ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΚΡΗΤΗΣ
ΠΑΡΑΡΤΗΜΑ ΡΕΘΥΜΝΟΥ
ΤΜΗΜΑ ΜΟΥΣΙΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ ΚΑΙ ΑΚΟΥΣΤΙΚΗΣ

Π Τ Υ Χ Ι Α Κ Η Ε Ρ Γ Α Σ Ι Α

Θ Ε Μ Α : Ακουστική ανίχνευση καταστάσεων εκτάκτου ανάγκης

ΟΝΟΜ/ΜΟ ΣΠΟΥΔΑΣΤΗ: ΜΠΑΚΑΣ ΔΗΜΗΤΡΙΟΣ

Α. Μ.: 773

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΠΟΤΑΜΙΤΗΣ ΗΛΙΑΣ

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της παρούσης πτυχιακής εργασίας, κύριο Ηλία Ποταμίτη, που οι απαντήσεις στα ερωτήματα μου και οι συμβουλές του ήταν πολύ χρήσιμες και ανεκτίμητες.

ΠΡΟΛΟΓΟΣ

Αυτή η διπλωματική εργασία αποτελεί την προσπάθειά όσον αφορά την ανάπτυξη και την, εκ βαθέων, μελέτη ενός ολοκληρωμένου και λειτουργικού συστήματος, για την αυτόματη, ακουστική ανίχνευση καταστάσεων εκτάκτου ανάγκης.

Ποιο συγκεκριμένα, γίνεται μία αναφορά στα ήδη υπάρχοντα, αυτόματα συστήματα αναγνώρισης φωνής και ήχου (ανθρώπινης και μη προέλευσης) και μελετάμε τα στάδια που αποτελούν ένα τέτοιο ολοκληρωμένο σύστημα, καθώς και τις μεθόδους επεξεργασίας του σήματος που χρησιμοποιείται σε κάθε υπομονάδα του. Στη συνέχεια ερευνούμε το κατά πόσο, οι προαναφερθείσες μέθοδοι και τεχνικές μπορούν να εφαρμοστούν σε ένα σύστημα ακουστικής ανίχνευσης καταστάσεων εκτάκτου ανάγκης. Χρησιμοποιώντας ηχογραφημένα δείγματα από μια ρεαλιστική βιβλιοθήκη όπου δημιουργήσαμε, καθώς και ειδικά προγράμματα επεξεργασίας σήματος. Αναλύουμε τρόπους μοντελοποίησης των δειγμάτων αυτών με τελικό σκοπό την αυτόματη επιβεβαίωση της ύπαρξης η όχι κατάστασης εκτάκτου ανάγκης.

Πρέπει να αναφερθεί ότι παρόμοιο του παραπάνω σύστηματος έχει υλοποιηθεί ως τμήμα του συστήματος PROMETHEUS ως project της ευρωπαϊκής ένωσης το οποίο όμως δίνει μια ποιο ολοκληρωμένη επιτήρηση ενός χώρου. Δηλαδή εκτός της ακουστικής επιτήρησης έχει την δυνατότητα και οπτικής επιτήρησης που περιλαμβάνει κάμερες, σενσόρες κ.τ.λ

Ο κύριος στόχος αυτής της εργασίας εδώ και 1 περίπου χρόνο όπου εκπονείται είναι η δημιουργία μιας πληρέστερης ποσοτικά αλλά συνάμα και ποιοτικά βιβλιοθήκης δειγμάτων μιας και είναι το σημαντικότερο τμήμα κάθε τέτοιου συστήματος καθώς και προτείνεται ένα σύστημα το οποίο εντοπίζει μη-τυπικές καταστάσεις σε περιβάλλον σταθμού μετρό.

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία με τίτλο " Ακουστική ανίχνευση καταστάσεων εκτάκτου ανάγκης " υποβλήθηκε στο Α.Τ.Ε.Ι. Μουσικής τεχνολογίας και ακουστικής στο Ρέθυμνο για τη εκπλήρωση των υποχρεώσεων όσον αφορά την πτυχιακή μου εργασία.

Εστιάζει σε αναδυόμενες εφαρμογές της τεχνολογίας αναγνώρισης γενικευμένου ακουστικού σήματος προτείνοντας καινοτόμες λύσεις. Η οργάνωση της όπως φαίνεται παρακάτω:

Στο *Κεφάλαιο 1* παρουσιάζεται μία γενική επισκόπηση της αυτόματης αναγνώρισης γενικευμένων ακουστικών γεγονότων. Επιπλέον συζητάμε τις εφαρμογές της τεχνολογίας αναγνώρισης ακουστικού σήματος και δίνουμε μία σύντομη περιγραφή του ιδανικού συστήματος (state of the art).

Στο *Κεφάλαιο 2* εισάγουμε τον αναγνώστη στο χώρο της επεξεργασίας ακουστικών σημάτων που δε περιλαμβάνουν ομιλία. Παρουσιάζονται οι σύγχρονες προσεγγίσεις όσον αφορά στις μεθοδολογίες εξαγωγής χαρακτηριστικών και αναγνώρισης προτύπων.

Στο *Κεφάλαιο 3* προτείνεται ένα καινοτόμο σύστημα αναγνώρισης ήχων ειδικά σχεδιασμένο για το χώρο των ηχητικών γεγονότων αστικού περιβάλλοντος και αναλύεται ο σχεδιασμός της αντίστοιχης βάσης δεδομένων. Δημιουργήθηκε μία βιβλιοθήκη δειγμάτων χωρισμένη σε δύο ομάδες ακουστικών παραμέτρων βασισμένη στην ιεραρχικά ανθρώπινη αντίληψη των ήχων ((α) μηχανικός, (β) μη-μηχανικός) που οδηγούν σε υψηλή ακρίβεια αναγνώρισης. Γίνεται περιγραφή ενός συστήματος αναγνώρισης καθώς και ο τρόπος λειτουργίας αυτού και εξάγονται τα αποτελέσματα του πειράματος.(πίνακες)

Στο *Κεφάλαιο 4* περιγράφεται ένα σύστημα το οποίο εντοπίζει μη-τυπικές καταστάσεις σε περιβάλλον σταθμού μετρό με στόχο να βοηθήσει το εξουσιοδοτημένο προσωπικό στην συνεχή επίβλεψη του χώρου.

Στο *Κεφάλαιο 5* προτείνεται ένα προσαρμοζόμενο σύστημα για ακουστική παρακολούθηση εν δυνάμει καταστροφικών καταστάσεων ικανό να λειτουργεί κάτω από διαφορετικά περιβάλλοντα. Δείχνουμε ότι το σύστημα επιτυγχάνει υψηλή απόδοση και μπορεί να προσαρμόζεται αυτόνομα σε ετερογενείς ακουστικές συνθήκες.

Στο *Κεφάλαιο 6* ερευνάται η χρήση της μεθόδου ανίχνευσης καινοτομίας για ακουστική επόπτευση κλειστών και ανοιχτών χώρων. Ηχογραφήθηκε μία βάση δεδομένων πραγματικού κόσμου και προτείνονται τρεις τεχνικές για την υλοποίηση αυτών.

Στο *Κεφάλαιο 7* παρουσιάζεται μία καινοτόμα μεθοδολογία για αναγνώριση γενικευμένου ακουστικού σήματος που οδηγεί σε υψηλή ακρίβεια αναγνώρισης. Αυτο επιτυγχάνεται με τα πλεονεκτήματα της χρονικής συγχώνευσης χαρακτηριστικών σε συνδυασμό με μία παραγωγική τεχνική κατηγοριοποίησης.

Τέλος, αναφέρουμε συμπεράσματα και την συνεισφορά της πτυχιακής.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Ο στόχος του ερευνητικού πεδίου της αναγνώρισης ακουστικού σήματος είναι η δημιουργία ενός συστήματος που χρησιμοποιεί μόνο το εισερχόμενο ακουστικό σήμα με στόχο την ανάλυση του περιβάλλοντα χώρου. Η ανάλυση του περιβάλλοντα χώρου περιλαμβάνει την καταμέτρηση των ηχητικών πηγών, τον διαχωρισμό τους και την αναγνώρισή τους. Η ιδιότητα αυτή έρχεται σε αντιστοιχία με το ανθρώπινο γνώρισμα το οποίο επιτρέπει τη πλήρη κατανόηση του περιβάλλοντος, απλώς και μόνο ακούγοντάς το (για παράδειγμα κάποιος στέκεται στο φανάρι και καταλαβαίνει ότι ένα αμάξι κορνάρει, ένας άνθρωπος μιλάει στο κινητό ενώ ένας σκύλος γαβγίζει). Μάλιστα το ανθρώπινο αυτί επιδεικνύει εξαιρετική ανεκτικότητα σε μια σειρά από "ανωμαλίες" που μπορεί να έχει υποστεί το ακουστικό σήμα, π.χ. θόρυβος, αντήχηση κ.τ.λ.

Η τεχνολογία της αυτόματης αναγνώρισης γενικευμένου ακουστικού σήματος (generalized sound recognition technology) αποτελεί μέρος της Υπολογιστικής Ακουστικής Ανάλυσης Σκηνής (Computational Auditory Scene Analysis - CASA), (Wang et al., 2006). Ο μακροπρόθεσμος στόχος αυτής της τεχνολογίας είναι η πλήρης κατανόηση και ερμηνεία ενός σκηνικού ενώ βασιζόμαστε μόνο στην ακουστική πληροφορία του χώρου. Βασίζεται στην εξαγωγή των ιδιαίτερων χαρακτηριστικών κάθε ηχητικής πηγής, τα οποία περιγράφουν τον τρόπο με τον οποίο κάθε πηγή κατανέμει την ενέργεια της πάνω στο φάσμα. Ο τρόπος αυτός είναι χαρακτηριστικός για κάθε πηγή και αποτελεί την "υπογραφή" του (audio signature). Στη συνέχεια γίνεται στατιστική σύγκριση αυτών των χαρακτηριστικών με εκείνα που εξάγονται από το σήμα αγνώστου ταυτότητος. Η τελική απόφαση παίρνεται με βάση το κριτήριο της μέγιστης πιθανότητας. Αυτός ο τρόπος αυτόματης αναγνώρισης ακουστικού σήματος ονομάζεται αναγνώριση ήχου με βάση το περιεχόμενο (content-based audio recognition) καθώς στηρίζεται μόνο στη πληροφορία του σήματος και σε καμία άλλη μετά-πληροφορία που ίσως να το συνοδεύει, όπως για παράδειγμα τα αρχεία τύπου ID3 (συμπληρωματικά αρχεία κειμένου που περιέχουν όνομα τραγουδιστή, δίσκου κ.α.) που προσφέρονται ως συμπλήρωμα των αρχείων μουσικής mp3.

Η ακουστική ταξινόμηση σημάτων αποτελεί έναν ερευνητικό τομέα που καλύπτει πολλά μικρότερα ερευνητικά πεδία. Η επεξεργασία σήματος, η εξαγωγή χαρακτηριστικών γνωρισμάτων, η αναγνώριση προτύπων και η ψυχοακουστική διαδραματίζουν έναν σημαντικό ρόλο. Στο σχεδιασμό οποιουδήποτε τέτοιου συστήματος, πρέπει αρχικά να αποφασιστεί ποιος συγκεκριμένος στόχος θα εκτελεστεί από το σύστημα, να επιλεγούν σχετικές ακουστικές παράμετροι αλλά και η κατάλληλη τεχνική κατηγοριοποίησης.

1.1. Εφαρμογές της Τεχνολογίας Αναγνώρισης Ήχων

Από τη σκοπιά της εφαρμοσμένης μηχανικής, ένας ήχος είναι μια δόνηση μέσω ενός ελαστικού μέσου (αέρας, στερεό, υγρό) που είναι ανιχνεύσιμη και ερμηνεύσιμη από βιολογικούς αισθητήρες ή έναν υπολογιστή. Ο λόγος για τον οποίο επεκτείνουμε τον ορισμό του ηχητικού γεγονότος πέρα από το όριο της ανθρώπινης ακοής είναι επειδή υπάρχουν βιολογικοί αισθητήρες (π.χ. στα έντομα) καθώς και μικρόφωνα που μπορούν να καταγράψουν τους ήχους που ξεπερνούν το φάσμα συχνότητας της ανθρώπινης ακοής. Η έξοδος αυτών των αισθητήρων μπορεί να γίνει είσοδος σε έναν αλγόριθμο αναγνώρισης προτύπων στα πλαίσια της αυτόματης αναγνώρισης γενικευμένου ακουστικού σήματος.

Σαν απόρροια της ευρείας γκάμας ηχητικών σημάτων, υπάρχει πληθώρα εφαρμογών που εμπεριέχουν την αναγνώρισης τους, οι οποίες εδώ αναφέρονται συνοπτικά:

- Εντοπισμός σήματος ομιλίας (Speech Activity Detection): Εδώ, ο πρωταρχικός σκοπός είναι η κατηγοριοποίηση του σήματος ως ομιλία ή μη-ομιλία. Επιτρέπει το μαρκάρισμα της αρχής και του τέλους ενός τμήματος ομιλίας.
- Η απόδοση ενός συστήματος αναγνώρισης ομιλίας/ομιλητή/συναισθήματος/γλώσσας ενισχύεται από αυτή τη διαδικασία, καθώς δύναται να λειτουργεί μόνο σε σήματα ομιλίας (ηχηρά και άηχα) και όχι σε άλλου είδους ηχητικά γεγονότα (Benyassine et al., 1997; Sohn et al., 1999, Cho et al., 2001).

Εφαρμογές σε σήματα *μουσικής*:

1. Αναγνώριση μουσικού είδους (π.χ. jazz, hip-hop, classical κτλ. - Tzanetakis et al., 2002; Gouyon et al., 2004)
2. Αναγνώριση μουσικού οργάνου (Eggink et al., 2004; Livshin et al.,2004; Peeters, 2003; Eggink et al., 2003; Liu et al., 2001)
3. Πλήρης αντιστοίχηση μουσικού σήματος σε νότες (music transcription- FitzGerald et al., 2002; Klapuri et al., 2006)
4. Αναγνώριση εκτελεστή μουσικού έργου (Widmer, 2006)
5. Συστηματοποιημένη ταξινόμηση και ανάκτηση μουσικού σήματος (music indexing and retrieval - Peeters et al., 2003; Wold et al., 1996; Slaney, 2002; Berenzweig et al., 2003).¨

Εφαρμογές προερχόμενες από την επεξεργασία *βιοακουστικού* σήματος όπως ανάπτυξη εξοπλισμού ικανού να εντοπίζει και να αναγνωρίζει έμβιους οργανισμούς,

επίβλεψη βιότοπου κ.α. (Arrigoni, 2003; Lee et al., 2006; Mitrovic et al., 2006; Gaston et al., 2004).

Ακουστική *μηχανών* (machine acoustics): Τα περισσότερα στερεά όπως το μέταλλο, το σκυρόδεμα, το κεραμικό κ.λπ., όταν υποβάλλονται σε πίεση εκπέμπουν ήχο λόγω της γρήγορης ενεργειακής απελευθέρωσης από μια εντοπισμένη πηγή μέσα στο υλικό. Η πηγή της ακουστικής εκπομπής μπορεί να δηλώνει εσωτερική παραμόρφωση, εξάρθρωση, σπάσιμο, και διάδοση ρωγμών. Διαφορές εφαρμογές προκύπτουν από την επεξεργασία αυτών των σημάτων, όπως μη-καταστροφικός έλεγχος μηχανημάτων, εντοπισμός δυσλειτουργιών κ.α. (Carolan et al., 1997; Diei et al., 1987; Dimla et al.,1997; Diniz et al., 1992; Iwata et al., 1977).

Αναγνώριση *περιβάλλοντος* και των ηχητικών γεγονότων μέσα σε αυτό (context awareness): ορίζεται ως η διαδικασία η οποία περιλαμβάνει την αυτόματη αναγνώριση του περιβάλλοντος γύρω από μια συσκευή αλλά και της δραστηριότητας του χρήστη. Επιτρέπει στις φορητές συσκευές να παρέχουν τις υπηρεσίες τους αποτελεσματικότερα καθώς μπορούν να *αυτοπροσαρμόζουν* το τρόπο λειτουργίας τους σύμφωνα με το περιβάλλοντα χώρο (Chu et al., 2006; Eronen et al., 2006; Clarkson et al., 1998). Επίσης, σε αυτή τη κατηγορία ανήκουν εφαρμογές όπως επίβλεψη χώρου, ενίσχυση της ανθρώπινης μνήμης κ.α. (Clavel et al., 2008; Vemuri et al., 2004).

1.2. Γενική επισκόπηση της διεθνούς βιβλιογραφίας

Είναι γεγονός ότι το ερευνητικό πεδίο της επεξεργασίας ακουστικών σημάτων και ειδικότερα των σημάτων εκείνων που δεν ανήκουν στη κατηγορία της ανθρώπινης ομιλίας (non-speech audio signal processing) έχει λάβει λιγότερη προσοχή απ' ότι το πεδίο της τεχνολογίας ομιλίας. Αυτό οφείλεται στον αδιαμφισβήτητο ρόλο που έχει η ομιλία ως τον πιο φυσικό και διαδεδομένο τρόπο επικοινωνίας μεταξύ των ανθρώπων. Επίσης χρησιμοποιείται και σε εφαρμογές οι οποίες εμπεριέχουν αλληλεπίδραση μεταξύ ανθρώπου και μηχανής (human-computer interaction) όπως οι φωνητικές πύλες. Παρόλα αυτά έχει σημειωθεί σημαντική πρόοδος σε διάφορες εφαρμογές της συγκεκριμένης τεχνολογίας, όσο υπάρχουν εφαρμογές οι οποίες είτε είναι νέες είτε δεν έχει βρεθεί μέχρι τώρα μία αποδεκτή λύση που να αντιμετωπίζει συνολικά το πρόβλημα.

Η πιο σημαντική προσπάθεια είναι η δημιουργία και καθιέρωση του πρωτοκόλλου *MPEG-7 Audio* (Motion Pictures Experts Group - Casey, 2001). Πρόκειται για προτυποποίηση μίας μεθοδολογίας γενικού σκοπού με στόχο την αξιόπιστη αναγνώριση ηχητικών σημάτων με αυτόματο τρόπο. Εμπεριέχει μία σειρά από ακουστικά χαρακτηριστικά ικανά να αντιπροσωπεύσουν μία μεγάλη γκάμα ήχων. Όσον αφορά το κομμάτι της αναγνώρισης προτύπων προτείνεται η χρήση κρυφών μοντέλων Markov (hidden Markov models). Κατά συνέπεια δημιουργείται ένα μοντέλο για κάθε

ηχητική κατηγορία χρησιμοποιώντας τις αντίστοιχες παραμέτρους. Η αναγνώριση ενός άγνωστου ήχου περιλαμβάνει της στατιστική σύγκριση των παραμέτρων του με εκείνες που έχουν ήδη εξαχθεί μέσω των κρυμμένων μοντέλων Markov. Το μοντέλο με τη μεγαλύτερη πιθανότητα αντιπροσωπεύει και την κατηγορία του άγνωστου σήματος. Αυτή η τεχνική προσέγγισης της συνάρτησης κατανομής είναι και η πιο διαδεδομένη καθώς μπορεί και μοντελοποιεί όχι μόνο τις στατικές ιδιότητες ενός ακουστικού σήματος αλλά και την μεταξύ τους αλληλουχία και σχέση. Άλλες εργασίες που αντιμετωπίζουν την αυτόματη αναγνώριση γενικευμένου ακουστικού σήματος είναι οι (Kim et al., 2004; Wold et al., 1996; Zhang et al., 1998; Umapathy et al., 2007).

Η συγκεκριμένη τεχνολογία έχει ακόμα να αντιμετωπίσει διαφόρων ειδών ανοιχτά ζητήματα. Ένα από τα κύρια προβλήματα είναι η δημιουργία μίας βάσης δεδομένων αναφοράς, έτσι ώστε μελέτες από διάφορους ερευνητές να μπορούν να συγκριθούν άμεσα όπως συμβαίνει σε άλλες περιοχές (π.χ. τεμαχιοποίηση ομιλίας όπου έχει καθιερωθεί η βάση TIMIT). Επιπλέον, η πλειονότητα της ερευνητικής δραστηριότητας έχει επικεντρωθεί σε βάσεις "καθαρές" από κάθε είδους παραμόρφωση, όπως π.χ. θόρυβος. Τέλος, στη βιβλιογραφία υπάρχει μικρός αριθμός εργασιών, οι οποίες πειραματίζονται πάνω σε βάσεις δεδομένων πραγματικού κόσμου (real-world).

Το επόμενο κεφάλαιο παρουσιάζει τις ακουστικές παραμέτρους αλλά και τις τεχνικές αναγνώρισης προτύπων που έχουν χρησιμοποιηθεί προς όφελος της αυτόματης ταξινόμησης ήχων.

Κεφάλαιο 2

Εξαγωγή Ακουστικών Χαρακτηριστικών και Αυτόματη Κατηγοριοποίηση

Σε αυτό το κεφάλαιο προσφέρεται μία γενική επισκόπηση των χαρακτηριστικών που χρησιμοποιούνται για να περιγράψουν ένα ακουστικό σήμα καθώς και των τεχνικών που οδηγούν στην αυτόματη κατηγοριοποίησή τους. Σε καμία περίπτωση η ανάλυση που ακολουθεί δεν μπορεί να θεωρηθεί διεξοδική αλλά δίνει μία συνοπτική εικόνα των περισσότερων τεχνικών που έχουν χρησιμοποιηθεί από την επιστημονική κοινότητα.

2.1. Ακουστικά Χαρακτηριστικά

Η φάση εξαγωγής των ακουστικών χαρακτηριστικών είναι το πρώτο βήμα της διαδικασίας ταξινόμησης και αποτελείται από τη μετατροπή του ακουστικού σήματος σε

μια σειρά διανυσμάτων μικρών διαστάσεων, όπου κάθε ένα συνοψίζει ένα τμήμα του σήματος. Ιδανικά, τα χαρακτηριστικά πρέπει να αντιπροσωπεύσουν όλες τις πληροφορίες που είναι σχετικές με την εφαρμογή ταξινόμησης και να απορρίπτουν τις άσχετες πληροφορίες. Η επιλογή των χαρακτηριστικών εξαρτάται από τη συγκεκριμένη εφαρμογή. Χαρακτηριστικά τα οποία περιλαμβάνονται σε μια εφαρμογή μπορούν να οδηγήσουν σε μείωση της απόδοσης αναγνώρισης εάν χρησιμοποιηθούν σε άλλη εφαρμογή. Σε μουσικές εφαρμογές, οι πληροφορίες που έχουν να κάνουν με το ρυθμό καθώς επίσης και οι δυναμικές ιδιότητες του σήματος είναι πολύ σημαντικές, ενώ σε άλλες δε βρίσκουν χρησιμότητα. Η διαδικασία εξαγωγής χαρακτηριστικών αποτελείται από έναν αριθμό διαδοχικών επιπέδων. Στη συνέχεια, θα αναφέρουμε εν συντομία αυτά τα επίπεδα που είναι σε μεγάλο βαθμό κοινά όσον αφορά τις διάφορες κατηγορίες εφαρμογών της αυτόματης αναγνώρισης ήχων.

Φάση Α: Ηχογράφηση και προεπεξεργασία σημάτων

Το φυσικό ηχητικό σήμα είναι αναλογικό και επομένως πρέπει να μετατραπεί σε ψηφιακό προκειμένου να υποβληθεί σε επεξεργασία από ένα υπολογιστικό σύστημα. Αυτό εκτελείται σε δύο βήματα: δειγματοληψία και κβαντοποίηση του πλάτους του. Σύμφωνα με το θεώρημα δειγματοληψίας των Nyquist-Shannon, προκειμένου να συντηρηθεί η πληροφορία του αρχικού αναλογικού σήματος, η δειγματοληψία πρέπει να γίνει ομοιόμορφα και με ρυθμό που να είναι τουλάχιστον δύο φορές υψηλότερος από την υψηλότερη συχνότητα στο αναλογικό σήμα. Στη συνέχεια, το σήμα κβαντοποιείται ως προς το πλάτος και τον χρόνο, μετατρέπεται δηλαδή σε έναν ψηφιακό κώδικα (όπως η κωδικοποίηση PCM, που χρησιμοποιείται ευρύτατα). Η επιλογή της ανάλυσης κβαντοποίησης εξαρτάται από την εκάστοτε εφαρμογή. Σε εφαρμογές αναγνώρισης ομιλίας ή/και περιβαλλοντικών ήχων, η ανάλυση 16 bit/δείγμα (με τη κωδικοποίηση PCM) παρέχει μια αποδεκτή ποιότητα. Ο θόρυβος κβαντοποίησης δε γίνεται αντιληπτός από το μέσο ακροατή και δεν επιφέρει αξιοσημείωτη μείωση της απόδοσης της αυτόματης διαδικασίας ταξινόμησης.

Στη συνέχεια αφαιρείται η μέση τιμή του ψηφιοποιημένου σήματος (DC offset), με στόχο την αφαίρεση της σταθερής συνιστώσας που ίσως έχει εμφανιστεί κατά τη διάρκεια της ηχογράφησης. Έπειτα εφαρμόζεται προσαρμοστικός έλεγχος κέρδους (adaptive gain control) που κρατά το εύρος του σήματος μέσα σε προκαθορισμένα πλαίσια. Η φάση προεπεξεργασίας μπορεί να περιλαμβάνει φιλτράρισμα των επιθυμητών ζωνών συχνοτήτων, καθώς και αποθορυβοποίηση.

Για την παραπάνω διαδικασία χρησιμοποιήθηκε το πρόγραμμα Sound Forge της Sony ώστε να γίνει ψηφιοποίηση και επεξεργασία σήματος των δειγμάτων όπου δημιουργήθηκε η βιβλιοθήκη της εργασίας. Ηχογραφήθηκαν τμήματα από ταινίες από την τηλεόραση καθώς και από το διαδίκτυο.

Φάση Β: Χωρισμός και ανάλυση μικρών χρονικών τμημάτων (Short-time analysis)

Ο μετασχηματισμός Fourier υποθέτει ότι οποιοσδήποτε ήχος μπορεί να δημιουργηθεί με κατάλληλο άθροισμα σημάτων ημιτόνου με διαφορετικές παραμέτρους. Ουσιαστικά ο συνεχής μετασχηματισμός Fourier αποσυνθέτει ένα δεδομένο ηχητικό σήμα σε έναν άπειρο αριθμό ημιτονοειδών συναρτήσεων που έχει συγκεκριμένες τιμές συχνοτήτων, πλατών και φάσεων.

Η κατηγοριοποίηση ήχων είναι βασισμένη στο γεγονός ότι κάθε ηχητική πηγή έχει έναν μοναδικό τρόπο να κατανείμει την ενέργειά της στις διάφορες περιοχές συχνοτήτων. Το σήμα ήχου συνήθως τεμαχιοποιείται σε μικρότερα κομμάτια (πλαίσια) και τα δείγματα κάθε κομματιού υποβάλλονται σε περαιτέρω επεξεργασία ανεξάρτητη των υπολοίπων πλαισίων. Υποθέτουμε ότι τα δεδομένα έχουν στάσιμες ιδιότητες μέσα στο ίδιο πλαίσιο. Αυτή η υπόθεση υποστηρίζεται από το ότι οι ηχητικές πηγές (π.χ. τα όργανα άρθρωσης, ο μηχανισμός παραγωγής ήχων των ζώων, των εντόμων, των μηχανών, των μουσικών οργάνων, κ.λπ.) χαρακτηρίζονται από αδράνεια, η οποία απαγορεύει τις στιγμιαίες αλλαγές στο περιεχόμενο συχνότητας του ήχου.

Ο βραχύχρονος μετασχηματισμός Fourier (short-time Fourier transform) αποκαλύπτει το περιεχόμενο συχνότητας όλων των πλαισίων με διαδοχικό τρόπο και είναι η αφετηρία των περισσότερων προσεγγίσεων εξαγωγής χαρακτηριστικών που είναι βασισμένες στο φασματικό περιεχόμενο. Το γεγονός ότι κάθε πλαίσιο εξάγεται από το αρχικό σήμα, εισάγει μια ασυνέχεια στην αρχή και το τέλος κάθε τμήματος. Αυτό το φαινόμενο αντιμετωπίζεται με την εφαρμογή ενός χρονικού παραθύρου πάνω στα «κομμένα» τμήματα το οποίο σταδιακά μειώνει το πλάτος πάνω στα όρια προς το μηδέν.

Η επίδραση της μείωσης του πλάτους των ορίων αντισταθμίζεται με την εφαρμογή ενός βαθμού επικάλυψης μεταξύ των παραθύρων. Αν και έχει διεξαχθεί αρκετή έρευνα πάνω στην ιδανική μορφή του παραθύρου, καμία σημαντική διαφορά δεν έχει αναφερθεί από την σκοπιά των αποτελεσμάτων ταξινόμησης. Συμπερασματικά μπορούμε να πούμε ότι ο βραχυπρόθεσμος μετασχηματισμός Fourier είναι το εργαλείο ανάλυσης χρόνου-συχνότητας που χρησιμοποιείται ευρύτατα στην επεξεργασία ακουστικού σήματος. Παρόλα αυτά στη παρούσα πτυχιακη ερευνάται και η αποτελεσματικότητα του διακριτού μετασχηματισμού *wavelet*, ο οποίος αν και προσφέρει υψηλή απόδοση σε άλλα ερευνητικά πεδία (όπως επεξεργασία εικόνας - Skodras et al., 2001), δεν έχει μελετηθεί διεξοδικά η χρησιμότητά του στο πεδίο της αυτόματης αναγνώρισης ήχων.

Φάση Γ: Υπολογισμός χαρακτηριστικών στο πεδίο της συχνότητας

Η περιοχή του φάσματος στην οποία πραγματοποιείται το μεγαλύτερο μέρος της ακουστικής δραστηριότητας είναι και η περιοχή ενδιαφέροντος. Για τη κάλυψη

αυτής της περιοχής, εφαρμόζεται μια τράπεζα-φίλτρων (filter-bank) σε ένα από τα εξής επίπεδα: α) στο φάσμα πλάτους, β) στο φάσμα ενέργειας είτε γ) στο λογαριθμικά συμπιεσμένο φάσμα ενέργειας. Τα κέντρα των φίλτρων (γραμμικών ή μη) είναι διακριτά μεταξύ τους, και χρησιμεύουν ως τα σημεία ορίου για τα αντίστοιχα γειτονικά φίλτρα. Αυτό οδηγεί στην εξομάλυνση των λεπτομερειών του φάσματος και μειώνει πολύ τις διαστάσεις του διανύσματος του φάσματος, διατηρώντας τις ουσιαστικές πληροφορίες και τα χαρακτηριστικά του συγκεκριμένου ακουστικού γεγονότος. Παραδείγματος χάριν για σήμα που έχει δειγματοληψία σε συχνότητα 8000 Hz, ο διακριτός μετασχηματισμός Fourier (discrete Fourier transform - DFT) με 512 δείγματα οδηγεί σε διάνυσμα φάσματος 256 διαστάσεων, το οποίο μετά την εφαρμογή ενός γραμμικού ή ενός Mel φίλτρου μειώνεται σε 40 επιτυγχάνοντας έτσι δραστική μείωση των διαστάσεων του φάσματος. Οι συντελεστές του μετασχηματισμού Fourier μπορούν άμεσα να χρησιμοποιηθούν ως χαρακτηριστικά αλλά συνήθως αποτελούν τη βάση για τη δημιουργία πιο σύνθετων χαρακτηριστικών. Τα πιο υποσχόμενα είναι οι cepstral συντελεστές συχνότητας του πρωτοκόλλου MPEG-7, τα MFCCs και οι παραλλαγές τους.

Το πρωτόκολλο MPEG-7 (Kim et al., 2005) περιλαμβάνει χαμηλού επιπέδου ακουστικούς περιγραφείς που έχουν σπουδαία σημασία για τη περιγραφή του ήχου. Υπάρχουν 17 χρονικοί και φασματικοί περιγραφείς που μπορούν να χρησιμοποιηθούν σε ποικίλες εφαρμογές. Οι σημαντικότερες ομάδες είναι οι εξής: φασματικές, timbral (χαρακτηρίζουν τη χροιά), χρονικές και timbral-φασματικές.

Οι τέσσερις βασικοί φασματικοί ακουστικοί περιγραφείς είναι όλοι βασισμένοι στη διακριτή ανάλυση σήματος που προσφέρει ο DFT. Αυτοί οι τέσσερις περιγραφείς είναι:

- *Audio Spectrum Envelope*: είναι ένα λογαριθμικά συμπιεσμένο διάνυσμα του φάσματος συχνοτήτων, οι μπάντες του οποίου ορίζονται από ένα διαιρέτη δύναμης του 2 ή ένα πολλαπλάσιο μιας οκτάβας. Αυτό το χαρακτηριστικό είναι γενικού σκοπού και περιγράφει την εξέλιξη του φάσματος ενέργειας ενός ακουστικού σήματος για κάθε πλαίσιο.
- *Audio Spectrum Centroid*: πρόκειται για τον υπολογισμό του κέντρου βάρους της ενέργειας των λογαριθμικά τοποθετημένων περιοχών του φάσματος. Επομένως είναι μια ένδειξη για το εάν στο φασματικό περιεχόμενο του σήματος κυριαρχούν οι υψηλές ή οι χαμηλές συχνότητες.
- *Audio Spectrum Spread*: επιστρέφει τη διασπορά (variance) της ενέργειας των λογαριθμικά τοποθετημένων περιοχών του φάσματος γύρω από το audio spectrum centroid. Αυτός το χαρακτηριστικό δείχνει το κατά πόσο η ενέργεια είναι συγκεντρωμένη ή εξαπλωμένη έξω από το φάσμα και επομένως μπορεί να βοηθήσει στη διάκριση μεταξύ τονικών ήχων και θορύβων.

- *Audio Spectrum Flatness*: είναι ο λόγος μεταξύ του γεωμετρικού και του αριθμητικού μέσου όρου του φάσματος του σήματος. Αυτό το χαρακτηριστικό εκφράζει τη διαφορά μεταξύ του θορύβου και των τονικών σημάτων δεδομένου ότι τα αρμονικά σήματα δίνουν μια τιμή κοντά στο 0, ενώ τα σήματα θορύβου δίνουν κοντά στο 1.

Το χαρακτηριστικό *Audio Fundamental Frequency* είναι η θεμελιώδης συχνότητα ενός ακουστικού σήματος ενώ το *Audio Harmonicity* αντιπροσωπεύει τη φασματική δομή ενός σήματος από την σκοπιά της αρμονικότητας. Επιτρέπει τη διάκριση ανάμεσα σε ήχους με ευδιάκριτο αρμονικό φάσμα (π.χ., μουσικά σήματα ή λεκτικά φωνήεντα), και ήχους με μη-αρμονικό φάσμα (π.χ. θόρυβος).

Τα χαρακτηριστικά που ανήκουν στο πεδίο του χρόνου εστιάζουν στη χροιά (η χροιά περιλαμβάνει όλα τα στοιχεία ενός ήχου π.χ. η ενεργειακή κατανομή στους αρμονικούς ήχους που το διαφοροποιούν από άλλους ήχους με ίδια θεμελιώδη συχνότητα, της ηχηρότητας και της διάρκειας). Το χαρακτηριστικό *Log Attack Time* είναι ο λογάριθμος του χρόνου που απαιτείται για το σήμα για να μεταβεί από τη σιωπή στο σταθερό μέρος του. Αυτό το χαρακτηριστικό γνώρισμα εκφράζει τη διαφορά μεταξύ ενός ξαφνικού και ομαλού ήχου και θεωρείται σημαντικό για τη διάκριση μεταξύ απομονωμένων κρουστικών ήχων. Το *Temporal Centroid* εντοπίζει εγκαίρως το σημείο στο οποίο συγκεντρώνεται η ενέργεια ενός σήματος. Οι επόμενοι πέντε περιγραφείς είναι φασματικά χαρακτηριστικά σε ένα γραμμικό πεδίο συχνοτήτων και συσχετίζονται με τη χροιά. Το *Spectral Centroid* είναι ο σταθμισμένος, σύμφωνα με την ενέργεια μέσος όρος της συχνότητας των μερών (bins) στο γραμμικό φάσμα ενέργειας. Το *Harmonic Spectral Centroid* είναι ο σταθμισμένος σύμφωνα με το πλάτος μέσος όρος των αρμονικών κορυφών (peaks) του φάσματος. Οι αρμονικές κορυφές αντιστοιχούν στις συχνότητες που είναι ένα ακέραια πολλαπλάσια της θεμελιώδους συχνότητας. Το χαρακτηριστικό *Harmonic Spectral Deviation* δείχνει τη φασματική απόκλιση των αρμονικών μερών από το φασματικό φάκελο (envelope) του σήματος. Το *Harmonic Spectral Spread* περιγράφει την σταθμισμένη σύμφωνα με το πλάτος απόκλιση των αρμονικών κορυφών του φάσματος από το Harmonic Spectral Centroid. Τέλος, το χαρακτηριστικό *Harmonic Spectral Variation* είναι η κανονικοποιημένη συσχέτιση μεταξύ του πλάτους των αρμονικών κορυφών ανάμεσα σε δύο διαδοχικά πλαίσια.

Mel-Frequency Cepstral Coefficients (MFCC): Αν και τα MFCC (Slaney, 1998) προέρχονται από το πεδίο της αυτόματης αναγνώρισης ομιλίας, χρησιμοποιούνται ευρέως σε πολλές εφαρμογές ταξινόμησης ακουστικών σημάτων. Τα MFCC αποτελούν μία 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.

Αρχικά εφαρμόζεται προέμφαση στο σήμα με το φίλτρο

$H(z) = 1 - az^{-1}$, όπου $a \in [0.95, 0.97]$,

και στη συνέχεια κόβεται σε πλαίσια διάρκειας 20-40 ms, χρησιμοποιώντας τη τεχνική Hamming. Στη συνέχεια, κάθε πλαίσιο υποβάλλεται στον DFT και περνούν από μία ομάδα τριγωνικών ζωνοδιαβατών φίλτρων. Στη συνέχεια, ένας μικρός αριθμός συνιστωσών υπολογίζεται με την εφαρμογή του DCT πάνω στις λογαριθμικά συμπιεσμένες εξόδους των φίλτρων.

Φάση Δ: Μετασχηματισμός των ακουστικών παραμέτρων

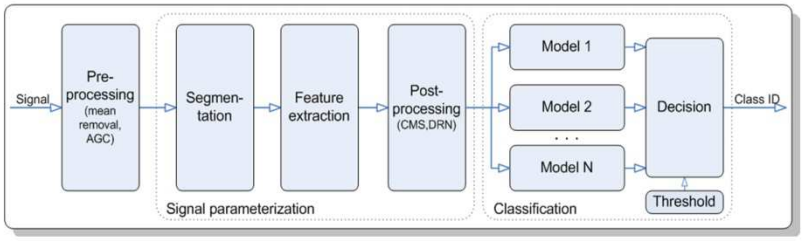
Η προβολή του φάσματος πριν ή μετά από την εφαρμογή κάποιας τράπεζας φίλτρων είναι μια κοινή στρατηγική προκειμένου να προβληθεί το αρχικό διάνυσμα μεγάλων διαστάσεων σε ένα συμπαγές διάνυσμα χαμηλότερης διάστασης με στόχο την αποδοτικότερη ταξινόμηση. Το μέγεθος του διανύσματος των χαρακτηριστικών δεν πρέπει να είναι μεγάλο επειδή αυτό οδηγεί στην αραιή αντιπροσώπευση των τμημάτων (clusters) στο χώρο υψηλών διαστάσεων (που απαιτεί μεγάλο όγκο δεδομένων για εκπαίδευση) αλλά ούτε και πολύ χαμηλό που ίσως οδηγήσει στο να απορριφθούν σημαντικές πληροφορίες.

Γενικά, το προκύπτον διάνυσμα χαρακτηριστικών χρησιμοποιεί πληροφορίες από όλες τις ζώνες φάσματος και οδηγείται σε ένα τελικό στάδιο αποσυσχέτισης (decorrelation) που επιτρέπει την αποδοτικότερη μοντελοποίηση με τη χρήση για παράδειγμα κρυμμένων μοντέλων Markov Γκαουσσιανών συναρτήσεων με διαγώνιου τύπου πίνακα συνδιασποράς. Τα διαγώνιου τύπου GMM απαιτούν τον υπολογισμό λιγότερων ελεύθερων μεταβλητών, και επομένως, εκπαιδεύονται καλύτερα για να αντιπροσωπεύσουν τις επιθυμητές κατηγορίες. Για αυτόν τον λόγο, διάφορες στρατηγικές αποσυσχέτισης έχουν προταθεί στη βιβλιογραφία. Οι δημοφιλέστερες είναι DCT, linear discriminant analysis, principal component analysis, factor analysis, subspace analysis, independent component analysis και singular value decomposition.

2.2. Μεθοδολογίες Αναγνώρισης Προτύπων

Η ταξινόμηση προτύπων στηρίζεται στη συνεισφορά του Thomas Bayes ο οποίος έθεσε τις δύο ακόλουθες υποθέσεις: α) το πρόβλημα λήψης μίας απόφασης μπορεί να εκφραστεί με πιθανολογικούς όρους και β) όλες οι σχετικές πιθανότητες είναι γνωστές ή μπορούν να εξαχθούν από τα δεδομένα παρατήρησης και/ή με τη χρήση εκ των προτέρων γνώσης σχετικά με το πρόβλημα. Αυτές οι δύο υποθέσεις είναι κοινές στις σύγχρονες εφαρμογές στατιστικής ταξινόμησης. Πλέον η θεωρία του Bayes αποτελεί τον ακρογωνιαίο λίθο ενός μεγάλου αριθμού προσεγγίσεων ταξινόμησης.

Δύο σημαντικές κατηγορίες ταξινομητών μπορούν να διακριθούν: οι διαχωριστικοί (discriminative) και οι μη-διαχωριστικοί (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 των κατηγοριών.



Σχήμα 2.1: Διάγραμμα αναγνώρισης N διαφορετικών κατηγοριών ήχων.

Οι διαχωριστικοί ταξινομητές περιλαμβάνουν: linear discriminant analysis (Fisher, 1936), Polynomial ταξινομητή (Specht, 1967), νευρικά δίκτυα χρονικής καθυστέρησης (Lang et al., 1988), ανατροφοδοτούμενα (recursive) νευρικά (Jordan, 1986; Elman, 1990), multilayer perceptron (Rosenblatt, 1958), και διανυσματικές μηχανές υποστήριξης (Vapnik, 1995).

Οι μη-διαχωριστικές προσεγγίσεις δεν στοχεύουν άμεσα στην ελαχιστοποίηση του λάθους ταξινόμησης. Μια σημαντική ομάδα μη-διαχωριστικών προσεγγίσεων καλείται παραγωγική (generative). Όπως το όνομά τους προτείνει, οι παραγωγικοί ταξινομητές προσπαθούν να χτίσουν την κατανομή στηριζόμενοι απόλυτα στα δεδομένα εκπαίδευσης. Η ομάδα των παραγωγικών προσεγγίσεων περιλαμβάνει: πιθανολογικό νευρωτικό δίκτυο (probabilistic neural network - PNN), το οποίο συνδυάζει τους Parzen εκτιμητές πυκνότητας πιθανότητας με τη στρατηγική Bayes για τους κανόνες απόφασης (Specht, 1988), γραμμικό συνδυασμό Γκαουσσιανών συναρτήσεων (Gaussian mixture models - GMM - Hansen, 1982) και τα κρυμμένα μοντέλα Markov (hidden Markov models - HMM - Baum et al., 1966) που αναπτύχθηκαν από τον Baum. Τα HMM έχουν τη δυνατότητα να μοντελοποιήσουν τη χρονική συμπεριφορά μιας ακολουθίας γεγονότων. Το GMM (ως HMM μίας κατάστασης) δεν είναι ευαίσθητο στη χρονική ακολουθία των εισόδων. Επίσης, το πρώτο PNN που προτάθηκε από τον Specht χειρίζεται τις εισόδους ανεξάρτητα τη μία από την άλλη και έτσι δε μπορεί να μοντελοποιήσει χρονικούς ή χωρικούς συσχετισμούς. Το χαρακτηριστικό που διαφοροποιεί τις παραγωγικές προσεγγίσεις είναι ότι επεξεργάζονται τα δείγματα κάθε κατηγορίας ανεξάρτητα από αυτά των άλλων κατηγοριών. Κατά κανόνα, όταν συγκρίνονται με τους διαχωριστικούς αναγνωριστές, οι παραγωγικοί ταξινομητές περιλαμβάνουν περισσότερες παραμέτρους προς υπολογισμό. Αυτό οφείλεται στο ότι προσπαθούν να αποτυπώσουν τις εναλλαγές των δεδομένων που ανήκουν στην ίδια κατηγορία, κάτι το οποίο δεν είναι απολύτως απαραίτητο. Εντούτοις, σε πραγματικές

εφαρμογές, παραδείγματος χάριν στην περίπτωση της τεχνολογίας αναγνώρισης ομιλητών, όπου ο σκοπός είναι η προστασία του μοντέλου ομιλητή από προσπάθειες απάτης, η ιδιότητα των παραγωγικών προσεγγίσεων να δημιουργούν ένα περιεκτικό μοντέλο ομιλίας του χρήστη βρίσκει πολύ μεγάλη χρησιμότητα.

Εκτός από τους παραγωγικούς ταξινομητές, η ομάδα των μη-διαχωριστικών ταξινομητών περιλαμβάνει και ταξινομητές που δεν μπορούν να χαρακτηριστούν παραγωγικοί, επειδή δεν δημιουργούν τις συναρτήσεις κατανομών των δεδομένων. Στην πραγματικότητα, δεν είναι εκτιμητές πιθανότητας, και επομένως, δεν ανήκουν στις παραγωγικές προσεγγίσεις. Παραδείγματα είναι οι Κ-κοντινότεροι γείτονες (K-NN - Cover et al., 1967) και η learning vector quantization (LVQ - Kohonen, 1986), οι οποίοι προσπαθούν να διαμορφώσουν την δομή των δεδομένων μέσω της ομαδοποίησης των γνωστών δειγμάτων και εκπροσώπησης της κάθε ομάδας από έναν μόνο αντιπρόσωπο. Στη φάση της αναγνώρισης αυτοί οι ταξινομητές εντοπίζουν τον κοντινότερο στους εκπροσώπους, που έχουν υπολογιστεί από τα δεδομένα εκπαίδευσης σύμφωνα με κάποιο μετρικό απόστασης. Αυτές οι μέθοδοι έχουν περιορισμένες ικανότητες και δεν χρησιμοποιούνται στα σύγχρονα συστήματα αναγνώρισης ήχων.

Οι διαχωριστικές, όπως και 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.

Οι υβριδικές μέθοδοι αποτελούνται από (α) τους υβριδικούς ταξινομητές, που συνδυάζουν το παραγωγικό μοντέλο με έναν διαχωριστικό ταξινομητή, και (β) διαχωριστικά εκπαιδευμένοι παραγωγικοί ταξινομητές, οι παράμετροι των οποίων ρυθμίζονται με τη βελτιστοποίηση μιας διαχωριστικής αντικειμενικής συνάρτησης (discriminative objective function). Οι περιπτώσεις που συνδυάζουν τις δύο κύριες κατηγορίες είναι:

• Radial basis function (RBF - Powell, 1987), η οποία ενώνει το παραγωγικό GMM (στην πραγματικότητα Γκαουσσιανές συναρτήσεις μίας διάστασης) με το διαχωριστικό MLP

• GMM-LR/SVM (Bengio et al., 2001), το οποίο χτίζει το παραγωγικό πρότυπο GMM στα πλαίσια ενός διαχωριστικού SVM

• HMM/MLP (Bourlard et al., 1994), που συνδυάζει τις χρονικές ικανότητες του HMM με τις διαχωριστικές του MLP. Δύο άλλες προσεγγίσεις, δηλαδή το HMM/RNN (Neto et al., 1995) και το εισόδου-εξόδου HMM (Bengio et al., 1996), εκμεταλλεύονται ίδια πλεονεκτήματα με αυτά της προσέγγισης HMM/MLP.

• Το διαχωριστικά εκπαιδευμένο HMM (Setlur et al., 1996) αποτελεί μία προσέγγιση που ανήκει στη δεύτερη κατηγορία των συνδυασμένων μεθόδων, δηλαδή οι διαχωριστκοί εκπαιδευμένοι παραγωγικοί ταξινομητές. Στο διαχωριστικά εκπαιδευμένο

HMM το κριτήριο μέγιστης πιθανότητας του παραδοσιακού HMM αντικαθίσταται από το μέγιστο αμοιβαίο κριτήριο πληροφοριών.

• PNN/RNN υβρίδια. Πρόσφατα προτάθηκαν δύο νέες προσεγγίσεις ταξινόμησης, το τοπικά επαναλαμβανόμενο PNN (Locally Recursive PNN - Ganchev et al., 2003) και η γενικευμένη LR PNN (GLR PNN - Ganchev et al., 2004). Αυτά τα νευρωνικά δίκτυα στηρίζονται στην υβριδική αρχιτεκτονική PNN-RNN, και προέρχονται από το αρχικό PNN με την ενσωμάτωση ενός πρόσθετου κρυμμένου επιπέδου, που καλείται επαναλαμβανόμενο επίπεδο. Το επαναλαμβανόμενο στρώμα, που είναι τοποθετημένο μεταξύ του στρώματος αθροίσματος και του ανταγωνιστικού στρώματος εξόδου της αρχικής δομής PNN, αποτελείται από τους νευρώνες που κατέχουν ανατροφοδοτήσεις. Η κύρια διαφορά μεταξύ του LR PNN και του GLR PNN είναι στο σύνδεσμο του επαναλαμβανόμενου στρώματος που ενσωματώνεται στο PNN.

Οι περισσότερες από τις προαναφερθείσες μεθόδους ταξινόμησης αρχικά υιοθετήθηκαν σε εφαρμογές σχετικές με επεξεργασία ομιλίας λόγω της μακροχρόνιας ιστορίας της, ιδίως όταν συγκρίνεται με τις άλλες εφαρμογές επεξεργασίας ήχων (non-speech audio processing). Εντούτοις, στην τελευταία δεκαετία μερικές νέες εφαρμογές, οι οποίες απαιτούν την αναγνώριση γενικευμένου ακουστικού σήματος, αρχίζουν να προσελκύουν το ενδιαφέρον πολλών ερευνητών.

Κεφάλαιο 3

Αυτόματη Αναγνώριση Ηχητικών Γεγονότων Αστικού Περιβάλλοντος

Στο κεφάλαιο αυτό παρουσιάζεται μία νέα εφαρμογή της τεχνολογίας αναγνώρισης ήχων . Ένας από τους άμεσους στόχους του συστήματος είναι η παρακολούθηση της κίνησης με σκοπό την αναπροσαρμογή του συστήματος οδικής κυκλοφορίας έτσι ώστε να μειωθεί ο χρόνος που απαιτείται για την κάλυψη μιας απόστασης. Επίσης, η μεθοδολογία μπορεί να ενσωματωθεί σε διάφορων τύπων συσκευές, των οποίων η λειτουργία μπορεί να αλλάζει συναρτήσει της απόκρισης του συστήματος το οποίο βασίζεται στην «ακουστική αίσθησή του» (auditory sense). Προτείνεται μία νέα μέθοδος για την αυτόματη αναγνώριση των αστικών ακουστικών σκηνών (στη βιβλιογραφία αναφέρονται ως soundscenes). Το σύστημά αυτό αποτελεί ένα ιεραρχικό σχήμα ταξινόμησης ενώ συγκρίνονται οι αποδόσεις δύο ομάδων ακουστικών χαρακτηριστικών. Ένας νέος αλγόριθμος μέτα-επεξεργασίας χρησιμοποιείται για να βελτιώσει την ικανότητα διάκρισης των MPEG-7 γνωρισμάτων και αποδεικνύεται ότι παρέχει βελτιωμένα αποτελέσματα. Η προσέγγισή μας εξετάζεται

ακολουθώντας μια προσεκτικά σχεδιασμένη διαδικασία και αποδεικνύεται ότι τα χαρακτηριστικά του πρωτοκόλλου MPEG-7 προσφέρουν υψηλότερα ποσοστά αναγνώρισης από τα MFCCs.

3.1. Εισαγωγή

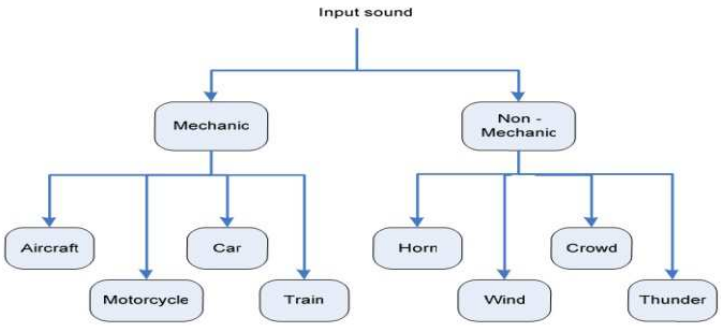
Καθημερινά ερχόμαστε σε επαφή με πολλούς διαφορετικούς τύπους αστικών ήχων (αυτοκίνητο, μοτοσικλέτα, πλήθος κ.λπ.). Οι άνθρωποι μπορούν αποτελεσματικά να τους διαφοροποιήσουν με χαρακτηριστική ευκολία χρησιμοποιώντας μόνο την ακουστική τους αίσθηση. Σκεφτείτε για παράδειγμα την κατάσταση κατά την οποία κάποιος στέκεται δίπλα σε έναν φωτεινό σηματοδότη. Χρησιμοποιώντας μόνο τους εισερχόμενους ήχους, είναι σε θέση να καταλάβει ότι ένα αυτοκίνητο περνά, ένας σκύλος γαυγίζει ενώ ακούγεται μία κόρνα. Ο γενικός στόχος της εργασίας μας είναι η ανάπτυξη ενός συστήματος που έχει τη δυνατότητα αυτόματα «να καταλαβαίνει» το περιβάλλοντα χώρο του, ενώ επεξεργάζεται μόνο τους ήχους που «ακούει» με τη χρήση ενός μικροφώνου. Ο τομέας της υπολογιστικής ακρόασης (computer audition) αντιμετωπίζει μια αυξανόμενη ζήτηση σε πολυάριθμες εφαρμογές (ρομποτική συνειδητοποίηση (robot awareness), περιβαλλοντικός έλεγχος, επισημείωση δεδομένων κ.λπ.) και έχει πλέον εξελιχθεί σε έναν ερευνητικό τομέα μεγάλου ενδιαφέροντος.

Κατά τη διάρκεια των προηγούμενων δεκαετιών, έχει διεξαχθεί αρκετή έρευνα στο χώρο της ακουστικής ταξινόμησης με βάση το περιεχόμενο. Στην εργασία (Eronen et al, 2006) ερευνάται ένα ακουστικό σύστημα αναγνώρισης που χρησιμοποιείται για την ταξινόμηση 24 αστικών σκηνών. Χρησιμοποιούν διάφορα απλοϊκά χαρακτηριστικά χαμηλών διαστάσεων καθώς επίσης και τυποποιημένους φασματικούς περιγραφείς σε συνδυασμό με μία HMM μεθοδολογία ταξινόμησης επιτυγχάνοντας ποσοστό αναγνώρισης 58%. Ένα σύστημα για την ταξινόμηση σε επίπεδο πλαισίων των θορύβων που ανήκουν σε πέντε κατηγορίες παρουσιάζεται στην εργασία (El-Maleh et al., 1999). Χρησιμοποιήθηκαν οι φασματικές συχνότητες γραμμών (linear spectral frequencies) μαζί με έναν ταξινομητή δέντρου απόφασης και οδήγησαν σε ακρίβεια ταξινόμησης 88.1%. Μια μέθοδος βασισμένη σε τρεις περιγραφείς χαμηλού επιπέδου του MPEG-7 (spectral centroid, spectrum spread και spectrum flatness) παρουσιάζεται στην (Wang et al, 2006). Για την αναγνώριση προτύπων προτείνεται μία μίξη των διανυσματικών μηχανών υποστήριξης και του k-κοντινότερου γείτονα και έτσι ορίζεται η κατηγορία ενός άγνωστου ήχου. Οι κατηγορίες περιλαμβάνουν ήχους οι οποίοι συχνά υπάρχουν σε ένα τυπικό περιβάλλον σπιτιού. Τέλος, ένα σύστημα βασισμένο σε μία μεθοδολογία "ανακαλύψης" παραμέτρων με στόχο την ανάλυση αστικών ηχητικών σκηνών περιγράφεται στην εργασία (Defrévill et al, 2006). Έξι κύριες κατηγορίες ήχων οργανώνονται ενώ εξετάζονται οι αποδόσεις των GMM και k-κοντινότερων γειτόνων.

Εμείς υιοθετήσαμε τα MFCCs και τις παραμέτρους του πρωτοκόλλου MPEG-7. Ο στόχος μας είναι να προσδιορίσουμε ποια ομάδα παραμέτρων περιέχει τις πιο χρήσιμες πληροφορίες όσον αφορά την αναγνώριση των αστικών ακουστικών σκηνών. Το υπόλοιπο του κεφαλαίου οργανώνεται ως εξής. Τα επόμενα τρία τμήματα περιγράφουν τη γενική αρχιτεκτονική του συστήματός μας, τη μεθοδολογία εξαγωγής ακουστικών γνωρισμάτων καθώς και τη διαδικασία αναγνώρισης. Η λεπτομερής ανάλυση της μεθόδου αξιολόγησης καθώς επίσης και των αποτελεσμάτων δίνεται στο τελευταίο μέρος του κεφαλαίου.

3.2. Περιγραφή του συστήματος

Εδώ περιγράφεται η αρχιτεκτονική του συστήματός που κάλλιστα κάνει δυνατή την αξιόπιστη αναγνώριση των εν λόγω ηχητικών γεγονότων. Προσεγγίζουμε το ζήτημα βασισμένοι στον τρόπο με τον οποίο οι άνθρωποι ταξινομούν υποσυνείδητα το περιβάλλον τους χρησιμοποιώντας μόνο τις ακουστικές πληροφορίες. Ο στόχος μας είναι να διακρίνουμε τις σκηνές που ανήκουν σε οκτώ διαφορετικές κατηγορίες: *αεροσκάφος, μοτοσικλέτα, αυτοκίνητο, πλήθος, βροντή, αέρας, τραίνο και κόρνα.*

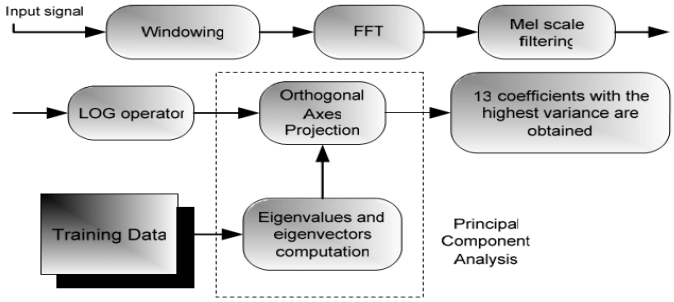


Σχήμα 3.1: Το προτεινόμενο δέντρο κατηγοριών

Χρησιμοποιούμε ένα ιεραρχικό σχήμα ταξινόμησης που αποτελείται από δύο στάδια και που προέρχεται από τον τρόπο που οι άνθρωποι αντιλαμβάνονται αυτές τις κατηγορίες (Σχ. 3.1). Η πρώτη φάση ταξινομεί τους ήχους σε δύο κατηγορίες (μηχανικοί και μη-μηχανικοί) ενώ ο δεύτερος ολοκληρώνει το υπόλοιπο της διαδικασίας ταξινόμησης προσδιορίζοντας την ακριβή κατηγορία. Επιπλέον η τοπολογία μας περιλαμβάνει την προεπεξεργασία των ακουστικών σημάτων, την εξαγωγή χαρακτηριστικών, την αφαίρεση των πλαισίων σιγής, PCA καθώς και διαφορετικών ειδών ταξινομητές, οι παράμετροι των οποίων αντιπροσωπεύουν την εκ των προτέρων γνώση που έχουμε διαθέσιμη για τις διάφορες κατηγορίες ήχων μέσω των δεδομένων εκπαίδευσης.

3.3. Εξαγωγή χαρακτηριστικών και αφαίρεση σιγής

Προκειμένου να αξιολογηθεί η απόδοση των επιλεγμένων χαρακτηριστικών γνωρισμάτων όσον αφορά την αναγνώριση περιβαλλοντικών ήχων πρέπει να ακολουθηθεί η ίδια μέθοδος προεπεξεργασίας. Κατά συνέπεια, χρησιμοποιήθηκαν ακριβώς οι ίδιες παράμετροι κατά τις δύο διαδικασίες εξαγωγής ακουστικών χαρακτηριστικών. Τα σήματα χωρίζονται σε πλαίσια των 30ms με την χρονική μετατόπιση μεταξύ δύο διαδοχικών πλαισίων να είναι 10ms ακολουθώντας τις οδηγίες του πρωτοκόλλου MPEG-7. Το μέγεθος του FFT είναι 512 ενώ εφαρμόζουμε την αφαίρεση μέσης τιμής στα ηχητικά κύματα για να αποβάλουμε οποιοδήποτε πιθανό DC- offset χωρίς να χρησιμοποιήσουμε προ-έμφαση.



Σχήμα 3.2 Η διαδικασία εξαγωγής παραμέτρων συμπεριλαμβανομένης της PCA

Στη παρούσα εφαρμογή η σιγή θεωρείται "θόρυβος", καθώς μπορεί να έχει αρνητική επίδραση στην τελική απόδοση. Συνεπώς, πριν από την εξαγωγή χαρακτηριστικών, η σιγή αφαιρείται χρησιμοποιώντας έναν ανιχνευτή δραστηριότητας φωνής (VAD) βασισμένο σε στατιστικά πρότυπα (Sohn et al, 1999). Στον Πίνακα 3.1 φαίνεται η ανάλυση της βάσης που χρησιμοποιήθηκε μετά από την επεξεργασία VAD.

Κατηγορίες ιεραρχικά	Πηγή ήχου	Αριθμός ηχογραφήσεων	Μέση διάρκεια (s)
Μηχανικοί	Αεροσκάφος	110	40.5
	Αυτοκίνητο	81	27.1
	Μοτοσικλέτα	79	14.8
	Τρένο	82	28.3
Μη-μηχανικοί	Πλήθος	60	58.6
	Αέρας	66	49
	Κόρνα	194	4.2
	Κεραυνός	60	16.3

Πίνακας 3.1. Ανάλυση των ηχητικών κατηγοριών.

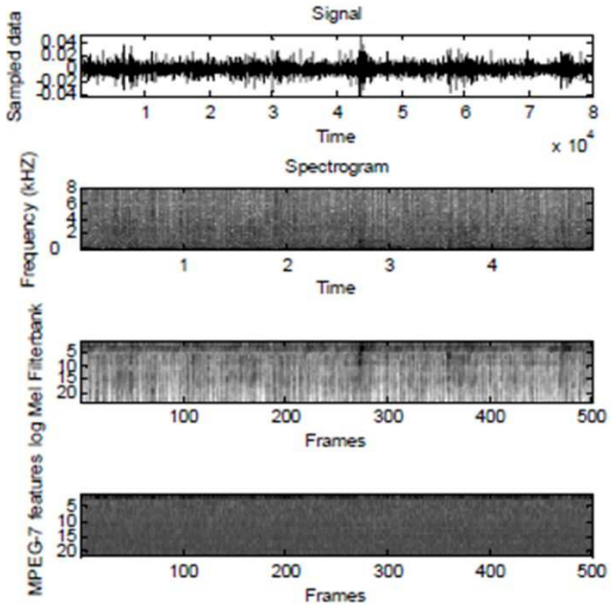
3.3.1. MFCs

Το πρώτο σύνολο χαρακτηριστικών αποτελείται από τους πρώτους 13 Mel cepstral συντελεστές συχνότητας. Τα διάφορα βήματα που οδηγούν στην εξαγωγή τους είναι: Αρχικά ο μετασχηματισμός STFT υπολογίζεται για κάθε πλαίσιο και στη συνέχεια φιλτράρεται χρησιμοποιώντας μια τριγωνική τράπεζα φίλτρων Mel. Έχει αποδειχθεί ότι αυτή η διαδικασία αναδεικνύει τα στοιχεία εκείνα που διαδραματίζουν σημαντικό ρόλο όσον αφορά την ανθρώπινη αντίληψη. Έπειτα χωρίζουμε λογαριθμικά τα δεδομένα και εκμεταλλευόμαστε τις ιδιότητες αποσυσχέτισης (decorrelation) του DCT έτσι ώστε η ενέργεια του κάθε πλαισίου να αντιπρωπεύεται από έναν μικρό σχετικά αριθμό των συντελεστών της. Τελικά λαμβάνονται υπόψη οι σημαντικότεροι 13 συντελεστές.

Σε αυτό το σημείο διερευνούμε τη χρήση μιας ορθογώνιας τεχνικής αποσυσχέτισης αντί του DCT. Η PCA υιοθετείται για να μειώσει τη διαστατικότητα των στοιχείων, προβάλλοντας τα σε άξονες οι οποίοι καθορίζονται από τα ίδια τα στοιχεία (MFCC-PCA feature set). Ο βασικός πυρήνας, που αποτελείται από όλα τα ιδιοδιανύσματα υπολογίζεται από τις τιμές χαρακτηριστικών που προέρχονται από ολόκληρο το σύνολο εκπαίδευσης. Στη συνέχεια κρατιούνται μόνο τα πρώτα 13 διανύσματα με τις υψηλότερες ιδιοτιμές. Με αυτήν την διαδικασία τα στοιχεία μετασχηματίζονται σε ένα νέο σύστημα συντεταγμένων που βασίζεται στις σχέσεις μεταξύ τους. Αυτή η διαδικασία φαίνεται στο Σχήμα 3.2. Τα άγνωστα διανύσματα χαρακτηριστικών μετασχηματίζονται σε διανύσματα χαμηλότερης διάστασης - βασισμένα στον πυρήνα που προέρχεται από τα δεδομένα εκπαίδευσης. Πρέπει να σημειωθεί ότι η PCA είναι μια διαδικασία που βασίζεται στα δεδομένα αντίθετα από τον DCT που συμπιέζει την ενέργεια των δεδομένων ακολουθώντας μια τυποποιημένη διαδικασία στάθμισης.

3.3.2. Παράμετροι του πρωτοκόλλου MPEG-7

Αυτό το σύνολο περιγραφέων έχει ως σκοπό την παραγωγή μιας συμπαγούς αντιπροσώπευσης του ακουστικού περιεχομένου. Η παραγωγή μεταδεδομένων γίνεται αυτόματα διευκολύνοντας τις τεχνικές έρευνας και ανάκτησης μέσα σε μεγάλες βάσεις δεδομένων.



Σχήμα 3.3 Mel-log filterbank και τα MPEG-7 χαρακτηριστικά για ένα δείγμα από την κατηγορία ήχων μοτοσικλέτας

Μια μεγάλη ποικιλία των ακουστικών περιγραφέων χαμηλού επιπέδου (LLDs) έχει κατηγοριοποιηθεί σε πρότυπα όπως φαίνεται το σύνολο MPEG-7 χαρακτηριστικών γνωρισμάτων στον Πίνακα 3.2. Αυτή η διαδικασία διαμορφώνει ένα διάνυσμα χαρακτηριστικών με 54 διαστάσεις. Ο αλγόριθμος μετα- επεξεργασίας που εξηγήθηκε προηγουμένως συμπεριλαμβανομένης της λογαριθμικής τμηματοποίησης εφαρμόζεται επίσης σε αυτό το σύνολο περιγραφέων (τα διανύσματα που αντιστοιχούν στις 13 υψηλότερες ιδιοτιμές λαμβάνονται - MPEG 7-PCA). Επιπλέον, για λόγους σύγκρισης, εξετάζεται επίσης και ο DCT. Η περιγραφή του σταδίου ταξινόμησης δίνεται στην επόμενη ενότητα. Ενδεικτικά στο Σχήμα 3.3 απεικονίζονται οι τιμές των δύο συνόλων χαρακτηριστικών για ένα δείγμα της κατηγορίας μοτοσικλέτας.

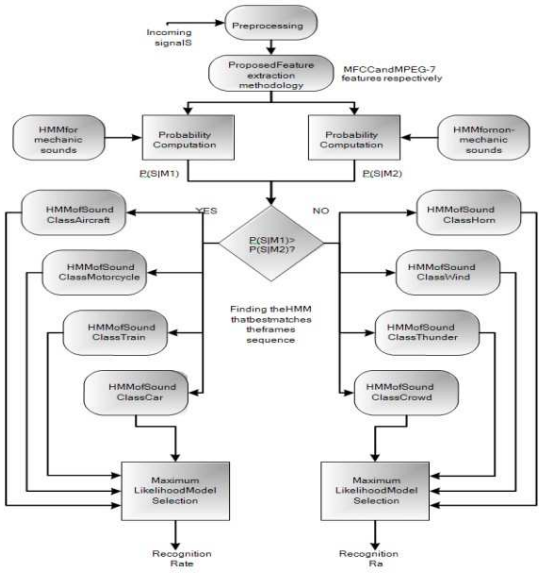
Περιγραφέας	**Διάσταση**	**Συντόμευση**
Audio Waveform	2	AW
Audio Power	1	AP
Audio Spectrum Envelope	27	ASE
Audio Spectrum Centroid	1	ASC
Audio Spectrum Spread	1	ASS
Audio Spectrum Flatness	19	ASF
Harmonic ratio	1	HR
Upper Limit of Harmonicity	1	ULH
Audio Fundamental Frequency	1	AFF

Πίνακας 3.2. Οι παράμετροι του πρωτοκόλλου MPEG-7 και οι διαστάσεις τους.

3.4. Διεξαγωγή πειραμάτων

3.4.1. Ιεραρχία Ταξινόμησης

Χρησιμοποιήσαμε ένα ιεραρχικό σχήμα ταξινόμησης που αποτελείται από δύο στάδια. Αρχικά ο εισερχόμενος ήχος ταξινομείται σε μια από τις επόμενες κατηγορίες: (α) μηχανικός, (β) μη-μηχανικός. Κατόπιν ένα διαφορετικό μέρος της διαδικασίας αναγνώρισης ενεργοποιείται ανάλογα με το πρώτο αποτέλεσμα και το οποίο παράγει την τελική κατηγορία.



Σχήμα 3.4 Ιεραρχική δομή ταξινόμησης της προτεινόμενης μεθοδολογίας.

Ένα ισχυρό εργαλείο που χρησιμοποιείται συνήθως στο πεδίο της επεξεργασίας ομιλίας, το HMM υιοθετήθηκε και για τα δύο βήματα ταξινόμησης. Για κάθε κατηγορία A ένα HMM δύο καταστάσεων δημιουργείται για τη μοντελοποίηση των δεδομένων εκπαίδευσης ενώ το όριο των επαναλήψεων του αλγορίθμου Baum-Welch που χρησιμοποιήθηκε είναι 50.

Οι καταστάσεις συνδυάζονται με μια από τα αριστερά προς τα δεξιά τοπολογία ενώ οκτώ Γκαουσσιανές συναρτήσεις χρησιμοποιούνται για να διαμορφώσουν κάθε κατάσταση. Απεικονίζουμε ολόκληρη τη διαδικασία ταξινόμησης στο Σχήμα 3.4. Οι παράμετροι επιλέχτηκαν μετά από μια σειρά πειραμάτων ενώ εξακριβώθηκε ότι ο ανωτέρω συνδυασμός επιτυγχάνει πολύ καλή απόδοση. Το HMM είναι βασισμένο στην υπόθεση ότι η διαδικασία που προσπαθούμε να μοντελοποιήσουμε ακολουθεί ένα πολύ διακριτό χρονικό σχέδιο/πρότυπο και ότι αυτό μπορεί να προσεγγιστεί και να προβλεφθεί. Κατά τη διαδικασία της μέτρησης της απόδοσης του συστήματος, οι ακολουθίες χαρακτηριστικών εισάγονται σε κάθε HMM με αποτέλεσμα ένα μέτρο της

πιθανότητας του συγκεκριμένου μοντέλου να παρήγαγε αυτήν την σειρά των χαρακτηριστικών. Πρέπει να σημειωθεί ότι ο αλγόριθμος Forward Backward χρησιμοποιήθηκε για τον υπολογισμό των πιθανοτήτων. Σε όλα τα πειράματα χρησιμοποιήθηκε η υλοποίηση σε περιβάλλον Matlab που διανέμεται ελεύθερα στην διεύθυνση http://www.npt.nuwc.navy.mil/Csf/.

3.4.2. Συλλογή δεδομένων και εκτίμηση απόδοσης

Εκτιμήσαμε την απόδοση και των δύο συνόλων γνωρισμάτων με και χωρίς τον προτεινόμενο αλγόριθμο μετα-επεξεργασίας. Η βάση που χρησιμοποιήθηκε περιέχει δεδομένα μεγάλης ποικιλομορφίας ακόμα και μεταξύ των στοιχείων της ίδιας κατηγορίας προκειμένου οι ακουστικές σκηνές να αντιπροσωπεύουν όσο το δυνατόν καλύτερα την πραγματικότητα. Συλλέχθηκε από διάφορες πηγές του παγκόσμιου δικτύου, αλλά κυρίως από τη BBC Sound Effects Library. Ο ρυθμός δειγματοληψίας των στοιχείων ήταν 16KHz με 16 bit ανάλυση (μονοφωνικό). Τα δεδομένα οργανώθηκαν σε οκτώ κατηγορίες με μέση διάρκεια 29.85 δευτερόλεπτα. Οι κατηγορίες είναι αεροσκάφος, μοτοσικλέτα, αυτοκίνητο, πλήθος, βροντή, αέρας, τραίνο και κόρνα καθεμία από τις οποίες αντιπροσωπεύεται από ένα HMM αφού έχει εφαρμοστεί η διαδικασία αφαίρεσης σιγής όπως περιγράφεται στην ενότητα 3.3.

Η διαδικασία αξιολόγησης περιλαμβάνει τα επόμενα βήματα: τα εισερχόμενα σήματα προεπεξεργάζονται και εξάγονται τα χαρακτηριστικά τους σύμφωνα με τις προαναφερθείσες τεχνικές. Το αποτέλεσμα εισάγεται στο ιεραρχικό δέντρο ταξινόμησης όπου εφαρμόζεται η μεθοδολογία μέγιστης πιθανότητας και λαμβάνεται η τελική απόφαση του συστήματος. Για να εξασφαλιστεί η αξιοπιστία των αποτελεσμάτων υιοθετήθηκε η προσέγγιση ten-fold cross validation (Witten et al., 2005) ενώ δόθηκε ιδιαίτερο βάρος ώστε μέρη του ίδιου ηχητικού δείγματος να μην περιλαμβάνονται ταυτόχρονα και στο σύνολο εκπαίδευσης και στο σύνολο εξέτασης.

3.5. Αποτελέσματα

Σε αυτή την ενότητα περιγράφονται τα αποτελέσματα της προτεινόμενης μεθοδολογίας. Η τυποποιημένη έκδοση του συνόλου MFCC κατορθώνει να φθάσει σε 81.7% για τη πρώτη φάση του ταξινομητή ενώ οι MPEG-7 περιγραφείς μετα-επεξεργασμένοι (με log και DCT) έφθασαν σε 93.5%. Σημαντικά καλύτερη απόδοση επιτεύχθηκε με τη χρησιμοποίηση της τεχνικής PCA: 90.6% για MFCC και 100% για MPEG -7 Low Level Descriptors (LLDs). Οι μήτρες σύγχυσης (confusion matrices) που αντιστοιχούν στο δεύτερο μέρος της ιεραρχίας ταξινόμησης απεικονίζονται στους επόμενους πίνακες (Πίνακες 3.3-3.6).

Αρχικά αναλύουμε τα αποτελέσματα χωρίς τη χρήση PCA. Μπορεί να παρατηρηθεί ότι και τα δύο σύνολα χαρακτηριστικών παρέχουν τα καλύτερα ποσοστά αναγνώρισης για τον προσδιορισμό της τελικής κατηγορίας όταν αυτή ανήκει στο σύνολο των μη- μηχανικών. Η ακρίβεια ταξινόμησης που επιτυγχάνονται για τις μηχανικές κατηγορίες είναι 77.75% και 76.3% ενώ για τις μη-μηχανικές κατηγορίες η απόδοση φτάνει 78.4% και 83.35% για τα MFCCs και τους περιγραφείς του MPEG-7 αντίστοιχα. Γενικά μπορούμε να πούμε ότι και τα δύο σύνολα καταλήγουν στα ίδια λάθη ταξινόμησης, κάτι το οποίο δείχνει ότι συλλαμβάνουν σχετικά όμοιες πληροφορίες. Η κατηγορία της κόρνας αναγνωρίζεται με την καλύτερη ακρίβεια ενώ ο προσδιορισμός των ήχων αεροσκαφών, αυτοκινήτων και αέρα έδωσε τα πιο φτωχά αποτελέσματα. Ένας πιθανός λόγος είναι η μεγάλη ποικιλομορφία μεταξύ των ηχητικών δειγμάτων που ανήκουν στην ίδια κατηγορία (High within-class variability).

Επίσης διάφορα ηχητικά δείγματα είναι ηχητικά παρόμοια αν και ανήκουν σε διαφορετικές κατηγορίες. Για παράδειγμα πολλά δείγματα αυτοκινήτων ακούγονται παρόμοια με ήχους τραίνων ενώ το ίδιο πράγμα ισχύει μεταξύ αρκετών ήχων των κατηγοριών πλήθους και αέρα.

Πολλές λάθος ταξινομήσεις που εμφανίστηκαν, διορθώθηκαν από την PCA όπως φαίνεται στους Πίνακες 3.5 και 3.6. Τα ποσοστά αναγνώρισης κάθε μερικού προβλήματος βελτιώθηκαν δραστικά. Αν και η ίδια ακουστική ασυνέπεια ισχύει εδώ, το σύστημα προσφέρει υψηλή αξιοπιστία και άριστη απόδοση με μέσες ακρίβειες 89.75% και 97.15% για MFCC και MPEG-7 χαρακτηριστικά αντίστοιχα. Εξάγεται το συμπέρασμα ότι και τα δύο σύνολα γνωρισμάτων οδηγούν στα ίδια λάθη με το υψηλότερο ποσοστό λάθους στην κατηγορία του αέρα. Πρέπει να υπογραμμίσουμε ότι τα MPEG-7 LLDs φτάνουν στο υψηλότερο ποσοστό σωστής ταξινόμησης, κάτι που καθιστά την χρησιμοποίησή τους στην εφαρμογή της αυτόματης αναγνώρισης ηχητικών γεγονότων αστικού περιβάλλοντος, απαραίτητη.

Απόκριση	Αεροσκ.	Μοτοσυκλ.	Αυτοκίν.	Τρένο
Αεροσκάφος	**75**	0	18.7	6.3
Μοτοσυκλέτα	7.7	**84.7**	7.6	0
Αυτοκίνητο	7.6	7	**70**	15.4
Τρένο	12.5	0	6.2	**81.3**
Απόκριση	**Αέρας**	**Κεραυνός**	**Πλήθος**	**Κόρνα**
Αέρας	**69.4**	10.5	20.1	0
Κεραυνός	0	**85.7**	0	14.3
Πλήθος	0	9.1	**72.8**	18.1
Κόρνα	5.7	8.5	0	**85.8**

Πίνακας 3.3. Μήτρες σύγχυσης - τυπικά MFCCs.

Απόκριση	Αεροσκ.	Μοτοσυκλ.	Αυτοκίν.	Τρένο
Αεροσκάφος	**93.8**	6.2	0	0
Μοτοσυκλέτα	0	**86.4**	0	13.6
Αυτοκίνητο	0	0	**100**	0
Τρένο	6.2	0	0	**93.8**
Απόκριση	**Αέρας**	**Κεραυνός**	**Πλήθος**	**Κόρνα**
Αέρας	**71**	14	0	15
Κεραυνός	0	**100**	0	0
Πλήθος	5.1	0	**82.8**	12.1
Κόρνα	9.7	0	0	**90.3**

Πίνακας 3.5. Μήτρες σύγχυσης – MFCCs με PCA.

Απόκριση	Αεροσκ.	Μοτοσυκλ.	Αυτοκίν.	Τρένο
Αεροσκάφος	**55.6**	0	25.7	18.7
Μοτοσυκλέτα	14.3	**71.1**	14.6	0
Αυτοκίνητο	0	0	**84.6**	15.4
Τρένο	0	0	6.2	**93.8**
Απόκριση	**Αέρας**	**Κεραυνός**	**Πλήθος**	**Κόρνα**
Αέρας	**83.4**	0	0	16.6
Κεραυνός	0	**85.7**	0	14.3
Πλήθος	20	10	**70**	0
Κόρνα	5.7	0	0	**94.3**

Πίνακας 3.4. Μήτρες σύγχυσης - MPEG-7 MPEG-7 με (συμπεριλαμβανομένων log και DCT).

Απόκριση	Αεροσκ.	Μοτοσυκλ.	Αυτοκίν.	Τρένο
Αεροσκάφος	**100**	0	0	0
Μοτοσυκλέτα	0	**100**	0	0
Αυτοκίνητο	0	0	**100**	0
Τρένο	0	7	0	**93**
Απόκριση	**Αέρας**	**Κεραυνός**	**Πλήθος**	**Κόρνα**
Αέρας	**84.4**	6	9.6	0
Κεραυνός	0	**100**	0	0
Πλήθος	0	0	**100**	0
Κόρνα	0	0	0	**100**

Πίνακας 3.6. Μήτρες σύγχυσης - PCA

Κεφάλαιο 4

Ακουστική Παρακολούθηση Καταστροφικών Καταστάσεων

Στο παρών κεφάλαιο παρουσιάζεται μια πρακτική μεθοδολογία για τον αυτόματο έλεγχο χώρου βασισμένο απλώς στις εισερχόμενες ακουστικές πληροφορίες. Εξετάζεται η0 περίπτωση όπου μη τυπικές καταστάσεις όπως κραυγές, εκρήξεις και πυροβολισμοί λαμβάνουν χώρα σε ένα περιβάλλον σταθμού μετρό. Η προσέγγισή είναι βασισμένη σε ένα σχήμα δύο επιπέδων αναγνώρισης, καθένα από τα οποία εκμεταλλεύεται την τεχνολογία HMM για να προσεγγίσει τη συνάρτηση πυκνότητας της αντίστοιχης ηχητικής κατηγορίας. Ο κύριος στόχος είναι να ανιχνευθούν τα ανώμαλα γεγονότα που πραγματοποιούνται σε ένα θορυβώδες περιβάλλον. Ακολουθεί μια λεπτομερή διαδικασία αξιολόγησης για διαφορετικούς λόγους σήματος προς θόρυβο και παρουσιάζονται υψηλά ποσοστά ανίχνευσης όσον αφορά τα ποσοστά ψευδών συναγερμών και λάθος ανιχνεύσεων.

4.1. Εισαγωγή

Η έρευνα στον τομέα των αυτόματων συστημάτων παρακολούθησης στρέφεται κυρίως στην ανίχνευση των ανώμαλων γεγονότων βασισμένων σε καταγραμμένες οπτικές πληροφορίες (Haritaoglou et al., 2000). Οι τρέχουσες υλοποιήσεις αποτελούνται τυπικά από έναν μεγάλο αριθμό καμερών που κατανέμονται σε μια περιοχή και συνδέονται σε ένα κεντρικό δωμάτιο ελέγχου. Ενώ αυτό το είδος ανάλυσης παρέχει πολύτιμες πληροφορίες, εμείς επικεντρωνόμαστε στην ανίχνευση των μη τυπικών γεγονότων εκμεταλλευόμενοι μόνο πληροφορίες ακουστικής μορφής. Αυτή η προσέγγιση προσφέρει διάφορα πλεονεκτήματα όπως: α) χαμηλό υπολογιστικό κόστος ενώ β) οι συνθήκες φωτισμού του χώρου προς έλεγχο ή/και πιθανά εμπόδια δεν έχουν άμεση επιρροή στο ηχητικό σήμα. Οι προηγούμενες προσεγγίσεις σχετικά με το θέμα του ακουστικού ελέγχου περιλαμβάνουν ένα αρκετά μεγάλο φάσμα εφαρμογών.

Ο κύριος σκοπός του προτεινόμενου συστήματος είναι να χαρακτηριστεί αποτελεσματικά ο περιβάλλοντας χώρος ως προς την ασφάλειά του χρησιμοποιώντας ένα απλό μικρόφωνο. Το σύστημα είναι σχεδιασμένο ώστε να βοηθήσει το εξουσιοδοτημένο προσωπικό να εκτελέσει τις κατάλληλες ενέργειες για την παρεμπόδιση εγκληματικής πράξης ή/και τη ζημιά ιδιοκτησίας. Για να είναι χρήσιμο ένα τέτοιο σύστημα πρέπει να προσφέρει πολύ χαμηλό ποσοστό εσφαλμένων συναγερμών κρατώντας παράλληλα την ακρίβεια ανίχνευσης όσο το δυνατόν υψηλότερη ακόμα και κάτω από θορυβώδεις συνθήκες. Η προσέγγισή μας βασίζεται στο γεγονός ότι ο ήχος παρέχει πληροφορίες που είναι δύσκολο ή ακόμα και αδύνατο να ληφθούν με

οποιοδήποτε άλλο μέσο. Συν τοις άλλης, μια τέτοια μέθοδος χαρακτηρίζεται από χαμηλό κόστος και σχετικά εύκολη διαδικασία εγκατάστασης. Σε αυτό το κεφάλαιο επικεντρωνόμαστε στην ανίχνευση των μη τυπικών ηχητικών γεγονότων (κραυγή, πυροβολισμός και έκρηξη) σε ένα περιβάλλον σταθμού μετρό. Η παρούσα μεθοδολογία εμπνέεται από την εργασία (Wilpon et al., 1990) σχετικά με την εύρεση συγκεκριμένων λέξεων (keyword spotting). Επεκτείνουμε αυτήν την ιδέα στον τομέα της εύρεσης keysound, όπου οι κραυγές, οι πυροβολισμοί και οι εκρήξεις θεωρούνται ως keysound γεγονότα. Στην περίπτωσή μας το αδιάφορο πρότυπο (garbage model) είναι η ακουστική σκηνή του σταθμού μετρό που παρουσιάζει ιδιαίτερα μη στατικές ιδιότητες (περιλαμβάνει κόρνες, το άνοιγμα/κλείσιμο πορτών, ανθρώπους που μιλούν στο υπόβαθρο, κίνηση τραίνων κ.λπ.).

Έχουν πραγματοποιηθεί εκτενείς πειραματισμοί σχετικά με το καλύτερο σύνολο χαρακτηριστικών που περιλαμβάνονται στη διαδικασία εξαγωγής. Το τελικό σύνολο αποτελείται από τους γνωστούς συντελεστές MFCC και από μια δεύτερη ομάδα παραμέτρων βασισμένων στο ακουστικό πρότυπο MPEG-7. Στη συνέχεια οι ακολουθίες των γνωρισμάτων διαμορφώνονται από συναρτήσεις πυκνότητας πιθανότητας που αντιπροσωπεύονται από GMMs και HMMs. Το επόμενο μέρος αυτού του κεφαλαίου είναι οργανωμένο ως εξής: στην παράγραφο 4.2 δίνεται μια συνοπτική επισκόπηση του συστήματος μαζί με μια περιγραφή των παραμέτρων. Η παράγραφος 4.3 εξηγεί το πειραματικό πρωτόκολλο που χρησιμοποιείται ενώ τα συμπεράσματά αναφέρονται στο τελευταίο τμήμα.

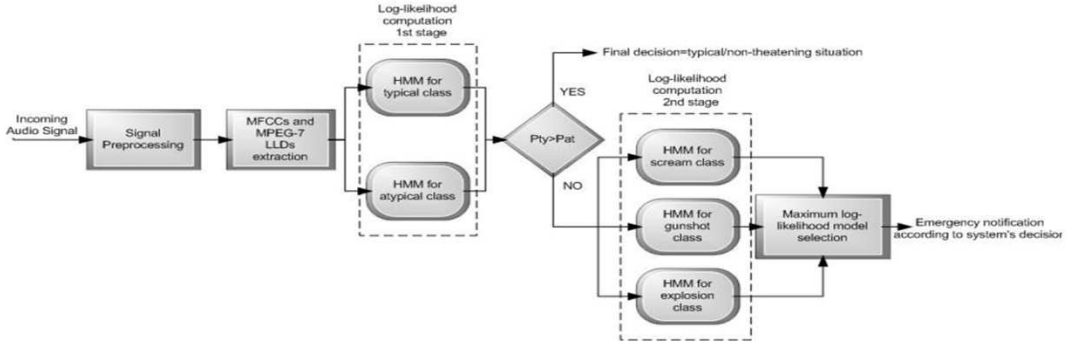
4.2. Περιγραφή του συστήματος

Το σύστημά έχει σχεδιαστεί ως μία τοπολογία δύο επιπέδων που αποδείχθηκε ότι παρέχει καλύτερα ποσοστά αναγνώρισης από αυτήν του ενός. Το εισερχόμενο σήμα ταξινομείται πρώτα ως τυπικό (περιβάλλον σταθμού μετρό) ή μη τυπικό (κραυγή, πυροβολισμός ή έκρηξη) και σε περίπτωση που το σύστημα "αποφασίζει" ότι είναι μη τυπικό ακολουθεί ένα δεύτερο στάδιο επεξεργασίας όπου προσδιορίζεται ο τύπος της ανωμαλίας. Η προτεινόμενη αρχιτεκτονική αποτελεί μια πλήρως πιθανοτική δομή βασισμένη σε εργοδικά HMMs (πλήρως συνδεδεμένα) για την αντιπροσώπευση κάθε ηχητικής κατηγορίας.

4.2.1. Ανάλυση της ακουστικής παραμετροποίησης

Σε αυτό το τμήμα σχολιάζουμε τις ομάδες περιγραφέων που χρησιμοποιήθηκαν προκειμένου να εκπαιδευθούν τα πιθανοτικά πρότυπα που αντιπροσωπεύουν την *a-priori* γνώση που έχουμε για τις ηχητικές κατηγορίες. Χρησιμοποιούμε τη τράπεζα φίλτρων Mel λόγω της δυνατότητάς της να ελαττώνει τον αριθμό των διαστάσεων του διανύσματος του μετασχηματισμένου Fourier. Στη συνέχεια περιλαμβάνουμε το

λογαριθμικό διαμοίρασμα των δεδομένων, μια διαδικασία που μιμείται τη φυσική επιλεκτικότητα των φασματικών ζωνών που επιδεικνύει το μέσο ανθρώπινο αυτί ως ένα ορισμένο βαθμό (βλέπε και Σελίδα 35). Επίσης υιοθετήσαμε το πρωτόκολλο MPEG-7 δεδομένου ότι αυτήν την περίοδο αποτελεί το state of the art στο χώρο της αυτοματοποιημένης αναγνώρισης γενικευμένου ακουστικού σήματος που βασίζεται στο περιεχόμενο. Υιοθετούμε τους επόμενους τέσσερις περιγραφείς χαμηλού επιπέδου:



Σχήμα 4.1: Μπλοκ διάγραμμα του προτεινόμενου συστήματος ακουστικής επόπτευσης.

Audio Fundamental Frequency, Waveform min και max και τον Audio Spectrum Flatness. Τα προαναφερθέντα σύνολα χαρακτηριστικών γνωρισμάτων αξιολογήθηκαν και τα δύο χωριστά αλλά και συνδυασμένα χρησιμοποιώντας διαφορετικές τιμές παραμέτρων κάθε φορά. Επιλέχτηκαν επειδή συλλαμβάνουν διαφορετικές πτυχές των πληροφοριών σε σχέση με αυτές που παρέχονται από τα MFCC. Τα MFCCs είναι μια Mel-βαθμισμένη προβολή του λογαρίθμου του φάσματος ενώ το ASF αποτελεί μια υψηλότερου επιπέδου περιγραφή του ακουστικού σήματος που δείχνει πόσο επίπεδο είναι το δεδομένο φάσμα. Πρέπει να σημειωθεί ότι εξετάστηκε η ενσωμάτωση ενός συνόλου παραμέτρων βασισμένων στην ανάλυση του εμβαδού των συναρτήσεων αυτοσυσχέτισης βασισμένη σε κρίσιμες ζώνες (Teager energy operator autocorrelation envelope area) η οποία προτείνεται στην εργασία Zhoun et al., 2001. Αυτές οι παράμετροι χρησιμοποιούνται για την ταξινόμηση πίεσης/άγχους από σήματα ομιλίας αλλά ο συνδυασμός τους με τα προαναφερθέντα χαρακτηριστικά δεν πρόσφερε βελτιωμένα αποτελέσματα.

Σε αυτήν την εργασία εξετάζουμε το keysound spotting σε περιβάλλον σταθμού μετρό όπου τα ακουστικά σήματα που χαρακτηρίζονται από μεγάλη διάρκεια πρέπει να υποβληθούν σε επεξεργασία με σκοπό την ανίχνευση μη τυπικών ηχητικών γεγονότων. Κατά συνέπεια η στιγμιαία τιμή κάθε χαρακτηριστικού υπολογίζεται από ένα μεγαλύτερο μέγεθος πλαισίων από αυτά που χρησιμοποιούνται συνήθως. Μετά από διάφορα πειράματα αποφασίστηκε ότι όλα τα ηχητικά δείγματα πρέπει να κοπούν σε

πλαίσια μεγέθους 200ms με επικάλυψη 50%. Επίσης χρησιμοποιούμε αφαίρεση μέσης τιμής και διαβάθμιση της διασποράς για την κανονικοποίηση των δεδομένων.

Κατηγορία	Αριθμός ηχογραφήσεων	Διάρκεια (s)
Εκρηξη	131	13.77
Πυροβολισμός	187	32.94
Κραυγή	270	4.04
Σταθμός μετρό	32	44.88
Σύνολο	620	23.9

Πίνακας 4.1: Οι κατηγορίες της βάσης των ακουστικών δεδομένων

4.2.2. Σχήματα κατηγοριοποίησης

Χρησιμοποιήσαμε διαγώνια GMM και HMM δύο διαφορετικών τοπολογιών (από τα αριστερά προς τα δεξιά και πλήρως-συνδεμένο). Στη συνέχεια τα προηγουμένως δημιουργημένα πρότυπα χρησιμοποιούνται για τον υπολογισμό ενός βαθμού ομοιότητας (π.χ. log likelihood) μεταξύ κάθε προτύπου και ενός άγνωστου σήματος εισαγωγής. Αυτός ο τύπος αποτελέσματος συγκρίνεται με τα υπόλοιπα και η τελική απόφαση γίνεται με τον προσδιορισμό της μέγιστης log-likelihood. Κατά τη διάρκεια ολόκληρης της διαδικασίας χρησιμοποιήθηκε η υλοποίηση Torch (για τα GMM και τα HMM) η οποία είναι γραμμένη σε C++. Ο μέγιστος αριθμός επαναλήψεων του k-means για την έναρξη ήταν 50 ενώ και οι EM και Baum-Welch αλγόριθμοι είχαν και ανώτερο όριο 25 επαναλήψεων με ένα κατώτατο όριο ίσο με 0.001 μεταξύ των πιθανοτήτων συνεχόμενων επαναλήψεων.

4.3. Πειραματική διαδικασία

Φυσικές βάσεις δεδομένων με ακραίες συναισθηματικές εκδηλώσεις και μη τυπικά γεγονότα ήχων για εφαρμογές επιτήρησης δεν είναι δημόσια διαθέσιμες λόγω του ιδιωτικού χαρακτήρα των δεδομένων, της έλλειψης τους και της μη προβλεψιμότητάς τους.

Η βάση μας αποτελείται από ήχους επαγγελματικών ηχητικών συλλογών. Αυτά τα είδη συλλογών αποτελούν μια τεράστια πηγή ηχητικών καταγραφών υψηλής ποιότητας που χρησιμοποιούνται κυρίως από τη κινηματογραφική βιομηχανία. Μια σημαντική λεπτομέρεια, που δεν είναι ευρέως γνωστή, είναι ότι ο ήχος σε μία ταινία δεν είναι ο ακριβής ήχος που καταγράφεται σε μια σκηνή αλλά υποβάλλεται σε επεξεργασία και στις περισσότερες περιπτώσεις προστίθεται χωριστά στην ακουστική ροή αργότερα. Επομένως, υπάρχει διαθέσιμη μεγάλη βάση πραγματικού φωνητικού και μη-φωνητικού ήχου για την κατάρτιση των εκπαιδευμένων πιθανοτικών προτύπων ταξινόμησης. Χρησιμοποιήσαμε ηχητικά δείγματα από τις ακόλουθες συλλογές: (i) BBC Sound Effects Library, (ii) Sound Ideas Series 6000, (iii) Sound Ideas: the art of Foley,

(iv) Best Service Studio Box Sound Effects και (v) ηχητικά δείγματα από αναζήτηση στο διαδίκτυο για την κατασκευή της τελικής βάσης.

4.3.1. Δημιουργία μοντέλων και ακρίβεια αναγνώρισης

Τα δεδομένα που ανήκουν σε κάθε κατηγορία χωρίστηκαν σε 75% για την εκπαίδευση και 25% για τη δοκιμή με τυχαίο τρόπο. Ένα πλήρως-συνδεμένο HMM χτίστηκε για κάθε κατηγορία για να συλλάβει τις ιδιότητές της ενώ η δοκιμή αποτελείται από μια απλή σύγκριση των log-likelihoods. Λόγω της αρχιτεκτονικής του συστήματος κατασκευάσαμε αρχικά δύο είδη προτύπων: τυπικό (ακουστική σκηνή σταθμού μετρό) και μη-τυπικό (που συμπεριλαμβάνει έκρηξη, πυροβολισμό και κραυγή). Μετά από εκτενείς πειραματισμούς χρησιμοποιήσαμε 6 καταστάσεις όπου καθεμία διαμορφώθηκε από 19 Γκαουσσιανές συναρτήσεις ενώ το μέσο ποσοστό αναγνώρισης που επιτεύχθηκε ήταν 98.87%. Όσον αφορά το δεύτερο στάδιο, χτίσαμε τρία HMMs για την περιγραφή κάθε μη τυπικής κατάστασης. Οι ίδιες παράμετροι παρείχαν την υψηλότερη μέση ακρίβεια αναγνώρισης - 93.05% - και η αντίστοιχη μήτρα σύγχυσης παρουσιάζεται στον Πίνακα 4.2.

Αποκρίθηκε Παρουσιάστηκε	**Έκρηξη**	**Πυροβολισμός**	**Κραυγή**
Εκρηξη	86.06	11.62	2.32
Πυροβολισμός	1.72	93.10	5.17
Κραυγή	0	0	100

Πίνακας 4.2: Μήτρα σύγχυσης για τις τρεις μη τυπικές ηχητικές κατηγορίες (%).

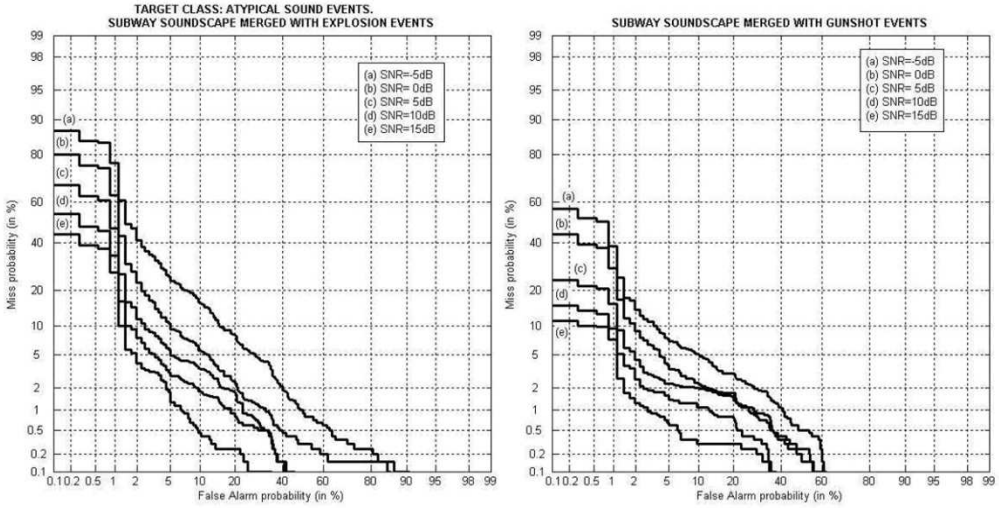
Παρατηρούμε ότι τα ηχητικά γεγονότα κραυγής αναγνωρίζονται με την καλύτερη ακρίβεια. Αυτό οφείλεται στη διαφορετική ενεργειακή κατανομή που έχουν οι μη τυπικές φωνητικές αντιδράσεις σε σύγκριση με τις υπόλοιπες κατηγορίες. Η χαμηλότερη ακρίβεια λαμβάνεται στα ηχητικά γεγονότα έκρηξης, εκ των οποίων το 11.62% κατηγοριοποιούνται λανθασμένα ως πυροβολισμοί. Πολλά από τα λάθη εμφανίζονται λόγω της μεγάλης μεταβλητότητας μεταξύ των ηχητικών δειγμάτων της ίδιας κατηγορίας.

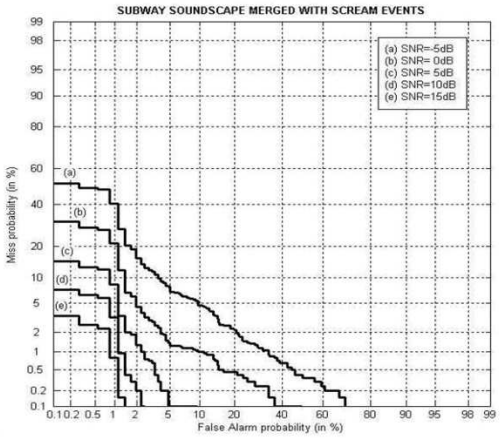
4.3.2. Εντοπισμός μη τυπικού ηχητικού γεγονότος σε σταθμό μετρό

Οι επείγουσες καταστάσεις που λαμβάνουν χώρα σε σταθμό μετρό δημιουργήθηκαν τεχνητά συγχωνεύοντας ανώμαλα ηχητικά γεγονότα με τις καταγραφές υπογείων σιδηροδρόμων σε διαφορετικά επίπεδα θορύβου (από -5dB μέχρι 15dB με βήμα 5dB SNR). Η προτεινόμενη αρχιτεκτονική εξετάστηκε χρησιμοποιώντας τις καμπύλες Detection Error Tradeoff (DET). Δύο σειρές πειραμάτων πραγματοποιήθηκαν αφιερωμένες σε κάθε στάδιο της υλοποίησής μας. Οι καμπύλες DET και για τα

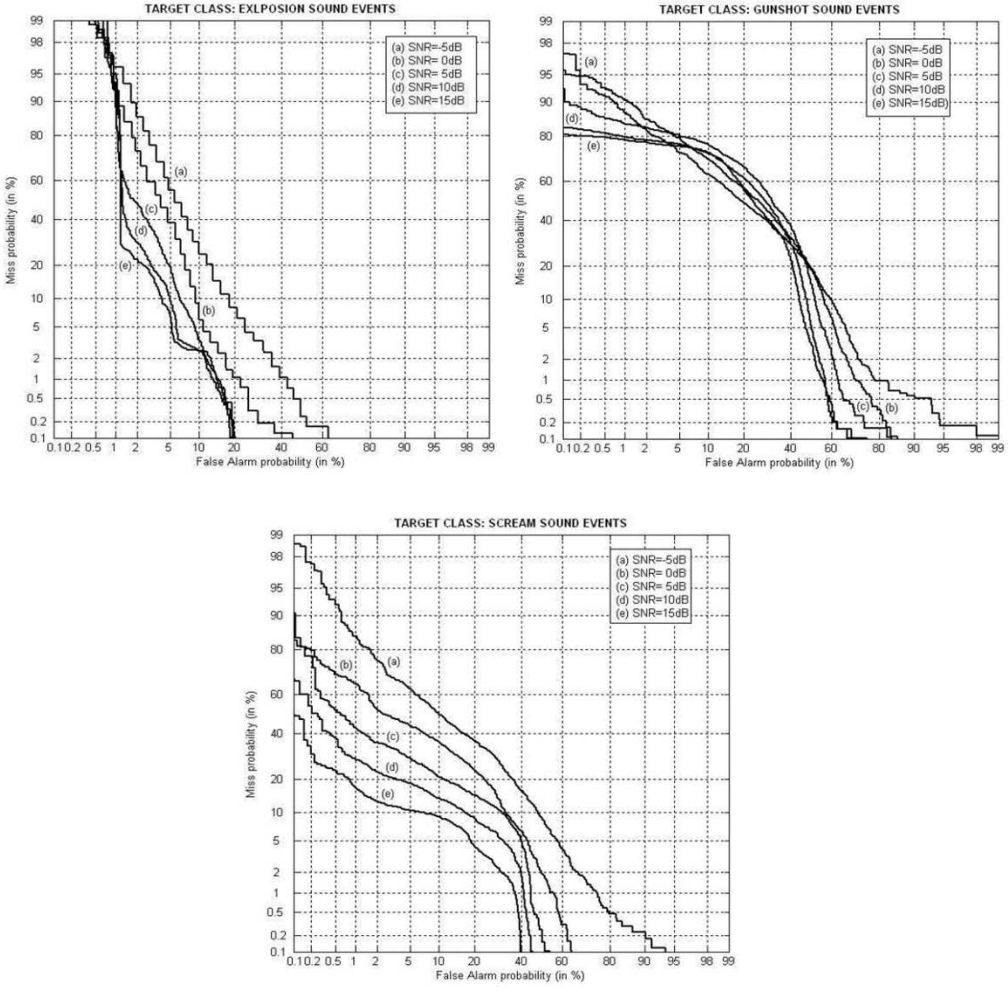
δύο στάδια απεικονίζονται στα Σχήματα 4.2 και 4.3 (στάδιο 1 και στάδιο 2 αντίστοιχα). Το Σχήμα 4.2 παρέχει τα αποτελέσματα της ανίχνευσης άτυπων γεγονότων περιλαμβάνοντας και τις τρεις κατηγορίες. Οι τιμές των log-likelihood των δύο στατιστικών προτύπων (τυπικών/μη τυπικών) χρησιμοποιήθηκαν για τη δημιουργία των καμπύλων DET. Τα αποτελέσματα ακολουθούν μια γρήγορη μείωση όσο μειώνεται το SNR. Οι ανώμαλοι ήχοι ανιχνεύονται επαρκώς ακόμη και σε εξαιρετικά χαμηλές τιμές του SNR. Το μέσο ποσοστό λάθους (EER) σχετικά με όλους τους τύπους γεγονότων σε SNR ίσο με -5dB είναι 8.53% ενώ η ελάχιστη τιμή του (καλύτερο ποσοστό ανίχνευσης) επιτυγχάνεται στην κατηγορία του πυροβολισμού. Τα ακουστικά σήματα που είναι πιο τρωτά στην πρόσθεση παρασιτικού θορύβου είναι αυτά της έκρηξης με EER ίσο με 12.88% σε -5dB SNR. Για τους στόχους επιτήρησης μια ενεργειακή αναλογία ίση με 0dB αντιπροσωπεύει κατάλληλα τις συνθήκες πραγματικού κόσμου. Το προτεινόμενο σύστημα επιδεικνύει πολύ καλή απόδοση στη συγκεκριμένη αναλογία, με μέσο EER ίσο με 4.8% και ρυθμό ψευδών συναγερμών ίσο με 1.83%, τιμή που είναι ιδιαίτερα σημαντική για αυτό το είδος των εφαρμογών.

Το Σχήμα 4.3 επεξηγεί τις ικανότητες του συστήματος σχετικά με την ανίχνευση κάθε μη τυπικής ηχητικής κατηγορίας που συγχωνεύεται με τις καταγραφές σταθμών μετρό σε διαφορετικά επίπεδα θορύβου. Οι λάθος ταξινομήσεις που εμφανίζονται σε αυτό το επίπεδο επεξεργασίας περιλαμβάνουν τα λάθη που είναι λιγότερο σημαντικά σε σύγκριση με το προηγούμενο.





Σχήμα 4.2: Οι καμπύλες DET 1ου σταδίου όσον αφορά στα μη τυπικά γεγονότα ως την κατηγορία στόχου κάτω από διαφορετικά SNRs. Κάθε σχήμα αντιστοιχεί σε αποτελέσματα που ελήφθησαν συγχωνεύοντας το σήμα του ακουστικού περιβάλλοντος μετρό με αυτό των μη τυπικών ηχητικών γεγονότων



Σχήμα 4.3: Οι καμπύλες DET 2ου σταδίου όπου η κάθε μία αντιστοιχεί σε ένα συγκεκριμένο μη τυπικό ηχητικό γεγονός ως την κατηγορία στόχου κάτω από διαφορετικά SNRs. Κάθε σχήμα αντιστοιχεί σε αποτελέσματα που ελήφθησαν όσον αφορά στον εντοπισμό εκρήξεων, πυροβολισμών και κραυγών.

Εδώ μια απειλητική κατάσταση έχει ανιχνευθεί και το σύστημα προσπαθεί να προσδιορίσει ποιος τύπος ανωμαλίας είναι παρών ενώ ένα εξουσιοδοτημένο πρόσωπο έχει ήδη ειδοποιηθεί προκειμένου να πάρει τα κατάλληλα μέτρα. Κατά συνέπεια, σ' αυτό το στάδιο της αναγνώρισης το κύριο ενδιαφέρον μας ποσοστό είναι να λάβουμε πολύ χαμηλό λάθος ανιχνεύσεων και έπειτα να προσπαθήσουμε να αποκτήσουμε όσο το δυνατόν χαμηλότερο ποσοστό ψευδών συναγερμών. Οι log-likelihoods που αποκτήθηκαν από τα πιθανοτικά πρότυπα, τα οποία περιγράφουν κάθε μη τυπική ηχητική κατηγορία, χρησιμοποιήθηκαν κατά τη διάρκεια αυτής της φάσης. Μπορούμε να παρατηρήσουμε ότι τα γεγονότα πυροβολισμού ανιχνεύονται με σχετικά χαμηλά EERs σε όλες τις τιμές SNR σε αντίθεση με τα δύο άλλα είδη μη τυπικών γεγονότων τα οποία ανιχνεύονται με απλώς ικανοποιητική ακρίβεια. Όπως αναμενόταν, η πιθανότητα miss detection πέφτει καθώς οι όροι SNR αυξάνονται από -5dB σε 15dB. Πιο ακριβέστερα οι ήχοι έκρηξης, που αλλοιώνονται από τον περιβαλλοντικό θόρυβο σταθμών μετρό με αναλογία ίση με -5dB ανιχνεύονται με EER 13.2%, οι ήχοι πυροβολισμού με 24.5% και οι ήχοι κραυγής με 28.2%. Επιπλέον, η εφαρμογή μας παρέχει πολύ καλή πιθανότητα false alarms με μέση τιμή 6% μεταξύ των τριών ηχητικών κατηγοριών για SNR ίσο με 0dB. Τα αντίστοιχα EERs που επιτυγχάνονται από το σύστημα όσον αφορά στην έκρηξη, στο πυροβολισμό και στην ανίχνευση κραυγής είναι 8.54%, 24.5% και 21.1%.

Κεφάλαιο 5

Προσαρμοζόμενο Σύστημα για Ακουστική Επόπτευση με Στόχο τον Εντοπισμό Μη- τυπικών Καταστάσεων

Η αξιόπιστη αναγνώριση γενικευμένων ακουστικών γεγονότων αποτελεί ένα θέμα της εντατικής έρευνας που αφορά την κοινότητα της επεξεργασίας σήματος. Σε αυτό το κεφάλαιο παρουσιάζεται μια αποδοτική μεθοδολογία για την ακουστική επιτήρηση μη- τυπικών καταστάσεων που μπορεί να χρησιμοποιηθεί κάτω από διάφορα ακουστικά περιβάλλοντα . Ο αρχικός στόχος είναι ο συνεχής ακουστικός έλεγχος μιας σκηνής για ενδεχομένως επικίνδυνα γεγονότα προκειμένου να βοηθηθεί ένας εξουσιοδοτημένος υπάλληλος ούτως ώστε να λάβει τις απαραίτητες ενέργειες με στόχο

την αποφυγή ανθρώπινων απωλειών ή/και ζημία ιδιοκτησίας. Σχεδιάστηκε ένας πιθανολογικός ιεραρχικός αναγνωριστής βασισμένος σε GMM και σε state of the art ακουστικές παραμέτρους που επιλέχθηκαν μετά από εκτενείς πειραματισμούς. Ένα χαρακτηριστικό του προτεινόμενου συστήματος είναι ο βρόχος ανατροφοδότησης που είναι ειδικά σχεδιασμένος για την προσαρμογή των μοντέλων και ο οποίος παρέχει προσαρμοστικότητα σε διαφορετικά ηχητικά περιβάλλοντα. Εκθέτουμε εκτενή πειραματικά αποτελέσματα βιβλιογραφίας συμπεριλαμβανομένης της εγκατάστασης του συστήματος σε πραγματικό περιβάλλον. Επίσης αναφέρονται ποσοστά ανίχνευσης που αφορούν στη λειτουργία του συστήματος για τρεις συνεχείς ημέρες. Επιπλέον υιοθετήσαμε μια αξιόπιστη διαδικασία αξιολόγησης ενώ η απόδοση του συστήματος έφτασε σε υψηλά επίπεδα όσον αφορά στο μέσο ποσοστό αναγνώρισης, στον ρυθμό πιθανότητας αποτυχίας (miss probability rate) καθώς και στον ρυθμό εσφαλμένων συναγερμών (false alarm rate).

5.1. Εισαγωγή

Τα τελευταία χρόνια τα αυτόματα συστήματα που εποπτεύουν ανθρώπινες καθημερινές δραστηριότητες γίνονται όλο και πιο κοινά (Park et. al., 2008; Haritaoglou et. al., 2000). Ο στόχος αυτής της εργασίας είναι να συμβάλει στην αστική ασφάλεια προτείνοντας ένα αυτόματο ακουστικό σύστημα παρακολούθησης που ελέγχει τους δημόσιους χώρους για ενδεχομένως επικίνδυνες καταστάσεις. Αυτού τους είδους τα γεγονότα είναι πιθανό να εμπεριέχουν απειλή προς ανθρώπινη ζωή ή ζημιά ιδιοκτησίας (π.χ., πυροβολισμός, έκρηξη και ανθρώπινη αντίδραση σε αυτό το είδος κατάστασης) και συνεπάγονται συνήθως μια ισχυρή ακουστική εκπομπή. Αυτή η εργασία περιγράφει ένα πρακτικό σύστημα που εκμεταλλεύεται μόνο το λαμβανόμενο ακουστικό σήμα. Αυτή η μορφή πληροφορίας καταγράφεται με αρκετά οικονομικό τρόπο σε σχέση με άλλου είδους αισθητήρες (π.χ. υπέρυθρες και οπτικές κάμερες ή ανιχνευτές λέιζερ) και μπορεί να χρησιμοποιηθεί είτε αυτόνομα ως ένα σύστημα αναγνώρισης ακουστικών γεγονότων είτε σε μια διαδικασία μίξης της πιθανότητας των ανιχνευμένων γεγονότων μαζί με τα συμπληρωματικά στοιχεία άλλων αισθητήρων.

Το ερευνητικό πεδίο της ακουστικής επιτήρησης έχει κερδίσει πολύ προσοχή τα τελευταία χρόνια εξετάζοντας διάφορους τύπους εφαρμογών. Είναι ένας κλάδος της τεχνολογίας γενικευμένης αναγνώρισης ηχητικού σήματος, δηλαδή της υπολογιστικής ακουστικής ανάλυσης σκηνής. Αυτή η ιδιαίτερη περιοχή προσπαθεί να ερμηνεύσει τον περιβάλλοντα χώρο χρησιμοποιώντας αποκλειστικά τον εισερχόμενο ήχο, εμπνεόμενη από την αντίστοιχη ανθρώπινη ιδιότητα. Οι προηγούμενες προσπάθειες στο χώρο της ανίχνευσης μη-τυπικού ηχητικού γεγονότος εξετάζουν ένα ευρύ φάσμα των ακουστικών χαρακτηριστικών, τα οποία συνδυάζονται με διάφορες τεχνικές ταξινόμησης. Η προηγούμενη ερευνητική δουλειά πάνω στο θέμα απέχει αρκετά από την συγκεκριμενοποίηση μίας κοινής μεθοδολογίας όπως π.χ. στην περίπτωση της

αναγνώρισης ομιλίας/ομιλητή όπου ο ταξινομητής και η διαδικασία εξαγωγής χαρακτηριστικών έχουν λίγο πολύ καθιερωθεί (δηλ. GMMs και HMMs ως ταξινομητές και ως είσοδος χρησιμοποιούνται παραλλαγές διάφορων φασματικών χαρακτηριστικών). Η δυσκολία έγκειται στα εξής γεγονότα: α) μια μη-τυπική κατάσταση δεν είναι μια καλά καθορισμένη κατηγορία (π.χ. γέλιο vs. κλάματος vs. κραυγής), β) υπάρχουν πολλές περιπτώσεις όπου υπάρχει μια λεπτή γραμμή μεταξύ μιας τυπικής και μιας μη-τυπικής κατάστασης (π.χ. πυροβολισμός vs. έκρηξης) και το γ) το μικρόφωνο ενδέχεται να είναι τοποθετημένο μακριά από την πηγή του ακουστικού γεγονότος και επομένως, η αντήχηση καθώς και ακουστικά γεγονότα που ανήκουν σε μια σχεδόν απεριόριστη σειρά κατηγοριών μπορούν να γίνουν είσοδος του μικροφώνου. Οι προηγούμενες προσεγγίσεις στο χώρο της ακουστικής επίβλεψης εστιάζουν στις διαφορετικές πτυχές του ταξινομητή, της διαδικασίας εξαγωγής χαρακτηριστικών γνωρισμάτων, των δεδομένων εκπαίδευσης και του αριθμού των κατηγοριών. Οι συντάκτες της (Clavel et al., 2008) χρησιμοποίησαν ένα σύστημα αναγνώρισης συναισθήματος για φοβικές συναισθηματικές εκδηλώσεις που πραγματοποιούνται κατά τη διάρκεια μη-τυπικών καταστάσεων. Τα εξαγόμενα χαρακτηριστικά περιέγραψαν την προσωδία και την ακουστική ποιότητα και συνδυάστηκαν με φασματικές και cepstral παραμέτρους για να εκπαιδεύσουν διαφορετικά GMM για τα ηχηρά και τα άηχα ακουστικά μέρη. Η βάση δεδομένων τους βασίστηκε σε ταινίες επιστημονικής φαντασίας και αποτελείται από επτά ώρες ηχητικών καταγραφών οι οποίες οργανώθηκαν σε 400 οπτικοακουστικές ακολουθίες (βάση δεδομένων SAFE). Ο στόχος κατηγοριοποίησης περιελάμβανε φόβος vs. ουδέτερη ομιλία και επιτυγχάνουν ρυθμό λάθους 30%. Οι (Valenzise et al, 2007) παρουσίασαν ένα σύστημα παρακολούθησης χώρου για ανίχνευση πυροβολισμών καθώς και εντοπισμό κραυγής σε μια δημόσια πλατεία. Σαράντα εννέα χαρακτηριστικά γνωρίσματα υπολογίστηκαν συνολικά και δόθηκαν ως είσοδος σε μια υβριδική μέθοδο επιλογής filter/wrapper. Η έξοδός του χρησιμοποιήθηκε για τη δημιουργία δύο παράλληλων GMMs για τον διαχωρισμό των κραυγών από το θόρυβο και των πυροβολισμών από το θόρυβο. Τα δεδομένα προήλθαν από τις ήχους ταινιών, το διαδίκτυο αλλά και από ανθρώπινες φωνές ενώ τα δείγματα θορύβου λήφθηκαν σε μια δημόσια πλατεία του Μιλάνου. Η τελική ακρίβεια ήταν 93% για false rejection rate 5% όταν ο λόγος σήματος προς θόρυβο είναι 10dB. Μια ενδιαφέρουσα εφαρμογή, ανίχνευση εγκλήματος μέσα σε ανελκυστήρες περιγράφηκε στην (Radhakrishnan et al., 2005). Η προσέγγισή τους στηρίχθηκε στην χρονική ανάλυση και κατάτμηση των σημάτων. Με την αυτόματη συγκέντρωση των ακουστικών δεδομένων, ανακαλύφθηκαν συνεπή πρότυπα και τα δεδομένα που συλλέχθηκαν χρησιμοποιήθηκαν για την κατάρτιση ενός GMM για καθεμία εκ των οκτώ κατηγοριών χρησιμοποιώντας χαρακτηριστικά χαμηλού επιπέδου.

αναφορά	Άτυπη κατηγορία ήχου	Προσαρμογή μοντέλων	Περιβάλλον	Ταξινομητής	Χαρακτηριστικά	Βάση δεδομένων
προτεινόμενη προσέγγιση	Scream, gunshot and explosion	*MAP* adaptation of GMMs	Metro station, urban and military setting	GMM	MFCC, MPEG-7, CB-TEO, Intonation	Large audio corpora from professional sound
Clavel et al.	Fear-type emotions	-	-	GMM	Prosody, audio quality, spectral,	SAFE corpus
Valenzise et al.	Scream and gunshot	-	Public square	GMM	Temporal, spectral, cepstral, correlation	Movie soundtracks, internet and people
Radhakrishnan et al.	Banging and non-neutral speech	-	Elevator	GMM	MFCC	Elevator recordings
Clavel et al.	Gunshot	-	Public space	GMM	MFCC, spectral moments	CDs for the national French public radio
Harma et al.	Interesting events in an office	Noise spectrum update	Office	Threshold, clustering	Temporal, spectral	Recordings from an office room
Rouas et al.	Shout	Adaptive threshold for sound activity detection	Railway	GMM, SVM	Energy, MFCC	Recorded during 4 scenarios
Vacher et al.	Scream and glass break	-	Apartment	GMM	Wavelet based cepstral coefficients	Laboratory recordings and RCWP
Atrey et al.	Shout	-	Office corridor	GMM	ZCR, LPC, LPCC, LFCC	Recorded in office

Πίνακας 5.1: Ερευνητικές προσεγγίσεις στο χώρο της ακουστικής παρακολούθησης

Τα δεδομένα τους περιείχαν καταγραφές ύποπτων δραστηριοτήτων σε ανελκυστήρες καθώς και μερικά δείγματα χωρίς γεγονότα ενώ αναφέρουν την ανίχνευση όλων των ύποπτων δραστηριοτήτων χωρίς καμία αποτυχία. Μια μέθοδος ανίχνευσης πυροβολισμού κάτω από θορυβώδη περιβάλλοντα αναλύεται στην (Clavel et al., 2005). Τα δεδομένα τους αποτελούνται από τα στοιχεία που δημιουργήθηκαν τεχνητά χρησιμοποιώντας ένα σύνολο διάφορων δημόσιων χώρων και ηχητικών γεγονότων πυροβολισμού που εξήχθησαν από το εθνικό γαλλικό δημόσιο ραδιόφωνο. Ευρέως χρησιμοποιημένα χαρακτηριστικά γνωρίσματα υιοθετήθηκαν, συμπεριλαμβανομένου των MFCC για την κατασκευή δύο GMMs που

αντιπροσωπεύουν την κατηγορία του πυροβολισμού και την κανονική κατηγορία χρησιμοποιώντας δεδομένα διαφόρων επιπέδων SNR. Τα οφέλη της εποπτευομένης και ανεπίβλεπτης ομαδοποίησης (supervised και unsupervised clustering) για την ακουστική επιτήρηση σε ένα τυπικό περιβάλλον γραφείου περιγράφονται στην (Harma et al., 2005). Το σύστημά τους βασίστηκε σε ένα συνεχώς ανανεούμενο φασματικό προφίλ παρασιτικού θορύβου που εξυπηρέτησε την ανίχνευση των γεγονότων ενδιαφέροντος. Και η μεθοδολογία των κ-μέσων και η χειρωνακτική επιλογή κέντρων των συστάδων χρησιμοποιήθηκαν για τη συγκέντρωση των ακουστικών αρχείων που ηχογρφήθηκαν σε ένα τυπικό δωμάτιο γραφείου για μια περίοδο 48 ημερών. Η ανίχνευση στηρίχθηκε σε δύο εναλλακτικά κριτήρια κάθε ένα από τα οποία έθετε ένα κατώτατο όριο επάνω σε δύο ποσότητες που είχαν ως σκοπό να ανιχνεύσουν δυνατά onsets και transients στο περιβάλλον. Στην (Rouas et al., 2006) εξετάζεται το ζήτημα της ανίχνευσης των ακουστικών γεγονότων σε οχήματα δημόσιων συγκοινωνιών ενώ γίνεται χρήση μιας παραγωγικής και μιας διαχωριστικής μεθόδου. Τα ακουστικά δεδομένα καταγράφηκαν χρησιμοποιώντας 4 μικρόφωνα κατά τη διάρκεια τεσσάρων διαφορετικών σεναρίων που περιέλαβαν τις σκηνές πάλης, μια βίαιη σκηνή ληστείας και τις σκηνές αρπαγής τσάντας ή κινητών τηλεφώνων. Χρησιμοποίησαν GMM και SVM ενώ το σύνολο χαρακτηριστικών γνωρισμάτων τους διαμορφώθηκε από τα πρώτα 12 MFCC, την ενέργεια και τις αντίστοιχες παραγώγους και επιταχύνσεις. Οι (Vacher et al., 2004) παρουσίασαν ένα πλαίσιο για την ανίχνευση ήχων και την ταξινόμηση τους κάτω από το πρίσμα της τηλεϊατρικής. Η βάση τους αποτελείται από τις καταγραφές που έγιναν στο εργαστήριο CLIPS και αρχεία της RCWP - Ιαπωνία2. (2 http://tosa.mri.co.jp/sounddb/indexe.htm.)

Χρησιμοποίησαν cepstral συντελεστές βασισμένοι στο πεδίο wavelet για να εκπαιδεύσουν GMMs για οκτώ ηχητικές κατηγορίες ενώ το σύστημά τους αξιολογήθηκε κάτω από διαφορετικούς λόγους σήματος προς θόρυβο. Τελευταίο αναφέρουμε, ένα ιεραρχικό σχέδιο ταξινόμησης που κατηγοριοποιούσε κανονικά και συναισθηματικά ηχητικά γεγονότα (Atrey et al., 2006).

Οι συντάκτες της χρησιμοποίησαν τέσσερα ακουστικά γνωρίσματα για την δημιουργία GMMs, καθένα από τα οποία αντιπροσωπεύει έναν κόμβο του δέντρου ταξινόμησης. Οι ήχοι καταγράφηκαν για περίπου δύο ώρες σε πραγματικό περιβάλλον (διάδρομος γραφείου) και συμπεριελάμβαναν συζήτηση, κραυγή, κτύπο και τα βάδισμα.

Κατά την άποψή μας, οι προηγούμενες προσεγγίσεις εστιάζουν στις διαφορετικές πτυχές της ταξινόμησης των γενικών ακουστικών γεγονότων προσπαθώντας να βελτιστοποιήσουν ένα συγκεκριμένο μέρος του προβλήματος και είναι συνήθως βασισμένα σε εργαστηριακά πειράματα, δηλ. ηχογραφημένες εκ των προτέρων και καλά καθορισμένες κατηγορίες παρουσιάζονται σε έναν αλγόριθμο ταξινόμησης. Η προσέγγιση μας στοχεύει σε ένα πρακτικό σύστημα που λειτουργεί σε πραγματικό χώρο, σε 24/7 βάση και δεν χρησιμοποιεί απομονωμένα γεγονότα αλλά μία συνεχή ροή ακουστικών γεγονότων όπως στην περίπτωση της πραγματικής ζωής. Η έμφασή μας

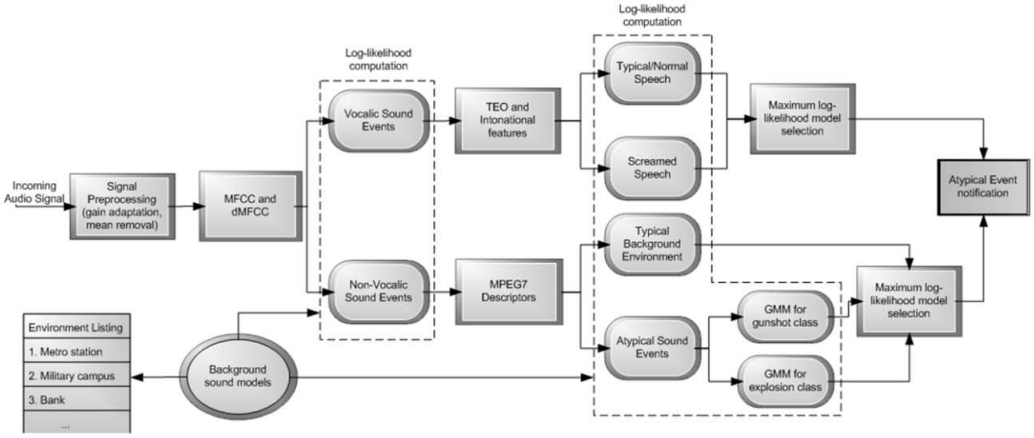
τοποθετείται πάνω στην κατασκευή ενός ολοκληρωμένου ακουστικού συστήματος παρακολούθησης που είναι αυτοπροσαρμοζόμενο σε διαφορετικά ακουστικά περιβάλλοντα (π.χ. σταθμός μετρό, αστικό περιβάλλον κ.λπ.). Η συγκεκριμένη εργασία στρέφεται στις μη-τυπικές καταστάσεις που χαρακτηρίζονται από τα συγκεκριμένα ηχητικά γεγονότα - κραυγές, εκρήξεις και πυροβολισμοί - που πραγματοποιούνται σε α) σταθμό μετρό, β) αστικό περιβάλλον και γ) ένα περιβάλλον που ταιριάζει σε στρατιωτικές εφαρμογές. Επιπλέον ειδική μέριμνα έχει ληφθεί ώστε το σύνολο των δεδομένων μας να είναι λεπτομερές και περιεκτικό, κάτι που έγινε δυνατό με τον συνδυασμό πολλών και καλά τεκμηριωμένων επαγγελματικών βιβλιοθηκών ήχων, οι οποίες περιέχουν σήματα υψηλής ποιότητας. Παρέχουμε μια λεπτομερή διερεύνηση για τα χαρακτηριστικά γνωρίσματα με στόχο την ανίχνευση τριών τύπων μη τυπικών ηχητικών γεγονότων καθώς και για την λειτουργία της αυτο-προσαρμογής με την οποία επανεκπαιδεύονται τα μοντέλα του συστήματος κατά τη διάρκεια της λειτουργίας του.

Ο Πίνακας 5.1 παρέχει μια σύγκριση των πτυχών που εξετάζονται σε αυτήν την εργασία με διάφορες υπάρχουσες προσεγγίσεις. Η ομιλία με κραυγές (screamed speech) είναι ένας φωνητικός ήχος άμεσα συσχετιζόμενος με τα ανθρώπινα αρνητικά συναισθήματα (π.χ. φόβος, πόνος, θυμός) και η ανίχνευσή της μπορεί να βοηθήσει στην ελαχιστοποίηση του κόστους ή ακόμα και στην αποφυγή επικίνδυνων καταστάσεων. Η μεθοδολογία μας εξετάστηκε σε ένα εσωτερικό χώρο όπου τυπικές και μη-τυπικές καταστάσεις αναπαράχθηκαν με τυχαίο τρόπο για τρεις διαδοχικές ημέρες μέσω των μεγάφωνων ενός Η/Υ ενώ ένας άλλος Η/Υ ανέλυε τον εκπεμπόμενο ήχο κάθε 2 δευτερόλεπτα. Το μέρος εξαγωγής χαρακτηριστικών γνωρισμάτων είναι γραμμένο σε Matlab© ενώ το μέρος ταξινόμησης σε C++. Το σύστημα είναι σχεδόν πραγματικού - χρόνου δεδομένου ότι αναφέρει ένα γεγονός με μια μέση καθυστέρηση 2 δευτερολέπτων. Για τον τύπο της εφαρμογής που ερευνάμε αυτή η καθυστέρηση δεν είναι κρίσιμη και, επιπλέον, η καθυστέρηση επεξεργασίας μπορεί να μειωθεί περαιτέρω δεδομένου ότι ο κώδικας δεν είναι βελτιστοποιημένος όσον αφορά στη διαδικασία εξαγωγής γνωρισμάτων.

Το υπόλοιπο αυτού του κεφαλαίου οργανώνεται ως εξής: στην ενότητα 5.2 δίνεται μια πλήρης επισκόπηση του συστήματος μαζί με μια σύντομη περιγραφή όλων των συνόλων ηχητικών παραμέτρων. Οι ενότητες 5.3 και 5.4 εξηγούν τις πειραματικές διαδικασίες και εκθέτουν λεπτομερώς τα αποτελέσματα ανίχνευσης που επιτυγχάνονται σε διαφορετικά επίπεδα SNR. Τα συμπεράσματά μας συνάγονται στο τελευταίο μέρος του κεφαλαίου.

5.2. Γενική αρχιτεκτονική του συστήματος

Ο κύριος σκοπός του συστήματός μας είναι να ανιχνευθούν οι ανθρώπινες φωνητικές αντιδράσεις (δηλ. κραυγές, εκφράσεις του πόνου) και τα μη-φωνητικά μη-τυπικά γεγονότα (πυροβολισμοί και εκρήξεις) που συνδέονται με επικίνδυνες καταστάσεις. Για αυτόν τον λόγο σχεδιάστηκε η δομή όπως απεικονίζεται στο Σχήμα 5.1. Υιοθετούμε μια ιεραρχία τριών διακριτών σταδίων, όπου κάθε στάδιο εξαρτάται από το προηγούμενο, για την επεξεργασία της εισερχόμενης ακουστικής ακολουθίας προτού να καθοριστεί η κατηγορία της. Εν συντομία, μετά από την προεπεξεργασία και την εξαγωγή των χαρακτηριστικών, ο ήχος ταξινομείται ως φωνητικό (κανονική ή κραυγή) ή μη-φωνητικό (περιβάλλον, πυροβολισμός ή έκρηξη) γεγονός. Με βάση την παρούσα απόφαση μια διαφορετική πορεία επιλέγεται για να χαρακτηρίσει περαιτέρω το ακουστικό σήμα. Σε περίπτωση που εντοπίζεται φωνητικό γεγονός, υπολογίζεται ένα διαφορετικό σύνολο περιγραφέων και ο ήχος ταξινομείται ως κανονική ομιλία ή κραυγή.



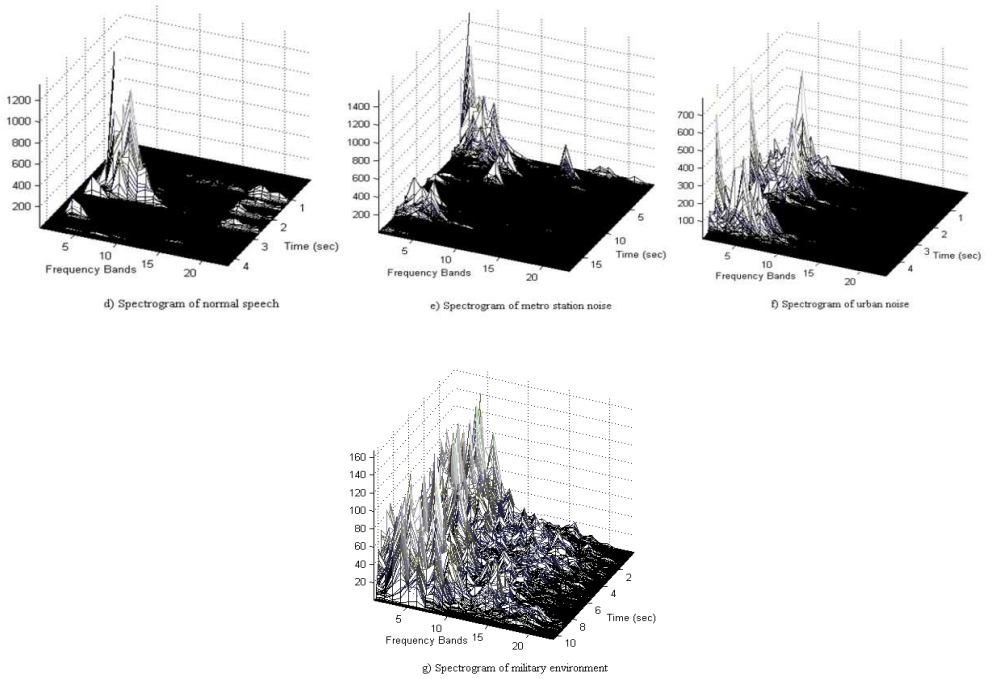
Σχήμα 5.1: Σχηματικό διάγραμμα του συστήματος ακουστικής παρακολούθησης.

Στην περίπτωση του μη-φωνητικού γεγονότος, μια πρόσθετη φάση εξαγωγής χαρακτηριστικών ακολουθεί και το σήμα ταξινομείται ως περιβαλλοντικός θόρυβος ή ως μη-τυπικό ηχητικό γεγονός, ενώ κατά τη διάρκεια του τρίτου σταδίου το σύστημα προχωρά στη διευκρίνιση του τύπου της επικίνδυνης κατάστασης. Οι κανονικές καταστάσεις καλύπτουν όλα εκείνες τις καταστάσεις που δεν εμπεριέχουν τον κίνδυνο ζωής ή/και ιδιοκτησίας.

5.2.1. Ανάλυση της εξαγωγής χαρακτηριστικών

Σε αυτήν την παράγραφο εξηγούμε τα χαρακτηριστικά χαμηλού επιπέδου που εξάγονται από τα ακουστικά σήματα για την κατασκευή των στατιστικών προτύπων που αντιπροσωπεύουν την *a-priori* γνώση που έχουμε για τις ακουστικές κατηγορίες. Επιλέχτηκαν επειδή συλλαμβάνουν διαφορετικές πτυχές της ακουστικής δομής.

Επιπλέον δεν είναι πολύ ευαίσθητα στις διάφορες συνθήκες θορύβου, όπως η ενέργεια ή η ηχηρότητα. Η πρώτη φάση κάνει διακρίσεις μεταξύ των φωνητικών και μη-φωνητικών γεγονότων, κατά συνέπεια χρησιμοποιήσαμε τους συντελεστές MFCC που παρέχουν μια συμπαγή περιγραφή για το πώς η ενέργεια διανέμεται πάνω στις μπάντες συχνοτήτων. Στη συνέχεια, τα φωνητικά γεγονότα χαρακτηρίζονται για το εάν είναι μη-τυπικά χρησιμοποιώντας την ανάλυση Teager energy operator (TEO) autocorrelation envelope area (Zhoun et al, 2001), την θεμελιώδη συχνότητα και την αναλογία αρμονικότητας: προς θόρυβο (Harmonicity to Noise Ratio - HNR).



Σχήμα 5.2: Αντιπροσωπευτικά σχήματα των περιβαλλοντικών θορύβων και μη-τυπικών ηχητικών γεγονότων στο φασματικό πεδίο Mel.

Είναι ενδεικτικά των παραλλαγών που επιδεικνύει η προσωδία στη περίπτωση της μη-τυπικής ομιλίας. Τα μη-φωνητικά γεγονότα υποβάλλονται σε επεξεργασία με τον υπολογισμό των Waveform Min, Waveform Max, Audio Fundamental Frequency και Audio Spectrum Flatness όπως καθορίζονται από το πρότυπο MPEG-7 (Quackenbush et al., 2001).

Πρόβλημα κατηγοριοποίησης	Αριθμός Γκαουσσιανών	Ομάδα χαρακτηριστικών	Ποσοστό αναγνώρισης (%)
Φωνητικά vs. Μη- φωνητικά ηχητικά γεγονότα (σταθμός μετρό)	64	MFCC+dMFCC	100
Φωνητικά vs. Μη- φωνητικά ηχητικά γεγονότα (αστικό περιβάλλον)	128	MFCC+dMFCC	99.85
Φωνητικά vs. Μη- φωνητικά ηχητικά γεγονότα (στρατιωτικό περιβάλλον)	128	MFCC+dMFCC +MPEG-7 LLDs	100
Τυπικά vs. Μη τυπικά μη φωνητικά ηχητικά γεγονότα (σταθμός μετρό)	128	MFCC+dMFCC +MPEG-7 LLDs	97.2 (87.6)
Τυπικά vs. Μη τυπικά μη φωνητικά ηχητικά γεγονότα (αστικό περιβάλλον)	128	MFCC+dMFCC +MPEG-7 LLDs	92.95 (88.2)
Τυπικά vs. Μη τυπικά μη φωνητικά ηχητικά γεγονότα (στρατιωτικό περιβάλλον)	32	MFCC+dMFCC +MPEG-7 LLDs	100 (91.6)
Εκρηξη vs. Πυροβολισμός	512	MFCC+dMFCC +MPEG-7 LLDs	83.9 (76.4)
Τυπική vs. Μη τυπική Ομιλία	128	MFCC+dMFCC +Intonation +CB-TEO-Auto-Env	100 (89.1)

Πίνακας 5.2: Μέσα ποσοστά αναγνώρισης που επιτυγχάνονται για κάθε στάδιο της τοπολογίας του συστήματος για διαφορετικά είδη περιβαλλοντικού θορύβου. Το σκορ χωρίς το επιπρόσθετο στάδιο εξαγωγής χαρακτηριστικών δίνεται σε παρένθεση για λόγους σύγκρισης

Αυτά συλλαμβάνουν τη μορφή του σήματος στο πεδίο του χρόνου, την περιοδικότητα καθώς και το πόσο επίπεδο είναι το φάσμα στις διαφορετικές ζώνες. Δεδομένου ότι προσπαθούμε να τοποθετήσουμε στο χρόνο συγκεκριμένα ηχητικά γεγονότα πειραματιστήκαμε με μεγαλύτερα μεγέθη πλαισίων από αυτά που χρησιμοποιούνται συνήθως (10-30ms) για την αναγνώριση ομιλίας/ομιλητών. Βασισμένοι στο κριτήριο του υψηλότερου ποσοστού αναγνώρισης και μετά από εκτενείς πειραματισμούς καταλήξαμε σε πλαίσια των 200ms με επικάλυψη 75%. Στη συνέχεια εξηγούμε εν συντομία τις διαδικασίες εξαγωγής των ακουστικών γνωρισμάτων.

Mel Frequency Cepstral Coefficients

Αρχικά υπολογίζεται η ενέργεια του STFT που συνδέεται με κάθε πλαίσιο η οποία έπειτα φιλτράρεται χρησιμοποιώντας την κλίμακα Mel που μικραίνει τον αριθμό των διαστάσεων και προσδίδει έμφαση τις φασματικές ζώνες που είναι σημαντικές όσον αφορά στην ανθρώπινη αντίληψη. Στη συνέχεια εφαρμόζεται ο λογάριθμος πάνω

στους φασματικούς συντελεστές και τελικά χρησιμοποιείται ο DCT για την αποσυσχέτιση τους. Διατηρούμε τα 13 σημαντικότερα διανύσματα, συμπεριλαμβανομένου του μηδενικού που εκφράζει την ενέργεια του σήματος. Επιπλέον, υιοθετήσαμε την τεχνική cepstral μέσης κανονικοποίησης (cepstral mean normalization) σε όλα τα πειράματα. Ο ρυθμός με τον οποίο αλλάζουν με την πάροδο του χρόνου υπολογίζεται επίσης και οδηγεί σε ένα διάνυσμα χαρακτηριστικών γνωρισμάτων με 26 διαστάσεις. Τα MFCCs χρησιμοποιούνται μόνο για τη διάκριση των φωνητικών γεγονότων από τα μη-φωνητικά καθώς επίσης και κατά τη διάρκεια των υπόλοιπων φάσεων της τοπολογίας του συστήματος συνδυαζόμενα με τους περιγραφείς που εξηγούνται στη συνέχεια. Επίσης χρησιμοποιήθηκαν οι ακουστικοί περιγραφείς του πρωτοκόλλου MPEG-7 που περιγράφηκαν στο προηγούμενο κεφάλαιο. Συγκριτικά αποτελέσματα που έχουν να κάνουν με την προσθήκη αυτών των ομάδων παραμέτρων απεικονίζονται στον Πίνακα 5.2.

Παράμετροι βασισμένες στον επιτονισμό και στον Teager Energy Operator

Η ανάλυση TEO πραγματοποιείται σε 16 κρίσιμες ζώνες. Τα φίλτρα Gabor χρησιμοποιούνται για να εστιάσουν σε μια συγκεκριμένη φασματική περιοχή ενώ ο συντελεστής TEO της καθεμίας αφορά πλαίσια των 200ms με επικάλυψη 75%. Στη συνέχεια το εμβαδόν των συναρτήσεων αυτοσυσχέτισης υπολογίζεται και ομαλοποιείται διαιρούμενη από το μισό του μήκους των πλαισίων. Το διάνυσμα χαρακτηριστικών γνωρισμάτων έχει 16 συντελεστές όπως και ο αριθμός των κρίσιμων ζωνών. Η ακουστική ανάλυση που στηρίζεται στον ενεργειακό χειριστή Teager μπορεί να αποκαλύψει πτυχές των λεκτικών ή μη λεκτικών ανθρώπινων αντιδράσεων που δεν συλλαμβάνονται από τα MFCC και συσχετίζονται με την έκφραση πίεσης (stress). Έχει αποδειχθεί ότι είναι ενδεικτικός των αλλαγών του τρόπου με τον οποίο μεταβάλλεται η ροή του αέρα σχετικά με τη παραγωγή ομιλίας κάτω από μη-τυπικές περιστάσεις. Αυτοί οι συντελεστές επισυνάπτονται στη θεμελιώδη συχνότητα, στη παράγωγο της και στο HNR (βασισμένα στην ανάλυση cross correlation) και απεικονίζουν τις παραλλαγές που υφίσταται η προσωδία σχετικά με τη τυπική και τη μη-τυπική ομιλία. Μαζί με τα ήδη υπολογισμένα MFCC διαμορφώνουν ένα διάνυσμα για τη διάκριση μεταξύ των τυπικών και μη-τυπικών φωνητικών ακουστικών γεγονότων. Για τον υπολογισμό τους χρησιμοποιήσαμε το λογισμικό PRAAT[3] του οποίου η λειτουργία είναι βελτιστοποιημένη για σήματα ομιλίας.

5.2.2. Διαδικασία αναγνώρισης προτύπων

Χρησιμοποιήσαμε διαγώνια GMM για τη διαμόρφωση της κατανομής της κάθε ηχητικής κατηγορίας. Είναι βασισμένα στην υπόθεση ότι η κατανομή των στοιχείων που ανήκουν σε κάθε κατηγορία μπορεί να περιγραφεί στατιστικά από έναν γραμμικό συνδυασμό Γκαουσσιανών συναρτήσεων. Η υλοποίηση μας βασίστηκε σε αυτή της βιβλιοθήκης μηχανικής μάθησης Torch[4], η οποία είναι γραμμένη σε C++. Ο μέγιστος

αριθμός επαναλήψεων του k-means για την έναρξη ήταν 50 και ο αλγόριθμος EM είχε και ανώτερο όριο 25 επαναλήψεων συνδυασμένο με ένα κατώτατο όριο 0.001 όσον αφορά στη διαφορά μεταξύ δύο διαδοχικών επαναλήψεων (κριτήριο ολοκλήρωσης). Τα πιθανοτικά πρότυπα αποθηκεύτηκαν και στη συνέχεια χρησιμοποιήθηκαν για τον καθορισμό της log-likelihood. Αυτή μας δείχνει το πόσο πιθανό είναι το συγκεκριμένο δείγμα να παρήχθη από το συγκεκριμένο μοντέλο. Τελικά υπολογίστηκε αυτό το είδος αποτελέσματος όσον αφορά κάθε πρότυπο και η προέλευση του ηχητικού δείγματος προσδιορίστηκε επιλέγοντας το πρότυπο με τη μέγιστη log-likelihood.

5.3. Περιγραφή της πειραματικής διαδικασίας

Αυτό το μέρος περιέχει την περιγραφή της οργάνωσης των πειραμάτων που πραγματοποιήθηκαν. Στην παράγραφο 5.3.1, εξηγούμε τις δοκιμές ταξινόμησης σχετικά με κάθε στάδιο του προτεινόμενου συστήματος όσον αφορά τον αριθμό των Γκαουσσιανών συναρτήσεων που προσφέρει το υψηλότερο ποσοστό αναγνώρισης. Στο μέρος 5.3.2 το σύστημα που ενσωματώνει τα μοντέλα που δημιουργήθηκαν προηγουμένως αξιολογήθηκε ως προς τους εσφαλμένους συναγερμούς και τις λάθος ανιχνεύσεις που επιδεικνύει μέσω ενός τεχνητού πειράματος για διάφορους λόγους SNR. Τα μη-τυπικά ηχητικά γεγονότα συγχωνεύθηκαν τυχαία με τον παρασιτικό θόρυβο και εξετάστηκαν για ανίχνευση. Η τελευταία πειραματική φάση, που συζητείται στην Παράγραφο 5.4 στόχευσε στη μίμηση μιας τρέχουσας άτυπης κατάστασης. Εξετάστηκε το σύστημα συμπεριλαμβανομένου του βρόχου ανατροφοδότησης που επιτρέπει την αναπροσαρμογή των μοντέλων επάνω στην ανίχνευση των προσομοιωμένων μη-τυπικών καθώς και των τυπικών καταστάσεων.

Τα μη-τυπικά ακουστικά δεδομένα συμπεριλαμβανομένων των ακραίων συναισθηματικών εκδηλώσεων και των ανώμαλων ηχητικών γεγονότων δεν είναι δημόσια διαθέσιμα λόγω του ιδιωτικού χαρακτήρα τους, της έλλειψης τους και της μη προβλεψιμότητάς τους (Clavel et al., 2004). Τα δεδομένα που δείχνουν τις καταστροφικές καταστάσεις όσον αφορά στο ιδιαίτερο είδος κινδύνων που αναφέρονται σε αυτό το άρθρο προσδιορίστηκαν και απομονώθηκαν από επαγγελματικές ηχητικές συλλογές μεγάλης κλίμακας. Αυτοί οι τύποι συλλογών χρησιμοποιούνται κυρίως από τη κινηματογραφική βιομηχανία λόγω της απέραντης ποικιλίας τους, της ογκώδης ποσότητας τους και της υψηλής τους ποιότητας. Η ακουστική ροή μιας κινηματογραφικής ταινίας (π.χ. βήματα, ήχοι πόρτας κ.λπ.) δεν είναι απαραίτητα αυτή που καταγράφηκε πραγματικά επί τόπου αλλά μπορεί να είναι κάποια διαφορετική η οποία να έχει αντικαταστήσει την πραγματική. Κατά συνέπεια, υπάρχει τεράστιος όγκος δεδομένων με σχεδόν οποιοδήποτε είδους ακουστικών γεγονότων συμπεριλαμβανομένων των μη-φωνητικών καθώς επίσης και των φωνητικών μη τυπικών αντιδράσεων για την κατάρτιση στατιστικών προτύπων αναγνώρισης. Η τελική βάση συγκροτήθηκε από τις ακόλουθες συλλογές: (i) BBC

Sound Effects Library, (ii) Sound Ideas Series 6000, (iii) Sound Ideas: the art of Foley, (iv) Best Service Studio Box Sound Effects, (v) TIMIT και (vi) ήχοι από διάφορες πηγές του διαδικτύου. Με τη χρησιμοποίηση αυτών των συνόλων δεδομένων ταυτόχρονα ένας υψηλός βαθμός παραλλαγής καθώς επίσης και ποικιλομορφίας σχετικά με την οντότητα των ακουστικών κατηγοριών ενσωματώθηκε στα πρότυπα.

3 http://www.praat.org

4 http://www.torch.ch.

Κατηγορία ήχων	**Αριθμός ηχογραφήσεων**	**Μέση διάρκεια (s)**
Εκρηξη	131	13.77
Πυροβολισμός	187	32.94
Κραυγή	270	4.04
Τυπική ομιλία	1680	3.08
Σταθμός μετρό	32	44.88
Αστικό περιβάλλον	106	83.35
Στρατιωτικό περιβάλλον	31	68.69
Σύνολο	2437	35.82

Πίνακας 5.3: Η βάση των ηχητικών δεδομένων

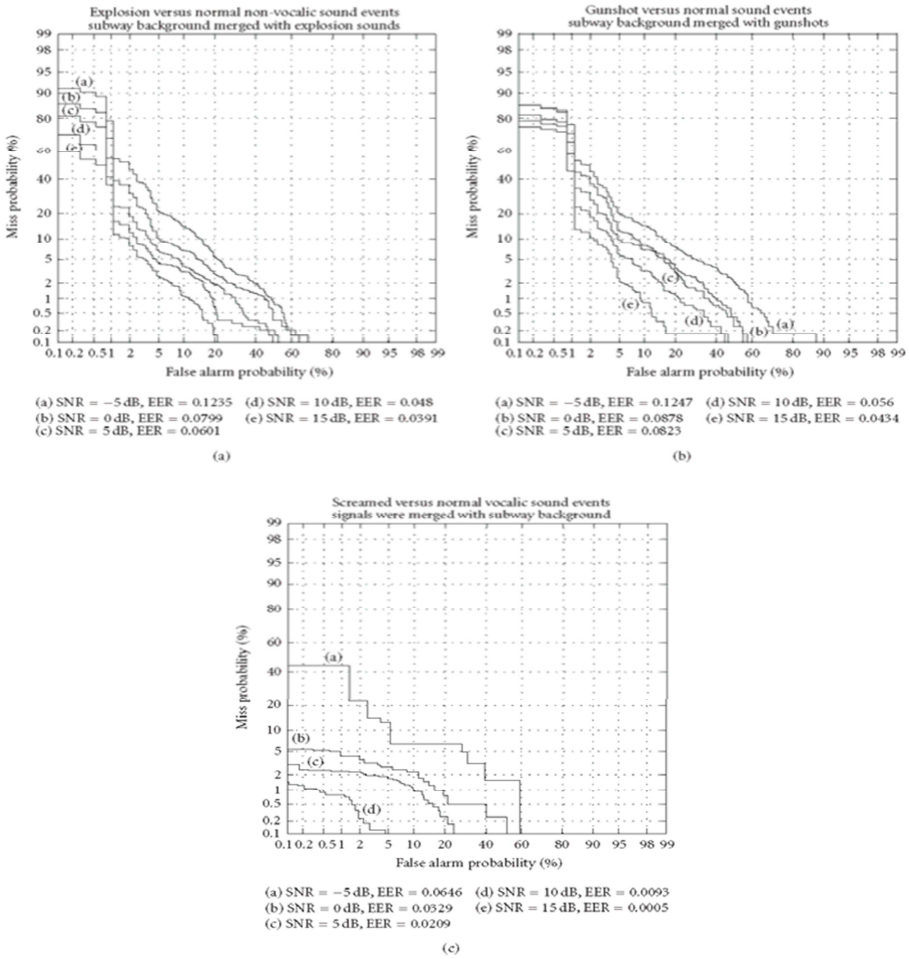
Ο ρυθμός δειγματοληψίας όλων των ηχητικών δειγμάτων ήταν 16 KHz με δεκαεξάμπιτη ανάλυση ενώ η μέση διάρκεια για κάθε κατηγορία αλλά και συνολικά δίνεται στον Πίνακα 5.3. Στο Σχήμα 5.2 φαίνονται φασματογραφήματα από αντιπροσωπευτικά δείγματα κάθε κατηγορίας. Τα μη τυπικά ηχητικά γεγονότα συγχωνεύθηκαν τεχνητά με τον ιδιαίτερα μη στατικό περιβαλλοντικό θόρυβο για τη μίμηση των ανώμαλων καταστάσεων και τη πραγματοποίηση των πειραμάτων ανίχνευσης. Οι λεπτομέρειες σχετικά με την αξιολόγηση του συστήματος υπό τους διαφορετικούς λόγους SNR δίνονται στην επόμενη παράγραφο.

5.3.1. Δημιουργία στατιστικών μοντέλων και πειράματα αναγνώρισης

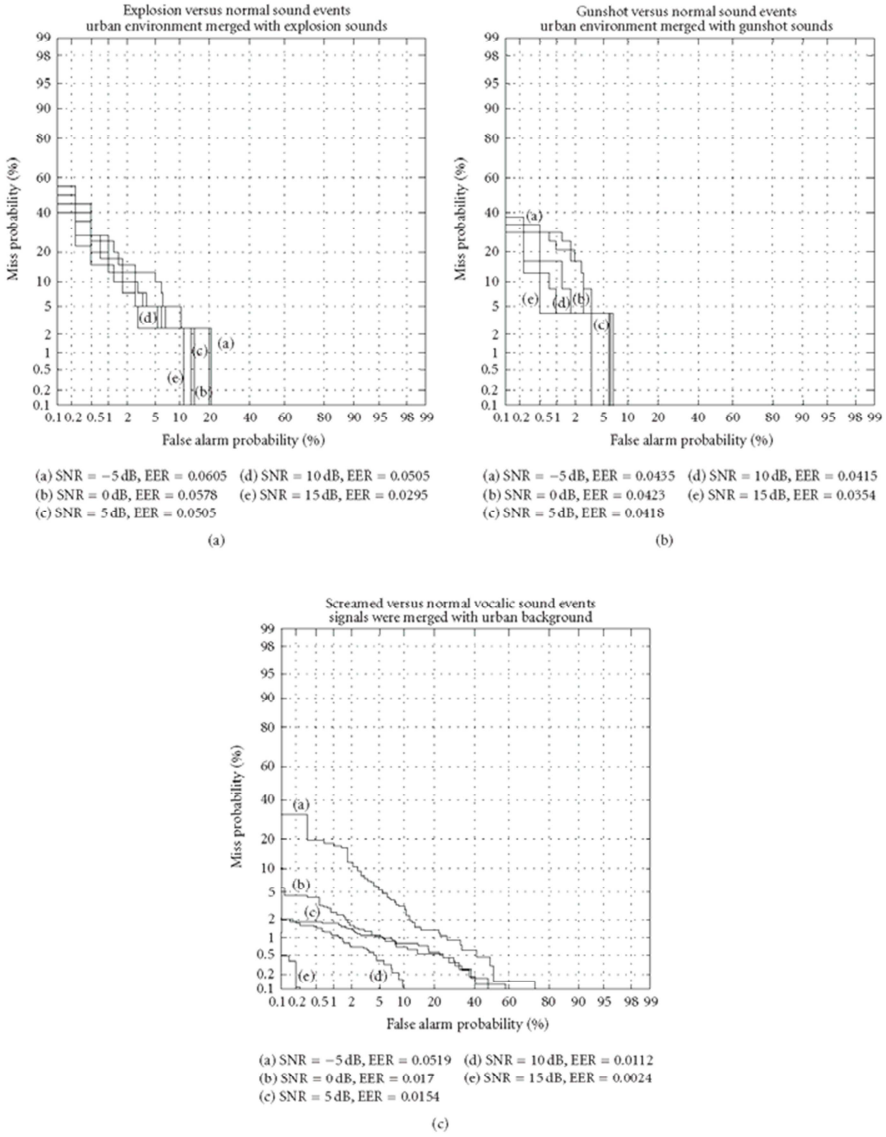
Χρησιμοποιήσαμε το 75% των δεδομένων που ανήκουν σε κάθε κατηγορία για την κατάρτιση καθενός στατιστικού προτύπου για να αντιπροσωπεύσει την αντίστοιχη ηχητική κατηγορία. Τα υπόλοιπα 25% χρησιμοποιήθηκαν για δοκιμή του συστήματος ενώ ο διαχωρισμός έγινε με τυχαίο τρόπο. Η ακουστική αναγνώριση προτύπων είναι βασισμένη στην ελλοχεύουσα υπόθεση ότι κάθε ακουστικό σήμα έχει έναν μοναδικό τρόπο με τον οποίο κατανέμει την ενέργειά του στις διαφορετικές ζώνες συχνοτήτων. Αυτό αποτελεί την αποκαλούμενη ακουστική υπογραφή του που μπορεί να αποκαλυφθεί και να προσδιοριστεί στη συνέχεια αυτόματα χρησιμοποιώντας στατιστικές τεχνικές ανάλυσης προτύπων. Ένα GMM με διαγώνια μήτρα συνδιακύμανσης χτίστηκε για κάθε κατηγορία ενώ η δοκιμή ουσιαστικά αποτελείται από μια απλή σύγκριση των log- likelihoods. Αρχικά δημιουργήθηκαν οι δύο τύποι προτύπων σύμφωνα με την τοπολογία

του συστήματος: *φωνητικό* (συμπεριλαμβανομένης της κανονικής και μη τυπικής ομιλίας) και *μη-φωνητικό* (συμπεριλαμβανομένης της έκρηξης, του πυροβολισμού και του αντίστοιχου ηχητικού περιβάλλοντος). Στη συνέχεια δημιουργήσαμε πρότυπα σχετικά με την *κανονική ομιλία, την μη τυπική ομιλία, το περιβάλλον θορύβου* και *τα μη-τυπικά ηχητικά γεγονότα* (συμπεριλαμβανομένης της *έκρηξης* και του *πυροβολισμού*). Μετά από εκτενείς πειραματισμούς σχετικά με τον αριθμό Γκαουσσιανών συναρτήσεων αλλά και τα σύνολα χαρακτηριστικών γνωρισμάτων που χρησιμοποιήθηκαν κατά τη διάρκεια κάθε προβλήματος ταξινόμησης, πετύχαμε τα ποσοστά αναγνώρισης που δίνονται στον Πίνακα 5.2 και τα οποία συνδέονται με κάθε στάδιο της αρχιτεκτονικής του συστήματος για όλα τα διαφορετικά είδη περιβάλλοντος. Κατά τη διάρκεια αυτής της πειραματικής φάσης ταξινομήσαμε κάθε ηχητικό δείγμα της βάσης (το μέρος του 25%) αθροίζοντας τις log-likelihoods που αποκτήσαμε από κάθε πρότυπο όσον αφορά όλα τα πλαίσια των συγκεκριμένων δειγμάτων. Τελικά το πρότυπο που παρουσίασε την υψηλότερη αθροισμένη log-likelihood επιλέχτηκε και η κατηγορία του χαρακτήρισε τον συγκεκριμένο ήχο.

Μπορούμε να πούμε ότι τα μέσα ποσοστά αναγνώρισης που επιτυγχάνονται κατά τη διάρκεια κάθε βήματος επεξεργασίας του συστήματος είναι σχετικά υψηλά και σε μερικές περιπτώσεις όπως τυπική vs. μη-τυπικής ομιλίας, το ποσοστό διάκρισης φθάνει το 100%. Στο συγκεκριμένο ποσοστό απεικονίζεται η ανομοιότητα σχετικά με τη φασματική/ενεργειακή κατανομή των δύο αυτών κατηγοριών αλλά και η ικανότητα των επιλεγμένων ακουστικών παραμέτρων να την καταγράψουν. Αυτό το είδος διαφοράς φαίνεται σαφέστατα στο Σχήμα 5.2 που περιλαμβάνει χαρακτηριστικά φασματογραφήματα για κάθε ακουστική κατηγορία. Η ταξινόμηση των εκρήξεων και των ηχητικών γεγονότων πυροβολισμού παρουσιάζει το χαμηλότερο ποσοστό αναγνώρισης που είναι 83.9%. Στο συγκεκριμένο στάδιο, τα λάθη εμφανίζονται λόγω της μεγάλης μεταβλητότητας/ποικιλίας μεταξύ των δειγμάτων της ίδιας κατηγορίας. Επιπλέον διάφορα ηχητικά δείγματα είναι παρόμοια ακόμα κι αν ανήκουν σε διαφορετικές κατηγορίες, κάτι που σημαίνει ότι κάποια έκρηξη ηχεί όπως ένας πυροβολισμούς και αντίστροφα. Η περαιτέρω ενσωμάτωση μιας σειράς ηχητικών περιγραφέων όχι μόνο δεν παρείχε καλύτερη απόδοση, αλλά αύξησε το υπολογιστικό κόστος. Μετά από τους εκτενείς πειραματισμούς, κατορθώσαμε να συλλάβουμε διακριτικές και χαρακτηριστικές πληροφορίες των ακουστικών κατηγοριών χρησιμοποιώντας ένα διάνυσμα χαρακτηριστικών γνωρισμάτων με σχετικά χαμηλό αριθμό διαστάσεων για κάθε φάση της τοπολογίας του συστήματος.



Σχήμα 5.3: Detection Error Tradeoff καμπύλες. Οι κατηγορίες στόχου είναι έκρηξη, πυροβολισμός, κραυγή και τυπική ομιλία. Ο background θόρυβος αντιστοιχεί στο περιβάλλον σταθμού μετρό



Σχήμα 5.4: Detection Error Tradeoff καμπύλες. Οι κατηγορίες στόχου είναι έκρηξη, πυροβολισμός, κραυγή και τυπική ομιλία. Ο background θόρυβος αντιστοιχεί σε αστικό περιβάλλον

Οι καμπύλες DET δείχνουν την ανίχνευση των μη-τυπικών καταστάσεων στο πλαίσιο τριών ειδών περιβαλλόντων και διαφορετικών όρων SNR. Πενήντα αντιπροσωπευτικά μη τυπικά ηχητικά γεγονότα επιλέχτηκαν από κάθε κατηγορία (κραυγή, έκρηξη και πυροβολισμός) τα οποία συγχωνεύθηκαν τεχνητά με ένα μέρος ίδιου μεγέθους τυχαία επιλεγμένο που ανήκει στη κανονική περιβαλλοντική ακουστική σκηνή. Αυτή η διαδικασία επαναλήφθηκε για κάθε μη τυπικό γεγονός 50 φορές για την παραγωγή αξιόπιστων αποτελεσμάτων.

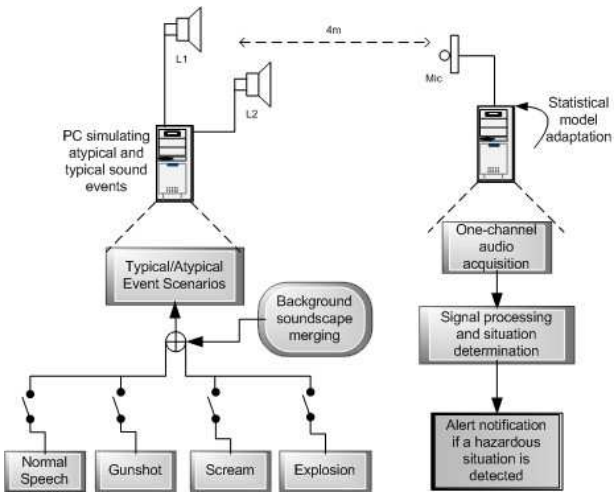
Τελικώς κάθε κατηγορία ανώμαλων καταστάσεων εξετάστηκε για τον προσδιορισμό της 2500 φορές (50x50). Δύο σειρές πειραμάτων πραγματοποιήθηκαν για την ανίχνευση μη τυπικών ακουστικών γεγονότων στο πλαίσιο του σταθμού μετρό και του αστικού περιβάλλοντος. Κάθε καταγραφή ομαλοποιείται με βάση τη μέγιστη αξία της (προσαρμογή κέρδους/gain adaptation). Οι καμπύλες DET και για τους δύο τύπους περιβάλλοντος παρουσιάζονται στα Σχήματα 6.3 και 6.4. Η log-likelihood που παρήχθη από το ανώμαλο μη-φωνητικό Γκαουσσιανό μίγμα υιοθετήθηκε στην περίπτωση ανίχνευσης της έκρηξης και του πυροβολισμού ενώ η log-likelihood που παρήχθη από το μη τυπικό φωνητικό μοντέλο χρησιμοποιήθηκε για την παραγωγή των DET καμπύλων σχετικά με το στάδιο κραυγή vs. τυπική ομιλία. Αυτές οι τιμές ομαλοποιήθηκαν κάθε φορά χρησιμοποιώντας την αντίστοιχη τυπική log-likelihood.

Στο Σχήμα 5.3 απεικονίζονται τα αποτελέσματα της ανίχνευσης μη τυπικών ηχητικών γεγονότων και για τις τρεις διαφορετικές ηχητικές κατηγορίες στο πλαίσιο του περιβάλλοντος σταθμών μετρό. Μια γρήγορη μείωση παρατηρείται όταν μειώνεται το SNR. Εντούτοις οι επείγουσες καταστάσεις ανιχνεύονται επαρκώς ακόμη και σε πολύ χαμηλές τιμές του SNR. Στην περίπτωση του -5dB SNR το μέσο EER όλων των ειδών γεγονότων είναι 8.29% ενώ το καλύτερο ποσοστό ανίχνευσης αφορά τα ανώμαλα φωνητικά ηχητικά γεγονότα με 6.46% EER. Αυτό είναι μια έκβαση της δομής της υλοποίησης μας, όπου κάθε στάδιο κάνει διακρίσεις ακουστικών σημάτων που έχουν τα διαφορετικά φασματικά πρότυπα και έχουν μόνο μερικά κοινά χαρακτηριστικά. Τα ακουστικά σήματα που είναι πιο τρωτά στην συγχώνευση παρασιτικού θορύβου είναι οι πυροβολισμοί με 12.47% EER σε SNR -5dB. Στην ενεργειακή αναλογία 0dB κατά την οποία αντιπροσωπεύονται ικανοποιητικά οι πραγματικές συνθήκες, το προτεινόμενο σύστημα καταδεικνύει υψηλή επίδοση με μέσο EER ίσο με 6.68%, το μέσο ποσοστό λάθος αναγνωρίσεων 16.4% και μέσο όρο εσφαλμένων συναγερμών για τα ανώμαλα γεγονότα 2.26%, κάτι που είναι μεγάλης σπουδαιότητας για συγκεκριμένο είδος εφαρμογών. Στο Σχήμα 5.4 παρουσιάζονται οι ικανότητες της υλοποίησής μας στο πλαίσιο του αστικού περιβάλλοντος. Σε αυτή τη φάση χρησιμοποιήσαμε τα στατιστικά πρότυπα που δημιουργήθηκαν με το συνυπολογισμό των αστικών ακουστικών στοιχείων. Όπως αναμένεται, η πιθανότητα λάθος ανιχνεύσεων πέφτει καθώς το SNR αυξάνεται από -5dB σε 15dB. Τα μη τυπικά ηχητικά γεγονότα ανιχνεύονται με σχετικά χαμηλά EERs σε όλες τις τιμές SNR όταν αλλοιώνεται το ακουστικό σήμα από τον αστικό περιβαλλοντικό θόρυβο. Παρατηρούμε ότι επιτυγχάνεται καλύτερη απόδοση με μέσο

EER ίσο με 5.19% για SNR -5db σε αντίθεση με το περιβάλλον μετρό. Ακριβέστερα, οι επείγουσες καταστάσεις σε SNR -5dB ανιχνεύονται με EERs 6.05%, 4.35% και 5.19% όπου η ανωμαλία αναφέρεται στα ηχητικά γεγονότα έκρηξης, πυροβολισμού και κραυγής αντίστοιχα. Τα γεγονότα που επηρεάζονται λιγότερο από τον παρασιτικό θόρυβο είναι οι ήχοι κραυγής ενώ η ανίχνευση έκρηξης παρουσιάζει τα υψηλότερα EERs σε όλους τους όρους SNR. Επιπλέον, η εφαρμογή μας παρέχει πολύ

χαμηλό ρυθμό εσφαλμένων συναγερμών με μέση τιμή 1% για τα τρία ανώμαλα ηχητικά γεγονότα σε SNR ίσο με 0dB ενώ το αντίστοιχο μέσο ποσοστό λάθος ανιχνεύσεων είναι 13.2%. Τα αντίστοιχα EERs που επιτυγχάνονται από το σύστημα όσον αφορά τις καταστάσεις έκρηξης, πυροβολισμού και κραυγής είναι 5.78%, 4.23% και 1.7% αντίστοιχα.

Οι αντίστοιχες καμπύλες DET που αφορούν στην περίπτωση ενός περιβάλλοντος που ταιριάζει σε στρατιωτικές εφαρμογές παρουσιάζονται στο Σχήμα 5.5. Όπως μπορεί να ελεγχθεί οπτικά, τα αποτελέσματα βελτιώνονται σημαντικά σε αυτήν την περίπτωση. Επιτυγχάνονται πολύ χαμηλά EERs όσον αφορά την ανίχνευση και των τριών διαφορετικών ειδών μη τυπικών καταστάσεων. Τα καλύτερα ποσοστά ανίχνευσης εμφανίζονται στην περίπτωση της κραυγής, κάτι το οποίο αναμενόταν λόγω της διαφορετικής δομής των συγκεκριμένων ακουστικών σημάτων έναντι των παραλλαγών που επιδεικνύει το συγκεκριμένο περιβάλλον. Οι ανώμαλες καταστάσεις ανιχνεύονται καλά ακόμη και σε χαμηλές τιμές του SNR. Πιο συγκεκριμένα στην περίπτωση των -5dB τα EERs είναι 1.93%, 5.24% και 1.03% για την έκρηξη, τον πυροβολισμό και την ανίχνευση γεγονότος κραυγής αντίστοιχα. Ένας σημαντικός στόχος ενός τέτοιου συστήματος είναι να περιοριστεί ο ρυθμός εσφαλμένων συναγερμών. Λήφθηκε ιδιαίτερη προσοχή σχετικά με αυτήν την πτυχή και η μέση πιθανότητα εσφαλμένων συναγερμών στην περίπτωση των 0dB SNR είναι 0.93% με τη μέση πιθανότητα λάθος ανιχνεύσεων ίση με 2.67% και για τα τρία είδη μη τυπικών καταστάσεων ενώ το μέσο EER είναι 2.24% (πρέπει να αναφερθεί ότι τα αντίστοιχα EERs για κάθε μη τυπικό ηχητικό γεγονός και στο πλαίσιο των τριών διαφορετικών τύπων περιβαλλόντων παρουσιάζεται κάτω από κάθε σχήμα).



Σχήμα 5.6: Διάγραμμα των πειραμάτων εντοπισμού και κατηγοριοποίησης σε πραγματικό εσωτερικό χώρο

Καταλήγουμε στο συμπέρασμα ότι τα αποτελέσματα που αναλύονται σε αυτό το τμήμα είναι πολύ ενθαρρυντικά και υπογραμμίζουν τη σημασία της επιλεγμένης

στατιστικής αρχιτεκτονικής στην οποία ενσωματώθηκαν ακουστικά γνωρίσματα τα οποία συλλαμβάνουν διαφορετικές πτυχές της ακουστικής δομής.

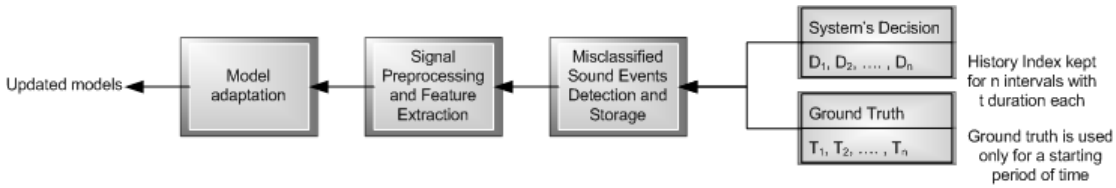
5.4. Πειράματα σε πραγματικούς εσωτερικούς χώρους

Ο βασικός σκοπός κατά τη διάρκεια της τρίτης πειραματικής φάσης ήταν να προσεγγίσουμε τις πραγματικές συνθήκες λειτουργίας και να αξιολογηθεί τη προσαρμογή των στατιστικών μοντέλων που παρέχεται μέσω του βρόχου ανατροφοδότησης (βλέπε Σχήμα 5.7). Οι μη τυπικές καταστάσεις δημιουργήθηκαν τεχνητά με τυχαίο τρόπο, όπως περιγράψαμε στην ενότητα 5.3. Στη συνέχεια αναπαράγονται μέσω μεγάφωνων ενώ ένα μικρόφωνο τοποθετήθηκε σε ένα άλλο μέρος ενός δωματίου διαστάσεων 6.75x4.9x3 με χρόνο αντήχησης 0.3 δευτερόλεπτα. Χρησιμοποιήσαμε δύο προσωπικούς υπολογιστές εκ των οποίων ο ένας αναπαρήγαγε τα ανώμαλα ηχητικά γεγονότα μέσω δύο συμβατικών μεγάφωνων σε προκαθορισμένες χρονικές στιγμές (ώστε να υπάρξει η *a-priori* γνώση της επισημείωσης) ενώ ο δεύτερος συνελάμβανε συνεχώς ακουστικά στοιχεία με ένα απλό μικρόφωνο. Στη συνέχεια αυτά τα δεδομένα υποβλήθηκαν σε επεξεργασία και ταξινομήθηκαν από το προτεινόμενο σύστημα. Αυτός το προσωπικός Η/Υ χρησιμοποιήθηκε επίσης για να πραγματοποιήσει και την επιβλεπόμενη καθώς και ανεπίβλεπτη προσαρμογή προτύπων. Ολόκληρη η οργάνωση του πειράματος απεικονίζεται στο Σχήμα 5.6.

5.4.1. Το πρόβλημα του κινούμενου παραθύρου

Διάφορα ζητήματα εμφανίστηκαν στο συγκεκριμένο τύπο πειράματος που περιλαμβάνουν τα κινούμενα παράθυρα αλλά και τη σπανιότητα με την οποία εμφανίζονται τα μη-τυπικά ηχητικά γεγονότα. Η παραθυροποίηση του εισερχόμενου ακουστικού σήματος σε κομμάτια ενός προκαθορισμένου μεγέθους δεν είναι επαρκής τρόπος ανάλυσης καθώς δε μπορεί να παρέχει ικανοποιητικά αποτελέσματα επειδή η διάρκεια των εκρήξεων, των πυροβολισμών και των κραυγών ποικίλει πολύ. Ούτε ο χρόνος έναρξης ούτε η διάρκεια ενός keysound γεγονότος είναι γνωστές στο σύστημα, κατά συνέπεια μπορεί το γεγονός ενδιαφέροντος να κοπεί σε μέρη που ανήκουν σε διαφορετικές ηχητικές κατηγορίες. Για αυτόν τον σκοπό αποφασίσαμε να επεξεργαστούμε το εισερχόμενο ακουστικό σήμα σε μια frame by frame ανάλυση (200ms με επικάλυψη 75%), ενώ τα διαδοχικά πλαίσια που καταχωρούνται στην ίδια κατηγορία συγχωνεύονται σε ένα τμήμα με το χρόνο έναρξης να αντιστοιχεί στο πρώτο πλαίσιο και τη διάρκεια στο συνολικό αριθμό των πλαισίων του τμήματος. Και οι δύο πληροφορίες αποθηκεύονται για τη βοήθεια πιθανής μελλοντικής έρευνας για τη σκηνή στην συγκεκριμένη χρονική περίοδο. Επίσης το άθροισμα των log-likelihoods ομαλοποιείται με βάση τον αριθμό πλαισίων κάθε ηχητικού γεγονότος. Επιπλέον, μια διαδικασία λείανσης εφαρμόστηκε στη συνέχεια για να αφαιρέσει τις αδικαιολόγητες

ασυνέχειες μεταξύ των γειτονικών πλαισίων. Ουσιαστικά αφαιρέσαμε τις μεμονωμένες ανιχνεύσεις (outliers) που αντιστοιχούν σε 200ms, η οποία δεν αποτελεί λογική διάρκεια όσον αφορά στα μη τυπικά ηχητικά γεγονότα.

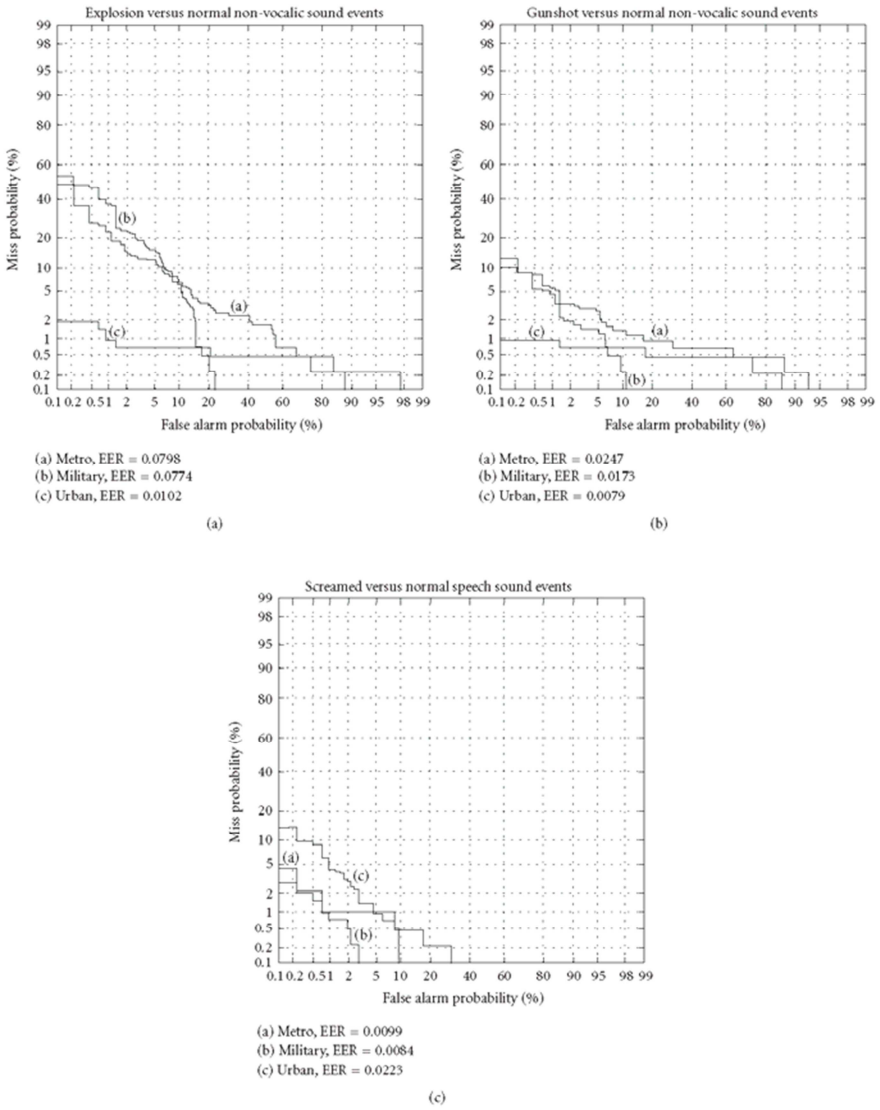


Σχήμα 5.7: Ο βρόχος ανατροφοδότησης για την προσαρμογή των GMM.

5.4.2. Αντιμετωπίζοντας την σπανιότητα των μη τυπικών γεγονότων

Το δεύτερο ζήτημα αποτελείται από όχι μόνο από τη σπανιότητα που χαρακτηρίζει την ύπαρξη ανώμαλων καταστάσεων αλλά και του γεγονότος ότι ένα άλλο μη ενδιαφέρον ηχητικό γεγονός μπορεί να λάβει χώρα και να προκαλέσει μία λάθος αναγνώριση (false alarm). Είναι μάλλον δύσκολο να δημιουργηθούν τα ακριβή πρότυπα τα οποία θα αντιπροσωπεύουν όλους αυτούς τους ήχους κυρίως επειδή η ύπαρξή τους είναι δύσκολο να είναι γνωστή a-priori (π.χ. ήχοι κλειδιών, κόρνες, γάβγισμα σκύλου κ.λπ.) ενώ είναι σημαντικό να μην παρερμηνευθούν από το σύστημα ως μη τυπικά γεγονότα. Προς την εκπλήρωση αυτού του στόχου, συλλέξαμε δεδομένα από τους διάφορους τύπους ηχητικών πηγών για να βεβαιωθούμε ότι η ακουστική σκηνή κάθε σεναρίου περιγράφεται από έναν υψηλό βαθμό ποικιλομορφίας (π.χ. ο υπόγειος σιδηρόδρομος περιλαμβάνει κόρνες, το άνοιγμα/κλείσιμο πόρτας, τους ανθρώπους που μιλούν στο background, τη μετακίνηση τρένων κ.λπ.).

Αν και τα ακουστικά δείγματα είναι αντιπροσωπευτικά των κατηγοριών που χρειαζόμαστε, δεν παρέχουν μια ακριβή περιγραφή *όλων* των πιθανών πραγματοποιήσεων τέτοιων γεγονότων. Συνεπώς ενσωματώνουμε μια τεχνική που δίνει στο σύστημά μας τη δυνατότητα να προσαρμοστεί στους ακουστικούς όρους του περιβάλλοντος λειτουργίας. Η προσαρμοστικότητα παρέχεται από την εφαρμογή της *Maximum a Posteriori* (MAP) μεθόδου στα στατιστικά πρότυπα κάθε κατηγορίας και εφαρμόζεται μέσω του βρόχου ανατροφοδότησης (Reynolds et al., 2000). Αρχικά μια εποπτευομένη (supervised) ημιαυτόματη διαδικασία ενεργοποιείται που λαμβάνει υπόψη της, το ground truth ως εισαγωγή από εξουσιοδοτημένο προσωπικό. Μετά από αυτήν την αρχική χρονική περίοδο το σύστημα προσαρμόζεται κατά τρόπο ανεπίβλεπτο (unsupervised) και εκμεταλλεύεται αποκλειστικά τις δικές του αποφάσεις.



Σχήμα 5.8: Detection Error Tradeoff καμπύλες που δείχνουν την ικανότητα του προσαρμοσμένου συστήματος να εντοπίζει μη φωνητικά και φωνητικά μη τυπικά ηχητικά γεγονότα κάτω από τρία διαφορετικά είδη ακουστικού περιβάλλοντος.

Ο βρόχος ανατροφοδότησης απεικονίζεται στο Σχήμα 5.7. Ένας δείκτης ιστορίας (history index) κρατιέται που περιέχει τη σειρά αποφάσεων που λαμβάνονται από το σύστημα σχετικά με το διάστημα v διάρκειας τ δευτερολέπτων, όπου κάθε διάρκεια εξαρτάται από τις τελικές αποφάσεις του συστήματος, δηλ. όταν προβλέπει το σύστημα την ίδια κατηγορία για διαδοχικά πλαίσια, αυτά τα συγκεκριμένα πλαίσια κατόπιν περιλαμβάνονται σε μια ακουστική ακολουθία ενώ μια νέα ακολουθία διαμορφώνεται όταν αλλάζει η πρόβλεψη. Το ground truth κρατήθηκε επίσης παράλληλα για τις ίδιες χρονικές περιόδους και τα ακουστικά δεδομένα αποθηκεύτηκαν σε 16 KHz με δεκαεξάμπιτη ανάλυση. Στη συνέχεια αναλύθηκαν τα στοιχεία που έχουν κατηγοριοποιηθεί λανθασμένα και οι αντίστοιχες ακολουθίες χαρακτηριστικών γνωρισμάτων χρησιμοποιήθηκαν για την προσαρμογή των αντίστοιχων μοντέλων. Αυτή η φάση πραγματοποιείται κατά τη διάρκεια μιας ανενεργού χρονικής περιόδου

(π.χ. κατά τη διάρκεια της νύχτας). Κατόπιν το σύστημα προσαρμόστηκε κατά τρόπο ανεπίβλεπτο χρησιμοποιώντας τις αποφάσεις του να αντικαταστήσει το ground truth. Πιο συγκεκριμένα η διαδικασία της ανεπίβλεπτης προσαρμογής λειτουργεί με τον ακόλουθο τρόπο: για μια δεδομένη χρονική περίοδο όλα τα τμήματα συμπεριλαμβανομένων των αντίστοιχων προβλέψεων αποθηκεύονται. Στη συνέχεια οι κατάλληλες ακουστικές παράμετροι εξάγονται από το σύστημα (π.χ. MFCC και dMFCC για την προσαρμογή του φωνητικού/μη-φωνητικού προτύπου). Έπειτα αυτές οι παράμετροι υιοθετούνται για να προσαρμόσουν το αντίστοιχο πρότυπο σύμφωνα με την πρόβλεψη του συστήματος. Με αυτό το τρόπο το σύστημα είναι ικανό για την αυτόνομη αναπροσαρμογή των προτύπων του, και άρα προσαρμόζεται ακόμα καλύτερα στις περιβαλλοντικές συνθήκες.

5.4.3. Οι καμπύλες DET του προσαρμοσμένου συστήματος

Το συγκεκριμένο πείραμα πραγματοποιήθηκε για τρεις συνεχόμενες ημέρες ενώ το ground truth ήταν γνωστό. Τα μισά από αυτά τα δεδομένα αναλύθηκαν για τα λάθη ταξινόμησης και έπειτα χρησιμοποιήθηκαν για την προσαρμογή των αντίστοιχων προτύπων (εποπτευομένη προσαρμογή MAP). Στη δεύτερη φάση τα Γκαουσσιανά πρότυπα προσαρμόστηκαν κατά τρόπο ανεπίβλεπτο βασισμένο στις αποφάσεις που λήφθηκαν αυτόματα από το σύστημα και χρησιμοποιώντας τα ακουστικά δεδομένα που συλλέχθηκαν κατά τη διάρκεια μιας ημέρας. Τα αποτελέσματα που αναφέρονται σε αυτό το τμήμα αποκτήθηκαν με τη χρήση των log-likelihoods που πάρθηκαν ως αποτελέσματα από τα πρότυπα τα οποία είχαν προσαρμοστεί μετά και από τις δύο φάσεις του πειράματος. Κατά τη διάρκεια της διαδικασίας προσαρμογής οι παράμετροι των Γκαουσσιανών συναρτήσεων (βάρη, συνδιασπορές και μέσοι όροι) μαθεύτηκαν από τα δεδομένα προσαρμογής ενώ η αξία του προγενέστερου βάρους (weight on prior) κατά τη διάρκεια της αναπροσαρμογής τέθηκε ίση με 0.5 (αλλαγές σε αυτήν την παράμετρο δεν παρείχαν καλύτερη απόδοση).

Φάση προσαρμογής	**Περιβάλλον**	**Μέσο EER (για τρία μη τυπικά γεγονότα)**	**Μέσο EER για τρία περιβάλλοντα (% βελτίωση)**
Χωρίς προσαρμογή	Μετρό	0.32461	0.26253 (-)
	Αστικό	0.2901	
	Στρατιωτικό	0.17289	
Supervised	Μετρό	0.11226	0.09676 (63.14%)
	Αστικό	0.12673	
	Στρατιωτικό	0.0513	
Supervised και Unsupervised	Μετρό	0.03786	0.02856 (70.48%)
	Αστικό	0.01346	
	Στρατιωτικό	0.03436	

Πίνακας 5.4: Τα Equal Error Rates για τις τρεις φάσεις του πειράματος

Σε αντίθεση με άλλους αλγορίθμους προσαρμογής (π.χ. γραμμική συμμεταβολή μέγιστης πιθανότητας - maximum likelihood linear regression), η συγκεκριμένη μεθοδολογία απαιτεί περισσότερα στοιχεία δεδομένου ότι λειτουργεί στο επίπεδο των συναρτήσεων. Εντούτοις λόγω αυτής της χαμηλού επιπέδου προσέγγισης, όταν στοιχεία μεγάλης ποσότητας είναι διαθέσιμα η μέθοδος *MAP* τείνει να αποδίδει καλύτερα. Αυτή η προϋπόθεση ισχύει όσον αφορά στην προσέγγισή μας δεδομένου ότι συλλέξαμε 72 ώρες (3x24h) δεδομένων. Τα μισά από αυτά τα στοιχεία χρησιμοποιήθηκαν για την εποπτευομένη προσαρμογή προτύπων. Οι επόμενες 24 ώρες εξυπηρέτησαν την ανεπίβλεπτη προσαρμογή ενώ το υπόλοιπο των στοιχείων (12h) χρησιμοποιήθηκε για τη δοκιμή του προσαρμοσμένου συστήματος. Πρέπει να αναφερθεί ότι αυτό το πειραματικό στάδιο εκμεταλλεύεται τις ιδιότητες μείωσης αντήχησης που προσφέρει η cepstral μέση κανονικοποίηση (cepstral mean normalization).

Οι καμπύλες DET για κάθε ένα εκ των τριών περιβαλλόντων απεικονίζονται στο Σχήμα 5.8 (έκρηξη, πυροβολισμός και ανίχνευση μη τυπικής ομιλίας). Όπως μπορούμε να δούμε τα καλύτερα ποσοστά ανίχνευσης επιτυγχάνονται στο αστικό περιβάλλον (μέσο EER=0.01346) ακολουθούμενα από το περιβάλλον που ταιριάζει σε στρατιωτικές εφαρμογές (μέσο EER=0.03436) και τον σταθμό μετρό (μέσο EER=0.03786). Επιπλέον τα ποσοστά των εσφαλμένων συναγερμών έχουν σχετικά χαμηλές τιμές όσον αφορά στις χαρακτηριστικές συνθήκες και των τριών περιβαλλόντων. Πρέπει να σημειωθεί ότι το σύστημα συνεχίζει την διαδικασία προσαρμογής κατά τρόπο ανεπίβλεπτο, επιτυγχάνοντας έτσι ακόμα καλύτερη απόδοση. Στον Πίνακα 5.4 είναι ταξινομημένα τα EERs που αντιστοιχούν σε κάθε στάδιο του συγκεκριμένου πειράματος. Όπως μπορούμε να δούμε υπάρχουν σημαντικές βελτιώσεις και στις δύο φάσεις προσαρμογής. Καταλήγουμε στο συμπέρασμα ότι τα αποτελέσματα του προσαρμοσμένου συστήματος είναι αρκετά ελπιδοφόρα και επιδεικνύουν τη φορητότητα αλλά και την ευελιξία που προσφέρεται από την προτεινόμενη δομή.

Κεφάλαιο 6

Πιθανολογική Ανίχνευση Καινοτομίας για Ακουστική Παρακολούθηση κάτω από Συνθήκες Πραγματικού Κόσμου

Η ανίχνευση καινοτομίας (novelty detection) ουσιαστικά αποτελεί τον προσδιορισμό άγνωστων/νέων στοιχείων, δηλ. στοιχεία που διαφέρουν πολύ από αυτά με τα οποία εκπαιδεύθηκε το σύστημα. Σε αυτό το κεφάλαιο ερευνάται η συγκεκριμένη τεχνική όπως εφαρμόζεται στην ακουστική επόπτευση ανώμαλων καταστάσεων.

Μια ευρεία ποικιλία ακουστικών παραμέτρων υιοθετείται προς τη διαμόρφωση ενός διανύσματος χαρακτηριστικών πολλών-πεδίων (multi-domain), το οποίο συλλαμβάνει διαφορετικά χαρακτηριστικά των ακουστικών σημάτων. Χρησιμοποιήσαμε τη τράπεζα φίλτρων Mel, το ακουστικό πρωτόκολλο MPEG-7, τον χειριστή ενέργειας Teager και την ανάλυση κυματιδίων. Στη συνέχεια οι συντελεστές γνωρισμάτων τροφοδοτούν τρεις πιθανολογικές μεθοδολογίες ανίχνευσης καινοτομίας. Η απόδοσή τους υπολογίζεται χρησιμοποιώντας δύο μέτρα που λαμβάνουν υπόψη τους, τους λάθος εντοπισμούς καθώς και τους εσφαλμένους συναγερμούς. Το σύνολο των δεδομένων καταγράφηκε κάτω από πραγματικές συνθήκες συμπεριλαμβανομένων τριών διαφορετικών τοποθεσιών όπου καταγράφηκαν διάφοροι τύποι τυπικών και μη τυπικών ηχητικών γεγονότων. Χρησιμοποιήθηκαν: α) ένα κλειστό περιβάλλον που προσομοιάζει το χώρο ενός τυπικού σπιτιού, β) ένας ανοικτός δημόσιος χώρος και γ) ένας διάδρομος γραφείων. Ο τελικός σκοπός του προτεινόμενου συστήματος είναι να ενισχυθεί η απόφαση ενός εξουσιοδοτημένου ατόμου προς τη λήψη των κατάλληλων αποφάσεων/ενεργειών για την παρεμπόδιση οποιασδήποτε απώλειας ζωής/ιδιοκτησίας. Τα αποτελέσματα δείχνουν ότι η πιθανοτική ανίχνευση καινοτομίας μπορεί να παρέχει μια ακριβή ανάλυση της σκηνής όσον αφορά στον προσδιορισμό ανώμαλων γεγονότων.

6.1. Εισαγωγή

Έχουν προταθεί αρκετά συστήματα για τις εφαρμογές επιτήρησης με κύριο στόχο να εξεταστούν γεγονότα που εμπεριέχουν βία, εγκλήματα κτλ. Η επόπτευση, γενικά, δεν είναι σε σύγκρουση με το νόμο και είναι κοινή πρακτική σε καταστήματα, υπηρεσίες, αερολιμένες κ.λπ., όπου η ανάγκη για αποτελεσματικούς όρους ασφάλειας δικαιολογεί την εγκατάσταση διάφορων αισθητήρων όπως κάμερες, μικρόφωνα κ.λπ. Η απαίτηση για ανεπίβλεπτη αξιολόγηση της παρούσας κατάστασης έχει παρακινήσει την κοινότητα της επεξεργασίας σήματος προς τον πειραματισμό με διάφορα αυτοματοποιημένα πλαίσια, π.χ. (Haritaoglu et al., 2000). Στόχος αυτών των

συστημάτων είναι η παροχή μιας ακριβούς περιγραφής μιας σκηνής ενδιαφέροντος και η χρησιμοποίηση αυτής σε μια διεπαφή υποστήριξης απόφασης (decision support interface) προκειμένου να ελαχιστοποιηθεί το φορτίο εργασίας του ανθρώπου χειριστή.

Η *ανίχνευση καινοτομίας* αναφέρεται στην αναγνώριση άγνωστων (ή νέων) δεδομένων, δηλ. δεδομένα που διαφέρουν αρκετά από εκείνα που επεξεργάστηκε το σύστημα κατά τη διάρκεια της εκπαίδευσης. Αυτός ο τύπος δεδομένων μπορεί να οδηγήσει το σύστημα στην κακή μεταχείρισή τους (π.χ. λανθασμένη κατηγοριοποίηση). Ο a-priori προσδιορισμός των δεδομένων που μπορούν να υποβληθούν σε αποτελεσματική επεξεργασία από το σύστημα αποτελεί προϋπόθεση για μια επιτυχημένη λειτουργία και αναγνώριση (Markou et al., 2003). Αυτή η τεχνική έχει αποδειχθεί ότι διευκολύνει πολλές εφαρμογές όπως η αναγνώριση ψηφίων γραμμένων με το χέρι (Tax et al., 1998), η ανίχνευση καρκίνου (Tarassenko, 1995), η ανίχνευση παρείσφρησης σε δίκτυα υπολογιστών (Wei et al., 2001), η ταξινόμηση εικόνων (Singh et al., 2001) κ.α. Όσον αφορά στον τομέα της ακουστικής επεξεργασίας σήματος, η ιδέα της ανίχνευσης καινοτομίας έχει υιοθετηθεί σε διάφορα άρθρα. Ένας αλγόριθμος διανυσματικών μηχανών υποστήριξης εισάγεται στην εργασία (Davy et al., 2002) για τον προσδιορισμό απότομων φασματικών αλλαγών. Συγκρίνεται με έναν μη παραμετρικό ανιχνευτή για την ακουστική κατάτμηση μουσικών σημάτων όπου και χρησιμοποιήθηκε αποτελεσματικά. Μια τεχνική βασισμένη σε νευρωνικό δίκτυο με εφαρμογή στη κατηγοριοποίηση μικτών ακουστικών γεγονότων παρουσιάζεται στην (Linares et al., 1997). Ένα σήμα πραγματικού περιβάλλοντος δίνεται στο σύστημα, το οποίο επιτυγχάνει την ανίχνευση ταυτόχρονων γεγονότων. Μια ενδιαφέρουσα εφαρμογή, ο μη καταστροφικός έλεγχος μηχανικών συστημάτων εξερευνείται στην (Emamian et al., 2000). Ο βραχύχρονος μετασχηματισμός Fourier εφαρμόζεται στα σήματα εισαγωγής και γίνεται είσοδος στο δίκτυο Kohonen. Τα αποτελέσματα παρουσιάζουν αποδοτική ανίχνευση ρωγμών ακόμα και σε ένα θορυβώδες περιβάλλον. Η ανίχνευση καινοτομίας αξιοποιείται από τους συντάκτες της (Flexer et al., 2005) για την ταξινόμηση μουσικού είδους. Οι (Davy et al.,2006) εξηγούν την κατασκευή μιας διανυσματικής μηχανής υποστήριξης με έναν διαδοχικό αλγόριθμο βελτιστοποίησης για την ανίχνευση ανώμαλων γεγονότων. Κατά τη διάρκεια της πειραματικής φάσης αυτή η προσέγγιση εφαρμόζεται σε ένα μουσικό σήμα όπου επιτυγχάνονται καλά αποτελέσματα κατάτμησης. Ο αυτόματος εντοπισμός των σημείων σημαντικών αλλαγών σε μουσική ή άλλων ειδών ήχους μέσω της αυτό- ομοίωσης αναλύεται στην εργασία (Foote, 2000). Η μέθοδος είναι βασισμένη στον υπολογισμό της μήτρας απόστασης διάφορων ακουστικών παραμέτρων, ο οποίος ακολουθείται από μια διαδικασία συσχετισμού πυρήνων. Αυτό εφαρμόζεται στο πρώτο λεπτό του *Animals Have Young* (video V14 από το σύνολο δεδομένων MPEG-7) για την κατάτμηση ομιλίας/μουσικής με αρκετά καλά αποτελέσματα. Αναφέρεται ότι διάφορες χρήσιμες εφαρμογές μπορούν να ωθηθούν μ' αυτό τον αλγόριθμό, όπως η ακουστική ευρετηρίαση και η περιληπτική παρουσίαση πληροφοριών. Η ανίχνευση καινοτομίας χρησιμοποιείται επίσης στην

εργασία (Richard et al., 2007) για την εξομάλυνση των αποφάσεων ενός SVM με εφαρμογή στη διάκριση ομιλίας/μουσικής. Οι συντάκτες ερεύνησαν τρεις τύπους εξομάλυνσης: χρησιμοποιώντας το Μπεϋζιανό κριτήριο πληροφοριών (BIC), το μίας-κατηγορίας SVM καθώς και πιθανολογικές αποστάσεις.

Η εκμετάλλευση της ακουστικής πληροφορίας με στόχο τον έλεγχο/επόπτευση του περιβάλλοντα χώρου είναι ένας σχετικά νέος ερευνητικός τομέας με πολλές ενδιαφέρουσες εφαρμογές. Η εστίαση τοποθετείται κυρίως στο σύνολο χαρακτηριστικών, στον ταξινομητή, στα στοιχεία κατάρτισης καθώς επίσης και στις ακουστικές κατηγορίες. Όσον αφορά στο σύνολο δεδομένων που χρησιμοποιείται, αποτελείται συνήθως από τις ηχογραφημένες εκ των προτέρων και καλά καθορισμένες ηχητικές κατηγορίες ενώ υπάρχει ανάγκη για μία βάση που να αντιπροσωπεύει ικανοποιητικά τις πραγματικές περιστάσεις. Επιπλέον, έχουν προταθεί διάφορα σύνολα χαρακτηριστικών (π.χ. MFCC, ZCR, LPC, LPCC, LFCC κ.λπ.) σε συνδυασμό με παραγωγικούς και μη-παραγωγικούς αλγορίθμους αναγνώρισης προτύπων. Σε αυτήν την παρουσίαση υπάρχει μετατόπιση από τις ελεγχόμενες εργαστηριακές συνθήκες και η έμφαση δίνεται επάνω στη χρήση μιας βάσης που καταγράφηκε κάτω από πραγματικές συνθήκες και διαφορετικούς τύπους περιβαλλόντων.

Παρουσιάζεται η έννοια της ανίχνευσης καινοτομίας στον τομέα του ακουστικού ελέγχου χώρων για την παρουσία επικίνδυνων καταστάσεων. Η χρήση της παρακινείται από το γεγονός ότι δεν είναι εφικτό να ληφθεί μια ευρεία ποικιλία δεδομένων, τα οποία να είναι αντιπροσωπευτικά των ανώμαλων καταστάσεων ώστε να δημιουργηθούν αξιόπιστα στατιστικά μοντέλα, ειδικά κάτω από πραγματικές συνθήκες. Αντίθετα, υπάρχει αφθονία δεδομένων για τη διαμόρφωση της κανονικής κατάστασης δεδομένου ότι η μη τυπικές καταστάσεις αποτελούν σπάνια γεγονότα. Η υπόθεση ότι τα ανώμαλα ακουστικά γεγονότα είναι κατά μεγάλο βαθμό διαφορετικά από τα κανονικά ισχύει. Βλέποντας το πρόβλημα μέσα από το πρίσμα της στατιστικής, η ανίχνευση καινοτομίας αναφέρεται στην απόφαση εάν μια άγνωστη ακουστική ακολουθία έχει παραχθεί από την κατανομή που ακολουθείται από τα μοντέλα κανονικών καταστάσεων. Εκθέτουμε πειραματισμούς σε ένα σύνολο δεδομένων που καταγράφηκε κάτω από πραγματικές συνθήκες κατά τη διάρκεια των γυρισμάτων τριών σεναρίων: α) έξυπνο περιβάλλον κλειστού χώρου, β) σενάριο ATM ανοικτού χώρου και γ) σενάριο ασφάλειας γενικού σκοπού κλειστού και ανοικτού χώρου. Επιπλέον γίνεται εκτενή χρήση πιθανοτήτων δομών (ομαδοποίηση (clustering) GMM, καθολικό GMM και καθολικό HMM) που υπηρετούν το στόχο της ανίχνευσης καινοτομίας. Εκπαιδεύονται με ακουστικές παραμέτρους διαφορετικών πεδίων (χρόνου, συχνότητας και κυματιδίου) που μπορούν να παρέχουν μια ακριβέστερη περιγραφή των ακουστικών σημάτων έναντι αντιπροσωπεύσεων που χρησιμοποιούν λιγότερα πεδία ειδικά όταν έχουμε να αντιμετωπίσουμε ιδιαίτερα θορυβώδεις συνθήκες, οι οποίες αντιμετωπίζονται συχνά όταν έχουμε να κάνουμε με πραγματικές συνθήκες (κάτι που πρέπει να αναμένεται). Τα πειράματα που πραγματοποιήθηκαν καταδεικνύουν την αποτελεσματικότητα της

προτεινόμενης προσέγγισης που όπως αποδεικνύεται αποδίδει πολύ καλά ακόμα και κάτω από περιβάλλοντα με διαφορετικές ακουστικές ιδιότητες.

Το επόμενο τμήμα παρέχει λεπτομέρειες σχετικά με το σχεδιασμό που οδήγησε στην καταγραφή της ακουστικής βάσης δεδομένων μας. Η ενότητα 6.3 περιγράφει τις πιθανοτικές στρατηγικές ανίχνευσης καινοτομίας που χρησιμοποιήθηκαν ενώ η ενότητα 6.4 εξηγεί το πειραματικό πρωτόκολλο και παρουσιάζει τα τελικά αποτελέσματα.

6.2. Σενάριο εφαρμογής

6.2.1. Η ιδέα πίσω από το σχεδιασμό των σεναρίου

Έγινε καταγραφή με διάφορα σενάρια απο διάφορα ηχογραφημένα ηχητικά τοπία (sound scapes) ή από διάφορα έργα και οργανώνονται σύμφωνα με το περιβάλλον που χρησιμοποιήθηκε για κάθε σύνοδο καταγραφής. Είχαν ως σκοπό να είναι όσο το δυνατόν πιο γενικού σκοπού για την εξυπηρέτηση ενός ευρέος φάσματος πραγματικών εφαρμογών.

Χρησιμοποιήθηκε η εξης τοποθεσία: δημόσιος υπαίθριος χώρος.

Τα ηχογραφημένα στοιχεία χρησιμοποιήθηκαν για τη γενίκευση των δραστηριοτήτων που καταγράφηκαν κατά τη διάρκεια του σεναρίου ασφάλειας μέσα σε ένα εικονίκο τυποποιημένο εσωτερικό χώρο. Αυτά τα περιβάλλοντα παρουσιάζουν μεγάλες διαφορές από την άποψη των ακουστικών όρων (ήρεμες συνθήκες στο εσωτερικό περιβάλλον, σχετικά ήρεμες στην εσωτερική δημόσια περιοχή και συνθήκες μη στατικού παρασιτικού θορύβου στην υπαίθρια δημόσια περιοχή).

6.2.2. Σενάριο δημόσιας ασφάλειας

Οι καταγραφές αυτού του σεναρίου πραγματοποιήθηκαν με δειγματα σε εσωτερικό και σε υπαίθριο περιβάλλον, διευκολύνοντας κατά συνέπεια ένα γενικής χρήσης σενάριο ασφάλειας.

Σενάριο	Είδος γεγονότος	Ακουστικό γεγονός	Συνολικός αριθμός	Συνολική διάρκεια (s)
Σενάριο ασφαλείας γενικού σκοπού	μη τυπικό	panic	10	128.6
		surprise	6	7.5
		anger	8	12.1
		scream	14	12.8
		people fighting	5	23.8
		pain	6	13.5
	τυπικό	background speech	101	841.8
		normal speech	156	1233.2
		background noise (wind, birds, other noise)	21	255.1

Πίνακας 6.1: Ακουστικά γεγονότα στη βάση δεδομένων PROMETHEUS.

Για να δημιουργήσουμε ρεαλιστικά σενάρια τα άτομα και οι ομάδες που ακούγονται περιλαμβανάνουν μία λογομαχία σε κισσέ, β) ένας τσακωμός ανθρώπων γ) ένα άτομο ικετεύει για χρήματα και δ) μια ομάδα ανθρώπων συγκεντρώνεται π.χ. διαδήλωση

6.2.3. Ανάλυση του ακουστικού μέρους της βάσης

Για πρακτικούς λόγους, κατά τη διάρκεια των πειραματισμών αυτής της εργασίας χρησιμοποιήσαμε ένα μικρόφωνο. Το ακουστικό σήμα δειγματολήφθηκε στα 32KHz με 32 bit ανάλυση και αποθηκεύτηκε σε μορφή WAV. Ο Πίνακας 6.1 συνοψίζει τα καταγραμμένα ακουστικά γεγονότα (τυπικά και μη τυπικά). Τα τυπικά ακουστικά γεγονότα είναι κανονική ομιλία, παρασιτικός θόρυβος (π.χ. αέρας, ομιλία στο υπόβαθρο κ.λπ.), κουδούνι πόρτας κ.λπ. ενώ τα ακουστικά γεγονότα ενδεικτικά των ανώμαλων καταστάσεων είναι πέσιμο αντικειμένων, σπάσιμο υλικού, συναγερμός πυρκαγιάς και των ανθρώπινοι φωνητικοί ήχοι σχετικοί με αρνητικά συναισθήματα. Υπάρχουν επίσης αρκετά μέρη με παρασιτικό θόρυβο. Σε αυτό το κεφάλαιο εξετάζουμε μόνο τις μη τυπικές καταστάσεις που είναι ανιχνεύσιμες μέσω του ακουστικού αισθητήρα. Αυτές οι σκηνές αντιστοιχούν περίπου στο 90% όλων των μη τυπικών καταστάσεων της καταγεγραμμένης βάσης δεδομένων.

6.3. Μεθοδολογία ανίχνευσης καινοτομίας

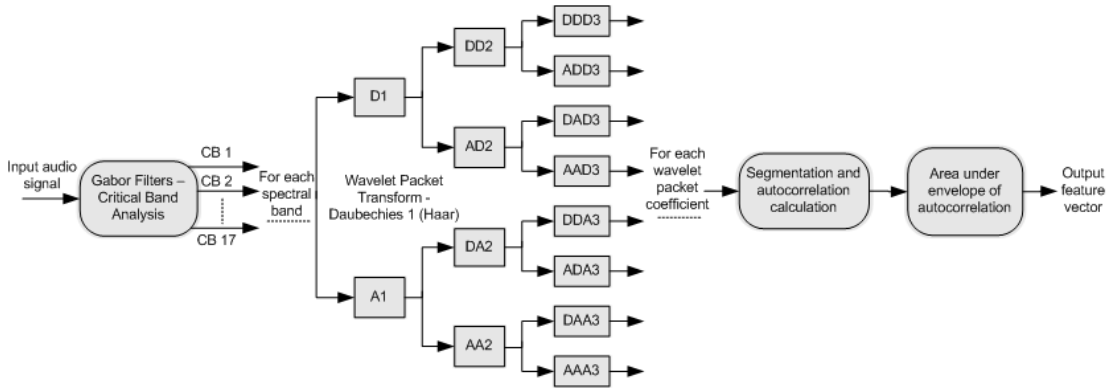
Είναι σχεδόν αδύνατο να συλλεχθούν όλες οι πιθανές εκδοχές κάθε μη τυπικού γεγονότος. Επομένως η κατάρτιση αντιπροσωπευτικών στατιστικών προτύπων δεν είναι εφικτή. Πειραματιστήκαμε προς την υπερνίκηση αυτού του εμποδίου με το σχεδιασμό μίας μεθοδολογίας μηχανικής μάθησης που είναι βασισμένη στην ανίχνευση καινοτομίας και έχει τη δυνατότητα να αναγνωρίσει ακουστικά δεδομένα που δεν έχει "δει" κατά τη διαδικασία εκπαίδευσης. Κατά συνέπεια, η προτεινόμενη αρχιτεκτονική

παρακινείται από το γεγονός ότι τα ανώμαλα ακουστικά γεγονότα είναι πολύ διαφορετικά από τα κανονικά. Με άλλα λόγια η πρόθεσή μας είναι να θεσπίσουμε ένα πλαίσιο που δεν προϋποθέτει τη διακριτή μοντελοποίηση της κατηγορίας που περιλαμβάνει τα μη τυπικά δεδομένα. Το τελικό ολοκληρωμένο σύστημα θα ωφεληθεί από την έκβαση αυτής της μεθοδολογίας για τον εντοπισμό και κανονικής αλλά και ανώμαλης ανθρώπινης συμπεριφοράς.

6.3.1. Ακουστικές παράμετροι

Αρχικά είναι ουσιαστικό να επιλεχτούν ακουστικές παράμετροι ικανές να αντιπροσωπεύσουν αποτελεσματικά τα χαρακτηριστικά των προαναφερθέντων ακουστικών κατηγοριών. Μετά από τα αποτελέσματα της προηγούμενης δουλειάς μας (βλέπε κεφάλαια 5 και 6) καταλήξαμε στη χρησιμοποίηση των επόμενων ακουστικών περιγραφέων:

• Ακουστικό πρωτόκολλο MPEG-7: Audio spectrum flatness, waveform min, waveform max και audio fundamental frequency. Αυτοί παρέχουν μια συμπαγή αντιπροσώπευση της μορφής του σήματος στο πεδίο του χρόνου, της περιοδικότητας του καθώς και το πόσο "επίπεδο" είναι το φάσμα σε διαφορετικές ζώνες. Η εξαγωγή τους έχει τυποποιηθεί κατά τη διάρκεια της δημιουργίας του ακουστικού πρωτοκόλλου MPEG -7, το οποίο θεωρείται η state of the art για την αναγνώριση ήχων (Quackenbush et al.,2001).



Σχήμα 6.1: Εξαγωγή των παραμέτρων από το πεδίο wavelet με τριφασικό φιλτράρισμα

• *MFCC*: Παρμένοι από τον τομέα της αυτόματης αναγνώρισης ομιλίας/ομιλητών όπου και είναι γνωστό ότι παρέχουν ικανοποιητική απόδοση. Η διαδικασία παραγωγής τους έχει ως σκοπό να μιμηθεί ως ένα ορισμένο βαθμό το ανθρώπινο σύστημα ακουστικής αντίληψης. Σε αυτήν την μελέτη χρησιμοποιούμε τους

πρώτους 13 συντελεστές (συμπεριλαμβανομένου του μηδενικού) οι οποίοι χρησιμοποιούνται σε συνδυασμό με τις αντίστοιχες παραγώγους τους.

• *Προσωδία και χαρακτηριστικά βασισμένα στον χειριστή ενέργειας Teager*: Τα επιλεγμένα χαρακτηριστικά γνωρίσματα μπορούν να συλλάβουν τις αλλαγές που η ομιλία και η προσωδία καταδεικνύουν στην περίπτωση ανώμαλων φωνητικών ακουστικών γεγονότων. Ο υπολογισμός της περιοχής φακέλων αυτοσυσχέτισης TEO βασισμένος σε ζώνες συχνοτήτων είναι χρήσιμος για την ταξινόμηση πίεσης/άγχους σε σήματα ομιλίας (Zhou et al., 2001). Τους χρησιμοποιούμε σε συνδυασμό με τη θεμελιώδη συχνότητα, την παράγωγο της και το harmonicity to noise ratio (HNR). Ο συγκεκριμένος συνδυασμός μπορεί να είναι αποτελεσματικός όσον αφορά τον εντοπισμό ανθρώπινων φωνητικών αντιδράσεων κάτω από μη τυπικές περιστάσεις.

Οι προαναφερθείσες ακουστικές παράμετροι ανήκουν είτε στο πεδίο του χρόνου είτε σε εκείνο της συχνότητας. Έχει αποδειχθεί ότι οι παράμετροι των διαφορετικών περιοχών μπορούν να παρέχουν βελτιωμένη απόδοση (βλέπε Κεφάλαιο 8). Κατά συνέπεια πειραματιστήκαμε με μια πρόσφατα εισαχθείσα ομάδα περιγραφέων που είναι βασισμένη στην επεξεργασία πολλαπλών αναλύσεων. Το κύριο πλεονέκτημα του μετασχηματισμού wavelet είναι ότι μπορεί να επεξεργαστεί χρονικές σειρές, οι οποίες χαρακτηρίζονται από μη στατική ισχύ σε πολλές διαφορετικές συχνότητες. Σε αυτό το κεφάλαιο χρησιμοποιήσαμε τη συνάρτηση Daubechies 1 (ή Haar) ως αρχικό/μητρικό κύμα. Ο DWT εφαρμόζεται τρεις διαδοχικές φορές και στην ουσία αποτελεί ένα τριφασικό φιλτράρισμα των ακουστικών σημάτων όπως μπορούμε να δούμε στο Σχήμα 6.1. Μετά από μια ανάλυση κρίσιμων ζωνών βασισμένη στα ζωνοπερατά φίλτρα Gabor, εξάγουμε τα τριών επιπέδων πακέτα κυματιδίων για κάθε φασματική ζώνη. Στη συνέχεια, "κόβουμε" τους εξαγμένους συντελεστές και υπολογίζεται η περιοχή κάτω από την συνάρτηση αυτοσυσχέτισης τους η οποία και ομαλοποιείται σύμφωνα το μισό μέγεθος ενός πλαισίου. Οι N ομαλοποιημένες παράμετροι ολοκλήρωσης υπολογίζονται για κάθε πλαίσιο, όπου το N είναι ο συνολικός αριθμός των ζωνών συχνότητας πολλαπλασιασμένος με τον αριθμό των συντελεστών κυματιδίων (17x8=136). Αυτές οι παράμετροι αναπαριστούν το βαθμό μεταβλητότητας ενός συγκεκριμένου συντελεστή κυματιδίου μέσα σε μια συγκεκριμένη ζώνη συχνοτήτων. Όταν τα τυπικά ηχητικά γεγονότα συγκρίνονται με τα μη τυπικά καταδεικνύουν μεγάλες διαφορές μέσα σε αυτές τις ζώνες και επομένως η χρησιμοποίησή τους μπορεί να είναι ωφέλιμη.

6.3.2. Τεχνικές αναγνώρισης προτύπων

Αυτό το τμήμα παρουσιάζει τις τρεις τεχνικές που αξιολογήθηκαν για την ανίχνευση καινοτομίας στην ακουστική βάση που αναλύθηκε προηγουμένως. Βασίζονται στον

υπολογισμό της ελλοχεύουσας συνάρτησης πυκνότητας πιθανότητας με τη χρησιμοποίηση ενός μίγματος Γκαουσσιανών συστατικών. Υπολογίζουμε τις παραμέτρους του μοντέλου χρησιμοποιώντας μόνο τυπικά δεδομένα κατάρτισης, κάτι που είναι αντίθετο με την προσέγγιση ταξινόμησης που χρησιμοποιεί και τυπικά αλλά και μη τυπικά δεδομένα κατά τη διαδικασία της κατάρτισης. Κατόπιν για να καθορίσουμε εάν το άγνωστο σήμα είναι αρκετά παρόμοιο με το κανονικό μοντέλο χρησιμοποιούμε ένα σχέδιο που στηρίζεται σε ένα κατώφλι. Οι τρεις μέθοδοι που υιοθετήθηκαν για την ανίχνευση καινοτομίας είναι: α) ένα καθολικό GMM, ένα β) ένα καθολικό HMM και γ) ομαδοποίηση GMM (GMM clustering). Υιοθετήσαμε επίσης τη μέγιστη a posteriori (MAP) μέθοδο προσαρμογής προτύπων για τις παραμέτρους των Γκαουσσιανών συναρτήσεων (Reynolds et al., 2000). Γενικά, τα τυπικά δεδομένα αναμένονται να υπάρχουν πολύ συχνά και έτσι θα συγκεντρώνονται εύκολα προκειμένου να βελτιωθεί το πρότυπο της τυπικής κατηγορίας.

6.3.3. Καθολική μοντελοποίηση

Το πρόβλημα της εκτίμησης μίας συνάρτησης πυκνότητας πιθανότητας μπορεί να αντιμετωπιστεί με την κατασκευή ενός GMM. Είναι ουσιαστικά ένας γραμμικός συνδυασμός Γκαουσσιανών κατανομών που χαρακτηρίζονται από διαφορετικές παραμέτρους (βάρη, μήτρες μέσων και συνδιακύμανσης). Μπορούν να προσεγγίσουν οποιαδήποτε κατανομή δεδομένων εφόσον ένα επαρκές ποσό τους είναι διαθέσιμο (McLachlan et al., 1988). Ο αλγόριθμος k-means χρησιμοποιείται για να προσδώσουμε αρχικές τιμές στις παραμέτρους οι οποίες επανυπολογίζονται από τον επαναληπτικό αλγόριθμο ΕΜ. Ένα GMM που αποτελείται από *M* Γκαουσσιανές συναρτήσεις δίνεται από τον ακόλουθο τύπο:

$$p(x_t) = \sum_{m=1}^{m=M} \pi_m N(x_t, \mu_m, \Sigma_m) \qquad (6.1)$$

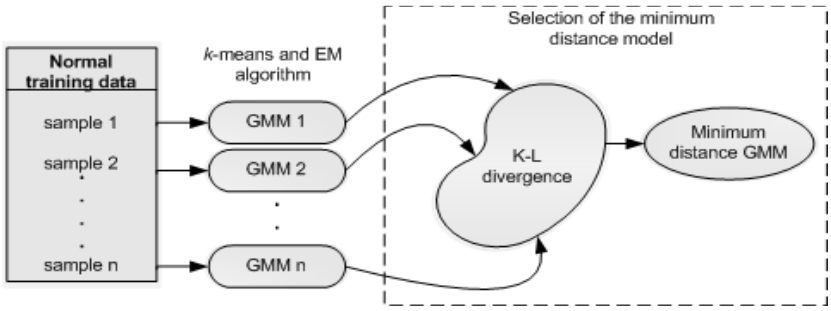
όπου το x_t αποτελεί τους συντελεστές των χαρακτηριστικών στο χρόνο *t*, το *N* είναι μία Γκαουσσιανή συνάρτηση με μέσο μ_m και μήτρα συνδιακύμανσης Σ_m ενώ π_m είναι η εκ των προτέρων πιθανότητα της συγκεκριμένης κατάστασης. Σε αυτό το στάδιο υποθέτουμε διαγώνια μήτρα συνδιακύμανσης. Ο αριθμός των Γκαουσσιανών συναρτήσεων λήφθηκε για το ακόλουθο σύνολο: {2, 4, 8, 16, 32, 64, 128, 256, 512}.

Στη πρώτη φάση δημιουργούμε ένα καθολικό GMM που αντιπροσωπεύει τα χαρακτηριστικά γνωρίσματα που εξάγονται από την κανονική/τυπική κατηγορία. Ένα κατώτατο όριο καθορίζεται από την ελάχιστη πιθανότητα των ακουστικών δειγμάτων

κατάρτισης. Εάν η πιθανότητα που παράγεται από το καθολικό πρότυπο όσον αφορά στην ομοιότητά της με ένα δείγμα δοκιμής είναι χαμηλότερη από το προκαθορισμένο κατώτατο όριο, το δείγμα είναι ταξινομημένο ως ανώμαλο γεγονός. Τα ακουστικά δεδομένα που ανήκουν μέσα σε αυτό το όριο πιθανότητας χρησιμοποιούνται περαιτέρω για την προσαρμογή του καθολικού προτύπου.

Η δεύτερη καθολική τεχνική διαμόρφωσης ήταν το από τα αριστερά προς τα δεξιά κρυμμένο πρότυπο Markov, μια τεχνική που χρησιμοποιείται εκτεταμένα αυτήν την περίοδο για τις ανάγκες της τεχνολογίας αναγνώρισης γενικευμένου ακουστικού σήματος (Kim et al., 2005). Παίρνουν υπό εξέταση τη συμπεριφορά του σήματος με την πάροδο του χρόνου, ιδιότητα που μπορεί να είναι μεγάλης σημασίας και κατά τη διάρκεια της μοντελοποίησης αλλά και της αναγνώρισης. Όσο ο χρόνος περνάει, κάθε ηχητικό γεγονός ακολουθεί ένα σχέδιο που παρουσιάζει συνεπή χαρακτηριστικά. Το HMM χωρίζει την ακολουθία τους σε έναν προκαθορισμένο αριθμό καταστάσεων, όπου καθεμία διαμορφώνεται από ένα διαγώνιο GMM. Στη συνέχεια οι σχέσεις μεταξύ κάθε κατάστασης μαθαίνονται χρησιμοποιώντας τον αλγόριθμο Baum-Welch, ο οποίος οδηγεί σε μια μήτρα των μεταβάσεων για όλες τις καταστάσεις. Στην από αριστερά προς τα δεξιά περίπτωση κάθε κατάσταση μπορεί να οδηγήσει είτε στην ίδια είτε στην επόμενη. Ο αριθμός καταστάσεων ποίκιλλε μεταξύ 3 και 5 ενώ ο αριθμός των Γκαουσσιανών συναρτήσεων που δοκιμάστηκαν κάθε φορά ήταν: {2, 4, 8, 16, 32, 64, και 128}.

Ένα καθολικό HMM δημιουργείται για την αντιπροσώπευση των κανονικών δεδομένων κατάρτισης. Το κατώφλι καινοτομίας τίθεται ως η ελάχιστη log-likelihood που παράγεται από τις ακολουθίες χαρακτηριστικών των τυπικών δεδομένων. Κατά τη διάρκεια της δοκιμής το άγνωστο σήμα κόβεται σε κομμάτια καθορισμένου μήκους τα οποία είναι ίσα με τον αριθμό των καταστάσεων του προηγουμένως δημιουργημένου προτύπου. Εάν η log-likelihood του σήματος ξεπερνά το κατώφλι, το συγκεκριμένο τμήμα είναι ταξινομημένο ως ξένο/καινοτόμο (outlier). Διαφορετικά οι παράμετροι χρησιμοποιούνται για την προσαρμογή του καθολικού προτύπου χρησιμοποιώντας τη μέθοδο *MAP*.



Σχήμα 6.2: Ομαδοποίηση GMM για επιλογή ενός απλού μοντέλου προς την αντιπροσώπευση ολόκληρης της ηχητικής κατηγορίας.

6.3.4. Ομαδοποίηση GMM

Η βασική ιδέα πίσω από την προτεινόμενη μέθοδο είναι να προσδιοριστεί ένα απλό πρότυπο που μπορεί να αντιπροσωπεύσει ένα πολύ μεγαλύτερο σύνολο ηχητικών γεγονότων. Αυτό το πρότυπο μπορεί να παρέχει μια περιγραφή των χαρακτηριστικών γνωρισμάτων που δεν είναι τόσο γενική όσο οι προηγούμενες. Πράγματι υπάρχουν περιπτώσεις όπου τα κανονικά ακουστικά δεδομένα παρουσιάζουν μικρές παραλλαγές σε σχέση με το ευρύ καθολικό πρότυπο. Αυτό το γεγονός μπορεί να οδηγήσει σε λάθος ταξινομήσεις. Η προτεινόμενη προσέγγιση δεν συγκεντρώνεται στα σφαιρικά χαρακτηριστικά που μοιράζονται μεταξύ τους τα δεδομένα εκπαίδευσης. Κατά συνέπεια τέτοιες περιπτώσεις μπορούν να εξεταστούν αποτελεσματικότερα από το σύστημα.

Αρχικά χτίζουμε ένα διαγώνιο GMM για καθένα εκ των δειγμάτων που ανήκουν στην κατηγορία των κανονικών ηχητικών γεγονότων. Κατόπιν υπολογίζουμε τις αποστάσεις όσον αφορά κάθε ζευγάρι του συνόλου. Η μήτρα απόστασης παρέχει μια περιγραφή της κατανομής που ακολουθείται από τα πρότυπα μέσα στο χώρο των ακουστικών χαρακτηριστικών. Τελικά επιλέγουμε το πρότυπο με την ελάχιστη απόσταση ως αυτό που μπορεί να αντιπροσωπεύσει ολόκληρη την κανονική ακουστική κατηγορία. Η επιλογή του κεντρικού προτύπου (centric model) καταδεικνύεται στο Σχήμα 7.2.

Προκειμένου να μετρηθεί η απόσταση μεταξύ δύο Γκαουσσιανών κατανομών χρησιμοποιούμε μια Monte Carlo προσέγγιση της απόστασης Kullback-Leibler (KL). Αυτός το μετρικό παρέχει έναν βαθμό ομοιότητας του ακουστικού περιεχομένου που αντιπροσωπεύεται από τα συγκεκριμένα πρότυπα. Η απόσταση KL μεταξύ των δύο συναρτήσεων πυκνότητας GMM, *pA* και *pB* ορίζεται ως

$$KL(A \| B) = \int p_A(x) \log \frac{p_B(x)}{p_A(x)} dx$$

$$KL(A \| B) \approx \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_B(x_i)}{p_A(x_i)} \qquad (6.2)$$

Αν και μέχρι τώρα δεν υπάρχει λύση κλειστής μορφής όσον αφορά τη περίπτωση των GMM, η απόσταση KL μπορεί να προσεγγιστεί από τον εμπειρικό μέσο, δηλαδή:

$$KL(A \| B) \approx \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_B(x_i)}{p_A(x_i)} \qquad (6.3)$$

όπου το n είναι ο αριθμός των δειγμάτων που παίρνουμε από την p_A ενώ πρέπει να αναφερθεί ότι αυτή η λύση προκύπτει από τα κεντρικά του οριακού θεωρήματος (central limit theorem). Αυτή η μεθοδολογία εξετάστηκε χρησιμοποιώντας το ακόλουθο σύνολο όσον αφορά στον αριθμό των Γκαουσσιανών συστατικών: {2, 4, 8, 16, 32}. Πρέπει να σημειωθεί ότι η απόσταση μεταξύ δύο GMMs υπολογίζεται αφότου λάβουμε n=2000 δείγματα.

Κατά τη διάρκεια της φάσης αξιολόγησης ταιριάζουμε το εισερχόμενο πλαίσιο με το κεντρικό πρότυπο. Το κατώφλι ορίζεται ίσο με τη μέγιστη log-likelihood των στοιχείων που υπάρχουν στο σώμα κατάρτισης. Εάν η log-likelihood υπερβαίνει αυτό το κατώτατο όριο, ο συγκεκριμένος ήχος προσδιορίζεται ως ανωμαλία.

Η προσέγγιση που περιγράφηκε παραπάνω μπορεί επίσης να χρησιμοποιηθεί για την αυτοματοποιημένη αναγνώριση ήχων η οποία θα είναι βασισμένη σε μια ολιστική αντιπροσώπευση κάθε δείγματος χωρίς την οικοδόμηση πιθανοτικών προτύπων για κάθε κατηγορία ξεχωριστά. Ένας γρήγορος τρόπος να βρεθεί η κατηγορία στην οποία ανήκει η άγνωστη ακουστική ακολουθία θα ήταν να επιλεχτεί η κατηγορία του πιο κοντινού (υπό την KL έννοια) με αυτό προτύπου.

6.4. Πειραματικό πρωτόκολλο και ανάλυση λαθών

Ο στόχος αυτών είναι να αξιολογήσουμε την απόδοση τριών πιθανοτικών μεθόδων ανίχνευσης καινοτομίας στο ίδιο σύνολο δεδομένων. Για την εξαγωγή των παραμέτρων που περιγράφηκαν στη παράγραφο 6.3.1 χρησιμοποιήσαμε πλαίσια των 30ms με βήμα ίσο με 10ms ακολουθώντας τις συστάσεις του πρωτοκόλλου MPEG-7. Τα αρχεία ήχου ήταν σε 16KHz με 16bit κβαντοποίηση ενώ προεπεξεργάστηκαν έτσι ώστε να ακυρωθεί οποιοδήποτε DC-offset. Όσον αφορά στα χαρακτηριστικά που ανήκουν στο πεδίο συχνότητας, το μέγεθος του FFT ήταν 512. Τα πειράματά μας βασίστηκαν στην υλοποίηση Torch (διαθέσιμη στην διεύθυνση http://www.torch.ch) των GMM και των HMM που γράφτηκε σε C++. Όσον αφορά την προσαρμογή *MAP* των παραμέτρων των Γκαουσσιανών συστατικών, η αξία του προγενέστερου βάρους κατά τη διάρκεια της αναπροσαρμογής τέθηκε ίση με 0.5. Πραγματοποιήσαμε εκτενή πειράματα για την επιλογή των παραμέτρων των στατιστικών τεχνικών διαμόρφωσης ώστε να πετύχουμε την υψηλότερη απόδοση. Επιπλέον, ο μέγιστος αριθμός επαναλήψεων του αλγορίθμου k- means για την έναρξη ήταν 50 ενώ και οι αλγόριθμοι EM και Baum-Welch είχαν και ανώτερο όριο 25 επαναλήψεις με ένα κατώτατο όριο ίσο με 0.001 μεταξύ διαδοχικών επαναλήψεων. Πρέπει να σημειωθεί ότι καμία τεχνική προεπεξεργασίας (όπως η κανονικοποίηση) δεν εφαρμόστηκε στα ακατέργαστα δεδομένα εισαγωγής με στόχο την εκτίμηση πυκνότητας, όπως προτείνεται στην εργασία (Bishop et al., 1994).

Τα δεδομένα που χρησιμοποιήθηκαν για την οικοδόμηση/προσδιορισμό του προτύπου της κανονικής κατηγορίας ποίκιλαν σύμφωνα με κάθε σενάριο. Παραδείγματος χάριν, στην περίπτωση του σεναρίου του έξυπνου σπιτιού τα κανονικά δεδομένα περιλαμβάνουν κανονική ομιλία, κουδούνι πορτών και ομιλία στο υπόβαθρο ενώ τα κανονικά στοιχεία του σεναρίου ασφάλειας είναι ομιλία υποβάθρου, κανονικοί ομιλία και παρασιτικός θόρυβος, παρόμοια με αυτά της περίπτωσης του σεναρίου του ATM. Κατά τη διάρκεια της κατάρτισης επιλέχθηκαν τυχαία τα μισά από τα κανονικά ακουστικά στοιχεία ενώ τα υπόλοιπα χρησιμοποιήθηκαν για τη δοκιμή του συστήματος. Επιπλέον, διαφορετικοί τύποι ανώμαλων ηχητικών γεγονότων υπόκεινται σε ανίχνευση για κάθε σενάριο.

Σενάριο	**Αλγόριθμος αναγνώρισης προτύπων**	**Οι παράμετροι με την καλύτερη απόδοση**	**TDR (%)**	**FDR (%)**
Σενάριο ασφαλείας γενικού σκοπού	Universal GMM	256 modes	92.5	9.1
	Universal HMM	3 states, 32 modes	96.4	3.2
	GMM clustering	16 modes (centric model)	89.8	4.8
	Classification approach - GMM	64 modes	81.5	8.9
Σενάριο έξυπνου σπιτιού	Universal GMM	256 modes	93.6	8.3
	Universal HMM	5 states, 16 modes	95.4	9.5
	GMM clustering	16 modes (centric model)	96.1	1.9
	Classification approach - GMM	128 modes	86.6	12.6
Σενάριο ATM	Universal GMM	128 modes	84.1	6.2
	Universal HMM	4 states, 32 modes	91.6	1.2
	GMM clustering	8 modes (centric model)	85.4	2.5
	Classification approach - GMM	64 modes	79.3	14.3

Πίνακας 6.2: Η απόδοση του συστήματος για κάθε σενάριο και στρατηγική ανίχνευσης καινοτομίας.

Τα σύνολα χαρακτηριστικών γνωρισμάτων που αναφέρθηκαν στη παράγραφο 6.3.1 συνδυαστήκαν με τους αλγορίθμους εκτίμησης pdf που περιγράφηκαν στη παράγραφο 6.3.2, και στη συνέχεια εξετάστηκαν για την ανίχνευση μη τυπικών ηχητικών γεγονότων χρησιμοποιώντας δύο κριτήρια απόδοσης. Οι παραδοσιακές μέθοδοι, όπως μέσο ποσοστό αναγνώρισης ή μήτρα σύγχυσης δεν εξετάζουν το διπλό

είδος λάθους που έχουμε να αντιμετωπίσουμε: αποτυχία αναγνώρισης μιας ανώμαλης κατάστασης ή ανίχνευση μίας όταν δεν είναι παρούσα. Και τα δύο είναι κρίσιμα και πρέπει να ληφθούν υπόψη όσον αφορά στη μέτρηση της απόδοσης των διαφορετικών υλοποιήσεων για να επιλέξουμε το σύστημα με εκείνη την αρχιτεκτονική που προσφέρει την καλύτερη απόδοση. Δύο μέτρα καθορίστηκαν: *true detection rate* (TDR) και *false detection rate*

(FDR):

$$TDR = \Pr(abnormal \mid abnormal) = \frac{\text{no. of abnormal testing sequences detected as abnormal}}{\text{no. of abnormal sequences in the test set}} \quad (6.4)$$

$$FDR = \Pr(abnormal \mid abnormal) = \frac{\text{no. of normal testing sequences detected as abnormal}}{\text{no. of normal sequences in the test set}} \quad (6.5)$$

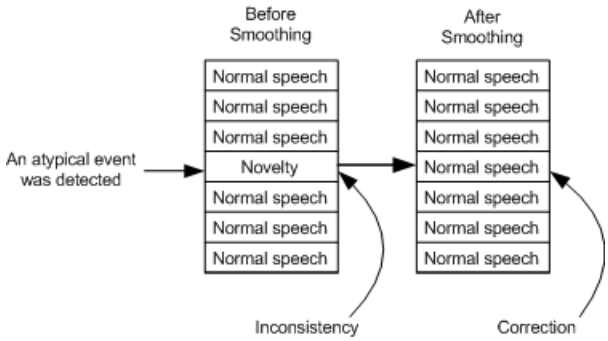
Το TDR ανακλά στην πιθανότητα που δείχνει εάν μια ανωμαλία ανιχνεύεται σωστά ενώ το FDR είναι η πιθανότητα ότι ένα κανονικό ακουστικό πλαίσιο προσδιορίζεται ψευδώς ως ανωμαλία.

Πραγματοποιήσαμε εξαντλητικά πειράματα χρησιμοποιώντας διαφορετικές τιμές παραμέτρων όσον αφορά στα σχήματα εκτίμησης πυκνότητας (καθολικό GMM, καθολικά HMM και ομαδοποίηση GMM) για την οικοδόμηση ενός ισχυρού συστήματος από την άποψη των TDR και FDR. Τα στοιχεία δοκιμής αποτελέσθηκαν από το α) το 50% των τυπικών ηχητικών δειγμάτων (που οι μεθοδολογίες δεν "βλέπουν" κατά τη διάρκεια της κατάρτισης) καθώς επίσης και β) όλα τα μη τυπικά ακουστικά στοιχεία (φωνητικά και μη-φωνητικά). Στο τέλος εφαρμόστηκε ένα σχήμα εξομάλυνσης πάνω στα αποτελέσματα όλων των μεθόδων ανίχνευσης καινοτομίας με τον ίδιο τρόπο. Αυτή η διαδικασία βασικά αφαιρεί τις μεμονωμένες ανιχνεύσεις πλαισίων, κάτι που θεωρείται λάθος εντοπισμός. Μπορούμε να δούμε ένα επεξηγηματικό παράδειγμα στο Σχήμα 6.3. Αυτό είναι βασισμένο στην υπόθεση ότι ένα πλαίσιο μεγέθους 30ms δεν μπορεί να περιλάβει μια επικίνδυνη κατάσταση λόγω της μικρής διάρκειάς του.

Στον Πίνακα 6.2 ταξινομούμε το TDR και το FDR όσον αφορά τις παραμέτρους στατιστικής μοντελοποίησης που παρείχαν την καλύτερη ακρίβεια. Παρατηρούμε ότι η καθολική μέθοδος HMM παρείχε την καλύτερη απόδοση και από την άποψη του TDR καθώς και του FDR όσον αφορά στα σενάρια ασφάλειας και ATM. Τα αντίστοιχα TDR και FDR είναι 96.4% και 3.2% για το γενικής χρήσης σενάριο ασφάλειας και 91.6% και 1.2% όσον αφορά το σενάριο του ATM.

Η ομαδοποίηση GMM κατέδειξε το χαμηλότερο FDR (1.9%) και υψηλότερο TDR (96.1%) όσον αφορά το σενάριο έξυπνου σπιτιού. Επιπλέον, κατά τη διάρκεια της ανάλυσης λαθών παρατηρήσαμε ότι η πλειοψηφία τους γίνεται από όλες τις

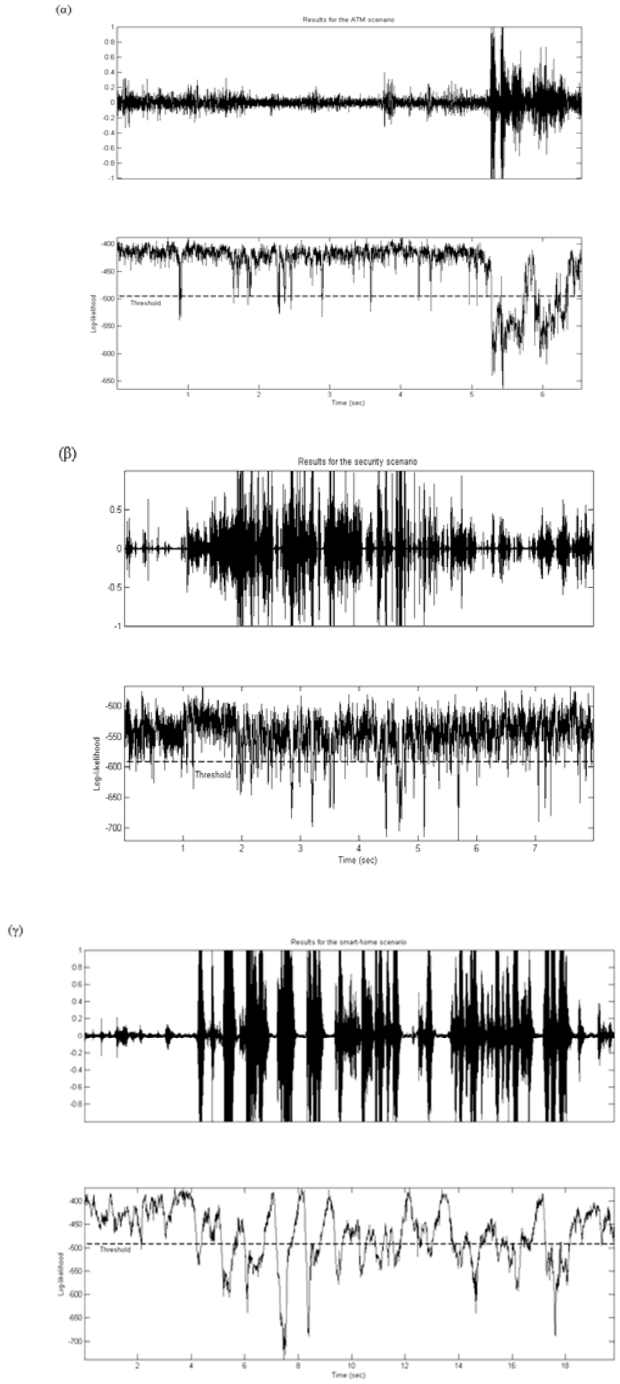
τεχνικές, δηλ. είναι αμοιβαίες. Παραδείγματος χάριν μια ακολουθία που δεν ταξινομείται σωστά από το καθολικό HMM είναι πιθανό να αναγνωριστεί επίσης λάθος από το καθολικό GMM. Αυτό δείχνει ότι τα διάφορα μη τυπικά ηχητικά γεγονότα καταχωρούνται λάθος από το προτεινόμενο πιθανοτικό σχήμα ανίχνευσης καινοτομίας.



Σχήμα 6.3: Ένα χαρακτηριστικό παράδειγμα του τρόπου με τον οποίο εφαρμόσαμε το σχήμα εξομάλυνσης πάνω στην ακολουθία προβλέψεων του συστήματος.

Κάποιος πρέπει επίσης να σκεφτεί παράλληλα ότι το πλαίσιο επιτήρησης λειτουργεί πάνω σε ακουστικά δεδομένα που συλλήφθηκαν υπό δυσμενείς πραγματικές σηνθήκες. Η καθολική τεχνική GMM επέδειξες την υψηλότερη FDR σε όλα τα σενάρια με λογικές μετρήσεις TDR. Αυτό μπορεί να οφείλεται στην ανικανότητά του να αποτυπώσει τη χρονική συμπεριφορά των ακουστικών σημάτων.

Παρέχονται επίσης συγκριτικά αποτελέσματα μεταξύ των μεθόδων ανίχνευσης καινοτομίας και της μεθοδολογίας που χρησιμοποιείται αυτήν την περίοδο από την πλειοψηφία των υπόλοιπων εργασιών, δηλ. η προσέγγιση της ταξινόμησης (Valanzise et al., 2007; Clavel et al., 2005). Αυτή η τεχνική αποτελείται ουσιαστικά από τη μοντελοποίηση της pdf κάθε κατηγορίας ήχων χρησιμοποιώντας ένα GMM. Υιοθετήθηκαν τα ίδια εκπαιδευτικά και εξεταστικά σύνολα. Ο αριθμός των Γκαουσσιανών συστατικών λήφθηκε από το επόμενο σύνολο: {2, 4, 8, 16, 32, 64, 128 και 256}. Τελικά επιλέχθει η παράμετρος που κατέδειξε την καλύτερη απόδοση υπό την έννοια των TDR και FDR. Όπως μπορούμε να δούμε στον Πίνακα 6.2 η προσέγγιση της ταξινόμησης παρουσίασε τα χειρότερα αποτελέσματα και για τα τρία σενάρια. Αυτό προκαλείται από την υψηλή μεταβλητότητα των δειγμάτων των ακουστικών κατηγοριών που εμποδίζει την αποδοτικότητα των διαδικασιών διαμόρφωσης και αναγνώρισης. Το κύριο μειονέκτημα είναι ότι το σχέδιο διαμόρφωσης εστιάζει στα χαρακτηριστικά που είναι κοινά μεταξύ των δειγμάτων μιας συγκεκριμένης κατηγορίας.



Σχήμα 6.4: Ένα δείγμα πανικού από το σενάριο ασφάλειας και οι αντίστοιχες log-likelihoods χρησιμοποιώντας τη μέθοδο του καθολικού HMM. Το δείγμα ξεκινά με τυπική ομιλία και συνεχίζει με αλλαγές μεταξύ τυπικής ομιλίας και εκφράσεις θυμού,3 παραδείγματα από τα ηχογραφημένα δείγματα

Οι μέθοδοι ανίχνευσης καινοτομίας μπορούν να ξεπεράσουν αυτό το μειονέκτημα δεδομένου ότι το διάστημα των log-likelihood που καθορίζεται από την τιμή των κατώτατων ορίων περιλαμβάνει όλα τα δείγματα από την κανονική ακουστική κατηγορία ανεξάρτητα από τα κοινά χαρακτηριστικά τους. Κατά συνέπεια τα

ανώμαλα ηχητικά γεγονότα μπορούν να αναγνωρίζονται ευκολότερα. Τα αποτελέσματα αποκαλύπτουν σαφώς την ανωτερότητα της τεχνικής ανίχνευσης καινοτομίας ενάντια σε εκείνη της ταξινόμησης.

Σαν ένα συμπληρωματικό τρόπο αξιολόγησης εφαρμόκεται η πιθανοτική μέθοδο που παρείχε την καλύτερη ακρίβεια σε τρεις σκηνές που επιλέχτηκαν από όλα τα σενάρια. Το Σχήμα 6.4 απεικονίζει τα αντίστοιχα αποτελέσματα. Οι επιλεγμένες σκηνές περιλαμβάνουν και τυπικά (π.χ. περιβαλλοντικός θόρυβος) και μη τυπικά ηχητικά γεγονότα (π.χ. κραυγή, πανικός). Όπως μπορούμε να δούμε όταν εμφανίζεται μια ανωμαλία, η log-likelihood βρίσκεται κάτω από το προκαθορισμένο κατώφλι και άρα προσδιορίζεται σωστά. Παρατηρούμε επίσης έναν μικρό αριθμό εσφαλμένων συναγερμών. Η πλειοψηφία τους απορρίπτεται από το σχέδιο εξομάλυνσης. Η απόκλιση της log-likelihood από την τιμή των κατώτατων ορίων μπορεί να θεωρηθεί ως επίπεδο ανωμαλίας. Σε περίπτωση που υπάρχει μια μικρή διαφορά από αυτή τιμή κατωφλίου, η κατάσταση χαρακτηρίζεται ως ανησυχητική, ενώ μια μεγάλη απόκλιση δείχνει μια κατάσταση εκτάκτου ανάγκης.

Κεφάλαιο 7

Εκμετάλλευση της Χρονικής Συγχώνευσης Χαρακτηριστικών για Κατηγοριοποίηση Γενικευμένου Ακουστικού Σήματος

Σε αυτό το κεφάλαιο παρουσιάζεται μια μεθοδολογία που ενσωματώνει τη χρονική συγχώνευση χαρακτηριστικών γνωρισμάτων με στόχο την αυτοματοποιημένη αναγνώριση γενικευμένου ακουστικού σήματος . Ένα τέτοιο σύστημα είναι μεγάλης χρησιμότητας όσον αφορά στην ανάλυση και την κατανόηση σκηνής ενώ βασιζόμαστε στις ακουστικές πληροφορίες. Αξιολογείται η απόδοση τριών συνόλων χαρακτηριστικών γνωρισμάτων βασισμένων στη τράπεζα φίλτρων Mel, το ακουστικό πρωτόκολλο MPEG-7 καθώς και την αποσύνθεση κυματιδίου. Επιπλέον ερευνήσαμε την εφαρμογή της χρονικής συγχώνευσης (temporal feature integration) χρησιμοποιώντας τις ακόλουθες τρεις διαφορετικές στρατηγικές: α) βραχύχρονες στατιστικές μετρήσεις (short term statistics), β) φασματικές στιγμές (spectral moments) και γ) αυτοπαλινδρομικά μοντέλα (autoregressive models). Η πειραματική οργάνωση εξηγείται λεπτομερώς και βασίζεται στην ταυτόχρονη χρήση επαγγελματικών συλλογών ακουστικών σημάτων. Με αυτόν τον τρόπο προσπαθούμε να διαμορφώσουμε μια αντιπροσωπευτική εικόνα των χαρακτηριστικών και των δέκα ηχητικών κατηγοριών. Κατά τη διάρκεια της πρώτης φάσης της υλοποίησης μας η διαδικασία της ακουστικής ταξινόμησης επιτυγχάνεται μέσω στατιστικών προτύπων (HMMs), ενώ

ένα σχέδιο μίξης που εκμεταλλεύεται τα πρότυπα που κατασκευάζονται από διαφορετικά σύνολα χαρακτηριστικών παρείχε το υψηλότερο μέσο ποσοστό αναγνώρισης. Το προτεινόμενο σύστημα όχι μόνο χρησιμοποιεί διαφορετικές ομάδες ηχητικών παραμέτρων αλλά υιοθετεί τα πλεονεκτήματα της χρονικής συγχώνευσης ακουστικών παραμέτρων.

7.1. Εισαγωγή

Οι άνθρωποι έχουν τη δυνατότητα να ανιχνεύσουν και να αναγνωρίσουν ένα ηχητικό γεγονός σχετικά εύκολα. Επιπλέον μπορούμε να επικεντρωθούμε σε ένα ιδιαίτερο ηχητικό γεγονός, απομονώνοντας το από τον παρασιτικό θόρυβο, π.χ. εστίαση σε μια συνομιλία ενώ παίζει δυνατή μουσική. Κατά τη διάρκεια των τελευταίων δεκαετιών έχει δοθεί έμφαση επάνω στις μεθόδους για την αυτοματοποιημένη αναγνώριση ομιλίας/ομιλητών. Αυτό οφείλεται στο γεγονός ότι η ομιλία διαδραματίζει έναν σημαντικό ρόλο όσον αφορά και στην αλληλεπίδραση ανθρώπου-με-άνθρωπο και ανθρώπου-με-μηχανή. Ενώ αυτή η περιοχή έχει φθάσει στην ωριμότητα της προώθησης εμπορικών προϊόντων, ο τομέας της μη-λεκτικής ακουστικής επεξεργασίας χρειάζεται ακόμα προσοχή δεδομένου ότι έχει τη δυνατότητα να παρέχει λύσεις σε διάφορες εφαρμογές. Η περιοχή της ακουστικής αναγνώρισης κυριαρχείται αυτήν την περίοδο από τις τεχνικές που εφαρμόζονται κυρίως στη τεχνολογία ομιλίας (Foote, 1999). Αυτό το γεγονός είναι βασισμένο στην υπόθεση ότι όλα τα ακουστικά σήματα μπορούν να υποβληθούν σε επεξεργασία κατά τρόπο κοινό, ακόμα κι αν εκπέμπονται από διαφορετικές πηγές. Γενικά, ο στόχος της τεχνολογίας αναγνώρισης γενικευμένου ακουστικού σήματος είναι η κατασκευή ενός συστήματος που μπορεί αποτελεσματικά να αναγνωρίσει το περιβάλλοντα χώρο απλώς εκμεταλλευόμενο τις εισερχόμενες ακουστικές πληροφορίες (υπολογιστική ακουστική ανάλυση σκηνής (Wang et al.,2006)). Κάθε ηχητική πηγή εκπέμπει ένα συνεπές ακουστικό σχέδιο που οδηγεί σε έναν συγκεκριμένο τρόπο κατανομής της ενέργειάς του στο περιεχόμενο της συχνότητας του. Αυτό το μοναδικό σχέδιο μπορεί να ανακαλυφθεί και να μοντελοποιηθεί με τη χρησιμοποίηση στατιστικών αλγορίθμων αναγνώρισης προτύπων. Εντούτοις υπάρχουν ποικίλα εμπόδια που πρέπει να αντιμετωπιστούν όταν λειτουργεί ένα τέτοιο σύστημα υπό όρους πραγματικού κόσμου. Όταν πρέπει να εξετάσουμε έναν μεγάλο αριθμό διαφορετικών ηχητικών κατηγοριών, η απόδοση αναγνώρισης μειώνεται. Επιπλέον, η κατηγοριοποίηση των ήχων σε ευδιάκριτες κατηγορίες είναι μερικές φορές διφορούμενη (μια ακουστική κατηγορία μπορεί να επικαλύψει μια άλλη) ενώ υπάρχει η πιθανότητα οι σύνθετες πραγματικές ακουστικές σκηνές να είναι πολύ δύσκολο να αναλυθούν. Αυτό το γεγονός έχει οδηγήσει σε λύσεις που στοχεύουν συγκεκριμένα προβλήματα ενώ ένα γενικό σύστημα είναι ακόμα ένα ανοικτό ερευνητικό αντικείμενο.

Τα τελευταία χρόνια, η τεχνολογία ταξινόμησης ακουστικών σημάτων έχει χρησιμοποιηθεί ευρέως για τις ανάγκες διάφορων αναδυόμενων πραγματικών

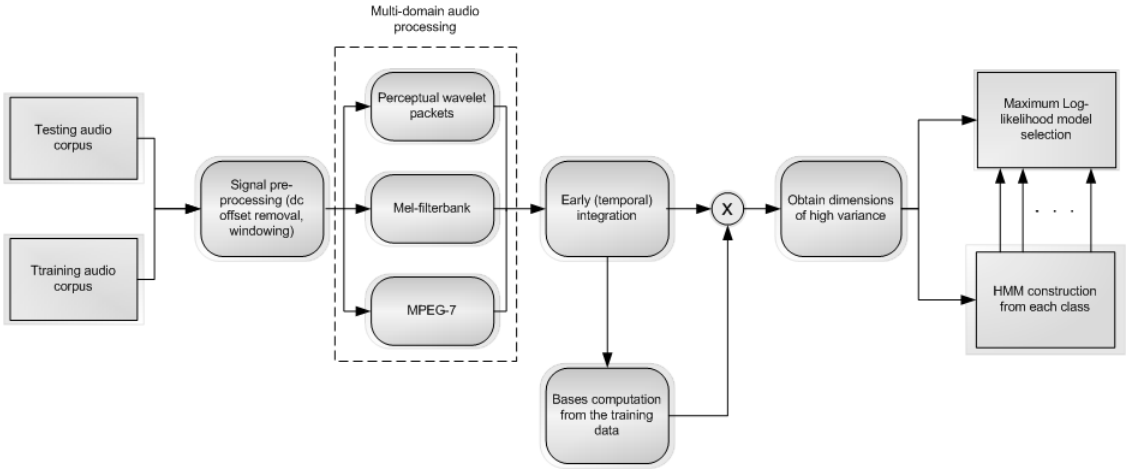
εφαρμογών, όπως ο περιβαλλοντικός έλεγχος, οι διάφορες εφαρμογές βιοακουστικής, η ακουστική επιτήρηση, οι εφαρμογές στη μουσική, η συνειδητοποίηση περιβάλλοντος από ρομπότ κ.λπ. (Watson et al., 2003; Gaston et al., 2004; Lee et al., 2006; Rouas et al., 2006; Tzanetakis et al., 2002; Chu et al., 2006). Ο σκοπός αυτής της εργασίας είναι η εκτενής αξιολόγηση ηχητικών παραμέτρων διαφορετικών περιοχών και των ιδιοτήτων τους για τον προσδιορισμό μιας ευρείας ποικιλίας ηχητικών κατηγοριών. Επιπλέον υιοθετούνται τρεις τύποι μεθοδολογιών οι οποίες επιτυγχάνουν χρονική συγχώνευση. Αναλύουμε αρχικά την απόδοσή τους πριν τις χρησιμοποιούμε για να λύσουμε ένα πραγματικό πρόβλημα. Το πιο κοντινό άρθρο στην εργασία μας είναι το (Casey, 2001) που εξετάζει το ακουστικό πρότυπο MPEG-7 σχετικά με την ταξινόμηση δέκα ηχητικών κατηγοριών. Εξηγείται ο περιγραφέας που δίνει μία χαμηλού επιπέδου προβολή του φάσματος. Στη συνέχεια συνδυάζεται με μια παραγωγική προσέγγιση (κρυμμένα μοντέλα Markov), ενώ η ακουστική βάση αποτελέστηκε από ηχητικά εφέ. Με μια συμβατική εκτίμηση της μέγιστης log-likelihood, ορίζεται μια κατηγορία σε όλα τα δείγματα δοκιμής και επιτυγχάνονται υψηλά ποσοστά αναγνώρισης. Στην εργασία (Kim et al., 2004) αξιολογούν την απόδοση των περιγραφέων του πρωτοκόλλου MPEG-7 σχετικά με την προβολή (Audio Spectrum Projection) ακουστικού φάσματος, η οποία εξάγεται με τη χρησιμοποίηση τριών μεθόδων αποσύνθεσης βάσης (principal component analysis, independent component analysis και η non negative matrix factorization) με στόχο την αυτόματη ταξινόμηση sound tracks ταινιών. Παράλληλα υιοθετούνται τα MFCC ενώ οι συναρτήσεις πυκνότητας πιθανότητας και των δύο συνόλων χαρακτηριστικών υπολογίζονται χρησιμοποιώντας τα συνεχή κρυμμένα μοντέλα Markov. Τα δεδομένα αποκτήθηκαν από μια λεκτική βάση δεδομένων καθώς και μια βιβλιοθήκη γενικών ηχητικών σημάτων. Καταλήγουν στο συμπέρασμα ότι τα MFCC καταδεικνύουν καλύτερη απόδοση κάτω από διάφορους πρακτικούς περιορισμούς, όπως απλότητα αλλά και κατανάλωση χρόνου και μνήμης.

Οι επόμενες δύο προσεγγίσεις δεν υιοθετούν μια παραγωγική τεχνική αναγνώρισης προτύπων αλλά είναι βασισμένες είτε σε μέτρα απόστασης είτε σε ευρετικά μέτρα. Οι (Wold et al., 1996) παρουσιάζουν ένα πλαίσιο για ακουστική ταξινόμηση χρησιμοποιώντας ποικίλα ακουστικά χαρακτηριστικά γνωρίσματα (ηχηρότητα, θεμελιώδης συχνότητα, φωτεινότητα (brightness), εύρος ζώνης και harmonicity). Οι μήτρες μέσων και συνδιακύμανσης τους υπολογίζονται από το σύνολο εκπαίδευσης ενώ οι ήχοι δοκιμής ταξινομούνται χρησιμοποιώντας δύο μέτρα απόστασης (σταθμισμένο L2 ή Ευκλείδεια απόσταση). Τα ακουστικά στοιχεία διάφορων ηχητικών πηγών σε συνδυασμό με βιβλιοθήκες που περιέχουν ήχους μουσικών οργάνων χρησιμοποιούνται για να την κατάρτιση διάφορων ηχητικών κατηγοριών που αντιπροσωπεύουν ζώα, μηχανές, μουσικά όργανα, ομιλία και περιβαλλοντικούς ήχους. Ένα online ακουστικό σύστημα ανάλυσης εξηγείται στην εργασία (Zhang et al., 1998) όπου οι ακουστικές καταγραφές προσδιορίζονται ως ομιλία, μουσική, σιγή καθώς και διάφοροι τύποι περιβαλλοντικών ήχων. Οι συντάκτες χρησιμοποίησαν τα στατιστικά και

μορφολογικά χαρακτηριστικά γνωρίσματα των χρονικών καμπύλων της ενεργειακής κατανομής, του ποσοστού zero-crossing και της θεμελιώδης συχνότητας ενώ ο προσδιορισμός βασίζεται σε μια ευρετική διαδικασία κατώτατων ορίων. Ένα άλλο είδος προσέγγισης που προσπαθεί να βελτιστοποιήσει το στάδιο εξαγωγής χαρακτηριστικών γνωρισμάτων όσον αφορά ένα δεδομένο πρόβλημα ταξινόμησης δίνεται στην (Umapathy et al., 2007). Οι συντάκτες εφαρμόζουν δύο μέτρα ανομοιότητας για τους υποχώρους χρόνου-συχνότητας οι οποίοι προσφέρουν την υψηλότερη δυνατότητα διάκρισης. Η έκβαση του αλγορίθμου τους είναι η κατασκευή ενός νέου δέντρου πακέτων wavelet. Στη συνέχεια, βάσει αυτού του δέντρου, εξάγονται τα χαρακτηριστικά γνωρίσματα και στέλνονται σε έναν γραμμικό διαχωριστικό ταξινομητή που περιλαμβάνει μια ιεραρχία τριών επιπέδων για τη ταξινόμηση των ακουστικών σημάτων σε δέκα κατηγορίες. Η ακουστική βάση δεδομένων αποτελείται από 213 ακουστικά σήματα που διαιρούνται σχεδόν εξίσου μεταξύ τεχνητών (113) και φυσικών (100) ήχων.

Αν και το ζήτημα της αναγνώρισης γενικευμένου ακουστικού σήματος έχει αντιμετωπιστεί από αρκετές μελέτες, η χρονική συγχώνευση των χαρακτηριστικών γνωρισμάτων έχει καλυφθεί μόνο από μερικές, οι οποίες στρέφονται κυρίως στην επεξεργασία μουσικών ακουστικών σημάτων. Στην εργασία (Meng et al., 2008) ερευνάται η χρήση δύο μεθόδων συγχώνευσης (απλές στατιστικές μετρήσεις και αυτοπαλινδρομικά μοντέλα) για την ταξινόμηση μουσικών ειδών. Το σύνολο δεδομένων τους διαιρείται σε δύο μέρη: α) 100 ηχητικά clips που διανέμονται εξίσου μεταξύ rock, κλασσικής, λαϊκής, τζαζ και techno μουσικής και β) 1210 ηχητικά clips αντιπροσωπευτικά 11 υφών μουσικής. Τέσσερις ταξινομητές χρησιμοποιήθηκαν (γραμμικός τρόπος, Γκαουσσιανό μοντέλο με πλήρη συνδιακύμανση, το GMM με πλήρη συνδιακύμανση και ένα γενικευμένο γραμμικό μοντέλο) που εκπαιδεύθηκαν με τους πρώτους έξι συντελεστές των παραμέτρων MFCC. Οι (Joder et al., 2009) χρησιμοποίησαν και early (στο επίπεδο των χαρακτηριστικών) και late (στο επίπεδο των ταξινομητών) μεθοδολογίες χρονικής συγχώνευσης απευθυνόμενοι στο πρόβλημα της αναγνώρισης μουσικών οργάνων σε σόλο μουσικές εγγραφές. Συνολικά υπολογίζονται 162 χαρακτηριστικά γνωρίσματα διαφορετικών περιοχών που τροφοδοτούν τον αλγόριθμο επιλογής χαρακτηριστικών του Fisher. Για την αναγνώριση προτύπων χρησιμοποίησαν τις διανυσματικές μηχανές υποστήριξης (SVM) και τα κρυμμένα μοντέλα Markov ενώ η βάση δεδομένων τους περιείχε καταγραφές 8 διαφορετικών οργάνων οι οποίες αντιπροσωπεύουν τις κύριες κατηγορίες οργάνων. Η κύρια συμβολή της παρούσας εργασίας είναι η εφαρμογή της χρονικής συγχώνευσης των ακουστικών γνωρισμάτων στην περίπτωση της αναγνώρισης γενικευμένου ηχητικού σήματος. Δέκα ακουστικές κατηγορίες οργανώθηκαν ενώ αξιολογήθηκαν σύνολα χαρακτηριστικών διαφορετικών περιοχών (multi domain). Η βάση δεδομένων μας είναι λεπτομερής και συνοπτική μετά από συνδυασμό διάφορων καλά τεκμηριωμένων επαγγελματικών ηχητικών συλλογών που περιέχουν τον ήχους υψηλής ποιότητας. Μια πλήρης εξήγηση δίνεται στην ενότητα 7.4 ενώ πιστεύουμε ότι υπάρχει ανάγκη για μια

γενική βάση ακουστικών δεδομένων αναφοράς προκειμένου να συγκρίνονται τα αποτελέσματα διαφορετικών προσεγγίσεων. Εκτός από το ακουστικό πρότυπο MPEG-7 και τη τράπεζα φίλτρων Mel, ερευνήσαμε μια νέα μέθοδο που βασίζεται στη χρήση πολλαπλής ανάλυσης των ακουστικών σημάτων χρησιμοποιώντας τα βασισμένα σε κρίσιμες ζώνες πακέτα wavelet. Το πειραματικό πρωτόκολλο σχεδιάστηκε προσεκτικά ενώ οι παράμετροι κάθε σταδίου επιλέχτηκαν μετά από εκτενείς πειραματισμούς. Τελικά προτείνεται ένα σχήμα μίξης που εκμεταλλεύεται και τις τρεις ομάδες χαρακτηριστικών, όπου η καθεμία συγχωνεύεται στο πεδίο του χρόνου με βέλτιστο τρόπο.



Σχήμα 7.1: Μπλοκ διάγραμμα του συστήματος ταξινόμησης ηχητικών σημάτων.

Ο κύριος στόχος μας είναι να μελετήσουμε και να καταλάβουμε την επίδραση της χρονικής συγχώνευσης των ακουστικών παραμέτρων που ανήκουν σε διαφορετικές περιοχές - συχνότητα και κυματίδιο - για την ταξινόμηση των γενικών ηχητικών γεγονότων. Χρησιμοποιώντας τα αποτελέσματα αυτής της μελέτης πρέπει να είμαστε σε θέση να εφαρμόσουμε τις τεχνικές αυτές σε διαφορετικές εφαρμογές που εμπεριέχουν την τεχνολογία αναγνώρισης γενικευμένου ακουστικού σήματος.

Το υπόλοιπο αυτού του κεφαλαίου οργανώνεται ως εξής: στην ενότητα 7.2 δίνεται μια πλήρης επισκόπηση του συστήματος μαζί με μια περιγραφή όλων των συνόλων των ακουστικών παραμέτρων. Η ενότητα 7.3 περιγράφει τις μεθόδους χρονικής συγχώνευσης και η ενότητα 7.4 εξηγεί το πειραματικό πρωτόκολλο και εκθέτει τα λεπτομερή αποτελέσματα της ταξινόμησης, ενώ τα συμπεράσματά μας συνάγονται στην παράγραφο 7.5.

7.2. Ανάλυση του σχεδιασμού του συστήματος

Σε αυτό το τμήμα παρέχουμε τις λεπτομέρειες σχετικά με τη σχεδίαση του συστήματος αναγνώρισης ακουστικών σημάτων. Στο Σχήμα 7.1 παρουσιάζονται οι διαδικασίες εκπαίδευσης και κατηγοριοποίησης του προτεινόμενου συστήματος. Μετά

από την αφαίρεση τη μέση τιμής, τα ηχητικά δείγματα τεμαχίζονται σε επικαλυπτόμενα πλαίσια

όπου πάνω στα οποία εφαρμόζονται οι αλγόριθμοι εξαγωγής χαρακτηριστικών γνωρισμάτων. Σε αυτό το κεφάλαιο εξετάζουμε την ακουστική ανάλυση χρησιμοποιώντας διαφορετικά πεδία, κατά συνέπεια υπολογίζονται παράμετροι που προέρχονται από τα πεδία του χρόνου, της συχνότητας και του μετασχηματισμού wavelet. Οι διαφορετικοί συντελεστές χαρακτηριστικών γνωρισμάτων δεν δημιουργούν ένα συγκεντρωμένο διάνυσμα αλλά χρησιμοποιούνται παράλληλα για την κατασκευή τριών χωριστών προτύπων για κάθε κατηγορία ήχων. Στη συνέχεια χρησιμοποιούνται τρεις τύποι χρονικής συγχώνευσης: α) στατιστικός β) φασματικός και γ) δύο αυτοπαλινδρομικές συναρτήσεις. Εφαρμόζονται και στις τρεις ομάδες γνωρισμάτων ενώ δημιουργούνται πιθανοτικά πρότυπα για κάθε ομάδα και για κάθε μεθοδολογία χρονικής ολοκλήρωσης. Σε αυτή τη φάση εφαρμόστηκε μια τεχνική μείωσης της διασταλτικότητας (ανάλυση κύριων τμημάτων - PCA) στις στιγμιαίες τιμές των χαρακτηριστικών μετά από την τυποποιημένη σύσταση του πρωτοκόλλου MPEG-7. Μια βάση, αποκαλούμενη ακουστική βάση φάσματος (Audio Spectrum Basis) δημιουργείται από τα στοιχεία κατάρτισης για όλες τις ηχητικές κατηγορίες. Αυτή η φάση εξυπηρετεί επίσης τη μείωση της υπολογιστικής πολυπλοκότητας που παρεμβάλλεται κατά τη διάρκεια της δημιουργίας των στατιστικών προτύπων. Η τεχνική PCA έχει τη δυνατότητα να διατηρήσει αποτελεσματικά τη διασπορά των δεδομένων χρησιμοποιώντας έναν σχετικά μικρό αριθμό από τους συντελεστές των γνωρισμάτων. Η προσέγγιση που περιγράφεται εδώ είναι ισοδύναμη με τον προσδιορισμό ενός συνόλου ανεξάρτητων ακουστικών παραμέτρων για τη συγκεκριμένη εφαρμογή, αντί της επιλογής των καλύτερων μεμονωμένων παραμέτρων και του μετέπειτα συνδυασμού τους.

Η συνάρτηση πυκνότητας πιθανότητας των ακουστικών χαρακτηριστικών κάθε κατηγορίας προσεγγίζεται από τα κρυμμένα μοντέλα Markov (Rabiner, 1999). Τα HMMs αποτελούν μια ισχυρή τεχνική όχι μόνο για τις στατικές πτυχές μιας ακολουθίας χαρακτηριστικών αλλά και της χρονικής συμπεριφοράς της. Τελικά η πρόβλεψη στα δείγματα δοκιμής γίνεται με την επιλογή της συναρτήσεις πυκνότητας που δίνει ως αποτέλεσμα την υψηλότερη πιθανότητα, κάτι που εκφράζει το πόσο πιθανό είναι να έχει παραγάγει την συγκεκριμένη ακολουθία γνωρισμάτων. Η επόμενη παράγραφος αναλύει τις διαδικασίες που ακολουθήθηκαν κατά τη διάρκεια των διαφορετικών μεθόδων εξαγωγής χαρακτηριστικών.

7.2.1. Οι μεθοδολογίες εξαγωγής των ακουστικών παραμέτρων

Υπολογίστηκαν τρεις τύποι ακουστικών παραμέτρων: α) επιλέχτηκε η τράπεζα φίλτρων Mel λόγω της δυνατότητάς της να δίνει έμφαση στις σημαντικότερες πληροφορίες όσον αφορά στην ανθρώπινη αντίληψη, β) το πρότυπο MPEG-7 το οποίο

θεωρείται αυτήν την περίοδο η state of the art μεθοδολογία για την αυτόματη αναγνώριση ήχων που βασίζεται στο περιεχόμενο ενώ το γ) το τρίτο σύνολο βασίζεται στην επεξεργασία πολλαπλής ανάλυσης (multiresolution). Οι παράμετροι που χρησιμοποιήθηκαν (μέγεθος πλαισίου, επικάλυψη, μέγεθος του FFT) ήταν ίδιες έτσι ώστε να επιτύχουμε μια αξιόπιστη σύγκριση μεταξύ των συνόλων. Εντούτοις μια άμεση σύγκριση μεταξύ MFCC και του περιγραφέα του MPEG-7 δεν θα ήταν δίκαιη δεδομένου ότι μια τεχνική που εξαρτάται από τα δεδομένα (PCA) περιλαμβάνεται κατά τη διάρκεια του υπολογισμού του ASP. Ως εκ τούτου, αλλάξαμε τον αλγόριθμο όσον αφορά στην εξαγωγή MFCC και αντικαταστήσαμε το συγκεκριμένο στάδιο (DCT) με την τεχνική PCA, κάτι το οποίο εμπνεύστηκε από την ακουστική βάση φάσματος (ASB). Η PCA χρησιμοποιήθηκε επίσης για την εξαγωγή της τρίτης ομάδας παραμέτρων. Χρησιμοποιήθηκαν οι επόμενες τρεις ομάδες ακουστικών χαρακτηριστικών:

- Διάνυσμα γνωρισμάτων βασισμένο στην τράπεζα φίλτρων Mel
- Audio Spectrum Projection
- Ανάλυση επιφάνειας των πακέτων κυματιδίων κρίσιμων ζωνών

7.2.2. Δημιουργία στατιστικών μοντέλων

Η αυτόματη αναγνώριση ήχων είναι βασισμένη στην υπόθεση ότι κάθε ηχητικό γεγονός ακολουθεί ένα ευδιάκριτο μοτίβο σε διαφορετικές συχνότητες, κάτι το οποίο καλείται συχνά η *ακουστική υπογραφή* του συγκεκριμένου σήματος. Τα προαναφερθέντα διανύσματα ακουστικών χαρακτηριστικών προσπαθούν να συλλάβουν αυτή την ιδιότητα και, στη συνέχεια, μπορούν να χρησιμοποιηθούν από στατιστικούς αλγορίθμους αναγνώρισης προτύπων προκειμένου να χρησιμοποιηθούν για να ταξινομήσουν άγνωστα ηχητικά γεγονότα. Μια ισχυρή τεχνική που προσεγγίζει τη συνάρτηση πυκνότητας πιθανότητας που ακολουθείται από τις τιμές των χαρακτηριστικών είναι τα κρυμμένα μοντέλα Markov. Με αυτήν την διαδικασία ένα πιθανοτικό πρότυπο κατασκευάζεται για κάθε κατηγορία ήχων χρησιμοποιώντας τα δεδομένα κατάρτισης. Αυτό το μοντέλο περιέχει την *a-priori* γνώση που έχουμε για την κατηγορία και εφ' όσον τα δεδομένα είναι αντιπροσωπευτικά της συγκεκριμένης κλάσης, το μοντέλο θεωρείται μια επαρκής περιγραφή τέτοιων ακουστικών γεγονότων. Αντίθετα από τα μίγματα Γκαουσσιανών προτύπων που δεν έχουν τη δυνατότητα να μοντελοποιήσουν τη χρονική εξέλιξη ενός ήχου, τα HMMs χωρίζουν την ακολουθία γνωρισμάτων σε έναν προκαθορισμένο αριθμό καταστάσεων και μαθαίνουν τις σχέσεις μεταξύ τους. Αυτό οδηγεί σε μια μήτρα μετάβασης *k*x*k* ενώ καθένα εκ των στοιχείων του αντιπροσωπεύει τη πιθανότητα μετάβασης ανάμεσα σε διαφορετικές καταστάσεις. Κατά συνέπεια, το στοιχείο (*i, j*) είναι η πιθανότητα της κίνησης προς την κατάσταση *j* στο χρόνο *t*+1 με δεδομένη την κατάσταση *i* στο χρόνο *t*. Στο τρέχον κεφάλαιο χρησιμοποιούμε από τα αριστερά προς τα δεξιά HMMs που σημαίνει ότι δεν υπάρχει κανένας κατευθυνόμενος βρόχος στην αυτοματοποίηση ενώ η κατανομή κάθε

κατάστασης μοντελοποιείται από ένα GMM με διαγώνια μήτρα συνδιακύμανσης. Κατά τη διάρκεια της ταξινόμησης τα εκπαιδευμένα πρότυπα χρησιμοποιούνται για τον υπολογισμό ενός βαθμού ομοιότητας

(π.χ. log- likelihood) μεταξύ του κάθε προτύπου και ενός άγνωστου σήματος εισόδου. Το πρότυπο που παράγει την υψηλότερη πιθανότητα αποτελεί την πρόβλεψη του συστήματος σχετικά με το συγκεκριμένο σήμα. Αυτή η τεχνική αναγνώρισης προτύπων ανήκει στις παραγωγικές προσεγγίσεις, των οποίων η κύρια ιδιότητα είναι ότι διαχειρίζονται τα δείγματα κάθε κατηγορίας ανεξάρτητα από αυτά των υπολοίπων.

Η υλοποίηση Torch (διαθέσιμη στη διεύθυνση http://www.torch.ch) των HMM, που γράφτηκε σε γλώσσα C++ χρησιμοποιήθηκε κατά τη διάρκεια της εκπαίδευσης καθώς και της δοκιμής. Ο μέγιστος αριθμός επαναλήψεων του αλγορίθμου *k-means* για την έναρξη ήταν 50 ενώ ο αλγόριθμος *Baum-Welch* είχε ανώτερο όριο ίσο με 25 επαναλήψεις ενώ το κατώτατο όριο μεταξύ δύο συνεχόμενων επαναλήψεων ήταν ίσο με 0.001. Εκτενείς πειραματισμοί πραγματοποιήθηκαν σχετικά με: α) την κατασκευή του προτύπου κάθε κατηγορίας ήχων όσον αφορά κάθε σύνολο χαρακτηριστικών γνωρισμάτων, β) την εξέταση κάθε μεθόδου χρονικής συγχώνευσης καθώς επίσης και γ) την αποτελεσματικότερη απόφαση σχετικά με το μέγεθος του χρονικού παραθύρου (ο αριθμός των πλαισίων που συγχωνεύθηκαν). Πιο συγκεκριμένα οι πιθανοί αριθμοί καταστάσεων ήταν μεταξύ 3 και 7 ενώ οι αριθμοί των Γκαουσσιανών συναρτήσεων που δοκιμάστηκαν αντίστοιχα ήταν: {2, 4, 8, 16, 32, 64, 128}. Οι τελικές τιμές αυτών των παραμέτρων επιλέχτηκαν χρησιμοποιώντας το κριτήριο του υψηλότερου ποσοστού αναγνώρισης.

7.3. Στρατηγικές χρονικής συγχώνευσης γνωρισμάτων

Πρόσφατα έχει γίνει κοινή πρακτική η εκπαίδευση και η δοκιμή ενός συστήματος ταξινόμησης ακουστικών σημάτων να γίνεται χρησιμοποιώντας ανάλυση πλαίσιο ανά πλαίσιο (π.χ. Aucoutourier et al., 2007). Αν και αυτό το είδος επεξεργασίας φαίνεται ότι παρέχει σχετικά επαρκή αποτελέσματα (Maleh et al., 1999), θα ήταν μεγάλου ενδιαφέροντος να πειραματιστούμε με πιο συμπαγή καθώς επίσης και βαθμιδωτά ακουστικά πλαίσια επεξεργασίας. Μια τέτοια τεχνική αντιπροσώπευσης σημάτων θα απαιτούσε τη λιγότερη μνήμη για αποθήκευση και την περαιτέρω επεξεργασία ενώ μπορεί να παρέχει μια χαρακτηριστικότερη δομή του σήματος που θέλουμε να αναλύσουμε. Είναι βασισμένο στην μετα-επεξεργασία των χαμηλού επιπέδου χαρακτηριστικών γνωρισμάτων που υπολογίζονται από πλαίσια μικρής διάρκειας. Μέσα σχεδόν σε κάθε ηχητικό δείγμα υπάρχουν μέρη που δεν είναι αντιπροσωπευτικά του συγκεκριμένου γεγονότος. Αυτά τα τμήματα είναι εκείνα που είναι τα πλέον πιθανά να κατηγοριοποιηθούν λανθασμένα. Προσπαθούμε να λύσουμε αυτό το πρόβλημα με την ενσωμάτωση της γνώσης που προσφέρεται από διάφορα

συνεχόμενα πλαίσια σε ένα. Επιπλέον πειραματιζόμαστε πάνω στη βέλτιστη τιμή των πλαισίων που ενσωματώνονται όσον αφορά σύνολα ακουστικών χαρακτηριστικών διαφορετικών πεδίων καθώς επίσης και διάφορων στρατηγικών συγχώνευσης.

Πιο συγκεκριμένα, μελετάμε την επίδραση της χρονικής συγχώνευσης των χαρακτηριστικών προκειμένου να επιτευχθεί μια σφαιρική αντιπροσώπευση της ακουστικής ακολουθίας χρησιμοποιώντας έναν μικρότερο αριθμό χρονικών αναφορών. Με την ενσωμάτωση των γνωρισμάτων υπό τη χρονική έννοια συλλαμβάνουμε μία χαρακτηριστικότερη - σφαιρική εικόνα του σήματος που μπορεί να είναι πιο αντιπροσωπευτική από τις τιμές των πλαισίων μικρής διάρκειας. Κατά συνέπεια η μέση μεταβλητότητα κάθε κατηγορίας μειώνεται, κάτι που οδηγεί στην αποτελεσματικότερη μοντελοποίηση των κοινών χαρακτηριστικών μεταξύ των δειγμάτων της ίδιας ηχητικής κατηγορίας. Το τμήμα του χρόνου στο οποίο πραγματοποιείται η συγχώνευση καλείται *texture window*. Αυτή η τεχνική ανήκει στην κατηγορία της πρόωρης συγχώνευσης δεδομένου ότι η ολοκλήρωση δεν πραγματοποιείται στο επίπεδο των ταξινομητών αλλά στο επίπεδο της εξαγωγής χαρακτηριστικών.Κάθε διαδικασία συγχώνευσης εφαρμόζεται πάνω σε έναν προκαθορισμένο αριθμό πλαισίων και τους μετασχηματίζει σύμφωνα με την ακόλουθη εξίσωση:

$$X_k = F(x_t, \ldots, x_{t+p-1}), \tag{7.1}$$

όπου το X_k δείχνει το ενσωματωμένο διάνυσμα του *K-οστού* texture window και X_i είναι η τιμή του χαρακτηριστικού X στο πλαίσιο t. Ο αριθμός των πλαισίων στον οποίο εφαρμόζεται η συγχώνευση δείχνεται ως p. Αυτή η εξίσωση παρέχει μια υψηλότερου επιπέδου περιγραφή της σειράς των ακουστικών παραμέτρων. Διάφορες στρατηγικές συγχώνευσης είναι βασισμένες στον υπολογισμό των στατιστικών πάνω στο texture window. Άλλες στρατηγικές είναι βασισμένες στην υπόθεση ότι η ακολουθία των γνωρισμάτων μπορεί να αντιμετωπισθεί ως μία τυχαία διαδικασία (π.χ. αυτοπαλινδρομικά πρότυπα). Οι τρεις διαφορετικές στρατηγικές συγχώνευσης που ερευνώνται σε αυτήν την εργασία εξηγούνται παρακάτω.

7.3.1. Υπολογισμός στατιστικών μεγεθών

Ένας σχετικά απλός τρόπος να συγχωνευθούν οι πληροφορίες που παρέχονται από πολλά διαδοχικά πλαίσια σε ένα είναι ο υπολογισμός διάφορων στατιστικών τους. Εξετάζουμε τις επόμενες πέντε στατιστικές μετρήσεις: *μέσος όρος* (ή αναμενόμενη τιμή), *διασπορά, διάμεσος* καθώς επίσης και το πρώτο και το τρίτο εκατοστημόριο (25^{th}

and 75th percentile) για κάθε texture window. Αν και είναι σχετικά απλός ο υπολογισμός τους, μπορούν να είναι αρκετά αντιπροσωπευτικά μιας συγκεκριμένης ακολουθίας χαρακτηριστικών. Εκτός από το μέσο όρο και τη διασπορά, που είναι εξαιρετικά σημαντικά (για παράδειγμα δείτε τις εργασίες Tzanetakis et al., 2001 και Khan et al.,2004) χρησιμοποιούμε και τα τρία εκατοστημόρια. Αντανακλούν στην τιμή που αποτελεί όριο ενός ορισμένου ποσοστού παρατηρήσεων. Το πρώτο, δεύτερο (διάμεσος) και τρίτο εκατοστημόριο αντιστοιχούν σε 25, 50 και 75 τοις εκατό αντίστοιχα. Η βραχυπρόθεσμη συνάρτηση συγχώνευσης με χρήση των στατιστικών μετρήσεων είναι η ακόλουθη:

$$F_{stat}(x_t,\ldots,x_{t+p-1})=[mean(x_t,\ldots,x_{t+p-1}),\mathrm{var}(x_t,\ldots,x_{t+p-1}),q1(x_t,\ldots,x_{t+p-1}),\ldots \\ median(x_t,\ldots,x_{t+p-1}),q3(x_t,\ldots,x_{t+p-1})];$$

(7.2)

Το αποτέλεσμά της είναι ένα διάνυσμα με μέγεθος πέντε φορές την αρχική διάσταση (R=5xD). Το κύριο μειονέκτημα των απλών στατιστικών είναι η ανεπάρκειά τους για να συλλάβουν τη δυναμικότητα ενός ακουστικού σήματος δεδομένου ότι ένας άλλος συνδυασμός διαφορετικών παρατηρήσεων μπορεί να οδηγήσει στο ίδιο ενσωματωμένο διάνυσμα. Οι επόμενες δύο στρατηγικές ολοκλήρωσης μοιράζονται το γεγονός ότι προσπαθούν να συλλάβουν τη χρονική συμπεριφορά μιας δεδομένης σειράς.

7.3.2. Φασματικές στιγμές (Spectral moments)

Η χρονική εξάρτηση μεταξύ των διαδοχικών παρατηρήσεων χαρακτηριστικών μπορεί να εξαχθεί χρησιμοποιώντας τις πληροφορίες που παρέχονται από το φάσμα αυτών των χαρακτηριστικών. Η μέθοδος που υιοθετήθηκε εδώ χρησιμοποιήθηκε και στην εργασία (Meng et al., 2007) με σκοπό την αυτόματη ταξινόμηση μουσικών ειδών ενώ αποτελεί επέκταση της διαμόρφωσης ενέργειας διάφορων χαρακτηριστικών που χρησιμοποιήθηκε από τους (Mc Kinney et al., 2003). Αρχικά υπολογίζεται ο STFT των ακουστικών παραμέτρων με βάση το texture window. Η έκβασή της αποτελεί τη βάση για τον υπολογισμό των φασματικών στιγμών και περιλαμβάνει όλες τις πληροφορίες που παρέχονται σχετικά με το φάσμα κάθε γνωρίσματος. Κατά αυτόν τον τρόπο μπορούμε να προσδιορίσουμε την ημιτονοειδή συχνότητα και το περιεχόμενο της φάσης των τοπικών τμημάτων μιας δεδομένης ακολουθίας χαρακτηριστικών καθώς αυτή αλλάζει με την πάροδο του χρόνου. Πρέπει να σημειωθεί ότι εδώ μια άλλη παράμετρος παρεμβάλλεται, το μέγεθος του FFT που είναι άσχετο με το FFT που υιοθετείται από τους αλγορίθμους εξαγωγής γνωρισμάτων και στην ουσία είναι ο αριθμός των πλαισίων που μπορούν να περιληφθούν στο texture window.

Καταρχάς υπολογίζεται το φάσμα ενέργειας της σειράς ενός ιδιαίτερου περιγραφέα σε dB και αποθηκεύεται η μέση αξία του *μ*. Στη συνέχεια υπολογίζονται οι επόμενες τέσσερις στατιστικές μετρήσεις του πλάτους του φάσματος ανά texture window: ο μέσος όρος *m*, η διασπορά *ν*, η ασυμμετρία (skewness) *γ* και η κύρτωση *κ*. Οι τελευταίες δύο μετρήσεις λαμβάνονται επειδή αυτοί εκφράζουν τη διασπορά των τιμών του χαρακτηριστικού γύρω από την αναμενόμενη τιμή τους. Εάν η ασυμμετρία είναι αρνητική, τα δεδομένα διαδίδονται περισσότερο προς τα αριστερά του μέσου όρου απ' ότι προς τα δεξιά. Εάν η εκτροπή είναι θετική, τα δεδομένα διαδίδονται περισσότερο προς τα δεξιά. Για μια τέλεια συμμετρική κατανομή η εκτροπή είναι μηδέν. Η κύρτωση περιγράφει την συνάρτηση πυκνότητας πιθανότητας μιας τυχαίας μεταβλητής ενώ δίνει έμφαση στην απόκλιση που εκθέτει η διασπορά της. Στη περίπτωση που η διασπορά παρουσιάζει σπάνιες ακραίες αποκλίσεις η κύρτωση είναι μεγαλύτερη του 3. Αντίθετα όταν η διασπορά παρουσιάζει συχνές μικρές αποκλίσεις, η κύρτωση χαρακτηρίζεται από χαμηλότερες τιμές. Το τελικό ενσωματωμένο διάνυσμα έχει πέντε φορές μεγαλύτερη διάσταση από την αρχική όπως και η προηγούμενη στρατηγική (*R*=5x*D*).

$$F_{spec}(x_t,\ldots,x_{t+p-1})=[\mu,m,v,\gamma,\kappa];(7.3)$$

7.3.3. Αυτοπαλινδρομικά μοντέλα (Autoregressive models)

Μια άλλη μεθοδολογία η οποία επιτυγχάνει την συγχώνευση των ακουστικών παραμέτρων, που προτάθηκε στην εργασία (Meng et al., 2007) χρησιμοποιεί AR πρότυπα για να συλλάβει την εξέλιξή τους στο χρόνο. Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι βασισμένοι σε μια σταδιακή προσέγγιση ελαχίστων τετραγώνων με μικρό υπολογιστικό κόστος ακόμα και όταν επεξεργαζόμαστε δεδομένα υψηλής διασταλτικότητας (Schneider et al., 2001). Επιπλέον τα διαστήματα εμπιστοσύνης για τις κατ' εκτίμηση παραμέτρους των μοντέλων δείχνουν το πόσο καλά ένα δημιουργημένο πρότυπο αντιστοιχεί στα συγκεκριμένα δεδομένα. Οι συντελεστές της αυτοπαλινδρομικής διαδικασίας υπολογίζονται για τη διαμόρφωση του συγχωνευμένου διανύσματος χαρακτηριστικών. Δύο τύποι διαδικασιών εξετάζονται σε αυτήν την εργασία: πολλών μεταβλητών αυτοπαλινδρόμηση (Multivariate AR) και διαγώνια αυτοπαλινδρόμηση (Diagonal AR). Ο τύπος για τον υπολογισμό των συντελεστών ενός αυτοπαλινδρομικού προτύπου τάξης Ο παρουσιάζεται παρακάτω

$$x[t]=w+\sum_{n=1}^{O} x[t-n]A_n+e_t,$$

(7.4)

όπου το *w* είναι το intercept vector, A_n είναι οι συνιστώσες-πίνακες *D*x*D* του αυτοπαλινδρομικού μοντέλου και ο e_t είναι το διάνυσμα λευκού θορύβου διάστασης *D*. Συνεπώς το συγχωνευμένου διάνυσμα χαρακτηριστικών είναι:

$$F_{MAR}(x_t,\ldots,x_{t+p-1})=[w,A_1,\ldots,A_O];$$

(7.5)

που είναι διάστασης *R*=*D* (*O*x*D*+1). Οι ίδιες προσεγγίσεις ελαχίστων τετραγώνων υπολογίζονται και στην περίπτωση του DAR αλλά γίνεται μια περαιτέρω υπόθεση: ότι οι ακουστικές παράμετροι είναι ανεξάρτητες η μία από την άλλη. Κατά συνέπεια προκύπτει ο περιορισμός ότι οι συντελεστές του μοντέλου πρέπει να είναι διαγώνιες μήτρες. Ως εκ τούτου, υπολογίζουμε τις παραμέτρους για κάθε χαρακτηριστικό γνώρισμα ξεχωριστά κάθε φορά και τα αποτελέσματα συνενώνονται. Σε αυτήν την περίπτωση έχουμε ένα διάνυσμα σημαντικά χαμηλότερης διάστασης, *R*=*D* (*O*+1).

$$F_{DAR}(x_t,\ldots,x_{t+p-1})=[w,D_1,\ldots,D_O];$$

(7.6)

7.4. Ανάλυση της πειραματικής διαδικασίας και συγκριτική αξιολόγηση

Αυτή η ενότητα καλύπτει τις λεπτομέρειες σχετικά με τις φάσεις δοκιμής και εκπαίδευσης που έλαβαν χώρα κατά τους πειραματισμούς μας. Ο στόχος μας είναι να αξιολογήσουμε την απόδοση των τριών διαφορετικών συνόλων χαρακτηριστικών στην ίδια βάση δεδομένων ενώ έχουν συγχωνευθεί στο πεδίο του χρόνου. Για κάθε στάδιο ταξινόμησης τα από τα αριστερά προς τα δεξιά HMMs βελτιστοποιήθηκαν

όσον αφορά τον αριθμό καταστάσεων και Γκαουσσιανών συναρτήσεων. Τα στοιχεία χωρίστηκαν σε 70% για κατάρτιση και 30% για δοκιμή με τυχαίο τρόπο ενώ αυτές οι ακολουθίες ήταν ίδιες για όλα τα στάδια. Η βάση μας αποτελείται από ηχητικά γεγονότα επαγγελματικών ηχητικών συλλογών υψηλής ποιότητας, οι οποίες υιοθετούνται κυρίως από τη κινηματογραφική βιομηχανία. Χρησιμοποιούνται για να επεξεργαστούν ή ακόμα και να αντικαταστήσουν την ακουστική ροή που καταγράφηκε πραγματικά στη σκηνή. Ο συνδυασμός αυτών των πηγών περιλαμβάνει μεγάλη ποικιλία φωνητικών και μη- φωνητικών ακουστικών γεγονότων που μπορούν να χρησιμοποιηθούν για την εκπαίδευση πιθανοτικών μοντέλων ταξινόμησης. Πρέπει να υπογραμμίσουμε το γεγονός ότι υπάρχει ανάγκη για μια κοινή βάση δεδομένων προκειμένου η κοινότητα να είναι σε θέση να συγκρίνει άμεσα την απόδοση μεταξύ διαφορετικών συστημάτων. Πιστεύουμε ότι η ακουστική βάση που χρησιμοποιήθηκε εδώ έχει τη δυνατότητα να γίνει μια βάση δεδομένων αναφοράς που είναι απαραίτητη για την αξιόπιστη σύγκριση σχετικών δημοσιεύσεων. Η βάση μας αποτελείται από τις ακόλουθες συλλογές: (i) BBC Sound Effects Library, (ii) Sound Ideas Series 6000, (iii) TIMIT και (iv) Sony Sound Effects Library. Οργανώθηκαν οι ακόλουθες δέκα ακουστικές κατηγορίες: *κελάηδισμα πουλιών, χειροκρότημα, γάβγισμα σκύλου, έκρηξη, βήμα, νιαούρισμα γάτας, πυροβολισμός, ομιλία και των δύο φύλων, γέλιο και χτύπος τηλεφώνου*. Η πρόθεσή μας ήταν να έχουμε όσο το δυνατόν περισσότερες κοινές κατηγορίες με προηγούμενες μελέτες. Ένα σύνολο δεδομένων που θα ήταν πλήρως ίδιο με άλλες δημοσιεύσεις δεν μπορούσε να διαμορφωθεί λόγω των διαφορετικών βάσεων δεδομένων που έχουν χρησιμοποιηθεί σε άλλες εργασίες ή/και της δύσκολης προσβασιμότητας τους. Η κύρια διαφορά είναι ότι αποφασίσαμε να μην χρησιμοποιήσουμε τη κατηγορία σπάσιμο γυαλιού (όπως χρησιμοποιείται στην εργασία Casey, 2001) δεδομένου ότι οι τέτοιου είδους ήχοι είναι παρόντες σε πολλά ηχητικά γεγονότα έκρηξης. Αντ' αυτού, αποφασίσαμε να προσθέσουμε μια άλλη ζωική κατηγορία, αυτή του νιαουρίσματος γάτας. Έχει ληφθεί προσοχή προκειμένου να περιληφθούν οι ήχοι από όλες τις βάσεις δεδομένων και στο σύνολο εκπαίδευσης αλλά και σε αυτό της δοκιμής έτσι ώστε τα πρότυπα να μην εξαρτώνται από τις συνθήκες ηχογράφησης της κάθε βάσης δεδομένων. Τα στατιστικά της τελικής βάσης είναι ταξινομημένα στον Πίνακα 8.1. Τα ηχητικά αρχεία ήταν σε ανάλυση 16 KHz με δεκαεξάμπιτη κβαντοποίηση ενώ προεπεξεργάστηκαν έτσι ώστε να αφαιρεθεί η μέση τιμή της κυματομορφής. Οι βάσεις δεδομένων αναζητήθηκαν εξαντλητικά για δείγματα που αντιστοιχούν στο πρόβλημά μας και όλα τα σχετικά μέρη προσδιορίστηκαν και απομονώθηκαν με σκοπό να συμπεριληφθούν στη τελική βάση. Ο βασικός περιορισμός μας ήταν το δείγμα να είναι «καθαρό» χωρίς οποιοδήποτε τύπο παρασιτικού θορύβου.

Κατηγορία ήχων	Αριθμός ηχογραφήσεων	Διάρκεια (s)
Κελάηδισμα πουλιού	55	7,913.4
Χειροκρότημα	64	1,467.5
Γαύγισμα	102	1,103.6
Έκρηξη	131	1,803.9
Βηματισμός	152	4,865.5
Νιαούρισμα	141	977.1
Πυροβολισμός	187	2,290.8
Ομιλία και των δύο φύλων	1680	5,174.4
Γέλιο	118	941.64
Τηλέφωνο	89	1,629.59
Σύνολο	2719	28,167.4

Πίνακας 7.1: Τα στατιστικά και οι κατηγορίες της τελικής βάσης ακουστικών σημάτων

Στο τέλος χρησιμοποιήθηκε ένας αλγόριθμος βασισμένος σε στατιστικά πρότυπα για την αφαίρεση σιγής που περιγράφεται στην εργασία (Sohn et al.,1999) έτσι ώστε οι τεχνικές εκτίμησης των pdf να μπορούν να επεξεργαστούν μόνο τη δομή ενός συγκεκριμένου ηχητικού γεγονότος.

7.4.1. Παράμετροι για την εξαγωγή χαρακτηριστικών και τη χρονική συγχώνευση

Ακολουθώντας την τυποποιημένη σύσταση του MPEG-7, το χαμηλού επιπέδου παράθυρο εξαγωγής χαρακτηριστικών είναι 30ms με την επικάλυψη στα 10ms, έτσι ώστε το σύστημα είναι ανεκτικό απέναντι σε πιθανά misalignments. Τα δεδομένα κόβονται σύμφωνα με το παράθυρο hamming για να λειανθούν οποιεσδήποτε ασυνέχειες ενώ το μέγεθος του FFT είναι 512. Όσον αφορά στον αριθμό κύριων συστατικών (principal components) που πρόκειται να εξεταστούν βάζουμε έναν περιορισμό ότι πρέπει να κρατηθεί τουλάχιστον το 95% της διασποράς. Ο μικρότερος αριθμός των συστατικών με τον οποίο τηρούνταν ο εν λόγω περιορισμός αποτέλεσε την τελική επιλογή. Με το τρέξιμο ενός πειράματος στα δεδομένα κατάρτισης για κάθε ομάδα περιγραφέων φθάσαμε στα ακόλουθα αποτελέσματα: 15 components για το σύνολο που είναι βασισμένο στη τράπεζα φίλτρων Mel, 16 για το ASP του MPEG-7 και 61 για την PWP

ανάλυση ολοκλήρωσης. Για κάθε πειραματική φάση ένας PCA πυρήνας προήλθε από τα στοιχεία εκπαίδευσης ο οποίος έπειτα χρησιμοποιήθηκε για το μετασχηματισμό των ακολουθιών εξέτασης σε ένα σύστημα συντεταγμένων το οποίο εξαρτάται από τα δεδομένα κατάρτισης.

Η εργαλειοθήκη ARfit (Schneider et al, 2001) χρησιμοποιήθηκε για να υπολογίσουμε τις παραμέτρους των διαδικασιών MAR και DAR. Η εργαλειοθήκη ARfit

είναι μία σειρά συναρτήσεων γραμμένες σε γλώσσα Matlab που έχουν σκοπό την ανάλυση της χρονικής σειράς πολλαπλών μεταβλητών χρησιμοποιώντας αυτοπαλινδρομικές διαδικασίες. Έπειτα, το μέγεθος του FFT σχετικά με την συγχώνευση μιας δεδομένης ακολουθίας χαρακτηριστικών σύμφωνα με τη στρατηγική των φασματική στιγμών τέθηκε ίσο με 128. Κατά αυτόν τον τρόπο το σύστημα μπορεί να ενσωματώσει μέχρι 128 πλαίσια που αντιστοιχεί σε διάρκεια ίση με περίπου 2.5 δευτερόλεπτα. Οι τιμές των πλαισίων που πρόκειται να ενσωματωθούν σε ένα texture window πάρθηκαν από το σύνολο: {10, 20, 30, 40, 50, 60, 90, 120} ενώ υιοθετήθηκε ένα σταθερό βήμα μεγέθους 10 πλαισίων, έτσι ώστε ο τελικός αριθμός των texture window να είναι ίδιος ανεξάρτητα από το συμπεριλαμβανόμενο αριθμό πλαισίων. Σε περίπτωση που το ηχητικό δείγμα είναι μικρότερης διάρκειας η συγχώνευση γίνεται μόνο σε ένα texture window. Πρέπει να αναφερθεί ότι για κάθε πειραματική φάση η απόδοση του συστήματος μετριέται χρησιμοποιώντας ανάλυση ανά πλαίσιο (ή ανά texture window). Για τη MAR μέθοδο το χαμηλότερο όριο των πλαισίων που πρόκειται να ενσωματωθούν είναι 30 δεδομένου ότι αυτή η μέθοδος απαιτεί έναν μεγαλύτερο αριθμό διαδοχικών παρατηρήσεων για τον υπολογισμό των συντελεστών των αντίστοιχων μοντέλων.

7.4.2. Αποτελέσματα κατηγοριοποίησης

Αυτό το τμήμα παρουσιάζει τα αποτελέσματα ταξινόμησης όσον αφορά τα διαφορετικά επίπεδα της μελέτης μας. Συγκρίνουμε αρχικά την απόδοση των συνόλων χαρακτηριστικών που υιοθετήθηκαν για να μοντελοποιήσουν κάθε κατηγορία ήχων καθώς επίσης και τη στρατηγική συγχώνευσης. Στη συνέχεια, συζητείται η επίδραση του μήκους του texture window. Τελικά συνάγουμε τα συμπεράσματά μας όσον αφορά στη στρατηγική συγχώνευσης που παρέχει την καλύτερη απόδοση και από την άποψη της υπολογιστικής πολυπλοκότητας αλλά και του ποσοστού αναγνώρισης.

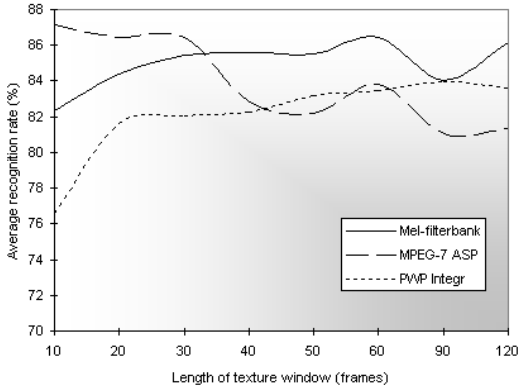
Ομάδα χαρακτηριστικών	**Στρατηγική συγχώνευσης (τάξης *O*)**	**Texture window (frames)**	**No. states**	**No. modes**	**Μέση ακρίβεια αναγνώρισης (%)**
Mel-filterbank	χωρίς συγχώνευση	-	4	64	80.21
	στατιστικά	60	5	128	86.44
	φασματικές στιγμές	90	6	8	79.2
	MAR (1)	50	5	16	71.1
	DAR (1)	60	5	128	85.29
	DAR (2)	90	5	128	83.55
	DAR (3)	10	6	32	76.86
MPEG-7 Audio	χωρίς συγχώνευση	-	5	64	82.06
	στατιστικά	10	3	32	87.13
	φασματικές στιγμές	90	3	64	81.98
	MAR (1)	50	5	32	67.21

spectrum projection	DAR (1)	120	3	16	79.78
	DAR (2)	40	4	16	80.59
	DAR (3)	120	5	32	80.26
PWP Integration analysis	χωρίς συγχώνευση	-	4	32	75.63
	στατιστικά	90	4	32	83.96
	φασματικές στιγμές	20	5	8	83.77
	MAR (1)	40	3	16	69.21
	DAR (1)	90	6	16	78.26
	DAR (2)	120	4	16	79.03
	DAR (3)	120	5	8	80
	DAR (4)	90	4	8	79.31

Πίνακας 7.2: Οι αποδόσεις του συστήματος για κάθε σύνολο ακουστικών παραμέτρων και το παράθυρο συγχώνευσης με τη μεγαλύτερη ακρίβεια.

Η απόδοση του συστήματος όσον αφορά κάθε σύνολο ακουστικών γνωρισμάτων με το αντίστοιχο μήκος texture window που παρείχε το καλύτερο μέσο ποσοστό αναγνώρισης είναι ταξινομημένη στον Πίνακα 7.2. Επίσης απεικονίζονται τα αντίστοιχα αποτελέσματα χωρίς τη χρησιμοποίηση κάποιας μεθοδολογίας συγχώνευσης για λόγους σύγκρισης. Επιπλέον οι παράμετροι των HMMs (αριθμός καταστάσεων και Γκαουσσιανών συναρτήσεων) δίνονται για κάθε περίπτωση. Όπως μπορούμε να παρατηρήσουμε ότι η καλύτερη γενική ακρίβεια επιτυγχάνεται από τον περιγραφέα ASP του MPEG-7 και αντιστοιχεί σε 87.13%. Το σύνολο που είναι βασισμένο στη τράπεζα φίλτρων Mel παρείχε τη δεύτερη καλύτερη απόδοση (86.44%) ενώ η ομάδα που εξήχθη από το πεδίο κυματιδίων κατέδειξε τη χειρότερη απόδοση (83.96%). Τα ποσοστά υπολογίστηκαν κατά μέσο όρο και για τις δέκα κατηγορίες ακουστικών σημάτων έτσι ώστε όλες οι κατηγορίες να συμβάλλουν εξίσου στο τελικό αποτέλεσμα ανεξάρτητα από τον αριθμό των δειγμάτων δοκιμής. Σκεπτόμενοι ότι τα πειράματά μας πάνω σε διάφορες βάσεις δεδομένων οι οποίες εμπεριέχουν δείγματα μεγάλης ποικιλομορφίας, μπορούμε να πούμε ότι τα αποτελέσματα είναι κάτι περισσότερο από ενθαρρυντικά. Πρέπει να σημειωθεί ότι πολλές από τις λάθος ταξινομήσεις εμφανίζονται λόγω της μεγάλης μεταβλητότητας μεταξύ των ηχητικών δειγμάτων της ίδιας κατηγορίας. Επιπλέον, διάφοροι ήχοι είναι ηχητικά παρόμοιοι ακόμα κι αν ανήκουν σε διαφορετικές κατηγορίες, π.χ. πολλές εκρήξεις ηχούν όπως μερικοί πυροβολισμοί και αντίστροφα. Τα αποτελέσματα επιβεβαιώνουν ότι το ακουστικό πρωτόκολλο MPEG-7 παρέχει για κάθε ακουστική κατηγορία μια αντιπροσώπευση που ακολουθεί ένα συνεπές μοτίβο το οποίο μπορεί να μοντελοποιηθεί από τα αριστερά προς τα δεξιά HMMs και στη συνέχεια να χρησιμοποιηθεί για την ταξινόμηση νέων δεδομένων.

Μια ιδιότητα που μοιράζεται από όλες τις ομάδες των ακουστικών παραμέτρων είναι ότι επιδεικνύουν την καλύτερη απόδοσή τους όταν υιοθετείται η βραχυπρόθεσμη μέθοδος στατιστικών για τη χρονική συγχώνευση. Αν και αυτή η μέθοδος είναι σχετικά απλή και δεν λαμβάνει υπόψη της τη πιθανότητα της χρονικής εξάρτησης μεταξύ των τιμών χαρακτηριστικών, αποδείχθηκε ικανή όσον αφορά την ενίσχυση της απόδοσης της αναγνώρισης με έναν τρόπο ο οποίος είναι ανεξάρτητος του πεδίου των



Σχήμα 7.2: Το μέσο ποσοστό αναγνώρισης συναρτήσει του μεγέθους του texture window όσον αφορά όλες τις ομάδες ακουστικών χαρακτηριστικών.

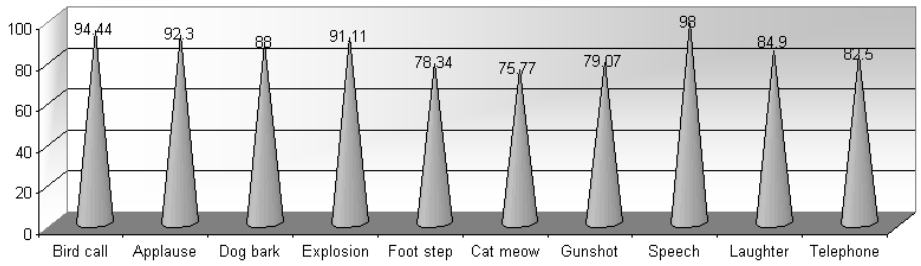
χαρακτηριστικών. Η μέθοδος των φασματικών στιγμών καταδεικνύει τα δεύτερα καλύτερα αποτελέσματα όσον αφορά στα γνωρίσματα PWP- Integration και MPEG-7ASP αντίθετα με το σύνολο της τράπεζας φίλτρων Mel όπου αυτό επιτυγχάνεται με την μέθοδο DAR πρώτης τάξης. Τα χαμηλότερα ποσοστά αναγνώρισης σε όλα τα σύνολα χαρακτηριστικών γνωρισμάτων δίνονται από μέθοδο MAR και είναι τα επόμενα: 71.1%, 67.21% και 69.29% για τις ομάδες Mel, MPEG-7 και PWP αντίστοιχα παρά τον υψηλό αριθμό της διάστασης του τελικού της διανύσματος.

Επιπλέον πρέπει να υπογραμμιστεί ότι οι ταξινομητές που εκπαιδεύονται σε χρονικά συγχωνευμένα δεδομένα αποδίδουν καλύτερα σχεδόν σε όλες τις περιπτώσεις (εξαίρεση αποτελεί η μέθοδος MAR). Αυτό σαφώς αποκαλύπτει ότι τα διανύσματα που ενσωματώνουν πληροφορίες από διαδοχικά πλαίσια συλλαμβάνουν καλύτερα **τ**ις ιδιότητες μιας ακουστικής κατηγορίας που απαιτούνται για την αναγνώριση. Η βελτίωση φθάνει σε ποσοστά 6.23%, 5.7% και 8.33% για τα Mel, MPEG-7 και το σύνολο PWP- Integration αντίστοιχα. Κατά συνέπεια η χρήση των τεχνικών χρονικής συγχώνευσης είναι χρήσιμη για την αναγνώριση γενικευμένου ακουστικού σήματος αντίθετα με τη ταξινόμηση μουσικών οργάνων όπου επιτυγχάνονται σχεδόν ίδια αποτελέσματα και χωρίς τη χρήση αυτών των τεχνικών (Joder et al., 2009).

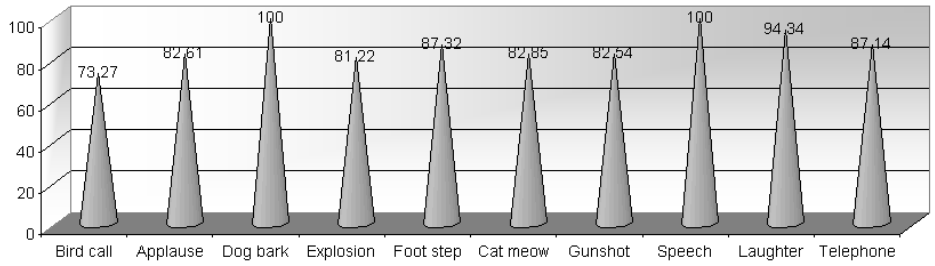
Μια ενδιαφέρουσα παρατήρηση είναι ότι όσο αυξάνουμε την τάξη των αυτοπαλινδρομικών διαδικασιών λαμβάνουμε χαμηλότερες ακρίβειες ταξινόμησης όσον αφορά στα σύνολα Mel και MPEG-7. Εντούτοις το σύνολο PWP-Integration παρείχε καλύτερες αποδόσεις και έτσι αποφασίσαμε να πραγματοποιήσουμε ένα πρόσθετο πείραμα χρησιμοποιώντας ένα μοντέλο DAR 4ης τάξης. Δυστυχώς αυτή η διαδικασία παρείχε χαμηλότερη ακρίβεια από την αντίστοιχη της 3ης τάξης. Αυτά τα γεγονότα οδηγούν στο συμπέρασμα ότι οι αυτοπαλινδρομικές συναρτήσεις δεν είναι ικανές να συλλάβουν τη χρονική συμπεριφορά της ακολουθίας χαρακτηριστικών ενός ακουστικού σήματος.

Κάποιος μπορεί να κάνει τη λογική υπόθεση ότι όσο μεγαλύτερη είναι η τιμή του παραθύρου συγχώνευσης, τόσο υψηλότερα θα είναι τα ποσοστά αναγνώρισης,

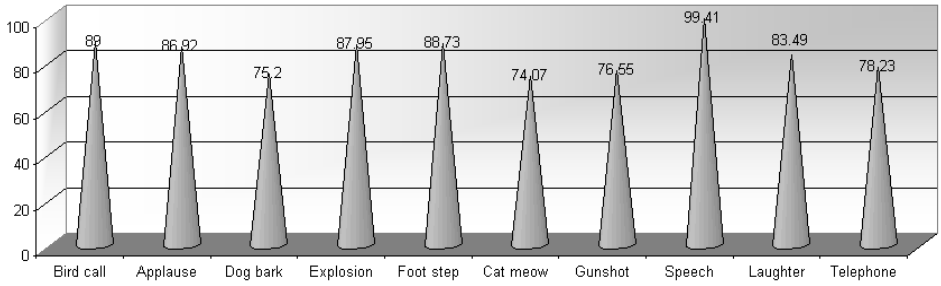
δεδομένου ότι μπορούν να χρησιμοποιηθούν περισσότερες πληροφορίες. Εντούτοις αυτή η υπόθεση δεν ισχύει σε όλες τις περιπτώσεις. Κατά τη διάρκεια αυτής της φάσης της πειραματικής ανάλυσης των αποτελεσμάτων απομονώσαμε το καλύτερο ποσοστό αναγνώρισης για κάθε texture window όσον αφορά κάθε ομάδα παραμέτρων σύμφωνα με τη μέθοδο βραχυπρόθεσμων στατιστικών.



(a) Mel-filterbank based feature set



(b) MPEG-7 Audio spectrum projection



(c) Perceptual wavelet packet integration analysis

Σχήμα 7.3: Ποσοστά αναγνώρισης για κάθε κατηγορία ήχων όσον αφορά τα σύνολα των επόμενων ακουστικών χαρακτηριστικών: a) Mel-filterbank, b) MPEG-7 ASP και c) PWP Integration χρησιμοποιώντας τις παραμέτρους (μήκος του texture window, αριθμός καταστάσεων και αριθμός Γκαουσσιανών συστατικών) που παρουσίασαν την μεγαλύτερη γενική ακρίβεια.

Το Σχήμα 7.2 επεξηγεί την παραλλαγή που εκθέτει το μέσο ποσοστό αναγνώρισης όταν αλλάζει το μήκος του texture window. Το σύνολο που είναι βασισμένο στην πολλαπλή ανάλυση παρουσιάζει μεγάλη βελτίωση καθώς το μήκος αυξάνεται μέχρι την τιμή των 90 πλαισίων και έπειτα η απόδοση πέφτει. Η αντίθετη

περίπτωση είναι αυτή του περιγραφέα ASP του MPEG-7 όπου η μέγιστη ικανότητα διάκρισής του εμφανίζεται όταν ενσωματώνονται πληροφορίες 10 πλαισίων. Η τράπεζα φίλτρων Mel αποτελεί την ενδιάμεση περίπτωση ενώ παρουσιάζει τη μέγιστη απόδοση όταν ενσωματώνονται 60 πλαίσια. Κατά τη διάρκεια των πειραματισμών μας ο αριθμός των texture windows που χρησιμοποιήθηκαν για την κατασκευή των HMMs κρατιόταν σταθερός λόγω της σταθερής τιμής του hop-size, ως εκ τούτου αποφύγαμε το overfitting των μοντέλων που οφείλεται σε ανεπαρκή ποσότητα δεδομένων.

Το Σχήμα 7.3 απεικονίζει λεπτομερέστερα τα αποτελέσματα ταξινόμησης όσον αφορά στην καλύτερη γενική απόδοση που πέτυχε κάθε σύνολο χαρακτηριστικών. Όπως μπορούμε να δούμε υπάρχουν μερικές κατηγορίες ήχων που αναγνωρίζονται με υψηλή ακρίβεια από όλες τις ομάδες περιγραφέων, όπως τα γεγονότα έκρηξης, ομιλίας και επιδοκιμασίας. Αφ' ετέρου διάφορα ηχητικά γεγονότα ταξινομούνται σωστά από ένα ή δύο σύνολα χαρακτηριστικών αλλά όχι από το τρίτο. Η τράπεζα φίλτρων Mel κατέδειξε την καλύτερη απόδοσή στην ταξινόμηση ομιλίας (98%) και στα κελαηδίσματα πουλιών (94.44%) ενώ δεν μπορεί να αναγνωρίσει ήχους γατών (75.77%) και βημάτων (78.34%) με ικανοποιητική ακρίβεια. Ο περιγραφέας ASP αναγνώρισε σωστά όλα (100%) τα ηχητικά δείγματα των κατηγοριών ομιλίας και ήχων σκυλιών αλλά επέδειξε το χαμηλότερο ποσοστό όσον αφορά στην κατηγορία με ήχους πουλιών. Το υψηλότερο αποτέλεσμα για την αναγνώριση ήχων βημάτων το λάβαμε από την ομάδα περιγραφέων που είναι βασισμένη στην ανάλυση με πακέτα wavelet. Αυτές οι παρατηρήσεις δείχνουν ότι τα σύνολα των ακουστικών παραμέτρων έχουν μερικά κοινά χαρακτηριστικά αλλά εκθέτουν πολλές διαφορές όταν πρόκειται να διακρίνουν κάποιες συγκεκριμένες ηχητικές κατηγορίες. Ως εκ τούτου αποφασίσαμε να ερευνήσουμε τη χρήση διάφορων τεχνικών μίξης που λειτουργούν πάνω στα αποτελέσματα των HMMs. Αυτή η πειραματική φάση περιγράφεται στην επόμενη παράγραφο.

Συμπερασματικά, διαπιστώσαμε ότι η χρονική συγχώνευση χαρακτηριστικών μπορεί να είναι ιδιαίτερου οφέλους όσον αφορά στη αναγνώριση γενικευμένου ακουστικού σήματος. Το σύνολο του ακουστικού πρωτοκόλλου MPEG-7 κατέδειξε την καλύτερη απόδοση ενώ καταλήξαμε στο συμπέρασμα ότι διαφορετικά σύνολα γνωρισμάτων μπορούν να ταξινομήσουν διαφορετικές ηχητικές κατηγορίες με διαφορετική ακρίβεια. Επιπλέον η αύξηση του αριθμού πλαισίων που πρόκειται να ενσωματωθούν μπορεί να παρέχει βελτιωμένα αποτελέσματα μόνο ως ένα ορισμένο βαθμό, οποίος εξαρτάται σε μεγάλο βαθμό από το συγκεκριμένο σύνολο χαρακτηριστικών. Μετά από ένα ορισμένο όριο οι ενσωματωμένες πληροφορίες δεν παρουσιάζουν ένα συνεπές σχέδιο, κάτι το οποίο εμποδίζει την διαδικασία κατασκευής προτύπων.

Μέθοδος μίξης	Ακρίβεια αναγνώρισης (%)
Majority voting simple	79.48
concatenation (no temporal integration)	85.12 (4 states, 32 modes)
J48 Tree	95.67
MLP	96.95

Πίνακας 7.3: Τα μέσα ποσοστά αναγνώρισης που επιτυγχάνουν οι τέσσερις μέθοδοι μίξης.

7.4.3. Μίξη των εξόδων των HMMs

Μετά από τα συμπεράσματα του προηγούμενου τμήματος, πειραματιστήκαμε πάνω στη ταυτόχρονη χρήση των αποτελεσμάτων των HMMs που χτίστηκαν με τη εξαγωγή διαφορετικών συνόλων χαρακτηριστικών τα οποία είναι συγχωνευμένα χρησιμοποιώντας το μήκος του texture window που παρείχε το υψηλότερο ποσοστό αναγνώρισης. Τα ίδια δεδομένα χρησιμοποιήθηκαν κατά τη διάρκεια και της εκπαιδευτικής καθώς και της εξεταστικής διαδικασίας. Η διαφορά είναι ότι αντί να επεξεργαστούμε τις ακολουθίες χαρακτηριστικών, αυτή τη φορά μοντελοποιήσαμε μόνο την πιθανότητα (log-likelihood) που παράγεται από κάθε HMM. Αξιολογήσαμε την απόδοση δύο σχεδίων μίξης: δέντρο J48 και πολυστρωματικό perceptron (Witten et al., 2005). Τα δέντρα απόφασης μπορούν να κατασκευαστούν εύκολα με έναν εποπτευόμενο τρόπο ενώ δεν γίνεται καμία a priori υπόθεση για τη κατανομή των δεδομένων. Το κύριο μειονέκτημά τους είναι ότι ακόμα και μικρές αλλαγές στο σύνολο κατάρτισης μπορούν να οδηγήσουν σε μια δομή δέντρου απόφασης που παρουσιάζει έναν πολύ μεγάλο αριθμό διαφορών. Εντούτοις μειώνουμε αυτόν τον κίνδυνο δεδομένου ότι μοντελοποιούμε τις πιθανότητες των ακολουθιών των χαρακτηριστικών και όχι τα χαρακτηριστικά τα ίδια. Η πολυστρωματική perceptron (MLP) μέθοδος ακολουθεί τη λογική του γραμμικού perceptron ενώ χρησιμοποιεί τους κόμβους με μη γραμμικές συναρτήσεις ενεργοποίησης για το διαχωρισμό δεδομένων που δεν είναι γραμμικά διαχωρίσιμα. Επιπλέον, τα τεχνητά νευρωνικά δίκτυα μπορούν να είναι πολύ χρήσιμα σε περιπτώσεις που τα σχέδια/μοτίβα δεν είναι εμφανή. Ο backpropagation αλγόριθμος χρησιμοποιήθηκε για να εκπαιδεύσει το νευρωνικό δίκτυο με ένα κρυμμένο στρώμα είκοσι κόμβων (το μισό σύνολο του αριθμού των γνωρισμάτων συν τον αριθμό των κατηγοριών) με ποσοστό εκμάθησης ίσο με 0.3.

Responded / Presented	Bird Call	Applause	Dog Barking	Explosion	Footstep	Cat Meowing	Gunshot	Speech	Laughter	Telephone
BirdCall	**95.3**	0	0	0	0	2.7	2	0	0	0
Applause	0	**100**	0	0	0	0	0	0	0	0
Dog Barking	0	0	**96.2**	0	0	0	0	0	4.8	0
Explosion	0	0	0	**94.89**	0	2.11	3	0	0	0
Footstep	0	0	0	0	**97.52**	0	2.28	0	0	0
Cat Meowing	0	1.1	1.5	0	0	**95.4**	0	1.3	0	0.7
Gunshot	0	0	0	2.1	0	0.1	**97.8**	0	0	0
Speech	0	0	0	0	0	0	0	**100**	0	0
Laughter	0	1.2	0	0	0	0.6	1.3	0	**96.9**	0
Telephone	0	0	0	0	0	3.2	1.3	0	0	**95.5**

Πίνακας 7.4: Μήτρα σύγχυσης του τελικού συστήματος με μίξη των πιθανοτήτων βασισμένη στο σχήμα του MLP.

Αυτές οι μέθοδοι επιλέχτηκαν λόγω της δυνατότητάς τους να χειριστούν τα περιττά (redundant) στοιχεία, κάτι που σημαίνει ότι στην περίπτωση που τα σύνολα χαρακτηριστικών συλλαμβάνουν επικαλυπτόμενες πληροφορίες σχετικά με το ακουστικό σήμα που αντιπροσωπεύουν, ο αλγόριθμος μπορεί αποτελεσματικά να τις εκμεταλλευτεί και η απόδοση του συστήματος αναγνώρισης συνήθως αυξάνεται (συγκρινόμενη με την υιοθέτηση κάθε ομάδας χαρακτηριστικών ξεχωριστά). Ένα περιττό σύνολο χαρακτηριστικών γνωρισμάτων μπορεί να παρέχει βελτιωμένη απόδοση υπό δυσμενείς ακουστικούς όρους (όπου κάποια από τα μέρη του φάσματος του σήματος είναι απόντα ή παραμορφωμένα) όπως στην περίπτωση της πραγματικής ζωής. Δύο απλούστερες προσεγγίσεις αξιολογήθηκαν επίσης: πλειοψηφική ψηφοφορία (majority voting) και απλός συνδυασμός όλων των παραμέτρων πριν από το στάδιο της χρονικής συγχώνευσης (simple concatenation). Όσον αφορά την περίπτωση της ψηφοφορίας με πλειοψηφία, εάν δεν είχαμε έστω και μία συμφωνία μεταξύ των κατηγοριοποιητών, δηλ. εάν οδηγούμασταν σε τρεις διαφορετικές αποφάσεις για την ίδια ακολουθία δοκιμής, η απόφαση παιρνόταν με έναν τυχαίο τρόπο.

Στον πίνακα 7.3 φαίνονται τα μέσα ποσοστά αναγνώρισης που επιτεύχθηκαν από τη μίξη των εξόδων των HMM με τέσσερις διαφορετικές μεθόδους. Η υψηλότερη απόδοση επιτυγχάνεται από την μίξη που είναι βασισμένα στον MLP (96.95%) ενώ η μέθοδος δέντρων απόφασης J48 κατέδειξε το δεύτερο καλύτερο αποτέλεσμα (95.67%). Τα HMMs που κατασκευάστηκαν με τη σύνδεση όλων των χαρακτηριστικών σε ένα διάνυσμα προσέφεραν μια σημαντική βελτίωση της τάξης του 3.06% συγκρινόμενα με το σύνολο του MPEG-7 (82.06%). Η μέθοδος της πλειοψηφικής ψηφοφορίας παρουσίασε

τη χειρότερη απόδοση (79.48%) δεδομένου ότι υφίσταται τις επιπτώσεις του γεγονότος ότι οι πρώτης φάσης ταξινομητές HMM τείνουν να διαφωνούν.

Μια άμεση σύγκριση του προτεινόμενου πλαισίου με άλλα συστήματα ταξινόμησης ήχων δεν είναι εφικτή λόγω των διαφορετικών δεδομένων που υιοθετούνται σε κάθε μια από αυτές τις εργασίες. Στην εργασία (Casey, 2001) χρησιμοποιείται το χαρακτηριστικό ASP του MPEG-7 χωρίς χρονική συγχώνευση. Το

συγκεκριμένο γνώρισμα φθάνει σε ποσοστό 82.06% στο σύνολο των δεδομένων μας, το οποίο είναι αρκετά χαμηλότερο από το ποσοστό αναγνώρισης που επιτυγχάνεται από το σχήμα που είναι βασισμένο στη μεθοδολογία MLP. Όσον αφορά τη μέθοδο MLP, παρατηρούμε ότι διόρθωσε πολλά από τα λάθη ταξινόμησης που παρήχθησαν από τα HMMs (η μήτρα σύγχυσης φαίνεται στον πίνακα 7.4).

Η υψηλή απόδοση του συστήματος οφείλεται στην ταυτόχρονη χρήση των διαφορετικών ομάδων ακουστικών παραμέτρων. Αυτό δείχνει (και προτρέπει) ότι το πρόβλημα της αναγνώρισης γενικευμένου ακουστικού σήματος αντιμετωπίζεται καλύτερα με χρησιμοποίηση μιας ομάδας περιγραφέων προερχόμενοι από διαφορετικά πεδία. Καταλήγουμε στο συμπέρασμα ότι τα τελικά αποτελέσματα ταξινόμησης είναι πολύ ικανοποιητικά δεδομένου ότι η βάση δεδομένων μας χαρακτηρίζεται από μεγάλη μεταβλητότητα ακόμα και σε δείγματα της ίδιας κατηγορίας (within class variability).

Συμπεράσματα και Μελλοντικές

Κατευθύνσεις

Αναφέρθηκαν διάφορες τεχνικές καθώς επίσης και αναδυόμενες μεθοδολογίες αυτόματης αναγνώρισης ήχων. Επίσης έγινε ανασκόπηση μερικών πρόσφατων τεχνικών στην περιοχή της επεξεργασίας σήματος και αναγνώρισης προτύπων που συσχετίζονται με τις εν λόγω εφαρμογές. Αυτό δείχνει ότι διαφορετικά υποπεδία της ταξινόμησης ήχων έχουν εξελιχθεί ανεξάρτητα και στη πραγματικότητα μοιράζονται τις ίδιες αρχές αναγνώρισης προτύπων και καταδεικνύουν ιδιαίτερα επικαλυπτόμενες μεθόδους εξαγωγής ακουστικών παραμέτρων.

Αν και, γενικά δεν είναι δυνατό να προσδιορίσουμε ένα σταθερό σύνολο χαρακτηριστικών ή αλγορίθμων ταξινόμησης που αποδίδουν βέλτιστα για όλες τις εφαρμογές καθώς τα χαρακτηριστικά που είναι κατάλληλα κάθε φορά εξαρτώνται από τις προδιαγραφές της εκάστοτε εφαρμογής, εμπειρικά προτείνουμε τους cepstral συντελεστές Mel-συχνότητας (MFCC), που έχουν αποδείξει τη χρησιμότητά τους και στους οποίους μπορούν να προστεθούν οι περιγραφείς του πρωτοκόλλου MPEG-7 για να ενισχύσουν την απόδοση αναγνώρισης. Επιπλέον, αν και η επιλογή της προσέγγισης ταξινόμησης εξαρτάται από τις λεπτομέρειες της εφαρμογής, ένας ταξινομητής βασισμένος σε HMM είναι η χαρακτηριστική επιλογή για αναγνώριση των ηχητικών γεγονότων που μεταβάλλονται στο χρόνο. Οι ταξινομητές που βασίζονται σε GMM (HMM μίας κατάστασης) παρέχουν μία καλή περιγραφή της κατανομής δεδομένων και αποτελούν μια λογική επιλογή, δεδομένου ότι αποδίδουν ικανοποιητικά σε πολλές εφαρμογές ταξινόμησης ακουστικών σημάτων. Άλλες προσεγγίσεις

ταξινόμησης, όπως SVM που δεν πάσχουν από το πρόβλημα του μεγέθους του διανύσματος των χαρακτηριστικών (συχνά αναφέρεται ως curse of dimensionality), παρέχουν μία εναλλακτική λύση καθώς δεν προσπαθούν να διαμορφώσουν τη εσωτερική κατανομή των δεδομένων εκπαίδευσης αλλά απλά αναζητούν το βέλτιστο χωρισμό μεταξύ των κατηγοριών. Τέλος, τα συμπληρωματικά προτερήματα των παραγωγικών (π.χ. GMM) και διαχωριστικών (π.χ. SVM) ταξινομητών υιοθετούνται συχνά με στόχο να μεγιστοποιηθεί η γενική απόδοση αναγνώρισης. Μία πολύ καλή επισκόπηση του πεδίου της αναγνώριση ήχων μπορεί να βρεθεί στην εργασία (Potamitis et al., 2008).

Η παρούσα συνεισφορά εστίασε στο στόχο του αυτόματου προσδιορισμού αστικών ηχητικών σκηνών. Εξερευνήθηκε η χρήση δύο συνόλων χαρακτηριστικών ενώ μια λεπτομερής σύγκριση έλαβε μέρος. Καταδείξαμε έναν αλγόριθμο μετα-επεξεργασίας συμπεριλαμβανομένου της PCA και η χρήση του αποδείχθηκε ότι προσφέρει βελτιωμένα αποτελέσματα. Παρουσιάσαμε μια προσέγγιση αναγνώρισης προτύπων βασισμένη στο state of the art ενώ τα τελικά ποσοστά αναγνώρισης ήταν περισσότερο από ικανοποιητικά. Η εργασία περιλαμβάνει τον διαχωρισμό ακουστικών σημάτων, την ενσωμάτωση περισσότερων ηχητικών κατηγοριών καθώς επίσης και την εξερεύνηση του συνδυασμού των δύο συνόλων χαρακτηριστικών προκειμένου να χρησιμοποιηθούν αποτελεσματικά οι πιο διακριτικές πληροφορίες που προσφέρουν.

Ακόμη παρουσιάσαμε και αξιολογήσαμε ένα δύο επιπέδων πιθανοτικό πλαίσιο για τον ακουστικό έλεγχο σε ένα περιβάλλον σταθμών μετρό. Ο κύριος στόχος του είναι να προσδιορίσει εγκαίρως καταστάσεις κινδύνου και να παραδώσει τα απαραίτητα μηνύματα ειδοποίησης σε έναν εξουσιοδοτημένο υπάλληλο. Η προτεινόμενη μεθοδολογία είναι πρακτική, μπορεί να λειτουργήσει σε πραγματικό χρόνο και μοντελοποιεί τρία ανώμαλα ηχητικά γεγονότα που αλλοιώνονται από τον ιδιαίτερα μη στατικό θόρυβο ενός σταθμού μετρό. Συμπεραίνουμε ότι τα αποτελέσματα αναγνώρισης κάτω από τον συγκεκριμένο είδος περιβαλλοντικού θορύβου είναι πολύ ενθαρρυντικά.

Αυτήν την περίοδο η επιτήρηση δημόσιων χώρων συνήθως πραγματοποιείται με κάμερες, ενώ ένας ανθρώπινος χειριστής πρέπει να προσέχει τις εξόδους τους σε 24ωρη βάση. Επιπλέον, η γνώση ότι οι δημόσιοι χώροι ασφαλίζονται με ευφυή έλεγχο αναμένεται να αποθαρρύνει την εκδήλωση επικίνδυνων πράξεων, όπως ληστείες κ.λπ. Πολύ έρευνα πραγματοποιείται από την κοινότητα επεξεργασίας σήματος προς τον έλεγχο χώρων χωρίς επίβλεψη. Πιστεύουμε ότι οι ακουστικές πληροφορίες μπορούν να είναι μεγάλου οφέλους προς την εκπλήρωση αυτού του στόχου.

Προτάθηκε ένα πλαίσιο που εκμεταλλεύεται τα πλεονεκτήματα των ακουστικών περιγραφών από διαφορετικά πεδία. Εφαρμόστηκαν τρεις πιθανολογικές μέθοδοι ανίχνευσης καινοτομίας και εξετάστηκαν σε ένα σύνολο δεδομένων που είχε ως σκοπό να είναι όσο το δυνατόν πιο κοντά στους πραγματικούς όρους και

περιλαμβάνει τρία διαφορετικά σενάρια. Λάβαμε υψηλή επίδοση όσον αφορά όλα τα σενάρια. Τα αποτελέσματα δείχνουν ότι η ακουστική μορφή μπορεί να λειτουργήσει και με αυτόνομο τρόπο όσον αφορά στον εντοπισμό μη τυπικών

Προτείναμε ένα ολοκληρωμένο σύστημα για την ακουστική επιτήρηση και ανίχνευση μη τυπικών καταστάσεων. Ερευνήσαμε έναν μεγάλο αριθμό ακουστικών χαρακτηριστικών προκειμένου να καταλήξουμε στην καλύτερη αντιπροσώπευση μιας μη τυπικής κατάστασης που περιλαμβάνει τις ακουστικές εκφράσεις του πόνου, της πίεσης, των πυροβολισμών και των εκρήξεων. Κατασκευάσαμε ένα ιεραρχικό σύστημα που είναι βασισμένο σε πιθανοτικά πρότυπα που εκπαιδεύθηκαν χρησιμοποιώντας μεγάλη ποσότητα ήχων υψηλής ποιότητας οι οποίοι πάρθηκαν από επαγγελματικές ηχητικές συλλογές. Το σύστημα αξιολογήθηκε υπό δυσμενείς όρους που περιέχουν ιδιαίτερα μη στατικό παρασιτικό θόρυβο στο πλαίσιο τριών διαφορετικών ειδών περιβαλλόντων. Το σύστημα μπορεί να εγκατασταθεί και να προσαρμοστεί σε εσωτερικό ή εξωτερικό χώρο χρησιμοποιώντας online προσαρμογή μοντέλων. Θεωρούμε ότι μπορεί να αποτελέσει ένα αναπόσπαστο μέρος ενός πρακτικού ακουστικού συστήματος παρακολούθησης.

Συμπεραίνουμε ότι αυτά τα αποτελέσματα είναι πολλά υποσχόμενα, λαμβάνοντας υπόψη ότι η μεταβλητότητα στο εσωτερικό κάθε κατηγορίας είναι πολύ υψηλή (π.χ. το κουδούνι πορτών και η κανονική ομιλία είναι μέρη της ίδιας κατηγορίας ενώ το ίδιο πράγμα ισχύει για την μη τυπικά ομιλία και το σπάσιμο υλικών). Επιπλέον πολλά απρόβλεπτα γεγονότα θορύβου αντιμετωπίστηκαν που δυσχεραίνουν τη λειτουργία του συστήματος αναγνώρισης. Συμπερασματικά, τα αποτελέσματα καταδεικνύουν την αποτελεσματικότητα της προσέγγισης ανίχνευσης καινοτομίας όσον αφορά στην περίπτωση της ακουστικής επόπτευσης για μη τυπικές καταστάσεις.

Τέλος προτείναμε ένα ολοκληρωμένο σύστημα για την αναγνώριση γενικευμένου ακουστικού σήματος που οδηγεί σε υψηλή ακρίβεια. Υιοθετήθηκε ένας συνδυασμός διάφορων καλά τεκμηριωμένων πηγών υψηλής ποιότητας για την οργάνωση ενός λεπτομερούς συνόλου δεδομένων. Επίσης εκτέθηκαν τα πλεονεκτήματα της χρονικής συγχώνευσης χαρακτηριστικών όσον αφορά ακουστικά γνωρίσματα διαφορετικών πεδίων. Το πειραματικό πρωτόκολλο σχεδιάστηκε προσεκτικά και όλες οι πτυχές της προτεινόμενης μεθοδολογίας αξιολογήθηκαν λεπτομερώς. Τα αποτελέσματα αποκαλύπτουν ότι διαφορετικά texture windows είναι κατάλληλα για κάθε γνώρισμα ενώ η στρατηγική της χρονικής συγχώνευσης που βασίζεται στον υπολογισμό βραχυπρόθεσμων στατιστικών κατέδειξε τα καλύτερα μέσα ποσοστά αναγνώρισης. Οι υπόλοιπες τεχνικές, αν και πιο σύνθετες και υπολογιστικά απαιτητικές δεν παρήγαγαν μια αντιπροσώπευση της δομής των ακουστικών σημάτων που να μπορεί να μοντελοποιηθεί και στη συνέχεια να προσδιοριστεί αποτελεσματικά. Η πρώτη φάση του συστήματος χρησιμοποιεί από τα αριστερά προς τα δεξιά HMMs για τον υπολογισμό της κατανομής των χαρακτηριστικών που ανήκουν σε κάθε κατηγορία ήχων. Μετά από τα αποτελέσματα των εκτενών πειραματισμών, ένα περαιτέρω βήμα ερευνήθηκε: η

ταυτόχρονη χρήση των ηχητικών παραμέτρων οι οποίοι είναι βασισμένοι στο φασματικό και στο πεδίο κυματιδίων. Το σχήμα μίξης MLP που διαμορφώθηκε χρησιμοποιώντας τις πιθανότητες που παρήχθησαν από τα προηγουμένως κατασκευασμένα HMMs παρείχε τα υψηλότερα ποσοστά ταξινόμησης όσον αφορά όλες τις ηχητικές κατηγορίες που εξετάστηκαν στη μελέτη μας. Αυτό δείχνει ότι η αυτοματοποιημένη αναγνώριση γενικευμένου ακουστικού σήματος αντιμετωπίζεται καλύτερα χρησιμοποιώντας ομάδες περιγραφέων από διαφορετικά πεδία. Οι κατηγορίες ήχων που τώρα δεν συμπεριλαμβάνονται στην εργασία μας μπορούν εύκολα να ενσωματωθούν εφόσον οργανώσουμε μία ικανοποιητική ποσότητα δεδομένων εκπαίδευσης. Η ίδια μεθοδολογία μπορεί να χρησιμοποιηθεί για την επεξεργασία των απαρατήρητων ακουστικών ακολουθιών (εξαγωγή χαρακτηριστικών συνδυασμένη με τεχνικές χρονικής συγχώνευσης) και στη συνέχεια μπορεί να κατασκευαστεί ένα πιθανοτικό πρότυπο για κάθε νέα κατηγορία ήχων. Η προτεινόμενη υλοποίηση είναι εύκολα προσαρμόσιμη και μπορεί να διευκολύνει πολλές εφαρμογές αναγνώρισης ήχων.

Ο στόχος αυτής της εργασίας ήταν η αξιολόγηση διάφορων τεχνικών συγχώνευσης με στόχο την αυτόματη ταξινόμηση ακουστικού σήματος. Ο στόχος τώρα είναι να χρησιμοποιηθούν τα αποτελέσματα που αναφέρονται σε αυτήν την εργασία για να δημιουργηθούν αυτόνομα συστήματα ικανά να διαμορφώσουν μια ακριβή περιγραφή του περιβάλλοντος χώρου βασισμένα απλώς στη «ακουστική αίσθησή τους». Τέτοια συστήματα θα μπορούσαν να διευκολύνουν τη καθημερινή μας ζωή με την παροχή λύσεων σε διάφορες πραγματικές εφαρμογές.

Το πεδίο της επεξεργασίας ακουστικών σημάτων που δε περιλαμβάνουν ομιλία (non speech audio signal processing) μπορεί να προσφέρει λύσεις σε διάφορα προβλήματα. Η συγκεκριμένη πτυχιακη πρότεινε μία σειρά από καινοτόμες προσεγγίσεις σε διαφορετικές εφαρμογές της τεχνολογίας αναγνώρισης γενικευμένου ακουστικού σήματος (αναγνώριση ακουστικών γεγονότων αστικού περιβάλλοντος, διάκριση ομιλίας/μουσικής και ακουστική επόπτευση μη-τυπικών καταστάσεων). Επίσης εισήγαγε στον συγκεκριμένο ερευνητικό πεδίο την ιδέα της συγχώνευσης ακουστικών χαρακτηριστικών στο πεδίου του χρόνου, διαδικασία η οποία οδηγεί σε σημαντικά υψηλότερη ακρίβεια αναγνώρισης σε σχέση με το πρωτόκολλο MPEG-7 Audio. Αυτή η μεθοδολογία είναι γενικής φύσης και η εφαρμογή της σε διαφορετικού τύπου προβλήματα (π.χ. ταξινόμηση βιοακουστικών σημάτων) αποτελεί αντικείμενο της μελλοντικής εργασίας μας.

Επίσης η δημιουργία συστήματος ακουστικής ανάλυσης που ενσωματώνει ακόμα περισσότερες ηχητικές πηγές, αποτελεί ένα πρόβλημα αυξημένης δυσκολίας που μπορεί πιθανά να αντιμετωπισθεί με ταξινομητές ιεραρχικής δομής. Επιπρόσθετα το πρόβλημα της αυτόματης καταμέτρησης και διαχωρισμού ηχητικών πηγών είναι μεγάλου ενδιαφέροντος.

Μία πολλά υποσχόμενη ερευνητική περιοχή είναι αυτή της μίξης/συνδυασμού του ακουστικού αισθητήρα με άλλους ετερογενείς αισθητήρες (π.χ. οπτικές κάμερες και IR). Αυτή η διαδικασία ίσως προσφέρει υψηλότερη ακρίβεια και βοηθήσει προς την κατασκευή ενός πλαισίου που διευκολύνει την ανίχνευση και την ερμηνεία διαφόρων τύπων ανθρώπινης συμπεριφοράς.

Συνεισφορά της πτυχιακης

Η παρούσα πτυχιακή παρουσιάζει και προτείνει λύσεις σε μία σειρά από νέες αλλά και υπάρχουσες εφαρμογές του πεδίου της αυτόματης αναγνώρισης γενικευμένων ακουστικών γεγονότων. Ο βασικός στόχος των αλγορίθμων που παρουσιάζονται είναι η ανάλυση του περιβάλλοντα χώρου χρησιμοποιώντας μόνο ακουστικά σήματα. Οι παρούσες μέθοδοι έχουν μεγάλη πρακτική αξία ενώ παράλληλα έχουν σχετικά μικρές απαιτήσεις σε οικονομικό καθώς και σε υπολογιστικό κόστος, χαρακτηριστικό που τις κάνει ιδιαίτερα ελκυστικές σε πολλών ειδών εφαρμογές. Πιο συγκεκριμένα η πτυχιακή εμπεριέχει:

- Στο *Κεφάλαιο 1* παρουσιάζεται μία γενική επισκόπηση της αυτόματης αναγνώρισης γενικευμένων ακουστικών γεγονότων. Επιπλέον συζητάμε τις εφαρμογές της τεχνολογίας αναγνώρισης ακουστικού σήματος και δίνουμε μία σύντομη περιγραφή του ιδανικού συστήματος (state of the art).
- Στο *Κεφάλαιο 2* εισάγουμε τον αναγνώστη στο χώρο της επεξεργασίας ακουστικών σημάτων που δε περιλαμβάνουν ομιλία. Παρουσιάζονται οι σύγχρονες προσεγγίσεις όσον αφορά στις μεθοδολογίες εξαγωγής χαρακτηριστικών και αναγνώρισης προτύπων.
- Κεφάλαιο 3: Σύστημα αναγνώρισης ηχητικών γεγονότων σε αστικό περιβάλλον με βασική εφαρμογή την παρακολούθηση της κίνησης οχημάτων. Επίσης αναλύεται ο σχεδιασμός βάσης δεδομένων που περιλαμβάνει μεγάλη ποικιλία των αντίστοιχων ηχητικών σημάτων. Συγκρίνουμε τις παραμέτρους του πρωτοκόλλου MPEG-7 με τις Mel Frequency Cepstral Coefficients (MFCC) όσον αφορά την ικανότητα ταξινόμησης τους σε μία νέα εφαρμογή της CASA. Η αναγνώριση ακουστικών προτύπων γίνεται με ιεραρχική χρήση στατιστικών μεθόδων και προτείνεται μία τεχνική μέτα-επεξεργασίας (post-processing) των ακουστικών παραμέτρων που οδηγεί σε μεγαλύτερη απόδοση .
- Κεφάλαια 4 και 5: Ολοκληρωμένο σύστημα εντοπισμού μη τυπικών καταστάσεων που διαδραματίζονται στο περιβάλλοντα χώρο. Το σύστημα βασίζεται στο σήμα ενός απλού μικρόφωνου ενώ έχει τη δυνατότητα να αυτοπροσαρμόζεται ώστε να λειτουργεί αποδοτικά σε ποικίλα ηχητικά περιβάλλοντα. Προτείνεται ο συνδυασμός ακουστικών χαρακτηριστικών τα

οποία μπορούν να εκφράσουν αποτελεσματικά τις ιδιότητες των σημάτων προς επεξεργασία. Αυτές είναι οι MFCC, Audio Spectrum Flatness, Audio Waveform Envelope, Teager energy operator, θεμελιώδης συχνότητα και αρμονικότητα. Ακόμα, προτείνεται μία ιεραρχική δομή αναγνώρισης με ξεχωριστά χαρακτηριστικά ανά επίπεδο, τα οποία είναι άμεσα συσχετιζόμενα με το εκάστοτε πρόβλημα αναγνώρισης. Η αυτοπροσαρμογή του συστήματος γίνεται με τη τεχνική της μεγαλύτερης εκ των υστέρων πιθανότητας (Maximum A Posteriori Probability). Γίνεται προσομοίωση πραγματικών συνθηκών με χρήση δύο υπολογιστών. Διάφορες περιβαλλοντικές συνθήκες δημιουργούνται (όπως αστικό περιβάλλον και σταθμός μετρό) και το σύστημα παρουσιάζει πολύ καλά αποτελέσματα .

- Κεφάλαιο 6: Εισαγωγή της τεχνικής "Ανίχνευση Καινοτομίας" (Novelty Detection) στο πρόβλημα της CASA με εφαρμογή σε δεδομένα πραγματικού κόσμου για τον εντοπισμό μη τυπικών καταστάσεων. Αυτή η τεχνική υιοθετήθηκε καθώς μας επιτρέπει να αποφανθούμε αν τα άγνωστα δεδομένα ανήκουν σε μία *a-priori* γνωστή κατανομή ή όχι. Με αυτό το τρόπο κατά τη διάρκεια σχεδιασμού του συστήματος απαλλαγόμαστε από την ανάγκη δημιουργίας στατιστικού μοντέλου για τουλάχιστον μία κατηγορία. Προτείνεται η ανίχνευση καινοτομίας βασισμένη σε ομαδοποίηση Γκαουσσιανών μοντέλων και εφαρμόζεται σε ενα σενάριο με διαφορετικές απαιτήσεις και διαφορετικά ακουστικά περιβάλλοντα. Η αξιολόγηση του συστήματος έδειξε καλύτερα αποτελέσματα από την προσέγγιση που βασίζεται στην κατηγοριοποίηση, που είναι και η πιο διαδομένη μέχρι τώρα .
- Κεφάλαιο 7: Καινοτόμο αλγόριθμο για την αυτόματη αναγνώριση γενικευμένου ακουστικού σήματος ο οποίος παρουσιάζει υψηλότερη απόδοση από το πρωτόκολλο MPEG-7 Audio. Επίσης προτείνεται ένας συνδυασμός βάσεων δεδομένων, που έχει τη δυνατότητα να αποτελέσει τη βάση αναφοράς όσον αφορά το συγκεκριμένο ερευνητικό πεδίο. Η καρδιά του αλγορίθμου είναι η ταυτόχρονη χρήση ακουστικών παραμέτρων από δύο πεδία (συχνότητας και κυματιδίου) σε συνδυασμό με τρεις τεχνικές για την αποτελεσματικότερη χρονική τους μοντελοποίηση.

Παράρτημα Α

Μέθοδοι αξιολόγησης και μετρικά απόδοσης για συστήματα αναγνώρισης

Η βασική αρχή για την αξιολόγηση ενός συστήματος είναι να συγκριθούν τα αποτελέσματά του με τη γνωστή ground truth, η οποία έχει δημιουργηθεί κατά τη διάρκεια του σχολιασμού των βάσεων των ηχητικών δεδομένων. Η ground truth είναι ένα είδος βέλτιστου αποτελέσματος, ο στόχος κάθε συστήματος είναι να φτάνει σε αποτελέσματα όσο το δυνατόν πιο κοντά σε αυτή. Το πόσο καλή είναι η απόδοση του συστήματος μετριέται με συγκεκριμένα μετρικά που συσχετίζουν την πραγματική έξοδο του συστήματος με τη βέλτιστη λύση. Υπάρχουν πολλά διαφορετικά μετρικά για αυτήν την σύγκριση. Όλα τους καλύπτουν μερικές συγκεκριμένες πτυχές της αξιολόγησης. Λόγω αυτού είναι σημαντικό να χρησιμοποιούνται προτυποποιημένα για:

• να επιτύχουμε συγκρισιμότητα μεταξύ των αποτελεσμάτων που δίνουν διαφορετικές υλοποιήσεις για ένα δεδομένο πρόβλημα

• να είμαστε βέβαιοι ότι καλύψαμε τις πολύ σημαντικές πτυχές του προβλήματος

• να καθιερωθεί μια ορισμένη αντικειμενικότητα.

Γενικά μετρικά για προβλήματα κατηγοριοποίησης

Σε αυτό το τμήμα περιγράφουμε τις γενικές στρατηγικές και μετρικά που χρησιμεύουν στο να αξιολογήσουμε γενικά τα συστήματα ανίχνευσης και ταξινόμησης. Θα αναλυθούν κοινά μετρικά όπως True/False negative/positive, ROC, μήτρα σύγχυσης, ακρίβεια, precision, recall, που δεν είναι εξαρτώμενα από κάποιο ερευνητικό πεδίο.

	Positive	Negative
detected	True Positive (TP)	False Positive (FP)
not detected	False Negative (FN)	True Negative (TN)

Πίνακας Α.1: Οι πιθανές περιπτώσεις στα προβλήματα δύο καταστάσεων.

Σε αυτό το τμήμα παρουσιάζουμε καθορισμένα μετρικά για προβλήματα ταξινόμησης δύο αλλά και περισσότερων κατηγοριών. Πολλά από αυτά χρησιμοποιούνται στην ίδια μορφή σε απολύτως διαφορετικές περιοχές της εφαρμοσμένης μηχανικής και είναι έτσι ένα είδος γενικού μέτρου για ποιοτικές εκτιμήσεις.

Α.1. Προβλήματα δύο καταστάσεων (εντοπισμός)

Τα προβλήματα δύο καταστάσεων συχνά καλούνται προβλήματα ανίχνευσης/εντοπισμού. Οι ταξινομητές αποφασίζουν εάν μια ορισμένη περίπτωση εμφανίζεται ή όχι. Χαρακτηριστικά οι δύο κατηγορίες χαρτογραφούνται στους όρους «ΘΕΤΙΚΗ» και «ΑΡΝΗΤΙΚΗ». Για την αξιολόγηση των προβλημάτων δύο κατηγοριών υπάρχουν τέσσερις διαφορετικές πιθανές περιπτώσεις για κάθε απλή ταξινόμηση. Η ground truth και άρα η σωστή κατηγορία μπορεί να είναι «ΘΕΤΙΚΗ» ή «ΑΡΝΗΤΙΚΗ». Τώρα το σύστημα ταξινόμησης μπορεί να βρει τη σωστή ή «true» λύση ή μπορεί να κάνει λάθος και να βρει έτσι μια «false» λύση.

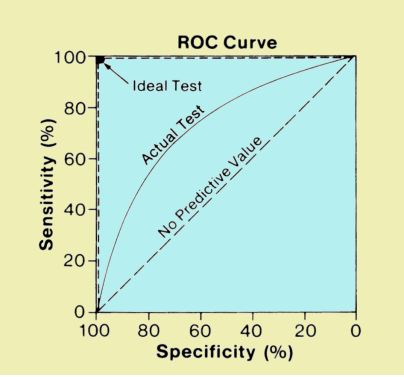
Με βάση τη συχνότητα των περιστατικών αυτών των τεσσάρων περιπτώσεων τα δημιουργούνται τα ακόλουθα τυπικά μετρικά για να περιγραφεί η απόδοση ενός ανιχνευτή:

- False-Positive Rate: $R_{F\bar{P}} \frac{FP}{FP+TN}$

ποσοστό των περιπτώσεων δοκιμής οι οποίες είναι «αρνητικές στη πραγματικότητα και κατηγοριοποιούνται λανθασμένα ως «θετικές» από τον αναγνωριστή.

- False-Negative-Rate $R_{FN} = \frac{FN}{FN+TP}$

ποσοστό των περιπτώσεων δοκιμής που είναι «θετικές» στη πραγματικότητα και κατηγοριοποιούνται λανθασμένα ως «αρνητικές» από τον αναγνωριστή.



Σχήμα Α.1: Η καμπύλη ROC. Όσο περισσότερο η καμπύλη βρίσκεται στη πάνω αριστερά προς τα δεξιά. Αν το κατώφλι είναι μηδέν, όλες οι περιπτώσεις δοκιμής κατηγοριοποιούνται ως ΑΛΗΘΕΙΣ και άρα η ευαισθησία (sensitivity) είναι ένα αλλά η ιδιομορφία (specificity) είναι μηδέν (κάτω αριστερά γωνία), αν το κατώφλι είναι μέγιστο, όλες οι περιπτώσεις δοκιμής κατηγοριοποιούνται σαν ΨΕΥΔΕΙΣ, και άρα η καμπύλη έχει το τελικό της σημείο στη πάνω δεξιά γωνία. Η βέλτιστη

- Specificity (True-Negative Rate)

$$R_{TN} = \frac{TN}{FP+TN}$$

ποσοστό των σωστά ταξινομημένων περιπτώσεων δοκιμής οι οποίες είναι «αρνητικές» στη πραγματικότητα.

- Sensitivity (True-Positive Rate)

$$R_{TP} = \frac{TP}{FN+TP}$$

ποσοστό των σωστά ταξινομημένων περιπτώσεων δοκιμής οι οποίες είναι «θετικές» στη πραγματικότητα.

- Detection Rate (Correct Classification Rate)

$$R_{T} = \frac{TP+FN}{TN+FN+TP+FP}$$

είναι το ποσοστό όλων των σωστά ταξινομημένων περιπτώσεων δοκιμής.

- False Detection Rate (Misclassification Rate)

$$R_{F} = \frac{FP+FN}{TN+FN+TP+FP}$$

είναι το ποσοστό όλων των λανθασμένα ταξινομημένων περιπτώσεων δοκιμής.

Μια χαρακτηριστική αντιπροσώπευση της απόδοσης ενός ανιχνευτή είναι η χρήση μιας καμπύλης αποκαλούμενης Receiver Operator Characteristics (ROC). Σε αυτή, το ποσοστό True-Positive σχεδιάζεται σε σχέση με το ποσοστό False-Positive σε εξάρτηση με μια παράμετρο, συνήθως ενός κατωφλίου (threshold). Αυτό γίνεται, εάν ένα προσαρμόσιμο κατώφλι T του συστήματος ανίχνευσης είναι αρμόδιο για την απόφαση «ανιχνεύθηκε» ή «δεν ανιχνεύθηκε» ενώ η βέλτιστη τιμή για αυτό το κατώφλι, ένα μέγιστο πηλίκο πρέπει να βρεθεί.

Οι καμπύλες Detection Error Tradeoff μπορούν να ειδωθούν σαν μία παρουσίαση της ανταλλαγής (tradeoff) μεταξύ δύο τύπων λάθους: missed detection και false alarms. Ένας σταθμισμένος μέσος όρος των ρυθμών των missed detection και false alarms μπορεί να χρησιμοποιηθεί ως είδος συνάρτησης κόστους. Υπάρχουν δύο στοιχεία που πρέπει να σημειωθούν για την καμπύλη DET. Κατ' αρχάς, εάν οι προκύπτουσες καμπύλες είναι ευθείες γραμμές, τότε αυτό παρέχει μια οπτική

επιβεβαίωση ότι οι ελλοχεύουσες κατανομές πιθανότητας του συστήματος είναι κανονικές. Δεύτερον, η διαγώνιος y =-x (με το X στην κανονική κλίμακα παρέκκλισης) αντιπροσωπεύει την τυχαία απόδοση. Με έναν μεγάλο αριθμό στόχων και τα κατά προσέγγιση ίσα περιστατικά όλων των στόχων, η γενική απόδοση αντιπροσωπεύεται αποτελεσματικά από αυτού του είδους τις καμπύλες. Η καμπύλη DET έχει ευδιάκριτα πλεονεκτήματα σε σχέση με τις τυποποιημένες καμπύλες ROC για την παρουσίαση αποτελεσμάτων απόδοσης όπου περιλαμβάνονται οι ανταλλαγές δύο τύπων λάθους.

Μια ειδική περίπτωση στόχων ανίχνευσης που είναι κοινή σε πολλές διαφορετικές εφαρμογές οι οποίες όχι μόνο αποφασίζουν εάν συμβαίνει μια ορισμένη περίπτωση αλλά αποφασίζουν επίσης και για ποιο μέρος του πληθυσμού ενός συνόλου συμβαίνει. Ένα χαρακτηριστικό παράδειγμα για αυτό είναι η κατάτμηση εικόνας (image segmentation). Εκεί το ζητούμενο δεν είναι μόνο να ανιχνευθεί εάν υπάρχει κάτι, αλλά και το που είναι. Για την αξιολόγηση τέτοιων εφαρμογών, όπου το αποτέλεσμα δεν είναι μια απλή δυαδική απόφαση αλλά μια δυαδική απόφαση για κάθε μέλος ενός συνόλου (στο δεδομένο παράδειγμα της κατάτμησης εικόνας: αποφασίζουμε για κάθε εικονοστοιχείο (pixel) εάν υπάρχει κάτι ή όχι) τα μέτρα Recall, Precision, Accuracy και F-measures χρησιμοποιούνται επιπρόσθετα με τα ήδη προαναφερθέντα.

Οι ορισμοί αυτών των μετρικών είναι οι παρακάτω:

- Precision

$$Precision = \frac{(groundtruth \cap detected)}{(detected)},$$

Το Precision είναι το ποσοστό του πληθυσμού που σωστά ανιχνεύεται σαν positive (True Positives), σε σχέση με το μέρος του πληθυσμού που ανιχνεύεται σαν positive (True Positive +False Positives).

- Recall

$$Recall = \frac{(groundtruth \cap detected)}{(groundtruth)},$$

Το Recall είναι το ποσοστό του πληθυσμού που σωστά ανιχνεύεται σαν positive (True Positives), σε σχέση με το μέρος του πληθυσμού που είναι πραγματικά positive (True positives + False Negatives).

- Accuracy

$$Accuracy = \frac{(groundtruth \cap detected) \cup (\overline{groundtruth} \cap \overline{detected})}{(groundtruth \cup \overline{groundtruth})}$$

Το Accuracy είναι το ποσοστό των πραγματικών αποτελεσμάτων εντοπισμού μέσα στο πληθυσμό (True Positives + True Negatives) σε σχέση με το μέγεθος του πληθυσμού (True Positives + True Negatives + False Positives + False Negatives)

Το F1-measure είναι ένα μετρικό απόδοσης που δημιουργήθηκε για την αντιπροσώπευση της ποιότητας ενός συστήματος εντοπισμού με ένα απλό αριθμό. Υπολογίζεται από τον αρμονικό μέσο των Precision και Recall:

- F_1-Measure (συχνά καλείται απλώς F-Measure)

$$F_1 = \frac{2(Precision \cdot Recall)}{Precision + Recall}$$

- Πέρα από το κλασικό F-Measure, στο οποίο τα Precision και Recall ισοσταθμίζονται, το αποκαλούμενο F_a Measure βάζει διαφορετικό βάρος πάνω στα Precision και Recall:

$$F_\alpha = \frac{(1+\alpha) \cdot (Precison \cdot Recall)}{(alpha \cdot Precision) + Recall}$$

Οι τιμές όλων των μετρικών που αναφέρθηκαν (Precision, Recall και F-Measures) κυμαίνονται μεταξύ μηδέν και ένα. Όσο πιο κοντά βρίσκονται στο ένα τόσο καλύτερο είναι το σύστημα.

Α.2. Προβλήματα πολλών καταστάσεων (multiple hypothesis)

Για μία εφαρμογή, στην οποία ο κατηγοριοποιητής πρέπει να χαρακτηρίσει μία σειρά από δεδομένα με μία από πολλές κατηγορίες (επίσης καλείται 1:N κατηγοριοποίηση) ένα τυπικό μέτρο απόδοσης είναι το Recognition Rate (RR), το οποίο συσχετίζει τον αριθμό των σωστά ταξινομημένων περιπτώσεων με τον συνολικό αριθμό αυτών.

$$RR = \frac{Number\ of\ correctly\ classified\ test\ sets}{Overall\ number\ of\ testsets}$$

Η καμπύλη Cumulative Matching curve (CMC) είναι μία διαφοροποιημένη εκδοχή του Recognition Rate. Υποθέτει ότι ο κατηγοριοποιητής δεν βρίσκει μόνο τη κατηγορία των δεδομένων δοκιμής αλλά προσδιορίζει και ένα είδος σκορ για κάθε μία κατηγορία. Οι αποκαλούμενες n-best λίστες, στις οποίες οι n-κατηγορίες με τα υψηλότερα σκορ κατηγοριοποίησης για κάθε περίπτωση δοκιμής μπορούν να προσδιοριστούν. Τώρα τα

CMC σχεδιάζονται συναρτήσει του n σχετικά με το σε ποιο ποσοστό της n-best λίστας βρίσκεται η σωστή κατηγορία.

$$CMC(n) = \frac{Number\ of\ correct\ classifications \quad \textbf{in - bestlist}}{Number\ of\ testdatan}$$

Η καμπύλη CMC αποτελεί μία μονοτονικά αυξανόμενη καμπύλη, ξεκινώντας με μηδέν για n=0 και καταλήγοντας στο ένα εάν το n είναι ο αριθμός όλων των πιθανών κατηγοριών. Αξίζει να προσέξουμε ότι για n=1 η CMC(1) είναι ίση με το Recognition Rate.

Μια άλλη χαρακτηριστική αντιπροσώπευση για τα προβλήματα πολλών κατηγοριών, ειδικά εάν η γενική διαστατικότητα του προβλήματος είναι μικρότερη από τις δώδεκα κατηγορίες είναι η *μήτρα σύγχυσης* (*confusion matrix*). Σε αυτήν την αντιπροσώπευση κάθε σειρά της μήτρας αντιπροσωπεύει τις περιπτώσεις μιας προβλεφθείσας ή αναγνωρισμένης κατηγορίας, ενώ κάθε στήλη αντιπροσωπεύει τις περιπτώσεις της πραγματικής κατηγορίας (ground truth). Η μήτρα σύγχυσης είναι χρήσιμη για την αξιολόγηση, επειδή είναι εύκολο να δει κανείς, εάν και το κατά πόσο το σύστημα συγχέει μερικές κατηγορίες.

Για παράδειγμα θα μπορούσε να είναι δυνατό από π.χ. σε ένα πρόβλημα με τρεις κατηγορίες, μία κατηγορία να αναγνωρίζεται τέλεια ενώ οι δύο άλλες κατηγορίες αναμιγνύονται τυχαία και να μη μπορούν να διακριθούν από το σύστημα.

Το γενικό ποσοστό αναγνώρισης ενός τέτοιου συστήματος, υποθέτοντας ομοιόμορφη κατανομή των κατηγοριών, θα ήταν RR=1/3 + 0.5 * 1/3 + 0.5 * 1/3 = 2/3. Αυτό το RR δεν αντιπροσωπεύει το γεγονός ότι μια από τις κατηγορίες αναγνωρίζεται τέλεια, κάτι που θα μπορούσε να ασκήσει σημαντική επίδραση σε μια πιθανή εφαρμογή. Μια μήτρα σύγχυσης ενός τέτοιου συστήματος θα έμοιαζε με την παρακάτω:

	Class1	Class2	Class3
Class1	1	0	0
Class2	0	0.5	0.5
Class3	0	0.5	0.5

όπου τα ποσοστά αναγνώρισης εξαρτώμενα από κάθε κατηγορία δηλώνονται στη κύρια διαγώνιο και η σύγχυση με τις άλλες κατηγορίες αντιπροσωπεύεται από τα στοιχεία που δεν βρίσκονται στην κύρια διαγώνιο.

Α.3. Τρόποι αξιολόγησης συστημάτων που επεξεργάζονται ακουστικά σήματα

Διάφορες υποκειμενικές και αντικειμενικές δοκιμές αξιολόγησης έχουν αναπτυχθεί για τη μέτρηση της ακουστικής ποιότητας. Αυτές οι δοκιμές εφαρμόζονται είτε ως υποκειμενικές αξιολογήσεις του ακούσματος είτε ως αντικειμενικές αξιολογήσεις (συχνά που δεν συμπεριλαμβάνουν την αντιληπτική ποιότητα, μη διαισθητικά, κ.λπ.). Το Mean Opinion Score (MOS) (ITU-T Recommendation P.800, 1996) αποτελεί ένα μετρικό στο οποίο η ποιότητα του ήχου αξιολογείται σε μια κλίμακα 5 σημείων (5: excellent, 4: good, 3: fair, 2: poor, 1: bad) και είναι η ευρύτατα χρησιμοποιημένη υποκειμενική δοκιμή.

Επιπλέον, το πολύ ευρέως χρησιμοποιημένο αντικειμενικό μέτρο απόδοσης, που χρησιμοποιείται για να αξιολογήσει την ποιότητα των ακουστικών σημάτων, είναι η αποσπασματική αναλογία σήματος προς θόρυβο (Signal to Noise Ratio). Το SNR ορίζεται ως η αναλογία μεταξύ της ενέργειας του καθαρού σήματος στόχου και της ενέργειας της διαφοράς μεταξύ των σημάτων παραγωγής και στόχου για το πλαίσιο m. Μετρημένο στη κλίμακα DB, το SNR ορίζεται ως:

$$SNR(t,c) = 10\log_{10}\frac{\sum_{n=1}^{N}\left|H_t^m(e^{j2\pi n/N})\right|^2}{\sum_{n=1}^{N}\left(\left|H_c^m(e^{j2\pi n/N})\right| - \left|H_t^m(e^{j2\pi n/N})\right|\right)^2}$$

όπου το N είναι το μέγεθος του μετασχηματισμού Fourier, το t είναι το αρχικό σήμα στόχου και το y είναι το σήμα παραγωγής/εξόδου. Οι υψηλότερες τιμές για το SNR δείχνουν μια καλύτερη ακουστική ποιότητα. Η δοκιμή MOS και τα μέτρα SNR είναι συμπληρωματικές η μια της άλλης και μπορούν να χρησιμοποιηθούν μαζί για την αξιολόγηση οποιουδήποτε συστήματος που εμπεριέχει ακουστική παραγωγή.

Ένα ακουστικό σύστημα αναγνώρισης τυπικά αξιολογείται με τον ακόλουθο τρόπο: συνήθως ένα μέρος της βάσης δεδομένων με γνωστό το ground truth χρησιμοποιείται για την κατάρτιση του συστήματος ενώ το υπόλοιπο των δεδομένων χρησιμοποιείται για τη δοκιμή του. Η δοκιμή εκτελείται είτε σε μια βάση πλαίσιο ανά πλαίσιο είτε για κάθε ηχητικό δείγμα. Οι τυποποιημένες μετρικές αξιολόγησης όπως τα ποσοστά αναγνώρισης και οι μήτρες σύγχυσης χρησιμοποιούνται κατά τη διάρκεια αυτού του σταδίου. Στην περίπτωσή μας εξετάζουμε επίσης το πρόβλημα της ανίχνευσης μη τυπικού γεγονότος, η οποία δεν μπορεί να αξιολογηθεί χρησιμοποιώντας αποκλειστικά αυτήν την διαδικασία. Σε μία τέτοια εφαρμογή, το σύστημα πρέπει να εξεταστεί για δύο τύπους λαθών: (α) να μην είναι σε θέση να ανιχνεύσει ένα ανώμαλο γεγονός ενώ υπάρχει και (β) να ανιχνεύσει ψευδώς ένα ανώμαλο γεγονός. Και τα δύο λάθη είναι κρίσιμα και πρέπει να αποφευχθούν όσο το δυνατόν περισσότερο. Λόγω αυτών των απαιτήσεων χρησιμοποιήσαμε τις καμπύλες DET που μπορούν αποτελεσματικά να μετρήσουν την απόδοση σε εφαρμογές ανίχνευσης και που περιλαμβάνουν μια ανταλλαγή των τύπων λάθους. Η απόδοση ενός τέτοιου συστήματος δεν μπορεί να αναλυθεί από ένα απλό ποσοστό αναγνώρισης λόγω του λάθους ανταλλαγής - η

ανίχνευση μιας άτυπης κατάστασης μπορεί να αποτύχει ή ένα τέτοιο γεγονός μπορεί να ανιχνευθεί ενώ δεν είναι παρόν.

Η έξοδος του συστήματος είναι μία πιθανότητα ότι το αντίστοιχο ακουστικό τμήμα είναι μέρος μίας κατηγορίας στόχου. Η κλίμακα της πιθανότητας είναι αυθαίρετη, αλλά πρέπει να είναι σύμφωνη σε όλες τις αποφάσεις, ενώ οι μεγαλύτερες τιμές δείχνουν μεγαλύτερη πιθανότητα της ύπαρξης ενός στόχου. Αυτές οι πιθανότητες χρησιμοποιούνται για να δημιουργήσουμε την καμπύλη απόδοσης που επιδεικνύει τη σειρά των πιθανών λειτουργικών χαρακτηριστικών.

Παράρτημα Β

Παρουσίαση του συστήματος PROMETHEUS

PROMETHEUS: Πρόβλεψη και ερμηνεία της ανθρώπινης συμπεριφοράς βασισμένη στις πιθανολογικές δομές και τους ετερογενείς αισθητήρες

Β.1. Η τρέχουσα εφαρμογη

Η τρέχουσα εφαρμογή αποτελεί το πρόγραμμα PROMETHEUS (FP7-214901) που στοχεύει στη θέσπιση ενός γενικού πλαισίου που συνδέει τους θεμελιώδεις στόχους αντίληψης με αυτοματοποιημένες διαδικασίες επίβλεψης, επιτρέποντας την ερμηνεία και τη βραχυπρόθεσμη πρόβλεψη των μεμονωμένων αλλά και συλλογικών ανθρώπινων συμπεριφορών σε ένα απρόβλεπτο περιβάλλον καθώς επίσης και τις σύνθετες ανθρώπινες αλληλεπιδράσεις σε αυτό.

Για να επιτύχει τους προαναφερθέντες στόχους, η κοινοπραξία PROMETHEUS λειτουργεί στους ακόλουθους επιστημονικούς και τεχνολογικούς στόχους:

1. διαμόρφωση αισθητήρων και συνδυασμό πληροφοριών από πολλαπλές, ετερογενείς αντιληπτικές μορφές

2. διαμόρφωση, εντοπισμός, και έλεγχος πολλών ανθρώπων

3. διαμόρφωση, αναγνώριση, και βραχυπρόθεσμη πρόβλεψη της σύνθετης συνεχούς ανθρώπινης συμπεριφοράς.

Β.2. τεχνολογία

Η τεχνολογία PROMETHEUS είναι βασισμένη στη χρήση ενός δικτύου ετερογενών αισθητήρων, οι οποίοι συλλέγουν στοιχεία κάτω από ένα πλήρως πιθανολογικό πλαίσιο. Αυτό το πλαίσιο εκτελεί ένα προσαρμοστικό συνδυασμό των ετερογενών πηγών πληροφορίας, η οποία περιλαμβάνει και πληροφορίες ενσωμάτωσης

- υπό την ευρύτερη έννοια

- για να ανιχνεύσει, να υπολογίσει και να προβλέψει με σφαιρική αντίληψη την κατάσταση της αλληλοεπίδρασης των ανθρώπων (αναφορά σε Fig.1 παρακάτω).

Ο πλεονασμός και η συμπληρωματικότητα των πληροφοριών που παρέχονται από τους ετερογενείς αισθητήρες διευκολύνουν μια καλή εκτίμηση κάθε βούλησης με συνέπεια την ερμηνεία των συμπεριφορών. Επιπλέον, στο πρόγραμμα PROMETHEUS το σύνολο των αισθητήρων επιλέχτηκε με έναν τρόπο που επιτρέπει την υπέρβαση των αδυναμιών από κάθε αντιληπτική μορφή όσον αφορά την κάλυψη της περιοχής και της απάντησής των κλειστών, θορυβωδών και διαφορετικών περιβαλλοντικών συνθηκών.

Η περιγραφή μιας πολλαπλών στάθμεων και πολύπλευρης διαδικασίας εξετάζεται με την αυτόματη ανίχνευση, την ένωση, το συσχετισμό, την εκτίμηση, και τον συνδυασμό στοιχείων από πολλαπλές αντιληπτικές μορφές που περιλαμβάνουν την προεπεξεργασία στοιχείων και την εξαγωγή χαρακτηριστικών γνωρισμάτων που ακολουθούνται από μια ιεραρχία τεσσάρων επίπεδων. Αυτά τα ποιο υψηλά επίπεδα επεξεργασίας είναι:

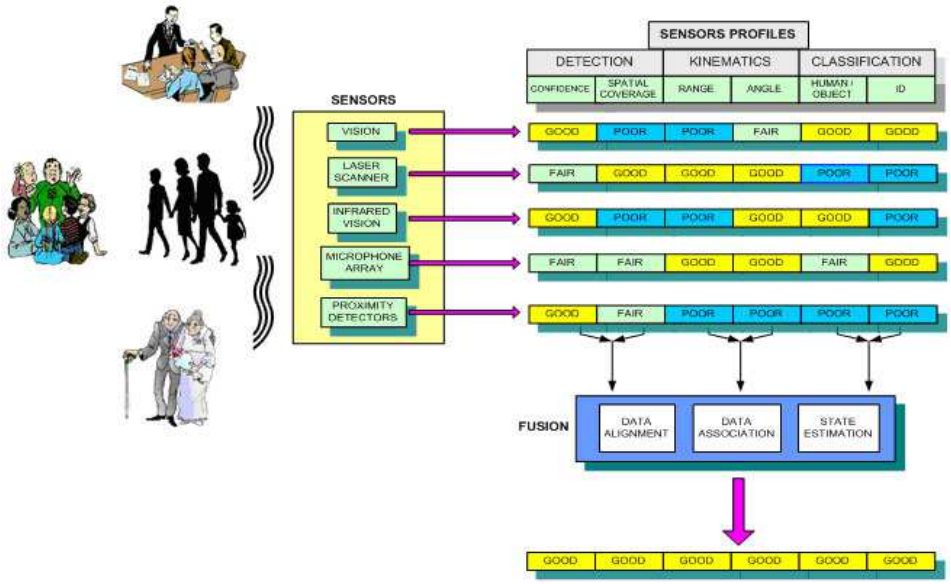
(ι) αξιολόγηση του αντικειμένου,

(ii) αξιολόγηση της κατάστασης,

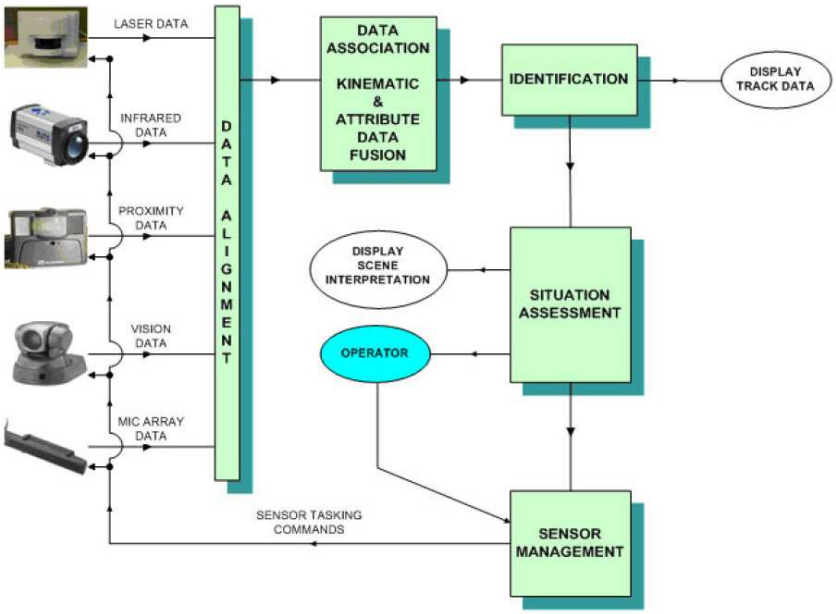
(iii) αξιολόγηση του αντίκτυπου

(iv) διαχείριση αισθητήρων.

Η λογική ροή μεταξύ των επίπεδων επεξεργασίας παρουσιάζονται παρακάτω



Σχήμα 1. Μίξη των ετερογενών αισθητήρων, που προσφέρουν τις συμπληρωματικές πληροφορίες, επιτρέπει τη γερή εκτίμηση



Σχήμα 2. Γενική αρχιτεκτονική συστημάτων

Β.3. εφαρμογές

Ενώ το προτεινόμενο πλαίσιο είναι γενικού σκοπού, η τεχνολογία που αναπτύσσεται στο πρόγραμμα PROMETHEUS θα εφαρμοστεί και θα καταδειχθεί σε διάφορα σενάρια σχετικά με την επιτήρηση των μεγάλων κοινών χώρων ή την φιλική προς τον άνθρωπο μέσω των μηχανών αλληλεπίδραση του οικογενειακού περιβάλλοντος. Ανάλογα με το συγκεκριμένο πρόβλημα το οποίο είναι προσιτό, ποικίλες εφαρμογές μπορούν να εκμεταλλευτούν την τεχνολογία PROMETHEUS.

Τα χαρακτηριστικά παραδείγματα, μπορούν να είναι στους τομείς της ασφάλειας (ανίχνευση από την μη φυσιολογική συμπεριφορά), στο σπίτι/υγειονομική περίθαλψη για τους ηλικιωμένους ανθρώπους, τηλεοπτική ανάλυση των αθλητικών δραστηριοτήτων σημεία πώλησης που διαφημίζουν, ή ανάλυση πελατών μέσω ακροατηρίου, κ.λπ.

Β.4. επίδειξη

Ένα σημαντικό μέρος στο πρόγραμμα PROMETHEUS στοχεύει στην ενίσχυση των στοιχείων μιας μεγάλης βάσης δεδομένων πολυαισθητήρων που μπορεί να χρησιμοποιηθεί για την ανάπτυξη των νέων αλγορίθμων και τη δημιουργία προτύπων του φυσικού περιβάλλοντος.

Εκτός από τυπικές από άνθρωπο σε άνθρωπο και από άνθρωπο προς αντικείμενο αλληλεπιδράσεις, η βάση δεδομένων περιέχει παραδείγματα των μη φυσιολογικών συμπεριφορών, όπως κλοπές σε δράση, καυγάδες στο δρόμο κ.λπ. Σύνθετες θεματικές ιστορίες και θεματικές εργασίες δημιουργούνται, έτσι ώστε οι ηθοποιοί να παίρνουν τις ακριβείς εντολές πού να πάνε, την όποια διαδρομή θα πάρουν και τι είδους δραστηριότητα θα παρουσιάζουν.

Ένα από τα πρώτα προβλήματα που αντιμετωπίζονται σε μια εφαρμογή επιτήρησης που χρησιμοποιεί ένα ετερογενές δίκτυο αισθητήρων είναι να ανιχνευθούν σχήμα 4 και να ακολουθηθούν σχήμα 2 διάφορα άτομα κάτω από αυστηρούς κλειστούς κανόνες. Για να εξετάσει αυτές τις δυσκολίες, η βάση δεδομένων PROMETHEUS καταγράφεται σε υπαίθρια και εσωτερικά περιβάλλοντα, που χρησιμοποιούν ένα ετερογενές σύνολο αισθητήρων:

(ι) φωτογραφικές μηχανές υψηλής ανάλυσης για την επισκόπηση και τη λεπτομέρεια της σκηνής,

(ii) τρισδιάστατες φωτογραφικές μηχανές,

(iii) θερμική υπέρυθρη φωτογραφική μηχανή,

(iv) σειρές μικροφώνων.

Το υπαίθριο σενάριο επιλέχτηκε για παρουσίαση γιατί παρέχει μια στενή αντιστοιχία από πραγματικές καταστάσεις, με μεταβαλλόμενους όρους φωτισμού, κινούμενο υπόβαθρο και έναν ποικίλο αριθμό ανθρώπων

μέσα σε μια σκηνή όπου οι παρεμβάσεις διαμορφώνουν το περιβάλλον. Μια πρόσθετη πρόκληση προέρχεται από τις υψηλές περιβαλλοντικές θερμοκρασίες, οι οποίες παρεμποδίζουν τη λειτουργιά των θερμικών αισθητήρων. Το σχήμα 3 επεξηγεί δύο απόψεις της υπαίθριας περιοχής καταγραφής στοιχείων.

Σχήμα 3. Άποψη παραδείγματος μιας υπαίθριας περιοχής καταγραφής στοιχείων

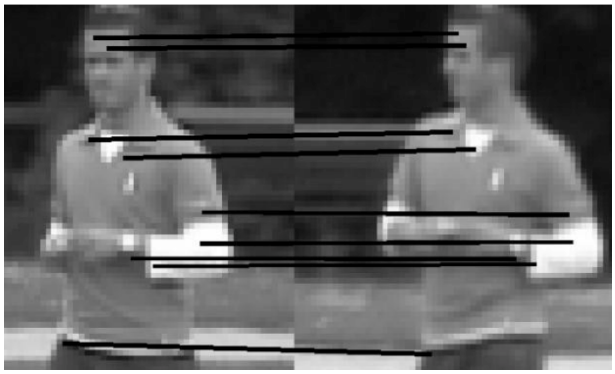
Αυτά τα ρεαλιστικά στοιχεία θα χρησιμοποιηθούν για να καταδείξουν την βάση δεδομένων των αλγόριθμων καθώς και τυχόν βελτιώσεις αυτών. Για παράδειγμα, η εργασία που παρουσιάζεται στο εργαστήριο ΚΑΤΟΙΚΙΔΙΩΝ ΖΩΩΝ το 2007 [1] είναι διευκρινισμένη στο σχήμα 4).

Συνεπώς, παρουσιάζουμε τα πρώτα αποτελέσματα για το πώς λειτουργεί υπό πραγματικές συνθήκες, η τεχνολογία έλεγχου όπως εκείνη καταγράφεται κατά τη διάρκεια της αρχικής καταγραφής. Ένα δεύτερο σημείο είναι να αντιμετωπιστεί η περίπτωση όπου ένα πρόσωπο εκ νέου λειτουργεί και καταγράφεται ως πραγματικό στοιχείο, σε μια σκηνή μετά από κάθε στιγμή. Ο στόχος είναι βρίσκοντας έναν αποδοτικό αλγόριθμο, ο οποίος θα είναι ικανός να προσθέσει ένα νέο στοιχειό σε μια υπάρχουσα βάση δεδομένων, ενημερώνοντας το πρωτότυπο για ότι διαφορετικό θέτοντας και αναγνωρίζοντας τους ανθρώπους που επανεμφανίζονται. Επομένως τα αμετάβλητα χαρακτηριστικά γνωρίσματα πρέπει να εξαχθούν από τη παρουσία του κάθε προσώπου και να αποθηκευτούν σε ένα πρότυπο.

Σχήμα 4. Παράδειγμα από το εργαστήριο 2007 ΚΑΤΟΙΚΙΔΙΩΝ ΖΩΩΝ

Οι πρώτες δοκιμές αναλύουν τα χαρακτηριστικά γνωρίσματα που παρουσιάζονται από το κάθε πρόσωπο και μας δίνουν ένα αποτέλεσμα σε κάθε διαφορετικό οπτικό πεδίο. Σε μια ιδανική εφαρμογή δοκιμάζεται να δημιουργηθεί ένα τρισδιάστατο πρότυπο από κάθε πρόσωπο και από το καθένα, ένα ή περισσότερα οπτικά πεδία. Ένα δείγμα για το ταίριασμα των σημείων το σώματος ενός προσώπου επιδεικνύεται στο σχήμα 5. Επιπλέον, μια αυτόματη ανάλυση των χαρακτηριστικών σκηνών εκτελείται χρησιμοποιώντας τις

πιθανολογικές προσεγγίσεις, όπως παρουσιάζονται στο σχήμα 3. Τα αποτελέσματα από μια προσέγγιση για την ανίχνευση αποσκευών παρουσιάζονται στο σχήμα 4.



Σχήμα 5. Ταίριασμα των χαρακτηριστικών γνωρισμάτων

Τεχνική Ορολογία

- Bayesian Information Criterion (BIC) - Μπαεζιανό κριτήριο πληροφοριών
- Computational Auditory Scene Analysis (CASA) - υπολογιστική ακουστική ανάλυση σκηνής
- Discrete Cosine Transform (DCT) – διακριτός μετασχηματισμός συνημίτονου
- Discrete Fourier Transform (DFT) - διακριτός μετασχηματισμός Fourier
- Discrete Wavelet Transform (DWT) - διακριτός μετασχηματισμός wavelet
- Filter bank – τράπεζα φίλτρων
- Gaussian Mixture Model (GMM) - Μοντέλο Γκαουσσιανών κατανομών
- Hidden Markov Model (HMM) – κρυμμένο μοντέλο Markov
- Low level descriptors (LLD) - Περιγραφείς χαμηλού επιπέδου
- Maximum a-posteriori probability (MAP) – μέγιστη εκ των υστέρων πιθανότητα
- Maximum Likelihood (ML) – μέγιστη πιθανοφάνεια
- Mel frequency cepstral coefficients (MFCC) – cepstral συντελεστές της κλίμακας Mel
- Multilayer Perceptron (MLP) - Πολυστρωματικό perceptron
- Novelty Detection - Ανίχνευση Καινοτομίας
- Pitch – θεμελιώδης συχνότητα
- Principal Component Analysis (PCA) - Ανάλυση κύριων συνιστωσών
- Radial basis function – συνάρτηση ακτινικής βάσης
- Short time Fourier Transform (STFT) - Βραχύχρονος μετασχηματισμός Fourier
- Signal to Noise Ratio (SNR) - Λόγος σήματος προς θόρυβο
- Support vector machines (SVM) – μηχανές υποστήριξης διανυσμάτων
- Soundscene - Ακουστική σκηνή
- Voice Activity Detection (VAD) - Εντοπισμός σήματος ομιλίας

Βιβλιογραφία

Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis, "An Adaptive Framework for
Acoustic Monitoring of Potential Hazards", *EURASIP* Journal on Audio, Speech, and Music Processing Volume 2009 (2009), Article ID 594103, doi:10.1155/2009/594103 (impact factor: 0.341).

Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis, "Exploiting Temporal Feature
Integration for Generalized Sound Recognition", *EURASIP* Journal on Advances in Signal Processing, Volume 2009 (2009), Article ID 807162, doi:10.1155/2009/807162, (impact factor: 1.012).

Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis, "A Practical System for Acoustic Surveillance of Hazardous Situations", *International Journal of Artificial Intelligence Tools*, vol. 20(1), doi: 10.1142/S021821301100005X (impact factor: 0.320).

Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis, "Probabilistic Novelty Detection for Acoustic Surveillance Under Real-World Conditions", *IEEE Transactions on Multimedia*, vol. 13, no. 4, August 2011 (impact factor: 1.684).

Stavros Ntalampiras, Ilyas Potamitis, Nikos Fakotakis, and Spyros Kouzoupis, "Automatic
Recognition of an Unknown and Time-Varying Number of Simultaneous Environmental Sound Sources", Journal of World Academy of Science, Engineering and Technology, vol.59,
2011 (invited). https://www.waset.org/journals/waset/v59/v59-393.pdf

Stavros Ntalampiras and Nikos Fakotakis, "Modeling the Temporal Evolution of Acoustic
Parameters for Speech Emotion Recognition", *IEEE Transactions on Affective Computing*, Jan.-March 2012, vol. 3, no. 1, pp. 116-125.
http://www.computer.org/csdl/trans/ta/2012/01/tta2012010116-abs.html.

Stavros Ntalampiras, Dejan Arsic, Martin Hofmann, Maria Andersson and Todor Ganchev,
"PROMETHEUS: Heterogeneous sensor database in support of research on human Behavioral patterns in unrestricted environments", *Signal, Image and Video Processing*, Springer, June 2012, DOI 10.1007/s11760-012-0346-9 (impact factor 0.617).
http://link.springer.com/article/10.1007%2Fs11760-012-0346-9

Stavros Ntalampiras, Ilyas Potamitis, Nikos Fakotakis, "Acoustic Detection of Human
Activities in Natural Environments", *Journal of Audio Engineering Society*, Vol. 60, No. 9, 2012 September (impact factor 0.537). http://www.aes.org/e-lib/browse.cfm?elib=16373

Stavros Ntalampiras, "A Novel Holistic Modeling Approach for Generalized Sound Recognition", *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 185-188, Feb. 2013,http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6403508 (impact factor 1.388).

Arrigoni, J.E., "An Evaluation Of Amphibian Monitoring Approaches in the Maya Forest", Chapter 3: An assessment of the vocalization survey method for monitoring anuran populations in the Maya Forest, *Master thesis*, pp. 21–42, February, 2003.

Arsic D., Schuller B., Rigoll G., "Multiple Camera Person Tracking in Multiple Layers Combining 2D and 3D Information", *M2SFA2' 2008*, Marseille, France, 2008.

Arsic, D., Hofmann, M., Schuller B., Rigoll G., "Multi-Camera Person Tracking and Left Luggage Detection Applying Homographic Transformation", *PETS 2007*, IEEE, 2007.

Atrey P. K., Maddage N. C., Kankanhalli MS., "Audio based event detection for multimedia surveillance," *in International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, May 2006.

Aucoutourier J. J., Defreville B., Pachet F., "The bag of frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *In Journal of Acoustical Society of America*, vol. 122, no. 2, pp. 881-891, 2007.

Baluja S., Covell M., "Waveprint: Efficient wavelet-based audio fingerprinting," *In Pattern Recognition*, vol. 41, pp. 3467-3480, 2008.

Baum L.E., Petrie T., "Statistical Inference for Probabilistic Functions of Finite State Markov Chains", In *Annals of Mathematical Statistics*, vol. 37, pp. 1554–1563, 1966.

Bengio Y., Frasconi P., "Input-output HMM's for sequence processing". In *IEEE Transactions on Neural Networks*, vol. 7, no. 5, pp. 1231–1249, 1996.

Bengio S., Mariethoz J., "Learning the decision function for speaker verification". *Technical* Report, IDIAP Research Report 00-40, IDIAP, January 2001.

Benyassine, A., Shlomot, E., Su, H.-Y., "ITU Recommendation G.729 Annex B: A Silence Compression Scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," In *IEEE Communications Magazine*, pp. 64–73, 1997.

Berenzweig, A., Ellis, D.P.W., Lawrence, S., "Anchor space for classification and similarity measurement of music", In *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 29–32, 2003.

Bishop C., "Novelty detection and neural network validation", *in Proc. IEEE Conf. on Vision and* Image Signal Processing, pp. 217-222, 1994.

Bourlard H.A., Morgan, N., “Connectionist Speech Recognition: A Hybrid Approach”, Kluwer Acadmeic Publishers, 1994.

Burred J. J., Haller M., Jin S., Samour A., Sikora T., *Audio content analysis*, book chapter in P. Hobson and Y. Kompatsiaris (Eds.), Semantic Multimedia and Ontologies: Theory and Applications, Springer, 2008.

Carolan, T.A., Kidd, S.R., Hand, D.P., Wilcox, S.J., Wilkinson, P., Barton, J.S., Jones, J.D.C., Reuben, R.L., “Acoustic emission monitoring of tool wear during the face milling of steels and aluminium alloys using a fiber optic sensor energy analysis”. In *Proceedings of the Institution of Mechanical Engineers, 211(B)*, pp. 299–309, 1997.

Casey M., “MPEG-7 sound recognition tools,” In IEEE Transactions on Circuits and Systems for *Video Technology*, vol. 11, no. 6, pp. 737–747, 2001.

Chen, F., *Speech technology in military applications*, book chapter of Designing Human Interfaces in Speech Technology, Springer US publisher, March 2006.

Cho, Y.D., Kondoz, A., “Analysis and improvement of a statistical model-based voice activity detector,” In *IEEE Signal Processing Letters*, vol. 8, pp. 276–278, 2001.

Chu S., Narayanan S., Jay Kuo C.-C., “Content Analysis for Acoustic Environment Classification in Mobile Robots,” In Proceedings of AAAI 2006 Fall Symposium, Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Systems, Arlington, VA, 2006.

Clarkson, B., Sawhney, N., Pentland, A., “Auditory Context Awareness via Wearable Computing”, In Proceedings of the Workshop on Perceptual User In-terfaces, November 1998.

Clavel C., Ehrette T., Richard G., “Event detection for an audio-based surveillance system,” *in* Proceedings of the IEEE Int. Conference on Multimedia and Expo, Amsterdam, 2005, pp. 1306-1309.

Clavel C., Vasilescu I., Devillers L., Ehrette T., “Fiction database for emotion detection in abnormal situations”, *ICSLP’ 2004*, Jeju, Korea, 2004.

Clavel C., Vasilescu I., Devillers L., Richard G., Ehrette T., “Fear-type emotion recognition for future audio-based surveillance systems,” *Speech Communication*, vol. 50, no. 6, pp. 487- 503, June 2008.

Cover T., Hart, P., “Nearest Neighbour Pattern Classification”, In *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, 1967.

Daubechies I., "The wavelet transform, time-frequency localization and signal analysis", *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 961-1005, 1990.

Davy M., Desobry F., Gretton A., Doncarli C., “An online support vector machine for abnormal events detection,” *Signal Processing*, Elsevier, vol. 86, no. 8, pp. 2009-2025, Aug. 2006.

Davy M., Godsill S., “Detection of abrupt spectral changes using support vector machines: an application to audio signal segmentation,” *in Proc. International Conference on Acoustics, Speech and Signal Processing*, Orlando, 2002, pp. 1313-1316.

Defreville B., Roy P., Rosin C., Pachet F., "Automatic recognition of urban sound sources", *AES 120th Convention*, Paris, France, 2006.

Didiot E., Illina I., Mella O., Fohr D., Haton J.-P., “A wavelet-based parameterization for speech/music segmentation,” *In INTERSPEECH-06*, Pittsburg, Pennsylvania, 17-21 September, 2006.

Diei, E.N., Dornfeld, D.A., “Acoustic emission sensing of tool wear in face mill-ing”, In Transactions of ASME, Journal of Engineering for Industry, vol. 109, pp. 234–240, 1987.

Dimla, D.E., Jr, Lister, P.M., Leighton, N.J., “Neural network solutions to the tool condition monitoring problem in metal cutting. A critical review of meth-ods”. In *International Journal of Machine Tools Manufacturing*, vol. 37, no. 9, pp. 1219–1240, 1997.

Diniz, A.E., Liu, J.J., Dornfeld, D.A., “Correlating tool life, tool wear and sur-face roughness by monitoring acoustic emission in turning”, In *Wear*, vol. 152, pp. 395–407, 1992.

EBU, "SQAM - CD: Sound quality assessment material", Polygram Cat. No 422 204-2, European Broadcasting Union (EBU), 1988.

EBU, "Sound quality assessment material", Recordings for subjective tests – Users/ handbook for the EBU-SQAM Compact Disc, Tech 3253, European Broadcasting Union (EBU), 1988.

Eggink, J., Brown, G.J., “A missing feature approach to instrument identification in polyphonic music”, In *Proceedings of the ICASSP’03*, Hong Kong, pp. 553–556, April 2003.

Eggink, J., Brown, G.J., “Instrument recognition in accompanied sonatas and concertos,” In *Proceedings of the ICASSP’04*, Montreal, Canada, pp. 217–220, May 2004.

El-Maleh K., Klein M., Petrucci G., Kabal P., "Speech/music discrimination for multimedia applications", *In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol.6, pp. 2445-2448, 2000.

El-Maleh K., Samouelian A., Kabal P., “Frame level noise classification in mobile environments,” *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 237-240, 1999.

Elman J.L., “Finding structure in time”, In *Cognitive Science*, vol. 14, pp. 179–211, 1990. Emamian V., Kaveh M., Tewfik A. H., “Robust clustering of acoustic emission signals using the Kohonen network,” in Proc. International Conference on Acoustics, Speech and Signal *Processing*, Istanbul, 2000, pp. 3891-3894.

Erne M., Moschytz G., Faller C., “Best wavelet-packet bases for audio coding using perceptual and rate-distortion criteria,” *In ICASSP 1999*, Phoenix, Arizona, 15-19 March, 1999.

Eronen A.J., Peltonen V.T, Tuomi J.T., Klapuri A.P., Fagerlund S., Sorsa T., Lorho G., Huopaniemi J., “Audio-based context recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 321 – 329, January 2006.

Farge M., "Wavelet Transforms and their Applications to Turbulence", *Ann. Rev. of Fluid Mech.*, vol. 24, pp. 395 - 457, 1992.

Ferryman J., Tweed D, “An Overview of the PETS 2007 Dataset”, *PETS 2007*, IEEE, 2007.

Fisher R.A., “The Use of Multiple Measurements in Taxonomic Problems”, In *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

FitzGerald, D., Coyle, E., Lawlor, B., “Sub-band Independent Subspace Analysis for Drum Transcription,” In *Proceedings of the DAFX’02*, pp. 65–69, 2002.

Flexer A., Pampalk E., Widmer G., “Novelty detection based on spectral similarity of songs,” *in* Proc. 6th International Conference on Music Information Retrieval, London, 2005, pp. 252-259.

Foote J. T., “An Overview of Audio Information Retrieval,” *In ACM-Springer Multimedia Systems*, vol. 7, no. 1, pp. 2-11, 1999.

Foote J., “Automatic audio segmentation using a measure of audio novelty,” *in Proc.* International Conference on Multimedia and Expo, New York, 2000, pp. 452-455.

Ganchev T., Tasoulis D.K., Vrahatis M.N., Fakotakis N., “Locally Recur-rent Probabilistic Neural Network for Text-Independent Speaker Verification”, In *Proceedings of the Eurospeech’03*, Geneva, Switzerland, vol. 3, pp. 1673–1676, September 1-4, 2003.

Ganchev T., Tasoulis D.K., Vrahatis M.N., Fakotakis N., “Generalized Locally Recurrent Probabilistic Neural Networks for Text-Independent Speaker Verification”, In *Proceedings of the ICASSP’04*, Montreal, Quebec, Canada, vol. 1, pp. 41–44, May 17-21, 2004.

Gaston K., O'Neill M.A., “Automated species identification -- why not?,” *In Philosophical transactions-Royal Society of London, Biological sciences*, vol. 359, no. 1444, pp. 655–667, 2004.

Gouyon, F., Dixon, S., Pampalk, E., Widmer, G., “Evaluating rhythmic descrip-tors for musical genre classification”, In *Proceedings of the AES 25th Interna-tional Conference*, London, United Kingdom, June 17-19, 2004.

Hansen L.P., “Large Sample Properties of Generalized Method of Moments Estimation”, In *Econometrica*, vol. 50, pp. 1029–1054, 1982.

Haritaoglou I., Harwood D., Davis L. S., “W4: Real-time surveillance of people and their activities,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809-830, August 2000.

Harma A., McKinney M.F., Skowronek J., “Automatic surveillance of the acoustic activity in our living environment,” *in IEEE International Conference on Multimedia and Expo*, Amsterdam, Holland, July 2005.

Hyoung-Gook K., Nicolas M., Thomas S., MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval, Wiley, 2005.

ITU Standard, “Methods for subjective determination of transmission quality”, Tech. Rep. ITU-T Recommendation P.800, ITU, Switzerland, 1996.

Iwata, K., Moriwaki, T., “An application of acoustic emission measurements to in process sensing of tool wear”, In *Annals of the CIRP*, vol. 25, no. 1, pp. 21–26, 1977.

Joder C., Essid S., Richard G., “Temporal integration for audio classification with application to musical instrument classification,” *In IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 174-186, 2009.

Jordan M.I., “Serial order: A parallel distributed processing approach”. *Insti-tute for Cognitive Science, Report 8604*, University of California, San Diego, 1986.

Kashif Saeed Khan M., Al-Khatib W. G., Moinuddin M., "Automatic classification of speech and music using neural networks", *In Proceedings of the 2nd ACM international workshop on Multimedia databases*, pp. 94-99, 2004.

Khan M. K. S., Al-Khatib W. G., Moinuddin M., “Automatic classification of speech and music using neural networks,” *In Proc. of the 2nd International Workshop on Multimedia databases*, Washington, DC, USA, 13 November, 2004.

Khan S.M., Shah M., “A Multiview Approach to Tracking People in Crowded Scenes Using a Planar Homography Constraint”, *ECCV 2006*, Graz, Austria, pp. 133-146, 2006.

Kim H. G., Sikora T., “Comparison of MPEG-7 audio spectrum projection features and MFCC applied to speaker recognition, sound classification and audio segmentation,” *In ICASSP 2004*, Canada, Montreal, 17-21 May, 2004.

Kim H.-G., Moreau N., Sikora T., *MPEG-7 Audio and Beyond: audio content indexing and retrieval*. Wiley Publishers, October 2005.

Kipp M., “Anvil - A Generic Annotation Tool for Multimodal Dialogue”, *Eurospeech '01*, pp. 1367-1370, 2001.

Klapuri, A., Davy, M., (Editors), “Signal Processing Methods for Music Tran-scription”, Springer-Verlag, New York, 2006.

Kohonen T., “Learning Vector Quantization for Pattern Recognition”, *Techni-cal Report TKK-F- A601*, Helsinki University of Technology, 1986.

Lang K.J., Hinton, G.E., “A time delay neural network architecture for speech recognition”, *Technical Report CMU-cs-88-152*, Carnegie Mellon University, Pittsburgh PA, 1988.

Lee C.-H., Chou C.-H., Han C.-C., Huang R.-Z., “Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis,” *In Pattern Recognition Letters*, vol. 27, pp. 93–101, 2006.

Linares G., Nocera P., Meloni H., “Mixed acoustic events classification using ICA and subspace classifier,” *in Proc. International Conference on Acoustics, Speech and Signal Processing*, Munich, 1997, pp. 3365-3368.

Liu, M., Wan, C., “Feature selection for automatic classification of musical instrument sounds”. In Proceedings of the 1st ACM/IEEE-CS Joint conference on Digital libraries, pp. 247–248, 2001.

Livshin, A.A., Rodet, X., “Musical Instrument Identification in Continuous Recordings”, In *Proceedings of the DAFX’04*, Naples, Italy, October 5-8, 2004.

Lo D., Goubran R.A., Dansereau R.M., “Multimodal talker localization in video conferencing environments”, *HAVE 2004*, pp. 195-200, 2004.

Maleh K. E., Samouelian A., Kabal P., “Frame level noise classification in mobile environments,” *In ICASSP 1999*, Phoenix, Arizona, 15-19 March, 1999.

Mariano V.Y., Min J., Park J.-H., Kasturi R., Mihalcik D., Doermann D., Drayer T., “Performance Evaluation of Object Detection Algorithms”, *ICPR’2002*, pp. 965-969, 2002.

Markou M., Singh S., “Novelty detection: a review,” *Signal Processing*, Elsevier, vol. 83, no. 12, pp. 2481-2497, Dec. 2003.

Martin, A., Doddington, G., Kamm T., Ordowski M., Przybocki M., “The DET curve in assessment of detection task performance,” *in Eurospeech*, Rhodos, Greece, September 1997.

Mc Kinney M. F., Breebart J., “Features for audio and music classification,” *In Proceedings of* International Symposium on Music Information Retrieval, pp 151–158, 2003.

McLachlan G. J., Basford K. E., Mixture Models: Inference and applications to clustering, New York: Marcel Dekker Inc., 1988.

Meng A., Ahrendt P., Larsen J., Hansen L. K., “Temporal feature integration for music genre classification,” *In IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1654-1664, 2007.

Mertins A., Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and *Applications*, Wiley Publishers, 1999.

Mitrovic, D., Zeppelzauer, M., “Discrimination and Retrieval of Animal Sounds”, In *Proceedings of the IEEE Multimedia Modelling Conference*, Bei-jing, China, pp. 339–343, 2006.

Neto J., Almeida L., Hochberg M., Martins C., Nunes L., Renals S., Robinson T., “Speaker adaptation for hybrid HMM/ANN continuous speech recog-nition system”, In *Proceedings of the Eurospeech’95*, pp. 2171–2174, 1995.

Park S., Kautz H., “A hierarchical recognition of activities in daily living using multi-Scale, multi-perspective vision and RFID,” *in IET International Conference on Intelligent Environments*, Seattle, USA, 2008.

Peeters, G., “Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization”, In *Proceedings of the AES 115th convention*, New York, USA, October 10-13, 2003.

Peeters, G., Rodet, X., “Automatically selecting signal descriptors for sound classification”. In *Proceedings of the ICMC’02*, Goteborg, Sweden, September 2002.

Pinquier J., Rouas J.-L., Andre-Obrecht R., "A fusion study in speech/music classification", *In* Proceedings of International Conference in Multimedia and Expo, vol. 2, pp. 409-412, 2003.

Potamitis I., Ganchev T., “Generalized Recognition of Sound Events: Approaches and Applications” (invited chapter), *Multimedia Services in Intelligent Environments*, Springer Verlag, 2008.

Powell M.J.D., “Radial Basis Functions for multivariable interpolation: A review”, In Mason, J., Cox, M. (eds.), *Algorithms for approximation*, Oxford, Clarendon Press, pp. 143–167, 1987.

Quackenbush S., Lindsay A., “Overview of MPEG-7 Audio,” *IEEE Transactions on Circuits andSystems for Video Technology*, vol. 11, no.6, pp. 725-729, June 2001.

Quan X., Zhang H., “Perceptual criterion fragile audio watermarking using adaptive wavelet packets,” *In ICPR 2004*, Cambridge, United Kingdom, 23-26 August, 2004.

Rabiner, L. R., “A tutorial on hidden Markov models and selected applications in speech recognition,” *In Proceedings of the IEEE*, vol. 77, mo. 2, pp. 257-286, 1989.

Radhakrishnan R., Divakaran A., “Systematic acquisition of audio classes for elevator surveillance,” *SPIE Image and Video Communications Processing*, vol. 5685, pp. 64-71, March 2005.

Ren Y., Johnson M. T., Tao J., “Perceptually motivated wavelet packet transform for bioacoustic signal enhancement,” *In Journal of Acoustic Society of America*, vol. 124, no. 1, pp. 316-327, 2008.

Reynolds D. A., Quatieri T. F., Dunn R. B., “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, no.1, pp. 19-41, Jan. 2000.

Richard G., Ramona M., Essid S., “Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio signals,” *in Proc. International Conference on Acoustics, Speech and Signal Processing*, Honolulu, 2007, pp. 461-464, vol. 2.

Rosenblatt F., “The perceptron: a probabilistic model for information storage and organization in the brain”, In *Psychological Review*, vol. 65, pp. 386–408, 1958.

Rouas J.-L., Louradour J., Ambellouis S., “Audio events detection in public transport vehicles,” in IEEE Intelligent Transportation System Conference, Toronto, Canada, September 2006.

Scharf B., Critical Bands. In Foundations of Modern Auditory Theory, J. V. Tobias, Ed. New York: Academic, vol. 1, pp. 157-202, 1970.

Scheirer E., Slaney M., "Construction and evaluation of a robust multifeature speech/music discriminator", *In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol.2, pp. 1331-1334, 1997.

Schneider T., Neumaier A., “ARFIT: A Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models,” *In ACM Transactions on Mathematical Software*, vol. 27, no. 1, pp. 58-65, 2001.

Senac C., Ambikairajh E., "Audio classification for radio broadcast indexing: feature normalization and multiple classifiers decision", *Advances in Multimedia Information Processing - PCM 2004*, Springer Berlin/Heidelberg, 2004.

Setlur A.R., Sukkar R.A., Jacob J., “Correcting recognition errors via dis-criminative utterance verification”, In *Proceedings of ICSLP'96*, Philadelphia, USA, vol. 2, pp. 602–605, 1996.

Siafarikas M., Mporas I., Ganchev T., Fakotakis N., "Speech Recognition Using Wavelet Packet Features", *Wavelet Theory and Applications*, vol. 2, no. 1, pp. 41-49, 2009, 2008.

Singh S., Markou M., “An approach to novelty detection applied to the classification of image regions,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp.396-407, Apr. 2004.

Skodras A., Christopoulos C., Ebrahimi T., "The JPEG 2000 still image compression standard", *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36-58, Sep. 2001.

Slaney M. “Auditory Toolbox. Version 2”, *Technical Report #1998-010*, Interval Research Corporation, 1998.

Slaney M., “Mixtures of probability experts for audio retrieval and indexing”, In *Proceedings of the IEEE International Conference on Multimedia and Expo*, Lausanne, Switzerland, vol. 1, pp. 345–348, August 2002.

Sohn J., Kim N. S., Sung W., "A statistical model-based voice activity detection", *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, January 1999.

Specht D.F., “Generation of polynomial discriminant functions for pattern recognition”. In *IEEE Transactions on Electronic Computers*, vol. 16, pp. 308–319, 1967.

Specht D.F., “Probabilistic Neural Networks for classification, mapping, or associative memory”, In *Proceedings of the IEEE Conference on Neural Net-works*, San Diego, vol. 1, pp. 525–532, July 1988.

Tarassenko L., “Novelty detection for the identification of masses in mammograms,” *in Proc. 4th* IEEE International Conference on Artificial Neural Networks, Cambridge, 1995, vol. 4, pp. 442-447.

Tax D. M. J., Duin R.P.W., “Outlier detection using classifier instability,” *in Proc. Advances in Pattern Recognition*, the Joint IAPR International Workshops, Sydney, 1998, pp. 593-601.

Thirde D., Li L., Ferryman J., “An Overview of the PETS 2006 Dataset”, *PETS’ 2005*, pp. 317- 324, 2005.

Torch Machine Learning Library, διαθέσιμο στη διεύθυνση http://www.torch.ch.

Torrence C.., Compo G. P., “A practical guide to wavelet analysis,” *In Bulletin of the American Meteorological society*, vol. 79, no. 1, pp. 61-78, 1998.

Tzanetakis G., Cook P., “Musical Genre Classification of Audio Signals,” *In IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, 2002.

Tzanetakis G., Essl G., Cook P. R., "Audio analysis using the wavelet transform", In Proceedings of International Conference on Acoustics and Music: Theory and Applications (AMTA), 2001.

UK Home Office, “Multiple-camera tracking scenario (MCTS)”, October 2008. Available: http://scienceandresearch.homeoffice.gov.uk/hosdb/publications/cctvpublications /MCTS_Sce nario_Definition_Ma1.-pdf? view=Binary

Umapathy K., Krishnan S., Rao R. K., “Audio signal feature extraction and classification using local discriminant bases,” *In IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1236-1246, 2007.

Vacher M., Istrate D., Besacier L., Serignat J.-F., Castelli E., “Sound detection and classification for medical Telesurvey,” *in International Conference of Biomedical Engineering*, Innsburg, Austria, February 2004.

Valenzise G., Gerosa L., Tagliasacchi M., Antonacci F., Sarti A., "Scream and gunshot detection and localization for audio-surveillance systems," *in Proceedings of Advanced Video and Signal-based Surveillance,* London, England, September 2007.

Vapnik V.N., "The Nature of Statistical Learning Theory", Springer, 1995.

Valenzise G., Gerosa L., Tagliasacchi M., Antonacci F., Sarti A., "Scream and gunshot detection and localization for audio-surveillance systems," *in Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, London, 2007, pp. 21–26.

Vemuri, S., Schmandt, C., Bender, W., Tellex, S., Lassey, B., "An Audio-Based Personal Memory Aid", In *Proceedings of the 6th International Conference Ubiquitous Computing, Ubicomp'04*, pp. 400–417, 2004.

Verma B., Blumenstein M., Pattern Recognition Technologies and Applications: Recent *Advances,* Information Science Reference Publishers, 2008.

Wallhoff F., Ru? M., Rigoll G., Gobel J., Diehl H., "Diehl Improved Image Segmentation Using Photonic Mixer Devices", *ICIP 2007*, vol. 6, pp. 53–56, 2007.

Wang D., Brown G. J., Computational Auditory Scene Analysis: Principles, Algorithms and *Applications,* Wiley Blackwell Publishers, 2006.

Wang J. F., Chen S. H., "A voice activity detection algorithm based on perceptual wavelet packet transform and Teager energy operator," *In International Symposium on Chinese Spoken Language Processing (ICSLP 2002),* Taipei, Taiwan, 23-24 August, 2002.

Wang J.-C., Wang J.-F., Kuok W.H., Hsu C.S., "Environmental Sound Classification Using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor," *International Joint Conference on Neural Networks,* 2006.

Watson A.T., O'Neill M. A., Kitching I. J., "A qualitative study investigating automated identification of living macrolepidoptera using the Digital Automated Identification System (DAISY)," *In Systematics and Biodiversity,* vol. 1, no. 3, pp. 287-300, 2003.

Wei F., Miller M., Stolfo S.J., Wenke L., Chan P.K., "Using artificial anomalies to detect unknown and known network intrusions," *in Proc. IEEE International Conference on Data Mining,* San Jose, 2001, pp. 123–130.

Widmer, G. (Ed.), "Special Issue on Machine Learning in Music", In *Machine Learning,* vol. 65, no. 2-3, December 2006.

Wilpon J. G., Rabiner L. R., Lee C.-H., Goldman E. R., "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing,* pp. 1870-1878, November 1990.

Witten I. H., Frank E., Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco, 2005.

Wold E., Blum T., Keislar D., Wheaton J., “Content-based classification, search and retrieval of audio,” *In IEEE Multimedia*, vol. 3, no. 3, pp. 27-36, 1996.

Yost W. A., *Fundamentals of Hearing*, 3rd Edition, New York Academic,” pp: 153-167, 1994. Zhang T., Kuo C.-C. J., “Content-based classification and retrieval of audio,” *in SPIE’s 43rd* Annual Meeting - Conference on Advanced Signal Processing Algorithms, Architectures, and *Implementations VIII*, San Diego, USA, July 22-24, 1998.

Zhoun G., Hansen J. H. L., Kaiser J.F., “Nonlinear feature based classification of speech under stress,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 201-216, March 2001.