος επιλέγοντας το καλύτερο βάσει SNR κανάλι ως κανάλι αναφοράς

Συνεδρία Δοκιμής	MFCC	MFCC+TDOA
1	118/125	120/125
2	138/146	137/146
3	138/144	137/144
4	142/152	144/152
5	184/191	184/191
6	215/223	215/223
7	218/222	218/222
8	272/286	271/286
ΣΥΝΟΛΟ	1425/1489	1426/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)	95,7	95,8

ποσοστό 95,7% (Πίνακας 5.9). Με τη συμμετοχή και των TDOA χαρακτηριστικών η επίδοσή του παρουσιάζει μια πολύ μικρή βελτίωση της τάξης του 0,1%. Η βελτίωση αυτή αντιστοιχεί στο μέγιστο ποσοστό επιτυχίας του μοντέλου που συνδυάζει τα MFCC με τα TDOA χαρακτηριστικά, το οποίο είναι 95,8% όπως φαίνεται και στον Πίνακα 5.13. Το ποσοστό αυτό επιτυγχάνεται για βάρη $W_1 = 0,8$ και $W_2 = 0,2$.



Σχήμα 5.6: Σχέση του ποσοστού επιτυχίας του μοντέλου και του W_2, δηλαδή του βάρους συμμετοχής των TDOA χαρακτηριστικών κατά την φάση της κατηγοριοποίησης

Πίνακας 5.14: Αποτελέσματα του συνδυασμού των MFCC και TDOA χαρακτηριστικών στο μοντέλο μονού καναλιού το οποίο κάνει χρήση μόνο των καλύτερων βάσει SNR καναλιών

Συνεδρία Δοκιμής	MFCC	MFCC+TDOA
1	119/125	120/125
2	142/146	142/146
3	140/144	140/144
4	141/152	144/152
5	184/191	183/191
6	217/223	217/223
7	218/222	218/222
8	275/286	275/286
ΣΥΝΟΛΟ	1436/1489	1439/1489
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ (%)	96,5	96,7

Η τελευταία μας προσπάθεια αφορά το Μοντέλο Μονού Καναλιού στο οποίο όμως σε κάθε συνεδρία χρησιμοποιείται μόνο το καλύτερο κανάλι με βάση τη μετρική του SNR. Το μοντέλο αυτό επιτυγχάνει ποσοστό 96,5% και παρουσιάζεται στον Πίνακα 5.5. Με τη συμμετοχή και των TDOA χαρακτηριστικών κατά την κατηγοριοποίηση το μοντέλο βελτιώνει την επίδοσή του κατά 0,2% και το ποσοστό επιτυχίας του φτάνει έως και 96,7%, όπως φαίνεται στον Πίνακα 5.14. Μπορεί η βελτίωση αυτή να φαντάζει πολύ μικρή, ωστόσο αποτελεί τη μία από τις δύο καλύτερες επιδόσεις που έχουμε λάβει από τα μοντέλα που παρουσιάσαμε στην παρούσα διπλωματική. Η επίδοση αυτή επιτυγχάνεται για βάρη $W_1 = 0,6$ και $W_2 = 0,4$. Αναλυτικά τα ποσοστά επιτυχίας του μοντέλου σε σχέση με τα βάρη συμμετοχής των MFCC και TDOA χαρακτηριστικών φαίνονται στο Σχήμα 5.6.

Κεφάλαιο 6

Συμπεράσματα

6.1 Συμβολή της διπλωματικής εργασίας

Η κύρια συνεισφορά της συγκεκριμένης διπλωματικής έγκειται στη συστηματική μελέτη του προβλήματος της κατηγοριοποίησης ακουστικών γεγονότων τα οποία λαμβάνουν χώρα κατά τη διάρκεια μιας συνεδρίασης και η αξιολόγηση των διαφορετικών μοντέλων που υλοποιήθηκαν για τον σκοπό αυτό. Πιο συγκεκριμένα οι επιστημονικές της συνεισφορές συνοψίζονται στους εξής άξονες:

- Στη μελέτη μιας σειράς από τρόπους συνδυασμού των 24 καναλιών που διαθέτει η βάση δεδομένων μας για την καταγραφή των ακουστικών γεγονότων που λαμβάνουν χώρα σε μια συνεδρίαση. Γίνεται προσπάθεια συνδυασμού των καναλιών στο επίπεδο της απόφασης, δηλαδή κατά τη φάση της κατηγοριοποίησης των ακουστικών γεγονότων, αλλά και στο επίπεδο του ηχητικού σήματος, όπου για κάθε συνεδρία ένας αριθμός ηχητικών σημάτων συνδυάζεται μέσω του λογισμικού BeamformIt για την παραγωγή ενός ενισχυμένου σήματος. Με το συνδυασμό της καναλικής πληροφορίας των μικροφώνων τα μοντέλα κατηγοριοποίησης παρουσιάζουν μια πολύ σημαντική βελτίωση σε σχέση με το μοντέλο που έκανε χρήση μόνο του πρώτου καναλιού, η οποία φτάνει μέχρι και το 1,4%.

- Στη χρησιμοποίηση μετρικών εκτίμησης της ποιότητας της καναλικής πληροφορίας για την επιλογή και το συνδυασμό των καλύτερων καναλιών. Τα μέτρα που χρησιμοποιούνται είναι το SNR, η διαφορά της λογαριθμικής πιθανοφάνειας (log-likelihood difference) και η διασπορά της λογαριθμικής πιθανοφάνειας (log-likelihood dispersion). Τα μοντέλα αυτά που χρησιμοποιούν

τις μετρικές αυτές για την επιλογή των καλύτερων καναλιών για κάθε συνεδρία επιτυγχάνουν ιδιαίτερα υψηλά ποσοστά. Με τη χρήση του SNR το μοντέλο επιτυγχάνει 96,5%, με χρήση της διασποράς της λογαριθμικής πιθανοφάνειας 96,6% ενώ με τη χρήση της διαφοράς της λογαριθμικής πιθανοφάνειας 96,7%. Η τελευταία αυτή επίδοση είναι και η μία από τις δύο καλύτερες που λαμβάνουμε από τα πειράματα της παρούσας διπλωματικής. Ακόμη, οι μετρικές αυτές χρησιμοποιούνται για το συνδυασμό μόνο των N καλύτερων καναλιών στο επίπεδο του ηχητικού σήματος με τη χρήση του BeamformIt.

- Στο συνδυασμό των χρονικών καθυστερήσεων της άφιξης των ηχητικών σημάτων κάθε καναλιού (TDOA χαρακτηριστικά) με τα κλασικά MFCC χαρακτηριστικά. Ο συνδυασμός των χαρακτηριστικών αυτών κατά τη φάση της κατηγοριοποίησης των ακουστικών γεγονότων επέφερε βελτίωση στα μοντέλα που έχουμε δημιουργήσει, η οποία κυμαίνεται από 0,1% έως και 0,3%. Το μοντέλο, το οποίο χρησιμοποιεί για κάθε συνεδρία το καλύτερο κανάλι σύμφωνα με την μετρική του SNR και συνδυάζει τα TDOA και MFCC χαρακτηριστικά κατά την κατηγοριοποίηση επιτυγχάνει ποσοστό 96,7%. Η επίδοση αυτή είναι ιδιαίτερα σημαντική καθώς αποτελεί την δεύτερη από τις δύο καλύτερες επιδόσεις που λαμβάνουμε από τα μοντέλα κατηγοριοποίησης της παρούσας διπλωματικής.

6.2 Μελλοντικές ερευνητικές κατευθύνσεις

Παρόλο που στην παρούσα διπλωματική εργασία δημιουργήθηκαν αρκετά μοντέλα κατηγοριοποίησης με αρκετά ικανοποιητικά αποτελέσματα, υπάρχουν ακόμα πολλά περιθώρια βελτίωσης και επέκτασης των μοντέλων αυτών. Ορισμένες κατευθύνσεις στις οποίες θα μπορούσε να επικεντρωθεί μια μελλοντική έρευνα προτείνονται παρακάτω:

- Όσον αφορά τα ακουστικά χαρακτηριστικά που χρησιμοποιήθηκαν, θα μπορούσε να γίνει περαιτέρω έρευνα στη δημιουργία μοντέλων που χρησιμοποιούν τα AM-FM χαρακτηριστικά κατά τη φάση της εξαγωγής χαρακτηριστικών. Η απόδοση των χαρακτηριστικών αυτών στα πειράματα της παρούσας διπλωματικής ήταν ιδιαίτερα χαμηλή και απείχε αρκετά από την επίδοση των άλλων χαρακτηριστικών. Ακόμη, θα μπορούσε να γίνει προσπάθεια συνδυασμού των κλασικών MFCC χαρακτηρισικών με τα RASTA-PLP κατά τη φάση της κατηγοριοποίησης με την

δημιουργία δύο ανεξάρτητων μοντέλων, ακριβώς όπως έγινε ο συνδυασμός MFCC και TDOA χαρακτηριστικών. Τα MFCC όσο και τα RASTA-PLP είχαν ιδιαίτερα υψηλές επιδόσεις στα μοντέλα στα οποία δοκιμάστηκαν.

- Άλλο ένα πεδίο μελλοντικής έρευνας αφορά τα μοντέλα που συνδυάζουν τα κανάλια καταγραφής των ακουστικών γεγονότων στο επίπεδο του ηχητικού σήματος με την χρήση του BeamformIt. Τα αποτελέσματα κατηγοριοποίησης των μοντέλων μας με τη χρήση του BeamformIt έχουν περιθώρια βελτίωσης. Η βελτίωση αυτή μπορεί να επιτευχθεί με συστηματική έρευνα των παραμέτρων του λογισμικού, οι οποίες επηρεάζουν κατά πολύ τα παραγόμενα αποτελέσματα της κατηγοριοποίησης. Στα μοντέλα της παρούσας διπλωματικής αρκετές από τις παραμέτρους του BeamformIt διατήρησαν τις προεπιλεγμένες τιμές τους.

- Σχετικά με τη δημιουργία μοντέλων που συνδυάζουν την ηχητική πληροφορία των καναλιών που καταγράφουν τις συνεδριάσεις, θα μπορούσε να γίνει χρήση και των υπόλοιπων ακουστικών χαρακτηριστικών και ταξινομητών. Σε αυτά τα μοντέλα της παρούσας διπλωματικής γίνεται χρήση μόνο των MFCC χαρακτηριστικών και του GMM ταξινομητή. Η απόδοση των μοντέλων αυτών με τη χρήση των RASTA-PLP ή των AM-FM χαρακτηριστικών καθώς και του SVM ή του kNN ταξινομητή μπορεί να αποτελέσει αντικείμενο μελλοντικής έρευνας.

- Αντικείμενο μελλοντικής έρευνας μπορεί να αποτελέσει και ο έλεγχος της απόδοσης των μοντέλων που συνδυάζουν τα κανάλια στο επίπεδο της απόφασης αλλά και του ηχητικού σήματος σε συνεδριάσεις της βάσης δεδομένων που περιέχουν ακουστικά γεγονότα, τα οποία μπορεί και να επικαλύπτονται. Ακόμη, μπορεί να γίνει προσπάθεια βελτίωσης της επίδοσης των μοντέλων αυτών με το συνδυασμό των οπτικοακουστικών πληροφοριών για κάθε συνεδρίαση.

Bibliography

[1] A. Tempko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," in *Multimodal Technologies for Perception of Humans (CLEAR 2006)*, 2006, pp. 311–322.

[2] D. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley-IEEE Press, 2006.

[3] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proc. European Signal Processing Conference*, 2010.

[4] D. R. Reddy, "Speech recognition by machine: A review," in *Proc. of the IEEE*, 2005, pp. 501–531.

[5] E. Marcheret, G. Potamianos, K. Visweswariah, and J. Huang, "The IBM RT06s evaluation system for speech activity detection in CHIL seminars," in *MLMI*, 2006, pp. 323–335.

[6] C. Boukis and L. Polymenakos, "The acoustic event detector of AIT," in *Multimodal Technologies for Perception of Humans*, 2007, pp. 328–337.

[7] X. Zhuang, X. Zhou, T. S. Huang, and M. Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 17–20.

[8] T. Butko and C. Nadeu, "Detection of overlapped acoustic events using fusion of audio and video modalities," in *VI Jornadas en Tecnologia del Habla and II Iberian SLTech Workshop*, 2010, pp. 165–168.

[9] A. Temko and C. Nadeu, "Detection of acoustic events in interactive seminar data with temporal overlaps," in *Proc. Interspeech*, 2008, pp. 2594–2597.

[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[11] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[12] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[13] S. Young *et al.*, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[14] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an AM-FM modulation model," *Speech Communication*, vol. 28, pp. 195–209, Jan. 1999.

[15] D. Dimitriadis and P. Maragos, "Robust AM-FM features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 621–624, 2005.

[16] X. Zhuang, J. Huang, G. Potamianos, and M. Hasegawa-Johnson, "Acoustic fall detection using Gaussian mixture models and GMM supervectors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 69–72.

[17] B. Schölkopf and A. Smola, *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[18] I. Kononenko and M. Kukar, *Machine Learning and Data Mining : Introduction to Principles and Algorithms*. Horwood, 2007.

[19] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2011–2023, 2007.

[20] X. Anguera, "Robust speaker diarization for meetings," Ph.D. dissertation, UPC, Barcelona, 2006.

[21] M. Vondrasek and P. Pollak, "Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency," *Radioengineering*, vol. 14, pp. 6–11, 2005.

[22] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," in *Proc. Interspeech*, 2000, pp. 746–749.

[23] D. Vijayasenan, F. Valente, and H. Bourlard, "Integration of TDOA features in information bottleneck framework for fast speaker diarization," IDIAP, Martigny, Switzerland, Tech. Rep., 2008.

[24] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: mixing acoustic features and inter-channel time differences," in *Proc. Interspeech*, 2006.

[25] X. Anguera, C. Wooters, J. M. Pardo, and J. Hernando, "Automatic weighting for the combination of TDOA and acoustic features in speaker diarization for meetings," *Proc. International Conference on Acoustics, Speech, and Signal Processing*, vol. 14, 2007.

[26] A. Temko, D. Macho, C. Nadeu, and C. Segura, "UPC-TALP database of isolated acoustic events," UPC, Barcelona, Spain, Tech. Rep., 2005.