Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

Πολυ-Αισθητηριακή Ακουστική Ανάλυση Περιβάλλοντος

Διδακτορική Διατριβή

του

ΠΑΝΑΓΙΩΤΗ ΓΙΑΝΝΟΥΛΗ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π.

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Πολυ-Αισθητηριακή Ακουστική Ανάλυση Περιβάλλοντος

Διδακτορική Διατριβή

του

ΠΑΝΑΓΙΩΤΗ ΓΙΑΝΝΟΥΛΗ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π.

Συμβουλευτική Επιτροπή: Καθ. Πέτρος Μαραγκός (Επιβλέπων), ΣΗΜΜΥ, ΕΜΠ
Αναπλ. Καθ. Γεράσιμος Ποταμιάνος, ΤΗΜΜΥ, Παν. Θεσσαλίας
Αναπλ. Καθ. Κωνσταντίνος Τζαφέστας, ΣΗΜΜΥ, ΕΜΠ

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 26η Νοεμβρίου 2021.

.....	.....	.....
Πέτρος Μαραγκός	Γεράσιμος Ποταμιάνος	Κωνσταντίνος Τζαφέστας
Καθηγητής	Αναπληρωτής Καθηγητής	Αναπληρωτής Καθηγητής
Ε.Μ.Π.	Παν/μίου Θεσσαλίας	Ε.Μ.Π.
.....	.....	.....
Στέφανος Κόλλιας	Παναγιώτης Τσανάκας	Αθανάσιος Κατσαμάνης
Καθηγητής	Καθηγητής	Ερευνητής Β΄
Ε.Μ.Π.	Ε.Μ.Π.	Ε.Κ. Αθηνά
.....		
Ευΐτα-Σταυρούλα Φωτεινέα		
Ερευνήτρια Α΄		
Ε.Κ. Αθηνά		

Αθήνα, Νοέμβριος 2021

(Υπογραφή)

.....
Παναγιώτης Γιαννούλης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Στη Διατριβή μας εξετάζουμε το πρόβλημα του εντοπισμού ακουστικών γεγονότων σε «έξυπνα» περιβάλλοντα με πολλαπλά μικρόφωνα. Ο εντοπισμός ακουστικών γεγονότων αποτελεί σημαντικό τμήμα του ευρύτερου πεδίου της υπολογιστικής ανάλυσης ακουστικής σκηνής, και στόχος του είναι ο αυτόματος εντοπισμός στον χρόνο και η αναγνώριση των ακουστικών γεγονότων που περιέχονται σε ένα ηχητικό στιγμιότυπο. Στην έρευνά μας εστιάζουμε στην ανάπτυξη μεθόδων για την αξιοποίηση της πληροφορίας από πολλαπλά μικρόφωνα για τον εντοπισμό γεγονότων σε απαιτητικές συνθήκες με φαινόμενα επικάλυψης. Αρχικά, δίνουμε έμφαση στο πρόβλημα του εντοπισμού της ανθρώπινης φωνής, και στα πλαίσια ενός «έξυπνου» οικιακού περιβάλλοντος με πολλαπλά δωμάτια αναπτύσσουμε ένα σύστημα χωρο-χρονικού εντοπισμού φωνής δύο σταδίων, κατάλληλο για διαλογικά συστήματα φωνητικών εντολών. Στο πρώτο στάδιο, το σύστημά μας συνδυάζει αποτελεσματικά τα σήματα από πολλαπλά μικρόφωνα για να πετύχει τον χρονικό εντοπισμό της φωνής, και στο δεύτερο, καινοτόμα πολυκαναλικά χαρακτηριστικά εξάγονται για τον χωρικό εντοπισμό της φωνής σε επίπεδο δωματίου. Το σύστημά μας επιδεικνύει εύρωστη απόδοση και συγκρίνεται ευνοϊκά με μεθόδους βαθιάς μηχανικής μάθησης. Στη συνέχεια, στο ευρύτερο πρόβλημα του εντοπισμού ακουστικών γεγονότων, δίνουμε έμφαση στο απαιτητικό σενάριο των επικαλυπτόμενων γεγονότων και πειραματιζόμαστε με μεθόδους παραγοντοποίησης μη-αρνητικών πινάκων (NMF). Στα πλαίσια αυτής της έρευνας, διερευνούμε μεθόδους για την βελτίωση του σταδίου εντοπισμού σε βασικές μεθόδους NMF, την αύξηση της αποδοτικότητας σε δύσκολες επικαλυπτόμενες συνθήκες συστημάτων NMF που συνδυάζονται με ταξινομητές, και τέλος την ανάπτυξη αποτελεσματικών πολυ-καναλικών συστημάτων NMF για προβλήματα εντοπισμού γεγονότων. Τέλος, πειραματιζόμαστε με μεθόδους βαθιάς μηχανικής μάθησης για τον εντοπισμό επικαλυπτόμενων γεγονότων σε περιπτώσεις όπου υπάρχει μεγάλη ποικιλία πιθανών κλάσεων. Σε αυτή την κατεύθυνση, προτείνουμε τον συνδυασμό και την από κοινού εκπαίδευση ενός πολυ-καναλικού νευρωνικού δικτύου διαχωρισμού γεγονότων με ένα νευρωνικό δίκτυο ταξινόμησης ακουστικών γεγονότων, πετυχαίνοντας βελτιωμένη απόδοση σε σχέση με παραδοσιακές τεχνικές. Για την αξιολόγηση των μεθόδων μας, χρησιμοποιούμε διάφορες συνθετικές και πραγματικές βάσεις δεδομένων που δημιουργήθηκαν/ηχογραφήθηκαν σε κατάλληλα πολυ-καναλικά «έξυπνα» περιβάλλοντα.

Λέξεις Κλειδιά: Εντοπισμός ακουστικών γεγονότων, Πολυ-καναλικές μέθοδοι, Επικαλυπτόμενα γεγονότα, «Εξυπνα» περιβάλλοντα

Abstract

In our Dissertation we examine the problem of Acoustic Event Detection (AED) in multi-channel smart-space environments. AED constitutes a major part of the computational auditory analysis field, and its main goal is the automatic end-pointing and classification of each sound event present in an audio clip. In our research we focus on developing methods for exploiting the information from multiple microphones for detecting events under challenging and overlapping conditions. At first, we focus on the detection of human speech events in smart homes consisting of multiple rooms, equipped with multiple microphones. For this purpose, we develop a novel two-step room-localized Speech Activity Detection (SAD) system, appropriate for voice-enabled applications. In its first step, our system efficiently combines the signals from multiple microphones to produce temporal speech segmentation, and in the second step it extracts novel room-discriminant multi-channel features to locate the speaker at the room-level. Our system performs robustly and compares favorably to deep-learning based alternatives. Then, for the general AED task, we focus on the challenging overlapping scenario and experiment with non-negative matrix factorization (NMF)-based approaches. In the process, we investigate ways to improve the detection step of well-known NMF baselines, to increase the robustness of classifier-based NMF systems in highly overlapping conditions, and finally to develop an efficient multi-channel NMF system for detection tasks. Finally, we employ deep-learning methods for overlapping AED when the number of different event classes is large. In this direction, we propose the combination and joint training of a multi-channel sound source separation network with a multi-label AED network, achieving improved results over traditional neural network approaches. For our experiments, we employ several synthetic and real databases recorded in suitable multi-microphone smart-space environments.

Index Terms: Acoustic event detection, Multi-channel, Overlapping events, Smart homes

Πρόλογος

Με την ολοκλήρωση αυτής της διατριβής κλείνει ένας μεγάλος κύκλος, κατά τη διάρκεια του οποίου νιώθω ότι βγαίνω ιδιαίτερα κερδισμένος σε γνώσεις και εμπειρίες. Θα ήθελα να ευχαριστήσω όλους όσους με στήριξαν και συνεργάστηκαν μαζί μου σε αυτή την μακρόχρονη διαδρομή.

Αρχικά θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Πέτρο Μαραγκό για την συνεργασία που είχαμε όλα αυτά τα χρόνια και για την καθοδήγηση που μου παρείχε. Ήταν εκείνος που μέσα από την διδασκαλία του στα προπτυχιακά μαθήματα της σχολής με ενέπνευσε να ακολουθήσω τις ερευνητικές περιοχές της επεξεργασίας σημάτων και της μηχανικής μάθησης. Κατά την διάρκεια της διατριβής, μέσα από την συνεργασία μας σε ερευνητικά προγράμματα, τις ερευνητικές συζητήσεις που είχαμε, αλλά και μέσα από την διδασκαλία μεταπτυχιακών μαθημάτων, με διαμόρφωσε καθοριστικά σαν ερευνητή και επιστήμονα. Επίσης θα ήθελα να ευχαριστήσω τον συνεπιβλέποντα καθηγητή κ. Γεράσιμο Ποταμιάνο για τη μακρόχρονη συνεργασία που είχαμε. Η συνεισφορά του ήταν πολύτιμη τόσο στη διεξαγωγή της έρευνας όσο και στη συγγραφή των διάφορων ερευνητικών άρθρων που προέκυψαν κατά τη διάρκεια της Διδακτορικής διατριβής. Ευχαριστώ επίσης τους κ. Κώστα Τζαφέστα, κ. Αθανάσιο Κατσαμάνη, κ. Στέφανο Κόλλια, κ. Ευΐτα Φωτεινέα, και κ. Παναγιώτη Τσανάκα που δέχθηκαν να είναι μέλη της επταμελούς μου επιτροπής.

Θα ήθελα ακόμα να ευχαριστήσω όλα τα μέλη του εργαστηρίου Ρομποτικής και Αυτοματισμού για το ευχάριστο και δημιουργικό κλίμα και τις κοινές εμπειρίες που είχαμε όλα αυτά τα χρόνια. Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου για την συνεχή στήριξή της από την αρχή μέχρι το τέλος αυτής της διαδρομής.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κατάλογος Συντομογραφιών

AED ..... Acoustic Event Detection - Εντοπισμός Ακουστικών Γεγονότων

ASR ..... Automatic Speech Recognition - Αυτόματη Αναγνώριση Φωνής

CNNs ..... Convolutional Neural Networks - Συνελικτικά Νευρωνικά Δίκτυα

DNNs ..... Deep Neural Networks - Βαθιά Νευρωνικά Δίκτυα

FBEs ..... FilterBank Energies - Ενέργειες Συστοιχίας Φίλτρων

GCC ..... Generalized Cross Correlation - Γενικευμένη Ετερο-Συσχέτιση

GMMs ..... Gaussian Mixture Models - Μοντέλα Μείγματος Γκαουσσιανών

HMMs ..... Hidden Markov Models - Κρυφά Μαρκοβιανά Μοντέλα

LTSV ..... Long-Term Signal Variability - Μακρόχρονη Μεταβλητότητα Σήματος

MFCCs ..... Mel-Frequency Cepstral Coefficients - Φασματικοί Συντελεστές Μελ Συχνοτήτων

MLD ..... Mixture of Local Dictionaries - Μείγμα από Τοπικά Λεξικά

NMF ..... Non-negative Matrix Factorization - Μη-Αρνητική Παραγοντοποίηση Πινάκων

RIR ..... Room Impulse Response - Κρουστική Απόκριση Δωματίου

SAD ..... Speech Activity Detection - Εντοπισμός Φωνής

SNR ..... Signal-to-Noise Ratio - Σηματοθορυβικός Λόγος

SRP ..... Steered-Response Power - Ισχύς Εστιασμένης Απόκρισης

SVMs ..... Support Vector Machines - Μηχανές Διανυσματικής Υποστήριξης

TDOA ..... Time Difference Of Arrival - Διαφορά Χρόνου Άφιξης

VQT ..... Variable Q-Transform - Μεταβλητός Μετασχηματισμός-Q

Εκτεταμένη Περίληψη

Εισαγωγή

Ο χωρο-χρονικός εντοπισμός ακουστικών γεγονότων (AED) αποτελεί ένα σημαντικό τμήμα του ευρύτερου πεδίου της υπολογιστικής ανάλυσης ακουστικών σημάτων, το οποίο έχει προσελκύσει ιδιαίτερο ενδιαφέρον στην ερευνητική κοινότητα τα τελευταία χρόνια. Τυπικές εφαρμογές του πεδίου αυτού περιλαμβάνουν συστήματα για «έξυπνα» σπίτια [1–5], για παρακολούθηση της υγειονομικής περίθαλψης [6], για εύρεση και ανάκτηση πολυμέσων [7], καθώς και συστήματα ασφαλείας και παρακολούθησης χώρων [8, 9].

Ο κύριος σκοπός του εντοπισμού ακουστικών γεγονότων είναι ο αυτόματος χρονικός εντοπισμός και η κατηγοριοποίηση των διάφορων ακουστικών γεγονότων μέσα σε ένα ηχητικό απόσπασμα, αποκαλύπτοντας έτσι σημαντική πληροφορία για ανθρώπινες και μη δραστηριότητες στο περιβάλλον. Ανάλογα με το συγκεκριμένο πρόβλημα και το αντίστοιχο ακουστικό περιβάλλον, μπορεί να προκύψει ένας μεγάλος αριθμός από πιθανά ακουστικά γεγονότα [10,11]. Στην συγκεκριμένη διατριβή, δόθηκε έμφαση σε ακουστικά γεγονότα που συνήθως πραγματοποιούνται σε «έξυπνα» περιβάλλοντα («έξυπνα» σπίτια ή χώρους εργασίας). Μερικά παραδείγματα τέτοιων ακουστικών συμβάντων είναι η «φωνή», το «περπάτημα», η «ραδιοφωνική μουσική», το «χτύπημα της πόρτας», η «πληκτρολόγηση» κλπ. Ένα ακουστικό γεγονός με ιδιαίτερο ενδιαφέρον είναι η ανθρώπινη «φωνή», αφού αποτελεί το κύριο μέσο για επικοινωνία ανθρώπου-μηχανής και επομένως διαδραματίζει σημαντικό ρόλο σε πολλές εφαρμογές.

Ιδιαίτερη έμφαση δόθηκε στον εντοπισμό της φωνής (SAD) σαν μια ειδική περίπτωση του εντοπισμού ακουστικών γεγονότων. Συγκεκριμένα μελετήθηκε ο εντοπισμός της φωνής στα πλαίσια συστημάτων φωνητικών εντολών μέσα σε «έξυπνα» περιβάλλοντα. Τέτοιου είδους συστήματα συνήθως περιλαμβάνουν μια ακολουθία από υπο-συστήματα, με αυτό του εντοπισμού φωνής να είναι από τα σημαντικότερα, αφού παρέχει την απαραίτητη είσοδο για πολλά άλλα υπο-συστήματα, όπως π.χ. τον χωρικό εντοπισμό ομιλητή, τον καθαρισμό/βελτίωση της φωνής, τον εντοπισμό λέξεων-κλειδιών για τις φωνητικές εντολές, και την αυτόματη αναγνώριση φωνής. Εκτός των συστημάτων φωνητικών εντολών, ο εντοπισμός φωνής βρίσκει εφαρμογή και σε άλλα προβλήματα/περιοχές, όπως στις τηλεπικοινωνίες, στην κωδικοποίηση φωνής, στη αναγνώριση ομιλητή μέσω ομιλίας, και άλλα.

Στο υπόλοιπο μέρος του κεφαλαίου, θα γίνει μια επισκόπηση των σχετικών εργασιών στα πεδία του εντοπισμού φωνής και των ακουστικών γεγονότων. Στη συνέχεια θα γίνει μια περιγραφή των κύριων συνεισφορών της ερευνητικής μας εργασίας, και θα δοθεί συνοπτικά η

δομή της Διατριβής.

Σχετικές Εργασίες

Πληθώρα μεθόδων έχουν προταθεί τα τελευταία χρόνια στη βιβλιογραφία για το πρόβλημα του εντοπισμού ακουστικών γεγονότων, με μεγάλη ποικιλία στην επιλογή των αλγορίθμων και των ακουστικών χαρακτηριστικών που χρησιμοποιούνται. Μπορούμε να διαχωρίσουμε τις σχετικές εργασίες με βάση το αν αναπτύχθηκαν και αξιολογήθηκαν για τον εντοπισμό μεμονωμένων ή επικαλυπτόμενων ακουστικών γεγονότων, όπως επίσης και με βάση το αν στηρίζονται στην ύπαρξη ενός ή πολλαπλών μικροφώνων για την καταγραφή του ακουστικού περιβάλλοντος. Στην περίπτωση του προβλήματος εντοπισμού μεμονωμένων γεγονότων (isolated AED), εργασίες που χρησιμοποιούν κλασσικές μεθόδους εντοπισμού και ταξινόμησης, όπως π.χ. κρυφά Μαρκοβιανά μοντέλα (HMMs), σε συνδυασμό με παραδοσιακά ακουστικά χαρακτηριστικά (π.χ. MFCCs), έχουν δείξει ότι μπορούν να πετύχουν ικανοποιητική απόδοση [2].

Στην περίπτωση ωστόσο του πιο απαιτητικού προβλήματος εντοπισμού επικαλυπτόμενων γεγονότων (overlapping AED), χρειάζεται να χρησιμοποιηθούν μέθοδοι οι οποίες να επιτρέπουν τον εντοπισμό πολλαπλών γεγονότων. Για παράδειγμα, στην εργασία [12], για την επίλυση του προβλήματος της επικάλυψης, προτάθηκε μια μέθοδος που κάνει χρήση του αλγόριθμου αποκωδικοποίησης Viterbi σε πολλαπλά μονοπάτια. Άλλες εργασίες για τον εντοπισμό επικαλυπτόμενων γεγονότων περιλαμβάνουν βαθιά νευρωνικά δίκτυα (DNNs) [13], μοντέλα ανάλυσης πιθανοτικών συνιστωσών [14], συστήματα βασισμένα στον γενικευμένο μετασχηματισμό Hough [15], καθώς και στη μη-αρνητική παραγοντοποίηση πινάκων (NMF) [16].

Μεταξύ των διαφόρων μεθόδων που έχουν αναπτυχθεί για τον εντοπισμό ακουστικών γεγονότων, οι προσεγγίσεις που βασίζονται στην τεχνική NMF έχουν προσελκύσει αρκετό ενδιαφέρον, και ιδιαίτερα στα σενάρια με επικάλυψη. Αυτό οφείλεται στην γενικότερη ευρωστία τους αλλά και στην δυνατότητά τους να επιτρέπουν τον εντοπισμό πολλαπλών γεγονότων που συμβαίνουν ταυτόχρονα, δεδομένου ότι κατάλληλες μη-αρνητικές και γραμμικές αναπαραστάσεις τους είναι διαθέσιμες. Οι NMF μέθοδοι μπορούν να διακριθούν σε αυτές που χρησιμοποιούν άμεσα τις τιμές του πίνακα ενεργοποιήσεων (activation matrix) για να επιτύχουν τον εντοπισμό των γεγονότων [16, 17], και σε αυτές που χρησιμοποιούν έναν ταξινομητή ο οποίος εκπαιδεύεται με βάση τον πίνακα ενεργοποίησης [18, 19]. Στην εργασία [16], μετά την κατασκευή ενός αρκετά μεγάλου λεξικού NMF, οι ενεργοποιήσεις χρησιμοποιούνται άμεσα για τον εντοπισμό του κάθε ακουστικής κλάσης. Σχετικά με τις μεθόδους που χρησιμοποιούν επιπλέον ταξινομητές, στην εργασία [18], ένα αρκετά μικρού μεγέθους λεξικό κατασκευάζεται αυτόματα με χρήση της μεθόδου αραιής συνελικτικής μη-αρνητικής παραγοντοποίησης πινάκων (CNMF), και στη συνέχεια οι παραγόμενες ενεργοποιήσεις χρησιμοποιούνται ως είσοδος για την εκπαίδευση ενός (HMM) για τις διάφορες ακουστικές κλάσεις. Επίσης στην εργασία [20], οι συγγραφείς, εμπνευσμένοι από το γεγονός ότι οι μέθοδοι NMF μπορούν να επωφεληθούν από την δημιουργία ενός μείγματος από τοπικά λεξικά (Mixture of Local Dictionaries (MLD)) [21], προτείνουν ένα συνδυασμό ταξινομητή και συστήματος NMF, χρησιμοποιώντας MLD για να πετύχουν βελτιωμένη απόδοση στον εντοπισμό γεγονότων.

Παρόλο που οι μέθοδοι NMF αποτελούν μια φυσική επιλογή για τα σενάρια επικαλυπτόμενων γεγονότων, έχουν επίσης κάποια μειονεκτήματα, που κυρίως αφορούν στην μειωμένη υπολογιστική τους αποδοτικότητα (ταχύτητα υπολογισμού λύσης), αλλά και στην ικανότητα διάκρισης σχετικά με την ταξινόμηση (ειδικά σε περιπτώσεις με μεγάλο αριθμό ακουστικών γεγονότων). Σε περίπτωση που αρκετά δεδομένα είναι διαθέσιμα για εκπαίδευση, μέθοδοι βασισμένες στην βαθιά μηχανική μάθηση μπορούν να πετύχουν καλύτερη απόδοση σε προβλήματα εντοπισμού γεγονότων [22], λόγω της καλύτερης ικανότητας διάκρισης που παρέχουν. Γενικά, διάφορες μέθοδοι βασισμένες σε βαθιά νευρωνικά δίκτυα έχουν προταθεί με ιδιαίτερη επιτυχία τα τελευταία χρόνια, περιλαμβάνοντας DNNs [23], συνελικτικά νευρωνικά δίκτυα (CNNs) [24], συνελικτικά αναδρομικά νευρωνικά δίκτυα [22], και transformers [25].

Όλες οι μέθοδοι που αναφέρθηκαν παραπάνω έχουν κυρίως εφαρμοστεί στην περίπτωση του μονο-καναλικού (καταγραφή από ένα μικρόφωνο) εντοπισμού γεγονότων. Ωστόσο, όταν η δεδομένη εγκατάσταση το επιτρέπει, η εκμετάλλευση πληροφορίας από πολλαπλά μικρόφωνα μπορεί να φανεί πολύτιμη. Στην εργασία [2], διάφορες τεχνικές για συνδυασμό πολλαπλών μικροφώνων προτάθηκαν στα πλαίσια ενός συστήματος βασισμένου σε HMMs, ενώ στην εργασία [26] χαρακτηριστικά βασισμένα σε σύνολα λέξεων (bag-of-words) από διαφορετικά μικρόφωνα χρησιμοποιήθηκαν για την εκπαίδευση ενός ταξινομητή τύπου random forest. Σχετικά με μεθόδους βασισμένες σε νευρωνικά δίκτυα, στην εργασία [27] εκμετάλλευση της πολυ-καναλικής πληροφορίας πραγματοποιήθηκε είτε τροφοδοτώντας το δίκτυο με εισόδους από πολλαπλά μικρόφωνα, είτε εξάγοντας πολυ-καναλικά χωρικά χαρακτηριστικά. Σε μεθόδους από την κατηγορία NMF, πολυ-καναλικές επεκτάσεις έχουν επίσης προταθεί, αλλά ήταν κυρίως στοχευμένες στο πρόβλημα του «τυφλού» διαχωρισμού πηγών (blind source separation) [28–32].

Παρόμοια με την περίπτωση του εντοπισμού ακουστικών γεγονότων, ο εντοπισμός φωνής (SAD) επίσης αποτελεί ένα πεδίο έντονης ερευνητικής δραστηριότητας, με μεγάλη ποικιλία αλγορίθμων να έχουν προταθεί στην βιβλιογραφία για περισσότερο από τέσσερις δεκαετίες, όπως φαίνεται και από την επισκόπηση της εργασίας [33]. Μερικές από τις πιο θεμελιωμένες μεθόδους περιλαμβάνουν αλγορίθμους ενσωματωμένους σε πρότυπα (standards) [34, 35], τον αλγόριθμο των Sohn et al. που βασίζεται σε στατιστικά μοντέλα [36], και τον αλγόριθμο των Ramirez et al. που βασίζεται στην ιδέα της φασματικής απόκλισης [37], μεταξύ άλλων. Τυπικά, οι μέθοδοι για εντοπισμό φωνής εξάγουν διάφορα χαρακτηριστικά από την κυματομορφή τα οποία είναι, για παράδειγμα, σχετικά με την ενέργεια ή τον ρυθμό μηδενικής-διέλευσης (zero-crossing rate) [34, 35, 38, 39], την αρμονικότητα και την τονικότητα της φωνής (harmonicity and pitch) [40–42], την δομή των κύριων συχνοτικών συνιστωσών (formants) [34, 43–45], τον βαθμό της στασιμότητας των σημάτων φωνής και θορύβου [46–48], την διαμόρφωση (modulation) [49–51], ή τα MFCCs [45]. Η εξαγωγή χαρακτηριστικών στη συνέχεια ακολουθείται από παραδοσιακή στατιστική μοντελοποίηση, ή πιο πρόσφατα, από ταξινομητές βαθιάς μηχανικής μάθησης, όπως για παράδειγμα DNNs [52, 53], επαναλαμβανόμενα νευρωνικά δίκτυα (RNNs) [54, 55], ή CNNs [56–58], συχνά σε συνδυασμό και με αυτοκωδικοποιητές (autoencoders) [59]. Επίσης, έχουν προταθεί με επιτυχία [60] «από άκρη-σε-άκρη» (end-to-end) μέθοδοι βαθιάς μηχανικής μάθησης οι οποίες ενεργούν απευθείας στο ηχητικό σήμα.

Συγκεκριμένα στον τομέα των «έξυπνων» σπιτιών, διάφορα συστήματα εντοπισμού φωνής έχουν αναπτυχθεί την τελευταία δεκαετία, έπειτα από την συλλογή βάσεων δεδομένων από αντίστοιχα οικιακά περιβάλλοντα [61–65]. Για παράδειγμα στην εργασία [66], φασματικοί συντελεστές γραμμικών συχνοτήτων (linear-frequencies cepstral coefficients) εξάγονται σαν χαρακτηριστικά και συνδυάζονται με ταξινομητές βασισμένους σε μοντέλα μείγματος Γκαουσσιανών (GMM) και HMM για να εντοπίσουν φωνή σε συνθήκες άγχους, ή ακουστικά γεγονότα μέσα σε ένα «έξυπνο» σπίτι για ηλικιωμένους ανθρώπους. Σε ένα παρόμοιο εγχείρημα στα πλαίσια του ερευνητικού έργου Sweet-Home [67], πραγματοποιείται αρχικά εντοπισμός ακουστικών γεγονότων με χρήση χαρακτηριστικών που εξάγονται από τον διακριτό μετασχηματισμό κυματιδίων (discrete wavelet transform) και μιας στρατηγικής προσαρμοστικού κατωφλίου (adaptive thresholding), και στη συνέχεια πραγματοποιείται κατηγοριοποίηση μεταξύ φωνής και άλλων γεγονότων με χρήση ταξινομητή της κατηγορίας μηχανών διανυσματικής υποστήριξης (SVM) σε συνδυασμό με GMM υπερ-διανύσματα (supervectors) βασισμένα σε MFCC χαρακτηριστικά. Στην εργασία [68], ένας απλός αλγόριθμος εντοπισμού φωνής βασισμένος στην ενέργεια του ηχητικού σήματος, προηγείται μιας μεθόδου αναγνώρισης φωνητικών εντολών βασισμένης σε HMM. Στην εργασία [69], ο εντοπισμός φωνής πραγματοποιείται κάνοντας χρήση ήχου που καταγράφεται από μικρόφωνο προσαρμοσμένο σε σετ ακουστικών με σκοπό την παρακολούθηση της ανθρώπινης δραστηριότητας μέσα σε ένα «έξυπνο» σπίτι, με το προτεινόμενο σύστημα να περιλαμβάνει ανιχνευτή ενέργειας και ένα νευρωνικό δίκτυο εκπαιδευμένο σε συντελεστές γραμμικής πρόβλεψης (linear predictive coding) και άλλα συχνοτικά χαρακτηριστικά.

Τα συστήματα που προαναφέρθηκαν έχουν ως στόχο τον εντοπισμό ανθρώπινης ομιλίας γενικά μέσα στα πλαίσια ενός «έξυπνου» περιβάλλοντος, χωρίς όμως να λαμβάνουν υπόψιν την τυπική δομή πολλαπλών δωματίων που συνήθως συναντάμε σε τέτοιους χώρους. Μερικές μόνο προσεγγίσεις στη βιβλιογραφία εξετάζουν το πρόβλημα του χωρο-χρονικού εντοπισμού φωνής σε ένα οικιακό περιβάλλον με πολλά δωμάτια, το οποίο αποτελεί και το κύριο αντικείμενο της έρευνάς μας για τον εντοπισμό φωνής. Σε αυτό το σενάριο, ζητούμενο είναι η χρονική κατάτμηση σε φωνή/σιωπή ξεχωριστά για κάθε ένα από τα δωμάτια ενός «έξυπνου» σπιτιού.

Η πλειοψηφία των συστημάτων χωρο-χρονικού εντοπισμού φωνής λειτουργούν σε δύο στάδια. Συνήθως, στο πρώτο στάδιο εντοπίζονται τα πιθανά χρονικά όρια των φωνητικών τμημάτων για όλο το σπίτι ή για κάθε δωμάτιο, τα οποία στο δεύτερο στάδιο επανεξετάζονται, διορθώνονται και αντιστοιχίζονται στα σωστά δωμάτια. Συγκεκριμένα, στην εργασία [70], στο πρώτο στάδιο του προτεινόμενου αλγόριθμου, για κάθε εντοπισμένο τμήμα φωνής, ο σηματοθορυβικός λόγος (SNR) και χαρακτηριστικά βασισμένα στην έννοια της συνάφειας σημάτων (coherence-based) εξάγονται για όλα τα δωμάτια και στη συνέχεια συνενώνονται σε ένα ενιαίο διάνυσμα που τροφοδοτεί έναν ταξινομητή γραμμικής διακριτικής ανάλυσης (LDA) για την αντιστοίχιση του φωνητικού τμήματος στο σωστό δωμάτιο. Στην εργασία [71], στο πρώτο στάδιο, ένα σύστημα εντοπισμού φωνής βασισμένο σε στατιστικά μοντέλα εφαρμόζεται σε κάθε μικρόφωνο, και στη συνέχεια μια πλειοψηφική στρατηγική (majority voting) εφαρμόζεται μεταξύ των διαθέσιμων μικροφώνων του κάθε δωματίου, για να παραχθούν τα χρονικά τμήματα φωνής σε κάθε δωμάτιο. Στο δεύτερο στάδιο, η έξοδος ενός υπο-συστήματος χωρικού εντοπισμού ομιλητή χρησιμοποιείται ως είσοδος σε έναν ταξινομητή (SVM ή νευρωνικό δίκτυο)

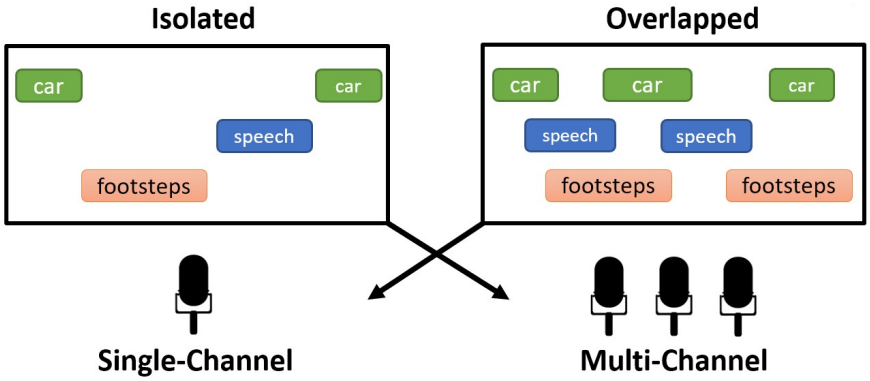
για την περαιτέρω εξέταση των φωνητικών τμημάτων του δωματίου και την διαγραφή αυτών που προέρχονται από άλλα δωμάτια. Στην εργασία [72], στο πρώτο στάδιο, πολυ-επίπεδα perceptrons χρησιμοποιούνται για κάθε μικρόφωνο, και ο χρονικός εντοπισμός φωνής επιτυγχάνεται σε κάθε δωμάτιο μέσω πλειοψηφικής στρατηγικής. Στη συνέχεια, για τα τμήματα τα οποία αντιστοιχήθηκαν σε περισσότερα από ένα δωμάτια, ένα χαρακτηριστικό βασισμένο στην μέτρηση της απόκλισης της περιβάλλουσας του φωνητικού σήματος (envelope distortion measure) χρησιμοποιείται για να αποφασιστεί τελικά το σωστό δωμάτιο. Στην εργασία [73], τρία διαφορετικά χαρακτηριστικά εξετάζονται για τον χωρο-χρονικό εντοπισμό φωνής, και συγκεκριμένα ο σηματοθορυβικός λόγος, η περιοδικότητα, και το πεδίο καθολικής συνάφειας (global coherence field). Τα χρονικά όρια των τμημάτων φωνής για κάθε δωμάτιο υπολογίζονται με απλή κατωφλίωση των τιμών των ανωτέρω χαρακτηριστικών και κάνοντας χρήση ενός ευριστικού κανόνα για συνεχόμενα ενεργά χρονικά πλαίσια.

Επιπρόσθετα, μέθοδοι ενός μόνο σταδίου έχουν επίσης αναπτυχθεί για το πρόβλημα του χωρο-χρονικού εντοπισμού φωνής. Συγκεκριμένα, στην εργασία [74], ένα βαθύ νευρωνικό δίκτυο παίρνει σαν είσοδο ένα 176-διάστατο διάνυσμα αποτελούμενο από ένα σύνολο διαφορετικών χαρακτηριστικών, όπως MFCCs, RASTA-PLPs, διακύμανση περιβάλλουσας, τονικότητα, κλπ. Παρόμοιο σύνολο χαρακτηριστικών (αλλά 187-διάστατο τελικό διάνυσμα) και νευρωνικά δίκτυα επιλέγονται επίσης στην εργασία [75], όπως επίσης και επιπλέον ταξινομητές, περιλαμβανομένων και 2D-CNN. Η μέθοδος αυτή, επεκτείνεται και σε μια 3D-CNN εκδοχή στην [76], όπου ως χαρακτηριστικά χρησιμοποιούνται 40-διάστατα διανύσματα με τις λογαριθμικές ενέργειες υπολογισμένες από συστοιχίες συχνοτικών φίλτρων στην Mel κλίμακα (log-FBEs), η χρονική πληροφορία εισάγεται με την συνένωση διανυσμάτων χαρακτηριστικών από γειτονικά χρονικά πλαίσια, και τέλος οι 2D αναπαραστάσεις που προκύπτουν για κάθε μικρόφωνο συνενώνονται από όλα τα μικρόφωνα, οδηγώντας έτσι σε έναν 3D πίνακα χαρακτηριστικών. Τέλος, στην εργασία [77], το προαναφερθέν 3D-CNN σύστημα συνδυάζεται με έναν ταξινομητή CNN που χρησιμοποιεί τα χωρικά χαρακτηριστικά γενικευμένης ετερο-συσχέτισης (GCC-PHAT), επιτυγχάνοντας από κοινού χρονικό εντοπισμό φωνής και χωρικό εντοπισμό του ομιλητή.

Συνεισφορά Διατριβής

Το αντικείμενο αυτής της Διδακτορικής διατριβής αφορά στις ερευνητικές περιοχές της επεξεργασίας ακουστικών σημάτων και της μηχανικής μάθησης, με κύριους άξονες τον χωρο-χρονικό εντοπισμό και την κατηγοριοποίηση ακουστικών γεγονότων.

Σχετικά με το ευρύτερο θέμα του εντοπισμού ακουστικών γεγονότων, στην διατριβή αυτή θεωρήθηκαν διάφορες παραλλαγές του προβλήματος, περιλαμβάνοντας σενάρια με μεμονωμένα ή και επικαλυπτόμενα γεγονότα, καθώς και την ύπαρξη ενός ή πολλαπλών μικροφώνων για την καταγραφή του ακουστικού περιβάλλοντος (Σχήμα 1). Στα πλαίσια αυτά αναπτύχθηκαν και αξιολογήθηκαν διάφορες μέθοδοι για τον εντοπισμό ακουστικών γεγονότων. Ακολουθεί μια σύντομη περιγραφή των σημαντικότερων συνεισφορών της παρούσας διατριβής.



Σχήμα 1: Σχηματική απεικόνιση των διαφορετικών σεναρίων ως προς τα φαινόμενα επικάλυψης και τον αριθμό μικροφώνων. Στην μελέτη μας θεωρήθηκαν όλοι οι δυνατοί συνδυασμοί σεναρίων.

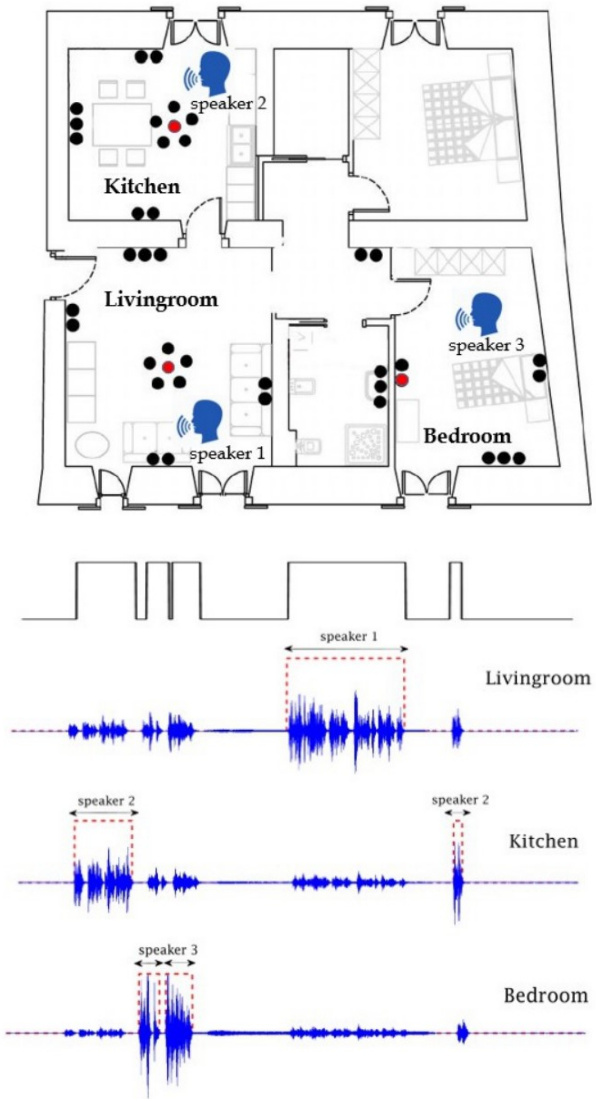
Χωρο-χρονικός εντοπισμός φωνής σε πολυ-καναλικά περιβάλλοντα

Αρχικά, στα πλαίσια ενός περιβάλλοντος «έξυπνου» σπιτιού, αναπτύχθηκε ένα σύστημα χωρο-χρονικού εντοπισμού φωνής, το οποίο έχει την δυνατότητα να εντοπίζει τα χρονικά όρια της φωνής («πότε») αλλά και την θέση του ομιλητή («πού») σε επίπεδο δωματίου, όπως φαίνεται στο Σχήμα 2. Το σύστημα αυτό μπορεί να διευκολύνει την επικοινωνία πολλαπλών ομιλητών σε διαφορετικά δωμάτια με την διεπαφή φωνής του «έξυπνου» σπιτιού. Επιπλέον, για να επιτευχθεί ικανοποιητική απόδοση του συστήματος σε ρεαλιστικές και απαιτητικές ακουστικές συνθήκες, το σύστημα που προτάθηκε εκμεταλλεύεται την πληροφορία από πολλαπλά μικρόφωνα που είναι εγκατεστημένα στον «έξυπνο» χώρο, ενώ ακολουθεί μια αποτελεσματική λογική δύο σταδίων. Όπως φαίνεται στο διάγραμμα του Σχήματος 3, το πρώτο στάδιο αποτελείται από ένα πολυ-καναλικό σύστημα μηχανικής μάθησης το οποίο εντοπίζει τα χρονικά όρια της φωνής στο σπίτι, ενώ το δεύτερο στάδιο βασίζεται σε καινοτόμα και πολυ-καναλικά ακουστικά χαρακτηριστικά για να εντοπίσει χωρικά τον ομιλητή σε επίπεδο δωματίου.

Αρχικά για το πρώτο στάδιο, ο πυρήνας του συστήματος βρίσκεται στην μοντελοποίηση φωνής/σιωπής για ένα μικρόφωνο. Για κάθε μικρόφωνο, εξάγεται ένα 39-διάστατο διάνυσμα παραδοσιακών χαρακτηριστικών MFCCs σε παράθυρα των 25 ms. Στη συνέχεια, ένας ταξινομητής βασισμένος σε GMMs εκπαιδεύεται σε αυτά τα χαρακτηριστικά για τον διαχωρισμό των δύο κλάσεων («φωνή», «σιωπή»). Όσον αφορά στον συνδυασμό των πολλαπλών μικροφώνων, δοκιμάστηκαν αρκετές μέθοδοι σε επίπεδο απόφασης. Στην μέθοδο που αποδείχθηκε η πιο αποδοτική, για κάθε χρονικό παράθυρο, υπολογίζεται ένα σταθμισμένο άθροισμα των τιμών λογαριθμικής πιθανοφάνειας για τις δύο κλάσεις, όπως προκύπτουν από τους GMM ταξινομητές των διαφόρων μικροφώνων ενός δωματίου:

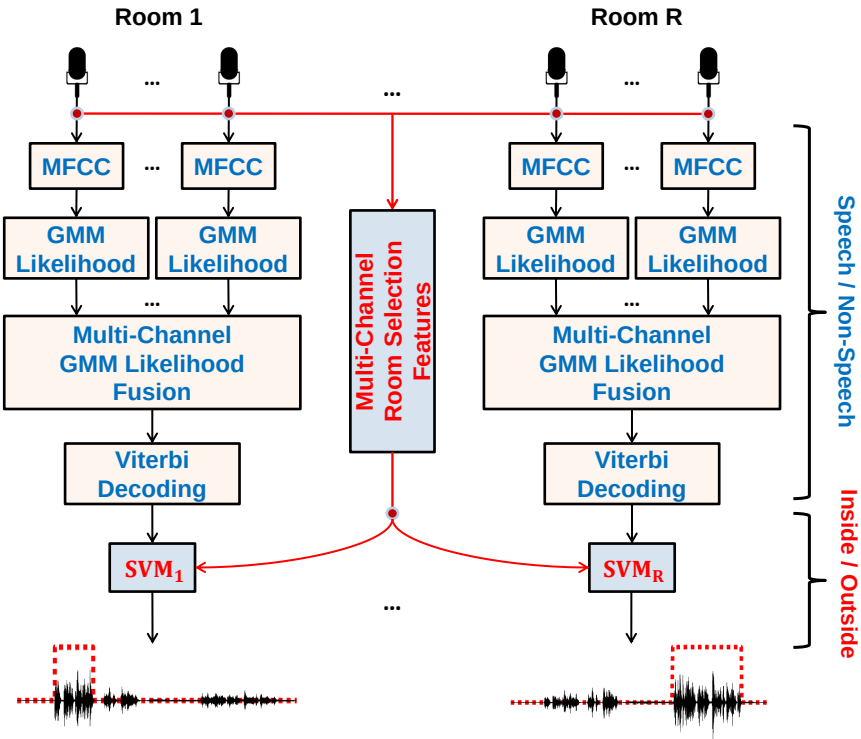
$$c_{\mathcal{M},j}(\mathbf{o}_{\mathcal{M},t}) \;=\; \sum_{m \in \mathcal{M}} w_{m,t}\, b_{m,j}(\mathbf{o}_{m,t}) \;,$$

όπου $\mathcal{M}$ είναι το σύνολο των μικροφώνων του δωματίου, $b_{m,j}(\mathbf{o}_{m,t})$ είναι η λογαριθμική πιθανοφάνεια των GMMs για το μικρόφωνο m, δεδομένων των ακουστικών χαρακτηριστικών



Σχήμα 2: Παράδειγμα χρονικού και χωρο-χρονικού εντοπισμού φωνής σε ένα «έξυπνο» σπίτι πολλών δωματίων, εξοπλισμένο με πολλαπλά μικρόφωνα. Σε αυτό το παράδειγμα, τρεις ομιλητές δίνουν φωνητικές εντολές σε τρία δωμάτια. Πάνω: Κάτοψη του «έξυπνου» σπιτιού που χρησιμοποιήθηκε στα πλαίσια του ερευνητικού προγράμματος ***DIRHA*** [100], με τις μαύρες κουκίδες να υποδηλώνουν τα διάφορα μικρόφωνα, εγκατεστημένα στους τοίχους και τις οροφές. Κάτω: ηχητικές κυματομορφές διάρκειας ενός λεπτού, ηχογραφημένες από τα μικρόφωνα με κόκκινο χρώμα (ένα για κάθενα από τα τρία ενεργά δωμάτια), με επισημειωμένες τις αντίστοιχες χρονικά ενεργές περιοχές για κάθε δωμάτιο. Στην κορυφή φαίνεται επίσης η επισημείωση των χρονικά ενεργών περιοχών φωνής για όλο το σπίτι.

$\mathbf{o}_{m,t}$, και της κλάσης $j \in \{\text{sp}, \text{sil}\}$ («φωνή», «σιωπή»). Τα βάρη μπορούν να είναι σταθερά και ίδια για όλα τα μικρόφωνα ($w_{m,t} = 1 \,/\, |\mathcal{M}|$), είτε να αλλάζουν δυναμικά σε κάθε χρονικό παράθυρο t και ανάλογα με την εμπιστοσύνη στην απόφαση του κάθε μικροφώνου, οπότε υπολογίζονται ως:



Σχήμα 3: Σχηματικό διάγραμμα του προτεινόμενου συστήματος για τον χωρο-χρονικό εντοπισμό φωνής. Οι διεργασίες του πρώτου σταδίου απεικονίζονται με μπλέ χρώμα, ενώ του δεύτερου σταδίου με κόκκινο.

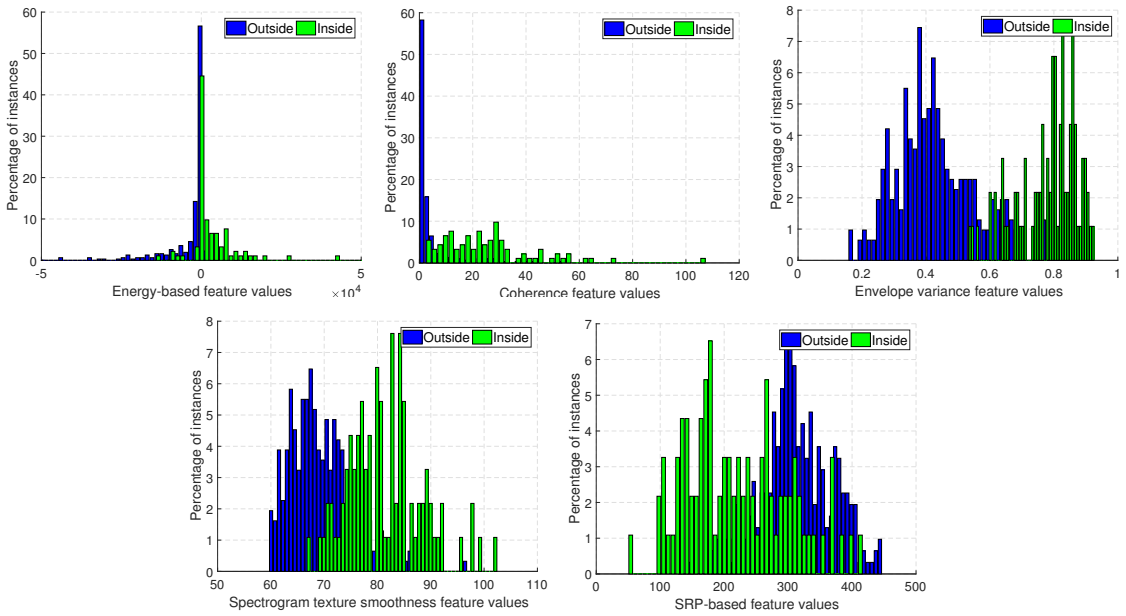
$$w_{m,t} \;=\; \frac{|\; b_{m,\,\mathrm{sp}}(\mathbf{o}_{m,t}) \;-\; b_{m,\,\mathrm{sil}}(\mathbf{o}_{m,t}) \;|}{\sum\limits_{m' \in \mathcal{M}} |\; b_{m',\,\mathrm{sp}}(\mathbf{o}_{m',t}) \;-\; b_{m',\,\mathrm{sil}}(\mathbf{o}_{m',t}) \;|} \;,$$

Μετά τον υπολογισμό των πολυ-καναλικών πιθανοφανειών, η τελική χρονική κατάτμηση σε περιοχές «φωνής»/«σιωπής» γίνεται με χρήση του αλγορίθμου Viterbi ως εξής:

$$\delta_{\mathcal{M},j}(t) \;=\; \max_{j'}\{\delta_{\mathcal{M},j'}(t-1) \;+\; \log(a_{j'j})\} \;+\; c_{\mathcal{M},j}(\mathbf{o}_{\mathcal{M},t}) \;,$$

όπου $\delta_{M,j}(t)$ δηλώνει το σκορ του καλύτερου μονοπατιού που καταλήγει στην κλάση j μετά από t παράθυρα παρατήρησης. Με κατάλληλο ορισμό των πιθανοτήτων μετάβασης $a_{j'j}$ μεταξύ των κλάσεων, ρυθμίζουμε την ευκολία μετάβασης μεταξύ «φωνής»/«σιωπής».

Στο δεύτερο στάδιο, πέντε καινοτόμα χαρακτηριστικά εξάγονται για κάθε τμήμα φωνής που εντοπίστηκε από το πρώτο στάδιο, έτσι ώστε να το κατηγοριοποιήσουν ως «εντός» ή «εκτός» δωματίου. Ο σχεδιασμός αυτών των ειδικών χαρακτηριστικών είχε σαν αφετηρία τρείς βασικές ιδέες: (α) Τα σήματα φωνής εκτός δωματίου θα έχουν μικρότερη ακουστική «ενέργεια» (ή επίσης μικρότερο SNR) σε σχέση με αυτά που παράγονται μέσα από το δωμάτιο, (β) τα σήματα που έρχονται από γειτονικά δωμάτια, θα χαρακτηρίζονται από μεγαλύτερη «αντήχηση» καθώς θα έχουν ανακλαστεί περισσότερες φορές, (γ) η δίοδος αύξησης των ηχητικών σημάτων από εξωτερικά δωμάτια θα είναι μέσω της πόρτας του δωματίου. Η διακριτική ικανότητα των



Σχήμα 4: Τα ιστογράμματα των πέντε ειδικά σχεδιασμένων χαρακτηριστικών για τον εντοπισμό του δωματίου του ομιλητή. Είναι εμφανής η ικανότητα τους για την κατηγοριοποίηση ενός σήματος φωνής ως «εντός» ή «εκτός» δωματίου.

διαφόρων χαρακτηριστικών φαίνεται στο Σχήμα 4. Επίσης, αξίζει να σημειωθεί ότι ενώ τα χαρακτηριστικά του πρώτου σταδίου εξάγονται ανά μικρόφωνο, τα χαρακτηριστικά του δεύτερου σταδίου είναι πολυ-καναλικά, και εξάγονται ανά δωμάτιο. Στη συνέχεια, τα χαρακτηριστικά από τα διάφορα δωμάτια συνδυάζονται μεταξύ τους για να προκύψει ένα συνολικό διάνυσμα χαρακτηριστικών, το οποίο τροφοδοτεί έναν ταξινομητή SVM για την κατηγοριοποίηση ενός τμήματος φωνής ως εντός ή εκτός δωματίου. Συγκεκριμένα εκπαιδεύεται ένας ταξινομητής SVM για κάθε δωμάτιο, οπότε ένα εντοπισμένο τμήμα φωνής μπορεί τελικά να αποδοθεί σε ένα, σε περισσότερα από ένα, ή και σε κανένα δωμάτιο.

Το προτεινόμενο σύστημα, το οποίο δεν απαιτεί μεγάλο όγκο δεδομένων εκπαίδευσης, αξιολογείται εκτενώς σε συνθετικές αλλά και πραγματικές πολυ-καναλικές βάσεις (ηχογραφημένες στο «έξυπνο» σπίτι του προγράμματος DIRHA), και σημειώνει σημαντικές βελτιώσεις έναντι εναλλακτικών παραδοσιακών μεθόδων. Επιπλέον, επιδεικνύει ευρωστία σε σενάρια με μικρότερο αριθμό διαθέσιμων μικροφώνων, ενώ επίσης συγκρίνεται ευνοϊκά με σύγχρονες μεθόδους βαθιάς μάθησης, όπως φαίνεται στον Πίνακα 1.

Μονο-καναλική μέθοδος NMF για εντοπισμό επικαλυπτόμενων ακουστικών γεγονότων

Στη γενικότερη περίπτωση του εντοπισμού ακουστικών γεγονότων, ιδιαίτερη έμφαση δόθηκε σε σενάρια με χρονική επικάλυψη. Η αρχική μας προσέγγιση για την αντιμετώπιση της επικάλυψης βασίζεται σε μεθόδους σχετικές με NMF, καθώς διακρίνονται για την ευρωστία τους σε συνθήκες θορύβου, και για τη δυνατότητά τους να επιτρέπουν τον εντοπισμό ταυτόχρονων γεγονότων, δεδομένου ότι κατάλληλες μη-αρνητικές και γραμμικές αναπαραστάσεις τους είναι διαθέσιμες.

Στόχος των μέθοδων NMF είναι η μη-αρνητική παραγοντοποίηση ενός πίνακα χαρακτηρι-

Πίνακας 1: Αποτίμηση του προτεινόμενου συστήματος έναντι εναλλακτικών συστημάτων βασισμένων σε μεθόδους βαθιάς μάθησης, για τον χωρο-χρονικό εντοπισμό φωνής στην βάση DIRHA-sim-evalita [112].

method		SAD error (%)
deep-learning	DNN [75]	5.8
	3D-CNN [76]	7.0
	3D-CNN [77]	5.2
	3D-CNN (SAD+SLOC) [77]	3.5
proposed		4.7

στικών $\mathbf{V} \in \mathbb{R}_{\geq 0}^{P \times N}$, από το γινόμενο $\mathbf{V} \approx \mathbf{\Lambda} = \mathbf{W} \cdot \mathbf{H}$, όπου $\mathbf{W} \in \mathbb{R}_{\geq 0}^{P \times R}$ είναι ο πίνακας λεξικό, και $\mathbf{H} \in \mathbb{R}_{\geq 0}^{R \times N}$ είναι ο πίνακας ενεργοποιήσεων. Το P συμβολίζει την διάσταση του διανύσματος χαρακτηριστικών, το N τον αριθμό των χρονικών πλαισίων και το R τον συνολικό αριθμό των λέξεων/μοτίβων στον πίνακα λεξικό. Ένα παράδειγμα εφαρμογής NMF φαίνεται στο Σχήμα 5. Η ελαχιστοποίηση μιας κατάλληλης συνάρτησης κόστους $D(\mathbf{V}||\mathbf{\Lambda})$ οδηγεί σε εξισώσεις υπολογισμού των $\mathbf{W}$ και $\mathbf{H}$ [118]. Στην περίπτωση μας χρησιμοποιούμε την Kullback-Leibler (KL) divergence συνάρτηση κόστους η οποία ορίζεται ως εξής:

$$D(\mathbf{V}||\mathbf{\Lambda}) = ||\mathbf{V} \odot \log(\mathbf{V} \oslash \mathbf{\Lambda}) - \mathbf{V} + \mathbf{\Lambda}|| \ ,$$

όπου τα $\odot$ και $\oslash$ συμβολίζουν πολλαπλασιασμό και διαίρεση στοιχείο-προς-στοιχείο μεταξύ πινάκων. Επίσης στην περίπτωση του προβλήματος εντοπισμού γεγονότων, συχνά χρησιμοποιείται η μέθοδος sparse-NMF, η οποία εισάγει στην αντικειμενική συνάρτηση έναν όρο αραιότητας του πίνακα ενεργοποιήσεων:

$$G\,(\,\mathbf{V}\,||\,\mathbf{\Lambda}\,)\ =\ D\,(\,\mathbf{V}\,||\,\mathbf{\Lambda}\,)\ +\ \lambda\,\|\mathbf{H}\|_1 \ ,$$

όπου η παράμετρος λ ελέγχει την ισορροπία μεταξύ ακριβούς ανακατασκευής και αραιότητας της λύσης. Σε αυτή την δουλειά χρησιμοποιούμε τη συνελικτική επέκταση CNMF [124] η οποία επιτρέπει την ανασύνθεση των γεγονότων από λέξεις/μοτίβα με χρονική εξέλιξη. Στην περίπτωση του CNMF ο πίνακας $\mathbf{V}$ προσεγγίζεται από το συνελικτικό άθροισμα του πίνακα λεξικού και του πίνακα ενεργοποιήσεων:

$$\mathbf{V}\ \approx\ \mathbf{\Lambda}\ =\ \sum_{t=0}^{T-1} \mathbf{W}_t \cdot \overset{t\rightarrow}{\mathbf{H}} \ ,$$

όπου ο τελεστής $\overset{t\rightarrow}{\bullet}$ μετατοπίζει τις στήλες ενός πίνακα κατά t θέσεις προς τα δεξιά, T είναι η χρονική διάρκεια των λέξεων, και $\mathbf{W}_t \in \mathbb{R}_{\geq 0}^{P \times R}$ είναι το τμήμα του πίνακα λεξικού που σχετίζεται με τη χρονική στιγμή t των λέξεων.

Στην περίπτωση εντοπισμού ακουστικών γεγονότων, έχοντας αποκτήσει έναν πίνακα λεξικό από δεδομένα εκπαίδευσης, η μέθοδος sparse-CNMF παίρνει ως είσοδο τον πίνακα πα-

ρατηρήσεων $\mathbf{V}$ και δίνει στην έξοδο τον πίνακα ενεργοποιήσεων $\mathbf{H}$. Στη συνέχεια, στην παραδοσιακή μέθοδο εντοπισμού, για κάθε χρονικό πλαίσιο, αθροίζονται οι ενεργοποιήσεις που αντιστοιχούν σε κάθε ακουστικό γεγονός, δημιουργώντας έτσι έναν νέο πίνακα $\mathbf{H'} \in \mathbb{R}_{\geq 0}^{C \times N}$ συνολικών ενεργοποιήσεων για κάθε γεγονός:

$$H'(\,i\,,n\,) \,=\sum_{r\,\in\,\{i\}} H(\,r\,,n\,) \;\;,$$

όπου το i συμβολίζει το γεγονός ($i = 1,\cdots,C$), $\{i\}$ είναι το σύνολο των γραμμών του πίνακα $\mathbf{H}$ που αντιστοιχούν σε λέξεις/μοτίβα του i-οστού γεγονότος, και $n \in \{1,\cdots,N\}$ συμβολίζει το χρονικό πλαίσιο. Τέλος, ένα γεγονός θεωρείται ενεργό κατά το χρονικό πλαίσιο n, εαν $H'(i,n) > \theta_H$, όπου θ_H είναι ένα κατάλληλο κατώφλι ενεργοποίησης. Ωστόσο αυτή η παραδοσιακή προσέγγιση εντοπισμού είναι ευπαθής σε ψευδείς ενεργοποιήσεις (false alarms), καθώς κάποιες φορές μπορεί να προκύψουν εσφαλμένα ενεργοποιήσεις μεγάλης έντασης για κάποιες λέξεις ενός γεγονότος, χωρίς να υπάρχει στην πραγματικότητα αυτό το γεγονός.

Για να αυξήσουμε την ευρωστία της μεθόδου εντοπισμού, προτείνουμε την εισαγωγή ενός επιπλέον κριτηρίου το οποίο σχετίζεται με το υπόλοιπο ανακατασκευής. Συγκεκριμένα, για κάθε ακουστική κλάση, και για ένα χρονικό διάστημα seg, μετράμε το KL-divergence κόστος $D\,(\,\mathbf{V}_{seg}\,||\,\mathbf{\Lambda}_{seg}^{(i,bg)}\,)$, όπου:

$$\mathbf{\Lambda}_{seg}^{(i,bg)} = \sum_{t\,=\,0}^{T-1} \mathbf{W}_{t}^{(i,bg)} \cdot \overset{t\,\rightarrow}{\mathbf{H}}_{seg}^{(i,bg)} \;\;,$$

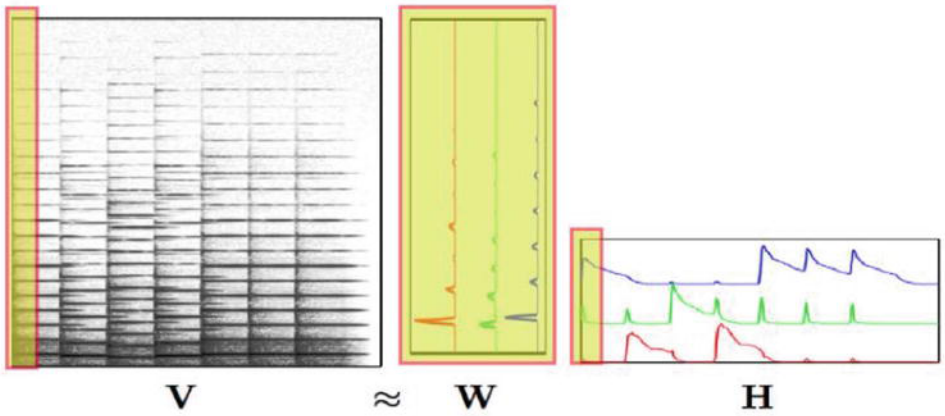
δηλαδή το υπόλοιπο ανακατασκευής, εαν χρησιμοποιηθούν για την ανακατασκευή μονάχα οι λέξεις που αντιστοιχούν στο γεγονός-i (και στο γεγονός περιβαλλοντικού θορύβου bg). Στη συνέχεια υπολογίζουμε το «λόγο υπολοίπου» («residual ratio») του i-οστού γεγονότος, ως τον λόγο:

$$\mathcal{E}(i\,,n) = \frac{D\,(\mathbf{V}_{seg}\,||\,\mathbf{\Lambda}_{seg}^{(i,bg)})}{D\,(\mathbf{V}_{seg}\,||\,\mathbf{\Lambda}_{seg})} \;\;,\quad \text{for all} \quad n \in seg \;\;.$$

όπου $D\,(\,\mathbf{V}_{seg}\,||\,\mathbf{\Lambda}_{seg}\,)$ είναι το υπόλοιπο ανακατασκευής όταν χρησιμοποιούνται για την ανακατασκευή λέξεις από όλα τα γεγονότα για το δεδομένο διάστημα seg. Τελικά, η απόφαση για τον εντοπισμό ενός γεγονότος προκύπτει με συνδυαστική κατωφλίωση των δύο κριτηρίων:

$$H'(i\,,n) > \theta_H \quad \text{and} \quad \mathcal{E}(i\,,n) < \theta_{\mathcal{E}} \;\;.$$

Για τα πειράματα χρησιμοποιήθηκε η βάση "Sound event detection in synthetic audio" από τον διαγωνισμό DCASE'16 [79], η οποία περιέχει έντεκα ακουστικά γεγονότα σχετικά με ένα περιβάλλον γραφείου («καθαρισμός λαιμού», «βήχας», «χτύπημα πόρτας», «κλείσιμο πόρτας», «συρτάρι», «γέλιο», «πληκτρολόγηση», «ήχος κλειδιών», «ξεφύλλισμα βιβλίου», «ήχος κλήσης κινητού», «φωνή»). Στον Πίνακα 2 φαίνονται τα αποτελέσματα για την μέθοδο CNMF με το προτεινόμενο στάδιο εντοπισμού (activations&residuals), την μέθοδο CNMF με το παραδοσιακό στάδιο εντοπισμού (activations-only), καθώς και μια παραδοσιακή μέθοδο NMF. Οι τρείς μέθοδοι αξιολογούνται σε τρία διαφορετικά υπο-σύνολα της βάσης, και με δύο διαφο-



Σχήμα 5: Παράδειγμα εφαρμογής NMF σε μια ηχογράφηση πιάνου για το τραγούδι "Mary had a little lamb". Ο πίνακας λεξικό **W** περιλαμβάνει λέξεις που είναι σχετικές με την φασματική πληροφορία της κάθε νότας, και ο πίνακας ενεργοποιήσεων **H** εντοπίζει τις χρονικές στιγμές που χρησιμοποιήθηκε η κάθε νοτα (από [119]).

Πίνακας 2: Αποτίμηση της προτεινόμενης μεθόδου και των παραδοσιακών μεθόδων.

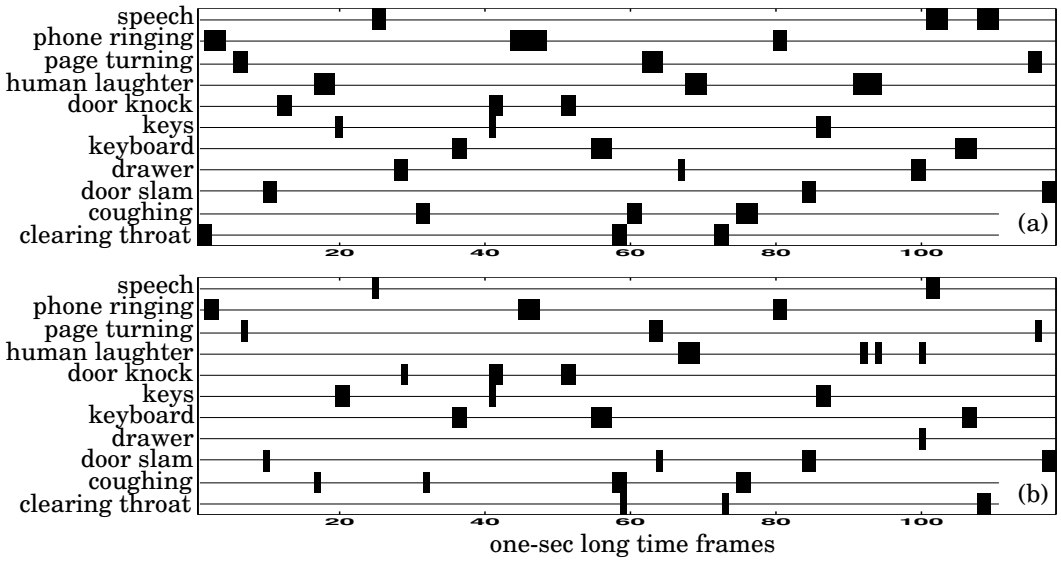
system	setup #1		setup #2		test	
	F-score	ER	F-score	ER	F-score	ER
NMF-baseline	0.42	0.79	0.32	0.87	0.37	0.89
activations-only	0.83	0.30	0.43	0.79	–	–
activations&residuals	0.84	0.29	0.55	0.63	0.56	0.68

ρετικές μετρικές [79] (υψηλές τιμές για το F-score και χαμηλές τιμές για το ER σημαίνουν καλύτερη απόδοση). Παρατηρούμε ότι η προτεινόμενη μέθοδος καταφέρνει να ενισχύσει την ευρωστία του σταδίου εντοπισμού, οδηγώντας στα καλύτερα συνολικά αποτελέσματα. Τέλος, στο Σχήμα 6 βλέπουμε ένα παράδειγμα της εξόδου του προτεινόμενου συστήματος εντοπισμού γεγονότων, μαζί με τις πραγματικές ενεργοποιήσεις γεγονότων, για μια ηχογράφηση διάρκειας δύο λεπτών.

Πολυ-καναλικές μέθοδοι NMF για εντοπισμό επικαλυπτόμενων ακουστικών γεγονότων

Στην περίπτωση ενός περιβάλλοντος με πολλαπλά μικρόφωνα, η κατάλληλη αξιοποίηση της πολυ-καναλικής πληροφορίας μπορεί να οδηγήσει σε σημαντικά βελτιωμένα αποτελέσματα. Σε αυτή την ενότητα εξετάζουμε δύο πολυ-καναλικές επεκτάσεις της μεθόδου NMF, κατάλληλες για εντοπισμό επικαλυπτόμενων ακουστικών γεγονότων. Η μονο-καναλική μέθοδος πάνω στην οποία στηριζόμαστε για τον σχεδιασμό των επεκτάσεων είναι μια μέθοδος sparse-NMF (δηλαδή NMF με «αραιό» πίνακα ενεργοποιήσεων).

Στην πρώτη πολυ-καναλική μέθοδο, συνδυάζουμε τα διάφορα μικρόφωνα σε επίπεδο απόφασης, αναμένοντας ότι θα πάρουμε πιο αξιόπιστα αποτελέσματα σε σύγκριση με ένα μονο-καναλικό σύστημα. Συγκεκριμένα, όπως φαίνεται στο Σχήμα 7, αρχικά κάθε μικρόφωνο m λειτουργεί ανεξάρτητα από τα υπόλοιπα, υπολογίζοντας μέσω sparse-NMF τον δικό του πίνακα



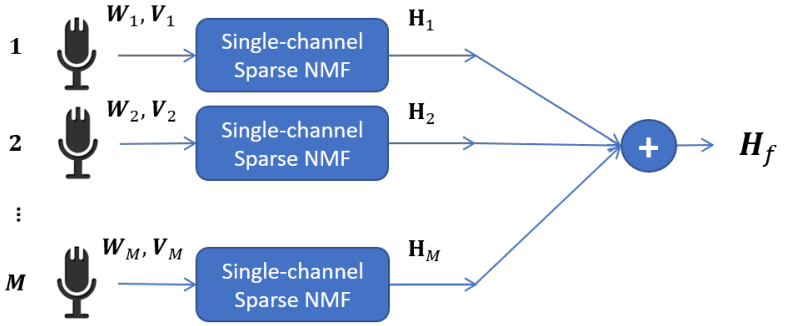
Σχήμα 6: Παράδειγμα εντοπισμού ακουστικών γεγονότων από το προτεινόμενο CNMF συστήμα. (α) πραγματικές ενεργοποιήσεις γεγονότων (β) έξοδος του προτεινόμενου συστήματος.

ενεργοποιήσεων $\mathbf{H}_m$, χρησιμοποιώντας το δικό του λεξικό $\mathbf{W}_m$ και τον πίνακα παρατηρήσεων $\mathbf{V}_m$. Στη συνέχεια ένας τελικός πίνακας ενεργοποιήσεων υπολογίζεται ως ο μέσος όρος των επιμέρους πινάκων από τα διάφορα μικρόφωνα.

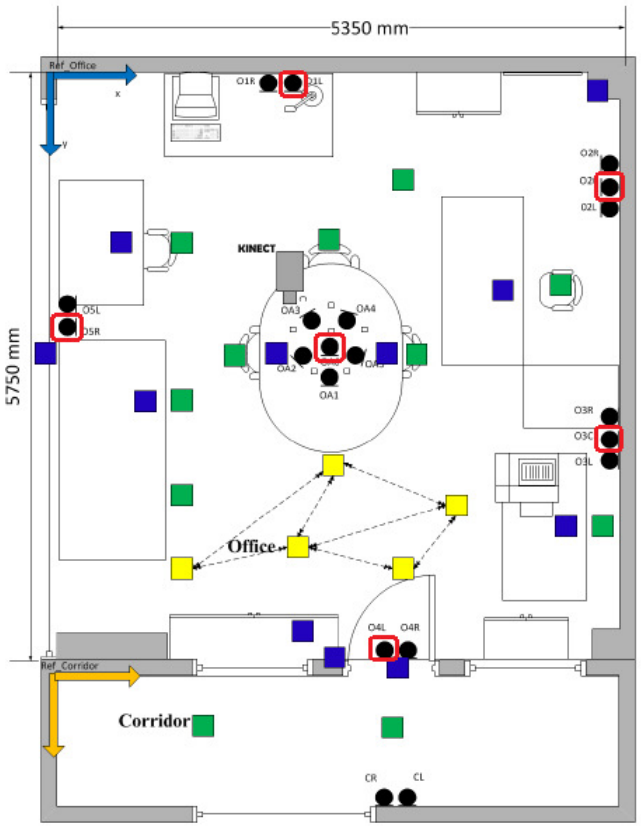
Στη δεύτερη μέθοδο, θεωρούμε την βελτιστοποίηση μιας καινοτόμου αντικειμενικής συνάρτησης, η οποία αποτελείται από έναν πολυ-καναλικό όρο ανακατασκευής και ένα πολυ-καναλικό όρο αραιότητας:

$$J = \sum_{m=1}^{M} D(\mathrm{V}_m|\mathrm{WH}_m) + \lambda \sum_{n=1}^{N} \Omega(h_{1,n}, ..., h_{M,n}) \ ,$$

όπου M είναι ο αριθμός των μικροφώνων, N ο αριθμός των χρονικών πλαισίων, $\mathbf{W}$ ένα πολυ-καναλικό λεξικό που προκύπτει από τον συνδυασμό των επιμέρους λεξικών των διάφορων μικροφώνων, ενώ με $D(\mathbf{V}_m|\mathbf{WH}_m)$ συμβολίζεται η KL-divergence συνάρτηση κόστους ανακα-



Σχήμα 7: Σχηματικό διάγραμμα της πρώτης πολυ-καναλικής μεθόδου με άθροισμα των πινάκων ενεργοποιήσεων από τα διάφορα μικρόφωνα.



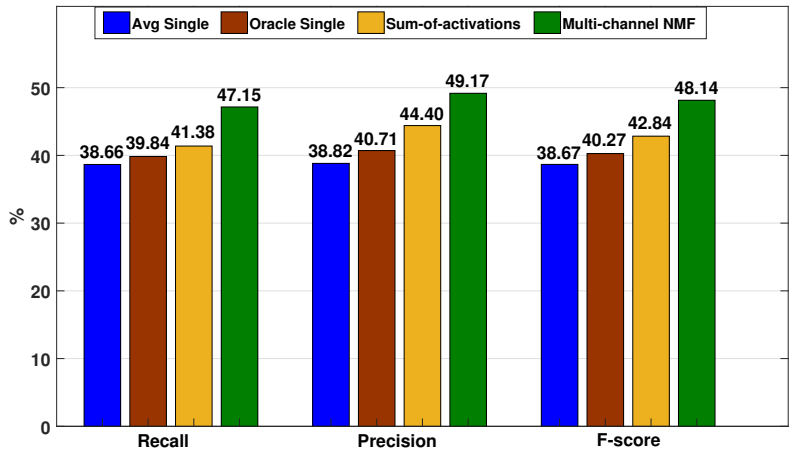
Σχήμα 8: Σχέδιο κάτοψης του «έξυπνου» γραφείου της βάσης ATHENA. Φαίνονται οι θέσεις για τα μικρόφωνα (μαύρο), τους ομιλητές (πράσινο και κίτρινο), και τα ακουστικά γεγονότα (μπλε). Τα έξι μικρόφωνα που χρησιμοποιήθηκαν για τις πολυ-καναλικές μεθόδους NMF επισημειώνονται με κόκκινο πλαίσιο.

τασκευής για το μικρόφωνο m. Επίσης, $\mathbf{H}_m = [h_{m,1}, ..., h_{m,N}]$ και $h_{m,n} = [h_{m,n}^{(1)\,T}, ..., h_{m,n}^{(C)\,T}]^T$, δηλαδή, με $h_{m,n}$ συμβολίζεται η n-οστή στήλη του πίνακα ενεργοποιήσεων $\mathbf{H}_m$. Τέλος, η συνάρτηση πολυ-καναλικής αραιότητας Ω ορίζεται ως:

$$\Omega(h_{1,n}, ..., h_{M,n}) = \sum_{c=1}^{C} \log(\epsilon + \sum_{m=1}^{M} \|h_{m,n}^{(c)}\|_1) \ ,$$

όπου ο όρος $h_{m,n}^{(c)}$ συμβολίζει το τμήμα της στήλης του πίνακα ενεργοποιήσεων που σχετίζεται με το ακουστικό γεγονός c, και C είναι ο αριθμός των διαφορετικών ακουστικών γεγονότων.

Με την εισαγωγή αυτής της αντικειμενικής συνάρτησης, στόχος μας είναι η συνεργασία των διαφορετικών μικροφώνων σε επίπεδο NMF για την εύρεση μιας καλύτερης λύσης. Συγκεκριμένα, ο όρος ανακατασκευής ελαχιστοποιεί το σφάλμα ανακατασκευής για όλα τα μικρόφωνα θεωρώντας επιπλέον ένα βελτιωμένο συνολικό λεξικό που περιέχει μοτίβα/λέξεις από όλα τα μικρόφωνα. Επίσης ο όρος αραιότητας, είναι ουσιαστικά μια πολυ-καναλική επέκταση του log/l1 group-sparsity, όπου σε εμάς τα groups είναι τα ακουστικά γεγονότα. Στόχος του είναι να επιτρέψει να ενεργοποιηθούν μόνο ενας μικρός αριθμός από γεγονότα από όλα τα μικρόφωνα σε κάθε χρονικό πλαίσιο, το οποίο έχει ως συνέπεια τα μικρόφωνα να τείνουν να



Σχήμα 9: Συγκριτική αποτίμηση των μονο-καναλικών και πολυ-καναλικών μεθόδων NMF για τον εντοπισμό επικαλυπτόμενων ακουστικών γεγονότων στην βάση ATHENA. Με Avg. Single συμβολίζεται η μέση απόδοση της μονο-καναλικής μεθόδου για τα διάφορα μικρόφωνα, ενώ με Oracle Single συμβολίζεται η απόδοση του καλύτερου μικροφώνου.

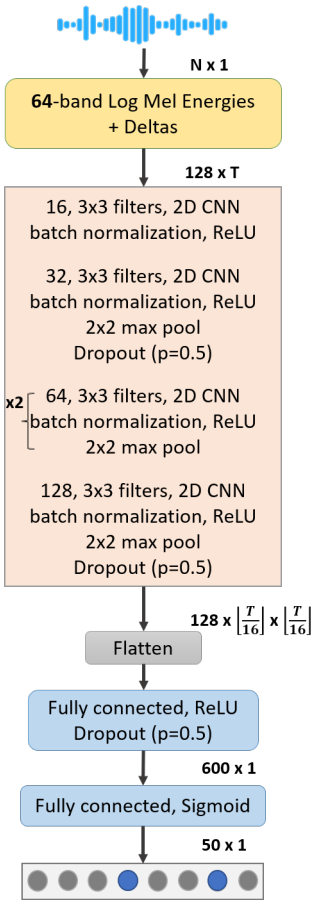
βρούν λύσεις στις οποίες συμφωνούν.

Για τα πειράματα μας χρησιμοποίηθηκε η πολυ-καναλική βάση ATHENA [80] (Σχήμα 8), η οποία περιέχει 4 ώρες ηχογραφήσεων από ένα πολυ-καναλικό «έξυπνο» χώρο γραφείου. Η βάση αυτή είναι κατάλληλη για πολυ-καναλικό εντοπισμό επικαλυπτόμενων γεγονότων, αφού περιέχει δεκαέξι διαφορετικά ακουστικά γεγονότα, τα οποία αρκετά συχνά επικαλύπτονται μεταξύ τους χρονικά.

Όσον αφορά την αξιολόγηση των μεθόδων, στο Σχήμα 9 βλέπουμε πως συγκρίνονται οι δύο προτεινόμενες πολυ-καναλικές μέθοδοι, σε σχέση με αποτελέσματα της μονο-καναλικής μέθοδου NMF. Παρατηρούμε ότι η δεύτερη μέθοδος πολυ-καναλικού NMF πετυχαίνει αισθητά βελτιωμένα αποτέλεσματα σε σύγκριση με όλες τις άλλες εναλλακτικές.

Βαθιά μάθηση για εντοπισμό επικαλυπτόμενων ακουστικών γεγονότων σε πολυ-καναλικά περιβάλλοντα

Έως τώρα, θεωρήσαμε μεθόδους βασισμένες σε NMF για το πρόβλημα των επικαλυπτόμενων γεγονότων. Πράγματι, οι NMF μέθοδοι αποτελούν μια ταιριαστή επιλογή για σενάρια με επικάλυψη, αφού έχουν εγγενώς τη δυνατότητα να εντοπίζουν γεγονότα που συμβαίνουν ταυτόχρονα. Επιπλέον μπορούν να εκπαιδευτούν με χρήση λίγων μόνο δεδομένων, καθώς και να δώσουν λύσεις που μπορούν να ερμηνευτούν εύκολα. Ωστόσο στις μέρες μας, έχει επικρατήσει η κατηγορία μεθόδων βαθιάς μάθησης [22–25], αφού υπερέχει σημαντικά ως προς την ταχύτητα υπολογισμού λύσης, αλλά κυριότερα, ως προς την διακριτική ικανότητα. Αξίζει ωστόσο να σημειωθεί ότι οι μέθοδοι βαθιάς μάθησης, απαιτούν αρκετά μεγαλύτερο όγκο δεδομένων για εκπαίδευση, όπως επίσης και την ύπαρξη στιγμιοτύπων επικάλυψης κατά την εκπαίδευση για την επίλυση σεναρίων με επικαλυπτόμενα γεγονότα.

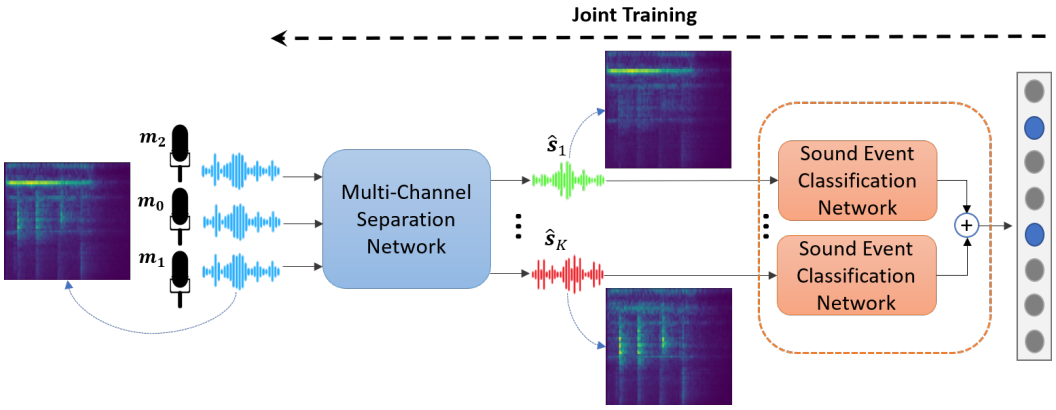


Σχήμα 10: Αρχιτεκτονική μονο-καναλικού νευρωνικού δικτύου για ταξινόμηση ακουστικών γεγονότων.

Πράγματι, η κυρίαρχη προσέγγιση εκπαίδευσης συστημάτων βαθιάς μάθησης για τον εντοπισμό επικαλυπτόμενων γεγονότων, είναι η τροφοδότηση ενός νευρωνικού δικτύου με πληθώρα στιγμιότυπων επικάλυψης που είτε υπάρχουν στα διαθέσιμα δεδομένα, είτε παράγονται με συνθετικό τρόπο από υπάρχοντα μεμονωμένα στιγμιότυπα. Ωστόσο σε περιπτώσεις που είτε ο αριθμός των πιθανών ακουστικών γεγονότων είναι μεγάλος, είτε ο βαθμός επικάλυψης (πολυφωνίας) αυξάνεται, ο παραδοσιακός αυτός τρόπος εκπαίδευσης γίνεται προβληματικός, καθώς απαιτεί την ύπαρξη ποικίλων στιγμιότυπων επικάλυψης για έναν σημαντικά μεγάλο αριθμό συνδυασμών γεγονότων.

Μια εναλλακτική προσέγγιση η οποία περιορίζει τα παραπάνω προβλήματα, είναι η χρήση ενός δικτύου διαχωρισμού γεγονότων σαν ένα πρώτο στάδιο, στοχεύοντας στην μετατροπή του προβλήματος ταξινόμησης επικαλυπτόμενων γεγονότων σε πρόβλημα ταξινόμησης μεμονωμένων γεγονότων. Εμπνευσμένες από την σημαντική πρόοδο που έχει σημειωθεί τα τελευταία χρόνια στο ερευνητικό πεδίο διαχωρισμού ακουστικών πηγών [132–137], κάποιες πρόσφατες μελέτες εντάσσουν τη χρήση τέτοιων δικτύων στα συστήματα τους και αναφέρουν βελτιωμένα αποτελέσματα για το πρόβλημα του μονο-καναλικού προβλήματος εντοπισμού ακουστικών γεγονότων με επικάλυψη.

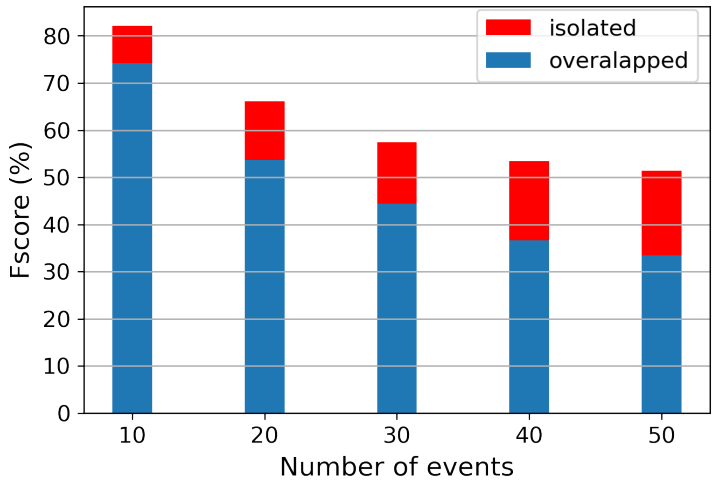
Σε αυτή την ενότητα, παρουσιάζουμε τη δουλειά μας στο πρόβλημα του εντοπισμού ε-



Σχήμα 11: Σχηματική απεικόνιση της αρχιτεκτονικής της προτεινόμενης μεθόδου για τον πολυ-καναλικό εντοπισμό επικαλυπτόμενων γεγονότων.

πικαλυπτόμενων ακουστικών γεγονότων σε πολυ-καναλικά περιβάλλοντα με χρήση μεθόδων βαθιάς μάθησης. Προτείνουμε για πρώτη φορά, το συνδυασμό ενός πολυ-καναλικού δικτύου διαχωρισμού ακουστικών γεγονότων με ένα δίκτυο ταξινόμησης ακουστικών γεγονότων, για την επίλυση του προβλήματος στο απαιτητικό σενάριο όπου ο αριθμός των πιθανών διαφορετικών ακουστικών κλάσεων είναι μεγάλος. Συγκεκριμένα, χρησιμοποιούμε ένα πολυ-καναλικό δίκτυο διαχωρισμού ηχητικών πηγών [141], έτσι ώστε να μπορέσουμε να εκμεταλλευτούμε, εκτός από την φασματική πληροφορία, και την χωρική πληροφορία για την διαφοροποίηση των ταυτόχρονων ακουστικών συμβάντων. Ως δίκτυο ταξινόμησης γεγονότων χρησιμοποιούμε ένα δίκτυο βασισμένο σε CNNs, η αρχιτεκτονική του οποίου φαίνεται στο Σχήμα 10. Το τελικό προτεινόμενο σύστημα, όπως φαίνεται στο Σχήμα 11, συνδυάζει τα δίκτυα διαχωρισμού και ταξινόμησης. Πιο αναλυτικά, αρχικά το δίκτυο διαχωρισμού παίρνει ως είσοδο το ακουστικό μείγμα των επικαλυπτόμενων γεγονότων, και δίνει στην έξοδο του K διαχωρισμένα σήματα, τα οποία αποτελούν την είσοδο του δικτύου ταξινόμησης. Η ιδέα είναι ότι, δεδομένης μιας ικανοποιητικής ποιότητας διαχωρισμού, το πρόβλημα επικαλυπτόμενων γεγονότων μπορεί να μετατραπεί προσεγγιστικά σε πρόβλημα ταξινόμησης ενός συνόλου μεμονωμένων ηχητικών γεγονότων, το οποίο αποτελεί ευκολότερο πρόβλημα. Το δίκτυο ταξινόμησης που χρησιμοποείται, εφαρμόζει το δίκτυο του Σχήματος 10, για κάθε ένα από τα K διαχωρισμένα σήματα, και στη συνέχεια υπολογίζει τον μέσο όρο των εξόδων τους. Για την εκπαίδευση του προτεινόμενου συστήματος, δοκιμάσαμε δύο διαφορετικές προσεγγίσεις, την «ακολουθιακή» (Sequential training), όπου τα δύο δίκτυα εκπαιδεύονται ξεχωριστά μεταξύ τους, και την «από κοινού» (Joint training), όπου τα δίκτυα επανα-εκπαιδεύονται από κοινού, έτσι ώστε οι παράμετροι και των δύο δικτύων να ρυθμιστούν και συντονιστούν ως προς το τελικό ζητούμενο της ταξινόμησης γεγονότων. Τέλος, στα συγκριτικά πειράματα μας θεωρήθηκε και ο συνδυασμός του προτεινόμενου δικτύου με το βασικό δίκτυο του Σχήματος 10 σε επίπεδο απόφασης.

Για τα πειράματα μας χρησιμοποιήθηκε η βάση δεδομένων ESC50 [81], η οποία περιέχει 2000 ηχητικά στιγμιότυπα διάρκειας 5 δευτερολέπτων από 50 διαφορετικά ακουστικά γεγονότα. Στη συνέχεια, δημιουργήσαμε μια πολυ-καναλική βάση, συνελίσοντας τα στιγμιότυπα αυτά με πραγματικές κρουστικές αποκρίσεις από την βάση δεδομένων DIRHA. Δημιουργήσαμε έτσι με



Σχήμα 12: Απόδοση του δικτύου ταξινόμησης για τα προβλήματα μεμονωμένων και επικαλυπτόμενων γεγονότων, για διαφορετικές τιμές του συνολικού αριθμού κλάσεων.

συνθετικό τρόπο μια βάση με πληθώρα στιγμιοτύπων επικάλυψης, όπως θα είχαν ηχογραφηθεί από μια συστοιχία 3 μικροφώνων. Ο μέσος χρόνος αντήχησης $\overline{T}_{60}$ κυμαίνεται μεταξύ 0.58 και 0.83 δευτερολέπτων, ενώ οι αποστάσεις των γεγονότων από το κεντρικό μικρόφωνο μεταξύ 0.72 και 3.2 μέτρων.

Προχωρώντας στα αποτελέσματα, αρχικά επιβεβαιώσαμε την δυσκολία που αντιμετωπίζουν τα νευρωνικά δίκτυα που εκπαιδεύονται με τον παραδοσιακό τρόπο, στην περίπτωση που ο αριθμός των πιθανών γεγονότων είναι μεγάλος. Όπως φαίνεται στο Σχήμα 12, καθώς ο αριθμός των ακουστικών κλάσεων αυξάνεται, η απόδοση του δικτύου ταξινόμησης χειροτερεύει και στην περίπτωση του προβλήματος μεμονωμένων γεγονότων (**isolated**), και στην περίπτωση των επικαλυπτόμενων γεγονότων (**overlapped**). Ωστόσο η διαφορά απόδοσης μεταξύ τους γίνεται όλο και μεγαλύτερη, αφού ο βαθμός πολυπλοκότητας του προβλήματος επικάλυψης αυξάνεται σημαντικά με την αύξηση του αριθμού των γεγονότων. Το προτεινόμενο σύστημα στοχεύει να αντιμετωπίσει αυτό το πρόβλημα και να πετύχει απόδοση παρόμοια με την περίπτωση των μεμονωμένων γεγονότων.

Τέλος στον Πίνακα 3, φαίνεται η συγκριτική αποτίμηση των προτεινόμενων μεθόδων και των παραδοσιακών μεθόδων για δύο διαφορετικά σενάρια αντήχησης. Παρατηρούμε γενικά ότι στην περίπτωση του σεναρίου μεσαίας αντήχησης ($\overline{T}_{60}$=0.61s), και οι δύο παραλλαγές της προτεινόμενης μεθόδου (**C** και **D**), πετυχαίνουν σημαντικές βελτιώσεις σε σχέση με τις παραδοσιακές μεθόδους, ενώ καλύτερη αποδεικνύεται η μέθοδος με το σχήμα «από κοινού» εκπαίδευσης (**D**). Επίσης, επιπλέον βελτιώσεις σημειώνονται με τον συνδυασμό των προτεινόμενων και των παραδοσιακών μεθόδων σε επίπεδο απόφασης (**Late Fusion**). Ωστόσο, στην περίπτωση του σεναρίου υψηλής αντήχησης ($\overline{T}_{60}$=0.80s), οι προτεινόμενες μέθοδοι αποτυγχάνουν να βελτιώσουν τα αποτελέσματα, λόγω της ανεπαρκούς απόδοσης του δικτύου διαχωρισμού. Το γεγονός αυτό έρχεται σε συμφωνία με πρόσφατες μελέτες για την απόδοση δικτύων διαχωρισμού σε συνθήκες υψηλής αντήχησης [148]. Παρόλα αυτά, ο συνδυασμός των προτεινόμενων μεθόδων με τις παραδοσιακές, έδωσε βελτιωμένα αποτελέσματα και σε αυτή την περίπτωση.

Πίνακας 3: Αποτίμηση των διαφόρων μεθόδων για το πρόβλημα των επικαλυπτόμενων ακουστικών γεγονότων.

System	F-score (%)	
	$\overline{T}_{60}$=0.61s	$\overline{T}_{60}$=0.80s
(A) Baseline (1 channel)	41.26	39.05
(B) Baseline (3 channels)	41.45	39.33
(C) Proposed - Sequential	44.72	38.41
(D) Proposed - Joint	47.46	38.75
Late Fusion (B+C)	46.20	41.52
Late Fusion (B+D)	48.95	41.95

Δομή Διατριβής

Η διατριβή ακολουθεί την παρακάτω δομή:

- Το Κεφάλαιο 1 εισάγει τις βασικές έννοιες για τα δύο προβλήματα που μελετούνται, τον εντοπισμό ακουστικών γεγονότων και τον εντοπισμό φωνής. Επιπλέον παρουσιάζει μια επισκόπηση των ερευνητικών μεθόδων που έχουν αναπτυχθεί στην σχετική βιβλιογραφία των τελευταίων ετών.

- Το Κεφάλαιο 2 παρουσιάζει την έρευνά μας στο πρόβλημα του χωρο-χρονικού εντοπισμού φωνής σε «έξυπνα» περιβάλλοντα. Συγκεκριμένα, στα πλαίσια ενός «έξυπνου» σπιτιού εξοπλισμένου με πολλαπλά μικρόφωνα, αναπτύσσουμε έναν αλγόριθμο δύο σταδίων, ο οποίος με την χρήση ειδικά σχεδιασμένων πολυ-καναλικών χαρακτηριστικών στο δεύτερο στάδιο, κατηγοριοποιεί τα τμήματα φωνής που έχουν ανιχνευθεί στο πρώτο στάδιο ως εντός ή εκτός δωματίου. Για τα πειράματά μας χρησιμοποιούμε τρείς διαφορετικές βάσεις δεδομένων που περιέχουν πολυ-καναλικά δεδομένα από το «έξυπνο» σπίτι του ερευνητικού έργου DIRHA. Η προτεινόμενη μέθοδος έχει καλύτερη απόδοση από διάφορες εναλλακτικές μεθόδους της σχετικής βιβλιογραφίας, ενώ επίσης παρουσιάζει ευρωστία και σε σενάρια με μειωμένο αριθμό από διαθέσιμα μικρόφωνα.

- Το Κεφάλαιο 3 παρουσιάζει την δουλειά μας στο πρόβλημα του εντοπισμού μεμονωμένων ακουστικών γεγονότων, εστιάζοντας σε μεθόδους συνδυασμού της πληροφορίας από πολλαπλά μικρόφωνα. Συγκεκριμένα, πειραματιζόμαστε με πολυ-καναλικό συνδυασμό σε επίπεδο σήματος, χαρακτηριστικών, απόφασης, ή πολυ-καναλικής εκπαίδευσης, χρησιμοποιώντας παραδοσιακά στατιστικά μοντέλα (HMMs, GMMs). Για τα πειράματά μας, γίνεται χρήση μιας πολυ-καναλικής βάσης με πραγματικές ηχογραφήσεις ακουστικών γεγονότων από έναν «έξυπνο» χώρο συσκέψεων (UPC-TALP corpus) [78]. Τα αποτελέσματα δείχνουν ότι με τον κατάλληλο συνδυασμό της πολυ-καναλικής πληροφορίας μπορεί κανείς να πάρει σημαντικά βελτιωμένη απόδοση συγκριτικά με την χρήση ενός μόνο μικροφώνου.

- Το Κεφάλαιο 4 επεκτείνει την δουλειά μας στο πιο απαιτητικό σενάριο των επικαλυπτόμενων γεγονότων, εστιάζοντας σε μεθόδους NMF για την μονο-καναλική περίπτωση. Στο πρώτο μέρος, εστιάζουμε στην βελτίωση του σταδίου εντοπισμού γεγονότων στα πλαίσια της μεθόδου CNMF. Τα πειράματα, για τα οποία χρησιμοποιήθηκε η πραγματική βάση από τον ερευνητικό διαγωνισμό DCASE'16 Task 2 [79], δείχνουν σημαντική βελτίωση στον εντοπισμό γεγονότων με την προτεινομενη μέθοδο έναντι συμβατικών σχημάτων εντοπισμού με χρήση της CNMF μεθόδου. Στο δεύτερο μέρος, εξετάζουμε την περίπτωση συνδυασμού NMF μεθόδων με ταξινομητές. Δείχνουμε τα προβλήματα που μπορεί να προκύψουν σε περιπτώσεις επικάλυψης, και προτείνουμε μια μέθοδο για την μετρίαση αυτών των προβλημάτων. Για τα πειράματά μας χρησιμοποιούμε δύο βάσεις, μια με συνθετικά γεγονότα, και μια με πραγματικά ακουστικά γεγονότα (υποσύνολο της βάσης DCASE'16 Task 2).

- Το Κεφάλαιο 5 επεκτείνει τις NMF μεθόδους του προηγούμενου κεφαλαίου για τον εντοπισμό επικαλυπτόμενων γεγονότων, στην περίπτωση που διαθέτουμε πολλαπλά μικρόφωνα. Συγκεκριμένα, προτείνουμε μια πολυ-καναλική μεθόδο NMF η οποία βασίζεται στην ελαχιστοποίηση μιας καινοτόμου αντικειμενικής συνάρτησης που περιέχει έναν πολυ-καναλικό όρο αραιότητας (sparsity term). Τα πειράματα διεξάγονται στην πολυκαναλική πραγματική βάση ATHENA [80] που περιέχει ηχογραφήσεις από ένα «έξυπνο» χώρο συσκέψεων. Τα αποτελέσματα επιβεβαίωνουν την ανωτερότητα της προτεινόμενης μεθόδου έναντι κλασσικών μονο-καναλικών μεθόδων NMF ή απλούστερων πολυκαναλικών NMF μεθόδων.

- Το Κεφάλαιο 6 παρουσιάζει την μέθοδο που αναπτύξαμε βασιζόμενοι σε τεχνικές βαθιάς μάθησης για τον εντοπισμό επικαλυπτόμενων ακουστικών γεγονότων σε πολυ-καναλικά περιβάλλοντα, στην περίπτωση που ο αριθμός των πιθανών γεγονότων είναι μεγάλος. Το προτεινόμενο σύστημα συνδυάζει και εκπαιδεύει από κοινού ένα νευρωνικό δίκτυο διαχωρισμού ακουστικών συμβάντων με ένα δίκτυο ταξινόμησης γεγονότων. Για τα πειράματα χρησιμοποιήθηκε μια συνθετική βάση η οποία δημιουργήθηκε συνδυάζοντας πραγματικές ηχογραφήσεις 50 διαφορετικών ακουστικών γεγονότων από την βάση ESC50 [81] με πραγματικές κρουστικές αποκρίσεις από το «έξυπνο» σπίτι του ερευνητικού έργου DIRHA. Τα αποτελέσματα δείχνουν ότι σε δύσκολα σενάρια επικάλυψης, ο προτεινόμενος συνδυασμός μπορεί πράγματι να πετύχει σημαντικές βελτιώσεις έναντι νευρωνικών δικτύων εκπαιδευμένων με τον παραδοσιακό τρόπο, δεδομένου όμως ότι η αντήχηση στο ακουστικό περιβάλλον δεν είναι υψηλή.

- Τέλος, το Κεφάλαιο 7 συνοψίζει την ερευνητική μας δουλειά και προτείνει πιθανές μελλοντικές κατευθύνσεις για το πρόβλημα του εντοπισμού ακουστικών γεγονότων.

Συμπεράσματα και μελλοντική έρευνα

Γενικά ο εντοπισμός ακουστικών γεγονότων είναι μια αναπτυσσόμενη ερευνητική περιοχή, ενώ διάφορες παραλλαγές του προβλήματος έχουν αποτελέσει το αντικείμενο πολλών ερευνητικών

διαγωνισμών στην βιβλιογραφία τα τελευταία χρόνια [11]. Στην Διδακτορική μας διατριβή, δώσαμε έμφαση κυρίως στις κατευθύνσεις της πολυ-καναλικής επεξεργασίας και των σεναρίων με επικαλυπτόμενα ακουστικά γεγονότα.

Σχετικά με πιθανές προεκτάσεις της δουλειάς μας, πιστεύουμε ότι οι παρακάτω κατευθύνσεις μπορούν να φανούν εποικοδομητικές. Αρχικά, σχετικά με το σύστημα χωρο-χρονικού εντοπισμού φωνής, σκοπεύουμε να βελτιώσουμε περαιτέρω την απόδοση εισάγοντας στοιχεία από βαθιά μηχανική μάθηση (π.χ. αντικατάσταση των GMM ταξινομητών με CNNs). Επιπλέον θεωρούμε ως καλή κατεύθυνση τον σχεδιασμό ενός συστήματος για χωρο-χρονικό εντοπισμό επικαλυπτόμενων ακουστικών γεγονότων, επεκτείνοντας έτσι την δουλειά μας στις διεπαφές «έξυπνων» σπιτιών στην περίπτωση γενικότερων ακουστικών συμβάντων.

Όσον αφορά στο θέμα των ιδιαίτερα απαιτητικών σεναρίων επικάλυψης (μεγάλος αριθμός από πιθανές κατηγορίες γεγονότων, ή μεγάλος βαθμός πολυφωνίας στις επικαλύψεις), σκοπεύουμε να ερευνήσουμε επιπλέον αρχιτεκτονικές νευρωνικών δικτύων για καλύτερο συνδυασμό των λειτουργιών του διαχωρισμού και εντοπισμού γεγονότων, όπως και να πειραματισθούμε με χωρικά κατανεμημένες συστοιχίες μικροφώνων οι οποίες ενδεχομένως να αντιμετωπίσουν πιο αποτελεσματικά το πρόβλημα της αντήχησης.

Chapter 1

Introduction

Acoustic event detection (AED) constitutes a major part of the computational auditory analysis field, a research topic that has recently attracted significant interest in the literature. Typical applications of AED include smart home environments [1–5], multimedia indexing and retrieval [7], monitoring for healthcare [6], and security and surveillance systems [8, 9].

The main goal of AED is the automatic end-pointing and classification of each sound event present in an audio clip, revealing information about human or other activity. The "sound event" term refers to the audio part of any meaningful event with a noticeable acoustic impact. Depending on the application of interest and the corresponding environment, there can exist a large variety of possible acoustic events [10, 11]. In our work, we focus on acoustic events that can usually occur inside smart-space environments, primarily domestic ones. Some examples of such acoustic events can be "speech", "walking", "radio music", "door knocking", "keyboard typing", etc. Of particular interest is the acoustic event of speech, as it constitutes the primary mode of human-to-human and human-to-machine communication, and therefore plays a significant role in many applications. Hence, we focus part of our work on Speech Activity Detection (SAD), as a special case of the general AED problem.

SAD focuses on detecting the time boundaries of human speech present in an audio recording. In our work, we specifically focus on the SAD task in the context of voice-enabled systems for smart-home environments. Such systems typically contain a sequence of modules in their architecture, with SAD being a crucial one, as it provides input to other pipeline components, for example, speaker localization, speech enhancement, keyword spotting, and automatic speech recognition (ASR) [82–84], as well as contributing to the timing of the dialog management [85]. Further to voice-based interaction, SAD has found additional applications, such as telecommunications [34,35,86], variable-rate speech coding [87], and voice-based speaker recognition [45,88], among others.

In the rest of this chapter, we overview related work in the SAD and general AED fields. Then, we list the primary contributions of our work, and we overview the structure of this thesis.

1.1 Related work

1.1.1 Acoustic event detection

In the literature, several approaches have been proposed over the last years for the task of AED, varying in the algorithms and acoustic features employed. We can discriminate them depending on whether they were designed and evaluated for the isolated or the overlapped AED case, as well as on whether they employ single or multiple microphones in their setup. In the case of isolated AED, a number of conventional detection and classification approaches, such as ones based on hidden Markov models (HMMs) in conjunction with traditional audio features (for example Mel-frequency cepstral coefficients (MFCCs)), can achieve satisfactory performance [2].

In the case of the more challenging overlapping AED however, different approaches need to be employed in order to allow multiple event detection. For example, in [12], multiple-path Viterbi decoding is proposed to deal with the overlapping scenario. Other works for overlapping AED include multi-label deep neural networks (DNNs) [13], temporally-constrained probabilistic component analysis models [14], generalized Hough-transform based systems [15], and non-negative matrix factorization (NMF) [16].

Among such methods, NMF-based approaches and their variants have attracted significant interest in the field of both isolated and overlapping AED in recent years. This is due to both their robustness and their natural ability to detect multiple events occurring simultaneously, as far as appropriate non-negative and linear representations of them are available. NMF-related methods can be separated in those that exploit the NMF activations directly to perform event detection [16, 17], and in those that employ a classifier trained on these activations [18,19]. In [16], after building a quite large NMF-dictionary, NMF activations are directly exploited to perform detection for each event class. Regarding classifier-based NMF methods, in [18], a rather small dictionary of events is automatically built using sparse convolutive NMF (CNMF), and subsequently the activations produced are used as input for HMM training of each class. Also, in [20], exploiting the fact that NMF-based approaches can benefit from the creation of a Mixture of Local Dictionaries (MLD) [21], the authors propose a classifier-based NMF system using MLDs for improved detection performance.

Although NMF methods present a natural choice for overlapping AED scenarios, they have some disadvantages too, mainly including running-time efficiency and discriminative capability (when it comes to a large number of event classes). When enough training data are available, deep-learning based approaches can achieve superior performance in AED tasks [22], due to their better discriminative power. In general, several deep-learning based methods have been successfully proposed in recent years, including DNNs [23], convolutional neural networks (CNNs) [24], convolutional recurrent neural networks [22], and transformers [25].

All aforementioned approaches have been primarily applied to single-channel AED. However, whenever available, exploiting information from multiple channels can be valuable. In [2] various channel fusion methods were proposed within an HMM-based framework, while in [26] bag-of-words based features from different channels were used to train a global random forest classifier. Regarding neural network based methods, in [27] multi-channel exploitation was performed either

by feeding the network with inputs from multiple channels or by extracting multi-channel spatial features. In NMF related approaches, multi-channel extensions have also been considered, but mostly targeting blind source separation [28–32].

1.1.2 Speech activity detection

Similarly to AED, SAD has also been a topic of intense research activity, with numerous algorithms proposed in the literature over more than four decades, as for example overviewed in [33]. Some of the most established methods include algorithms incorporated into standards [34, 35], the statistical model-based approach by Sohn et al. [36], and the spectral divergence proposed by Ramírez et al. [37], among others. Typically, SAD methods extract various features from the waveform that are, for example, related to energy or zero-crossing rate [34, 35, 38, 39], harmonicity and pitch [40–42], formant structure [34, 43–45], degree of stationarity of speech and noise [46–48], modulation [49–51], or MFCCs [45]. Feature extraction is subsequently followed by traditional statistical modeling, or, more recently, by deep-learning based classifiers, for example DNNs [52, 53], recurrent neural networks [54, 55], or CNNs [56–58], often in conjunction with autoencoders [59]. Further, end-to-end deep-learning approaches applied directly to the raw signal have also been proposed [60].

Specifically for the smart-home domain, several SAD systems have been developed over the last decade, following the collection of appropriate corpora in domestic environments [61–65]. For example, in [66], linear-frequencies cepstral coefficients are employed as features in conjunction with Gaussian mixture model (GMM) and HMM classifiers to detect distressed speech or acoustic events inside a smart apartment for elderly persons. In a similar task under the Sweet-Home project in [67], sound event detection is first performed by discrete wavelet transform features and an adaptive thresholding strategy, followed by speech/event classification using support vector machines (SVMs) with GMM supervectors based on MFCCs. In [68], a simple energy-based SAD precedes the HMM-based recognition of sounds and spoken words. In [69], SAD is performed on headset microphone audio to track human behavior inside a smart home, with the proposed system employing an energy detector and a neural network trained on linear predictive coding coefficients and band-crossing features.

The aforementioned SAD systems aim to detect speech activity over the entire smart home, without however considering its typical multi-room layout. Only few recent approaches in the literature focus on the task of room-localized SAD in multi-room domestic environments that constitutes the focus of our work on SAD, yielding a speech/non-speech segmentation for each individual room of the smart home.

The majority of such systems operate in two stages. Typically, the first stage generates speech segment hypotheses over the entire home or for each specific room, which are further examined, refined, and assigned to the proper room at a second stage. Specifically, in [70], at the first stage of the proposed algorithm, DNN-based single-channel SAD is performed in each room. Then, at the second stage, for each detected speech segment, signal-to-noise ratio (SNR) and coherence-based features are extracted from all rooms and concatenated to feed a linear discriminant analysis

classifier that yields the segment room allocation. In [71], at the first stage, statistical-based SAD is performed for each microphone, and then majority voting over the room microphones provides the speech segments of each room. At the second stage, speaker localization output feeds a classifier (SVM or neural network) to further examine speech segments and delete those originating in other rooms. In [72], at the first stage, multi-layer perceptrons are employed for each microphone, and speech/non-speech segmentation is achieved via majority voting for each room. Then, in case of segments assigned to multiple rooms, a speech envelope distortion measure is employed to decide the correct room. In [73], three different features are investigated for room-localized SAD, namely SNR, periodicity, and the global coherence field. Speech boundaries for each room are computed by simple thresholding of these feature values and by using a heuristic rule over consecutive active frames.

In addition to the above, single-stage approaches have also been pursued for room-localized SAD. Specifically, in [74], a DNN is employed taking as input 176-dimensional vectors composed of a variety of features, such as MFCCs, RASTA-PLPs, envelope variance, pitch, etc. Similar features (but 187-dimensional) and DNNs are again considered in [75], as well as alternative classifiers, including a 2D-CNN. The latter is extended to a multi-channel 3D-CNN system in [76], where Log-Mel filterbank energies (40-dimensional) are employed as features, temporal context is exploited by concatenating adjacent time-frames, and the resulting 2D single-microphone feature matrices are stacked across channels. Finally, in [77], the aforementioned 3D-CNN is combined with the generalized cross-correlation (GCC-PHAT) [89] based CNN of [90] to yield a joint SAD and speaker localization network.

1.2 Contribution of this Thesis

In our work on AED we consider several variants of the task, including isolated and overlapped event scenarios, as well as single-channel or multi-channel setups available in the domestic smart-space environment (as depicted in Figure 1.1). Under this framework, we develop and evaluate several different approaches for AED. In particular, at first we study the isolated AED problem, providing several ways for combining the information from multiple microphones, and developing a multi-channel statistical based system with improved performance compared to single-channel baselines. Then, in the case of overlapped AED scenarios, we employ NMF methods, and by examining several variants of them, we propose methods for improved detection, for multi-channel expansion of the existing baselines, as well as modifications on state-the-art methods that combine NMF with classifiers, in order to improve their performance under highly overlapped conditions. Finally, we employ deep learning methods for overlapping AED when the number of event classes is large. In this direction, we propose the combination and joint training of a multi-channel sound source separation network with a multi-label AED network. Under moderate reverberation conditions, our proposed pipeline achieves significant improvements over traditional neural network approaches. For the evaluation of our proposed methods, suitable smart-space datasets, both real and simulated, are employed.

Regarding SAD, under the framework of smart-space environments, we develop an enriched

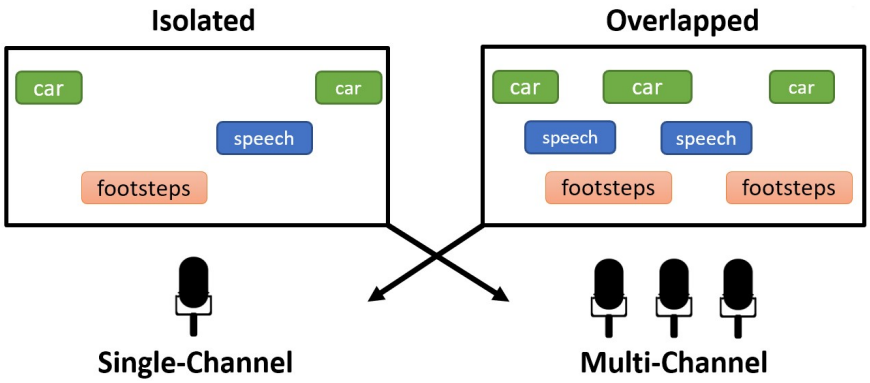


Figure 1.1: Different variants of the AED task regarding the number of available microphones, and the existence of overlap (or not) between the events. In our work, we considered all the possible combinations of scenarios.

SAD module named "room-localized" SAD, which is able to provide both the time boundaries of speech events ("when") and the coarse speaker position ("where") at the room level. This module can facilitate the communication of multiple speakers in different rooms with the smart-home voice interface. In addition, to achieve robust performance in the challenging acoustic conditions of a real smart space, our system takes advantage of the multiple microphones installed in it. The system follows a two-step approach, with the first step being a multi-channel statistical based module giving as output the temporal segmentation of speech. The second step, employing novel multi-channel based, hand-crafted features, provides the spatial intelligence of our SAD system, localizing the speaker at room level. The proposed approach is extensively evaluated on both simulated and real data recorded in a multi-room, multi-microphone smart home, significantly outperforming alternative baselines. Further, it remains robust to reduced microphone setups, while also comparing favorably to deep-learning based alternatives.

1.3 Structure of this Thesis

The rest of this Thesis is structured as follows:

- Chapter 2 presents our research on the SAD task for smart-space environments, providing details for our proposed systems, alongside experimental comparisons with alternative baselines and state-of-the-art systems.
- Chapter 3 presents our work on the isolated AED task, focusing on multi-channel fusion methods.
- Chapter 4 extends our work to the more challenging overlapped AED task, focusing on NMF-based approaches for the single-channel case.
- Chapter 5 extends our NMF-based approach of Chapter 4 to the multi-channel overlapped AED case.

- Chapter 6 presents our deep-learning based work on overlapped AED in multi-channel scenarios with a large number of possible event classes.
- Chapter 7 concludes the Thesis by summarizing our research work and discussing possible future efforts on the AED task.

Chapter 2

Speech Activity Detection in Multi-room Smart Spaces

2.1 Introduction

Smart-home technology has been attracting increasing interest lately, mainly in assistive scenarios for the disabled or the elderly, but also in "edutainment", home monitoring, and automation applications, among others [91–95]. Given that interaction with users must be convenient and natural, and motivated by the fact that speech constitutes the primary means of human-to-human communication, voice-enabled interaction systems have been progressively entering the field. Indeed, multiple smart-home projects have been focusing on voice-based interaction [96–103], and a number of commercial voice-assistant home devices have recently been introduced in the market [104].

In practice, domestic environments contain multiple rooms, where one or more users may be located wishing to interact with the smart-home voice interface. This scenario can be facilitated if the SAD module provides not only time boundaries of speech events ("when"), but also coarse speaker position ("where") at the room level, i.e., assigning room "tags" to the detected speech activity, thus yielding separate speech/non-speech segmentation outputs, one per room of the smart home (see also Figure 2.1). Enriching the traditional "room-independent SAD" to such "room-localized SAD" variant can be useful in multiple ways: It can help disambiguate user commands for voice-control of devices or appliances present in multiple rooms (e.g., light switches, windows, temperature control units, television sets, etc.); enable room-localized system feedback, for example via a loudspeaker or visual display at the room where speech activity takes place; and allow parallel voice interaction sessions by multiple subjects inside different rooms, engaging separate system pipelines, one per room [83]; finally, ASR itself can benefit significantly from room localization [70].

Designing a robust SAD system in domestic environments is a hard task due to the challenging acoustic conditions encountered. Such involve speech at low SNR, presence of reverberation, and multiple background noise sources often overlapping with speech activity. In the case of room-localized SAD, these difficulties are further exacerbated due to acoustic interference be-

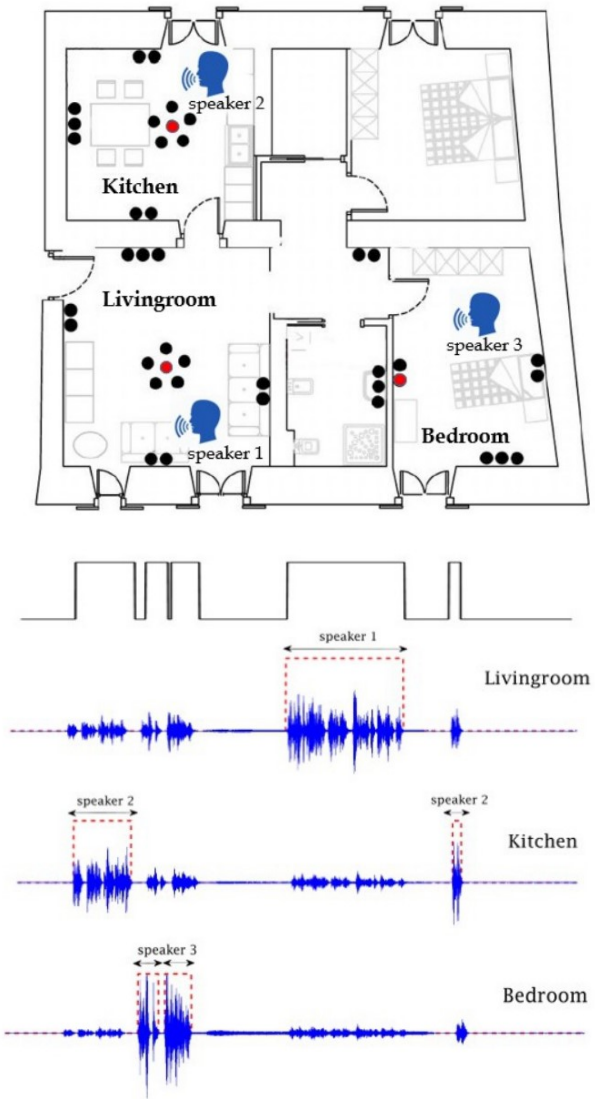


Figure 2.1: An example of room-independent vs. room-localized SAD in multi-room domestic environments equipped with multiple microphones. Here, three speakers are active in three rooms. *Top:* Floorplan of the smart home used in the DIRHA project [100] (see also Section 2.6.1 and Figure 2.6), with dots indicating microphone locations on the apartment walls and ceiling. *Bottom:* 1-minute long waveforms, captured by the red-colored microphones (one per room with an active speaker), shown together with the corresponding ground-truth of room-localized SAD. The room-independent speech/non-speech segmentation is also depicted at the top.

tween rooms. To counter these challenges, smart homes typically employ multiple microphones to capture the acoustic scene and "cover" the large multi-room interaction area. This allows exploiting multi-channel processing techniques, for example fusion of the microphone information at the signal, feature, or decision level, in order to facilitate the analysis of the acoustic scene of interest.

Several efforts have been reported recently on room-localized SAD in multi-room environments [70–77]. As already overviewed in Section 1.1.2, these approaches vary in the kind of features, classifiers, and number of microphones used per room. Depending on their design, they

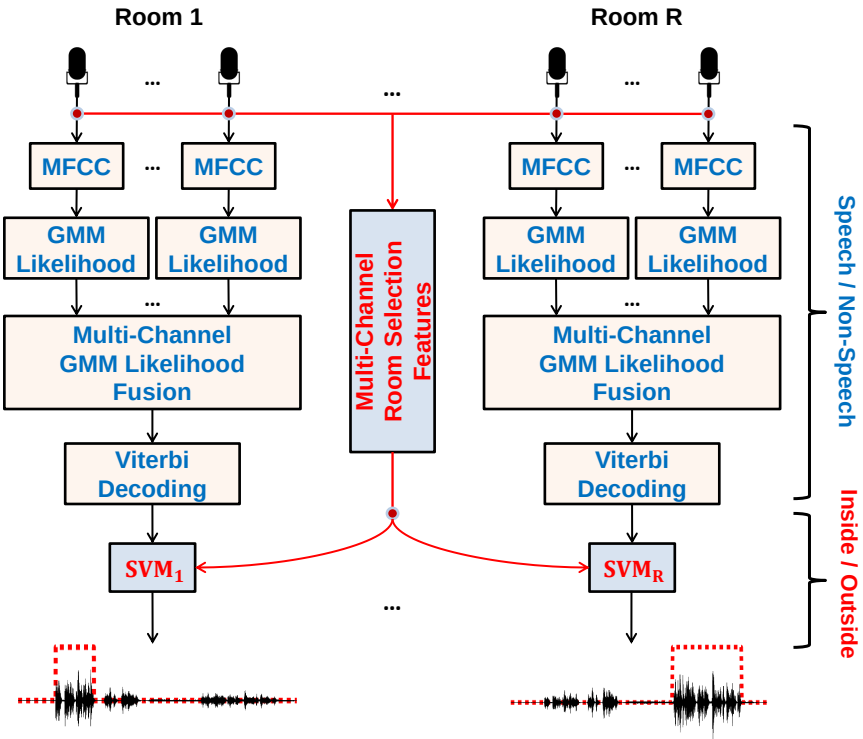


Figure 2.2: Block-diagram of the proposed room-localized SAD system. The first-stage algorithmic components are depicted in blue color and the second-stage ones in red.

typically consist of one or two algorithmic stages, and may or not allow the detection of simultaneously active speakers located in different rooms.

In this chapter, we investigate a room-localized SAD system for smart homes equipped with multiple microphones distributed in multiple rooms. The system employs a two-stage algorithm, incorporating a set of hand-crafted features specially designed to discriminate room-inside vs. -outside speech at its second stage, refining SAD hypotheses obtained at its first stage by traditional statistical modeling and acoustic front-end processing. Both algorithmic stages exploit multi-microphone information, combining it at the signal, feature, or decision level. The proposed approach is extensively evaluated on both simulated and real data recorded in a multi-room, multi-microphone smart home, significantly outperforming alternative baselines. Further, it remains robust to reduced microphone setups, while also comparing favorably to deep-learning based alternatives.

The remainder of the chapter is organized as follows: The overview of the proposed system is provided in Section 2.2, with its two algorithmic stages further detailed in Sections 2.3 and 2.4; alternative baselines are presented in Section 2.5; the datasets and experimental framework are discussed in Section 2.6; the evaluation is reported in Section 2.7; and, finally, conclusions are drawn in Section 2.8.

2.2 Notation and system overview

Let us denote by R the number of rooms inside a given smart home that is equipped with a set of microphones $\mathcal{M}_{\text{all}}$. This is partitioned into subsets $\mathcal{M}_r$, for $r = 1, 2, ..., R$, each containing the microphones located inside room r. Let us also denote by $\mathbf{o}_{m,t}$ the short-time acoustic feature vectors (e.g., MFCCs) extracted from the signal of microphone m, and by $\mathbf{o}_{\mathcal{M},t}$ their concatenation over microphone set $\mathcal{M} \subseteq \mathcal{M}_{\text{all}}$, with t indicating time indexing at the frame level (typically at a 10 ms resolution).

We are interested in room-localized SAD, seeking speech/non-speech segmentations for each room r, detecting speech events occurring inside it but ignoring speech originating in other rooms or any other non-speech events. As also shown in Figure 2.1, this differs from room-independent SAD, where a single speech/non-speech segmentation is produced, including speech events occurring inside any of the R rooms of the smart home.

As already discussed in the previous sections and also depicted in the block diagram of Figure 2.2, our proposed system for room-localized SAD operates in two stages. The first stage, detailed in Section 2.3, is based on single-channel GMM classifiers, each trained on an individual room microphone, employing MFCC features and operating at the frame level. An appropriate decision fusion scheme follows, combining GMM likelihood scores across all room microphones and, by means of Viterbi decoding, providing a crude speech/non-speech segmentation for the given room. Then at the second stage, presented in detail in Section 2.4, for the speech segments detected for each room, an SVM classifier is employed on a number of hand-crafted room-localization features, specially designed to discriminate room-inside vs. room-outside speech. Various feature fusion schemes across rooms are considered for this purpose, accompanied by different options for their SVM-based modeling.

2.3 First stage: speech segment generation

We now proceed with a detailed description of the first stage of the developed room-localized SAD system. This stage generates individual speech/non-speech segmentations for every room using the specific room microphones only, thus providing initial room-localized SAD hypotheses to be refined later. To accomplish this, it employs traditional acoustic front-end processing and statistical modeling at the microphone level as discussed in Section 2.3.1, followed by decision fusion across microphones as detailed in Section 2.3.2, and appropriate decoding schemes that are presented in Section 2.3.3. Variations on the choices of microphones and classes considered are discussed in Section 2.3.4.

2.3.1 Single-microphone system core

At the core of the system lies the single-microphone speech/non-speech modeling. Specifically, for each microphone of the smart home, a traditional 39-dimensional MFCC-plus-derivatives acoustic front-end is employed, with features extracted over 25 ms Hamming-windowed signal frames with a 10 ms shift. Subsequently, two-class microphone-specific GMMs are trained on these features

(32 Gaussian mixtures with diagonal covariance matrices are used in our implementation), with the set of classes being $\mathcal{J} = \{\mathrm{sp}_r, \mathrm{sil}_{\mathrm{all}}\}$, where sp_r denotes speech originating in room r where the given microphone is located, and $\mathrm{sil}_{\mathrm{all}}$ indicates lack of speech in all rooms. Alternative class choices for set $\mathcal{J}$ are discussed in Section 2.3.4.

2.3.2 Multi-microphone decision fusion

The developed system performs multi-microphone fusion at the decision level, where the GMM log-likelihood scores of different channels are combined at the frame level for each class of interest, potentially also incorporating channel decision confidence. In particular, the following approaches for decision fusion over microphone set $\mathcal{M} \subseteq \mathcal{M}_{\mathrm{all}}$ are considered, which were investigated in our work in [2], but for room-independent SAD only:

• **Log-likelihood summation**, where the fused log-likelihoods (log class-conditionals) at frame t become

$$c_{\mathcal{M},j}(\mathbf{o}_{\mathcal{M},t}) \;=\; \sum_{m \in \mathcal{M}} w_{m,t}\, b_{m,j}(\mathbf{o}_{m,t}) \;, \tag{2.1}$$

where $b_{m,j}(\mathbf{o}_{m,t})$ denotes the log-likelihoods of the GMMs for microphone m given its acoustic features $\mathbf{o}_{m,t}$ at time frame t, and class $j \in \mathcal{J}$. The individual microphone scores in (2.1) can be uniformly weighted by setting $w_{m,t} = 1 \,/\, |\mathcal{M}|$ (where $|\bullet|$ denotes set cardinality), in which case the scheme will be referred to as *unweighted log-likelihood summation* ("u-sum"), or adaptively weighted at any given time frame t, according to channel decision confidence that is estimated as

$$w_{m,t} \;=\; \frac{|\; b_{m,\,\mathrm{sp}_r}(\mathbf{o}_{m,t}) \;-\; b_{m,\,\mathrm{sil}_{\mathrm{all}}}(\mathbf{o}_{m,t}) \;|}{\displaystyle\sum_{m' \in \mathcal{M}} |\; b_{m',\,\mathrm{sp}_r}(\mathbf{o}_{m',t}) \;-\; b_{m',\,\mathrm{sil}_{\mathrm{all}}}(\mathbf{o}_{m',t}) \;|} \;, \tag{2.2}$$

in which case the method will be termed *weighted log-likelihood summation* ("w-sum"). Weighting by (2.2) is motivated by intuition that large log-likelihood differences between the classes imply higher classification confidence.

• **Log-likelihood selection**, where, at each time frame t, a microphone $\hat{m}_t \in \mathcal{M}$ is selected to provide all fused class log-likelihoods, i.e.,

$$c_{\mathcal{M},j}(\mathbf{o}_{\mathcal{M},t}) \;=\; b_{\hat{m}_t,j}(\mathbf{o}_{\hat{m}_t,t}) \;, \quad \text{for all} \;\; j \in \mathcal{J} \;. \tag{2.3}$$

Such microphone can be chosen as the one achieving the highest frame log-likelihood over all channels and over all classes, i.e.,

$$\hat{m}_t \;=\; \arg\max_{m \in \mathcal{M}} \{\; \max_{j \in \mathcal{J}} \; b_{m,j}(\mathbf{o}_{m,t}) \;\} \;,$$

in which case the scheme will be referred to as *log-likelihood maximum selection* ("u-max"), or as the channel with the highest confidence (2.2), i.e.,

$$\hat{m}_t \;=\; \arg\max_{m \in \mathcal{M}} w_{m,t} \;,$$

in which case the method will be termed *log-likelihood confidence selection* ("w-max").

• **Majority voting**, where, at each time frame t, single-channel decisions, computed as $\hat{j}_{m,t} = \arg\max_{j\in\mathcal{J}} b_{m,j}(\mathbf{o}_{m,t})$, are accumulated over microphone set $\mathcal{M}$, and the class with the highest decision incidence is chosen. Such accumulation can be computed uniformly over the channels, in which case the scheme will be termed *unweighted majority voting* ("u-vote"), or scaled by means of (2.2), resulting in *weighted majority voting* ("w-vote").

Among the above approaches, based on the experimental results of Section 2.7, the developed room-localized SAD system employs the "w-sum" scheme computed over the set of microphones inside one room at a time, i.e., $\mathcal{M} = \mathcal{M}_r$. Alternative choices for set $\mathcal{M}$ are discussed in Section 2.3.4.

2.3.3 Speech/non-speech segmentation

Following GMM training and multi-channel fusion, two speech detection implementations are developed: The first operates on mid-sized sliding windows, thus resulting in low latency, whereas the second performs Viterbi decoding over longer sequences, providing superior accuracy (as demonstrated in Section 2.7), but being more suitable for off-line processing.

• **GMM-based scoring over sliding window**: This scheme performs sequential classification over sliding windows of fixed duration and overlap (400 ms and 200 ms, respectively, are used). Specifically, for a given time-window $\mathcal{T} = [\, t_s \,, t_e \,]$ and microphone m, the log-likelihoods for each class $j \in \mathcal{J}$ are first computed by adding all frame scores within the window. This results in scores $b_{m,j}(\mathbf{o}_{m,\mathcal{T}}) = \sum_{t=t_s}^{t_e} b_{m,j}(\mathbf{o}_{m,t})$, where $\mathbf{o}_{m,\mathcal{T}}$ denotes all feature vectors within window $\mathcal{T}$. Microphone fusion is then carried out as in Section 2.3.2, but employing the window log-likelihoods instead.

• **HMM-based Viterbi decoding over sequence**: In this scheme, HMMs are built with a set of fully connected states $\mathcal{J}$, state transition probabilities $\{\, a_{jj'} \,,$ for $j \,, j' \in \mathcal{J} \,\}$, and class-conditional observation probabilities provided by the class GMMs of Section 2.3.1. Then, Viterbi decoding is performed over an entire sequence of observations (in our data, such are of 1 min length, as discussed in Section 2.6.1), in order to provide the desired speech/non-speech segmentation. Specifically, for the single-microphone case, the well-known recursion [105]

$$\delta_{m,j}(t) \;=\; \max_{j'}\{\delta_{m,j'}(t-1) \;+\; \log(a_{j'j})\} \;+\; b_{m,j}(\mathbf{o}_{m,t}) \;, \tag{2.4}$$

is used, where $\delta_{m,j}(t)$ denotes the score of the best decoding path ending at state j and accounting for the first t frame observations of microphone m. This can be readily extended to the fusion schemes of (2.1) and (2.3) over microphone set $\mathcal{M}$ as

$$\delta_{\mathcal{M},j}(t) \;=\; \max_{j'}\{\delta_{\mathcal{M},j'}(t-1) \;+\; \log(a_{j'j})\} \;+\; c_{\mathcal{M},j}(\mathbf{o}_{\mathcal{M},t}) \;, \tag{2.5}$$

whereas majority voting fusion schemes "u-vote" and "w-vote" are modified to be applied over best-path scores $\delta_{m,j}(t)$ instead of log-likelihoods $b_{m,j}(\mathbf{o}_{m,t})$.

Between the two aforementioned decoding schemes, the proposed system follows the HMM-

based approach due to its superior performance, with a number of fine-tuned parameters incorporated in it. Specifically, these are the state transition penalty that tunes the flexibility of the decoder to change states, as well as the speech class prior that favors or not the selection of the speech state.

2.3.4 Variations in sets of classes and microphones

As already discussed, to obtain the first stage of the speech/non-speech segmentation hypothesis for room r, only the particular room microphones are considered ($\mathcal{M} = \mathcal{M}_r$). A number of variations however are possible for the set of classes $\mathcal{J}$, which are investigated in the experiments of Section 2.7.2:

• $\mathcal{J} = \{\,\mathrm{sp}_r\,, \mathrm{sil}_{\mathrm{all}}\,\}$, where sp_r denotes speech inside room r, and $\mathrm{sil}_{\mathrm{all}}$ indicates absence of speech in all rooms of the smart home. This set is used in the proposed room-localized SAD algorithm.

• $\mathcal{J} = \{\,\mathrm{sp}_r\,, \mathrm{sil}_r\,\}$, where sil_r indicates absence of speech in room r. This set is used in our work in [106].

• $\mathcal{J} = \{\,\mathrm{sp}_r\,, \mathrm{sp}_{\bar{r}}\,, \mathrm{sil}_{\mathrm{all}}\,\}$, where $\mathrm{sp}_{\bar{r}}$ indicates speech inside any of the other rooms, excluding room r.

In addition, in [107], the first-stage of the algorithm provides room-independent SAD output. That system uses the "w-sum" decision fusion scheme with all smart-home microphones contributing to (2.1), i.e., $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$. Further, the set of classes employed is $\mathcal{J} = \{\,\mathrm{sp}_{\mathrm{all}}\,, \mathrm{sil}_{\mathrm{all}}\,\}$, where $\mathrm{sp}_{\mathrm{all}}$ denotes speech occurring in any of the smart-home rooms.

2.4 Second stage: room assignment

Following the generation of initial room-localized SAD hypotheses, the second stage of the developed algorithm performs the final selection of active segments for each room. For this purpose, five hand-crafted features are proposed as detailed in Section 2.4.1, extracted at the segment level for each room, and capable of segment discrimination as originating from inside vs. outside a given room. These features are then fused within and across rooms as presented in Section 2.4.2, and are fed to SVM classifiers that perform room assignment as detailed in Section 2.4.3, temporally operating on the given segment as discussed in Section 2.4.4. Various options for the above are presented.

2.4.1 Room discriminant features

As mentioned above, for any first-stage speech segment $\mathcal{T} = [\,t_s\,, t_e\,]$ starting at time-frame t_s and ending at frame t_e, segment-level features are extracted for each room. The design of these hand-crafted features is motivated by intuition concerning: (a) the energy; (b) the reverberation; and (c) the arrival direction of the microphone signals. For example, microphones located inside the room where a speech segment originates are expected to yield signals with higher energy and lower reverberation than microphones located outside it. Likewise, the room door region typically appears as the speech source for room-outside segments. In particular, five scalar features

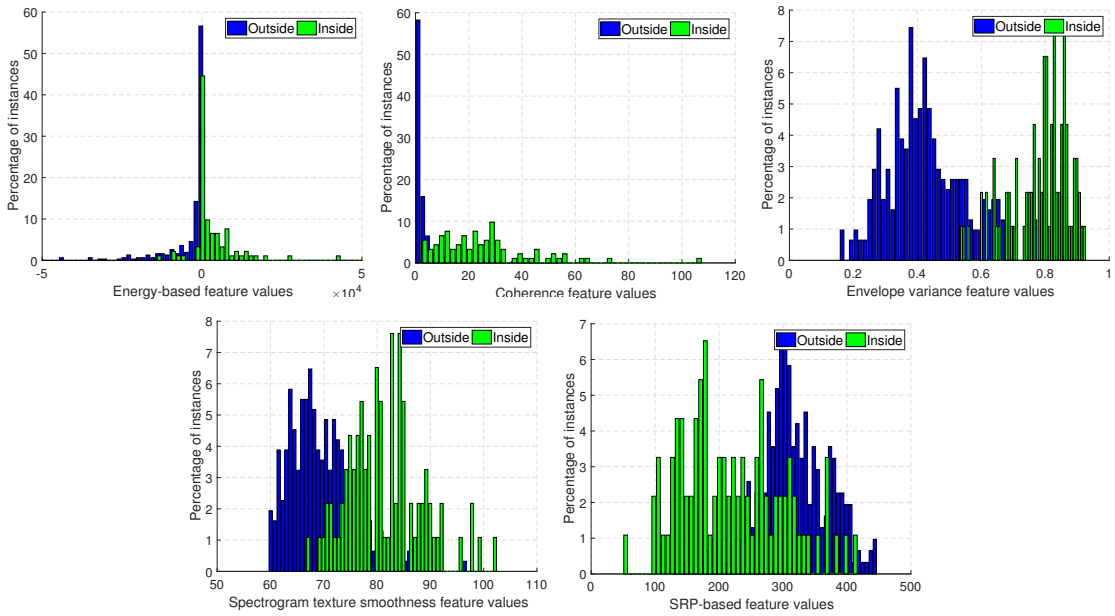


Figure 2.3: Histograms of the five hand-crafted scalar features of Section 2.4.1, demonstrating their ability to discriminate room-inside vs. room-outside speech. Histograms are computed over the development set of the simulated dataset of Section 2.6.1, for the case of the smart-home Bedroom (see also Figure 2.1). *Upper row, left-to-right:* Energy-based feature, coherence feature, and envelope variance one; *Lower row, left-to-right:* Spectrogram texture smoothness feature and SRP-based one.

that are specially designed to provide room-inside vs. -outside segment source discrimination are proposed, as also depicted in the histograms of Figure 2.3.

It should be noted that in contrast to the acoustic front-end of the first stage of the algorithm that extracts microphone-dependent features, the features of the second stage are instead room-dependent. Indeed, their estimation typically involves all microphones located in a room (or in the entire smart home), performing in a sense fusion of their information at the signal level. Such derivation requires of course knowledge of the microphone room-membership, but in the case of the coherence features also of additional information concerning which microphones lie adjacent to each other, and in the case of the SRP-based features further knowledge of the microphone topology and room layout. Details are provided next.

• **Energy-based feature**: Originally proposed in [106], this feature is motivated by intuition that microphones inside the room where speech activity occurs will exhibit, on average, higher SNRs compared to ones outside it. For its computation, given detected speech segment $\mathcal{T}=[\,t_s\,,t_e\,]$, the energy ratio (ER) of speech over non-speech is first computed for all smart-home microphones. For this purpose, the initial part of the speech segment, as well as the trailing part of non-speech preceding it, both of length $\Delta\tau$, are utilized to yield

$$\mathrm{ER}_{m,\mathcal{T}} \;=\; \left(\sum_{\tau=Lt_s}^{Lt_s+\Delta\tau-1} x_m(\tau)^2\right) \Bigg/ \left(\sum_{\tau=Lt_s-\Delta\tau}^{Lt_s-1} x_m(\tau)^2\right)\,, \tag{2.6}$$

for all microphones $m \in \mathcal{M}_{\text{all}}$. In (2.6), $x_m(\tau)$ denotes the signal captured by microphone m, with τ indicating indexing at the sample level. The latter is related to frame-level indexing by $\tau = L\,t$, where L is the number of signal samples over the short-time window shift. Following computations (2.6), the ERs are sorted across all smart-home microphones, and the microphone set with the K largest values is derived, denoted by $\mathcal{M}^{(K)}$. Finally, the desired energy-based feature for room r is extracted as the difference between the sum of the ERs of the microphones in set $\mathcal{M}^{(K)}$ that are located inside room r and the ER sum of the ones in $\mathcal{M}^{(K)}$ but located in other rooms, namely

$$f_{r,\mathcal{T}}^{(\text{en})} = \sum_{m \in \mathcal{M}^{(K)} \cap \mathcal{M}_r} \text{ER}_{m,\mathcal{T}} - \sum_{m \in \mathcal{M}^{(K)} \setminus \mathcal{M}_r} \text{ER}_{m,\mathcal{T}} ,$$

for all rooms $r = 1\,, 2\,, ...\,, R$. In our implementation, $K = 5$ and, in (2.6), $\Delta\,\tau$ corresponds to a 0.5 s interval.

• **Coherence feature**: Originally proposed in [70] and re-used in [106], this feature is motivated by intuition that signals captured by pairs of adjacent microphones located outside a speech-active room will exhibit higher reverberation and thus lower cross-correlation than pairs inside it. To compute the coherence feature for room r, the set of adjacent pairs of microphones inside the room is first determined, denoted by $\{\mathcal{M}_r \times \mathcal{M}_r\}_{\text{adj}}$. Such pairs typically consist of neighboring microphones in larger arrays (see also Section 2.6.1). Then, for every time-frame t within detected speech segment $\mathcal{T}$, the maximum cross-correlation of the signal frames of adjacent microphone pair (m, m') is computed, denoted by $\text{C}_{m,m'}(t)$. This is repeated for all pairs $(m, m') \in \{\mathcal{M}_r \times \mathcal{M}_r\}_{\text{adj}}$ and the maximum retained. Finally, the result is averaged over the entire segment $\mathcal{T}$, yielding the coherence feature for room r, as

$$f_{r,\mathcal{T}}^{(\text{coh})} = \underset{t \in \mathcal{T}}{\text{avg}} \left\{ \max_{\substack{(m,m') \\ \in \{\mathcal{M}_r \times \mathcal{M}_r\}_{\text{adj}}}} \text{C}_{m,m'}(t) \right\} .$$

Note that this feature employs the un-normalized cross-correlation function in order to also "capture" signal attenuation. In our implementation, signal cross-correlation is computed over fixed size sliding windows of 100 ms in length and a 25 ms shift.

• **Envelope variance feature**: Originally proposed in [108] for ASR channel selection and used in [72, 74, 106] for room-localized SAD, this feature is motivated by intuition that higher reverberation (indicative of room-outside speech) results in smoother short-time speech energy, also observed as reduced dynamic range of the corresponding envelope. To compute the envelope variance feature, we follow the derivations in [108]. Briefly, for each microphone m, the short-time filterbank energy, denoted by $X_m(n\,, t)$, is obtained for time-frames $t \in \mathcal{T}$, where, as above, $\mathcal{T}$ is the detected speech segment and n denotes the sub-band (20 linear filters are used here). Then

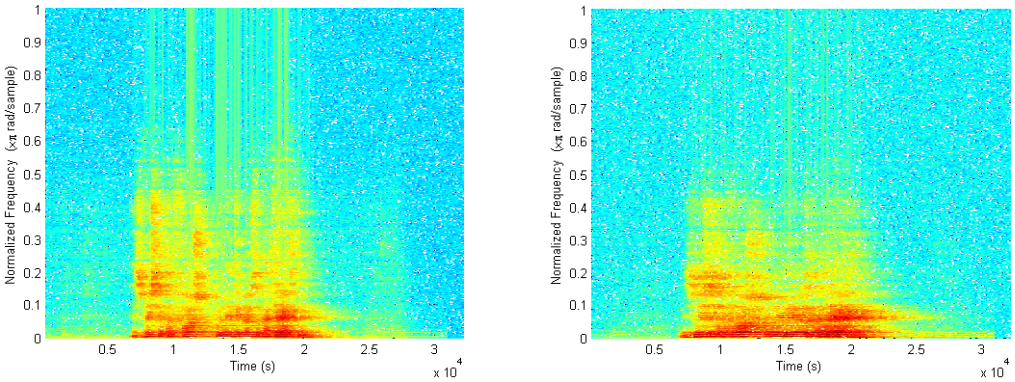


Figure 2.4: Motivation for the spectrogram texture smoothness feature. *Left:* Spectrogram from a microphone located inside the active-speaker room (Bedroom in the apartment of Figure 2.1); *Right:* Spectrogram from a microphone outside it (Kitchen).

the n^{th} sub-band envelope of microphone m is computed as

$$\hat{X}_m(n\,,t) \;=\; \exp\Big\{\,\log[X_m(n\,,t)] \;-\; \underset{t\,\in\,\mathcal{T}'}{\text{avg}}\big\{\,\log[X_m(n\,,t)]\,\big\}\Big\}\,,$$

where $\mathcal{T}'$ denotes medium-sized windows sliding over segment $\mathcal{T}$, the time-progression of which will be indexed by t' (600 ms long windows with a 50 ms shift are used). Then, the variance of each sub-band envelope is computed (following cube root compression) as

$$V_m(n\,,t') \;=\; \underset{t\,\in\,\mathcal{T}'}{\text{var}}\big\{\,\hat{X}_m(n\,,t)^{1/3}\,\big\}\,,$$

subsequently normalized over all smart-home microphones, and its average over all sub-bands obtained:

$$\text{EV}_m(\,t'\,) \;=\; \underset{n}{\text{avg}}\left\{\,\frac{V_m(n\,,t')}{\max\limits_{m'\,\in\,\mathcal{M}_{\text{all}}} V_{m'}(n\,,t')}\,\right\}\,. \tag{2.7}$$

In this work, we define the envelope variance feature of segment $\mathcal{T}$ for room r as the average over all mid-sized shifting windows within $\mathcal{T}$ of the maximum value of (2.7) over the set of all room microphones $\mathcal{M}_r$, i.e.,

$$f^{(\text{ev})}_{r,\mathcal{T}} \;=\; \underset{t'\,\in\,\mathcal{T}}{\text{avg}}\Big\{\,\max_{m\,\in\,\mathcal{M}_r}\text{EV}_m(\,t'\,)\Big\}\,. \tag{2.8}$$

• **Spectrogram texture smoothness feature**: For measuring the degree of reverberation, we propose an additional feature that is based on the "smearing" effect that reverberant conditions cause to the speech signal spectrogram. An example is shown in Figure 2.4: There, for a speech occurrence inside the Bedroom of the smart home of Figure 2.1, the spectrograms of two signals captured by a microphone located in the Bedroom and one in the Kitchen are depicted, showing that the latter (located outside the speech-active room) is much smoother (smeared). To measure this effect, the proposed feature considers the signal spectrogram as a 2D image, and attempts to quantify its texture smoothness by applying to it the 2D discrete Teager energy operator of [109],

yielding

$$\Phi_m(n\,,t) \;=\; 2\,\big(S_m(n\,,t)\big)^2 \;-\; S_m(n\,,t-1)\;S_m(n\,,t+1) \\ -\; S_m(n-1\,,t)\;S_m(n+1\,,t)\;,$$

where $S_m(n\,,t)$ denotes the signal spectrogram of microphone m at short-time frame $t \in \mathcal{T}$, and n is the frequency index (40 ms long Hamming windows with a 20 ms shift and 960 frequency bins are used here). Then, as for the envelope variance case, medium-sized windows $\mathcal{T}'$ sliding over segment $\mathcal{T}$ are considered, the time-progression of which is indexed by t' (600 ms long windows with a 50 ms shift are used). The values of $\Phi_m(n\,,t)$ are then averaged over a part of the resulting 960×30-sized spectrogram image, as

$$\Phi_m(\,t') \underset{n=1,\ldots,200}{=} \underset{t\,\in\,\mathcal{T}'}{\text{avg}}\,\big\{\,\Phi_m(n\,,t)\,\big\}\;,$$

where the frequency-domain averaging is carried out over the 200 lower-frequency bins that correspond to the 0–5 kHz frequency range of the 48 kHz-sampled signal, focusing on speech content. Finally, the spectrogram texture smoothness feature for room r and segment $\mathcal{T}$ is obtained by maximizing over all room microphones and averaging the result over all medium-sized windows, namely

$$f^{(\text{ts})}_{r,\mathcal{T}} \;=\; \underset{t'\in\mathcal{T}}{\text{avg}}\,\Big\{\,\max_{m\,\in\,\mathcal{M}_r}\Phi_m(\,t')\Big\}\;. \tag{2.9}$$

• **SRP-based feature**: The final feature considered for room assignment of detected speech segments is based on the steered-response-power (SRP-PHAT) approach of [110], and it is proposed for the first time in this research work for room-localized SAD. Employing SRP allows the creation of an acoustic map, by computing the signal power when steering microphone arrays in the direction of a specific location. The position of the sound source corresponds to that with the maximum SRP value over all possible locations. In the case of multi-room smart homes, one expects that speech originating from outside a given room will likely exhibit high SRP values at the door region that connects that room to the rest of the apartment. In contrast, for room-inside speech, the actual source location should yield the highest SRP instead. An example for this motivation is depicted in Figure 2.5. To compute the SRP-based feature for room r, a 3D region is first defined, denoted by $\mathcal{A}_r$, that corresponds to cylindrically-shaped volume(s) covering the room door(s). Specifically, on the floor plane, this lies inside room r, delineated by a 0.7 m radius semicircle around the door center, while also containing all points above it. Using a 10 cm spatial resolution for each dimension, and depending on the number of doors of the room, this scheme yields approximately between $2k$ and $4.3k$ points, denoted as $\vec{y} \in \mathcal{A}_r$, expressed in the 3D room coordinate system (see also Figure 2.5). Then, for all points $\vec{y} \in \mathcal{A}_r$, the corresponding SRP-PHAT values for time-frame $t \in \mathcal{T}$ are computed (200 ms long frames with a 100 ms shift are used), by summing the generalized cross-correlations over all pairs of adjacent microphones in room r, as

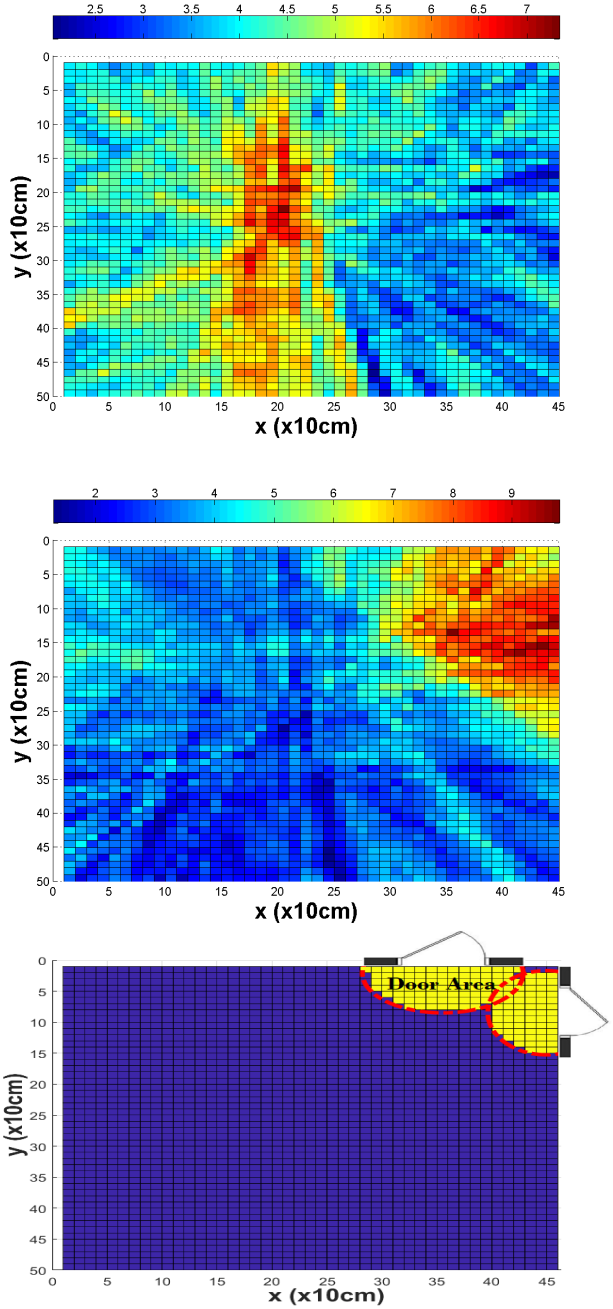


Figure 2.5: Motivation for the SRP-based feature (the acoustic maps are shown in 2D, obtained after summing SRP-PHAT values over the z-axis). *Top:* Acoustic map example for speech inside the Livingroom of the apartment of Figure 2.1. *Middle:* Acoustic map example for speech outside the Livingroom. *Bottom:* Livingroom door area (in yellow color) employed for the SRP-based feature computation (2.10) of this room.

$$\mathrm{P}_r(t\,,\vec{y}\,) = \sum_{\substack{(m,m')\\ \in\{\mathcal{M}_r\times\mathcal{M}_r\}_{\mathrm{adj}}}} \int_0^{2\pi} \frac{X_m(\omega,t)\; X^*_{m'}(\omega,t)}{\left|\, X_m(\omega,t)\; X^*_{m'}(\omega,t)\,\right|}\; e^{\,j\omega\tau_{mm'}(\vec{y}\,)} d\,\omega\ ,$$

where $X_m(\omega,t)$ denotes the DTFT of the m^{th} microphone signal frame, and $\tau_{mm'}(\vec{y}\,)$ is the

time-difference-of-arrival at point $\vec{y}$ between the signals of adjacent microphones m and m'. Finally, the SRP-based feature is computed by summing all above values and averaging them over all windows $t \in \mathcal{T}$, i.e.,

$$f_{r,\mathcal{T}}^{(\mathrm{srp})} \;=\; \underset{t\,\in\,\mathcal{T}}{\mathrm{avg}} \Big\{ \sum_{\vec{y}\,\in\,\mathcal{A}_r} \mathrm{P}_r(t\,,\vec{y}\,) \Big\}\,. \tag{2.10}$$

Clearly, computation of this feature requires knowledge of the microphone topology and room layout.

2.4.2 Intra- and inter-room feature fusion

Using the above framework in the proposed system, for each candidate speech segment $\mathcal{T}$, five features are extracted for each room r. The features are then combined by intra-room feature fusion (plain concatenation), resulting in 5-dimensional feature vectors

$$f_{r,\mathcal{T}}^{(\mathrm{all})} \;=\; \big[\, f_{r,\mathcal{T}}^{(\mathrm{en})},\; f_{r,\mathcal{T}}^{(\mathrm{coh})}, f_{r,\mathcal{T}}^{(\mathrm{ev})},\; f_{r,\mathcal{T}}^{(\mathrm{ts})},\; f_{r,\mathcal{T}}^{(\mathrm{srp})} \,\big]\,, \tag{2.11}$$

for each room $r = 1\,, 2\,, \ldots\,, R\,$.

In addition, inter-room feature fusion can be beneficial to room-inside vs. -outside speech discrimination. Two schemes are considered for this purpose:

• **Inter-room feature concatenation**, where vectors from all R rooms are concatenated, resulting in a single $5R$-dimensional feature vector for segment $\mathcal{T}$,

$$f_{\mathrm{home}\,,\mathcal{T}}^{(\mathrm{all})} \;=\; \big[\, f_{1\,,\mathcal{T}}^{(\mathrm{all})},\; f_{2\,,\mathcal{T}}^{(\mathrm{all})},\; \ldots\;, f_{R\,,\mathcal{T}}^{(\mathrm{all})} \,\big]\,. \tag{2.12}$$

• **Inter-room feature averaging**, where vectors from each room are augmented by the feature average across the remaining $R-1$ rooms, resulting in 10-dimensional representations of segment $\mathcal{T}$,

$$f_{r,\,\mathrm{avg}\,,\mathcal{T}}^{(\mathrm{all})} \;=\; \big[\, f_{r,\mathcal{T}}^{(\mathrm{all})}\,,\, \underset{r'\neq r}{\mathrm{avg}}\big\{ f_{r',\mathcal{T}}^{(\mathrm{all})} \big\} \,\big]\,, \tag{2.13}$$

for each room $r = 1\,, 2\,, \ldots\,, R\,$. This way, feature vector dimensionality is no longer a function of R. Alternatives to (2.13) can also be designed, for example employing feature extrema instead of averages.

2.4.3 SVM classification

The fused feature vectors are fed to appropriately designed classifiers, in order to determine the room of origin for a given segment. In our work, linear SVMs are employed for this purpose, due to the two-class nature of the problem (room-inside vs. room-outside segment classification), as well as the relatively small corpus size (see also Section 2.6.1).[1] Specifically, two SVM modeling

[1] All SVMs are trained in `Matlab`®, by its `svmtrain.m` function. By default, the regularization parameters are set taking into account the unbalanced nature of the two classes of interest. For this purpose, different penalties are set for misclassifying each class samples, with their ratio being equal to the inverse ratio of the two class sample sizes.

approaches are considered, resulting to a total of five different models, as discussed next.

• **Room-specific SVM models**, where a separate classifier is built for each smart-home room. Each training segment thus provides data to a total of R SVMs as a room-inside or -outside class sample, while during testing, a candidate segment is fed to the SVM of the room in which it was detected by the first stage. The SVMs can be built on any of the three feature vectors of Section 2.4.2, given by (2.11), (2.12), or (2.13), thus resulting in three different systems.

• **Global SVM models**, where a single SVM is developed being applicable to all rooms, thus removing dependence of the number of SVM models on R. Each training segment provides its data to the global SVM a total of R times (once as a room-inside sample and $R-1$ times as a room-outside one). During testing, candidate segments are fed to this global SVM. In both cases, room-dependent features are used, provided by (2.11) or (2.13), yielding two different systems. Features (2.12) are not used, as they would have re-introduced dependency on R.

Among the above modeling options, the proposed system employs room-specific SVMs on inter-room concatenated features (2.12). Note also that, since each room decides for its own final segments, it is possible that a segment gets assigned to multiple rooms or to no rooms at all.

2.4.4 Temporal operation and post-processing

In practice, the SVM classification of speech segments can be performed at two different temporal scales:

• **Over the entire segment**, where a single scalar feature is extracted for the segment for each of the five categories of Section 2.4.1, providing a single sample for SVM training or testing. Thus, assignment to a given room is made for the whole segment.

• **Over segment sliding windows**, where features are extracted on medium-sized windows sliding over the given segment. As a result, each segment provides multiple data points for SVM training or testing (per window). The scheme allows segment breakup and selective assignment of its parts into the room that it was detected in by the first stage of the algorithm.

The proposed room-localized SAD system employs the sliding-window approach, using windows of 600 ms in size advancing by a 100 ms shift. This necessitates minor modifications to the feature extraction methodology of Section 2.4.1. In particular, there is no longer the need of averaging in (2.8) and (2.9), since the medium-sized window sizes coincide, thus trivially allowing for one window only. Further, in (2.6), the non-speech energy is computed over the 0.5 s interval preceding the first window of the segment.

As a final step, post-processing is also applied to the results. Specifically, speech segments with less than 0.7 s distance between them are unified, whereas speech segments of less than 0.4 s duration are deleted.

2.5 Baseline approaches

Two additional, simpler systems are presented in this section, both following a two-stage architecture, to serve as baselines against the developed room-localized SAD system. The first method

employs MFCC features and GMM classifiers in both its algorithmic stages, while the second extends the well-known statistical model-based approach of Sohn et al. [36] to room-localized SAD, by incorporating a simple SNR-based room-assignment criterion. Details follow.

2.5.1 MFCC/GMM-based system

This baseline follows [106], and it is mainly considered in order to evaluate a system based entirely on a standard acoustic front-end (MFCC features), aiming also to demonstrate the value of the room discriminant features of Section 2.4.1.

Its first stage is identical to that of the proposed system. Namely, for every smart-home room, it performs weighted log-likelihood summation of MFCC/GMM-based scores by means of (2.1) and (2.2) over all room microphones ($\mathcal{M} = \mathcal{M}_r$) for classes $\mathcal{J} = \{\,\mathrm{sp}_r\,, \mathrm{sil}_{\,\mathrm{all}}\,\}$ (see also Section 2.3.4).

At the second stage, segments generated by the first stage are further examined and classified as room-inside or room-outside speech. For this purpose, room-specific GMMs are trained for each class $\mathcal{J} = \{\,\mathrm{sp}_r\,, \mathrm{sp}_{\bar{r}}\,\}$, and unweighted log-likelihood summation of MFCC/GMM-based scores is performed over all room microphones ($\mathcal{M} = \mathcal{M}_r$), followed by averaging over all short-time frames in the segment. Segments classified as room-outside speech are then deleted from the SAD output of the given room.

2.5.2 Sohn's algorithm with SNR criterion

The first stage of this baseline employs the well-known and effective SAD algorithm of Sohn et al. As they detail in [36], the method is based on a likelihood ratio test between speech and noise models, considered as Gaussians in the frequency domain under an i.i.d. assumption in frequency and that of additive uncorrelated noise. Following noise model estimation using observed noise and of the necessary SNRs by a decision-directed approach, the likelihood ratio test is performed, and decision results are smoothed by means of an HMM-based hang-over scheme [36].

In the designed baseline, Sohn's SAD is employed for each smart-home room r, using a single ad-hoc selected room microphone $m \in \mathcal{M}_r$. Then, at the second stage, for a first-stage generated segment in room r, the SNR of microphone m is compared to a global threshold; if below it, the particular segment is deleted from the room's SAD output. This baseline thus presents a well-established and relatively simple to implement approach for room-localized SAD.

2.6 Databases and experimental framework

We now proceed to describe the databases where the proposed system, its variations, and baselines are evaluated, as well as to discuss the adopted experimental framework and evaluation metrics used. In particular, the presentation refers to the experiments of Sections 2.7.1-2.7.4. An additional dataset and a slightly modified evaluation framework, necessary to allow comparisons with recent deep-learning based works, are detailed in the corresponding Section 2.7.5.

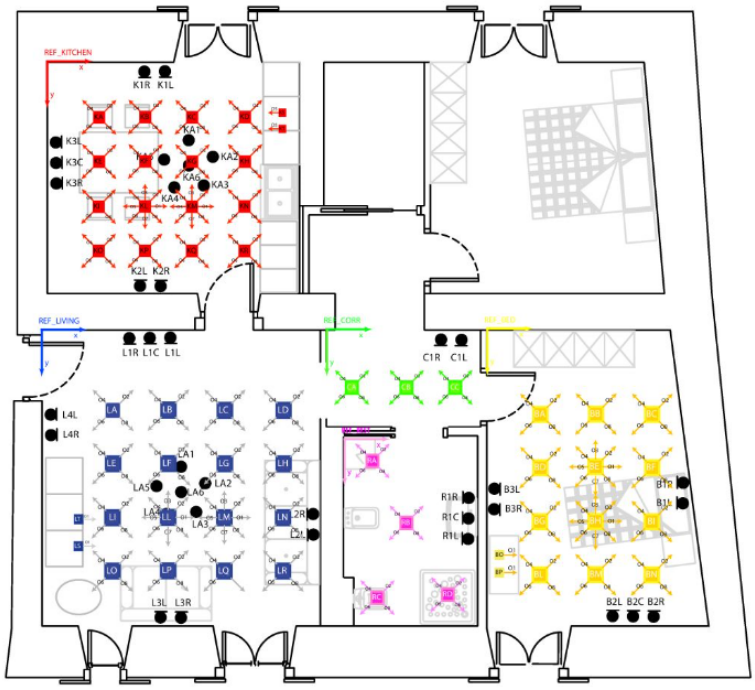


Figure 2.6: Floorplan of the multi-room DIRHA apartment where the datasets of Section 2.6.1 are simulated or recorded. Black circles indicate the 40 microphones installed inside five rooms on their walls or ceiling. Colored squares and arrows indicate possible positions and orientations of speech and other acoustic event sources (figure from [62]).

2.6.1 The DIRHA corpora

The experiments in Sections 2.7.1-2.7.4 are conducted on two databases: the Greek-language part of DIRHA-simcorpora II [61], hereafter referred to as "DIRHA-sim", and the "DIRHA-real" Greek corpus [83].[2] The datasets are either simulated or recorded inside a smart-home apartment (with an average reverberation time of 0.72 s), developed for the purposes of the DIRHA research project [100]. Its floorplan is depicted in Figures 2.1 and 2.6, showing that five of its rooms (Livingroom, Kitchen, Bathroom, Corridor, and Bedroom) are equipped with a total of 40 microphones grouped in 14 arrays. Most arrays consist of two or three microphones (with linear topology) located on the room walls, while, for each of the Livingroom and Kitchen, a six-element pentagon-shaped array is also located at the ceiling. As a result, concerning the set of adjacent microphone pairs used in calculating the coherence and SRP-based features, the two-element arrays provide one such pair, the three-element arrays two, and the pentagon-shaped arrays five, with all latter pairs containing the central array microphone. The Corridor thus yields the least pairs (one), while the Livingroom the most (ten).

As indicated by its name, the DIRHA-sim dataset contains audio simulations, produced as detailed in [61]. Briefly, first, about $9k$ room impulse responses are measured at each of 40 smart-home microphones from 57 possible source locations uniformly distributed in the rooms of interest

[2] DIRHA-sim is found at `https://dirha.fbk.eu/simcorpora`, whereas DIRHA-real is available on request to the author.

Table 2.1: Characteristics and statistics of the DIRHA-sim and DIRHA-real corpora, used in our experiments.

data characteristics	databases	
	DIRHA-sim	DIRHA-real
speech source	loudspeaker	human
1 min long sequences (#)	150	60
total speech (min)	47	19
overlapped speech (min)	22	0
non-speech events (#)	72	untranscribed
background noises (#)	10	untranscribed
subjects (#)	20	5
average SNR (dB)	13	15

and with up to 8 source orientations for each (as shown in Figure 2.6). These are then used to convolve high-quality, close-talk speech by 20 subjects (recorded at a 48 kHz sampling rate and an SNR average of 50 dB), while real, long-duration background noises and shorter acoustic events are also added to the resulting simulations. In total, 150 one-minute simulation sequences containing speech and noise are available. In contrast, the DIRHA-real set contains actual recordings of 5 subjects acquired by the 40 microphones inside the smart home under realistic noise conditions [83]. In total, 60 one-minute recorded sequences of speech and noise are available. Statistics of the two sets are summarized in Table 2.1.

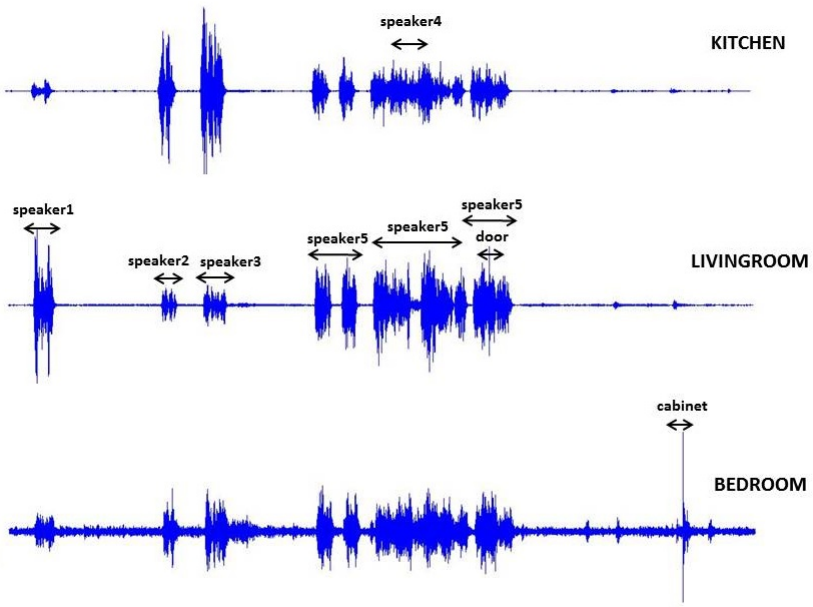
Apart from the main difference concerning the nature of the two sets (simulated vs. real), there exist two additional variations, as can be also observed in the waveform examples of Figure 2.7. First, DIRHA-sim is characterized by more adverse noise conditions, containing more background noises and acoustic events besides speech. Further, in DIRHA-sim, speech often overlaps with other acoustic events or speech in different rooms of the smart home. Indeed, as listed in Table 2.1, speech overlap there reaches 47% (22 out of 47 min). These facts deem DIRHA-sim much more challenging for room-localized SAD than DIRHA-real.

2.6.2 Experimental framework and metrics

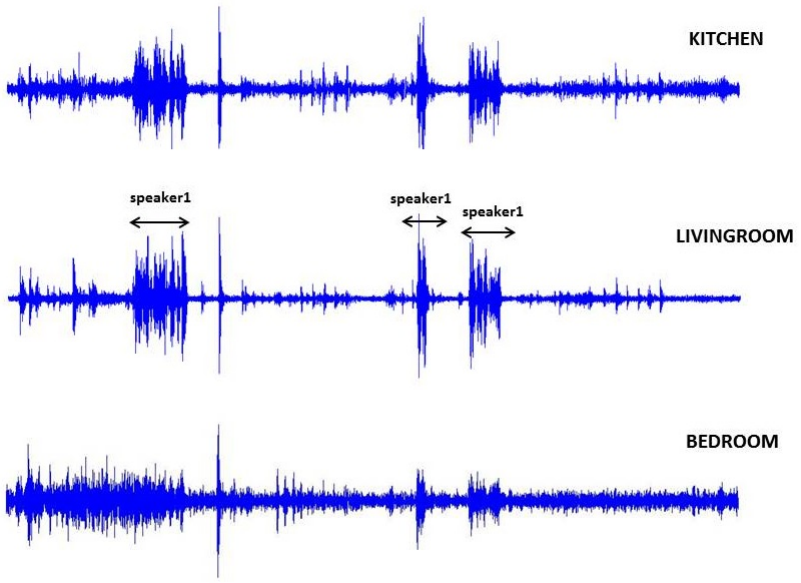
In the experiments of Sections 2.7.1-2.7.4, the DIRHA-sim dataset of 150 simulations is partitioned into a training set containing 75 of them and a test set with the remaining 75. Optimization of the first-stage algorithmic parameters of Section 2.3.3 (i.e., the transition penalty and constant prior added to the speech-class log-likelihood), as well as of the global threshold used in conjunction with Sohn's baseline, are performed on the training set. In the case of DIRHA-real, all 60 recordings are used for testing systems developed on the DIRHA-sim training data. This framework allows to also gauge the usefulness of simulated databases for training models and developing features and systems that can perform well in real-case scenarios, even when differences between the sets are significant.

For evaluation, the recall, precision, and F-score metrics are used, all computed at the frame

level with a 10 ms time resolution and reported %. Evaluation of room-localized SAD differs somewhat to the traditional room-independent case, as can be easily inferred from Figure 2.1. In traditional SAD, the aim is to detect speech anywhere in the smart home, and, as a result, each test-set sequence is evaluated only once (75 sequences for DIRHA-sim and 60 for DIRHA-real). In contrast, in the room-localized case, for each sequence, a total of $R = 5$ SAD outputs are



Dirha-sim



Dirha-real

Figure 2.7: Examples of multi-microphone data of the DIRHA corpora used in this work. Microphone waveforms in three rooms are shown. *Top:* A multi-speaker acoustic scene in the DIRHA-sim dataset. *Bottom:* A single-speaker scene in the DIRHA-real data.

Table 2.2: Room-independent SAD results on the DIRHA-sim (left) and DIRHA-real (right) test sets, further discussed in Section 2.7.1.

method		DIRHA-sim						DIRHA-real					
		recall		precision		F-score		recall		precision		F-score	
		GMM	HMM	GMM	HMM	GMM	HMM	GMM	HMM	GMM	HMM	GMM	HMM
oracle-best		96.94	94.67	94.01	96.82	95.45	95.73	93.01	95.49	95.91	96.46	94.44	95.97
channel avg.		87.86	82.26	76.64	83.13	81.82	82.69	65.56	71.57	89.47	87.42	75.37	78.34
best act.-SNR		94.56	92.36	83.85	87.95	88.88	90.10	88.77	90.33	88.95	86.87	88.86	88.57
best est.-SNR		96.60	93.63	66.56	73.54	78.81	82.38	92.43	93.41	74.38	74.02	82.43	82.59
Sohn's		81.22		58.91		68.29		78.05		61.51		68.80	
decision fusion	"u-sum"	94.39	91.08	83.60	90.97	88.67	91.01	74.76	89.11	96.54	91.70	84.26	**90.39**
	"w-sum"	95.00	91.78	83.57	91.82	88.92	**91.80**	76.87	87.37	96.58	93.37	85.67	90.27
	"u-max"	74.17	82.51	75.28	73.69	74.72	77.85	45.66	68.40	97.21	95.01	62.14	79.54
	"w-max"	95.44	95.53	82.34	87.16	88.41	91.15	79.76	89.66	95.77	88.70	87.03	89.18
	"u-vote"	92.55	88.92	84.18	92.24	88.16	90.55	69.12	83.39	96.61	95.02	80.58	88.82
	"w-vote"	91.37	91.83	87.39	90.40	89.34	91.11	74.76	85.03	96.54	94.82	84.26	89.66

evaluated (one for each room), with ground-truth each time considering only speech occurring inside the given room. Thus, $75 \times 5 = 375$ and $60 \times 5 = 300$ SAD outputs in total are evaluated for the DIRHA-sim and DIRHA-real test sets, respectively. This affects the evaluation metrics: for example, recall for room-localized SAD is computed as the ratio between the number of correctly detected room-inside speech frames and the total number of such frames in the ground-truth. In total, the test set contains 447 room-inside and 1788 room-outside speech segments in the DIRHA-sim case, and 232 and 928 segments respectively in DIRHA-real.

2.7 Experimental results

Next, we report our experiments. We first focus on room-independent SAD results, subsequently covering the room-localized case extensively. We also provide an error analysis of the proposed system, as well as a study on its robustness to the number of available microphones. We conclude the section with a comparison to recent deep-learning based approaches.

2.7.1 Room-independent SAD results

Room-independent SAD is evaluated first, primarily to showcase its easier nature compared to the room-localized task, as well as to benchmark differences between the various techniques of Sections 2.3 and 2.5 and simple channel selection schemes. Results are reported in Table 2.2 for both DIRHA-sim and DIRHA-real sets in terms of recall, precision, and F-score.

Specifically, in the lower part of Table 2.2, both the GMM- and HMM-based decoding schemes of Section 2.3.3 are presented in conjunction with the six fusion techniques of Section 2.3.2, but for the room-independent SAD system variant discussed at the end of Section 2.3.4 that uses all 40 smart-home microphones ($\mathcal{M} = \mathcal{M}_{\text{all}}$ and $\mathcal{J} = \{\text{sp}_{\text{all}}, \text{sil}_{\text{all}}\}$). These results are compared to two single-channel systems where microphone selection is driven by the best SNR per test-set sequence (actual based on ground-truth segmentation, or estimated), as well as the oracle-best channel result (that with the maximum F-score per sequence) and the average of all channel results. Finally, Sohn's algorithm is also considered, applied for each room (a single room microphone is used for each room), with the union of the results across rooms obtained.

For DIRHA-sim (left side of Table 2.2), we immediately observe the superiority of HMM-based Viterbi decoding over frame-based GMM segmentation. The best result is obtained by multi-channel fusion using log-likelihood summation scheme "w-sum", achieving an F-score of 91.80%. This is significantly higher than Sohn's method (68.29%), and it represents a 53.5% relative error reduction in F-score compared to the best estimated-SNR single-channel system (91.80% vs. 82.38%). Note that the latter performs similarly to the average of all channel results (82.69%), while it lags the ideal actual-SNR case (90.10%) where channel SNR computations employ ground-truth information. These comparisons confirm that the challenging nature of DIRHA-sim adversely affects SNR estimation. Note finally that the best multi-channel system still lags the oracle-best channel one (95.73%), showing potential for further improvements.

Similar observations hold in the DIRHA-real case (right side of Table 2.2). The best system again employs log-likelihood summation, with scheme "u-sum" reaching an F-score of 90.39%. This corresponds to a 44.8% relative F-score error reduction compared to the best estimated-SNR single-channel system (90.39% vs. 82.59%). The latter performs now better than the average of all channel results (78.34%), and it lies somewhat closer to the best actual-SNR system (88.57%) than in the DIRHA-sim case, due to the less adverse DIRHA-real environment. Note finally that, as above, the best multi-channel system lags the oracle-best channel result (95.97%).

2.7.2 Room-localized SAD results

We now switch focus to the room-localized SAD task. Our experiments are organized as follows: First, we evaluate the several possible choices of the system's first stage discussed in Section 2.3.4. Next, we investigate its second stage and the performance of the room discriminant features of Section 2.4.1. Finally, we present comparative results between our proposed system and the alternative baselines of Section 2.5.

The first experiment, reported in Table 2.3, compares the various design choices concerning the possible classes and microphones used in the first stage of the room-localized SAD system, as summarized in Section 2.3.4. In all cases, decision fusion by means of log-likelihood summation scheme "u-sum" is employed across microphones. For consistency in the comparisons, the various first stages considered are always followed by an identical second stage, namely that of the MFCC/GMM baseline of Section 2.5.1.

It is clear from Table 2.3 that the room-independent scheme leads to the worst performance, trailing all room-localized variants. The basic reason is that, in the latter schemes, the first stage can achieve high recall for room-inside speech and produces less room-outside segments compared to the room-independent case, thus the second stage has an easier task. The second line of the table corresponds to the classes and microphone set options chosen in the proposed system. These yield the highest recall (72.07%) among the room-localized SAD variants, with an F-score second, but very close, to the three-class modeling approach of the last line (66.12% vs. 66.43%).

The second experiment, reported in Table 2.4, concentrates on the proposed room discriminant features of Section 2.4.1, as well as their feature fusion schemes of Section 2.4.2 and the SVM modeling approaches of Section 2.4.3 operating over entire segments. The evaluation is conducted

Table 2.3: Effect of the various choices in the design of the system's first stage (discussed in Section 2.3.4) to the room-localized SAD performance on the DIRHA-sim test set. For consistency, the first stage is always followed by the second stage of the MFCC/GMM baseline of Section 2.5.1. Below, RI denotes room-independent operation ("oper") of the first stage and RL room-localized one.

oper	$\mathcal{M}$	classes $\mathcal{J}$	recall	precision	F-score
RI	$\mathcal{M}_{\text{all}}$	$\{\text{sp}_{\text{all}}, \text{sil}_{\text{all}}\}$	72.30	56.63	63.51
RL	$\mathcal{M}_r$	$\{\text{sp}_r, \text{sil}_{\text{all}}\}$	72.07	61.08	66.12
		$\{\text{sp}_r, \text{sil}_r\}$	71.20	60.39	65.35
		$\{\text{sp}_r, \text{sp}_{\bar{r}}, \text{sil}_{\text{all}}\}$	71.00	62.40	**66.43**

for the room-inside vs. room-outside speech classification task of the second stage of the developed algorithm. For this purpose, the ground-truth speech boundaries are used, thus decoupling the comparisons from the first stage. Further, results include four rooms of the smart home, excluding the Corridor ($R = 4$). Importantly, in addition to single features and their intra-room fusion (2.11), various feature subsets are also considered. Specifically, in Table 2.4 the best two, three, and four feature combinations are listed, as selected by wrapper-based sequential forward feature selection [111, ch. 5.7.2] that is conducted on DIRHA-sim (based on the corresponding proposed system F-scores). In addition, the three-feature subset of [107] is evaluated. Notice that the notation in (2.12) and (2.13) is slightly extended to allow inter-room fusion of single features and subsets.

Concerning DIRHA-sim (Table 2.4, top), in the case of room-specific SVMs we observe that for most individual features of Section 2.4.1 (with the exception of the energy-based one) performance improves by inter-room fusion. The best feature is the proposed spectrogram texture smoothness, achieving an F-score of 84.01% after fusion by (2.12). In contrast, the energy-based feature performs the worst at a 52.65% F-score after fusion by (2.13). For the entire feature vector ("all") obtained by intra-room fusion (2.11), small differences are observed between no room combination and inter-room fusion by (2.12) or (2.13), with the best F-score reaching 88.30%. Global SVM modeling performs slightly worse (85.46% F-score with fusion (2.13)).

Regarding feature subsets, the best two-feature set consists of the spectrogram texture smoothness and the SRP-based feature; envelope variance is then added to yield the best three-member set; and subsequently the coherence-based one is chosen. All subsets demonstrate better performance than individual features, when fused by (2.12) or (2.13). Also, we can observe that energy does not boost performance further, as the best four-feature set slightly outperforms the "all" set, achieving an 88.92% vs. 88.30% F-score with fusion (2.12). Finally, compared to [107], the "all" set achieves a 17.5% relative error reduction in F-score (88.30% vs. 85.81% with (2.12)).

In the less challenging DIRHA-real set (Table 2.4, bottom), the coherence, envelope variance, and spectrogram texture smoothness features take advantage of inter-room combination, whereas the energy- (as also on DIRHA-sim) and SRP-based ones fail to do so. The highest performing feature is the envelope variance with an F-score of 98.96% after fusion by (2.12), closely followed by spectrogram texture smoothness at 96.33% after fusion by (2.13). For the entire feature vector

Table 2.4: Performance of the room discriminant features of Section 2.4.1 and their combinations, in conjunction with inter-room fusion (Section 2.4.2) and SVM modeling (Section 2.4.3) for the room-inside vs. room-outside speech classification task of the second stage of the proposed algorithm. Results are reported on $R = 4$ rooms of the DIRHA smart home (excluding the Corridor) on the DIRHA-sim (top) and DIRHA-real (bottom) test sets using ground-truth speech segment boundaries. All SVMs operate over entire segments.

set	SVM models	feature ($\bullet$)	recall			precision			F-score		
			$f_{r,\mathcal{T}}^{(\bullet)}$	$f_{r,\,\mathrm{avg},\mathcal{T}}^{(\bullet)}$	$f_{\mathrm{home},\mathcal{T}}^{(\bullet)}$	$f_{r,\mathcal{T}}^{(\bullet)}$	$f_{r,\,\mathrm{avg},\mathcal{T}}^{(\bullet)}$	$f_{\mathrm{home},\mathcal{T}}^{(\bullet)}$	$f_{r,\mathcal{T}}^{(\bullet)}$	$f_{r,\,\mathrm{avg},\mathcal{T}}^{(\bullet)}$	$f_{\mathrm{home},\mathcal{T}}^{(\bullet)}$
DIRHA-sim	room-specific	(en)	63.97	37.93	40.06	50.51	86.03	86.92	56.45	52.65	54.84
		(coh)	47.46	87.41	88.66	67.90	77.01	76.05	55.87	81.88	81.87
		(ev)	82.89	90.81	90.38	78.01	74.85	76.28	80.37	82.06	82.74
		(ts)	71.91	86.00	89.35	52.21	74.46	79.28	60.50	79.82	84.01
		(srp)	76.76	79.85	79.25	53.94	56.44	60.94	63.36	66.13	68.90
		(ts,srp)	80.67	89.33	90.58	66.72	79.37	82.97	73.03	84.05	86.61
		(ts,srp,ev)	91.74	90.74	91.86	85.20	83.26	85.27	88.35	86.84	88.44
		(ts,srp,ev,coh)	90.62	90.42	92.27	83.65	84.96	85.80	86.99	87.61	**88.92**
		(en,coh,ev) [107]	89.48	87.65	90.37	78.90	81.16	81.69	83.86	84.28	85.81
		(all)	91.14	89.65	91.40	83.93	85.30	85.40	87.39	87.42	**88.30**
	global		91.12	92.21	n/a	78.49	79.63	n/a	84.34	85.46	n/a
DIRHA-real	room-specific	(en)	63.65	24.39	27.68	55.30	100.00	100.00	59.18	39.22	43.36
		(coh)	5.61	71.35	78.99	100.00	61.67	57.22	10.62	66.16	66.71
		(ev)	99.02	99.73	99.73	97.40	98.07	98.21	98.21	98.89	98.96
		(ts)	68.94	97.44	97.94	81.42	95.25	93.41	74.67	96.33	95.62
		(srp)	85.36	87.91	80.75	75.50	77.98	75.29	80.13	82.65	77.93
		(ts,srp)	90.28	94.52	97.33	91.58	95.32	86.76	90.92	94.92	91.74
		(ts,srp,ev)	99.90	98.82	97.81	99.82	97.87	97.24	99.86	98.34	97.53
		(ts,srp,ev,coh)	98.52	98.99	98.11	99.94	98.37	87.09	99.23	98.68	92.27
		(en,coh,ev) [107]	98.25	99.73	99.50	99.60	98.64	90.84	98.92	99.18	94.98
		(all)	98.89	98.85	95.68	99.94	98.46	80.21	99.42	98.66	**87.26**
	global		99.33	100.00	n/a	100.00	99.84	n/a	99.66	**99.92**	n/a

("all") obtained by intra-room fusion (2.11), small differences are observed between no room combination and inter-room fusion by (2.13), regardless of the SVM models used. However, concatenation across all rooms by (2.12) fails to improve matters (an F-score of only 87.26% is attained). This is probably due to the high dimensionality of the resulting vector and the use of multiple SVMs, in conjunction with the mismatch between the DIRHA-sim trained models and DIRHA-real test conditions. This seems also supported by the fact that inter-room fusion by means of (2.13) in most cases outperforms (2.12). Nevertheless, the best "all" feature system reaches an almost perfect F-score of 99.92%, obtained by global SVMs and fusion (2.13). Note also that this is very close to the 99.86% F-score of the spectrogram texture smoothness - SRP - envelope variance combination with no inter-room fusion.

As a complement to this experiment and to further gain insights into the room discriminant features, their correlation is investigated. For this purpose, the Pearson correlation coefficient is computed among all features over the speech segments of the DIRHA-sim test set, resulting in the matrix of Figure 2.8. As expected, the envelope variance, spectrogram texture smoothness, and coherence-based features demonstrate high correlation between them, as they are all related to reverberation. On the contrary, the energy- and SRP-based ones exhibit low correlation with all features.

In the third experiment, reported in Table 2.5, once again ground-truth segments are considered as input to the second stage. The aim here is three-fold: first, to showcase the superiority of the proposed room discriminant feature approach over the baselines of Section 2.5; second, to highlight performance differences among the various smart-home rooms; and, third, to further

Table 2.5: Comparison of the two baselines of Section 2.5 (upper part) and the room discriminant feature based approach (lower part) for the room-inside vs. -outside speech classification task. F-scores are reported for each room, as well as over $R = 4$ rooms (excluding the Corridor) and all $R = 5$ rooms of the DIRHA smart home, on both DIRHA-sim (left) and DIRHA-real (right) test sets using ground-truth speech segment boundaries. Room-specific SVMs are employed, operating over entire segments.

Features	DIRHA-sim							DIRHA-real					
	single room					multi-room		single room				multi-room	
	Liv.	Kitch.	Bath.	Bed.	Corr.	$R = 4$	$R = 5$	Liv.	Kitch.	Bath.	Bed.	$R = 4$	$R = 5$
MFCCs	70.96	72.52	61.25	76.94	39.03	72.32	70.49	46.93	71.50	80.98	58.11	68.91	68.91
SNR	55.59	57.57	17.38	50.75	8.80	48.59	41.22	41.47	72.99	53.42	31.63	53.74	45.54
$f^{(\text{all})}_{r,\mathcal{T}}$	83.74	92.83	83.33	86.76	34.28	87.39	80.63	97.17	100.00	99.89	100.00	**99.42**	**93.34**
$f^{(\text{all})}_{r,\text{avg},\mathcal{T}}$	84.80	92.83	84.48	85.05	38.11	87.42	81.46	99.50	97.51	99.89	99.16	98.66	89.94
$f^{(\text{all})}_{\text{home},\mathcal{T}}$	86.29	89.96	91.67	88.25	39.66	**88.30**	**84.26**	97.88	79.23	95.00	93.63	87.26	79.19

compare the fusion schemes of Section 2.4.2. Specifically, the MFCC/GMM-based second stage of the baseline of Section 2.5.1 is listed in the first line of Table 2.5, followed by the SNR-based room-assignment scheme of Section 2.5.2, as well as room-specific SVM modeling on (2.11), (2.13), and (2.12) operating over entire segments. F-scores are reported for each room separately (no Corridor F-score is shown for DIRHA-real, as there are no ground-truth room-inside segments there), as well as for all four (excluding the Corridor) or five rooms.

It is clear from Table 2.5 that the proposed approach dramatically outperforms the baselines: e.g., for $R = 5$, on DIRHA-sim the best result (84.26%) represents a 46.7% and 73.2% relative error reduction over the baselines of Sections 2.5.1 and 2.5.2 respectively, while on DIRHA-real the corresponding reductions of the best result (93.34%) stand at 78.6% and 87.8% relative. It is also clear that the Corridor is a challenging room, as seen by its low DIRHA-sim F-scores and the performance drop from the $R=4$ to the $R=5$ case. This is primarily due to its central location in the smart-home floorplan (see also Figure 2.6) exposing it to sounds coming from all other rooms, as well as the small number of microphones in it (only two). Regarding the multi-room results of the feature fusion schemes of Section 2.4.2, inter-room feature concatenation (2.12) performs best on DIRHA-sim, followed by (2.13). This can be expected as (2.12) captures more detailed

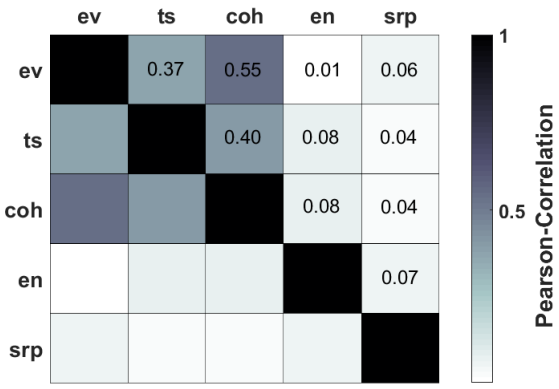


Figure 2.8: Pearson correlation coefficients between the room discriminant features of Section 2.4.1, computed on DIRHA-sim (ev: envelope variance; ts: spectrogram texture smoothness; coh: coherence-based; en: energy-based; srp: SRP-based).

Table 2.6: Performance of various approaches for the full task of room-localized SAD on DIRHA-sim (left) and DIRHA-real (right).

method		DIRHA-sim			DIRHA-real		
		recall	precision	F-score	recall	precision	F-score
single-stage	best RI	92.22	19.49	32.18	92.71	16.25	27.66
	MFCC/GMM	89.87	41.60	56.87	88.02	57.06	69.24
	Sohn's	73.17	17.33	28.02	73.40	17.71	28.53
two-stage baselines	MFCC/GMM	72.07	61.08	66.12	78.94	76.87	77.89
	Sohn's	43.14	21.96	29.11	46.39	22.26	30.08
proposed	seg ($R = 5$)	82.16	77.35	79.68	88.27	89.30	**88.78**
	win ($R = 5$)	83.09	78.96	**80.98**	86.51	88.87	87.68
	win ($R = 4$)	84.65	86.10	85.37	86.51	94.03	90.11

information (albeit at higher dimensionality). Similarly, fusion (2.13) is superior to the lack of inter-room combination in (2.11). On DIRHA-real, however, the above are reversed, as features (2.11) outperform (2.13) and, in turn, fusion by (2.12). This is primarily due to the mismatch of the DIRHA-sim trained SVMs to the DIRHA-real conditions, thus favoring lower-dimensional representations that generalize better, as also observed in Table 2.4.

Finally, Table 2.6 reports on the full task of room-localized SAD. Its upper part covers single-stage methods, namely the best room-independent approach ("best RI"), as well as the first stages of the MFCC/GMM baseline of Section 2.5.1 (recall that this is identical to the proposed system's first stage) and Sohn's algorithm (Section 2.5.2). The complete two-stage baselines are evaluated next, followed by the proposed algorithm employing room-specific SVMs on features (2.12) operating over the entire segments ("seg") or over sliding windows ("win"), where results both with and without the Corridor are reported.

As shown in Table 2.6, the proposed system operating over sliding windows reaches satisfactory performance, namely a 80.98% F-score on DIRHA-sim and 87.68% on DIRHA-real, which are further improved if the Corridor is excluded. The algorithm clearly outperforms the two-stage baselines dramatically, resulting to relative error reductions of 43.9% and 73.2% on DIRHA-sim compared to the methods of Sections 2.5.1 and 2.5.2, respectively. The corresponding improvements stand at 44.3% and 82.4% on DIRHA-real. The single-stage systems considered perform even worse. Not surprisingly, the addition of the second stage helps both baselines, especially the MFCC/GMM system.

Concerning the operation of the second stage over entire segments vs. sliding windows, it can be observed in Table 2.6 that the latter scheme fares slightly better on the more challenging DIRHA-sim dataset. An example of its superiority is provided in Figure 2.9 (same as in Figure 2.7 (left)). There, the Kitchen SAD results are shown for a case of two overlapping speakers located inside different rooms ("speaker 5" in the Livingroom and "speaker 4" in the Kitchen). The first stage of the system returns a segment containing both. Then, at the second stage, the segment-operating scheme classifies it entirely as Kitchen-inside speech, whereas the sliding window one

allows to only keep the part belonging to "speaker 4". Further, both schemes delete three erroneous first-stage segments, but fail to do so for two that originate in the Livingroom. However, on the less challenging DIRHA-real set, a slightly worse performance for the sliding-window scheme is observed in Table 2.6. This can be attributed to the lack of overlapping speech segments originating in different rooms, in conjunction with the obvious fact that window-based decisions rely on less data than entire segments.

2.7.3 Error analysis

This section attempts to provide additional insights into the performance of the various room discriminant features of Section 2.4.1. In particular, the focus lies on how such is affected by the speech source location and the amount of overlap in the detected segment.

Figure 2.10 concentrates on the two novel features proposed, namely the spectrogram texture smoothness (upper part) and the SRP-based feature (lower figure). There, in the case of segments with ground-truth boundaries and no overlap, performance of the features for Livingroom-inside vs. -outside classification on DIRHA-sim data is visualized by an appropriate coloring scheme within circular sectors that correspond to the various speech source positions and orientations in the smart home (blue indicates low misclassification rates, while red high ones). It can be observed that errors mostly occur around the Livingroom boundaries, but differ across features. For example, the spectrogram texture smoothness misclassifies mainly segments of adjacent rooms with orientation towards the Livingroom doors, as they reach its microphones with less reverberation. In contrast, the SRP feature classifies such correctly, as they produce high acoustic energy at the Livingroom doors. However, it misclassifies room-inside segments near these doors.

Finally, Figure 2.11 aims to quantify the effects of overlap to the room discrimination performance of the various features. For this purpose, two cases are considered: "low overlap" concerning speech segments with less than 30% of overlap with acoustic events of other rooms, and "high overlap" with more than 30%. Performance is measured in frame-based F-score, using ground-truth first-stage (room-independent) speech boundaries. In all single-feature sets, 5-dimensional

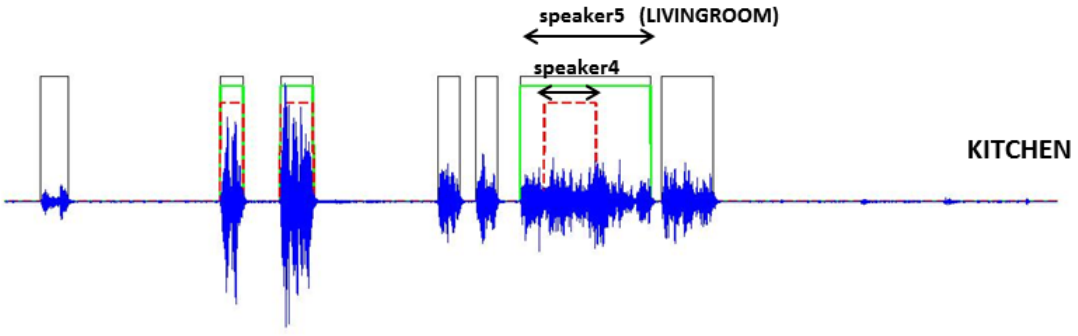


Figure 2.9: Example of the proposed room-localized SAD system output for the Kitchen of the DIRHA smart home, shown in green when operating over the entire first-stage segments (depicted in black), or in red dashed line when operating over shifting windows. The example is that of Figure 2.7 (left).

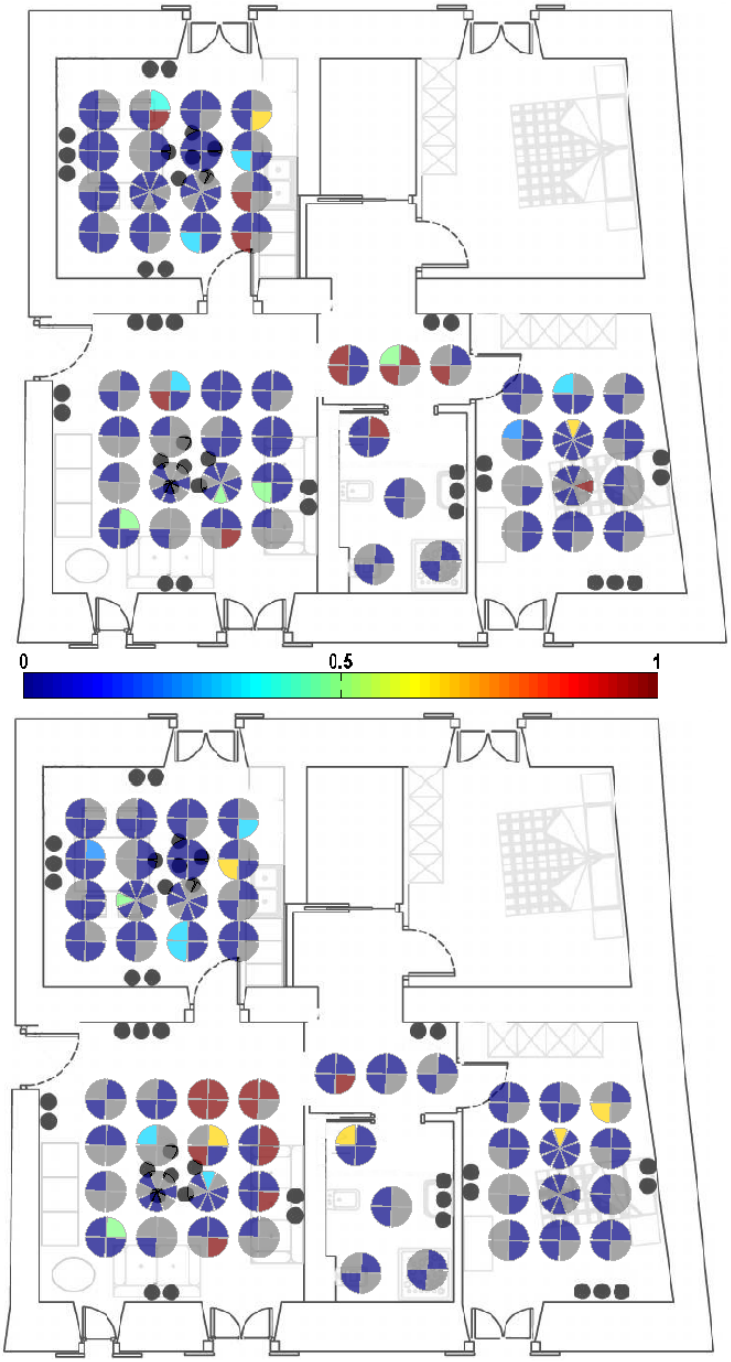


Figure 2.10: Visualization of error rates in Livingroom-inside vs. -outside segment classification for each speech source position and orientation on the DIRHA-sim test set, using two features of Section 2.4.1. *Upper:* spectrogram texture smoothness; *Lower:* SRP-based feature. Blue and red circular sectors indicate low and high percentage of errors, respectively, while gray sectors indicate unused orientations.

vectors are produced (one feature per room). Clearly, most sets exhibit low performance in the high overlap condition, with some (spectrogram texture smoothness, energy-based, and fused features) affected more.

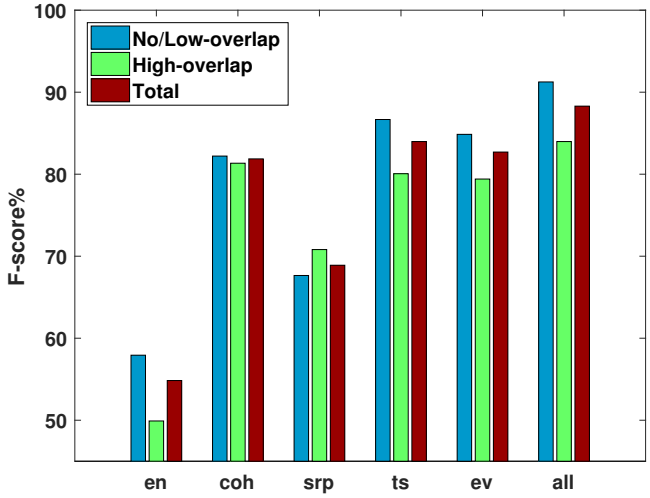


Figure 2.11: Performance of the room discriminant features of Section 2.4.1 in classifying speech segments exhibiting low or high overlap with audio events in other rooms for the DIRHA-sim test set (en: energy-based; coh: coherence-based; srp: SRP-based; ts: spectrogram texture smoothness; ev: envelope variance; all: intra-room fusion (2.11)). In all cases, inter-room fusion (2.12) and room-specific SVMs are used.

2.7.4 Robustness to reduced microphone setups

The proposed room-localized SAD algorithm relies on the availability of multiple microphones in the multi-room DIRHA apartment. As this installation includes 40 microphones, the question naturally arises as to how dependent the system is on such an expensive setup.

To investigate this, four reduced "nested" setups are considered, gradually decreasing the number of smart-home microphones $|\mathcal{M}_{\text{all}}|$ from 40 down to 5, specifically to $|\mathcal{M}_{\text{all}}| = 25$, 16, 10, and 5, as depicted in Figure 2.12 (compare to the original configuration of Figures 2.1 and 2.6). Note that the $|\mathcal{M}_{\text{all}}| = 10$ setup includes one microphone pair in each room, while the $|\mathcal{M}_{\text{all}}| = 5$ configuration only one microphone per room. For the latter, coherence- and SRP-based features cannot be computed due to the absence of microphone pairs, thus reducing the set of available features to three per room (see also Figure 2.13).

The first experiment, summarized in Table 2.7, quantifies the effects of reduced microphone setups to the GMM-HMM based SAD module. Specifically, room-independent SAD performance on the DIRHA-sim test set is reported (see also Table 2.2), employing HMM-based Viterbi decoding and "w-sum" decision fusion over the microphones of the various setups. To further reduce system complexity, a simplified modeling approach is also evaluated, where only a single GMM is trained on data of a specific microphone, in place of microphone-specific models. In particular, the Livingroom ceiling central microphone, available in all configurations, is used for this purpose. In that case, (2.1) and (2.2) are slightly modified by setting $b_{m,j}(\mathbf{o}_{m,t}) \leftarrow b_{M,j}(\mathbf{o}_{m,t})$, for all $m \in \mathcal{M}$, where M denotes the specific GMM-training microphone.

Concerning SAD performance, it is evident from Table 2.7 that it remains robust to the number of available microphones. In particular, the F-score degrades gracefully and monotonically as the installation becomes leaner: In the microphone-specific modeling case, the full-setup F-score of 91.80% reduces to 89.60% for $|\mathcal{M}_{\text{all}}| = 5$, exhibiting an absolute drop of only 2.2%.

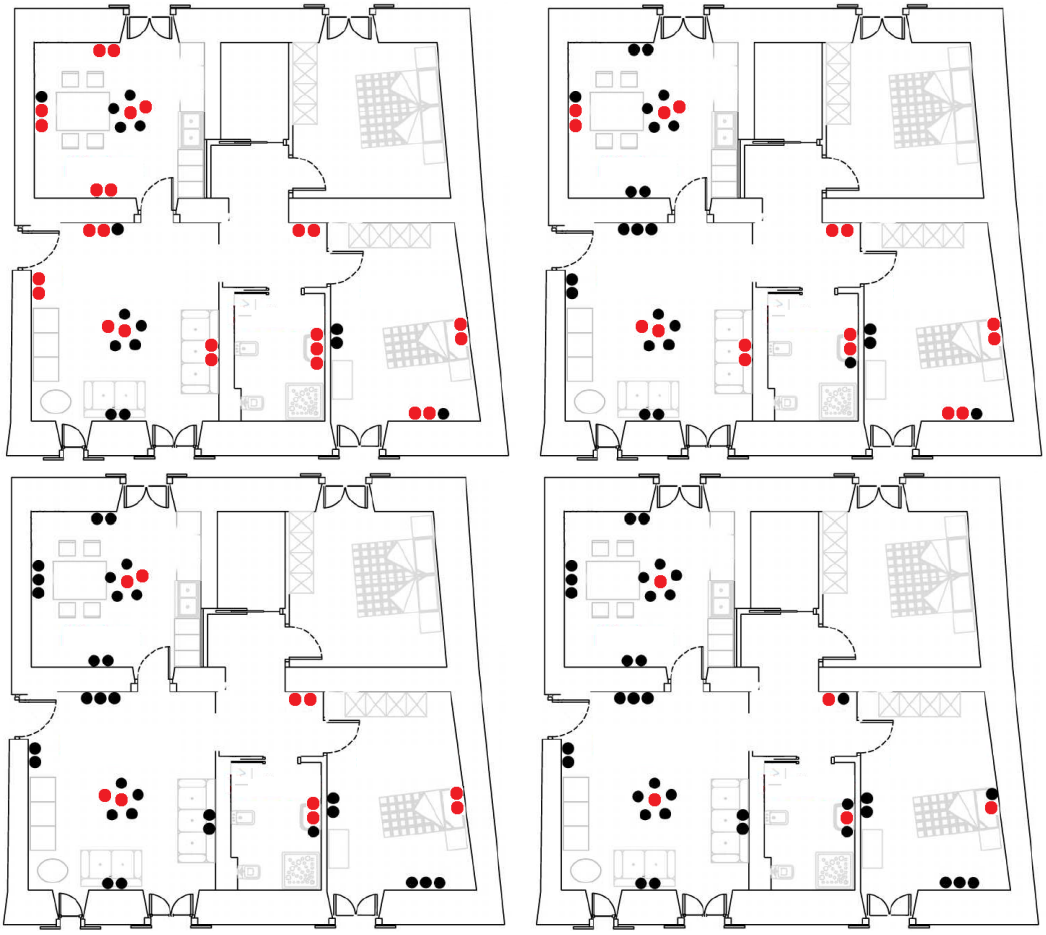


Figure 2.12: Reduced microphone setups of Section 2.7.4. *Left-to-right, Top-to-bottom:* $|\mathcal{M}_{\text{all}}| =$ 25, 16, 10, 5 microphones used (shown in red).

A similar trend is also observed in the single-GMM case. Further, comparing the two modeling approaches, the single-GMM one yields small only F-score absolute degradations within the 1.7% to 2.6% range (depending on the setup). Thus, in the lack of multi-channel training data, a single-microphone model constitutes a viable approach leading to satisfactory results.

Table 2.7: Room-independent SAD results on the DIRHA-sim test set, employing all available microphones ($|\mathcal{M}_{\text{all}}| = 40$) or the reduced setups of Figure 2.12. In all cases, HMM-based Viterbi decoding and "w-sum" decision fusion are used, where the combined log-likelihoods result from microphone-specific GMMs (left) or a GMM trained on a single microphone (right).

$\lvert\mathcal{M}_{\text{all}}\rvert$	microphone-specific GMMs			single microphone GMM		
	recall	precision	F-score	recall	precision	F-score
40	91.78	91.82	**91.80**	91.21	87.25	**89.19**
25	91.45	91.41	91.43	90.66	87.58	89.09
16	91.50	90.89	91.19	87.51	90.69	89.07
10	90.84	89.39	90.11	90.33	86.42	88.33
5	88.22	91.02	89.60	89.21	86.61	87.89

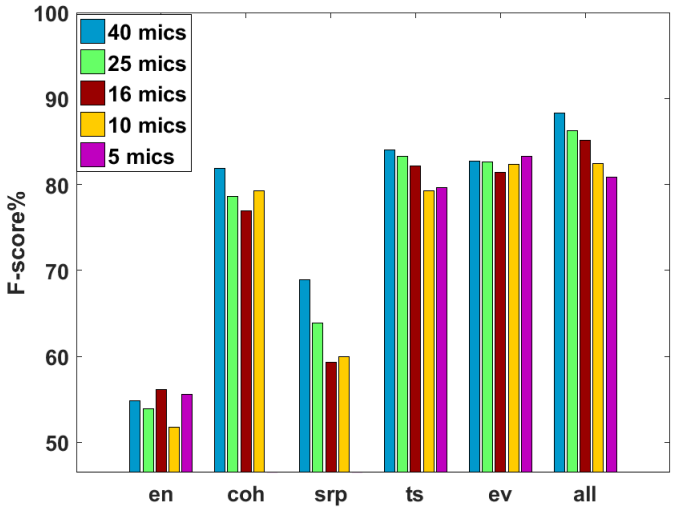


Figure 2.13: Performance of the room discriminant features of Section 2.4.1 for the speech-inside vs. -outside classification task with ground-truth segmentation on the DIRHA-sim test set, using various numbers of microphones (en: energy-based; coh: coherence-based; srp: SRP-based; ts: spectrogram texture smoothness; ev: envelope variance; all: intra-room fusion (2.11)). Inter-room fusion (2.12) and room-specific SVMs are used.

In the second experiment, depicted in Figure 2.13, the performance of the room discriminant features of Section 2.4.1 is examined as a function of the number of available microphones. For this purpose, the room-inside vs. -outside classification task (second stage of the algorithm) is considered with ground-truth segmentation on the DIRHA-sim test set. It can be readily noted that reduced setups have a noticeable, albeit not dramatic, effect on the performance of the intra-room fused features ("all"), degrading the full-setup F-score of 88.30% ($|\mathcal{M}_{\text{all}}| = 40$) to 85.10% for $|\mathcal{M}_{\text{all}}| = 16$ and 80.86% for $|\mathcal{M}_{\text{all}}| = 5$. Thus, the multi-channel second stage can benefit from larger microphone numbers, but can also perform satisfactorily with fewer microphones. Regarding individual feature performance in reduced setups, the envelope variance seems the most robust, while the SRP-based feature the least.

2.7.5 Comparison to deep-learning approaches

As overviewed in Section 1.1.2, a number of works on room-localized SAD have appeared recently, proposing single-stage algorithms based on deep-learning methods [74–77]. In this section, a performance comparison to our developed system is provided.

To enable such comparison, the experimental framework of these works is followed, deviating from that of Section 2.6. In particular, the corpus used is the Italian-language part of the DIRHA simcorpora (DIRHA-sim-evalita), first introduced as part of the SASLODOM evaluation campaign at the EVALITA'14 workshop [112]. This contains 80 one-minute simulations, generated in the DIRHA apartment as discussed in Section 2.6.1. Experiments are conducted by ten-fold cross-validation to reduce performance variance, with each test fold containing eight simulations. Results are reported in terms of the "overall SAD detection error" metric, as defined in [112], considering only two rooms ($R = 2$) of the DIRHA apartment, i.e., Livingroom and Kitchen.

Comparative results between the best deep-learning results of [74–77] and our proposed algorithm are presented in Table 2.8. In particular, the best system of [74, 75], employing a DNN over 187-dimensional features of various types that are extracted from the best microphone per room, yields a SAD error of 5.8%. Further, the 3D-CNN system of [76], operating on 40-dimensional Log-Mel filterbank energies after temporal splicing and combining information from the three best microphones per room, exhibits a 7.0% SAD error. This improves to 5.2%, when employing all microphones available in the two rooms [77]. Finally, a 3.5% SAD error is reported in [77], when the aforementioned 3D-CNN is extended incorporating 51-dimensional GCC-PHAT patterns [89] to jointly provide SAD and speaker location estimates (marked as "SAD+SLOC" in the table). However it should be noted that this system employs additional information during its training, in the form of ground-truth speaker positions (in 2D room coordinates).

In comparison, our proposed algorithm exhibits SAD errors of 5.7% and 4.7%, when operating over entire segments ("seg") or sliding windows ("win"), respectively. The latter represents a 19% relative SAD error reduction over the DNN of [75] and 10% over the 3D-CNN of [77], proving better than segment-based operation in the challenging and noisy DIRHA-sim-evalita data (as also observed in Table 2.6 for DIRHA-sim). These comparisons highlight the competitiveness of our two-stage system and the suitability of the five room discriminant features of its second stage. Of course, it is possible that the deep-learning methods could have gained advantage, had more training data been available in the DIRHA corpora.

2.8 Conclusions

In this chapter, we have presented an efficient multi-channel, two-stage approach to address speech activity detection in multi-room smart-home environments, equipped with multiple microphone arrays distributed inside them. In the general scenario, possibly concurrent speech activity in different rooms needs to be detected and the effect of cross-room interference suppressed. For this purpose, the proposed room-localized SAD system first employs a multi-channel speech/non-speech segmentation module per room, and it subsequently determines whether detected speech activity occurs inside or outside each room by utilizing a novel set of room discriminant features. Experiments on a suitable multi-room, multi-channel dataset demonstrate satisfactory performance

Table 2.8: Performance (in overall SAD detection error [112]) of deep-learning based approaches vs. the proposed algorithm for room-localized SAD on the DIRHA-sim-evalita corpus.

method		SAD error (%)
deep-learning	DNN [75]	5.8
	3D-CNN [76]	7.0
	3D-CNN [77]	**5.2**
	3D-CNN (SAD+SLOC) [77]	3.5
proposed	seg	5.7
	win	**4.7**

on both simulated and real data, reaching F-scores of 81.0% and 87.7% respectively, while significantly outperforming alternatives that combine well-known baselines and features (MFCCs, Sohn's SAD, SNR), as well as comparing favorably to deep-learning based approaches (DNNs, CNNs). The evaluation results verify the robustness of the two-stage system and the suitability of the devised hand-crafted features, while also highlighting the realistic design and value of the current simulated database for developing algorithms that generalize well to real recorded data.

Chapter 3

Isolated Acoustic Event Detection in Smart Spaces

3.1 Introduction

In the previous chapter, we studied SAD, which focuses on detecting the time boundaries of human speech present in an audio recording. In fact, "speech" belongs to the broad category of acoustic events, and SAD constitutes a special case of the more general AED problem.

In our work on AED we have considered several variants of the task, including isolated and overlapped event scenarios, as well as single-channel or multi-channel setups available in the given smart-space environment. Under this framework, we have developed and evaluated several different approaches for the AED task. In this chapter we focus on the isolated AED case.

Isolated AED refers to the case where there is no overlap between the occurrences of the different acoustic events in the audio clip. In this chapter we examine the problem of detecting acoustic events in smart indoors environments, equipped with multiple microphones. In particular, we focus on channel combination strategies, aiming to take advantage of the multiple microphones installed in the smart space, capturing the potentially noisy acoustic scene from the far-field. Towards this end, we investigate channel fusion at the signal level, employing beamforming techniques to produce enhanced signals, at the feature level, utilizing time-difference-of-arrival (TDOA) between channel signals as additional informative features, and at the decision level, appropriately integrating detection decisions to yield the final one. Further, "multi-style" training is also considered, utilizing observations from all available microphones to produce more robust models.

The above are investigated using two related detection systems that are based on appropriately trained GMMs on traditional audio front-end features. The first is a frame-based GMM that operates over sliding windows of fixed duration, whereas the second employs Viterbi decoding over the entire observation sequence, based on an HMM composed of the trained GMMs over the classes of interest. Experimental results are reported on a multi-microphone corpus containing isolated acoustic events of twelve types occurring in a single room that is appropriate for AED. In the evaluation results, multi-channel approaches are demonstrated to significantly outperform single-channel baselines.

The rest of this chapter is organized as follows: Section 3.2 presents the multi-channel methods for fusion and information extraction; Section 3.3 is devoted to the experiments and results; and, finally, Section 3.4 concludes this chapter.

3.2 Multi-channel information extraction and fusion

A number of channel combination approaches at different levels are investigated in this chapter, as discussed next.

3.2.1 Multi-channel training

In this approach, observations from all available microphones, or from an appropriate subset of them, are used during the training process in order to obtain the statistical model (GMM) of each class of interest. This is akin to the "multi-style" training procedure, often employed in ASR and other machine learning problems to improve robustness of the produced models. The obtained models can then be used during testing on one or more microphones, in the latter case using the decision fusion framework discussed below.

3.2.2 Signal fusion

In this approach, a plain delay-and-sum beamformer with no post-filtering is employed to combine audio from multiple microphones into a single enhanced signal (typically, a subset of the available microphones is exploited that are closely located within microphone arrays). For this purpose, the "BeamformIt" software is used [113]. Depending on which channels are combined, one or more beamforming signals can be created, thus also allowing multi-channel training and/or decision fusion approaches to be employed.

3.2.3 Decision fusion

In this approach, the available class models are tested on the appropriate channels that are to be fused at the decision level. Typically, for example, a single-channel classifier is tested on the respective channel that it is trained on; a multi-channel model is tested on any channel within the set of microphones that is trained on; and a signal-fusion model is tested on its corresponding enhanced signal. Such tests provide sequences of log-likelihood scores for each class and channel of interest, which are then fused at the frame level by a decision fusion method. In particular, we employ the multi-microphone decision fusion methods already used for the SAD task earlier and presented in Section 2.3.2, namely the "u-sum", "w-sum", "u-max", "w-max", "u-vote" and "w-vote" methods. Here the set of classes $\mathcal{J}$ includes all the acoustic events considered instead of just speech and non-speech events.

Approaches "w-sum", "w-max", and "w-vote" require channel confidence estimation to yield necessary weights. Previously in Section 2.3.2 we already presented a way to compute the weights using the log-likelihood differences between the classes. Here, we propose additional ways to

compute the weights, by utilizing the following channel decision confidence or channel quality indicators (similarly to [114]).

• **N-best average log-likelihood difference**: For every channel, this is derived by computing the average of the differences in the log-likelihood score between the highest scoring class GMM and the $N-1$ following in descending order (where $N = |\mathcal{J}|$ is upper bounded by the number of available event classes in the set $\mathcal{J}$).

In particular, if we denote with $b_{m,j}(\mathbf{o}_{m,t})$ the sorted log-likelihoods of the GMMs for microphone m given its acoustic features $\mathbf{o}_{m,t}$ at time frame t, and event class $j \in \mathcal{J}$, the weights for this channel quality indicator are computed as:

$$w_{m,t} = \sum_{j=2}^{N} b_{m,1}(\mathbf{o}_{m,t}) - b_{m,\mathrm{j}}(\mathbf{o}_{m,t}). \tag{3.1}$$

Large values of this difference indicate high confidence.

• **N-best average log-likelihood dispersion**: This constitutes a modification of the above, where log-likelihood differences between all top N-scoring class pairs are summed. The weights $w_{m,t}$ for the microphone-m at time frame t are computed as:

$$w_{m,t} = \sum_{j=1}^{N-1} \sum_{j'=j+1}^{N} b_{m,\mathrm{j}}(\mathbf{o}_{m,t}) - b_{m,\mathrm{j}'}(\mathbf{o}_{m,t}). \tag{3.2}$$

As before, large values demonstrate high confidence.

• **Log-likelihood score entropy**: The entropy over the probability distribution of all class posteriors is computed. In this case the weights are computed as:

$$w_{m,t} = -\sum_{j=1}^{N} p_{m,\mathrm{j}}(\mathbf{o}_{m,t}) \log(p_{m,\mathrm{j}}(\mathbf{o}_{m,t})), \tag{3.3}$$

where

$$p_{m,\mathrm{j}}(\mathbf{o}_{m,t}) = \frac{e^{b_{m,\mathrm{j}}(\mathbf{o}_{m,t})}}{\sum_{j'=1}^{N} e^{b_{m,\mathrm{j}'}(\mathbf{o}_{m,t})}}, \tag{3.4}$$

is the softmax function used to map the log-likelihood scores into probabilities summing to 1. Small entropy values indicate high classification confidence.

• **Segmental signal-to-noise-ratio (SNR)**: This is a commonly used channel quality indicator, with high SNR values indicating good data quality.

We note that in all cases, before fusion, the weights are normalized to sum to 1 across the channels, for each time frame. After experimenting with the above channel confidence indicators, we converged to using "segmental SNR", yielding weights after their normalization over the channels fused.

3.2.4 Feature extraction

Regarding the features used, 13 MFCCs with Δ's and $\Delta\Delta$'s are extracted from the single-channel or fused signal, over 100ms duration frames with a 20ms shift. An example of the discrimina-

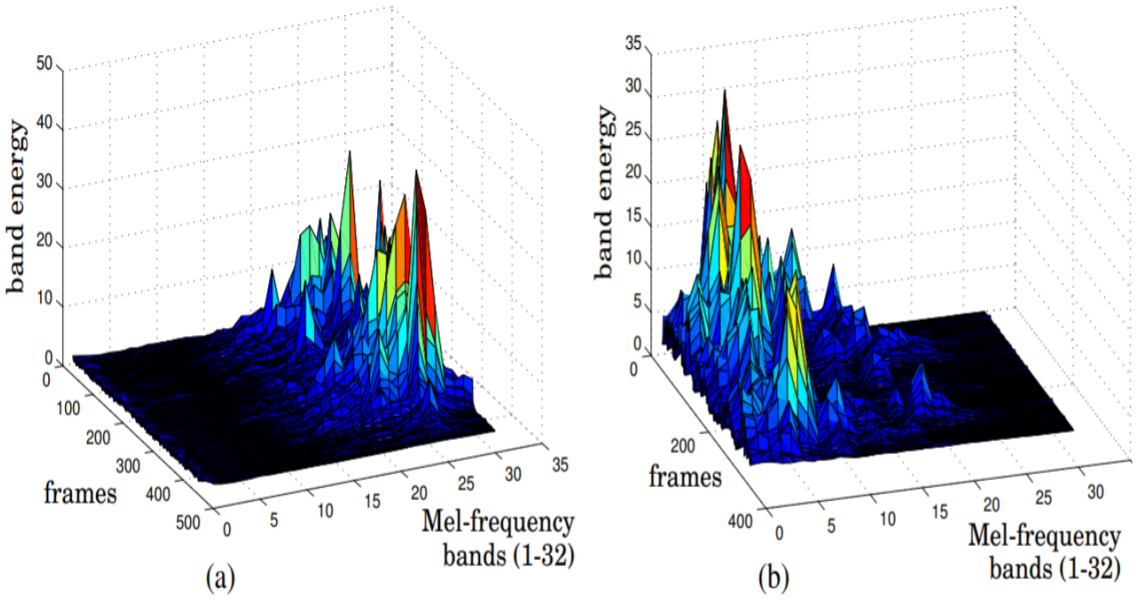


Figure 3.1: Mel-frequency band energy features (32 bands) depicted over time (frames) for an example occurrence of "key jingle" (left) and "speech" (right).

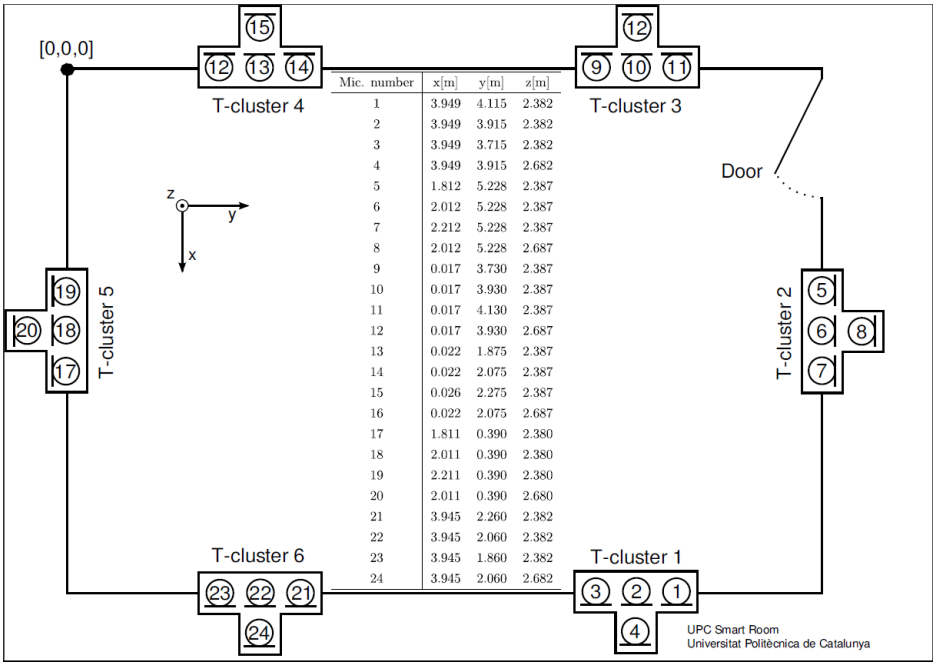
tive capability of the latter features (but without the discrete cosine transform operation included in MFCCs and the subsequent derivative computation) is given in Figure 3.1 for two different acoustic event type occurrences under consideration, namely "key jingle" and "speech". Clearly, the Mel-frequency band energies (which are the primitive features for the computation of the MFCCs) separate these two examples very well. In addition, and similarly to [115], we employ as features TDOAs between pairs of adjacent microphones, as these are related to the source location and possibly the class of certain acoustic events. Such features are used to train a separate GMM, which is then combined with MFCC-trained GMMs employing decision fusion.

3.2.5 Detection approaches

Two detection systems are developed, employing at their core the trained GMMs with multi-channel fusion. Following our speech detection implementations presented earlier in Section 2.3.3, we employ (a) the HMM-based Viterbi decoding over entire sequence, and (b) the GMM-based scoring over sliding window approaches. The only difference here is that the number of states is larger, as the number of acoustic events rises (not just speech/non speech events). Also the length of window and window shift in the GMM-based approach are 0.6 s and 0.4 s respectively in our experiments.

3.3 Experiments and results

The development and evaluation of the various approaches is performed on the UPC-TALP multi-microphone corpus of acoustic events [78]. This database contains a set of isolated acoustic events



Mic. number	x[m]	y[m]	z[m]
1	3.949	4.115	2.382
2	3.949	3.915	2.382
3	3.949	3.715	2.382
4	3.949	3.915	2.682
5	1.812	5.228	2.387
6	2.012	5.228	2.387
7	2.212	5.228	2.387
8	2.012	5.228	2.687
9	0.017	3.730	2.387
10	0.017	3.930	2.387
11	0.017	4.130	2.387
12	0.017	3.930	2.687
13	0.022	1.875	2.387
14	0.022	2.075	2.387
15	0.026	2.275	2.387
16	0.022	2.075	2.687
17	1.811	0.390	2.380
18	2.011	0.390	2.380
19	2.211	0.390	2.380
20	2.011	0.390	2.680
21	3.945	2.260	2.382
22	3.945	2.060	2.382
23	3.945	1.860	2.382
24	3.945	2.060	2.682

Figure 3.2: Floorplan of the meeting room used in the UPC-TALP database recordings. In total 24 microphones are employed, grouped in 6 T-shaped arrays. Their spatial coordinates (x,y,z) are also depicted.

that occur frequently in a meeting environment scenario. In our task, in addition to silence, we have 12 different events in total: knocks (door, table), door slams, steps, chair moving, spoon, paper work, key jingle, keyboard typing, phone ringing, applause, cough, and speech. Audio data from a total of 24 channels are available, provided by six T-shaped microphone arrays located on the room walls as shown in the floorplan of the meeting room in Figure 3.2. As the UPC-TALP database recordings are divided into 8 independent sessions, experiments have been conducted in a leave-one-out session fashion, keeping seven sessions for training and leaving one for testing.

The results for the AED task are depicted in Table 3.1. Performance of the various combination schemes considered is reported in terms of Diarization Error Rate (DER) [116], which in our case (isolated events) practically corresponds to frame misclassification. The results presented correspond to the best combination of parametres used (state transition penalty and number of Gaussians). As a baseline in our experiments, the "best estimated-SNR channel" selection strategy (per session) is considered. For a given session, the SNR for each channel is computed as the ratio between the total energy in the non-silence and silence segments detected. In the "best actual-SNR" method, segment boundaries are obtained from the ground-truth. In the "oracle best channel" method, in each session the channel with the lowest DER is selected. Finally "average over channels" refers to the mean DER of all the single-channel results in the leave-one-out experiment.

Concerning the results, at first we observe that Viterbi decoding (HMM) outperforms the sliding window approach (GMM). Regarding decision-level fusion, we can observe its superiority over the baseline systems. The best approach is "w-sum" that achieves a 8.10% relative error

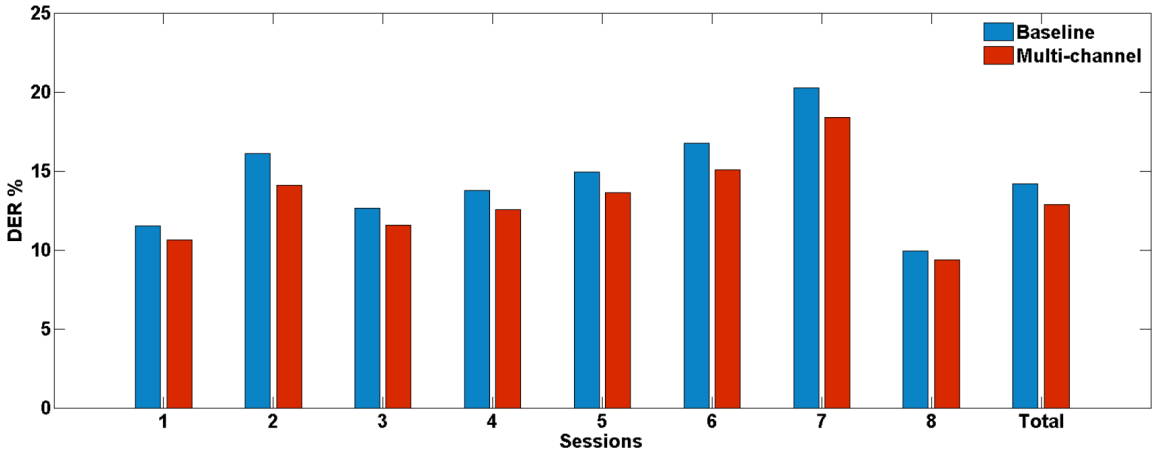


Figure 3.3: Performance (in DER %) of baseline "best estimated-SNR" and best multi-channel approach ("TDOAs & MFCCs") for the 8 sessions of AED problem.

reduction (from 14.20% to 13.05%) compared to the best SNR single-channel system.

The combination of decision fusion with multi-channel training and signal fusion yielded no improvement. Yet, the results remained better than the single-channel baseline. Finally, the combination of TDOAs with MFCCs in conjunction with GMMs at the decision level obtained the best overall result (Table 3.2). In particular, the combination of TDOAs with the "u-sum" method yielded a 12.88% DER, which corresponds to a 9.30% relative error reduction from the "best estimated-SNR channel" approach (Figure 3.3), and 11.20% from the "average over channels" DER. This can be explained by the fact that some events occur at similar locations in the various sessions. The best combination gave weight equal to 0.1 to the TDOA model and 0.9 to the MFCC model. Note that the DER of the TDOA model (without fusion) reaches 36.64%.

Table 3.1: Multi-channel fusion results for the AED problem. Results are depicted in DER %.

training style		single-channel		multi-channel	signal fusion
trained models (#)		24		1	6
channels tested (#)		24		24	6
model type		GMM	HMM	HMM	HMM
best estimated-SNR channel		18.54	14.20	14.59	14.53
best actual-SNR channel		18.43	14.16	14.43	14.48
average over channels		19.21	14.34	14.42	14.42
oracle best channel		17.50	12.71	13.04	13.40
decision fusion	u-max	18.94	13.76	14.19	14.37
	u-vote	18.09	13.13	13.42	13.40
	u-sum	17.91	13.21	13.36	13.50
	w-max	18.21	13.66	13.96	14.15
	w-vote	18.12	13.17	13.50	13.78
	w-sum	17.94	**13.05**	13.29	13.43

Table 3.2: Results for the fusion of MFCC and TDOA based models (with the HMM scheme) for AED.

TDOAs & MFCCs		DER%
decision fusion	u-sum	**12.88**
	w-sum	12.92

In order to verify that the improvement observed by the multi-channel approaches is statistically significant, we apply the Wilcoxon signed-rank test. In particular, a one-sided Wilcoxon test [117] is performed to compare the detection accuracies over all 8 leave-one-out experiments between the various multi-channel approaches and the baseline system. We also compare the significance of improvement between weighted and non-weighted approaches.

The outcomes of the tests are positive using the value $p < 0.05$. The improvements over the baseline observed are judged as significant in all approaches except the "u-max" one ("TDOAs", "w-sum", "u-sum", "w-vote", "u-vote", "w-max"). Also statistical significant improvement was observed between the "w-sum" and "u-sum" methods.

3.4 Conclusions

In this chapter, we investigated multi-channel combination approaches at different levels for the problem of isolated AED, outperforming the baseline single-channel system. Concerning the back-ends used, we can observe that Viterbi decoding is more appropriate for the detection task. It finds the most probable sequence of events in an optimal way. As for the decision fusion approaches, in general summation methods work better than majority based ones, and weighted better than unweighted ones. Finally, the extraction of the TDOAs and the training of a separate GMM on them improved further the performance of the overall system.

Chapter 4

NMF-based Single-channel Overlapped Acoustic Event Detection

4.1 Introduction

In the previous chapter we described the isolated case, which is the most common scenario under which the AED problem is examined. However, depending on the particular task and the environment conditions, there may exist low or high overlap between the occurrences of different acoustic events. In this work, we mainly employ NMF techniques in order to tackle the overlapped AED task for both single-channel and multi-channel setups.

The remainder of this chapter is organized as follows: First, in Section 4.2 we provide an introduction to the basic concepts and formulations of NMF-based AED. Then, in Section 4.3 we develop a CNMF-based system with an improved detection step, and in Section 4.4 we investigate ways to improve the robustness of classifier-based NMF systems in overlapping conditions. Finally, we summarize the chapter in Section 4.5.

4.2 Non-negative matrix factorization approaches

NMF-based approaches and their variants have begun to attract interest in the field of both isolated and overlapping AED in recent years. This is due to both their robustness and their natural ability to detect multiple events occurring simultaneously, as long as appropriate non-negative and linear representations of them are available. In this section, we will present the basic formulations of some of the most common NMF-based methods.

4.2.1 Basic NMF

The main idea behind the application of NMF for AED is the linear decomposition of acoustic events into spectral atoms. Given the representation of events with non-negative and approximately linear features (e.g., spectrogram, filterbank energies), overlapping events can be decomposed into atoms of individual events.

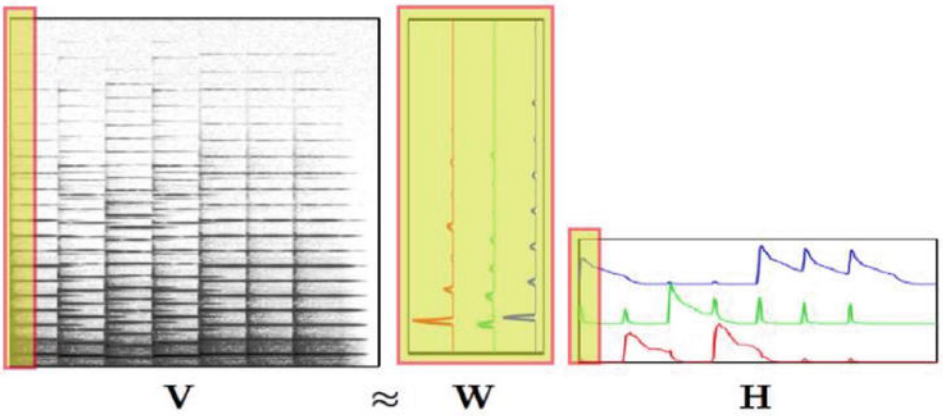


Figure 4.1: An NMF example of a piano performing "Mary had a little lamb". Dictionary matrix **W** captures the harmonic content of the three pitches of the passage and activation matrix **H** captures the time onsets and gains of the individual notes (figure from [119]).

NMF seeks to determine a linear non-negative approximate factorization of the observed feature matrix $\mathbf{V} \in \mathbb{R}_{\geq 0}^{P \times N}$, by the product $\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H}$, where $\mathbf{W} \in \mathbb{R}_{\geq 0}^{P \times R}$ denotes the non-negative dictionary matrix, and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{R \times N}$ represents the non-negative activation matrix. Here P denotes the feature dimensionality, N the number of time frames, and R the total number of event atoms in the dictionary matrix. An example of NMF is depicted in Figure 4.1. Minimization of a suitable error cost function $D(\mathbf{V}||\mathbf{WH})$ results in iterative estimation of **W** and **H** [118]. Most common choices for the cost function are the Euclidean distance, the Itakura-Saito divergence, and the Kullback-Leibler (KL) divergence. In the case of KL divergence, which is defined as:

$$D(\mathbf{V}||\mathbf{WH}) = ||\mathbf{V} \odot \log(\mathbf{V} \oslash \mathbf{WH}) - \mathbf{V} + \mathbf{WH}||, \tag{4.1}$$

matrices **W** and **H** are obtained by means of the following multiplicative updates:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left\{\mathbf{W}^T(\mathbf{V} \oslash (\mathbf{WH}))\right\} \oslash \left\{\mathbf{W}^T\mathbf{1}_V\right\} \tag{4.2}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left\{(\mathbf{V} \oslash (\mathbf{WH}))\mathbf{H}^T\right\} \oslash \left\{\mathbf{1}_V\mathbf{H}^T\right\}$$

where $\odot$ and $\oslash$ denote element-wise matrix multiplication and division, and $\mathbf{1}_V$ is a matrix with all elements equal to 1 and dimensions equal to those of **V**. Multiplicative update-based approaches have the most widespread usage for solving the NMF task, mainly due to their high reproducibility. Alternative efficient algorithms for solving the NMF task have also been proposed and applied successfully in the literature [120–122].

The dictionary **W** containing spectral atoms for each event is created during the training phase, either by employing the iterative updates presented above, or by using "exemplar" based methods. In the "exemplar" methods, representative spectral atoms are extracted directly from the training data of the given events.

For detection, assuming a given dictionary **W** that contains atoms of the various events of

interest, the estimated $\mathbf{H}$ provides the activations of each event through time. As in detection-related tasks the sparsity of $\mathbf{H}$ may become desirable or even crucial, sparse-NMF, a variant of NMF, is often employed, minimizing the following objective,

$$G\,(\,\mathbf{V}\,||\,\mathbf{WH}\,)\;=\;D\,(\,\mathbf{V}\,||\,\mathbf{WH}\,)\;+\;\lambda\,\|\mathbf{H}\|_1\;\;, \tag{4.3}$$

with parameter λ controlling the trade-off between sparseness on $\mathbf{H}$ and accurate reconstruction of $\mathbf{V}$ by the $\mathbf{WH}$ product. Depending on the cost function selected (KL divergence, Euclidean distance, etc.), different updating equations result for $\mathbf{W}$ and $\mathbf{H}$ [123].

4.2.2 Convolutive NMF

NMF is a linear non-negative approximate factorization of the observed feature matrix. CNMF [124] is its convolutive extension, making possible the decomposition of events into atoms with temporal evolution. It is formulated as follows: Given a non-negative data feature matrix $\mathbf{V} \in \mathbb{R}_{\geq 0}^{P\times N}$, the goal is to approximate $\mathbf{V}$ by matrix $\mathbf{\Lambda}$, derived as a temporal convolutive sum of a "dictionary" and "activations", namely

$$\mathbf{V}\;\approx\;\mathbf{\Lambda}\;=\;\sum_{t=0}^{T-1}\mathbf{W}_t\cdot\overset{t\rightarrow}{\mathbf{H}}\;\;, \tag{4.4}$$

where, operator $\overset{t\rightarrow}{\bullet}$ shifts the columns of its matrix argument t places to the right, $\mathbf{W}_t \in \mathbb{R}_{\geq 0}^{P\times R}$ denotes the non-negative dictionary matrix at time step t, $\mathbf{H} \in \mathbb{R}_{\geq 0}^{R\times N}$ represents the non-negative activation matrix, T is the number of time frames spanned by each dictionary atom, and R stands for the number of atoms in the dictionary. The i-th column of $\mathbf{W}_t$ describes the i-th atom, t time steps after its beginning. The dictionary thus contains R atoms of size $P\times T$ each. Minimization of a suitable error cost function $D(\mathbf{V}||\,\mathbf{\Lambda})$ results in the iterative estimation of $\mathbf{W}_t$ and $\mathbf{H}$ [124, 125].

Although CNMF tends to produce sparse activations, in order to ensure sparsity on $\mathbf{H}$, similarly with basic NMF, there is the sparse-CNMF variant, incorporating the sparsity term in its cost function [126].

$$G\,(\,\mathbf{V}\,||\,\mathbf{\Lambda}\,)\;=\;D\,(\,\mathbf{V}\,||\,\mathbf{\Lambda}\,)\;+\;\lambda\,\|\mathbf{H}\|_1\;\;. \tag{4.5}$$

With sparse-NMF and sparse-CNMF being the core methods, several other NMF variants and extensions have been developed and successfully applied in the literature [21, 127, 128].

4.3 Sparse-CNMF with improved detection and dictionary selection

In this section, we investigate sparse-CNMF for detecting overlapping acoustic events in single-channel audio, within the experimental framework of a suitable AED dataset provided by the DCASE'16 Challenge (Task 2) [79]. Our main focus lies in the efficient creation of the dictionary, as well as the detection scheme associated with the CNMF approach.

Specifically, for the dictionary creation stage, we propose an "exemplar"-based approach suitable for CNMF, by employing a shift-invariant method for the efficient size reduction of the dictio-

nary. The proposed method compares favorably to the standard CNMF-based dictionary building in our experiments. Further, for detection, we develop a novel algorithm that combines information from the CNMF activation matrix and atom-based reconstruction residuals, achieving significant improvements over conventional detection based on activations alone.

The structure of this section is as following: First the "Dictionary Building" and the "Detection Approaches" subsections present the proposed methods for exemplar-based dictionary creation and improvement of the ordinary detection scheme for CNMF, respectively. In the next subsections, details and experimental results of our system submitted to the DCASE'16 AED Challenge are provided.

4.3.1 Dictionary building

Dictionary building is a very important step in NMF-based methods. Representative atoms from each class must be contained in the dictionary matrix, capable of reconstructing unseen data. Using training data consisting of isolated event instances, a sufficient number of atoms is extracted and stored in the dictionary for each class of interest, resulting to matrices

$$\mathbf{W}_t = [\,\mathbf{W}_t^{(1)}, \ldots, \mathbf{W}_t^{(C)}\,]\,, \quad t \in [\,0\,, T{-}1\,]\ , \tag{4.6}$$

where C is the number of classes (events). In the case of CNMF-based methods, due to increased computational complexity, we need to create a rather compact dictionary. In the following, we present two alternatives for this task.

• **CNMF-based:** For each class of interest, the training instances are concatenated to form its data matrix, $\mathbf{V}^{(i)}$. Then, via sparse-CNMF iterative updates, matrices $\mathbf{W}_t^{(i)}$ and $\mathbf{H}^{(i)}$ are computed (as in [129]), and $\mathbf{W}_t^{(i)} \in \mathbb{R}_{\geq 0}^{P \times R_i}$ are stored in the dictionary. The duration, T, of each atom and their total number, R_i, are predefined. By extracting the same number of atoms for each class, their total number for all events in the dictionary becomes $R = C \cdot R_i$.

• **Shift-invariant dictionary reduction:** Here, we propose an alternative way for dictionary creation that selects a group of atoms from the original training data. For each class, first, a large number of atoms is extracted from its data matrix $\mathbf{V}^{(i)}$, using a sliding window of duration T (shifted by one feature frame at a time). Then, only R_i of them are selected by "uniformly sampling" the set of the resulting atoms, as explained next. The process aims at selecting different types of existing atoms based on a similarity measure, appropriate for CNMF. In our case, such similarity should be shift-invariant: i.e., two atoms are considered similar if the Euclidean distance between them, or between their temporally shifted versions, is small.

To achieve atom comparisons in a shift-invariant way, we first rearrange them into vectors of size $P \cdot T$, in a row-wise manner. This way, a time-shift of atoms results to shifts of their corresponding vectors. Then, atom similarity is measured as the Euclidean distance between the magnitudes of the Fourier transforms (DFTs) of the rearranged vectors, based on the well-known shift-invariant property of this transform. The available atoms are thus mapped to their Fourier-magnitude vectors, which are subsequently sorted based on their Euclidean distance from their mean. Finally, R_i atoms are selected by uniformly sampling the resulting sorted list.

The adopted sampling scheme represents a simple approach to desired dictionary size reduction. Alternatively, well-known clustering methods like k-means could also be used for the task.

4.3.2 Detection approaches

As stated earlier, having created the dictionary matrix $\mathbf{W}_t$, sparse-CNMF accepts as input the data matrix $\mathbf{V}$, and outputs the desired activation matrix $\mathbf{H}$ (following the approach in [126]). The final event detection can occur by exploiting the information in the above matrices. We present two main approaches for accomplishing this.

• **Using activations only:** Most of NMF-based approaches employ the information in $\mathbf{H}$ directly [16], or indirectly [18]. In our method, activations in $\mathbf{H}$ are directly used for detecting possible events. In particular, for each class, the activations are summed across all their atoms, for each frame, resulting in a new matrix $\mathbf{H'} \in \mathbb{R}_{\geq 0}^{C \times N}$, with elements

$$H'(\,i\,,n\,) \;=\sum_{r\,\in\,\{i\}} H(\,r\,,n\,) \;\;, \tag{4.7}$$

where i denotes the class ($i = 1,\cdots,C$), $\{i\}$ the set of row indices in $\mathbf{H}$ that correspond to the i-th event atoms, and $n \in \{1,\cdots,N\}$ the time frame. Then, at time n, a class is considered active if $H'(i,n) > \theta_H$, where θ_H is a suitably chosen activation threshold. A post-processing step can also be employed to yield smooth activations. Finally, as activation refers to atoms, $T-1$ additional frames following the detected activations are considered active.

• **Incorporating reconstruction residuals:** An alternative method to the above decides for an event activation, not by thresholding the elements of $\mathbf{H'}$, but by measuring KL-divergence between $\mathbf{V}$ and $\mathbf{\Lambda}$, when only the atoms of the event in question and of background noise are used in reconstruction (see Section 4.3.3 for details on background noise modeling). More specifically, the total reconstruction error of sparse-CNMF over a time-segment, seg, under consideration, is $D\,(\,\mathbf{V}_{seg}\,||\,\mathbf{\Lambda}_{seg}\,)$, whereas reconstruction error on basis of only the i-th event and noise is $D\,(\,\mathbf{V}_{seg}\,||\,\mathbf{\Lambda}_{seg}^{(i,bg)}\,)$, where,

$$\mathbf{\Lambda}_{seg}^{(i,bg)} = \sum_{t\,=\,0}^{T-1} \mathbf{W}_t^{(i,bg)} \cdot \overset{t\,\rightarrow}{\mathbf{H}}_{seg}^{(i,bg)} \;\;, \tag{4.8}$$

with $\mathbf{H}_{seg}^{(i,bg)}$ denoting the part of $\mathbf{H}$ that contains only rows corresponding to atoms of the i-th class or background noise and columns that correspond to the time frames of seg. Similarly, in the above, $\mathbf{\Lambda}_{seg}$ and $\mathbf{V}_{seg}$ contain the columns of (4.4) and of the data matrix, respectively, within the segment under consideration.

We define the "residual ratio" of the i-th event as the ratio between the residual on basis of (4.8) to the total one, using (4.4), namely

$$\mathcal{E}(i\,,n) = \frac{D\,(\mathbf{V}_{seg}\,||\,\mathbf{\Lambda}_{seg}^{(i,bg)})}{D\,(\mathbf{V}_{seg}\,||\,\mathbf{\Lambda}_{seg})} \;\;,\quad \text{for all} \quad n \in seg \;\;. \tag{4.9}$$

In computing (4.9), non-overlapping segments of 1 s in duration are used. Small residual ratio values for the i-th event in a given segment means that large percentage of the reconstruction in

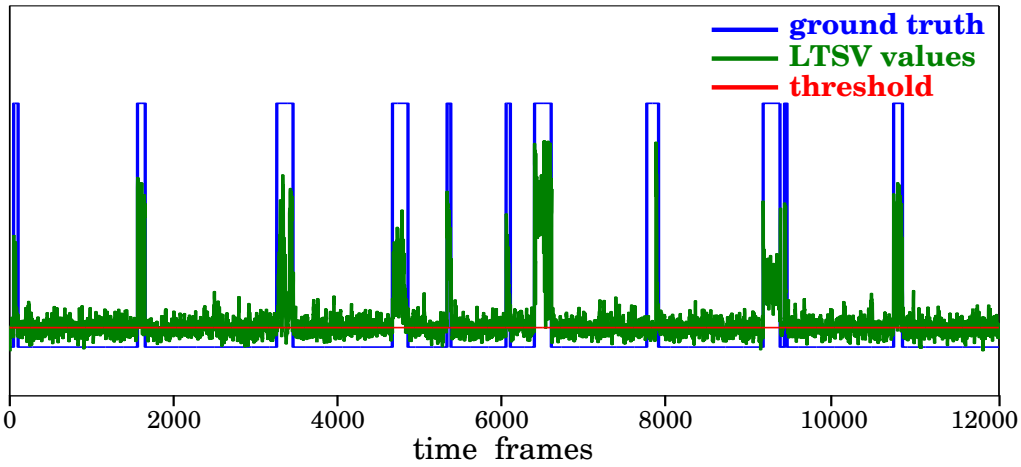


Figure 4.2: An example of applying the long-term signal variability (LTSV) measure to background noise detection (see Section 4.3.3). Ground-truth peaks correspond to acoustic events.

that segment is achieved using only the i-th event (together with background noise). Activations in $\mathbf{H}'$ with large magnitude are also often related with large percentage of reconstruction, but this is not always the case. From the minimization of (4.5), large magnitude activations may occur for a given event and a given time frame, but with a small corresponding reconstruction contribution.

In our first approach using activations only, the event detection criterion is the activation matrix $\mathbf{H}$ element magnitudes. In the residuals-based approach, instead, the criterion is the accuracy of reconstruction using only atoms and activations of a particular event. In our final system, submitted to the Challenge, we combine both. Thus, the i-th event is considered active at time frame n, if

$$H'(i\,,n) > \theta_H \quad \text{and} \quad \mathcal{E}(i\,,n) < \theta_{\mathcal{E}} \quad . \tag{4.10}$$

Thresholds θ_H and $\theta_{\mathcal{E}}$ are chosen as explained in Section 4.3.3.

4.3.3 System implementation details

• **Background noise modeling:** In addition to modeling the acoustic events by incorporating representative atoms in the dictionary, background noise modeling is necessary for robust AED. With the presence of background noise atoms in the dictionary, false alarm event activations are avoided in areas that events are not present. Also, more reliable reconstruction is possible in active areas, assuming additive noise.

In our approach, and following work in [16], we extract the background noise atoms from the observed data during decoding (on-the-fly). The advantage of this scheme is the adaptation of the background dictionary to slightly different conditions, possibly existing each time. However, instead of assuming background noise present at the beginning and end of the observed data, as in [16], we attempt to extract background atoms from various areas of the signal, by employing the long-term signal variability (LTSV) measure, described in [130]. This measure has been successfully used in voice activity detection, and it is based on the fact that background noise usually

exhibits smaller variability through time in its spectrum.

In our system, a frame is considered as noise if its LTSV value is lower than a fixed threshold, θ_L. As before, the shift-invariant dictionary reduction method is applied to areas that noise is detected to help provide background noise atoms. An example of the LTSV based approach is shown in Figure 4.2, where LTSV values for a Challenge corpus signal are depicted, together with ground-truth locations of acoustic events. As it can be seen, LTSV values and the chosen θ_L ensure that acoustic event time frames are avoided.

• **Features, system parameters, and post-processing:** We now provide some additional details of our implemented system. Concerning audio feature extraction, we have experimented with various feature sets that satisfy non-negativity and approximate linearity: Mel-filterbank energies, Gammatone-filterbank energies, DFT spectrogram, and the variable Q-Transform (VQT). The first three are computed using 30 ms long frames with a 10 ms shift, whereas VQT is obtained from the baseline system of [79]. Our final submitted system uses 150-dimensional Mel-filterbank energy features (M= 150).

Regarding dictionary building, atoms of 200 ms (T= 17 frames) in duration are used, and for the CNMF-framework, parameter λ in (4.3) is set to 0.7. Further, approximately 200 atoms per event class are used ($R_i \approx 200$), with $R \approx$ 2.4k total atoms (including background noise modeling).

Concerning the various thresholds employed, θ_H in (4.10) is computed as a percentage (15%) of the maximum value of matrix $\mathbf{H}'$ elements. Threshold $\theta_{\mathcal{E}}$ in (4.10) is computed as a percentage (106%) of the minimum of $\mathcal{E}(i\,,n)$ for a given segment. Such values are optimized on available development data (see Section 4.3.4).

Finally, as a post-processing stage in the detection system, one-dimensional dilation is performed on each row of matrix $\mathbf{H}'$, in order to broaden the intervals of high-peaked activations produced. In the case of the combined method, dilation is performed before the combination with the residuals approach. At the end, $T-1$ frames after each detected activation are also considered as active.

4.3.4 Database and experimental framework

We perform experiments on the DCASE'16 Challenge database designed for Task 2 – "Sound event detection in synthetic audio" [79]. The corpus contains recordings of eleven office-related acoustic events (see also Figure 4.3), consisting of three parts: The training set with 20 isolated recordings of each event; a development set with 18 two-minute long recordings of synthetic mixtures of audio events and noise at various SNRs and event overlap conditions ("density" and "polyphony"); and a test set of similar structure to the development set (54 recordings), only used in the Challenge evaluation, with its ground-truth publicly unavailable during the challenge time period.

Regarding the experimental setup, we report experiments on both the development and test sets (the latter as only provided by the Challenge organizers). Specifically, for the development set, due to its particularity of containing the same event instances as the training set, we use two different setups, described next.

Table 4.1: Performance of baseline and proposed systems.

system	setup #1		setup #2		test	
	F-score	ER	F-score	ER	F-score	ER
NMF-baseline	0.42	0.79	0.32	0.87	0.37	0.89
activations-only	0.83	0.30	0.43	0.79	–	–
activations&residuals	0.84	**0.29**	0.55	**0.63**	0.56	**0.68**

Table 4.2: Performance of different feature sets and dictionary sizes.

features	feat. dim.	dict. size	setup #1		setup #2	
			F-score	ER	F-score	ER
VQT	545	200	0.79	0.37	0.29	0.88
Gamma	150	200	0.82	0.33	0.35	0.86
Mel	150	200	0.83	**0.30**	0.43	**0.79**
Mel	150	100	0.81	0.36	0.42	0.85
Mel	100	100	0.83	0.30	0.42	0.82
DFT	545	100	0.78	0.42	0.41	0.83

Table 4.3: Performance of different dictionary building methods.

dictionary building method	setup #1		setup #2	
	F-score	ER	F-score	ER
sparse-CNMF	0.64	0.60	0.29	0.89
shift-invariant reduction	0.83	**0.30**	0.42	**0.82**

- Setup #1: This is identical to the default setup of Task 2. One dictionary is built using all isolated training data, and then AED is performed on all 18 development set recordings.
- Setup #2: Here, to allow testing on unseen event instances, we perform a 18-leave-one-out experiment. In total, 18 dictionaries are built, each tested on a single development set recording, by using each time all available training set instances, except those contained in the particular development set recording.

For the evaluation, we employ the adopted Challenge metrics [79], namely frame-based F-score and frame-based total error rate (ER). The latter is defined as $\text{ER} = (I + D + S)/N$, where I denotes acoustic event insertions, D deletions, S substitutions, and N the total number of ground-truth events at a given frame. ER is computed in frames of 1 s in length.

4.3.5 Results

In Table 4.1, the results using the Challenge-provided NMF baseline, our submitted system, and a variant of it are compared for the different experimental setups considered. Regarding the NMF-baseline, it builds the dictionary using the training data, and extracts 20 atoms per class. Atoms have single-frame duration, and are extracted from the VQT spectrogram (60 bins, 10 ms step). A post-processing stage applies median filtering to the output and allows up to five concurrent

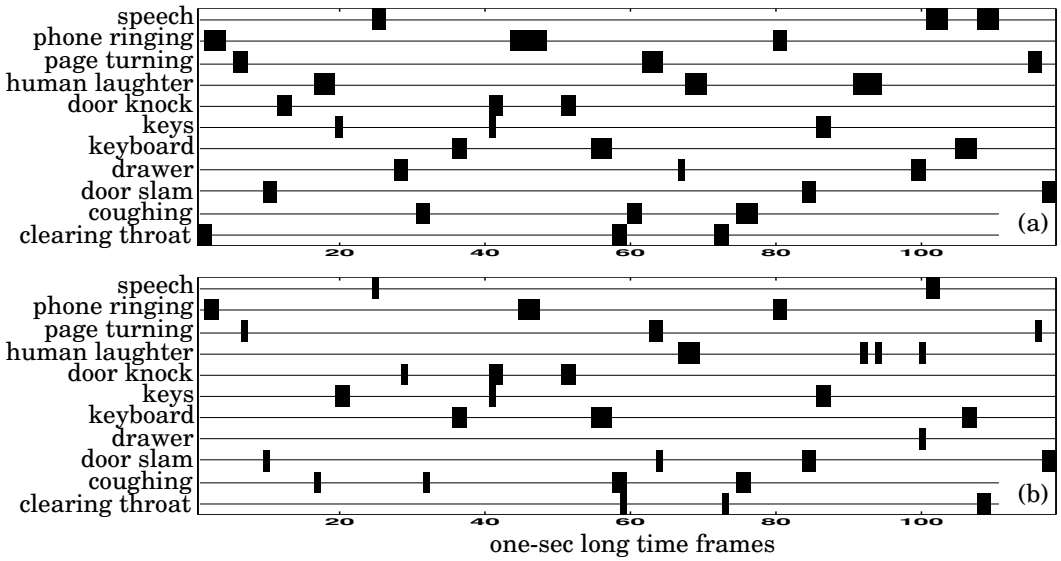


Figure 4.3: AED on the "dev_1_ebr_6_nec_3_poly_0.wav" Challenge recording: (a) ground-truth; (b) output of our submitted system. Acoustic event labels are also shown.

events [79].

Both our systems, depicted in Table 4.1, perform dictionary creation employing the shift-invariant reduction approach, and their details are provided in Section 4.3.3. It is obvious that both outperform the baseline in all setups. In particular, our submitted system ("activations & residuals") achieves 63.3%, 27.6%, and 23.6% relative reduction in ER over the baseline for setup #1, #2, and the test set, respectively. It seems that the extraction of more atoms per class (almost ten-fold over the baseline), combined with the incorporation of temporal structure under the CNMF-framework, lead to major improvements.

Comparing our two detection approaches, we can observe that the system using the combination of activations and reconstruction residuals (submitted to the Challenge) achieves a 20% ER relative reduction in setup #2, compared to the system using activations only. This highlights the complementarity of the two methods. The improvement is mainly due to the elimination of false activations, exhibiting large peaks in $\mathbf{H}'$ but also having a large residual ratio.

In Table 4.2, we show experimentation regarding different audio feature sets, together with variations in their dimensionality and dictionary size (number of atoms per class is depicted). We can observe that Mel-filterbank energies achieve the best performance among the different sets considered. It thus seems that they are more appropriate for the set of acoustic events considered in the Challenge. Also from the Mel feature results (150-dimensional), we can observe that increasing dictionary size leads to slight improvements.

A comparison of the different dictionary building methods is shown in Table 4.3, using the same detection system in both cases (a 100-dimensional Mel-filterbank, activations-only system, with 100 atoms per class). Clearly, the shift-invariant dictionary size reduction approach outperforms conventional CNMF-based dictionary building. This provides evidence that accurate representation of event atoms (instead of approximate) is beneficial to detection, as long as we

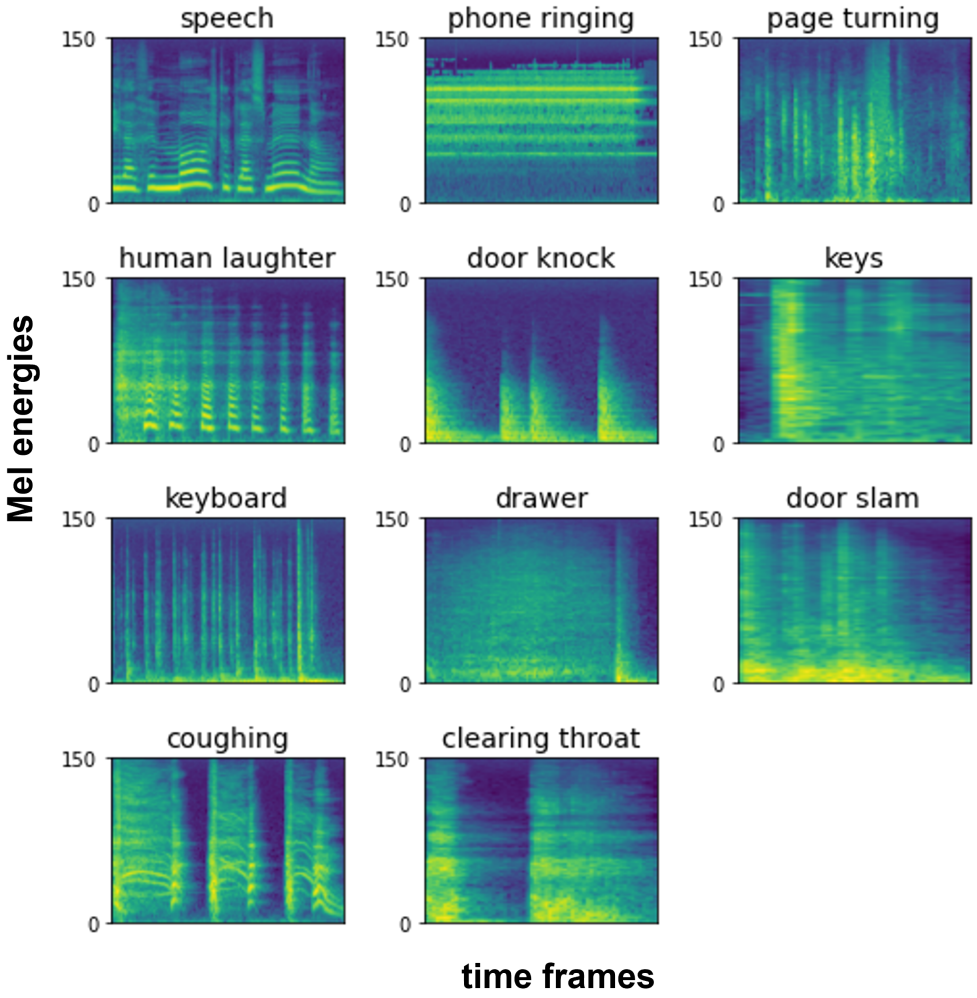


Figure 4.4: Mel energies representations for the different acoustic events. For better visibility purposes, the logarithm of the values is shown.

have a way to select appropriate atoms.

Finally, in Figure 4.3, the output of our system is shown against ground-truth for a particular audio recording of the development set, and in Figure 4.4 the Mel-energies representations for all the corresponding events are depicted, for comparative visual inspection of their spectral content.

4.4 Joint use of NMF and classification for overlapped AED

In general, NMF-related methods can be separated in those that exploit the NMF activations directly to perform event detection [16, 17] (like our approach in the previous section), and in those that employ a classifier trained on these activations [18, 19]. Based on the fact that NMF-based approaches can benefit from the creation of a Mixture of Local Dictionaries (MLD) [21], in [20] the authors propose a classifier-based NMF system using MLDs for improved detection performance, reaching the 1st place in the Task-2 of DCASE'16 Challenge.

In this section, we investigate the performance of classifier-based NMF methods for detecting

overlapping acoustic events. We provide evidence that the performance of classifier-based NMF systems deteriorates significantly in overlapped scenarios in case mixed observations are unavailable during training. To alleviate this problem, we propose the generation of mixed observations using the isolated ones available, and subsequently their incorporation in the training data. For the artificial mixing procedure, we use a k-means based method for each pair of events. The method of MLD is employed for the building of the NMF dictionary using both the isolated and artificially mixed data. Finally, an SVM classifier is trained for each of the isolated and mixed event classes, using the corresponding MLD-NMF activations from the training set. The proposed system, tested on two experiments with a) synthetic and b) real events, outperforms the state-of-the-art classifier-based NMF system in the overlapped scenarios.

The section is organized as following: Section 4.4.1 presents and discusses the drawbacks of the two NMF-based alternatives that are compared with our system; Section 4.4.2 describes the artificial generation of mixed data and the outline of the proposed method, and Section 4.4.3 reviews the experimental framework and reports our results.

4.4.1 Existing NMF-based methods for AED

We will present briefly two popular methods for NMF-based AED. The first can be considered as the baseline, as it is the simplest one: Sparse-NMF with thresholding. This is an approach similar to 4.3.2 (with activations only), but with the traditional NMF formulation instead of the convolutive one. The second is a classifier-based MLD-NMF method presented in [19, 20]. We will discuss the drawbacks of these two methods for isolated/overlapped acoustic event detection.

• **Sparse-NMF approach:** The Sparse-NMF method of eq. (4.3) is used here as a baseline. Regarding the building of the dictionary, by using training data consisting of isolated event instances ("exemplar" based approach), a sufficient number of atoms is extracted and stored in the dictionary for each class of interest, resulting in the total dictionary matrix $\mathbf{W}$. Then in the detection step, a simple thresholding on the activations of matrix $\mathbf{H}$ decides for the existence of each event in each frame. We can note two main disadvantages in this traditional method. The first is that the threshold-based decision in the detection step cannot be considered as the best choice in terms of robustness. The second and more important is that, as pointed out in [21], the convex cones created by the bases of the sub-dictionaries of the different classes may often overlap with each other. This means that new observations that fall in the overlapped regions can be reconstructed in many different ways (unstable activations), which may result to failures in the classification (e.g. false alarms).

• **SVM-based NMF approach with MLD dictionary:** This method essentially refers to the core system of [19, 20]. This system attempts to overcome the drawbacks of the aforementioned traditional sparse-NMF method, by employing an MLD dictionary framework and an SVM classifier for the final detection step. The MLD-based dictionary generation eliminates overlaps between convex cones and produces more stable activations, which are used for training robust SVM classifiers. As shown in the flow diagram in Figure. 4.5 (component blocks depicted in black), the

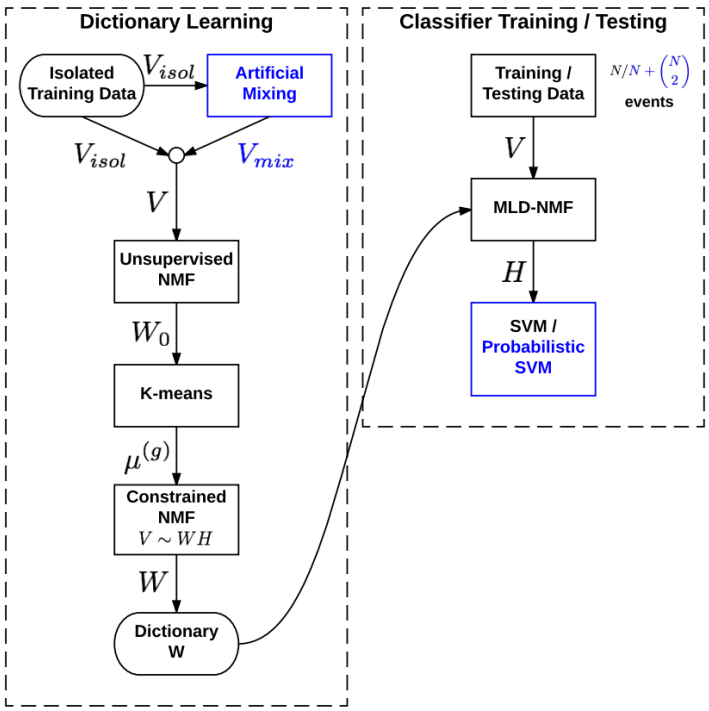


Figure 4.5: Block-diagram of the proposed AED method combining NMF with an SVM classifier.

method consists of two main parts: dictionary learning and classifier training.

For dictionary learning, the feature matrix $\mathbf{V}$ containing all training data is decomposed into an initial basis matrix $\mathbf{W}_0$ by basic unsupervised NMF. Next, by applying k-means to $\mathbf{W}_0$, G centroids $\mu^{(g)}$ are obtained, with $g \in \{1, ..., G\}$ denoting the centroid's index. The final MLD dictionary $\mathbf{W}$ consists of G sub-groups (of K_g bases each), which model acoustic atoms $\mathbf{W} = [\mathbf{W}^{(1)}...\mathbf{W}^{(G)}]$. The MLD dictionary is learned by minimizing the following objective:

$$D(\mathbf{V}||\mathbf{WH}) + \eta \sum_{g} D(\mu^{(g)}||\mathbf{W}^{(G)}) + \lambda \sum_{t} \Omega(\mathbf{h}_t) \quad ,$$

where $\mathbf{h}_t$ denotes the column vector of $\mathbf{H}$ at time frame t. The second term is a constraint that favors bases of sub-groups to be similar with $\mu^{(g)}$, so that the resulting convex cones are compact. The third term preserves group-sparsity in the solution.

Concerning classifier training, for each class considered, an activation matrix $\mathbf{H}_i$ is extracted from its corresponding training spectrogram $\mathbf{V}_i$ by MLD based NMF with the global dictionary $\mathbf{W}$. Then, the column vectors $\mathbf{h}_{t(i)}$ of $\mathbf{H}_i$ at each time frame t are used as feature vectors to train a linear SVM classifier. A multi-class SVM is trained using the one-against-all approach. This method seems to solve the problems of the traditional sparse-NMF approach in the isolated AED case. Nevertheless, we should point out one possible drawback in the case of overlapping scenarios: The classifiers are trained for each class of interest using its corresponding isolated data. This makes the classifier vulnerable in the presence of unseen mixed data. An observation of a mixed event containing classes i and j will not necessarily be classified correctly by both the classifiers of i-th and j-th event.

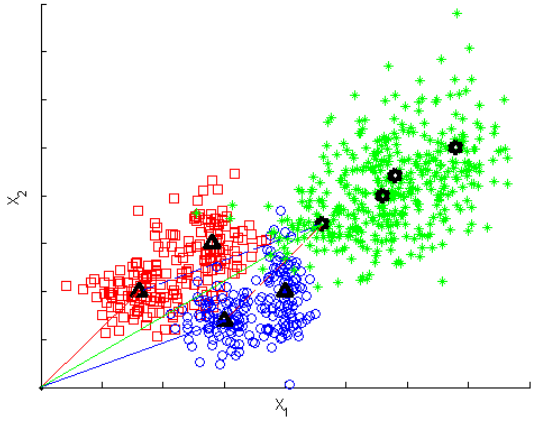


Figure 4.6: Generation of mixed data (green) from a pair of isolated events (blue and red). Toy example, with two features "x_1" and "x_2".

4.4.2 Proposed method

Our method attempts to solve the deficiency of the previous approach in overlapped scenarios, by considering mixed data in the training and testing stages. The block-diagram of the proposed method is depicted in Figure. 4.5 (black and blue component blocks).

• **Dictionary learning**: Our goal is to include mixed data in the dictionary learning procedure. Considering the difficulty of having enough mixed data available, we propose a method for artificial generation of such data. Assuming linearity of features, the method acts in the feature and not in the signal domain. The basic idea is shown in Figure. 4.6. In order to create representative observations of the mixed data, we try to combine (sum) representative observations from each of the two events considered. Given a number of centroids C and a percentage α, we first perform k-means clustering with C clusters in the feature space of each event. Then $\alpha\%$ of the samples in each cluster are selected. Finally, we consider all the combinations (addition) between the selected samples of the two classes. After mixed data generation, both isolated and mixed data are used as input to the MLD dictionary learning procedure. In this way, bases created in the final dictionary may correspond to overlapped events too.

• **Classifier training**: At the classifier training stage, instead of training N classifiers (N is the number of events), we train $N + \binom{N}{2}$. In addition, since we are modeling all possible events (isolated and mixed), we train linear probabilistic SVMs, and at the testing stage we choose the event with the highest score at each frame.

4.4.3 Databases and experimental framework

We perform our experiments on two datasets, with the one containing synthetic events and the other real events. In the case of the synthetic event dataset, we generated artificial spectral patches for 5 synthetic events, while in the real event case, we extracted spectral patches from 5 real events

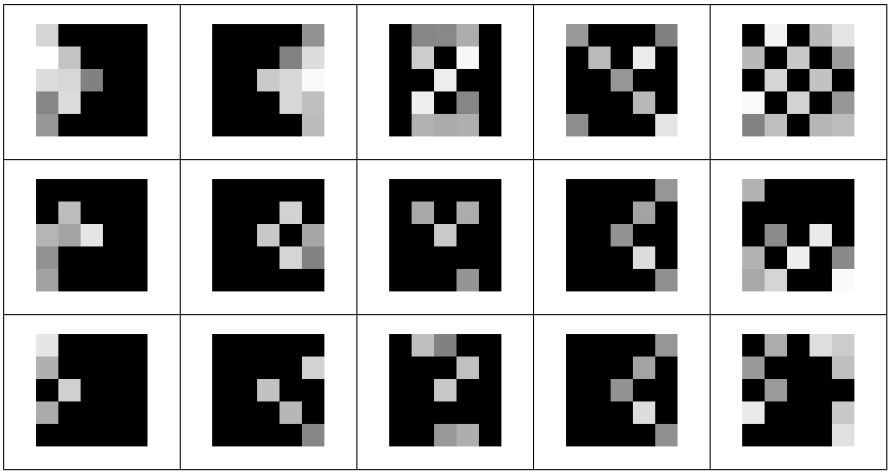


Figure 4.7: Different instances for each of the 5 synthetic events. In the first row, the most complete version of the spectral patch of each event is depicted. In the second and third rows, variations of them are shown for each event, where some active tiles may disappear.

contained in the database designed for Task 2 of the DCASE'16 Challenge (office-related events: drawer, phone, keys, speech, and doorslam).

In both datasets, the performance of different methods is evaluated in both isolated and overlapped scenarios. In the isolated case, testing sequences of isolated spectral patches are created, whereas in the overlapped case, sequences of mixed spectral patches are generated. A mixed spectral patch results from the superposition of two isolated spectral patches from the corresponding testing dataset. Regarding the spectral patch extraction, in the case of synthetic events, we generate 5x5 spectral patches with the following procedure: The spectral patches of each event are characterized by a particular pattern which is slightly varying its structure in the different instances (see Figure. 4.7). To introduce variability, each time some of the active "tiles" of the pattern can be missing (up to 5), while the active "tiles" take random positive values in the $[0.5, 1]$ interval. Random noise is also added after the generation of each spectral patch. In the case of real events, spectral patches have dimension 100x10 and are composed of 100 Mel-filterbank energies in 100 ms intervals (10 frames).

Finally, regarding the partition into training and testing sets, in the real event case, we partitioned the training data of DCASE'16 Challenge, so that 80% of event recordings is used for training and the remaining 20% for testing purposes. In the synthetic event case, we generated a small number of instances per event (30) for building the training set. For both databases, the testing sequences contain 1000 spectral patches for both isolated and overlapped scenarios. We should note that, due to the way we build our synthetic testing sequences, when overlap occurs, it affects the entire duration of the spectral patches involved. In this way, our problem can be also considered as classification of spectral patches of acoustic events with temporal information.

Table 4.4: Performance of the different systems for the synthetic data scenario in terms of F-score (%).

Method	Local opt.			Global opt.		
	Isol	Overl	Avg	Isol	Overl	Avg
sparse-NMF	95.10	95.82	95.46	95.21	93.53	**94.37**
SVM&MLD-NMF	96.78	77.23	87.00	94.39	77.23	85.81
Proposed	96.42	94.80	**95.61**	92.30	91.16	91.73

Table 4.5: Performance of the different systems for the real data scenario in terms of F-score (%).

Method	Local opt.			Global opt.		
	Isol	Overl	Avg	Isol	Overl	Avg
sparse-NMF	78.36	78.54	78.45	75.49	77.52	**76.51**
SVM&MLD-NMF	85.83	61.76	73.79	83.96	61.76	72.86
Proposed	85.79	74.49	**80.14**	82.00	68.86	75.43

4.4.4 Results

In Tables 4.4 and 4.5, the comparative results for the three different methods are presented in terms of F-score, for both isolated and overlapped scenarios and under two different experimental setups, for the two event datasets. In the first setup (Local opt.), optimization of the various parameters of the methods is performed in each scenario separately, while in the second (Global opt.) optimization is performed only once for the whole testing procedure. In fact, "Local opt." assumes prior knowledge of overlap existence.

In Table 4.4 we can draw three major conclusions: First of all, our proposed method clearly outperforms the state-of-the-art SVM&MLD-NMF based method in the overlapping scenarios, both in the "local" and "global" setups, achieving 77.16% and 61.18% relative error reductions respectively. In fact, SVM&MLD-NMF performance degrades significantly in the presence of mixed events. Next, we can observe that the performance of the baseline sparse-NMF approach is stable across the different scenarios and setups, achieving also the best F-score in the "global" optimization setup. We thus conclude that in the case of quite simple and discriminable events this baseline is a good option for both isolated and overlapped scenarios. Finally, only our proposed method seems to be affected significantly by using global optimization instead of the local one. It seems that parameter α, which controls the amount of mixing data included in the training phase, has strong influence on the behavior of our method.

In Table 4.5, corresponding results for the real-event scenario are presented. Similarly to the synthetic case, we can again notice a significant performance degradation of the SVM&MLD-NMF method when we move from the isolated to the overlapped scenario, as well as the superiority of the proposed method in the overlap case (33.29% and 18.57% relative error reduction in the "local" and "global" setups, respectively). Also, the baseline sparse-NMF method shows again stable performance across different scenarios. However, as expected, in this more challenging case of real events, both the SVM&MLD-NMF and proposed methods perform significantly better than the baseline in the isolated scenario. Finally, as before, among the three methods, our

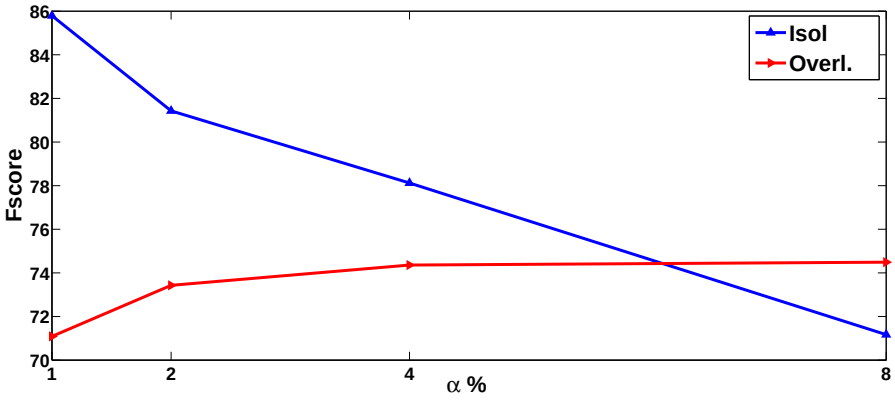


Figure 4.8: Performance of the proposed method (4.4.2) in both the isolated and overlapped scenarios, as the percentage α of mixing varies.

approach is affected the most by the switch from the "local" to the "global" optimization setup.

By summarizing the results, we can claim that the classifier-based SVM&MLD-NMF approach outperforms the baseline sparse-NMF based one in the isolated event scenario. This is important, as the fact is that the isolated scenario is by far the most frequent under realistic conditions. However, if we want to test the system under more challenging overlapping conditions, the performance of the existing method deteriorates. Our proposed method, by incorporating mixed data in the training phase, succeeds to improve performance significantly under overlapped conditions and to also provide better results overall. However, there exists one drawback: our method is strongly affected by the amount of mixed data employed for training. This is depicted also in Figure. 4.8, where the performance of the proposed method is shown on the real events dataset for both the isolated and overlapped cases, as the mixing parameter α varies. As α increases, performance improves also in the overlapping case, but at the same time degrades (with a higher rate) in the isolated case. With knowledge of the expected degree of overlap in our dataset, an optimal value of α could be chosen.

4.5 Conclusions

In this chapter, at first we presented a sparse-CNMF based system for single-channel overlapped audio event detection, employing an efficient dictionary building method and a novel detection approach. Attention was also given to background noise modeling and on experimentation with different possible feature sets for the CNMF framework. Results obtained on Task 2 of the DCASE'16 Challenge were satisfactory, significantly outperforming the NMF-baseline provided, and reaching the 5th place in the rankings of this task.

Next, we investigated the performance of state-of-the-art NMF approaches for single-channel overlapped acoustic event detection. We provided evidence of degradation of the existing method's performance under highly overlapped conditions, and we proposed a new method which tries to alleviate this problem by employing a module for artificial generation of mixed data in the training

phase. Probabilistic SVMs are also employed in the final classification step using all available classes (isolated and mixed). Results obtained on experiments with synthetic and real events were promising, outperforming the existing method in overlapping scenarios while also preserving good performance in the isolated ones.

Chapter 5

NMF-based Multi-channel Overlapped Acoustic Event Detection

5.1 Introduction

In the previous chapter, we have developed and presented NMF-based approaches that tackle the problem of overlapped AED by using single-channel audio. However, whenever the setup permits it, exploiting information from multiple channels can be valuable. In this chapter, we propose two multi-channel extensions of NMF, suitable for overlapping AED [127]. The single-channel baseline, upon which we build our methods, is again a sparse-NMF based approach, performing detection at the frame level. Our first method combines the different microphones at decision level, by summing their activation matrices to obtain an average confidence for the activation of each class. Our second method considers the optimization of a novel objective function, containing a multi-channel KL-divergence reconstruction term and a multi-channel class sparsity term. At each time frame, this class sparsity term forces the NMF solutions to contain only a small number of activated classes across all microphones. In this way, the updates of the activation matrix for each microphone at each iteration are informed by the activations from the other microphones too, and this leads to robust solutions where most of the microphones should agree. For our experiments we use the publicly available ATHENA database [80], which contains real multi-channel recordings from a smart-office environment including sixteen acoustic events and five types of background noise. The results confirm the superiority of the proposed multi-channel approaches over the single-channel baseline.

The chapter is organized as follows: Section 5.2 presents both the single-channel baseline and the two proposed multi-channel approaches; Section 5.3 describes the database and experimental framework employed and reports our results and, finally, Section 5.4 summarizes our conclusions.

5.2 Methods

5.2.1 Single-channel baseline

Sparse-NMF is employed as the single-channel baseline system. For the m^{th} channel, given the observed matrix $\mathbf{V}_m$ and the dictionary matrix $\mathbf{W}_m$ containing atoms for all acoustic events, sparse-NMF derives the activation matrix $\mathbf{H}_m$ by minimizing the objective function:

$$J_m = D(\mathbf{V}_m|\mathbf{W}_m\mathbf{H}_m) + \lambda\|\mathbf{H}_m\|_1 \ . \tag{5.1}$$

When employing the generalized KL-divergence error cost function D, the solution to (5.1) can be obtained by means of the iterative update:

$$\mathbf{H}_m \leftarrow \mathbf{H}_m \odot \left\{\mathbf{W}_m^T(\mathbf{V}_m \oslash (\mathbf{W}_m\mathbf{H}_m))\right\} \oslash \left\{\mathbf{W}_m^T\mathbf{1}_V + \lambda\mathbf{1}_H\right\} \ ,$$

where $\odot$ and $\oslash$ denote element-wise matrix multiplication and division, and $\mathbf{1}_V$ and $\mathbf{1}_H$ are matrices with all elements equal to 1 and dimensions equal to $\mathbf{V}_m$ and $\mathbf{H}_m$ respectively. Matrix $\mathbf{H}_m$ is initialized with random positive values, and for its computation we apply 100 iterations.

After obtaining matrix $\mathbf{H}_m$, for each time frame, the activations for each class are summed across all their atoms resulting in a new matrix $\mathbf{H}'_m \in \mathbb{R}_{\geq 0}^{C\times N}$, where C denotes the total number of event classes. Finally, detection is performed by thresholding, i.e., class c is considered active at time frame n, if $\mathbf{H}'_m(c,n) > \theta_H$, where θ_H is a suitably chosen threshold.

Regarding the creation of dictionary matrix $\mathbf{W}_m$, we use the "exemplar" based method: Using extracted isolated training instances from each event, we create the class-specific sub-dictionaries $\mathbf{W}_m^{(c)} \in \mathbb{R}_{\geq 0}^{P\times R_c}$, for $c = 1, ..., C$, by clustering the available isolated instances with the k-means algorithm (R_c centroids are selected). The total dictionary $\mathbf{W}_m$ is then created by concatenating the C sub-dictionaries, i.e., $\mathbf{W}_m = \left[\mathbf{W}_m^{(1)}, ..., \mathbf{W}_m^{(C)}\right] \in \mathbb{R}_{\geq 0}^{P\times R}$. Figure. 5.1 shows an example of a sparse-NMF decomposition resulting in the activation matrix $\mathbf{H}$. In particular, matrix $\mathbf{V}$ contains the log-Mel energies representation of a sound sample, containing three overlapping acoustic events ("cough", "spoon", and "Skype call"), and matrix $\mathbf{W}$ depicts the sub-part of the dictionary containing atoms for these three events. Red-dashed rectangles drawn on $\mathbf{H}$ indicate the ground-truth activation for each event. Clearly, atoms of each event are activated mostly in the areas where the corresponding event occurs.

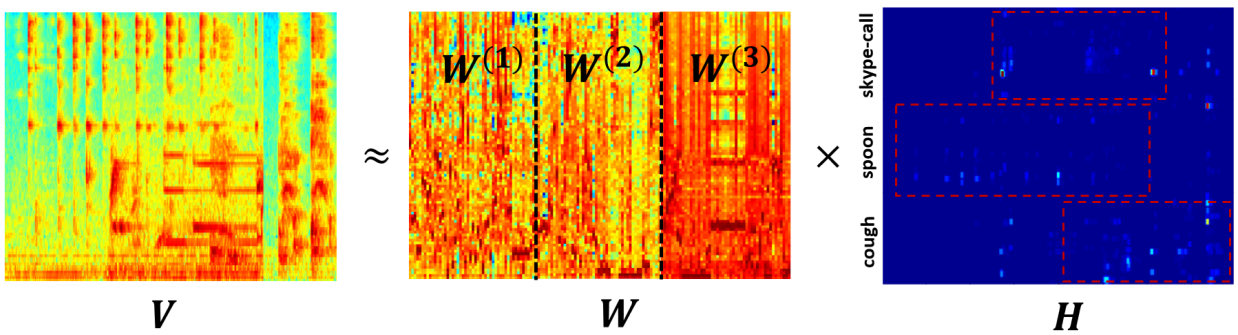


Figure 5.1: Sparse-NMF decomposition for an example of three overlapping acoustic events.

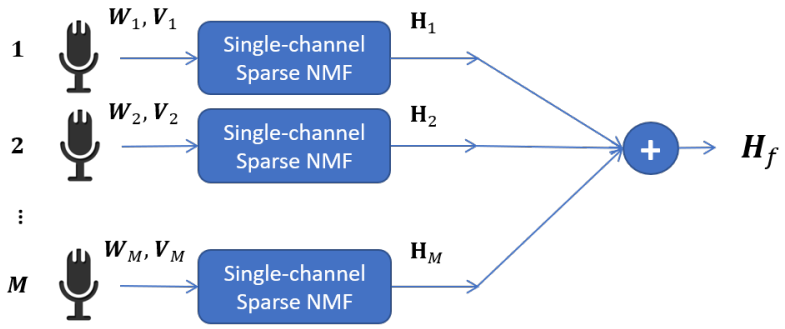


Figure 5.2: Block diagram for the sum of channel activations scheme for multi-channel sparse NMF.

5.2.2 Sum of channel activations

In NMF-based methods, activations produced for each class are directly related to the confidence about its existence. In this multi-channel approach, we combine the different channels at the decision level, expecting more reliable results compared to those based on a single channel alone.

At first, each channel m acts independently from the others, performing single-channel sparse-NMF by using its own observation matrix $\mathbf{V}_m$ and dictionary matrix $\mathbf{W}_m$ and outputs its activation matrix $\mathbf{H}_m$. Then, as shown in Figure. 5.2, the activations from all channels are averaged to obtain the final activation matrix $\mathbf{H}_f$:

$$\mathbf{H}_f = \frac{1}{M} \sum_{m=1}^{M} \mathbf{H}_m, \tag{5.2}$$

where M is the total number of channels considered. Finally, summing of activations per class and thresholding follow, as in the single-channel case.

5.2.3 Multi-channel NMF with class sparsity

In this approach, we extend the objective function of single-channel NMF in a multi-channel fashion. Towards this end, we first transform the reconstruction error term to contain the sum of KL-divergence errors from all channels. In this case, in each reconstruction term, each channel uses a global dictionary matrix $\mathbf{W}$ that is built similarly to the single-channel case approach discussed at the end of Section 5.2.1, but with the modification that atoms are sampled for each class from all training data across all channels. Further, we add a multi-channel class sparsity constraint as a second term. This constraint is used to regularize the NMF solutions so that, at each time frame, only a few classes are activated across all channels. As a consequence, the channels are forced to act in a collaborative way and find solutions to which they agree.

The multi-channel objective function J is defined as:

$$J = \sum_{m=1}^{M} D(\mathbf{V}_m | \mathbf{W}\mathbf{H}_m) + \lambda \sum_{n=1}^{N} \Omega(h_{1,n}, ..., h_{M,n}) \ , \tag{5.3}$$

where $\mathbf{H}_m$= $[h_{m,1}, ..., h_{m,N}]$ and $h_{m,n}$= $[{h^{(1)}_{m,n}}^T, ..., {h^{(C)}_{m,n}}^T]^T$, i.e., $h_{m,n}$ is the n^{th} column of the activation matrix $\mathbf{H}_m$. The class-sparsity function Ω is defined as:

$$\Omega(h_{1,n}, ..., h_{M,n}) = \sum_{c=1}^{C} \log(\epsilon + \sum_{m=1}^{M} \|h^{(c)}_{m,n}\|_1) \ , \tag{5.4}$$

where $h^{(c)}_{m,n}$ denotes the part of the activation column that is related to event class c. This function can be viewed as a multi-channel extension of the term used in [21, 131] to imply group sparsity. In our case, groups are the event classes considered.

By majorizing the second term of (5.3), we obtain the following updates for activation matrices $\mathbf{H}_m$, for all $m \in \{1, ..., M\}$, $c \in \{1, ..., C\}$, $n \in \{1, ..., N\}$:

$$\mathbf{H}_m \leftarrow \mathbf{H}_m \odot \left\{\mathbf{W}^T(\mathbf{V}_m \oslash (\mathbf{W}\mathbf{H}_m))\right\} \tag{5.5}$$

$$h^{(c)}_{m,n} \leftarrow h^{(c)}_{m,n} \oslash \left\{(\mathbf{W}^{(c)})^T \vec{1}_v + \lambda \vec{1}_{h_c} / (\epsilon + \sum_{m'=1}^{M} \|h^{(c)}_{m',n}\|_1)\right\} \ , \tag{5.6}$$

where ϵ is a small positive constant, while column vectors $\vec{1}_v$ and $\vec{1}_{h_c}$ have all their elements equal to 1 and dimensions $P \times 1$ and $R_c \times 1$ respectively. In our experiments, $\mathbf{H}_m$ is initialized with random positive values, and updates (5.5), (5.6) are applied iteratively for 100 iterations.

From (5.6) it can be seen that, at each iteration, the update for each channel is also affected by the activations of the other channels. In particular, when the total activation across all channels (as computed at the previous iteration) is low for a given class, the activations of the m^{th} channel at the current update are suppressed for that class. In (5.4), parameter λ tunes the size of the impact of this class-sparsity constraint: high values of λ will lead to solutions with only a few different classes activated at each frame.

After obtaining the M different activation matrices for all channels, we compute the final activation matrix $\mathbf{H}_f$ as in the previous method, using (5.2). Finally, we should note that dictionary matrix $\mathbf{W}$ that is used in updates (5.5) and (5.6) of $\mathbf{H}_m$ has its columns (atoms) normalized so that their elements sum to 1.

It is worth mentioning that in our work we employ the multiplicative updates approach for solving the NMF task, mainly because of their widespread use in related works and also due to their high reproducibility. Alternative efficient algorithms for solving the NMF task have also been proposed and applied successfully in the literature [120–122].

5.3 Experiments

5.3.1 Database

We perform our experiments on the ATHENA multi-modal database [80], captured in a smart office environment. In total, the dataset contains 240 one-minute long sessions of real recordings divided into a training and test set. This database is suitable for multi-channel overlapped AED,

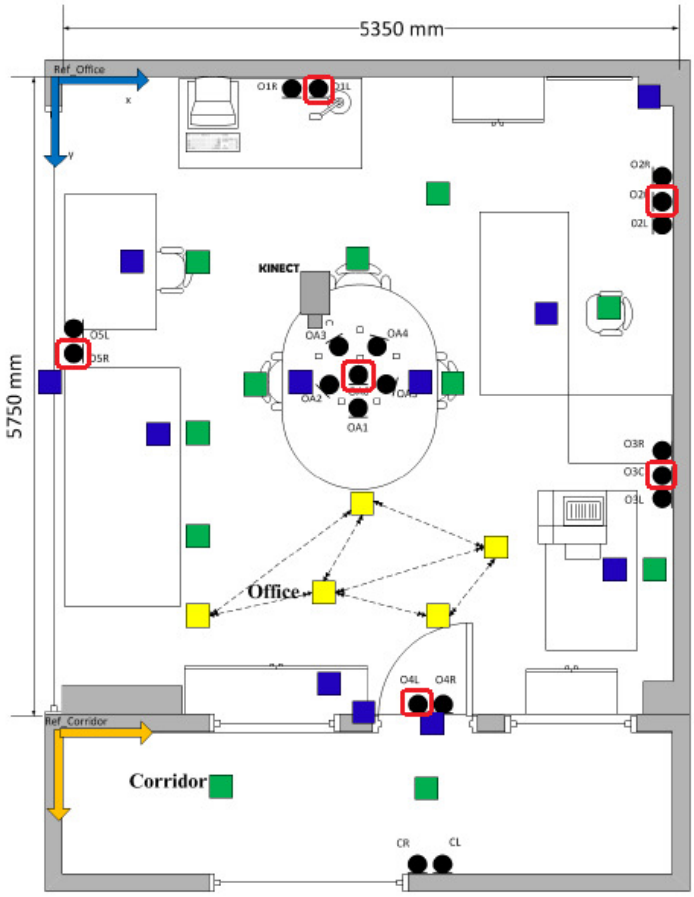


Figure 5.3: Floorplan of the smart office used in the ATHENA database recordings. Microphones (black), speaker (green and yellow) and event (blue) positions are depicted. Six microphones were used in our experiments, marked with a red square.

as it contains speech plus fifteen acoustic events captured from multiple microphones (20 in total) installed on the ceiling and walls of the smart space (see also Figure. 5.3). The acoustic events are categorized according to their average duration to long events ("walking steps", "cellphone ring", "keyboard", "glass fill", "coffee spoon", "Skype call", "cough", "paper work", "window open/close") and short events ("mouse click", "keys", "knock", "chair moving", "switch on/off", "door open/close"). To better approximate a realistic scenario, five different types of acoustic backgrounds are also considered in the various sessions (ambient noise, fan, radio music, vacuum cleaner, silence). Highly overlapped scenarios (40% of speech overlaps with other events) and adverse noise conditions make this dataset challenging for overlapped AED. The ATHENA database is publicly available [1].

5.3.2 System implementation details

Next, we provide details about the various parameters of the systems described. Regarding audio feature extraction, we employ 100 Mel-filterbank energies computed in windows of 30 ms

[1] http://cvsp.cs.ntua.gr/research/athenadb

duration and with a 10 ms shift. Concerning the number R_c of atoms selected per class in the dictionary and the sparsity parameter λ, we experimented with various combinations: $R_c \in \{20, 40, 60, 80, 100, 120, 150\}$ and $\lambda \in \{0.5, 1, 2, 4, 8, 16, 32\}$. Also for better background modeling, we extract and store in the dictionary R_c atoms for each type of background considered.

As a post-processing stage for the detection system, after thresholding the activations with θ_H, for each class, we unify active segments that occur with time distance less than t_u sec and delete active segments with duration shorter than t_d sec. All parameters were optimized on the development set (see Section 5.3.3).

Finally, for our multi-channel approaches we employ the six microphones that are highlighted with red marks in Figure. 5.3. The purpose of our selection was to uniformly sample the acoustic space.

5.3.3 Experimental setup

In our experiments, we have considered three types of acoustic backgrounds, namely ambient noise, fan, and silence, which are more common in real-life scenarios. These backgrounds cover roughly 1 hour of recordings in the training set and 1 hour in the test set. From the corresponding part of the training set, we select isolated instances of events and use them for dictionary building. We also divide the test set into development and evaluation sets, of 30 min duration each. The optimization of all system parameters was performed on the development set. The metrics used for evaluation and comparison of our methods are frame-based Recall, Precision, and F-score.

5.3.4 Results

The three methods are evaluated and compared in the 30 min long recordings of the evaluation set. As a baseline for our experiments we consider the average single-channel F-score, computed as the mean of the F-scores of the different single-channel NMF systems (6 in total). This corresponds to the expected performance we would get if we chose randomly a microphone in the smart space. As an alternative baseline, we also show the results of the oracle single-channel, i.e. the best-performing channel for the given evaluation set (central microphone of the ceiling in our case). In Figure. 5.4, the results in terms of Recall, Precision, and F-score are depicted for the baseline and the two multi-channel approaches. First, we can observe that both multi-channel approaches outperform the single-channel baseline, achieving 6.80% and 15.44% relative error reduction in terms of F-score (the sum-of-activations and multi-channel NMF methods, respectively). Further, both multi-channel methods show significant improvements over the oracle single-channel result. Also, multi-channel NMF with class sparsity performs better than the sum-of-activations method, achieving 9.27% relative error reduction. Finally, the multi-channel NMF approach shows the best results in all metrics, performing also at a slightly more balanced point between Recall and Precision than the sum-of-activations method.

We can also observe that, in general, AED performance is relatively low, indicating the challenging nature of the database. Such can be primarily attributed to the highly overlapped conditions, the adverse background noise, and the large variety of event classes considered.

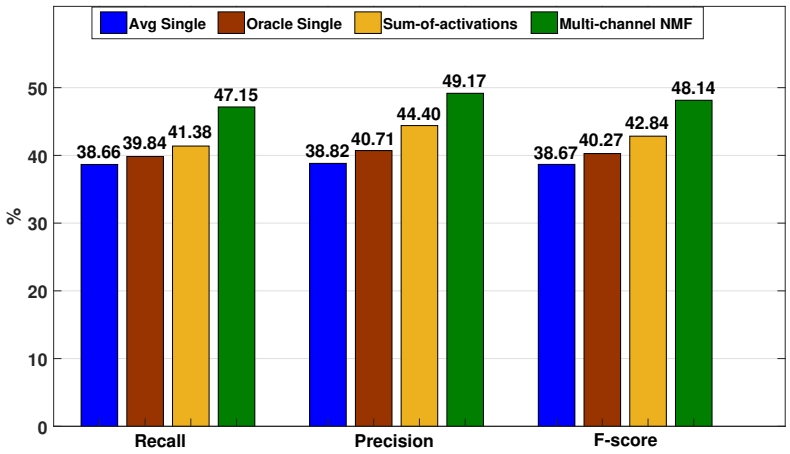


Figure 5.4: AED results on the evaluation set of the ATHENA database, depicted in terms of Recall, Precision, and F-score for the three different approaches of Section 5.2.

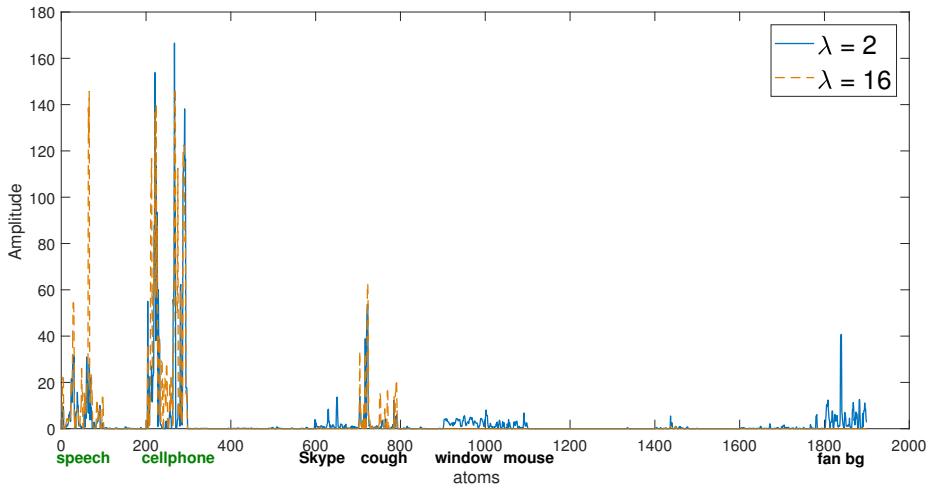


Figure 5.5: Activations across atoms in activation matrix $\mathbf{H}_f$ of the multi-channel NMF method. Activations are averaged for a given time interval of 2 sec in duration and shown for two different values of sparsity parameter λ. Events with green colors overlap in the ground truth.

Finally in Figure. 5.5, we can observe the effect of class sparsity parameter λ on the solutions for the activation matrices. In particular, we show the activations of the final activation matrix $\mathbf{H}_f$, averaged in time, for a given time interval where two acoustic events overlap ("speech" and "cellphone ring"). We can see that, when increasing the class sparsity parameter, the solutions become more concentrated on the atoms of the given events. When λ becomes lower, atoms from more classes become activated, leading to false alarms in the detection. In the given example, when λ=16 only one false alarm occurs for event "cough", while for λ=2 false alarms also occur for events "Skype call", "window open/close", and "mouse click".

5.4 Conclusions

In this chapter, we proposed two multi-channel NMF approaches for overlapped AED. The first method combines at decision level the independent sparse-NMF outputs from different channels. The second method considers the optimization of a novel multi-channel NMF objective function including a class sparsity term. Such term introduces robustness, as it forces the channels to activate only a few classes that they agree on. Both proposed multi-channel methods outperformed the single-channel baseline, with the second achieving satisfactory improvements.

Chapter 6

Deep Learning for Multi-channel Overlapped Acoustic Event Detection

6.1 Introduction

So far, we have considered NMF-based approaches for the overlapped AED task. Indeed, when it comes to overlapping scenarios, NMF constitutes a suitable choice, as it has the natural ability of detecting multiple events occurring simultaneously. Additional advantages of NMF techniques include their interpretability, their ability to be efficiently trained on small amounts of available data (even with no need for overlapped instances being available during training), and their robustness to noisy conditions. However, they have some disadvantages too, mainly including their running-time efficiency and discriminative capability (when it comes to a large number of event classes). An alternative family of approaches that have better discriminative power and also have the ability to model simultaneously activated events are the deep-learning based methods. Indeed, deep-learning approaches have been successfully applied to the AED task in recent years [22–25]. However, compared to NMF methods, deep-learning methods require much larger amounts of training data in order to perform well, as well as the existence of overlapped instances during training in order to operate in overlapped scenarios.

As a result, the standard training approach for overlapped AED in deep-learning based systems is to feed a multi-label neural network with overlapped instances that either exist in training or are artificially generated from the available isolated instances. However, the number of possible event combinations that need to be modeled grows rapidly as the number of event classes or the number of simultaneous events occurring (polyphony level) increase. In such cases, efficient training of the network can be problematic, as it depends on the existence or generation of sufficient and diverse overlapped data, thus rendering this approach not scalable.

An alternative approach that mitigates this issue is to employ a sound source separation network as a pre-processing step to AED, aiming in this way to approximately transform the overlapped task into the isolated one. Significant progress has been made in the domain of sound source separation in recent years, including mostly works on speech separation [132–135], and lately also on universal sound separation [136, 137]. Based on the above, some works employ

such systems, reporting improved results for the single-channel overlapped AED task [138, 139]. Also in [140], in a multi-channel setup, the authors train their network using beamformed signals from various directions of arrival with respect to the microphone array.

In this chapter, we present our work on the task of overlapped acoustic event classification which is a special case of AED when the time boundaries of active events are considered known a priori (classification only). In this context, we propose for the first time the combination of a multi-channel sound separation network with a multi-label AED system for addressing the overlapped AED task when the number of different event classes is large. In such a scenario, we examine how the proposed approach can reduce the performance gap of a AED system between the isolated and the more challenging overlapped cases. In particular, we employ a state-of-the-art multi-channel sound separation network in order to exploit, additionally to spectral content, the spatial discrimination of the events present in a mixture clip, while for the AED module we employ a CNN-based architecture suitable for AED. For the resulting pipeline, we examine both sequential and end-to-end joint re-training of the two modules, with the latter achieving the best performance. In addition, we propose the incorporation of a polyphony detection network, which can selectively apply the proposed system only to the overlapped instances during testing. Although our system is scalable to an arbitrary polyphony level, in this study, we examine the case of overlap with up to 2 simultaneous events. For our experiments we employ the ESC50 data collection [81], as it provides balanced data from a large variety of different event classes (50), and in order to design a multi-channel dataset, we combine it with real impulse responses from the DIRHA smart-home dataset [61]. Our results show that in this challenging overlapped scenario, and under moderate reverberation conditions, the proposed system can provide significant improvements over a baseline CNN-based AED network trained with the standard multi-label training approach.

This chapter is organized as follows: Section 6.2 provides the description of the several modules employed in our approaches; Section 6.3 describes the database and experimental framework used and reports our results; and, finally, Section 6.4 concludes the chapter.

6.2 System description

6.2.1 Baseline AED network

The architecture of the baseline AED network is depicted in Figure. 6.1. Given an input audio signal $\mathbf{s} \in \mathbb{R}^N$ (with N denoting the number of signal samples), the feature extraction stage computes 64-band Log-Mel filter-bank energies (logFBE) and their Deltas using 0.4 sec Hanning windows with 0.2 sec shift, producing the feature matrix $\mathbf{X^{(s)}} \in \mathbb{R}^{128\times T}$, where T is the number of resulting time frames. The feature matrix $\mathbf{X^{(s)}}$ is fed to the network that consists of a 5-layer CNN block, followed by 2 fully-connected linear layers. The output $\mathbf{y} \in \mathbb{R}^C$ has dimension equal to the number C of event classes and is expected to have high values at the indexes of activated events. For the multi-label training we employ the binary cross-entropy loss function. During testing, active sources are decided by applying a threshold on the output.

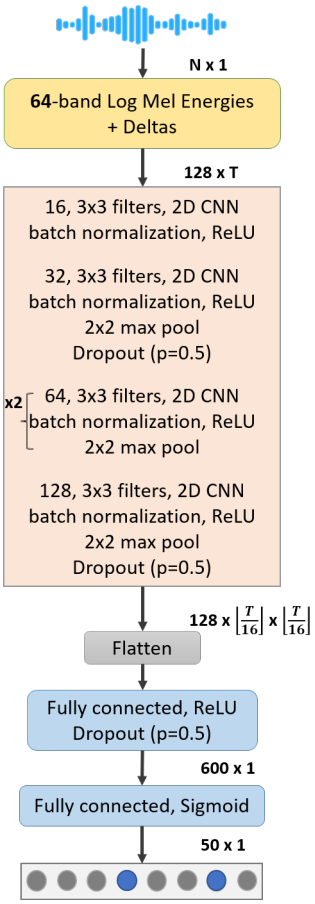


Figure 6.1: Single-channel deep-learning baseline architecture for acoustic event classification.

6.2.2 Multi-channel separation network

In our work we employ a multi-channel separation network originally proposed for speech separation in [141]. In this method the authors essentially improve their previous work on FaSNet [142], which is a multi-channel filter-and-sum neural beamforming network operating in the time domain. The improvements include: (a) the incorporation of a transform-average-concatenate (TAC) module that makes the network invariant to the permutation and the number of microphones, and (b) the transition to a single-stage architecture where the filters for all channels are jointly estimated.

The network takes as input time-domain mixture signals from M microphones and outputs K time-domain separated signals. Regarding the loss function, similarly to [142], we use the mean squared error between the FBE representations of the original sources, as captured by a reference microphone, and the reconstructed sources at the output of the network. In our case, as reference microphone we consider the central microphone of a 3-channel linear array (see Section 6.3.1).

6.2.3 Proposed system

The proposed system, as shown in Figure. 6.2, combines the separation and the AED networks in a cascade. In particular, we employ the separation network as a pre-processing step, which provides

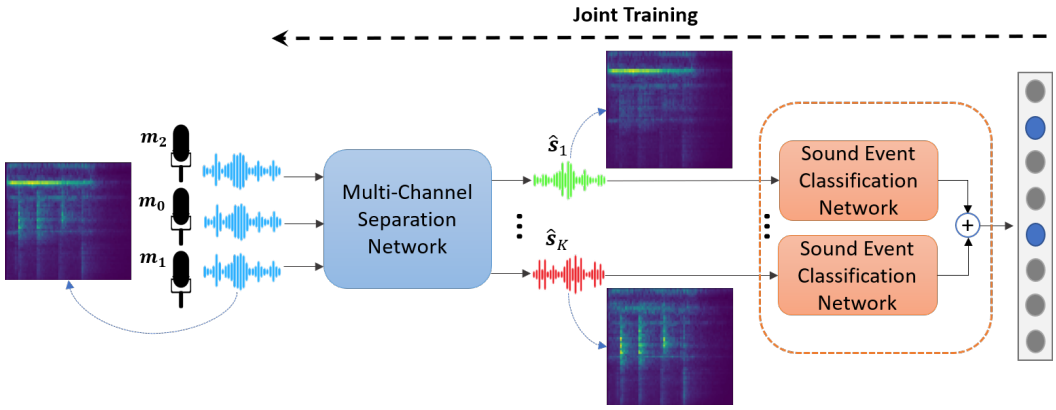


Figure 6.2: Pipeline of the proposed deep-learning system for multi-channel overlapped acoustic event classification.

the AED network with K separated signals in place of the original mixture. The idea is that, given a well-performing separation network, the overlapped task can be approximately reduced to classification of a set of isolated instances, therefore improving the performance of the system.

The AED network used in the proposed system applies an AED module with identical architecture with the baseline system for each of the K separated inputs $\hat{\mathbf{s}}_k$, $k = 1, \ldots, K$, and then averages their outputs. For training the proposed pipeline, we examine two approaches:

• **Sequential training:** In this case, we first train the separation network with mixtures that are artificially generated by the available isolated instances, as described in Section 6.3.2. Then, we train the AED network on the separated signals that result from the output of the separation network for the various mixtures.

• **Joint training:** In this case, the training consists of two stages. The first stage is the same with the sequential training, except that the AED network is trained on ground-truth separated signals. In the second stage, the two networks are jointly re-trained, using as input the mixture signals from the microphone array and as loss function the binary cross-entropy on the final output. In this way, the parameters of both networks can be fine-tuned towards the final objective of event classification.

Finally, we also examine the ensemble of the baseline AED network with the proposed system, by performing linear late fusion on their outputs, followed by thresholding.

6.2.4 Polyphony network

The proposed method is designed to operate on audio segments with overlapped events. In order to evaluate it in a realistic scenario with both isolated and overlapped instances, we need a module able to detect the polyphony level and selectively apply it only in the overlapped cases.

Polyphony classification modules based on deep learning have been recently employed with success in the literature [143, 144]. In our work, we implement a polyphony classification network that exploits both the spectral and the spatial information by using logFBE features in conjunction

with GCC-PHAT based features, computed for different pairs of microphones of the array. Similarly with [145], we consider the GCC-PHAT features as GCC spectrograms that are concatenated with the logFBEs to form the final feature matrix. In our case that we use a 3-microphone array, the network takes as input the feature matrix $\left[\mathbf{X}^{(\mathbf{s}_0)}; \mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_1)}; \mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_2)}\right] \in \mathbb{R}^{384\times T}$, where $\mathbf{s}_0$ is the signal captured by the central microphone m_0, and outputs a P-dimensional vector $\mathbf{y} \in \mathbb{R}^P$, where P denotes the maximum possible degree of polyphony (in this study, $P = 2$). For the polyphony network, we use the same architecture with the baseline AED network (just changed the output dimension of the last linear layer), and the cross-entropy as loss function.

6.3 Experiments

6.3.1 Database

For our experiments we employ the environmental sound classification (ESC50) dataset [81]. ESC50 contains 2000 5 sec-long audio clips from 50 different event classes, belonging to various sound categories such as animal sounds, natural soundscapes, human (non-speech) sounds, domestic sounds, and urban noises.

In order to create a multi-channel dataset, we convolve the audio clips with real room impulse responses (RIRs) from the DIRHA smart-home dataset [61]. In particular, we use a linear microphone array with 3 omni-directional microphones (spaced 15 cm apart) placed inside the living room of the DIRHA smart home, and 12 different locations with 2 possible orientations each for the event sound sources. With respect to the central microphone, the T_{60} reverberation times for the different source locations range from 0.58 to 0.83 sec, while their distances from 0.72 to 3.2 m.

6.3.2 Experimental setup

At first, all audio clips from ESC50 and RIRs from DIRHA are downsampled to 16 kHz. Before the convolution with the DIRHA RIRs, we pre-process the weakly-labeled audio clips of ESC50 as follows: similarly to [146], we first remove silent areas using an energy thresholding criterion, and then we split them to 1-sec segments with 80% overlap, thus producing about 34k clips in total. In this way, we obtain more samples to train our network, and also our system can operate at a finer temporal resolution. These audio clips are then split into training, validation, and test sets at a 8:1:1 ratio. In the split we ensure that different sets do not contain clips from the same recording.

In order to simulate a realistic scenario, we assume that for each set, 50% of their clips are observed as isolated instances and 50% as parts of overlapped instances. The audio clips are then convolved with RIRs to produce 1.5-sec long segments (by truncating longer parts). In the case of overlapped instances, we randomly choose a location and orientation for each event and mix them at SNRs between -2 and 2 dB. Overall, we end up with approximately 13.5k isolated and 6.5k overlapped instances in the training set, and 1.8k isolated and 0.8k overlapped for each of the validation and test sets. Also, by following the standard data augmentation paradigm, we further generate artificial mixtures from the observed isolated instances of each set by superposition. In

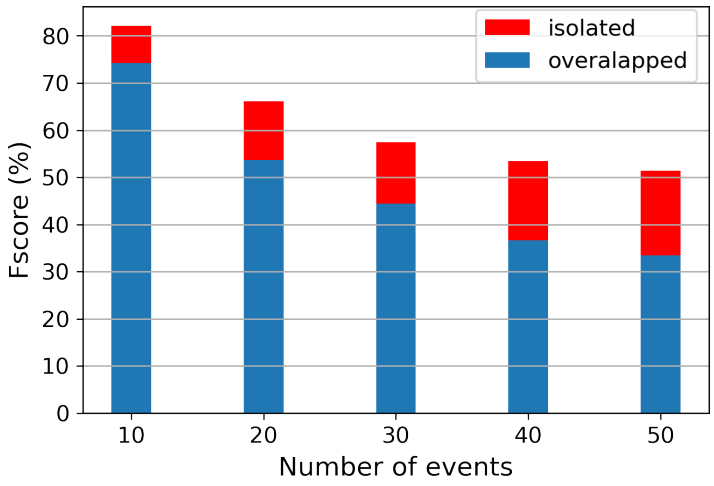


Figure 6.3: Performance of the AED baseline network for isolated and overlapped tasks for event sets of various numbers of classes.

this way, we also generate 30k overlapped instances for the training set (resulting in 36.5k total), and 2.2k for each of the validation and test sets (3k total each).

Regarding the evaluation metrics, for the multi-label AED task we employ the F-score metric, while for the performance of the polyphony network we use the classification accuracy.

6.3.3 Network training details

For training the networks, the Adam optimizer is used [147], with initial learning rate set to 0.001 and decreased to half every 30 epochs. All the networks are trained for 100 epochs, except the joint network that is re-trained for 30 epochs. In the end, the epoch with best performance on the validation set is kept. The batch size for the separation network is set equal to 20, while for the AED networks is set to 150. Finally, the separation network is trained on the set of 30k generated mixtures where separated ground-truth signals can be considered as known, and for the overlapped task the AED baseline network is trained on both 30k and 6.5k overlapped instances of the training set.

6.3.4 Results

In Figure. 6.3, we compare the performance of the baseline AED network for the overlapped and isolated tasks as the number of event classes considered increases. While the performance clearly degrades in both tasks, their gap progressively increases as the number of events adds complexity to the overlapped task. Given a well-performing separation network, our proposed pipeline aims to reduce this gap.

One way to improve performance in overlapped scenarios is to increase the training size. In Figure. 6.4, the performance of the AED baseline network for both isolated and overlapped tasks is depicted for different sizes of their training sets. As we can see, the performance in the overlapped scenario improves as we add more data to the training set, but at a decreased rate compared to

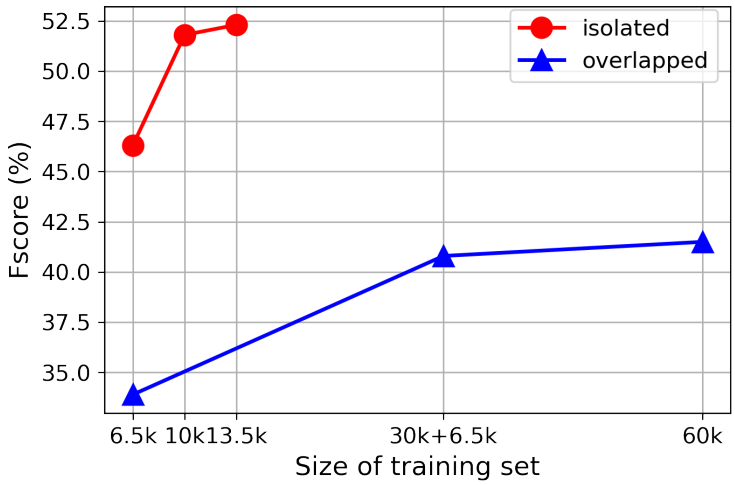


Figure 6.4: Performance of the AED baseline network for isolated and overlapped tasks for various sizes of the training set.

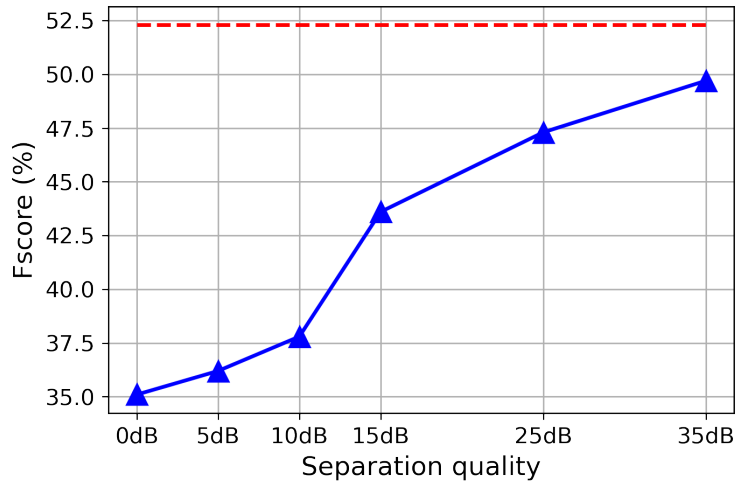


Figure 6.5: Performance of the AED network of Figure. 6.2 in the overlapped task (blue) for various levels of separation quality (measured in dB). The red dashed line corresponds to the performance of the baseline AED network on the isolated task, which can be considered as an upper limit.

the isolated scenario. Although we can artificially generate infinite overlapped examples, the contribution of the augmented data saturates at some point, as the diversity of produced mixtures from a given set is limited. On the other hand, in the isolated task higher F-score values are achieved for quite smaller training set sizes.

In Figure. 6.5, in an oracle experiment, we examine the performance of the AED network of Figure. 6.2 in the overlapped task (using 10k training samples and sequential training), in relation to several hypothetical levels of separation quality provided by the separation network. To simulate the outputs of the separation network, we artificially mix the isolated sources at different SNRs. While this experiment ignores the possible distortion artifacts that can be inserted by the separation network, it provides evidence that even when residuals of the undesired source are present in

Table 6.1: Performance of the various systems for the overlapped-event scenario, in terms of F-score.

System	F-score (%)	
	$\overline{T}_{60}$=0.61s	$\overline{T}_{60}$=0.80s
(A) Baseline (1 channel)	41.26	39.05
(B) Baseline (3 channels)	41.45	**39.33**
(C) Proposed - Sequential	44.72	38.41
(D) Proposed - Joint	**47.46**	38.75
Late Fusion (B+C)	46.20	41.52
Late Fusion (B+D)	**48.95**	**41.95**

each separated input signal, the separation module can significantly boost the performance of the AED network, provided that its separation quality exceeds a certain level ($\sim$10 dB). Indeed, it can be seen that as the separation quality increases, the overlapped task performance (in blue) approximates the isolated task performance (in red) of the baseline network.

Table 6.1 shows the performance of the various approaches for the overlapped task in terms of F-score for two different reverberation scenarios. In particular, the locations and orientations of the event sources are selected such as the mean reverberation time is 0.61s and 0.80s respectively. As a multi-channel extension of the baseline AED network, we perform decision level fusion on the outputs of the three single-channel networks. In both scenarios, we observe that this multi-channel version of the baseline is only slightly better than the single-channel one, as the logFBE features are expected to be similar in adjacent microphones. For the lower reverberation case, we observe that both of the proposed methods outperform the baseline, with the jointly trained variant achieving the best performance (47.46%). Further improvements are observed with the fusion schemes (46.20% and 48.95%), which indicates that the AED networks trained on the mixture signal and on the separated signals learn complementary information. This corresponds to 7.7% absolute improvement compared to the baseline (A). On the contrary, in the higher reverberation case, the proposed system (D) fails to improve the baseline, due to inadequate performance of the separation network. This is in agreement with the results of recent works on the performance of separation networks under high reverberation conditions [148], as well as with our results in Figure. 6.5, which indicate that separation needs to exceed a certain quality to boost the overall performance. Nevertheless, the fusion schemes still achieve improvements over the baseline.

Table 6.2 provides the polyphony level classification accuracy of the proposed polyphony network for various choices of feature sets. We observe that while all feature sets achieve good performance, the best option is to combine the logFBE features with the GCC-based ones, leading to 99.27% classification accuracy. With such performance, it is guaranteed that our pipeline will be applied almost only on overlapped instances during testing.

Table 6.2: Performance of the polyphony classification network for different feature sets, in terms of accuracy.

Features	Notation	Accuracy (%)
logFBE	$[\mathbf{X}^{(\mathbf{s}_0)}]$	95.59
GCC	$[\mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_1)}; \mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_2)}]$	98.68
logFBE + GCC	$[\mathbf{X}^{(\mathbf{s}_0)}; \mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_1)}; \mathbf{GCC}^{(\mathbf{s}_0,\mathbf{s}_2)}]$	**99.27**

6.4 Conclusions

In this chapter, we examined the combination of sound source separation with overlapped acoustic event classification in a multi-channel setup with a large variety of event classes. In particular, we combined and jointly re-trained a state-of-the-art multi-channel separation network with a CNN-based AED architecture, aiming to decompose the hard overlapped task into a classification of a set of isolated instances. Our results showcase the potential of incorporating separation methods in AED systems, albeit high reverberation scenarios can be a limiting factor for the performance of the proposed pipeline.

Chapter 7

Conclusions and Future Work

7.1 Thesis contributions

In this Dissertation, we study the problem of Acoustic Event Detection in multi-microphone smart-space environments. The contributions of our work are summarized below:

- At first, special focus is placed on the subtask of Speech Activity Detection, where in the framework of multi-room smart-spaces, we develop an efficient two-step room-localized SAD system, appropriate for voice-enabled applications. The developed system is a spatio-temporal SAD module that is able to provide both the time boundaries of speech events ("when") and the coarse speaker position ("where") at the room level. Also, the information from multiple microphones is exploited in several ways, both in the first step of temporal speech segmentation, as well as in the second step, in the extraction of the proposed novel room-discriminant features. The developed system is computationally efficient and can be trained without the need of a large amount of training data. Further, it remains robust to reduced microphone setups, while also comparing favorably to deep-learning based alternatives.

- Then, in the general AED task, we focus on the challenging overlapping scenario and, at first, experiment with NMF-based approaches. For the single-channel case, we first propose a method that improves the detection step of the well-known CNMF approach. Then we study the case of classifier-based NMF methods where we examine ways to increase their robustness in highly overlapping conditions.

- In the case of overlapped AED with multiple microphones available, we propose a multi-channel NMF system based on the minimization of a novel objective function containing a multi-channel sparsity term. The experiments, conducted in the ATHENA database which contains real multi-channel recordings, confirm the superiority of the proposed method compared to traditional NMF baseline methods.

- Finally, we focus on deep-learning based approaches for solving the overlapping AED task in the challenging scenario with a large number of different event classes. Deep-learning

systems with traditional multi-label training exhibit degraded performance when the number of possible event combinations increases. To this end, we propose a deep-learning pipeline that combines a multi-channel separation network with an event classification network, aiming to approximately transform the overlapped task into the isolated one. Our results showcase that the proposed direction is promising for the overlapped AED task, especially given the evolution of separation networks in recent years. For our experiments, we employ several synthetic and real databases recorded in suitable multi-microphone smart-space environments.

7.2 Future work

In general, AED is a rapidly growing research area, and in recent years, several variants of this task have been the subject of multiple evaluation campaigns in the literature [11]. In our Dissertation, we mostly focused on the aspects of multi-channel processing and overlapping scenarios. Regarding future work, some of the possible research directions are summarized below:

- Regarding the task of room-localized SAD, we aim to further improve the performance of our system by incorporating deep-learning based modules into our two-stage architecture (e.g. replacement of GMM classifiers with CNNs in the first step). In addition, we intent to develop a system for the task of room-localized overlapped AED, expanding in this way our work on smart-space interfaces in the more general case of arbitrary acoustic events.
- Regarding the aspect of specially challenging overlapped scenarios (large number of different event categories, or high polyphony level), we aim to investigate additional neural network architectures for better combination of the separation and detection concepts, as well as to experiment with spatially distributed microphone arrays in order to better tackle the reverberation issues.

Βιβλιογραφία

[1] S. Krstulovic, "Audio event recognition in the smart home," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. Plumbley, and D. Ellis, Eds. Springer, 2018, pp. 335–371.

[2] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos, "Multi-microphone fusion for detection of speech and acoustic events in smart spaces," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 2375–2379.

[3] M. Grassi, A. Lombardi, G. Rescio, P. Malcovati, A. Leone, G. Diraco, C. Distante, P. Siciliano, M. Malfatti, L. Gonzo, V. Libal, J. Huang, and G. Potamianos, "A hardware-software framework for high-reliability people fall detection." in *Proc. IEEE Conference on Sensors (SENSORS)*, 2008, pp. 1328–1331.

[4] V. Libal, B. Ramabhadran, N. Mana, F. Pianesi, P. Chippendale, O. Lanz, and G. Potamianos, "Multimodal classification of activities of daily living inside smart homes." in *Proc. International Workshop on Ambient Assisted Living (IWAAL)*, 2009, pp. 687–694.

[5] J. Huang, X. Zhuang, V. Libal, and G. Potamianos, "Long-time span acoustic activity analysis from far-field sensors in smart homes." in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 4173–4176.

[6] D. Hollosi, J. Schröder, S. Goetze, and J.-E. Appell, "Voice activity detection driven acoustic event classification for monitoring in smart homes," in *Proc. 3rd International Symposium in Applied Sciences in Biomedical and Communication Technologies (ISABEL)*, 2010, pp. 1–5.

[7] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 4, no. 2, pp. 1–23, 2008.

[8] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: a systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, pp. 1–46, 2016.

[9] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6460–6464.

[10] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[11] *DCASE: Detection and classification of acoustic scenes and events*, http://dcase.community/.

[12] A. Diment, T. Heittola, and T. Virtanen, "Sound event detection for office live and office synthetic AASP challenge," *Proc. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (WASPAA)*, 2013.

[13] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proc. International Joint Conference on Neural networks (IJCNN)*, 2015, pp. 1–7.

[14] E. Benetos, G. Lafay, M. Lagrange, and M. D. Plumbley, "Detection of overlapping acoustic events using a temporally-constrained probabilistic model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6450–6454.

[15] J. Dennis, H. Tran, and E. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised Hough transform," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1085–1093, 2013.

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

[17] P. Giannoulis, G. Potamianos, P. Maragos, and A. Katsamanis, "Improved dictionary selection and detection schemes in sparse-CNMF-based overlapping acoustic event detection," in *Proc. Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE)*, 2016, pp. 25–29.

[18] C. Cotton and D. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 69–72.

[19] T. Komatsu, Y. Senda, and R. Kondo, "Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2259–2263.

[20] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries," in *Proc. Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE)*, 2016, pp. 45–49.

[21] M. Kim and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 293–297, 2015.

[22] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 5, no. 6, pp. 1291–1303, 2017.

[23] I. Choi, K. Kwon, S. H. Bae, and N. S. Kim, "DNN-based sound event detection with exemplar-based approach for noise reduction," in *Proc. Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE)*, 2016, pp. 16–19.

[24] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *Signal Processing Letters (SPL)*, vol. 24, no. 3, pp. 279–283, 2017.

[25] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution augmented transformer for semi-supervised sound event detection," in *Proc. Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE), Tech. Rep.*, 2020.

[26] J. Kurby, R. Grzeszick, A. Plinge, and G. A. Fink., "Bag-of-features acoustic event detection for sensor networks," in *Proc. Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE)*, 2016, pp. 55–59.

[27] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Proc. Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE)*, 2016, pp. 6–10.

[28] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 281–295, 2018.

[29] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[30] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 261–264.

[31] N. Seichepine, S. Essid, C. Févotte, and O. Cappé, "Soft nonnegative matrix co-factorization," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5940–5949, 2014.

[32] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 1942–1945.

[33] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 91, pp. 1–15, 2015.

[34] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, 1997.

[35] "ETSI EN 301 708 V7.1.1: Digital cellular telecommunications system (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels (GSM 06.94 version 7.1.1 Release 1998)," 1999, France.

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

[37] J. Ramírez, J. C. Segura, C. Benítez, Á. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3–4, pp. 271–287, 2004.

[38] B. Kotnik, Z. Kacic, and B. Horvat, "A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 2001, pp. 197–200.

[39] C. Shahnaz, W.-P. Zhu, and M. O. Ahmad, "A multifeature voiced/unvoiced decision algorithm for noisy speech," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, 2006, pp. 2525–2528.

[40] R. Tucker, "Voice activity detection using a periodicity measure," *IEEE Proceedings I – Communications, Speech and Vision*, vol. 139, no. 4, pp. 377–380, 1992.

[41] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in *Proc. Conference of the International Speech Communication Association (Interspeech)*, 2005, pp. 369–372.

[42] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[43] L. R. Rabiner and M. R. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 4, pp. 338–343, 1977.

[44] J. A. Haigh and J. S. Mason, "A voice activity detector based on cepstral analysis," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 1993, pp. 1103–1106.

[45] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7229–7233.

[46] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[47] Y. Ma and A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 21, pp. 1–18, 2013.

[48] A. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, and S. S. Narayanan, "Multi-band long-term signal variability features for robust voice activity detection," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2013, pp. 718–722.

[49] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[50] G. Evangelopoulos and P. Maragos, "Multiband modulation energy tracking for noisy speech detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2024–2038, 2006.

[51] J.-H. Bach, B. Kollmeier, and J. Anemüller, "Modulation-based detection of speech in real background noise: Generalization to novel background classes," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 41–44.

[52] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[53] X.-L. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, 2016.

[54] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7378–7382.

[55] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 483–487.

[56] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 2519–2523.

[57] I. McLoughlin and Y. Song, "Low frequency ultrasonic voice activity detection using convolutional neural networks," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2015, pp. 2400–2404.

[58] S.-Y. Chang, B. Li, G. Simko, T. N. Sainath, A. Tripathi, A. van den Oord, and O. Vinyals, "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 5549–5553.

[59] Y. Jung, Y. Kim, Y. Choi, and H. Kim, "Joint learning using denoising variational autoencoders for voice activity detection," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 1210–1214.

[60] R. Zazo, T. N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform CLDNNs for voice activity detection," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2016, pp. 3668–3672.

[61] L. Christoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmüller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 2629–2634.

[62] M. Matassoni, R. F. Astudillo, A. Katsamanis, and M. Ravanelli, "The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2014, pp. 1613–1617.

[63] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, "The Sweet-Home speech and multimodal corpus for home automation interaction," in *Proc. International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 4499–4506.

[64] N. Bertin, E. Camberlein, E. Vincent, R. Lebarbenchon, S. Peillon, É. Lamandé, S. Sivasankaran, F. Bimbot, I. Illina, A. Tom, S. Fleury, and É. Jamet, "A French corpus for distant-microphone speech processing in real homes," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2016, pp. 2781–2785.

[65] N. Bertin, E. Camberlein, R. Lebarbenchon, E. Vincent, S. Sivasankaran, I. Illina, and F. Bimbot, "VoiceHome-2, an extended corpus for multichannel speech processing in real homes," *Speech Communication*, vol. 106, pp. 68–78, 2019.

[66] A. Fleury, N. Noury, M. Vacher, H. Glasson, and J.-F. Seri, "Sound and speech detection and classification in a health smart home," in *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2008, pp. 4644–4647.

[67] M. A. Sehili, B. Lecouteux, M. Vacher, F. Portet, D. Istrate, B. Dorizzi, and J. Boudy, "Sound environment analysis in smart home," in *Ambient Intelligence: Third International Joint Conference, AmI 2012 Proceedings*, F. Paternò, B. de Ruyter, P. Markopoulos, C. Santoro, E. van Loenen, and K. Luyten, Eds. Berlin, Heidelberg: Springer, 2012, vol. LNCS-7683, pp. 208–223.

[68] A. Karpov, L. Akarun, H. Yalçın, A. Ronzhin, B. E. Demiröz, A. Çoban, and M. Železný, "Audio-visual signal processing in a multimodal assisted living environment," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2014, pp. 1023–1027.

[69] O. Brdiczka, M. Langet, J. Maisonnasse, and J. L. Crowley, "Detecting human behavior models from multimodal observation in a smart home," *IEEE Transactions on Automation Science and Engineering*, vol. 6, no. 4, pp. 588–597, 2009.

[70] J. A. Morales-Cordovilla, H. Pessentheiner, M. Hagmüller, and G. Kubin, "Room localization for distant speech recognition," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2014, pp. 2450–2453.

[71] Y. Tachioka, T. Narita, S. Watanabe, and J. Le Roux, "Ensemble integration of calibrated speaker localization and statistical speech detection in domestic environments," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 162–166.

[72] A. Abad, M. Matos, H. Meinedo, R. F. Astudillo, and I. Trancoso, "The L2F system for the EVALITA-2014 speech activity detection challenge in domestic environments," in *Proc. Italian Conference on Computational Linguistics (CLiC-it) and International Workshop EVALITA*, 2014, pp. 147–152.

[73] A. Brutti, M. Ravanelli, P. Svaizer, and M. Omologo, "A speech event detection and localization task for multiroom environments," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 157–161.

[74] G. Ferroni, R. Bonfigli, E. Principi, S. Squartini, and F. Piazza, "A deep neural network approach for voice activity detection in multi-room domestic scenarios," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.

[75] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Deep neural networks for multi-room voice activity detection: Advancements and comparative evaluation," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 3391–3398.

[76] P. Vecchiotti, F. Vesperini, E. Principi, S. Squartini, and F. Piazza, "Convolutional neural networks with 3-D kernels for voice activity detection in a multiroom environment," in *Multidisciplinary Approaches to Neural Computing*, A. Esposito, M. Faudez-Zanuy, F. C. Morabito, and E. Pasero, Eds. Cham: Springer, 2018, vol. SIST-69, pp. 161–170.

[77] P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Deep neural networks for joint voice activity detection and speaker localization," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1567–1571.

[78] A. Temko, D. Macho, C. Nadeu, and C. Segura, "UPC-TALP database of isolated acoustic events," Polytechnic University of Catalonia (UPC), Barcelona, Spain, Internal Report, 2005.

[79] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[80] A. Tsiami, I. Rodomagoulakis, P. Giannoulis, A. Katsamanis, G. Potamianos, and P. Maragos, "ATHENA: A Greek multi-sensory database for home automation control," in *Proc. of 15th Annual Conference of the International Speech Communication Association (Interspeech)*, 2014, pp. 1608–1612.

[81] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM Int. Conference Multimedia*, 2015, pp. 1015–1018.

[82] E. Principi, S. Squartini, F. Piazza, D. Fuselli, and M. Bonifazi, "A distributed system for recognizing home automation commands and distress calls in the Italian language," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2013, pp. 2049–2053.

[83] I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, and P. Maragos, "Room-localized spoken command recognition in multi-room, multi-microphone environments," *Computer Speech and Language*, vol. 46, pp. 419–443, 2017.

[84] E. Principi, S. Squartini, R. Bonfigli, G. Ferroni, and F. Piazza, "An integrated system for voice command recognition and emergency detection based on audio signals," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5668–5683, 2015.

[85] R. C. Rose and H. K. Kim, "A hybrid barge-in procedure for more reliable turn-taking in human-machine dialog systems," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003, pp. 198–203.

[86] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the Pan-European digital cellular mobile telephone service," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1989, pp. 369–372.

[87] D. Enqing, Z. Heming, and L. Yongli, "Low bit and variable rate speech coding using local cosine transform," in *Proc. IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering (TENCOM)*, vol. 1, 2002, pp. 423–426.

[88] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[89] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 2814–2818.

[90] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Localizing speakers in multiple rooms by using deep neural networks," *Computer Speech and Language*, vol. 49, pp. 83–106, 2018.

[91] M. Chan, E. Campo, D. Estève, and J.-Y. Fourniols, "Smart homes – Current features and future perspectives," *Maturitas*, vol. 64, no. 2, pp. 90–97, 2009.

[92] M. P. Poland, C. D. Nugent, H. Wang, and L. Chen, "Smart home research: Projects and issues," *International Journal of Ambient Computing and Intelligence*, vol. 1, no. 4, pp. 32–45, 2009.

[93] D. Ding, R. A. Cooper, P. F. Pasquina, and L. Fici-Psquina, "Sensor technology for smart homes," *Maturitas*, vol. 69, no. 2, pp. 131–136, 2011.

[94] M. R. Alam, M. B. I. Reaz, and M. Mohd Ali, "A review of smart homes – Past, present, and future," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 6, pp. 1190–1203, 2012.

[95] M. Amiribesheli, A. Benmansour, and A. Bouchachia, "A review of smart homes in healthcare," *Journal of Ambient Intelligence and Humanized Computing*, vol. 6, no. 4, pp. 495–517, 2015.

[96] M. Matassoni, M. Omologo, R. Manione, T. Sowa, R. Balchandran, M. E. Epstein, and L. Seredi, "The DICIT project: an example of distant-talking based spoken dialogue interactive system," in *Proc. International Conference on Intelligent Information Systems (IIS)*, 2008, pp. 527–533.

[97] A. Badii and J. Boudy, "CompanionAble - integrated cognitive assistive and domotic companion robotic systems for ability and security," in *Proc. Congrès Société Fançaise des Technologies pour l'Autonomie et de Gérontechnologie (SFTAG)*, 2009, pp. 18–20.

[98] G. L. Filho and T. J. Moir, "From science fiction to science fact: a smart-house interface using speech technology and a photo-realistic avatar," *International Journal of Computer Applications in Technology*, vol. 39, no. 1/2/3, pp. 32–39, 2010.

[99] M. Vacher, D. Istrate, F. Portet, T. Joubert, T. Chevalier, S. Smidtas, B. Meillon, B. Lecouteux, M. Sehili, P. Chahuara, and S. Méniard, "The SWEET-HOME project: Audio technology in smart homes to improve well-being and reliance," in *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2011, pp. 5291–5294.

[100] "DIRHA: Distant-speech Interaction for Robust Home Applications," [Online] http://dirha.fbk.eu, (Accessed: 2019-04-22).

[101] J. F. Gemmeke, B. Ons, N. Tessema, H. Van hamme, J. van de Loo, G. De Pauw, W. Daelemans, J. Huyghe, J. Derboven, L. Vuegen, B. Van Den Broeck, P. Karsmakers, and B. Vanrumste, "Self-taught assistive vocal interfaces: An overview of the ALADIN project," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2013, pp. 2039–2043.

[102] M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, B. Lecouteux, and P. Chahuara, "Evaluation of a context-aware voice interface for ambient assisted living: qualitative user study vs. quantitative system evaluation," *ACM Transactions on Accessible Computing*, vol. 7, no. 2:5, pp. 1–36, 2015.

[103] M. Malavasi, E. Turri, J. J. Atria, H. Christensen, R. Marxer, L. Desideri, A. Coy, F. Tamburini, and P. Green, "An innovative speech-based user interface for smarthomes and IoT solutions to help people with speech and motor disabilities," *Studies in Health Technology and Informatics*, vol. 242, pp. 306–313, 2017.

[104] V. Këpuska and G. Bohouta, "Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home)," in *Proc. IEEE Annual Computing and Communication Workshop and Conference (CCWC)*, 2018, pp. 99–103.

[105] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, 1993.

[106] P. Giannoulis, A. Tsiami, I. Rodomagoulakis, A. Katsamanis, G. Potamianos, and P. Maragos, "The Athena-RC system for speech activity detection and speaker localization in the DIRHA smart home," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 167–171.

[107] P. Giannoulis, A. Brutti, M. Matassoni, A. Abad, A. Katsamanis, M. Matos, G. Potamianos, and P. Maragos, "Multi-room speech activity detection using a distributed microphone network in domestic environments," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1271–1275.

[108] M. Wolf and C. Nadeu, "On the potential of channel selection for recognition of reverberated speech with multiple microphones," in *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, 2010, pp. 574–577.

[109] P. Maragos and A. C. Bovik, "Image demodulation using multidimensional energy separation," *Journal of the Optical Society of America A*, vol. 12, no. 9, pp. 1867–1876, 1995.

[110] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin, Heidelberg: Springer, 2001, pp. 157–180.

[111] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Burlington, MA: Academic Press, 2009.

[112] A. Brutti, M. Ravanelli, and M. Omologo, "SASLODOM: Speech Activity detection and Speaker LOcalization in DOMestic environments," in *Proc. Italian Conference on Computational Linguistics (CLiC-it) and International Workshop EVALITA*, 2014, pp. 139–146.

[113] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[114] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition." in *Proc. International Conference on Spoken Language Processing (Interspeech)*, 2000, pp. 746–749.

[115] T. Butko, A. Temko, C. Nadeu, and C. C. Ferrer, "Fusion of audio and video modalities for detection of acoustic events," in *Proc. International Conference on Spoken Language Processing (Interspeech)*, 2008, pp. 123–126.

[116] R. Stiefelhagen, R. Bowers, and J. Fiscus, *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007*. Springer, 2008, vol. LNCS-4625.

[117] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[118] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[119] S. Vanambathina, "Speech enhancement using an iterative posterior NMF." in *New Frontiers in Brain-Computer Interfaces. IntechOpen*, 2019.

[120] N. Gillis and R. Luce, "Robust near-separable nonnegative matrix factorization using linear optimization," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1249–1280, 2014.

[121] N. Guan, D. Tao, Z. Luo, and B. Yuan, "NeNMF: An optimal gradient method for nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2882–2898, 2012.

[122] B. Recht, C. Re, J. Tropp, and V. Bittorf, "Factoring nonnegative matrices with linear programs," in *Advances in Neural Information Processing Systems*, 2012, pp. 1214–1222.

[123] J. L. Roux, F. Weninger, and J. Hershey, "Sparse NMF–half-baked or well done?" Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep. no. TR2015-023, 2015.

[124] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proc. International Conference on Independent Component Analysis and Signal Separation*, 2004, pp. 494–499.

[125] W. Wang, A. Cichocki, and J. Chambers, "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared Euclidean distance," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2858–2864, 2009.

[126] P. O'Grady and B. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in *Proc. 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, 2006, pp. 427–432.

[127] P. Giannoulis, G. Potamianos, and P. Maragos, "Multi-channel non-negative matrix factorization for overlapped acoustic event detection," in *Proc. 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 857–861.

[128] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 151–155.

[129] W. Wang, "Convolutive non-negative sparse coding," in *Proc. IEEE International Joint Conference on Neural Networks*, 2008, pp. 3681–3684.

[130] P. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[131] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 141–145.

[132] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. International Conference on Spoken Language Processing (Interspeech)*, 2016, pp. 545–549.

[133] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," *arXiv preprint arXiv:1910.06379*, 2019.

[134] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *arXiv preprint arXiv:2002.08933*, 2020.

[135] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," *arXiv preprint arXiv:2003.01531*, 2020.

[136] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. Ellis, "Improving universal sound separation using sound classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 96–100.

[137] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. L. Roux, and J. R. Hershey, "Universal sound separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 175–179.

[138] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8677–8681.

[139] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, "Improving sound event detection in domestic environments using sound separation," *arXiv preprint arXiv:2007.03932*, 2020.

[140] W. Xue, T. Ying, Z. Chao, and D. Guohong, "Multi-beam and multi-task learning for joint sound event detection and localization," in *Proc. Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE), Tech. Rep.*, 2019.

[141] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," *arXiv preprint arXiv:1910.14104*, 2020.

[142] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S. C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019, pp. 260–267.

[143] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of CRNN models," in *Proc. Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE), Tech. Rep.*, 2019.

[144] F.-R. Stöter, S. Chakrabarty, B. Edler, and E. A. P. Habets, "CountNet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 268–282, 2019.

[145] Y. Cao, T. Iqbal, Q. Kong, M. B. Galindo, W. Wang, and M. D. Plumbley, "Two-stage sound event localization and detection using intensity vector and generalized cross-correlation," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Tech. Rep.*, 2019.

[146] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted Boltzmann machine for environmental sound classification," in *Proc. International Conference on Spoken Language Processing (Interspeech)*, 2017, pp. 3107–3111.

[147] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[148] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and reverberant single-channel speech separation." in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 696–700.

Παράρτημα

Λίστα δημοσιεύσεων

Δημοσιεύσεις σε διεθνή περιοδικά με κριτές

- **P. Giannoulis**, G. Potamianos, and P. Maragos. "Room-localized speech activity detection in multi-microphone smart homes". *EURASIP Journal on Audio, Speech, and Music Processing*, 2019, 1:15.

- I. Rodomagoulakis, A. Katsamanis, G. Potamianos, **P. Giannoulis**, A. Tsiami, and P. Maragos. "Room-localized spoken command recognition in multi-room, multi-microphone environments". *Computer Speech & Language*, 2017, 46, pp. 419–443.

Δημοσιεύσεις σε διεθνή συνέδρια με κριτές

- **P. Giannoulis**, G. Potamianos, and P. Maragos. "Overlapped sound event classification via multi-channel sound separation neural network". *In Proc. 29th European Signal Processing Conference (EUSIPCO)*, 2021.

- **P. Giannoulis**, G. Potamianos, and P. Maragos. "Multi-channel non-negative matrix factorization for overlapped acoustic event detection". *In Proc. 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 857-861.

- **P. Giannoulis**, G. Potamianos, and P. Maragos, "On the joint use of NMF and classification for overlapping acoustic event detection". *In Proc. International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM) held in conjunction with the 25th European Signal Processing Conference (EUSIPCO)*, ser. MDPI Proceedings, vol. 2, no. 2, 2018.

- A. Dometios, A. Tsiami, A. Arvanitakis, **P. Giannoulis**, X. S. Papageorgiou, C. S. Tzafestas, and P. Maragos. "Integrated speech-based perception system for user adaptive robot motion planning in assistive bath scenarios". *In Proc. Workshop on Multimodal processing, modeling and learning for human-computer/robot interaction applications (MultiLearn) held in conjuction with the 25th European Signal Processing Conference*, 2017.

- **P. Giannoulis**, G. Potamianos, P. Maragos, and A. Katsamanis, "Improved dictionary selection and detection schemes in sparse-CNMF-based overlapping acoustic event detection". *In Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pp. 25–29, 2016.

- **P. Giannoulis**, A. Brutti, M. Matassoni, A. Abad, A.Katsamanis, M. Matos, G. Potamianos, and P. Maragos, "Multi-room speech activity detection using a distributed microphone network in domestic environments". *In Proc. 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1271–1275.
- **P. Giannoulis**, G. Potamianos, A. Katsamanis, and P. Maragos. "Multi-microphone fusion for detection of speech and acoustic events in smart spaces". *In Proc. 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 2375–2379.
- **P. Giannoulis**, A. Tsiami, I. Rodomagoulakis, A. Katsamanis, G. Potamianos, and P. Maragos, "The ATHENA-RC system for speech activity detection and speaker localization in the DIRHA smart home". *In Proc. 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014, pp. 167–171.
- A. Tsiami, I. Rodomagoulakis, **P. Giannoulis**, A. Katsamanis, G. Potamianos, and P. Maragos. "ATHENA: a Greek multi-sensory database for home automation control". *In Proc. 15th Annual Conference of the International Speech Communication Association (Interspeech)*, 2014, pp. 1608–1612.
- I. Rodomagoulakis, **P. Giannoulis**, Z.-I. Skordilis, P. Maragos, and G. Potamianos, "Experiments on far-field multichannel speech processing in smart homes". *In Proc. 18th International Conference on Digital Signal Processing (DSP)*, 2013, pp. 1–6.