Robust sound event detection

Michel Olvera

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Robust sound event detection

THÈSE

présentée et soutenue publiquement le 15 décembre 2022

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Michel Olvera

Composition du jury

Présidente :	Anne Boyer	Professeure, Université de Lorraine
Rapporteurs :	Mathieu Lagrange	CR CNRS, LS2N
	Juan Pablo Bello	Professeur, New York University
Examinateurs :	Daniel P. W. Ellis	Chargé de recherche, Google
	Gilles Gasso	Professeur, INSA Rouen Normandie
Directeur de thèse :	Emmanuel Vincent	DR Inria, Inria Nancy - Grand Est

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Acknowledgements

First and foremost, to my wife Mirian. Little did we know that this journey would profoundly impact every fiber of our beings. Throughout it all, you have provided me with unwavering moral support and patience beyond measure. I consider myself so fortunate to have you by my side, and I am endlessly grateful for your presence in my life.

Gilles and Emmanuel, I am beyond grateful for the incredible influence you have had on my life and career. Your kind support, understanding, and warmth have been the pillars that sustained me throughout this work. I cannot express enough my appreciation for the countless hours of wise discussions, feedback, and guidance that you generously offered.

Emmanuel, your exceptional intellect has always left me in awe. Your comments were always loaded with knowledge and precision, and your vision and clarity of thought were an inspiration to me. I feel truly honored to have had the opportunity to learn from you and work under your guidance.

Gilles, your compassion and empathy have been a constant source of comfort and encouragement. You have been there for me every step of the way, providing me with the trust, freedom, and support I needed to complete this challenging intellectual endeavor.
I will forever cherish the teachings and lessons I have learned from both of you, and I hope to apply them throughout my research career. This childhood dream of mine has been made possible thanks to your contribution, and for that, I am eternally grateful.

Thanks to my jury members: Anne, Mathieu, Dan and Juan Pablo, for the time and effort you have dedicated to reviewing my thesis, as well as for the insightful comments and suggestions you provided during the defense. Your critical evaluation has challenged me to think deeply about my work and to refine my arguments. Your expertise, wisdom, and commitment to excellence is a source of inspiration and motivation for me. As someone who has long admired your contributions in the field, I feel privileged to have had the opportunity to have you as my jury members.

To the Multispeech research group, I owe a debt of gratitude. Through their ever-revolving door of team members, I found true long-lasting friendships. Bittersweet were the farewells to those departing, but sweet were the welcomes to those arriving. Thanks for the memories we shared, and the bonds we formed, they will always hold a special place in my heart.

My heartfelt thanks go to to Lou Lee, Tulika Bose, Laura Zanella, Hélène Zganic, Nicolas Furnon, Sewade Ogun, Joris Cosentino, Louis Abel, Shakeel Sheikh, Manuel Pariente, Nicolas Turpault, Anastasiia Tsukanova, Pierre Champion, Sandipana Dowerah, Seyed Hosseini, Prerak Srivastava, Vinicius Ribeiro, Can Cui, Marina Krémé, Marie-Anne Lacroix, Romain Serizel, Paul Magron, Louis Delebecque, Hugo Bergerat, Ashwin Geet D'Sa, Ali Golmakani, Soklay Heng, Mathieu Hu, Hubert Nourtel, Francesca Ronchini, Sunit Sivasankaran, Brij Srivastava, Georgios Zervakis, Ajinkya Kulkarni, Raphaël Duroselle, Ioannis Douros, Guillaume Carbajal, Mostafa Sadeghi, Antoine Deleforge, Slim Ouini and Denis Jouvet.

I would also like to express my gratitude to all people in the Loria/Inria laboratory that foster a culture of inclusivity and for recognizing the value of diversity in the workplace. The rich experience of working alongside colleagues from diverse backgrounds has been nothing short of life-changing. Through exposure to different perspectives, customs, and ways of thinking I can see the world in new and exciting ways. I am particularly grateful to my coworkers who have generously shared their culture with me through language, music, festivals, arts and crafts

and food workshops. This acknowledgement is a testament to the profound impact that multiculturalism in the workplace can have on an individual's personal and professional development.

Lastly, I want to take a moment to acknowledge myself for the strength and perseverance I demonstrated throughout my PhD journey, particularly towards its completion, where I suffered greatly from a depressive episode. I have shown remarkable courage, bravery and persistence in forging ahead, even in these stressful and challenging times. Despite the ups and downs and having lost all motivation to continue, I kept a positive attitude to complete my thesis. I recognize that completing a PhD is an extremely difficult endeavor and I am proud of myself for showing resilience and determination throughout the process. I am grateful for my own commitment and for never giving up on the dream. I have allowed myself to lose to win and start to heal.

Figure 1: Ali (left) and me (right) on our way to winning the ping pong tournament at the Loria. This trophy is dedicated to Aswhin, Hugo, Shakeel, Prerak and Sewade, for the countless matches we played together.

Contents

Glossary

ACC : accuracy
ADA : adversarial domain adaptation
AL : active learning
ASC : acoustic scene classification
ASR : automatic speech recognition
CADA : conditional adversarial domain adaptation
CRNN : convolutional neural network
DCASE : detection and classification of acoustic scenes and events
DFT : discrete Fourier transform
DNN : deep neural network
FB : foreground-background
FN : false negative
FP : false positive
GMM : Gaussian mixture model
HMM : hidden Markov model
IBM : ideal binary mask
IRM : ideal ratio mask
JDOT : joint distribution optimal transport
MAL : medoid-based active learning
MLP : multi layer perceptron
MM : moment matching
MN : mean normalization
MNVN : mean normalization and variance matching
MVN : mean and variance normalization
NMF : non-negative matrix factorization
OT : optimal transport
PCA : principal component analysis
PSDS : polyphonic sound detection score
RNN : recurrent neural network
SAR : signal-to-artifacts ratio
SDR : signal-to-distortion ratio
SED : sound event detection
SIR : signal-to-interference ratio
SNR : signal-to-noise ratio
STFT : short-time Fourier transform
TF : time-frequency
TN : true negative
TP : true positive

1

Introduction

In this chapter, we describe the research context in which this thesis takes place, particularly the adaptation of audio analysis systems under mismatched test conditions. We define the motivation and the objectives and scope of our research work. Then, we present our contributions and finally, we detail the organization of the document and list the publications resulting from this work.

1.1 Motivation

The sense of hearing in humans is remarkable for its ability to connect us to the world through sound. From the subtlest whisper to the loudest noise, we gather, process and make sense of the sounds that we continually encounter in our daily lives. They are essential for our well-being and survival. They provide us with joy through the sound of music, and we express displeasure perhaps through a rasping scratch. Whether pleasant or annoying, sounds convey meaningful information. They pave the way for communication through speech, alert us to danger through sirens and gunshots, and call us to action through baby cries and alarms.

More than ever before, our constant search for computational methods to understand the soundscapes of our surroundings has provided rewards across many interdisciplinary domains including robotics, urban monitoring, bio-acoustics and surveillance. From industry to general interest applications, computational analysis of sound scenes and events allows us to automatically analyze and interpret the continuous flow of everyday sounds.

Nowadays, consumer electronic products involving sound sensing and automatic sound analysis are being commercialized on a massive scale, despite the fact that they tend to run into serious trouble in real environments. In fact, one of the main degradations encountered when moving from laboratory conditions to the real world is due to the high diversity of soundscapes. Complex mixtures of sounds are not composed of isolated audio events but of multiple events interacting with one another simultaneously. Moreover, differences between training and test conditions also typically arise due to extrinsic factors such as distant microphone capture, recording locations, different acquisition hardware and settings, as well as intrinsic factors of sound events, such as their frequency of occurrence, duration and variability. Interest in these problems has grown substantially in recent years. It has been particularly fueled in academia by the well known DCASE Challenge series (DCASE, 2022), and special sessions in conferences such as ICASSP or EUSIPCO and the WASPAA workshop, often with keen participation and sponsorship of industry.

Increasing the robustness of automatic sound analysis systems to the adversities of real-world environments will enable applications to go beyond identifying acoustic scenes or isolated sounds

with high accuracy, allowing them to better understand the surroundings. Making sense of the complex relationships between sounds and the place and time in which they occur could provide a better description of acoustic environments, which could potentially lead to the emergence of new applications and give a twist to existing ones.

1.2 Context

Taking the analysis of environmental sounds to the next level, research has benefited from the advances made in more established areas such as those focused on the study of speech and music signals. An example task is the separation of ambient sounds, which borrowed state-of-the-art supervised speech separation methods to investigate the potential segregation of sound events of a wide variety of classes and of much less structured attributes. A study carried out by Kavalerov et al. (2019) named this task as universal sound separation and showed promising results towards the separation of arbitrary sounds. More recently, the limited availability of training data with ground-truth source signals motivated unsupervised approaches (Wisdom et al., 2020). Audio source separation applied to sound events beyond speech or music may become a potential tool for audio analysis tasks involving sound recognition. In principle, it may improve the robustness of classification and detection systems to target sounds of interest occurring simultaneously or in the presence of noise or other non-target interferences.

However, it is equally important to provide sound analysis systems with robustness to other real-world issues. In fact, when developing audio analysis systems there is an underlying assumption that the training samples are drawn i.i.d samples drawn from the desired test distribution. In practice, this assumption is violated: audio analysis systems are prone to domain shifts, which result from changing conditions from development to deployment. Correcting this problem is known as *domain adaptation* (Ben-David et al., 2006). The goal is to tailor a model to perform well on test samples from a *target* domain whose distribution is different from the *source* domain from which the training data was originally drawn. As such, domain adaptation countermeasures require access to an annotated dataset drawn from a source distribution, and a dataset drawn from the target distribution. Depending on the availability of annotations for the target domain data, the adaptation process can be carried out in various ways. If no labels are provided, then *unsupervised domain adaptation* is carried out. When only a few annotations are available for a subset of the target domain data, then the adaptation process is known as *semi-supervised domain adaptation*. In both cases, the goal is to reduce the existing gap between the training and the target distributions.

Kull and Flach (2014) categorize the shift in the data distributions into *label shift*, *covariate shift* and *concept shift*. Real-world scenarios possibly combine any of these shifts. Most works on domain adaptation for sound analysis tasks tackle the covariate shift. For instance, Task 1 of the DCASE challenge motivates the development of methods that generalize to mismatched recording devices in the acoustic scene classification (ASC) task. Many of them rely on data augmentation techniques, and only a handful explicitly reduce the distribution shift with domain adaptation methods. In matched recording conditions, distortions due to background noise, reverberation, distant capture of the target sources, and the presence of non-target sounds can also degrade the quality of the test acoustic signal. These sources of mismatch have been of particular interest in sound event detection (SED). In this case, simulated soundscapes mixing audio events with different background noises and non-target events (Ronchini et al., 2021), or convolving them with room impulse responses help achieve robustness to covariate shift (Turpault and Serizel, 2020). Despite the effort of simulating real-world conditions, a performance gap between simulated and

real test data prevails.

1.3 General problem and objectives

This thesis focuses on sound analysis systems performing audio source separation, audio classification and sound event detection and their applicability in real scenarios. The main objectives of this thesis are twofold. First, we aim to use deep learning to come up with a masking-based separation system that isolates ambient sounds categorized into foreground and background according to their spectro-temporal structure, as this task corresponds to a practical scenario where short sound events with fast transients occur on top of long duration stationary sounds. We respond to the following questions: Can a deep learning-based separation system discriminate between the rapidly varying spectro-temporal features of foreground events against the slowly varying features of background sounds? Can such a distinction between sounds be robust to unseen events? Second, we aim to improve the robustness of audio analysis systems under mismatched training and testing conditions. To achieve this goal we explore two different tasks affected by different sources of mismatch: acoustic scene classification with mismatched recording devices and training of sound event detection systems with synthetic and real data. For the first task we answer the following question: What is the impact of moment normalization and moment matching strategies in adversarial domain adaptation for acoustic scene classification? For the second task, despite the semi-supervised learning capability adopted for training sound event detection systems with synthetic labeled data and unlabeled or partially labeled real data, can we further reduce the data mismatch between synthetic and real data by aligning their empirical distributions? Furthermore, what effects does active learning have on system performance in the target domain?

Other secondary goals within the above tasks include the exploration of suitable acoustic front-ends for ambient sounds, as well as improving detection scores based on the use of auxiliary information regarding the frequency content of the analyzed sound events.

1.4 Contributions

1.4.1 Foreground-background ambient sound scene separation

Recent works exploring the separation of sounds other than speech and music have shown that deep neural networks can learn to separate mixtures of arbitrary sounds (Kavalerov et al., 2019). While this represents a big leap forward towards the separation of broad categories of sounds, real-life ambient sound analysis systems operate on recordings involving multiple short sound events in the presence of continuous background sounds. Motivated by this observation, we thus propose the foreground-background ambient sound scene separation task. In this new task of practical interest, we investigate whether a deep neural network can differentiate short sound events with fast transient and rapidly evolving spectral characteristics (i.e., foreground sounds) against long duration sounds with slowly varying spectro-temporal structure (i.e., background sounds). Moreover, we investigate the ability of the separation model to generalize to unseen sound classes whose frequency contents exhibit such categorization (e.g., under label shift). Our proposed source separation model relies on a deep neural network that estimates the soft time-frequency mask associated with foreground sounds. Inspired by speaker extraction networks (Žmolíková et al., 2019), we extend their application to situations when the sources in a mixture belong to distinct classes (here, ambient sounds). Thus, we investigate the use of an optional

auxiliary network to inform the main network about the background statistics. We also explore per-channel energy normalization (PCEN) of the mixture time-frequency representation (Wang et al., 2017) as a way to improve the separation performance. We conduct experiments on simulated foreground-background mixtures of isolated sounds from the DESED and Audioset datasets. We find that the use of adaptation segments to inform the network is detrimental to the separation process while the use of PCEN is beneficial. The similar improvements achieved by the proposed model on mixtures of seen and unseen sound classes show its generalization capabilities in terms of the signal-to-distortion ratio (SDR).

1.4.2 Unsupervised domain adaptation for acoustic scene classification

Acoustic scene classification with mismatched recording devices is a task of practical interest in the DCASE Challenge series, in which a large number of recordings from a source device and a limited amount of recordings from target devices are available. The primary goal is to improve generalization on the underrepresented devices. Such an imbalance problem has been addressed with supervised machine learning algorithms often combined with data augmentation, regularization and fine tuning approaches. A few works have investigated a more practical scenario where recordings of the source and target devices are available, but only the source recordings are labeled. Among these methods, the unsupervised adversarial domain adaptation (UDA) method proposed by Drossos et al. (2019) proves effective to improve generalization on the target devices. However, it requires a large number of recordings from the target devices to carry out the adaptation process. To overcome this limitation, Mezza et al. (2021) proposed band-wise statistics matching (BWSM) as a simple, linear UDA method that does not require any learning phase, and that is equally effective to achieve generalization. The combination of these methods has not yet been explored. We thus propose the integration of non-linear learning-based adaptation with linear feature-based adaptation to tackle the domain shift arising in ASC due to mismatched recording devices. We identified that previous works implicitly leverage recordings captured simultaneously by the source and target devices to compensate for the effects of their frequency responses. To address UDA in a fully blind setting with respect to the target domain, we do not assume the availability of parallel recordings since in a realistic scenario a pretrained ASC system must be deployed on devices with unknown microphone responses. We provide a thorough analysis of first- and second-order moment normalization and matching strategies (among which BWSM is a particular case) and their integration into the adversarial-based domain adaptation framework. Such a framework however only guarantees the alignment of the marginal distributions of the source and the target domain data. Since the discriminator and classifier are of different nature, they fail to fully capture the relationship between feature representations and class posteriors, which are crucial to domain adaptation (Long et al., 2018). To overcome this issue and to guarantee the conditional alignment of features/labels, we extend the adversarial adaptation formulation to a conditional formulation to fully match the source and target data distributions. The proposed integration is observed to reduce the impact of mismatched recording devices when moment normalization and/or matching strategies are applied at all stages of the classification pipeline, i.e., during pretraining, adaptation and inference. We find that in such a setting, the performance achieved in the target domain is comparable to that of matched conditions, i.e., the source domain, thus effectively correcting the effects of mismatched recording devices. The results achieved by such strategies show their individual scope and limitations and serve as design choices to counteract the performance degradation of ASC systems under domain mismatch.

1.4.3 Robust sound event detection

We investigate two research axes that aim to increase the robustness of sound event detection systems. In the first line of research we explore a trainable acoustic front-end based on per-channel energy normalization to boost the detection scores. We also propose to categorize domestic sounds into foreground and background sounds and use this broad categorization as auxiliary information to improve sound event detection. We design a multi-task learning approach that jointly learns a foreground-background classifier and a sound event detection branch. We find that such a training scheme helps improve sound event detection. Moreover, we investigate a conditioning approach that combines the aforementioned branches into an improved sound event detection branch that further improves robust sound event detection. In the second research axis we investigate domain adaptation to account for the existing mismatch between synthetic and real data. While the common semi-supervised learning capability adopted for training sound event detection systems with synthetic labeled data and unlabeled or partially labeled real data aims to learn invariant representations for both domains (Turpault et al., 2019), there is still a gap in performance when testing such systems in real environments. To further reduce this data mismatch, we explore semi-supervised domain adaptation and active learning. We conduct experiments to align the empirical distributions of the feature representations of active and inactive frames of synthetic and real data via optimal transport (Courty et al., 2014). We find that the proposed methods lead to enhanced performance in terms of the event-based macro f1-score on the DESED validation and public evaluation sets. Furthermore, we propose an active learning strategy that prompts the user with sound excerpts that a SED system fails to identify to obtain reference labels. We design an experimental protocol that further reduces the synthetic/real data mismatch with an iterative procedure.

1.5 Organization of the thesis

Chapter 2 provides an overview of the state of the art in the fields related to this thesis. As such, it introduces basic concepts related to the computational analysis of sound scenes and events, audio source separation and domain adaptation methods to improve robustness to test conditions.

Chapter 3 presents the foreground-background ambient sound scene separation task. We describe our proposed deep learning-based separation framework and investigate its ability to handle a wide variety of known or unknown sound events comprising either rapidly- or slowly-varying spectro temporal characteristics.

Chapter 4 deals with unsupervised domain adaptation for acoustic scene classification. We conduct a comprehensive analysis of moment normalization, moment matching and adversarial domain adaptation formulations to counteract the performance degradation experienced by ASC systems with mismatched recording devices.

Chapter 5 deals with robustness of sound detection systems in two axes. We first investigate how to improve classification scores with a dedicated acoustic front-end, as well as with additional information regarding the spectro-temporal categorization of domestic sounds. Secondly, we explore semi-supervised domain adaptation and active learning approaches for sound event detection with the use of optimal transport to reduce the discrepancy between simulated and real data.

Chapter 6 concludes this thesis by summarizing its contributions and by outlining some future research perspectives motivated by our work.

1.6 Publication list

Most of our contributions have been published in the following conference and workshop papers. We specify in each chapter the publication related to the content presented.

- **M. Olvera**, E. Vincent, R. Serizel and G. Gasso. *Foreground-background ambient sound scene separation.* In Proc. EUSIPCO 2021, pp. 281-285.
- S. Cornell, **M. Olvera**, M. Pariente, G. Pepe, E. Principi, L. Gabrielli and S. Squartini. *Domain-adversarial training and trainable parallel front-end for the DCASE 2020 task 4 sound event detection challenge.* In Proc. DCASE, 2020, pp. 26–30.
- **M. Olvera**, E. Vincent and G. Gasso. *Improving sound event detection with auxiliary foreground-background classification and domain adaptation.* In Proc. DCASE, 2021, pp. 231-235.
- **M. Olvera**, E. Vincent and G. Gasso. *On The impact of normalization strategies in unsupervised adversarial domain adaptation for acoustic scene classification.* In Proc. ICASSP 2022, pp. 631-635.

Other contributions related to this thesis, but out of the scope of our objectives, lead to the following secondary publications.

- M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, **M. Olvera**, F.R. Stöter, M. Hu, J.M. Martın-Doñas, D. Ditter, A. Frank, A. Deleforge and E. Vincent. *Asteroid: the PyTorch-based audio source separation toolkit for researchers.* Proc. Interspeech 2020, pp. 2637–2641.
- S. Cornell, **M. Olvera**, M. Pariente, G. Pepe, E. Principi, L. Gabrielli and S. Squartini. *Task-aware separation for the DCASE 2020 task 4 sound event detection and separation challenge.* In Proc. DCASE, 2020, pp. 31-35.

2

State of the art

In this chapter we will present the fundamental concepts and state-of-the-art methods surrounding ambient sound analysis systems that make possible the classification and detection of sounds composing an everyday acoustic scene such as the one illustrated in Figure 2.1. We will describe the major challenges that sound audio analysis systems face under conditions that prevent their generalization in real-world scenarios, particularly under domain mismatch settings and discuss current approaches that overcome these issues.

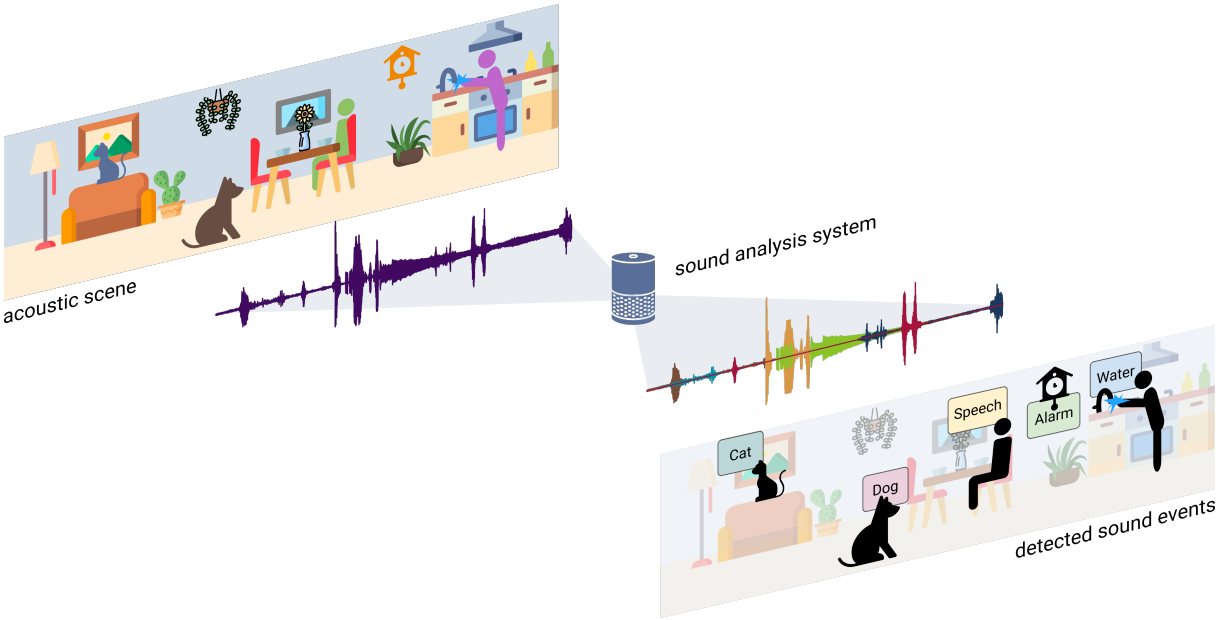


Figure 2.1: Analysis of sound scenes and events.

2.1 Sound analysis

2.1.1 Overview of sound classification tasks

The analysis of everyday acoustic scenes and events comprises various methodological tasks that imply classification or detection of sounds. Consider the audio recording of an acoustic scene shown in Figure 2.2. This soundscape serves as input to three major audio analysis tasks:

acoustic scene classification (ASC), *audio classification or tagging* and *sound event detection* (SED). ASC aims to abstract all interacting sound events composing the soundscape and assign it a meaningful class label. For instance "home", "living room", "indoor" are all suitable labels that categorize the acoustic scene of Figure 2.2 since all of them describe well the global contents of the soundscape. Sound event classification goes beyond the categorization of the acoustic context of the soundscape and aims to assign a label to all active sound events in fixed length audio samples that commonly span a few seconds. Since more than one class label or *tag* can be assigned to an audio clip, this task is also referred to as *tagging*. This tasks helps identify the constituent sound events of an acoustic scene regardless of their repetition or when they occur. Temporal information about sound events is provided by the sound event detection task, which in addition to providing class labels, provides start and end timestamps for each recognized sound event.

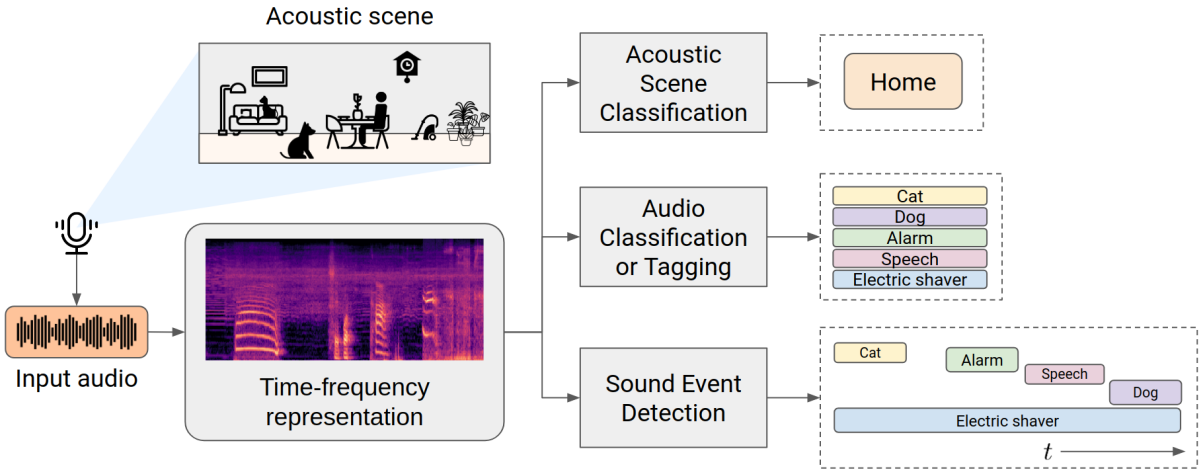


Figure 2.2: Audio analysis tasks.

In this section we'll provide a high level description of the processes involved in the analysis of everyday acoustic scenes. These processes are illustrated in Figure 2.3 and act upon a digital audio signal acquired with a recording device.

2.1.2 Sound analysis tasks

Acoustic scene classification

Acoustic Scene Classification (ASC) refers to the task of identifying the environment in which an audio recording was captured (Barchiesi et al., 2015). An ASC system can output a class label that describes the recording environment e.g., *metro station*, *shopping mall*, or the activity such as *children playing*, *machines working*. The contents of the acoustic scene can be categorized based on the hierarchical relationship between coarse- and fine-grained labels of a taxonomy. Coarse-grained labels can include for instance the high-level classes *indoor*, *outdoor* or *vehicle*, but commonly fine-grained labels, e.g., *grocery store*, *park*, or *bus* define the pool of labels. Classifying accurately two or more classes with similar acoustic properties, e.g., *shopping mall* vs. *airport* or *bus* vs. *tram*, represents a great challenge for ASC, which often results in high misclassification rates.

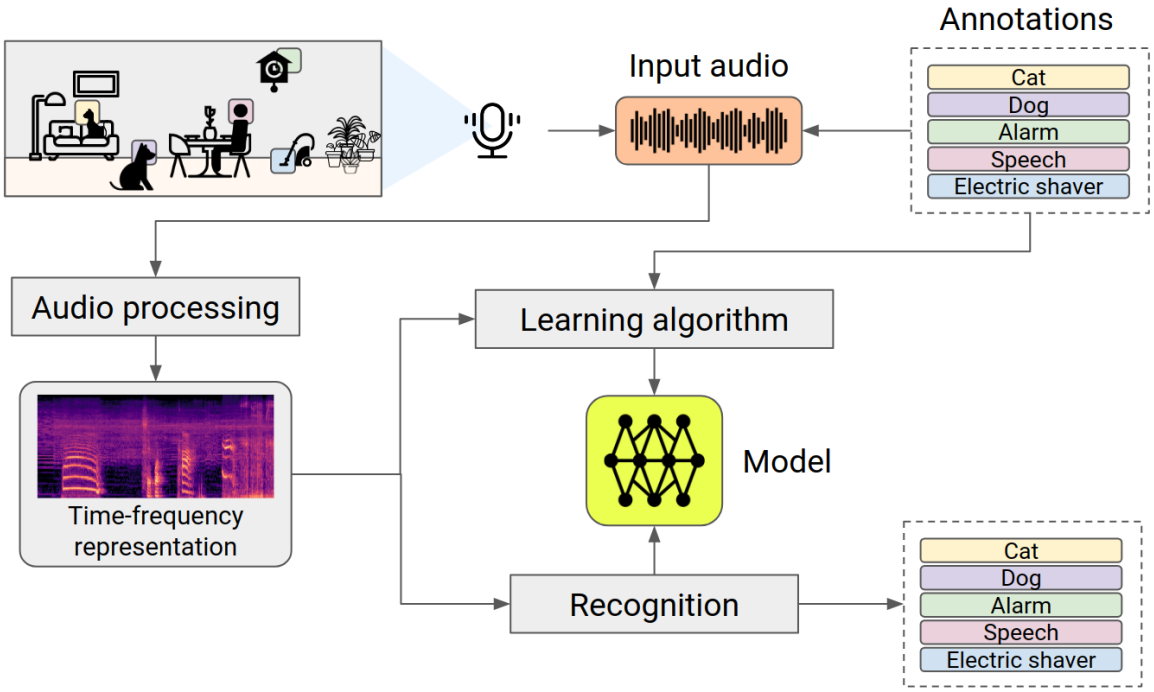


Figure 2.3: Audio analysis pipeline.

Audio tagging

Audio tagging deals with the recognition of the sound events composing a soundscape. The labels outputted by a tagging system describe the nature of the sound event instances. For an audio recording with an isolated sound event that may span the duration of the audio clip, the most simple tagging system assigns a single label. A more complex tagging system outputs as many labels as there are active classes in an audio recording. The task however does not take into consideration how many instances of a sound class appear in a recording.

Sound event detection

Sound Event Detection (SED) extends audio tagging by telling when sounds occur. Therefore, SED aims to identify both the class type and time boundaries of active sound events. In order to achieve this goal, SED first models the long term temporal context of sound events and then performs audio tagging over short consecutive audio segments. The ensemble of labels collected over time reveals activity patterns for each sound class, from which the temporal information of the sound event instances within a soundscape can be decoded.

Sound event localization and tracking

Sound event localization extends the sound event detection task by determining the physical locations where sound events are produced. Multichannel recordings captured with microphone arrays are generally required since spatial cues support the localization of sounds.

Audio captioning

Audio captioning is concerned with the description of general audio content using text. Generally, a system takes as input an audio recording and outputs the textual description of its contents, i.e., the caption. This task is often used for audio retrieval applications in which audio signals are retrieved using their textual descriptions.

2.1.3 Audio analysis applications of interest

In this section we present different sound analysis applications of interest that tackle real-world problems. Such tasks correspond to applications of general interest and in constant development.

Domestic environments

Audio analysis applications in domestic environments are intended to simplify domestic tasks. This is generally done through voice interaction, which is why nowadays many electronic devices enable automatic speech recognition by default. In addition to speech, there are plenty more sounds rich in information that take place in domestic scenarios. For instance, sounds like alarms, gunshots and glass breaking sounds can be indicators of danger. The analysis of such sounds is the core of audio surveillance applications, which ensure the security of the physical home environment by triggering an action, e.g., a phone call or notification. Besides, audio has a fundamental advantage over video in that it is less affected by lighting conditions and occlusion (Kolbæk et al., 2017; Alsina-Pagès et al., 2017).

Surveillance also extends to monitoring the behavior of individuals at home (van Hengel and Anemüller, 2009; Vacher et al., 2011; Huang et al., 2010) and center attention to human beings through the analysis of baby cries, falls and loss of autonomy (Torres et al., 2017; Zigel et al., 2009; Fleury et al., 2008). The difficulty posed to sound surveillance systems is associated with the frequency in which distress situations occur, which must nevertheless be detected on time (Chan and Eric, 2010; Doukas and Maglogiannis, 2010; Kim et al., 2020). The analysis of sounds that do not imply emergency or abnormal situations can also assist in the automation of tasks, for instance detecting footsteps at home to turn on the lights and recognizing a dog barking to feed them (She, 2004; Kim et al., 2018), or detecting running water to identify an eventual leak (Seyoum et al., 2017).

Urban sound analysis

Modern cities focus attention on technologies for building automation, smart mobility and digital management of infrastructure. This involves deployment of sensing equipment that analyzes the urban soundscapes as a cost effective solution or as a complement of other types of sensing modalities such as video (Bello et al., 2019).

Traditionally, recording urban soundscapes is performed through static sound monitoring stations. But since the number of sensing devices may be prohibitive for wide area coverage, the quality of sensing units must respond to a trade-off between cost and performance (Picaut et al., 2020; Vidaña-Vila et al., 2020). Maintenance costs, however, are much higher compared to mobile sensing alternatives. In this scenario, rather than collecting data with a fixed number of devices in a considerable number of measurement points, data collection is carried out by citizens using their smartphones (Ruge et al., 2013; Ventura et al., 2018). This approach has led to the development of several crowd-sourcing platforms that leverage the collaborative data to estimate noise density maps, which can then be used to track noise levels and sources of pollution

(Zappatore et al., 2017; Kanjo, 2010). While the collection of data is practical, it is prone to measurement errors. Moreover, a central hub is needed to run complex synchronization and calibration processes to handle incoming submissions, as well as various hardware specifications.

Beyond noise monitoring, the rich sounds of urban soundscapes can also serve as indicators of activity patterns associated with indoor or outdoor scenes such as parks or subway stations, as well as activities such as nightlife, tourism or construction sites. A more detailed analysis of the sounds that make up urban environments allows for closer monitoring of potentially dangerous situations, for instance, loud striking noise, people's cries and police sirens might be indicators of riots and manifestations (Ciaburro and Iannace, 2020; Tan et al., 2021).

Anomalous sound analysis

Anomalous sounds are indicators for irregular behaviors in plenty of tasks. The categorization of normal and abnormal sounds make each task unique (Mnasri et al., 2022). Industrial fault detection is an example application in which a careful inspection of sounds brings operational benefits (Duman et al., 2019). Automatic analysis of failure patterns is a cost effective maintenance solution that not only allows prompt detection of defective units, but also enhances operating equipment's reliability, which avoids long shutdown periods of equipment and instrumentation (Alaoui-Belghiti et al., 2019). Detecting abnormal sounds is however challenging due to their vast diversity and rare frequency of occurrence.

A simplest design choice for abnormal sound detection consists of modeling the narrow acoustic space of normal operating conditions, and detecting anomalies based on a specified criteria that measures deviation to the average expected behavior. This could be done, for instance, by computing scores through distances or a degree of similarity between training and test samples in some embedding space (Koizumi et al., 2020). Low deviation scores imply normal conditions, while high scores suggest faulty operations.

Other methods simulate abnormal sounds or rely on a small amount of data collected during the operational life cycle of industrial equipment (Purohit et al., 2019; Harada et al., 2021). In the latter approach, the data most likely contain abnormal sounds which can be used as a reference for modeling faulty behavior.

Some audio monitoring and surveillance tasks that follows the same principle as fault detection. For instance, traffic monitoring categorizes anomalies as sounds produced in hazardous events such as car accidents, tire skidding or harsh braking (Foggia et al., 2015; Socoró et al., 2017). Another example is in the field of human healthcare monitoring, where anomalous sounds assist in the early detection of diseases by identifying irregular patterns of biological audio signals, such as respiration and heartbeat (Rocha et al., 2017; Zabihi et al., 2016).

Bio-acoustics

Bioacoustics aims to study animal behavior and preserve wildlife through the acoustic signals they emit (Fletcher, 2014). A subject of interest for this cross-disciplinary science is the impact of human activities in waters. Seismic testing, pile driving, naval sonar, and maritime transport disrupt the acoustical environment of seas and oceans, compromising the ability of marine species to communicate, socialize, orient and navigate the waters (Parks et al., 2011). The anthropogenic noise generated by such activities also alters fundamental activities for their survival, such as feeding, mating, breeding and detecting predators (Gendron et al., 2020). Underwater noise pollution causes serious consequences for sea wildlife, which suffers from stress, physical injuries and strandings in a forced attempt to adapt (Popper and Hawkins, 2019).

Bioacoustics can also help track and preserve species through the recognition of their vocalizations (Fletcher, 2014; Stowell et al., 2016). Recording wildlife is thus key to analyze migration patterns, habitat displacement, climate change or if a species is at risk of disappearing (Kloepper and Simmons, 2014; Borker et al., 2014). Animal recordings are not only proof of biological activity, but can work as well as a time capsule (Sugai and Llusia, 2019). Evidence of past and current species represents invaluable historical records for future generations. In fact, there are even attempts to reproduce the sounds of extinct species based on morphology and behavioral features (Seilacher, 2007).

Bioacoustics implies a better understanding of historical changes in ecosystems and puts into perspective ongoing changes to preserve natural ecosystems. Thanks to bioacoustics, careful analysis of animal behavior assesses the impact of current ecosystem disruptions and calls for actions to interact with nature in a sustainable way.

2.1.4 Data

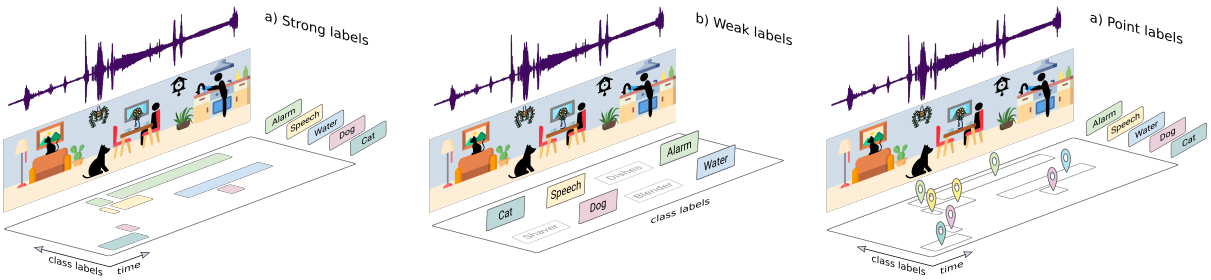


Figure 2.4: Labeling data for audio analysis systems.

Labeling data for audio analysis systems

Modern data-driven approaches to sound event detection require training acoustic models with some sort of supervision. In the context of sound analysis, labels can be of different nature depending on what elements or properties of the acoustic scene have reference annotations. These are illustrated in Figure 2.4

The type of annotations that indicate which sound classes are active in an audio sample are referred to as *weak* labels. As such, they don't provide information about how many times a sound event of a given class appears in the audio nor when it occurs. On the other hand, temporal information is provided by *strong* labels, which indicate both class information and when each sound event instance occurs. This is generally provided by timestamps which are used to know the duration of a sound event instance. Unlike weak labels, strong labels allow counting sound events with the same or different class label occurring in the acoustic scene.

In addition to weak and strong labels, there is another type of annotations referred to as *point labels* (Kim and Pardo, 2019). Such labels bear close similarity with strong labels, but they are not adopted in the development of sound analysis systems. Point labels relax the specification of having precise time boundaries imposed by strong labels and provide instead a single time mark (or point) that can be anywhere within the temporal limits of a sound event. Kim and Pardo (2019) showed that training sound event systems with point labels results in better performance than using weak labels alone, as well as comparable performance to using strong annotations.

Soundscape synthesis

Annotating sound recordings is often a subjective task for humans (Lafay et al., 2016). To ease manual categorization, human annotators follow a reference taxonomy, yet choosing the proper label for a sound event from a large pool of candidate labels is subjected to the annotator's own perception and experience.

Moreover, for sound event detection tasks, human annotators must as well identify precise temporal information of sound events, which is as well prone to high subjectivity. Generally, marking the starting time of a sound event is easier than identifying its ending time. Transients often give a clear cue about onset times (e.g., gun shot), but offsets tends to be fuzzy due to the gradual decay of sound energy levels (e.g, a car driving away). Furthermore, annotating temporal information of sound events is far less obvious in low signal-to-noise scenarios, or in the presence of high reverberation.

Soundscape synthesis (Salamon et al., 2017; Turpault et al., 2019) rises as an alternative way to manual labeling, avoiding labeling subjectivity to a great extent. Given a collection of isolated sound events, a synthesis process generates multiple soundscapes following desired specifications. Each synthetic soundscape is composed of a random mixture of sounds drawn from distributions with properties including event duration, start and end times, event-to-background ratio and heuristic signal transformations. Since a large number of unique soundscapes can be generated, soundscape synthesis is a suitable procedure for generating large-scale annotated datasets.

Synthetic soundscapes require, however, a certain degree of similarity with the acoustic scenes that could potentially be found in the deployment scenario, otherwise a mismatch between the training and testing conditions will occur. In fact, the challenge of simulating realistic soundscapes lies in replicating the acoustic conditions and distribution of sources of everyday acoustic scenes.

Data augmentation

Current datasets used to develop audio analysis systems are limited in size. This poses a problem to data-driven approaches since they require huge amounts of data to ensure generalization. Data augmentation techniques alleviate the data limit constraint by generating new training samples from manipulation of the available data. The enlarged dataset benefits from increased variability while meeting the model requirements to avoid overfitting.

Heuristic approaches perform transformations on physical properties of the original audio signals. Such transformations can take place in the time domain or in the time-frequency plane. Time domain transformations comprise transformations such as time stretching, which compresses or expands the signal in time or pitch shifting, which changes the perceived pitch of a sound by increasing or decreasing its fundamental frequency. A widely known computer package for soundscape synthesis and audio augmentation is Scaper (Salamon et al., 2017). On the other hand, transformations in the time-frequency plane include masking some time-frequency regions which force a model to regress the missing information, e.g., the SpecAugment method (Park et al., 2019a).

Datasets to carry out domain adaptation reseach

Audioset (Jansen et al., 2017) was introduced in 2017 to foster audio analysis research. It comprises weakly labeled audio recordings of over 500 sound classes from web videos. A recent upgrade of Audioset comprises the release of strong annotations for a subset of acoustic scenes comprising around 5% of the original data (Hershey et al., 2021). This labeling effort is intended

to show that fine-tuning a classifier previously trained on weakly labeled data with a small amount of strong labels that complete the original weak annotations delivers significant classification improvement. Several datasets have been built around Audioset, particularly for developing models for applications that require analyzing only a fraction of classes. A specific example is the Domestic Environment Sound Event Detection (DESED) dataset (Turpault et al., 2019), which focuses on domestic sounds. An alternative to Audioset is the Freesound dataset (FSD50K) (Fonseca et al., 2021), which contains more than 50k audio clips totalling over 100 h of manually labeled audio. It comprises 200 classes drawn from the Audioset Ontology. Unlike Audioset, FSD50K provides original audio sources rather than pre-computed features.

Smaller datasets have been curated from the Freesound project, such as the CURE dataset (Dubey et al., 2019), which allows the study of specific audio events relevant for hearing impaired people.

To address specific domain mismatches, several datasets have been introduced for different audio analysis tasks. The most relevant dataset for anomaly detection is the large-scale dataset called "ToyADMOS2" (Harada et al., 2021). It consists of two sub-datasets for machine condition inspection: fault diagnosis of machines with geometrically fixed tasks and fault diagnosis of machines with moving tasks. Domain shifts are represented by introducing several differences in operating conditions, such as the use of the same machine type but with different machine models and part configurations, different operating speeds, microphone arrangements and simulating environmental noise.

On generating realistic synthetic data

The DCASE Challenge 2021 Task 4 uses a heterogeneous setup that includes both recorded and synthetic soundscapes. Until recently only target sound events were considered when synthesizing the soundscapes. However, recorded soundscapes often contain a substantial amount of non-target events that may affect the performance. Ronchini et al. (2021) analyzed the impact of these non-target events in the synthetic soundscapes and to what extent using non-target events alternatively during the training or validation phase (or none of them) helps the system to correctly detect target events. They analyzed to what extent adjusting the signal-to-noise ratio between target and non-target events at training improves the sound event detection performance. Results showed that using both target and non-target events for only one of the phases (validation or training) helps the system to properly detect sound events, outperforming the baseline (which uses non-target events in both phases).

2.1.5 Features

Feature representations

Modern audio analysis systems can take various forms of the audio signal as input data. Modeling the relationship between inputs and labels is generally carried out using hand-crafted features extracted from the audio signals. Nevertheless, few works have attempted to conceive end-to-end classifications and detection models that learn features from the raw time domain signal (Lee et al., 2017; Sang et al., 2018; Schmitt and Schuller, 2019). Although proven successful for speech enhancement and separation tasks, their performance for sound classification or detection tasks is limited. The main drawback in such works is lack of enough data that complies with the high complexity that such modeling requires. For this reason, time frequency representations such as spectrograms, remain the preferred input features for developing audio analysis systems. The

compact dimensionality of such representations is a good trade-off between performance and model complexity.

Classifiers and detection systems can also be learned from general audio representations such as Audioset or L^3 pretrained features (Cramer et al., 2019). These representations are high level abstractions known as audio embeddings, since they are extracted from embedded layers of a neural network-based model. These audio embeddings which are learned from large scale datasets, serve as general purpose features for several downstream tasks where data is limited. A common use case is transfer learning, which re-purposes knowledge from pretrained audio features to a new application domain.

STFT Analysis

In this section we describe how to compute time-frequency representations from audio signals. The process is illustrated in Figure 2.5. Through Fourier analysis using the discrete Fourier transform (DFT), the time domain signal $x(t)$ is projected onto a basis of complex exponentials with linearly spaced frequencies as

$$x(f) = \sum_{m=0}^{L-1} x(m) e^{-2j\pi mf/F}, \quad f \in 0, \ldots, F-1 \tag{2.1}$$

where F is the number of frequency bins, L is the length of the signal, and j is the imaginary unit. The complex representation $x(f)$ can be decomposed into its *magnitude* $|x(f)|$ and *phase* $\angle x(f)$ spectra.

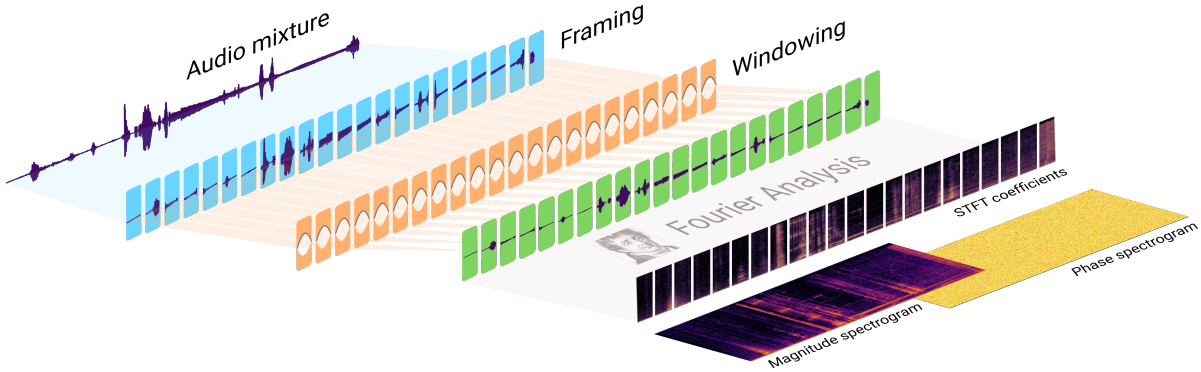


Figure 2.5: Signal analysis.

When processing long audio signals, the Fourier analysis is rarely carried out over the whole signal but rather over small overlapped segments since the frequency content of audio signals changes rapidly over time. This is achieved by the short time Fourier transform (STFT) which produces time-frequency representation referred to as the *spectrogram* of the audio signal.

First, the input mixture signal is segmented into fixed-length frames that vary according to the signal properties of the sources, e.g, between 10 and 50 ms for speech and 120 ms for music signals. Commonly, frames overlap by 50 or 75%. After segmentation, each frame is multiplied by a window function. The segmented and windowed signal $x(n,t)$ in frame $n \in 0, \ldots, N-1$ is defined as

$$x(n,t) = x(t+nM)h_{\mathrm{a}}(t), \quad t \in 0, \ldots, T-1 \tag{2.2}$$

where N is the number of time frames, T is the number of samples in a frame, M is the hop size between adjacent frames in samples, and $h_a(t)$ is the analysis window.

After windowing, the discrete Fourier transform (DFT) is applied to each windowed frame which results in complex-valued STFT coefficients

$$x(n,f) = \sum_{t=0}^{T-1} x(n,t)e^{-2j\pi tf/F}, \quad f \in 0,\dots F-1 \tag{2.3}$$

where F is the number of frequency bins, f is the frequency bin index.

Typically $F = T$. The frequency in Hz associated with the positive frequency bins $f \in 0,\dots,[F/2]$ is defined as $v_f = \frac{f}{F} f_s$, where f_s is the sampling frequency, that is a linear function of the frequency bin index f.

The resulting complex-valued STFT coefficients can be decomposed into phase $\angle x(n,f)$, magnitude $|x(n,f)|$ or power $|x(n,f)|^2$ coefficients. The magnitude or power spectrogram are commonly selected for further processing in audio analysis tasks.

Long windows give high frequency resolution but poor temporal resolution yielding a *narrow-band* spectrogram. Conversely, short windows produces the opposite effect yielding a *broadband* spectrogram. This is an inherent limitation of time-frequency analysis known as *Gabor limit*, analogous to Heisenberg's uncertainty principle, which realizes that the exact time and frequency of a signal cannot be known simultaneously (Gabor, 1946).

2.1.6 Models

Deep learning architectures

The great success of neural networks in domains such as computer vision extended their use to solve audio-related tasks. Shortly after the great success in speech separation and enhancement, deep neural networks established themselves as the preferred choice for the development of sound recognition systems. Thanks to mass availability of data and extensive computational power, neural networks outperformed well-established traditional methods, and prevail now as the state-of-the art approaches.

Convolutional neural networks

The ability of deep convolutional networks (CNN) to capture temporal and spectral characteristics of audio signals makes them a powerful tool for a wide range of audio analysis tasks. Convolutional layers perform filtering operations by convolving the image with a series of local filters. Filtering and pooling operations allow learning global representations of the image by extracting local features. These operations make CNNs powerful tools for learning high-level data representations. Their use in audio applications is straightforward when analyzing time-frequency representations as input features. The first applications of CNNs to environmental sounds concern classification and detection problems (Piczak, 2015). To date, DCASE is marked by the proposal of a vast number of classification and detection systems that employ CNNs in some part of their processing pipeline. Furthermore, large-scale datasets such as Audioset and FS50K of promote the exploration of more complex CNN architectures.

Recurrent neural networks

Recurrent neural networks (RNNs) allow the classification of sequences of observations, where each neuron depends on the input at the current time step and its output at the previous step. By using long-short term memory (LSTM) cells (Hochreiter and Schmidhuber, 1997) or

gated-recurrent units (GRU) (Cho et al., 2014), recurrent networks can model much longer time dependencies than Hidden Markov models (HMMs), which results in huge performance improvements. For such a reason a vast majority of neural network-based systems model dependencies of sequential data with recurrent networks. Parascandolo et al. (2016) introduced for the first time recurrent neural networks for sound event detection.

Convolutional recurrent neural networks

The most successful models to date for audio analysis tasks comprise an ensemble of convolutional layers and recurrent layers, i.e., convolutional recurrent neural networks (CRNNs) (Cakır et al., 2017; Xu et al., 2017). The convolutional layers learn representations from modeling the local characteristics of the time-frequency representations, while the recurrent layers model the temporal dependencies of the acoustic scene. These models are the default choice for sound event detection and offer strong improvements over individual convolutional or recurrent layers.

Transformer architectures

In the past two years, Transformer-based models have emerged for audio analysis tasks. The great success of recent attention mechanisms in computer vision (Guo et al., 2022) questions whether current CRNN approaches are still the best for audio classification and detection tasks. Since such architectures are capable of modeling longer feature dependencies than LSTMs and support parallel processing, they are used to replace the RNN component of CRNNs showing competitive performance (Miyazaki et al., 2020; Kong et al., 2020; Ye et al., 2021).

2.1.7 Training

Semi-supervised frameworks for sound classification and detection

The limitation of large-scale annotated data to train SED systems motivated the use of unlabeled data. For audio classification, Elizalde et al. (2017) proposed to combine labeled audio and unlabeled data from the web to improve audio classification models with self-training. In this semi-supervised learning approach, the large pool of unlabeled data passes through a set of classification models to generate *pseudo-labels*, which are then used to retrain the classifiers. Since this approach only selects samples with high-confidence predictions for retraining, it is susceptible to model bias, thus not necessarily more robust for classification.

For sound event detection, a popular and now well-established semi-supervised model that exploits unlabeled data is the Mean Teacher framework (Tarvainen and Valpola, 2017). Since Delphin-Poulat et al. (2020) introduced it in the DCASE 2019 challenge, and after ranking second among 58 systems, this semi-supervised framework has prevailed to leverage unlabeled data.

Alternative semi-supervised frameworks adapted to SED include Interpolation Consistency Training (ICT) and Shift Consistency Training (SCT) (Koh et al., 2021). The former frameworks is suitable for learning from ambiguous smaples, whereas the latter framework encourages the prediction of time-shifted inputs to be consistent with the time-shifted predictions to improve the temporal localization of sound events. Both approaches can be viewed as well as data augmentation approaches that increase the diversity of the training data.

Exploiting unlabeled data

Martín-Morató et al. (2021) presented a method that produces strong annotations for weakly labeled data with high precision. Labels are acquired based on the estimation of the annotator

competence, aggregation and processing of the weak labels. The resulting aggregated annotation is objective, being composed of multiple opinions. The experiment was carried over synthetic data in order to verify the quality of resulting annotations through comparison with ground-truth annotations. Unsupervised methods for sound analysis have been also carried out. For instance Jansen et al. (2017) proposed a method to learn audio representations in an unsupervised way by sampling triplets that elicits semantic structures in the absence of labeled data. The method produces representations invariant to additive noise and translations in time, and embeddings that inherit the categories of the constituents.

2.1.8 Performance evaluation of audio analysis systems

Several computational metrics have been proposed to assess the performance of sound event detection systems (Mesaros et al., 2016; Bilen et al., 2020). These metrics take the counts of correct and wrong predictions made by the system known as *intermediate statistics* (Virtanen et al., 2018). For a sound class c of interest and in accordance with some ground-truth reference they are defined as follows:

- *True positive (TP)*: an output which correctly indicates that class c is active.
- *True negative (TN)*: an output which correctly indicates that class c is inactive.
- *False positive (FP)* or *insertion*: an output which wrongly indicates that class c is active.
- *False negative (FN)* or *deletion*: an output which wrongly indicates that class c is inactive.

Some metrics define a *substitution* output in the particular case when the system outputs a false positive and a false negative at the same time, i.e., an output which indicates that a class (other than class c) is active and whose activity pattern matches that of reference class c.

When the evaluation is carried out by comparing the system outputs against the reference annotations on a fixed temporal grid with a time resolution as short as the analysis frame or as long as 1 s, the evaluation is referred to as *segment-based*. With this metric, the performance shows the ability of the system to correctly detect the activity patterns of sound events. Conversely, if the evaluation consists of comparing the system outputs against the reference annotations in terms of individual sound event instances, the evaluation is referred to as *event-based*. With this metric, the performance shows the ability of the system to detect the onset and offset of sound event instances.

For event-based metrics, the activity patterns of the detected events may extend or fall short of the temporal boundaries defined by the ground-truth. In order to avoid penalizing the system for small timing error, it is common to allow some degree of misalignment or *collar* between the system outputs and the reference annotation. An output is thus counted as a true positive if the temporal boundaries of a correctly identified sound event class lie within either a fixed length collar (e.g., 100 ms) or a proportion of the sound event duration (e.g., 50 %) with respect to the reference annotations.

Common metrics

Accuracy

Accuracy (ACC) is computed as the ratio of correct system outputs to the total number of outputs.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.4}$$

Accuracy measures how often the system yields correct outputs. As it is influenced by the class balance, intermediate statistics for rare classes may lead to paradoxically high accuracy classes.

Precision, recall, and F-score

Precision (P), *recall* (R), *F-score* are common metrics used in many sound classification and detection tasks. Precision and recall are defined as:

$$P = \frac{TP}{TP + FP}, \tag{2.5}$$

$$R = \frac{TP}{TP + FN}, \tag{2.6}$$

$$\tag{2.7}$$

and based on their harmonic mean, the F-score is calculated as:

$$F = \frac{2PR}{P + R}. \tag{2.8}$$

$$\tag{2.9}$$

An F-score of 1 indicates perfect precision and recall, while the lowest possible value is 0, if either P or R is 0. Its main drawback is that its value is influenced by the choice of averaging and similarly to accuracy, by the class balance.

Error rate

The *error rate (ER)* is an alternative metric that counts the substitution, deletion and insertion errors made by the system. Such intermediate statistics are defined as:

$$S = \min(FN, FP) \tag{2.10}$$

$$D = \max(0, FN - FP) \tag{2.11}$$

$$I = \max(0, FP). \tag{2.12}$$

The error rate is calculated as the ratio of the overall S, D and I values and the total reference counts N as:

$$ER = \frac{S + D + I}{N}. \tag{2.13}$$

An ER value of 0 means perfect classification and it can exceed 1 in particular cases.

For segment-based evaluation the intermediate statistics are calculated segment by segment and accumulated for all evaluation data, whereas for event-based evaluation the intermediate statistics are computed based on the class label and temporal location of the system's outputs.

Averaging

In order to provide robustness to class imbalance, the computation of the intermediate statistics can be carried out either globally (*micro-averaging*) or separately for each class (*macro-averaging*). The intermediate statistics in the former choice are accumulated over all available data, while for the latter they are first accumulated for each class and then averaged to get the final score.

PSDS

Since the use of collars emphasizes the detection of the temporal boundaries of sound events, it is not robust to the subjectivity of the human labeling of these timings. To alleviate this constraint, the recent *polyphonic sound detection score* (PSDS) proposed by Bilen et al. (2020) redefines true and false positives based on the intersection between the system outputs and the reference annotations. The PSDS allows fragmented outputs as long as they intersect sufficiently with the ground-truth. Furthermore, the score is defined as the normalized area under the receiving operating characteristic (ROC) curve, which makes the evaluation independent of a specific decision threshold thus providing a full picture of the system's performance over all decision thresholds.

However, the PSDS-ROC curve is only approximated using a finite set of thresholds and the choice of the thresholds can have a severe impact on the resulting score. Recently, Ebbers et al. (2022) proposed a methodology to efficiently compute the system's performance for all possible decision thresholds jointly, which allows to accurately compute the PDSDS.

2.2 Ambient source separation

The outstanding ability of human beings of directing attention to a sound of interest in the presence of interfering sources and noise has long inspired audio source separation research. Efforts have mainly focused on separating two types of signals: speech and music. Since speech is essential for communication, it is important to isolate it from unwanted interfering signals such as noise, as well as from other speech signals (Loizou, 2007). As for music, the interest lies in decomposing it into separated stems such as vocals, bass and drums (Cano et al., 2018). Recent exceptional results achieved in the segregation of speech and music have attracted attention to the separation of the wide variety of arbitrary sounds composing everyday acoustic scenes. In this section we will describe the foundations of source separation and will provide an overview of current methods that address the separation of ambient sounds.

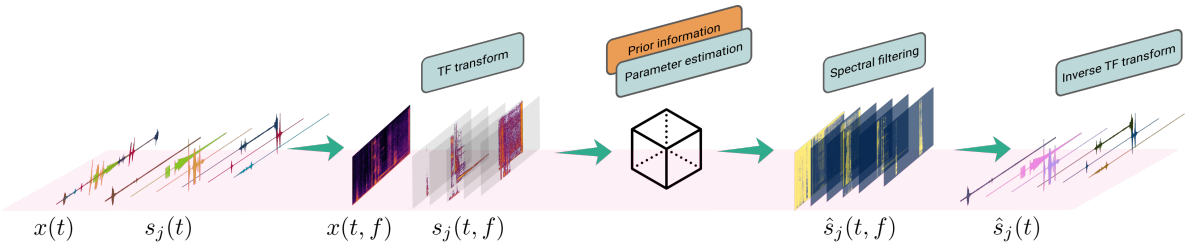


Figure 2.6: General source separation scheme.

2.2.1 General separation scheme

Source separation is addressed following a general processing flow that applies to single-channel settings and can be extended to multi-channel scenarios. As depicted in Figure 2.6, the time-domain mixture signal $x(t)$ and the sources $s_j(t)$ are represented in the time-frequency domain through complex-valued time-frequency coefficients $x(t,f)$ and $s_j(t,f)$, respectively. The parameters of the separation model are estimated either from the mixture $x(t,f)$ or from a separate training corpus. If additional prior information is available, it can be used in parameter

estimation. Given these parameters, a time-varying single-output complex-valued filter is derived and applied to the mixture $x(t, f)$ in order to obtain an estimate of the complex-valued time-frequency coefficients of the sources $\hat{s}_j(t, f)$. Lastly, the time-frequency representations are inverted producing time-domain source estimates $\hat{s}_j(t)$.

From the above general source separation processing scheme, Vincent et al. (2018) distinguish the following ways to characterize source separation methods based on the nature of the training data and how it is used to estimate the model parameters:

- *learning free* methods in which the model parameters are not estimated through a training stage but rather estimated from the test mixtures,
- *unsupervised source modeling* methods in which models are learned for each source using unlabeled isolated signals of that source type,
- *supervised source modeling* methods in which models are learning with additional supervision compared with unsupervised source modeling methods, e.g., pitch information,
- *separation based training* in which the model parameters are learned using a training corpus with mixture signals and their constituent true source signals.

2.2.2 The mixing process

Source separation aims to estimate the constituent sound sources of a complex audio mixture (Vincent et al., 2018). The separation of arbitrary sounds in a complex mixture such as the one illustrated in 2.7 is commonly referred to as *universal sound separation* (Kavalerov et al., 2019). Since everyday acoustic scenes are composed of overlapping sounds and are often corrupted by noise and reverberation, this task is relevant for the development of audio analysis models for its potential use as a pre-processing step prior to sound recognition.

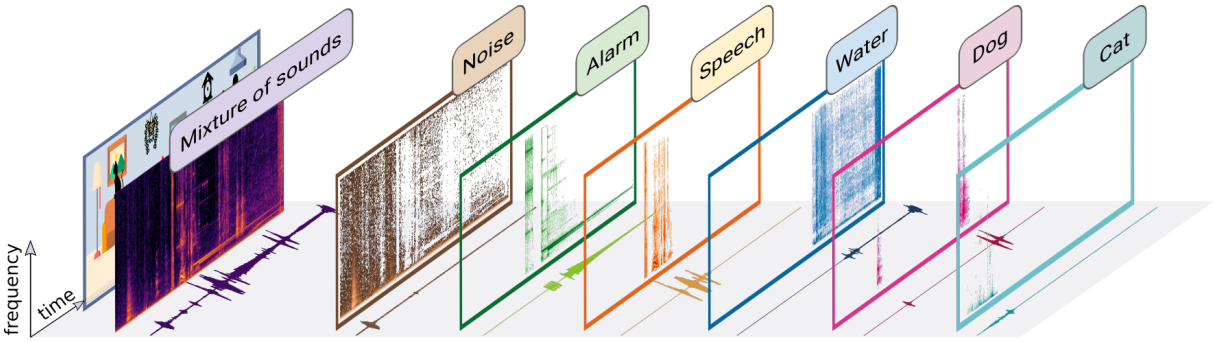


Figure 2.7: A complex mixture of sounds and its constituent sound sources.

Essentially, complex audio mixtures result from a linear process in which multiple scaled and delayed (or reverberated) versions of the sources are combined linearly. An audio mixture captured by multiple microphones results in a *multichannel* signal or in a *single-channel* signal if captured with a single microphone. The works in this thesis deal with single-channel signals only.

In the time domain, a single-channel complex audio mixture $x(t)$ is modeled as

$$x(t) = \sum_{j=1}^{J} s_j(t) \tag{2.14}$$

where each source $s_j(t)$, depending on the goal of the desired task, constitutes a *target* signal or a noise signal, and J indicates the number of sources.

The signal-to-noise ratio (SNR) of a target signal s in the mixture compares its power level to the power level of an interfering signal n as

$$\text{SNR} = 10\log_{10}\frac{||s||^2}{||n||^2} \tag{2.15}$$

and is expressed in decibels (dB). A SNR level higher than 0 dB means more of the signal than the unwanted interfering signal. Although recent works show that end-to-end source separation in the time domain leads to good performance in universal sound separation (Kavalerov et al., 2019), in this thesis we focus on the separation of ambient sounds in the time-frequency domain.

As discussed in Section 2.1.5, the analysis of a time domain acoustic signal by the STFT results in a time-frequency representation known as the *spectrogram*. Since the STFT is a linear transformation, the resulting signal in the time-frequency domain satisfies

$$x(t,f) = \sum_{j=1}^{J} s_j(t,f). \tag{2.16}$$

The complex values of the STFT coefficients allow easy treatment of the magnitude $|x(t,f)|$, power $|x(t,f)|^2$ or phase $\angle x(t,f)$ spectrograms. In this thesis we work with magnitude or power spectrograms and re-use the phase information from the input mixture for conceiving source separation models.

2.2.3 Masking-based source separation

Source separation approaches in the time-frequency domain estimate a *mask* whose values weight the time-frequency representation of the audio mixture. The mask aims to emphasize the time-frequency regions dominated by the target source while suppressing the regions dominated by other sources. Figure 2.8 shows a general overview of masking-based source separation.

Masks can be classified into *binary*, *soft* and *complex* depending on the type of values they can take. The mask is denoted as $m_j(t,f)$, and the masking operation is written as

$$\hat{s}_j(t,f) = m_j(t,f)x(t,f). \tag{2.17}$$

Masks can establish a reference separation performance according to some metric when the sources composing an audio mixture are available. A benchmark of this type is known as the *ideal* or *oracle* mask, which was suggested by Wang (2005) as the computational goal of CASA. It denotes the target energy in a time-frequency unit as $s(t,f)$ and the interference energy as $n(t,f)$. The ideal binary mask is given by

$$\text{IBM}_j(t,f) = \begin{cases} 1 & \text{if } \frac{|s_j(t,f)|}{|n(t,f)|} > \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.18}$$

Contrary to the binary mask, the soft mask takes on continuous values between 0 and 1, corresponding to the ratio of magnitudes between target interference signals. The ideal soft mask is defined as

$$\text{IRM}_j(f,t) = \frac{|s_j(f,t)|}{\sum_{i=1}^{J}|s_i(f,t)|}. \tag{2.19}$$

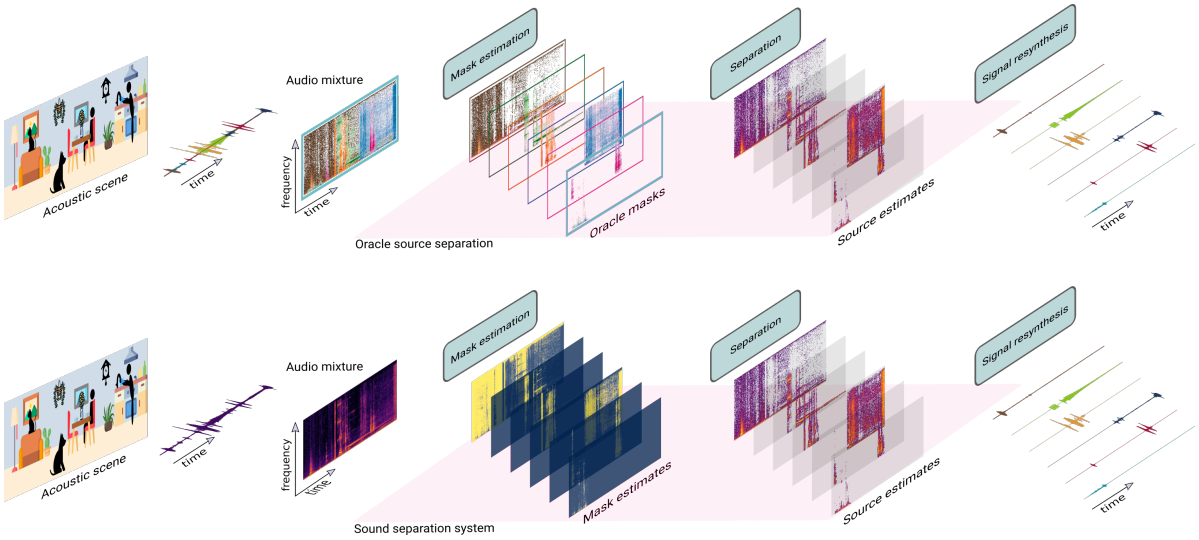


Figure 2.8: Masking-based separation by oracle separation (top) vs. masking-base source separation by the estimation of the sources (down). In oracle source separation masks are estimated perfect knowledge of the sources.

The IRM shares all properties of the IBM laid out by Wang (2005), matching them or exceeding them. According to Hummersone et al. (2014), the IRM is identically flexible in the sense that any source can be designated as the target while the sum of the remaining sources acts as interference. The IRM is also well-defined since the interference component can contain an arbitrary number of sources. However, the IRM agrees better than the IBM with psychoacoustic principles, particularly when dealing with complex mixtures of sound sources (e.g., when dealing with multiple sources or with noisy and reverberant environments).

Figure 2.9 shows an example of oracle separation in the logarithmic time-frequency domain by ideal binary and ratio masking to isolate a dog bark from a mixture of domestic sounds. The figure shows the spectrograms of the mixture and clean signals as well as the IBM and IRM masks. The masked mixtures correspond to estimates of the target sound which are close to the clean target.

Resynthesis

A resynthesis path allows a time-frequency representation of an audio signal to be converted back to the time domain. This favors a systematic evaluation that measures the quality of the separation model according to some criteria. After masking-based separation, resynthesis is achieved by inverting the STFT of the estimate $\hat{s}(n, f)$ of the target source as

$$\hat{s}(n,t) = \frac{1}{F}\sum_{f=0}^{F-1} \hat{s}(n,f)e^{2j\pi tf/F}, \quad t \in 0,\ldots,T-1. \tag{2.20}$$

The filtering process used to estimate the target source may introduce *musical noise* the temporal signal that is typically most audible at the frame boundaries. To avoid such artifacts, each frame $\hat{s}(n,t)$ is multiplied by a synthesis window $h_{\mathrm{s}}(t)$. The entire time domain signal $\hat{s}(t)$ is obtained as

$$\hat{s}(t) = \sum_{n=0}^{N-1} \hat{s}(n, t-nM)h_{\mathrm{s}}(t-nM). \tag{2.21}$$

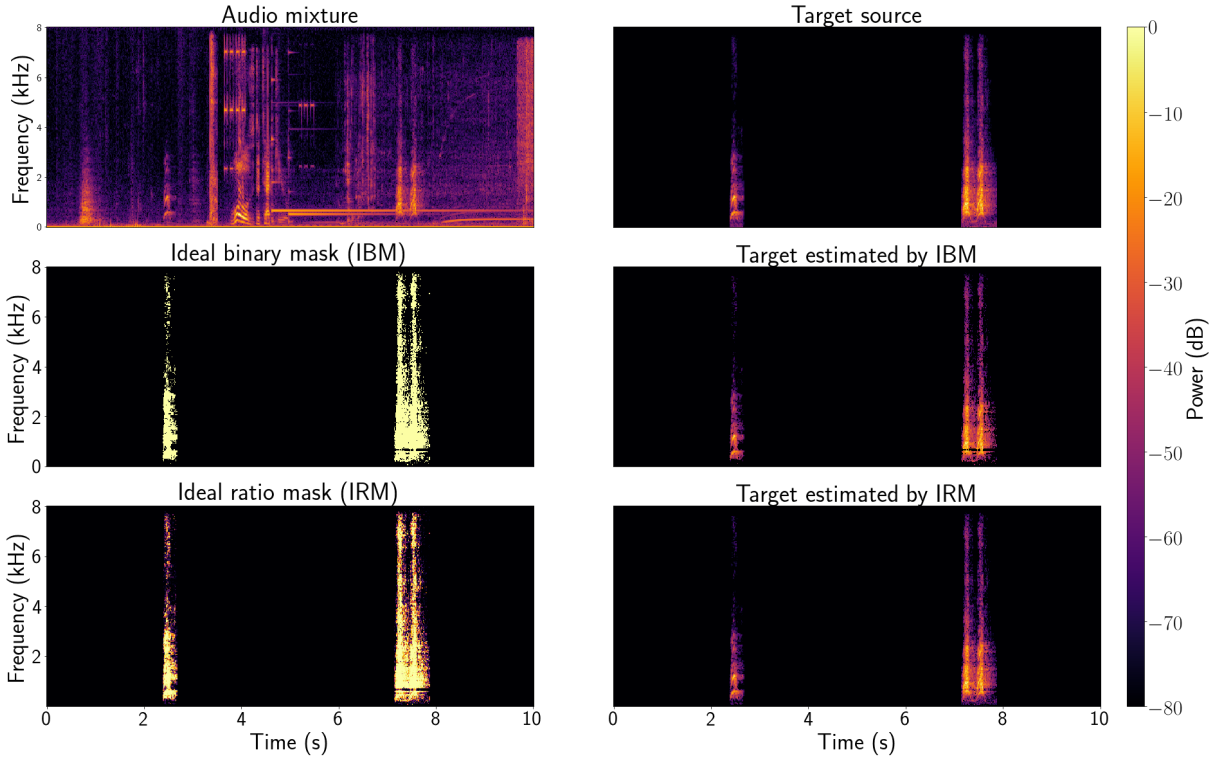


Figure 2.9: Separation of a dog bark (top right) from a mixture of domestic sounds composed of vacuum cleaner, dishes, alarms and speech sounds (top left). The middle row shows the IBM of the dog bark and its application to the mixture. The bottom row shows the sames as the middle row for the IRM. The masks correspond to oracle masks.

The synthesis window, and the analysis window must however satisfy the *perfect reconstruction* property: if no filtering is applied to the time-frequency representation of the mixture signal $x(t, f)$, the original time domain signal $x(t)$ is recovered if the analysis-synthesis window pair satisfies the completeness condition (Smith, 2011)

$$\sum_{n=0}^{N-1} h_a(t - nM)h_s(t - nM) = 1, \quad \forall t. \tag{2.22}$$

In such a case a rectangular synthesis window $h_s(t) = 1$ suffices when common bell-shaped analysis windows such as Hamming or Hann windows are used (Heinzel et al., 2002). In practice, however, time-frequency masking applies gains to the spectral coefficients of the mixture signal, hence the use of a synthesis window is required to avoid aliasing artifacts.

2.2.4 Evaluation

The evaluation metrics used to measure the performance of source separation systems include objective evaluation metrics which compare the estimated target to the clean (unmixed) signal, and subjective scores resulting from formal listening tests.

When ground-truth signals are available through an evaluation corpus, a number of objective evaluations can be defined. The goal of objective metrics is to quantify distortions of the estimated source $\hat{s}$, which can be decomposed into interferences e_{inter}, i.e., remaining portions of

non-target signals and *musical* noise artifacts as

$$\hat{s} = s_{\text{target}} + e_{\text{inter}} + e_{\text{artif}}. \tag{2.23}$$

Based on such decomposition Vincent et al. (2006) proposed the source-to-distortion ratio (SDR), source-to-interference ratio (SIR) and source-to-artifacts ratio (SAR) metrics, which are defined as

$$\text{SDR} = 10\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{inter}} + e_{\text{artif}}||^2} \tag{2.24}$$

$$\text{SIR} = 10\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{inter}}||^2} \tag{2.25}$$

$$\text{SAR} = 10\log_{10}\frac{||s_{\text{target}} + e_{\text{inter}}||^2}{||e_{\text{artif}}||^2}. \tag{2.26}$$

Based on the above metrics, Le Roux et al. (2019) proposed the scale-invariant signal-to-distortion ratio (SI-SDR) as well as the scale-invariant signal-to-interference ratio (SI-SIR) which allow scale invariance. Today, these metrics are widely used in the evaluation of end-to-end source separation systems (Pariente et al., 2020; Zeghidour and Grangier, 2021; Wisdom et al., 2020).

2.2.5 End-to-end ambient sound source separation

Soon after the achievement of outstanding results in speech enhancement and separation tasks, some researchers began to turn their attention towards the separation of sounds other than speech or music. The main motivation was to investigate the potential of current speech separation methods for the separation of sounds of a wide variety of classes and of a much less structured attributes. In an initial work carried out by Kavalerov et al. (2019), the source separation task applied to ambient sounds was referred to as *universal sound separation*. This novel task inherited the state-of-the-art models in speech separation and enhancement such as the TasNet architecture (Luo and Mesgarani, 2018), as well as a considerable fraction of their surprising performance, achieving an average SI-SDR of almost 10 dB. This results was further improved by Tzinis et al. (2020) with an iterative setup that conditions the separation model on semantically rich audio embeddings extracted from a pre-trained sound classifier. In a later work, Tzinis et al. (2022) proposed efficient models to carry on universal sound separation in low resource settings.

More recently, Wisdom et al. (2020) proposed an unsupervised approach to universal sound separation that can be trained by mixing only real-world audio mixtures (mixtures of mixtures) as opposed to isolated source signals. In a similar line, Pishdadian et al. (2020) designed a separation method that also avoids the need for ground truth data, albeit relying on weak supervision to learn to separate sounds. The key idea is to use a pre-trained audio classifier to enforce that each source estimate contains only the sound of a predefined class.

2.2.6 Target sound extraction

A task closely related to source separation is *target sound extraction*. It consists of selecting and extracting a specific source from an audio mixture rather than estimating all constituent sources. This comes in handy for situations where the focus is on a particular sound source while ignoring others. A common issue of source separation methods is the requirement of a priori knowledge

of the maximum number of sources in a mixture and the lack of source selection capabilities according to user-specified classes. Although the latter limitation could be addressed with a later stage of audio classification, the process remains sub-optimal. Target sound extraction addresses these limitations optimally by using auxiliary clues that inform the sound separation model. A few works have been proposed to extract the voice of a target speaker from a mixture of speakers (Wang et al., 2019a; Žmolíková et al., 2019), or speech from non-speech sounds (Huang et al., 2014; Weninger et al., 2015). Besides the speech domain, some works have proposed to extract individual instruments from music (Slizovskaia et al., 2019, 2021), including conditioning to specific instrument classes (Slizovskaia et al., 2019).

The extraction of arbitrary sounds has recently gained popularity with the applications of methods inspired from previous tasks. Similarly to the extraction of speech or music, most approaches rely on two schemes: a conditioning vector that encodes the sound class to be extracted, and a vector that summarizes the acoustic information of an excerpt of the target sound. In the first scheme, Ochiai et al. (2020) proposed the *SoundSelector* method to extract a desired sound event from a mixture of arbitrary sounds given a vector encoding the class of interest. This approach can potentially extract unlimited classes. An example of the second scheme is the *SoundFilter* method proposed by Gfeller et al. (2021), which filters out sounds from a mixture using only a short sample of the target sound. This method is even capable of extracting unseen sounds. In the same line of works, *SoundBeam* proposed by Delcroix et al. (2022) combines enrollment and one-hot encoders with multi-task learning to enforce a shared embedding space that improves target extraction. The strength of both conditioning approaches translates into performance improvement for the extraction of multiple classes as well as new classes. It also allows handling inactive target sound classes, which previous works have not evaluated thoroughly. A recent extraction method proposed by Wang et al. (2022a) combines target sound extraction with SED to guide the localization and extraction of target sounds using time information.

Other modalities such as text can also support the target sound extraction task, leading to interesting applications. For instance, Kilgour et al. (2022) proposed a text-driven universal sound separation that combines a natural language model named *SoundWords* with SoundFilter. The multimodal training of this approach yields a shared embedding space that co-locates sounds with their underlying semantic textual descriptors. Another interesting approach that employs text information was proposed by Okamoto et al. (2022), whose method conditions the source separator with onomatopoeic words that specify the target sound to be extracted.

2.2.7 Sound separation for sound recognition tasks

Source separation is appealing to sound event detection since a pre-processing treatment of the input audio mixture can get rid of unwanted and irrelevant sources for the analysis task, for instance, noise and non-target sound events. Source separation can also improve the detection of overlapping sounds. Most works involving source separation and sound event detection, use the former task as a pre-processing stage for the latter task. In this scenario, both the separation and SED systems are first optimized individually and then connected in cascade. This integration is however sub-optimal since errors generated in each block tend to accumulate.

In order to investigate how to best combine both tasks, Turpault et al. (2020) and Sharma et al. (2020) explored different levels at which the source estimates from a separation system can be integrated into a sound detection model. Results showed a modest overall improvement in the detection task, which Turpault et al. (2020) attributed to the mismatch between the source separation and sound separation training conditions. A further study by Turpault et al. (2021) benchmarked submissions to the DCASE Challenge 2020 Task 4, a Challenge edition in which the

use of separation systems was motivated to improve the performance of a SED baseline system. They showed that source separation provides SED systems with robustness to non-target sound events, however they do not provide conclusive clues on the optimal way to combined both systems. In a recent work, Sudo et al. (2020) proposed to train a sound separation model and an audio classification model in a jointly fashion for environmental sound segmentation, but little improvement was achieved for the segmentation of overlapping sounds. This limitation was overcome by including spatial cues in the multi-channel setting (Sudo et al., 2021), nonetheless failing under high reverberant conditions (Giannoulis et al., 2021). Undoubtedly, the optimal integration of both tasks requires further research.

2.3 Tackling the training-test mismatch

Audio analysis systems tend to run into serious trouble when the test conditions differ from those considered at the time of their conception. During model development it is often assumed that the training and test data are created by the same data generating process. However, in practical scenarios this assumption is generally not true, which prevents the system to perform optimally once deployed.

Domain adaptation (Ben-David et al., 2006) proposes a systematic solution to counteract any discrepancies between the training (*source domain*) and test (*target domain*) data. The goal is to minimize the performance gap between the source and target domains based on the observation that a shift has occurred between the source and target data distributions.

In this section, we first introduce the type of distribution shifts. Then, we present a selection of the plethora of recent works concerning the adaptation of audio analysis systems. According to the way in which the adaptation is carried out, we categorize such methods into three main groups, including feature normalization strategies, deep domain (model-based) adaptation techniques and retraining approaches.

2.3.1 Categorization based on the type of distribution shift

According to Kull and Flach (2014) a shift in the data distribution can be categorized mainly into *label shift*, *covariate shift* and *concept shift*. Other possible configurations combine any of these shifts. For instance, a hypothesis of practical interest in real scenarios assumes that covariate and label shift exist. In the following we define the aforementioned distribution shifts.

Label shift

Label or target shift assumes that the classes are not observed in the same proportion in the source and target domains. This type of distribution shift has gained recent interest in many computer science tasks, as it becomes critical for real-world applications where some of the classes may not even be present. Despite its relevance for optimal performance in real scenarios, this type of distribution shift has been little explored in audio analysis tasks. A use case of interest would be for instance to tackle the mismatch between detection of sounds in short vs. continuous recordings.

Covariate shift

In this scenario, the distribution of input features shifts between the source and the target domains, but the relationship between the features and the class labels is preserved. Most of

the proposed domain adaptation techniques for sound analysis tasks aim to solve this type of distribution shift. For instance, in the acoustic scene classification task, a common cause of covariate shift is the use of different microphones at training and test time. Accounting for the differences between microphone specifications suffice to correct such a mismatch. Task 1 of the DCASE challenge has long motivated the development of methods that generalize to mismatched recording devices. In a recent edition of the Challenge series, Heittola et al. (2020) exhibited that many submissions relied on data augmentation techniques to boost performance in the target domain, and only a handful of submissions developed explicit domain adaptation methods to achieve generalization under mismatched conditions.

Covariate shift can still arise with matched recording devices if the recording conditions at training and test time differ. Distortions introduced by background noise, reverberation, variation of the physical distance between the sources of interest and the microphone, and the presence of non-target sounds degrade the quality of the acoustic signal. Such acoustic interference has a negative effect in the recognition of sound scenes and events in real environments, particularly because several of these factors are present or combined. A simple method to increase robustness to adverse conditions consists of the simulation of degrading factors during model development, for instance, by mixing recordings with different background noises at different signal-to-noise ratios (Turpault et al., 2019), by convolving them with different room impulse responses (Turpault et al., 2021), or by mixing sound events of interest with non-target sounds to improve classification against their interference (Ronchini et al., 2021). Despite the attempt of these methods to simulate real-world conditions, a performance gap between synthetic and real data prevails. Only few works have used domain adaptation methods to account for the distribution mismatched between synthetic and real data.

Concept shift

In this scenario, the distribution of the features remain the same in the source and the target domains. However, the relationship between the features and the class labels is not preserved. Since the conditional distributions differ between domains, it is not possible to learn a model that predicts the class label as a function of the input features. This would correspond to the case where the labeling of sound events has been carried out systematically in two different ways across domains, or when the acoustic properties of a sound event evolve with time such that the initial relationship between features and class labels changes abruptly. This type of distribution shift is not addressed in sound analysis tasks since current datasets do not present such problems.

2.3.2 Retraining

The development of modern audio analysis systems requires a large amount of training data to ensure generalization. The high complexity of such systems makes them prone to overfitting if this need is not satisfied. One way to overcome the data limitation constraint is to leverage the learned representations of a model trained on a different yet related task, where the data are abundant. The pretrained model from the base task serves as the starting point for the desired task, where the data are scarce. The transfer of knowledge between both tasks is carried out when the training stage of the pretrained model is resumed using the limited data of the target task. This may involve optimizing all or only a subset of the parameters of the model on the target domain data. In the following subsections we will provide an overview of adaptation methods involving the transfer of knowledge based on the availability of labels for the same or a different target task, and how often the model needs to be retrained to perform optimally in the

target domain.

Knowledge transfer

Kumar et al. (2018) explored two scenarios to transfer knowledge from a resource-rich task to a resource limited task. In the first scenario, they investigated cross-dataset (Audioset to ESC-50) adaptation for a sound classification task with labels common to both datasets. In the second scenario, they leveraged the representations learned from Audioset and transferred them to the ASC task. In a later work, Kumar et al. (2021) investigated the transfer learning capacity of the representations of a model pretrained on Audioset to analyze the elements contributing the most to several downstream classification tasks including environmental, urban and domestics sounds. Kumar et al. (2021) identified meaningful structures that reflect proximity relationships among sound classes, for instance among music genres, instruments, moods and human actions.

Similarly, Diment and Virtanen (2017) used transfer learning for tagging audio events whose nature differ from that of the sounds of the base task, which makes the transfer of knowledge challenging. They showed that even though the considered base and target sound classes do not share many acoustic properties (baby cries vs. glass breaks), the recognition rate in the target task is remarkable, outperforming a baseline system trained on the target (although limited) domain data with supervised learning. In a similar work, Arora and Haeb-Umbach (2017) hypothesized that good generalization in the target domain is due to the fact that sound events share the same inventory of acoustic units which differ only in their temporal order.

Active learning

Active learning (Settles, 2009) engages humans in the learning process of a machine learning model in a setting where there is a large amount of unlabeled data in the target domain and a budget has been set to carry out the labeling. Therefore, it establishes a semi-supervised framework that allows a human being to label samples prioritized by the model so that a high level of accuracy is achieved by querying the fewest number of samples. This process is carried out dynamically and incrementally through several iterations. After each iteration, the decision boundaries are updated based on the newly acquired labeled data. The process ends when the budget for the labeling of the data is reached.

How to best select samples for human labeling is the core question of active learning. The selection criteria evaluates the informativeness of unlabeled samples for which many *querying* strategies have been proposed. Two general (and popular) strategies are *uncertainty sampling* (Lewis and Gale, 1994) and diversity sampling (Brinker, 2003). The former strategy queries samples based on a confidence score. Monarch (2021) distinguishes four confidence measures:

- **Least confidence**: ranks samples based on the difference between the most confident output scores and 100% confidence.
- **Margin of confidence**: ranks samples based on the difference between the top two most confidence output scores.
- **Ratio of confidence**: ranks samples based on the ratio between the top two most confident output scores.
- **Entropy-based**: ranks samples based on the difference between all output scores based on entropy.

On the other hand, diversity sampling helps to identify which information the model is missing. This is particularly interesting since it allows to correct unwanted bias towards certain classes. Monarch (2021) distinguishes as well four different sampling approaches:

- **Outlier sampling**: selects samples that are confusing to the model based on low activation values in the output logits or hidden layers.
- **Cluster-based sampling**: selects samples based on unsupervised clustering algorithms.
- **Representative sampling**: selects the samples that are the most closely resemble the target domain compared to the current training data.
- **Real-world diversity sampling**: uses sampling strategies that ensure fairness to and reduce real-world bias.

Active learning is a framework of practical interest for the analysis of sound scenes and events in real environments, since data collection is generally simple and the hassle of labeling can be greatly reduced by annotating only a small number of sound instances that ensure optimal performance of various applications. In the following, we highlight several active learning works relying on confidence-based active learning.

Han et al. (2016) proposed a strategy that combines confidence-based sampling and semi-supervised learning for audio classification including human activity, animal, vehicle and environmental sounds. In this semi-supervised active learning approach, sample selection is partially performed by human annotators and by the model itself. First, the model assigns a confidence score to the samples based on the probability estimates outputted by the classifier. The samples with the lowest scores are presented to the annotator for labeling, while the samples with the highest scores preserve the label given by the model. These data are then used to retrain the model. Such a strategy outperformed reference models based on confidence and semi-supervised learning, showing the efficiency of the combined method. A similar active learning strategy was proposed by Wang et al. (2019b) in the context of the Sounds of New York City (SONYC) project for urban noise monitoring (Bello et al., 2019). Their strategy combines low-confidence sampling with occasional sampling of high-confidence samples to allow annotators to validate or correct classification errors resulting from mislabeling. Their proposed approach improved classification over least-confidence sampling and semi-supervised active learning.

In a different application domain, Qian et al. (2017a,b) proposed a sparse-instance-based and a least-confidence-based active learning strategies for bird sound recognition. The first strategy is particularly interesting for unbalanced scenarios as it clarifies the confusion between class boundaries in few iterations. The second strategy focuses on tracking the few (and sparse) instances of the minority class since they are less likely to get annotations. The latter strategy is similar to the active learning strategy proposed by Kim and Pardo (2018) to reduce the manual effort of labeling long audio recordings (e.g., over 20 hours) containing sparsely distributed sound events. They created a graphical interface where the user selects a portion of the recording that contains a target sound, and then the model segments automatically the regions of the recording that are likely to contain it. The user verifies all segments proposed by the system and provides feedback that helps the model propose a more accurate segmentation to the user, allowing a faster labeling. This mechanism performs sampling selection using a relevance score based on the nearest neighbor method involving positive and negative segments.

Regarding diversity sampling-based active learning methods, we highlight the *MAL*: Medoid-based Active Learning framework proposed by Shuyang et al. (2017). This framework for audio

classification is based on K-medoids clustering, where medoids are first presented to the annotator for labeling, and then such labels are propagated to cluster members. MAL requires 60% less samples to be annotated to achieve the same accuracy than reference methods based on random sampling, uncertainty and semi-supervised learning, which makes it ideal for scenarios with small labeling budgets. Shuyang et al. (2018) extended the MAL framework with a second stage performing sample selection based on the mismatched outputs from the base classifier and a nearest-neighbor classifier. This optimized version of MAL requires labeling only 20% of the training data to achieve the same performance as requesting all data to be labeled. In another recent work, Shuyang et al. (2020) explored the benefits of the MAL framework on the SED task. Since the MAL framework operates over short sound segments and the recordings for SED are typically longer than recordings used in audio classification, the acoustic scene was first segmented by a novel segmentation method based on change point detection. Experiments on a dataset with rare sound events showed that the performance of the system by labeling only 2% of the data is equivalent to using all the labeled data.

Continuous learning

In real world scenarios, in addition to making sound systems more robust against effects of distribution shifts, it is often desired to provide them with some flexibility so that they can learn information from new sound classes once deployed. In an online scenario, Ntalampiras (2016) designed a passive adaptation method for audio classification that is able to continuously detect class specific distribution shifts as well as new appearing sounds by analyzing the statistical properties of an evolving target data distribution. The model is then adapted to the new conditions with incremental usage of the incoming data. This capability is referred to as *continuous learning* and is suitable for scalable audio analysis applications. The difficulty associated with algorithms that learn information incrementally lies in how to acquire the knowledge of new classes without forgetting the already learned information of the known classes. This phenomena known as *catastrophic forgetting* leads to performance degradation of the source task. Continuous learning approaches for audio analysis tasks focus on detecting novel sound classes in scenarios where the data for retraining grows over time and is prohibitive to maintain (Wang et al., 2019c; Koh et al., 2020), and oftentimes without need of human supervision (Bayram and İnce, 2021; Wang et al., 2022b). Recently, Wang et al. (2021a) explored continuous learning in combination with few-shot learning. They jointly trained a classifier with a set of base sound classes together with an attention module capable of identifying novel classes. Excepting the continuous learning approach proposed by Ntalampiras (2016), no other works consider a distribution shift in the target domain where novel sounds arise.

Few shot adaptation

Few-shot learning aims to learn novel classes and recognize them with high accuracy with only a few samples collected in the target domain. It is suitable for real-world applications where capturing and label a large amount of data is unfeasible. For audio analysis tasks, few-shot strategies have been proposed for audio classification (Cantarini et al., 2022; Naranjo-Alcazar et al., 2020) and sound event detection (Wang et al., 2021b). Metric learning approaches (Cantarini et al., 2022; Pons et al., 2019) and the use of autoenconders (Naranjo-Alcazar et al., 2020) successfully integrate new sound classes to be recognized in data scarce scenarios. An in-depth study conducted by Wang et al. (2021b) investigated the impact of unique properties of sounds in real scenarios such as polyphony, sound overlap and varying signal-to-noise ratios in few-shot

learning scenarios. Their study lead to design guidelines for few-shot learning which depend on the expected application domain. However, most existing works do not consider discrepancies between the source and target domain, a scenario commonly referred to in the literature as *cross-domain few shot learning* (Guo et al., 2020).

2.3.3 Feature transformation

Transformations on the available source and target domain data aim to reduce certain variabilities that induce a domain mismatch due to covariate shift. Such transformations can be applied directly to the input features of both domains or of a single domain. They can also be applied to the learned feature representations of a model. In this section, we review moment normalization and moment matching techniques used in the adaptation of audio analysis systems, notably in the context of acoustic scene classification.

Normalization strategies

It is well known that the normalization of features provides stability and reduces training time of machine learning models (LeCun et al., 2012). In the adaptation of audio analysis systems, normalization plays a much more important role, since the practice of z-scoring the features is not limited to the input data, but can also take place in the intermediate layers of a deep neural network with the *BatchNorm* algorithm (Ioffe and Szegedy, 2015). As long as the statistical moments are computed across all input features or over a mini-batch (as in BatchNorm), the moments reveal biases in the data, which once removed, increase the generalization ability of a model (Dubey et al., 2019).

One such bias corresponds to the effects of the microphone's frequency response, and has been demonstrated than compensating microphone mismatch with feature space transformations lead to improved performances for applications such as speech recognition (Vincent et al., 2017) and speaker verification (Gravier et al., 2000).

In the context of the acoustic scene classification task, Kośmider (2020) proposed a straightforward approach that generalizes to data recorded with mismatched recording devices. Their method named *spectrum correction*, compensates for the different frequency responses of several target recording devices by subtracting the mean of the log-Mel spectrograms for each target device and frequency separately. However, the efficacy of spectrum correction has only been explored in supervised learning settings where labeled source and target domain data are available for training.

Correction of the target domain data

As long as the source of mismatch can be corrected at the feature level, transformations applied only to the target domain data can make such data more similar to the source domain data.

In particular scenarios where the source of mismatch can be tackled at the feature level, transformations applied to the target domain data may be sufficient to make its distribution similar to that of the source domain. This can be thought of as a *correction* step, which takes places prior to the analysis task and that helps bridge a distribution gap. The advantage of this procedure is that an adaptation phase involving the retraining of a model is often not necessary. Since the transformations are based solely on the statistics of the data from both domains, such a correction constitutes an unsupervised domain adaptation method, since no annotations of the target domain data are required. In the context of acoustic scene classification, Mezza et al. (2021) proposed the *band-wise statistics matching* method to match the frequency-wise first- and

second-order statistics of the log-Mel spectrograms of the target domain data to the statistics of the source domain data to improve robustness to mismatched recording devices. The simplicity of this method yields similar performance than more complex methods involving retraining such as adversarial learning (Drossos et al., 2019). In a related work, Mezza et al. (2020) projected the spectro-temporal features from both the source and the target domains onto the principal subspace spanned by the eigenvectors of the covariance matrix of the source domain data. They trained a classifier on the projected features of the source domain data and test it on the target domain data projected onto the source domain subspace. Such a method outperformed previous unsupervised domain adaptation approaches. A recent in-depth study by Mezza et al. (2022) analyzed the combination of standardization and projection-based transformations, showing that robust adaptation can be achieved by providing as few as 90 seconds of target domain data and that the method can serve as a feature extraction stage for low complexity models.

2.3.4 Deep domain adaptation

Domain mismatch can also be reduced within deep neural networks during model training through an objective function that explicitly minimizes the gap between the source and the target data distributions. The use of an adaptation-based cost function leads to the construction of domain-agnostic representations, in which a common subspace is shared by both the source and target domain. According to Farahani et al. (2021), deep domain adaptation techniques are categorized into *discrepancy-based*, *reconstruction-based* and *adversarial-based adaptation*. We particularly distinguish between discrepancy-based and adversarial-based domain adaptation methods for audio analysis tasks. We however highlight the reconstruction-based approach of Mun and Shon (2019) that performs channel conversion using a factorized hierarchical variational autoencoder for device adaptation in acoustic scene classification.

Discrepancy-based

Discrepancy-based methods measure the distribution difference between the source and target domains and minimize it through an objective function that tends to align the marginal distributions across domains. Among the most common discrepancy criteria are: *maximum mean discrepancy* (MMD) (Yan et al., 2017), *Kullback-Leibler divergence* (KL) (Kullback and Leibler, 1951), *correlation alignment* (CORAL) (Sun and Saenko, 2016) and the *Wasserstein distance* (Panaretos and Zemel). These techniques however have been little explored for the adaptation of audio analysis tasks. We highlight, the work of Zhao et al. (2021), in which the domain shift problem is addressed for the diagnosis of mechanical faults. The mismatch between the training and test recordings of faulty sounds is accounted for with the accumulation of MMD losses over different views of the source domain data. The MMD and other alignment losses including KL, mean square error (MSE) and L1 loss were investigated by Zhao et al. (2022) to reduce the mismatch due to using different recording devices in the ASC task. In the same context, Hu et al. (2020) proposed to use the KL divergence as the core of a relational teacher student learning framework with neural label embedding (NLE).

A different type of loss functions that account for data distribution shifts rely on metric learning. The fundamental idea is to define a metric under which the features from the same classes are clustered together and those from different classes are located far apart. In this regard, prototypical networks have been used to adapt an ASC system to acoustic scenes recorded in different cities (Singh et al., 2021), and the mismatch between synthetic and real data for training sound event detection systems has been tackled by means of an inter-frame distance loss (Huang

et al., 2020b), which minimizes the distance of feature embeddings of the same class while maximizing the distance between those of different classes.

Adversarial-based

The combination of adversarial learning with discriminative feature learning is a popular training strategy that can be used for domain adaptation, particularly in the absence of target domain labels. The main idea of this strategy, called *adversarial domain adaptation* (Ganin et al., 2016), is to learn representations through an adversarial interplay between the feature extractor and a domain classifier or *discriminator*. Specifically, the discriminator is optimized by minimizing the classification error of differentiating between the source from the target features, while the feature extractor is optimized to produce representations that are indistinguishable by the discriminator. As training progresses, adaptation takes place with the emergence of domain-invariant features.

Several audio analysis task rely on the use of adversarial domain adaptation to improve generalization performance under mismatched conditions. Gharib et al. (2018) introduced the first unsupervised domain adaptation approach for the ASC task reduce the domain discrepancy due to using different recording devices. Their proposed adversarial training strategy was improved by Drossos et al. (2019), whose method based on the Wasserstein generative adversarial networks (WGAN) formulation learned intermediate feature embeddings that follow a more similar distribution for both the source and target domain data. In a recent work Kacprzak and Kowalczyk (2021) analyzed the impact of a wider set of generative adversarial networks (GAN) including Cycle GAN trained with recordings captured simultaneously by the source and target devices, i.e., *paired* data as well as with unpaired data.

For the SED task and particularly for the detection of domestic sounds, adversarial learning has been proposed to reduce the mismatch between synthetic and real data. For instance, Yang et al. (2020) designed a two-stage procedure adaptation strategy that first trained a SED system with a customary semi-supervised training scheme and secondly, performs adaptation by feeding synthetic and real data into a domain discriminator optimized through adversarial learning. A similar method proposed by Zheng et al. (2021) trains mutually a semi-supervised SED model with a specific SED model with a gradient reversal layer for domain adaptation. Adversarial learning has been also employed in other SED tasks such as the detection of bird sounds where robust representations are needed to distinguish background variations (Tang et al., 2021), as well as in the detection of abnormal sounds of industrial machines whose operating sounds tend to vary across time thus producing a shift between the source and target domains (Gu et al., 2021).

2.3.5 Discrete optimal transport

The Optimal Transport (OT) problem was first introduced by Monge (1781) as a way to minimize the effort spent in moving a pile of sand to a target location. A couple of hundred years later, Kantorovich re-framed OT as a linear programming problem. Its application as a domain adaptation method is motivated by its remarkable ability to find correspondences between data samples by exploiting the underlying geometry of the space in which the data are defined (Courty et al., 2017b). OT for domain adaptation, as its name indicates, follows a least effort principle that aligns arbitrary probability distributions in an optimal way with respect to a transportation cost or metric. The properties of such an efficient strategy based on a least costly alignment of the data have been explored in computer vision (Courty et al., 2014; Ge et al., 2021) and natural language processing (Alvarez-Melis et al., 2019; Chen et al., 2019b).

The original formulation of the OT problem introduced by Monge minimizes the cost of transporting a distribution μ^{s} to another distribution μ^{t} using a map T as

$$\min_{T} \int_{\mathcal{X}} c(x, T(x)) d_{\mu^{\mathrm{s}}}(x), \quad T\#\mu^{\mathrm{s}} = \mu^{\mathrm{t}}, \tag{2.27}$$

where μ^{s} and μ^{t} are measures in $\mathbb{R}^d$, the map $T : \mathbb{R}^d \to \mathbb{R}^d$ is a measurable function, $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ a cost function, and $\#$ is the push forward operator.

When the distributions of the source and target domain can be accessed through the available training data, the corresponding empirical distributions μ_{s} and μ_{t} are defined as

$$\mu_{\mathrm{s}} = \sum_{i=1}^{n_{\mathrm{s}}} a_i \delta_{x_i^{\mathrm{s}}}, \quad \mu_{\mathrm{t}} = \sum_{j=1}^{n_{\mathrm{t}}} b_j \delta_{x_j^{\mathrm{t}}}, \tag{2.28}$$

where δ_{x_i} is the Dirac function at position $x_i \in \mathbb{R}^d$ and $\mathcal{X} = \{x_i \in \mathbb{R}^n\}_{i=1}^{n_{\mathbf{s}}}$ and $\mathcal{Y} = \{x_j \in \mathbb{R}^n\}_{j=1}^{n_{\mathbf{t}}}$ are point clouds in the feature space. The coefficients a_i and b_i are uniform weights in the probability simplex, i.e., $\sum_{i=1}^{n_{\mathrm{s}}} a_i = \sum_{i=1}^{n_{\mathrm{t}}} b_i = 1$. The space of joint probability distributions $\Gamma(\mu_{\mathrm{s}}, \mu_{\mathrm{t}})$ with marginals μ_{s} and μ_{t} can be written as

$$\Gamma(\mu_{\mathrm{s}}, \mu_{\mathrm{t}}) = \{\gamma \in \mathbb{R}_{+}^{n_{\mathrm{s}} \times n_{\mathrm{t}}} | \gamma \mathbf{1}_{n_{\mathrm{t}}} = \mu_{\mathrm{s}}, \gamma^{\top} \mathbf{1}_{n_{\mathrm{s}}} = \mu_{\mathrm{t}}\}, \tag{2.29}$$

where $\mathbf{1}_n$ is an n-dimensional vector of ones. The Monge-Kantorovich formulation of the optimal transport problem seeks the optimal transportation coupling γ^* that minimizes the cost of moving x_i to x_j w.r.t. a metric $c(x_i, x_j)$ as

$$\gamma^* = \arg \min_{\gamma \in \Gamma(\mu_{\mathrm{s}}, \mu_{\mathrm{t}})} \langle \gamma, \mathbf{C} \rangle_F, \tag{2.30}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product, $\mathbf{C} \leq 0$ is the cost matrix $\in \mathbb{R}^{n_{\mathrm{s}} \times n_{\mathrm{t}}}$, whose term C_{ij} represents the pairwise cost $c(x_i, x_j)$. Whenever c is a norm, e.g. Euclidean, the minimum of this optimization problem is known as the *Wasserstein distance*

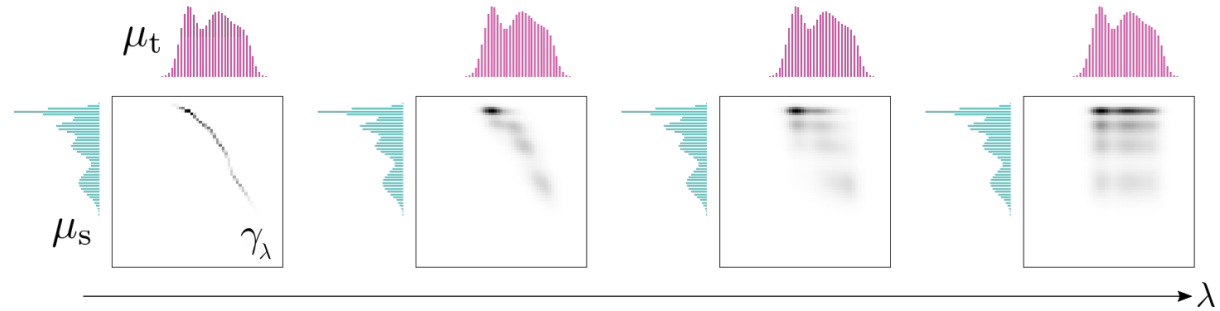
$$W(\mu_{\mathrm{s}}, \mu_{\mathrm{t}}) = \min_{\gamma \in \Gamma(\mu_{\mathrm{s}}, \mu_{\mathrm{t}})} \langle \gamma, \mathbf{C} \rangle_F, \tag{2.31}$$

which can be used to measure the distance between distributions. Equation (2.30) can be solved with combinatorial techniques, e.g., the Hungarian algorithm (Kuhn, 1955). In practice, a fast computation for the transport plan γ includes a regularization term, such as square L2 or entropic regularization (Cuturi, 2013).

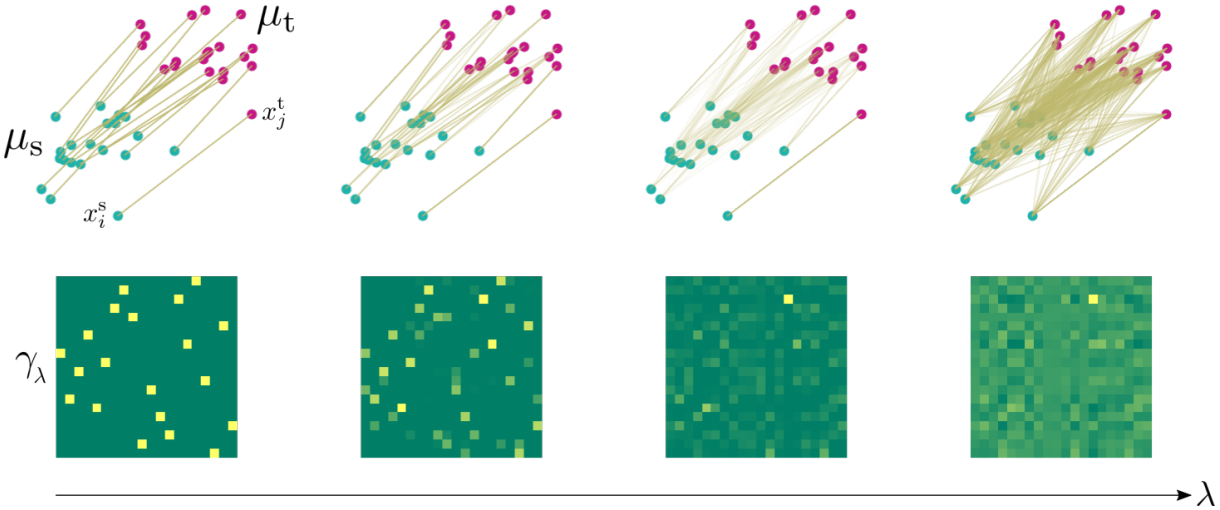
The regularized optimal transport based on the computation of the entropy of γ leads to the following optimization problem:

$$\gamma_\lambda^* = \arg \min_{\gamma \in \Gamma(\mu_{\mathrm{s}}, \mu_{\mathrm{t}})} \langle \gamma, \mathbf{C} \rangle - \frac{1}{\lambda} \Omega(\gamma), \tag{2.32}$$

where $\Omega(\lambda) = -\sum_{ij} \gamma_{ij} \log \gamma_{ij}$. The hyper-parameter λ controls the sparsity of the transportation coupling and can be optimized stochastically or with the Sinkhorn-Knopp algorithm (Knight, 2008) and variants. A high regularization value leads to a dense coupling between the source and target distributions while a low value leads to a less uniform coupling. This is illustrated in Figure 2.10.



(a) Optimal transport between histograms.



(b) Optimal transport between cloud points.

Figure 2.10: Marginal source and target distributions and regularized optimal transport plan. As the entropic regularization parameter λ increases, the transport plan γ_λ is more diffuse (top) and becomes less and less sparse (down).

2.3.6 Joint distribution optimal transport

The optimal transport problem through equations (2.30) and (2.32) does not however take into account the class information within the source and target domains which might be useful to correct a change in the conditional distributions. In fact, there is no clear reason to think that the conditional distributions remain unchanged, particularly when multiple real-world factors cause a distribution drift. Therefore, a more general approach is to align both the marginal feature and class-conditional distributions by minimizing the discrepancy between them.

In the *joint distribution optimal transport* (JDOT) framework, Courty et al. (2017a) proposed to handle a change in the marginal and conditional distributions with a joint cost measure combining both the distances between the samples and a loss function $\mathcal{L}$ measuring the dissimilarity between the source and target domain class labels as

$$d(x_i^s, y_i^s; x_j^t, y_j^t) = \alpha c(x_i^s, x_j^t) + \beta L(y_i^s, y_j^t), \tag{2.33}$$

where scalar parameters α and β weight the contribution of the cost metric $c(\cdot, \cdot)$ in the feature space and the cost $\mathcal{L}$ in the label space, respectively. In the unsupervised domain adaptation

problem, the target domain labels y_i^{t} are not available, and as such they are replaced by a proxy version $f(x_j^{\mathrm{t}})$, which depends on a classier $f : \mathcal{X} \to \mathcal{Y}$. The coupling that accounts for the joint features and label distributions reads

$$\min_{f,\gamma\in\Gamma(\mu_{\mathrm{s}},\mu_{\mathrm{t}})} \langle \gamma, \mathbf{D}_f \rangle_F, \tag{2.34}$$

where $\mathbf{D}_f$ depends on f and comprises all the pairwise costs $d(x_i^{\mathrm{s}}, y_i^{\mathrm{s}}; x_j^{\mathrm{t}}, f(x_j^{\mathrm{t}})) = \alpha c(x_i^{\mathrm{s}}, x_j^{\mathrm{t}}) + \beta L(y_i^{\mathrm{s}}, f(x_j^{\mathrm{t}}))$.

2.3.7 Learning with JDOT in deep embedding spaces

The JDOT framework described in the above section encompasses neural networks. Courty et al. (2017a) proved that equation (2.34) minimizes a learning bound on the target error of f, and that domain adaptation is possible as long as we can predict accurately in both the source and the target domains. Damodaran et al. (2018) discusses two issues of the JDOT framework. Firstly, solving for the optimal γ is computationally expensive, making it unsuitable for large datasets; secondly, the cost $c(x_i^{\mathrm{s}}, x_j^{\mathrm{t}})$ is computed in the feature space which may not be informative of the difference between samples. In order to address the latter issue, Damodaran et al. (2018) proposes to minimize the Wasserstein distance between the joint distributions within a neural network model, i.e., using the empirical distributions of the learned embeddings (extracted from a neural layer) instead of the input feature distributions. On the other hand, the intractability of γ is alleviated by an approximation method based on stochastic optimization using minibatches. The resulting improvements lead to the DeepJDOT framework, which consists of two parts: an embedding function $g : \mathcal{X} \to \mathcal{Z}$ by which the input features are mapped to a high-dimensional latent space $\mathcal{Z}$, and the classifier $f : \mathcal{Z} \to \mathcal{Y}$. The optimization of such model components, as well as the computation of the transport plan γ are perform through a two step process:

In the first step, the optimal coupling matrix γ is computed with fixed model parameters f and g:

$$\min_{\gamma\in\Gamma(\mu_{\mathrm{s}},\mu_{\mathrm{t}})} \sum_{i,j}^{m} \gamma_{ij} \left(\alpha c(g(x_i^{\mathrm{s}}), g(x_j)^{\mathrm{t}})) + \beta \mathcal{L}\left(y_i^{\mathrm{s}}, f(g(x_j^{\mathrm{t}}))\right)\right). \tag{2.35}$$

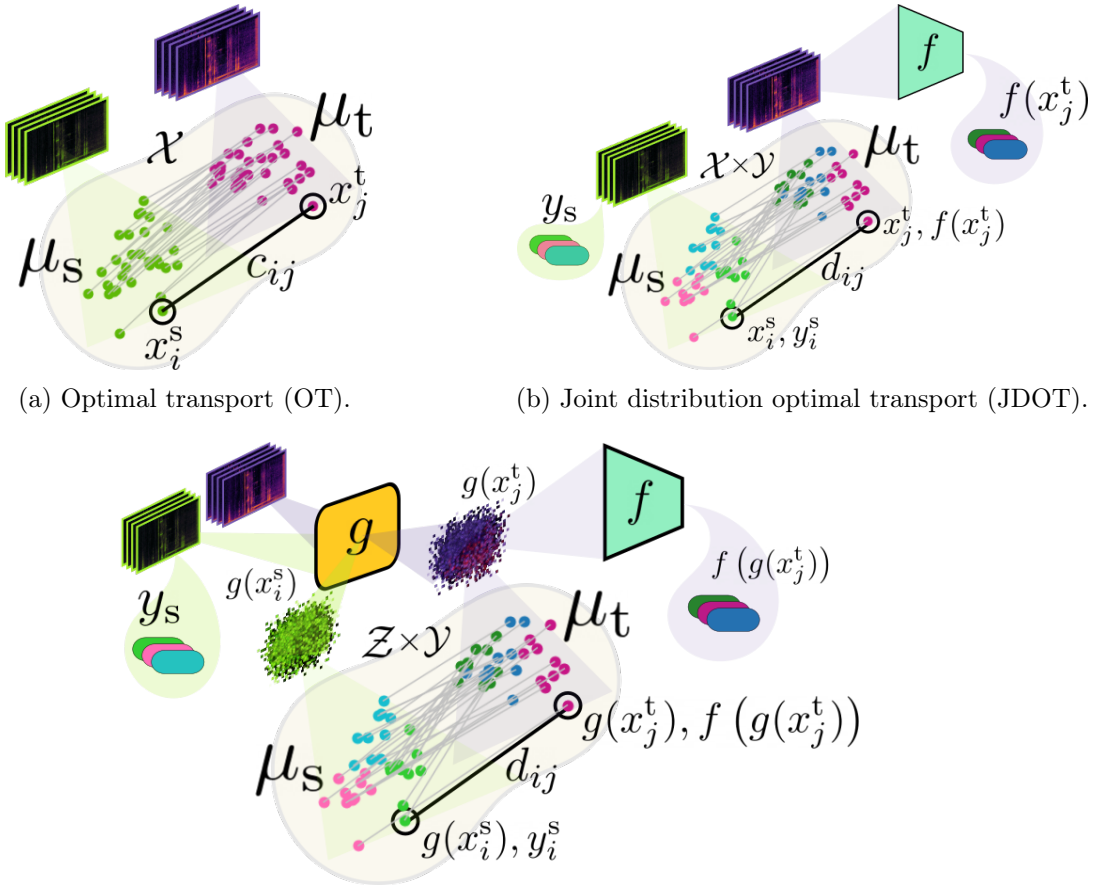
In the second step, with fixed γ, the embedding function g and classifier f are updated as

$$\min_{g,f} \quad \mathcal{L}_s\left(y_i^{\mathrm{s}}, f\left(g(x_j^{\mathrm{s}})\right)\right) + \sum_{i,j}^{m} \gamma_{ij} \left(\alpha c(g(x_i^{\mathrm{s}}), g(x_j^{\mathrm{t}})) + \beta \mathcal{L}\left(y_i^{\mathrm{s}}, f(g(x_j^{\mathrm{t}}))\right)\right), \tag{2.36}$$

where $\mathcal{L}_s$ correspond to the classification cost on the source domain to avoid catastrophic forgetting of the source domain data. Figure 2.11 shows the different optimal transport formulation presented in this section.

2.3.8 Summary

This chapter introduced the basic building blocks of modern computational audio analysis systems such as features, existing data and labeling, model architectures, and evaluation metrics. We focus on applications such as acoustic scene classification and sound event detection. We then formalized the source separation problem and discussed the works revolving around its application to separate sounds other than speech and music, such as ambient sounds. The last section of this chapter brought into discussion the distribution shift problem, distinguishing various scenarios in which such a problem arises. We described current domain adaptation methods applied



(a) Optimal transport (OT).

(b) Joint distribution optimal transport (JDOT).

(c) Deep joint distribution optimal transport (DeepJDOT).

Figure 2.11: Optimal transport formulations. (a) Marginal distribution alignment in the input space $\mathcal{X}$. (b) Marginal and class conditional distribution alignment in the product space $\mathcal{X} \times \mathcal{Y}$. (c) Joint distribution alignment of embedded features and labels in the product space $\mathcal{Z} \times \mathcal{Y}$.

to audio analysis tasks which aim to increase robustness to diverse factors impeding optimal performance in real-world scenarios. Since one goal of this thesis is to integrate optimal transport to sound event detection to bridge the gap between synthetic and real data, we emphasized on various formulations of optimal transport.

3

Ambient sound source separation

The recent advances in audio source separation make this field a potential support for audio analysis tasks that require robust identification of sound events in the presence of other interfering sounds. In practical scenarios, audio surveillance systems could benefit from increased performance if, prior to the recognition of audio events, source separation were used to isolate the sound events of interest. In the same way, hearing aid devices could improve sound selectivity with prior extraction or enhancement of the target sounds. Separation of ambient sounds is nevertheless challenging, particularly since research efforts have mainly concentrated in isolating speech and music signals.

Drawing from state-of-the-art speaker separation models based on deep neural networks, Kavalerov et al. (2019) explored the separation of arbitrary sounds: a task they referred to as *universal sound separation*, making a big leap towards separating sounds other than speech or music. Recent works such as *Sound Selector* (Ochiai et al., 2020) and (Delcroix et al., 2022) also tackle this task, but from a different perspective, namely the explicit extraction of sounds based on conditional clues. Inspired from the target speaker extraction task, these methods rely on an auxiliary network which is provided with the class label of the target sound or an audio sample that belongs to that class or reassembles the sound to be extracted. However, in real-world settings, it is impractical for audio analysis systems to isolate a wide variety of sound classes. Instead, they should be constrained to analyze a narrow pool of sounds that make sense for the intended application. Furthermore, audio analysis systems do not operate on short audio segments but on continuous recordings that involve multiple short audio events co-occurring with long audio events of a somewhat stationary structure. We thus study the separation of all short sound events (considered as a whole) from a stationary background, a task that we refer to as *foreground-background ambient sound source separation*. This scenario has not been widely explored, nonetheless, some works provide insights about the potential applications of this task. For instance, Serizel et al. (2020) pointed out that when the loudness of foreground sounds is higher than that of background sounds, or vice versa, sound detection models tend to be biased toward the loudest class. Separation of short vs. long events may help prevent biasing the separation models towards either class in varying SNR conditions. In another relevant work, carried out concurrently to this work, Varma et al. (2021) investigated the effectiveness of classifying urban soundscapes by using the foreground or background components alone. Their work provides a definition similar to ours for foreground and background sounds (sparse vs. stationary sounds). Separation is however carried out in the input feature space through robust principal component analysis (RPCA) and not modeled with deep neural networks.

In this chapter we address three goals. First, investigate whether a deep learning-based

separation system is able to differentiate the rapidly varying spectro-temporal features of short audio events against the more slowly varying features of background sounds encountered in real-life environments. Second, we explore the application of per-channel energy normalization (PCEN) to the input time-frequency representations to improve the separation performance. Lastly, we investigate if a source separation model of this nature is able to generalize to unseen foreground and/or background sounds. This is an important capability in real-world scenarios, where the application domain may have new and unseen sound classes that were not present in the development of the model, thus achieving robustness to label shift. We propose the use of a deep neural network to estimate the soft time-frequency masks associated with short and long events, and drawing from speaker extraction models, we explore the integration of an auxiliary network to inform the separation model about the statistics of the background sounds.

The structure of this chapter is the following. We first present the formulation of the foreground-background ambient source separation task in Section 3.1. Next, the experimental setup, quantitative and qualitative results are presented respectively in Sections 3.2, 3.3 and 3.4. Section 3.5 finally concludes the chapter. The contents of this chapter have been published at EUSIPCO 2020 (Olvera et al., 2021b).

3.1 Model

3.1.1 Problem formulation

For single-channel audio recordings we define the problem of foreground-background separation as the task of recovering the foreground component of the mixture. In real-world scenarios, the foreground stream is usually composed of sounds with rapidly varying spectral characteristics occurring in presence of a background stream whose spectral characteristics vary slowly over time. Figure 3.1 shows an example of the decomposition of the time-frequency representation of a mixture of sounds into their foreground and background components. Thus, the input mixture $x(t)$ is modeled as

$$x(t) = f(t) + b(t), \tag{3.1}$$

where

$$f(t) = \sum_{i=1}^{I} f_i(t), \tag{3.2}$$

with I the total number of foreground events in the mixture and $\{f_i(t)\}_{i=1..I}$ each individual foreground event in the presence of a background sound $b(t)$ whose spectro-temporal characteristics are assumed to be stationary. Given the mixture $x(t)$, our goal is to estimate the streams containing foreground and background sound events $f(t)$ and $b(t)$, respectively.

In this chapter, we focus only on the recovery of the foreground stream $f(t)$ containing sound events from a single class, i.e., $I = 1$ that overlaps with a long duration background sound.

The foreground-background separation task is performed in the short-time Fourier transform (STFT) domain. The STFT coefficients $X(n, f)$ of the mixture $x(t)$ in time frame n and frequency bin f satisfy

$$X(n, f) = F(n, f) + B(n, f), \tag{3.3}$$

where $F(n, f)$ and $B(n, f)$ are the STFT coefficients of $f(t)$ and $b(t)$, respectively. We will use the notation $\mathbf{X}$, $\mathbf{F}$, $\mathbf{B}$ for the $N_x \times F$ matrices comprising all complex-valued coefficients $X(n, f)$, $F(n, f)$ and $B(n, f)$, with N_x being the number of time frames and F the number of frequency bins. The STFT-domain components $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{B}}$ estimated by the separation process

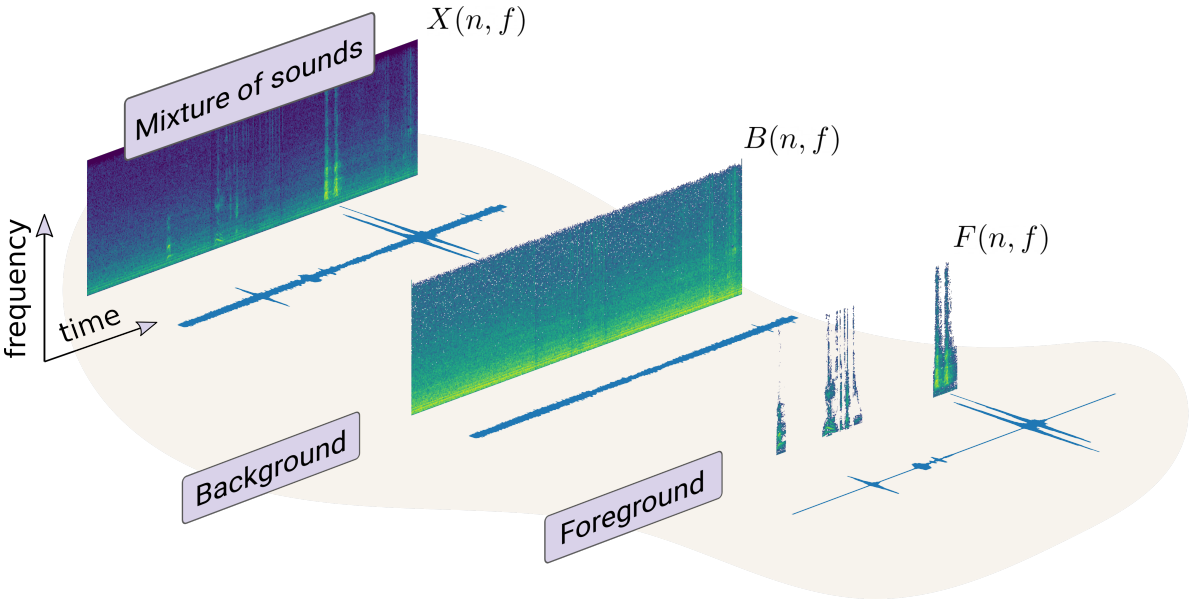


Figure 3.1: Mixture of foreground and background sound events.

are transformed back into time-domain signals $\widehat{\mathbf{f}}(t)$ and $\widehat{\mathbf{b}}(t)$ by computing the inverse STFT. Figure 3.2 shows a general overview of the separation process.

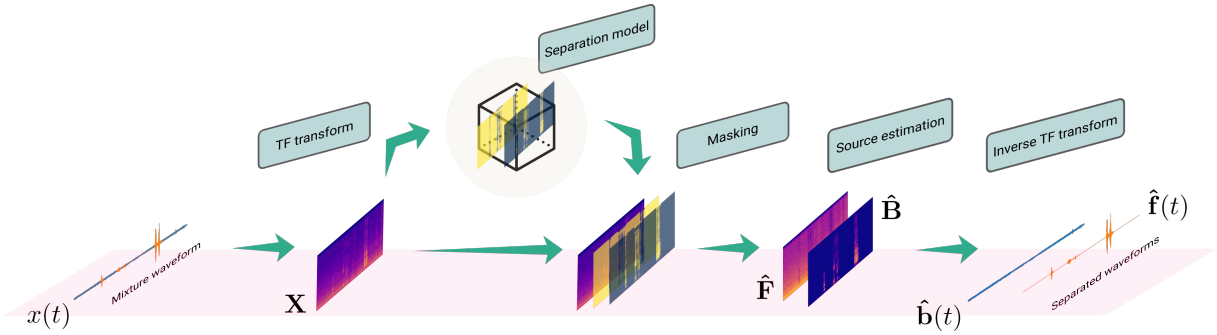


Figure 3.2: General overview of the separation process.

3.1.2 Separation framework

The proposed foreground-background separation framework is depicted in Figure 3.3. Following *SpeakerBeam*: the separation approach proposed by Žmolíková et al. (2019) for single-channel target speaker extraction, the separation framework relies on a main deep neural network and an optional auxiliary network to locate background and foreground components in the time-frequency plane.

Classically, the two networks operate on the nonlinear Mel frequency scale. In the following, we denote by F' the number of Mel bands and $|\mathbf{X}|^{\text{Mel}}$ the mixture Mel spectrogram which is obtained by multiplying the magnitude STFT $|\mathbf{X}|$ by the $F' \times F$ Mel filterbank matrix.

The auxiliary network can be active or inactive in the separation process. When it is *inactive*, the main network takes the mixture log-Mel spectrogram $\log |\mathbf{X}|^{\text{Mel}}$ as input and outputs a time-frequency mask $\mathbf{M}^{\text{Mel}}$, i.e., an $N_x \times F'$ matrix with real-valued entries in $[0, 1]$ that quantify

the proportion of foreground sound in each time-frequency bin. This matrix is then projected back to the STFT-domain to obtain a $N_x \times F$ mask $\mathbf{M}$[1]. Using the mask, the estimated STFT magnitudes of the foreground and background components are obtained as

$$|\widehat{\mathbf{F}}| = \mathbf{M} \odot |\mathbf{X}| \quad \text{and} \quad |\widehat{\mathbf{B}}| = (\mathbf{1} - \mathbf{M}) \odot |\mathbf{X}|, \tag{3.4}$$

where $\odot$ denotes an element-wise multiplication and $\mathbf{1}$ is a matrix of ones.

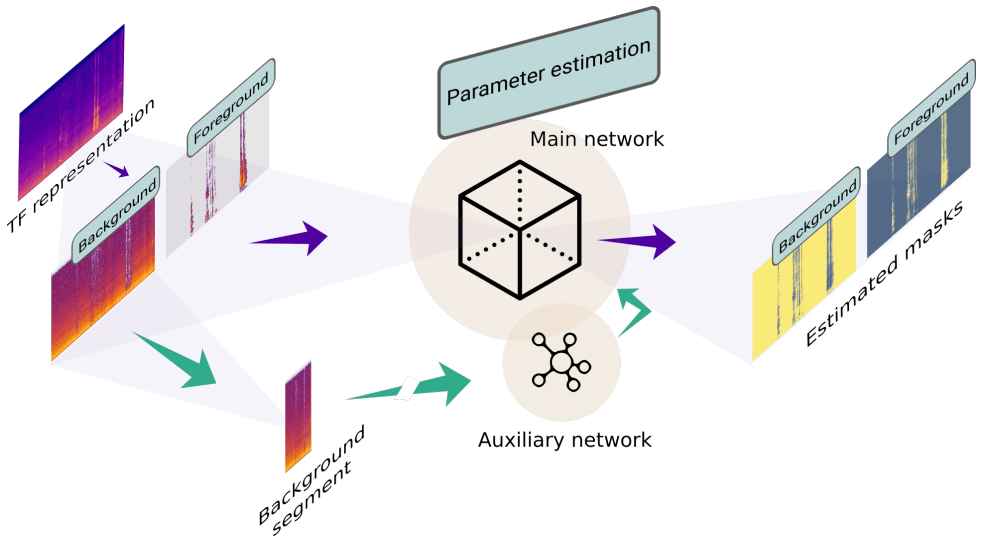


Figure 3.3: General diagram of the proposed foreground-background separation scheme.

When the auxiliary network is *active*, prior information about the background sound is assumed to be available in the form of an adaptation segment $a(t)$, which could be a short sample of the mixture to be processed or a preceding time interval which has been classified with high confidence as background-only, i.e., without any overlapping foreground sounds. The auxiliary network compresses the $N_a \times F'$ log-Mel spectrogram of the adaptation segment as $\log |\mathbf{A}|^{\text{Mel}}$, with N_a the corresponding number of frames, into a fixed-size vector $\boldsymbol{\lambda}$ which is used together with the mixture log-Mel spectrogram $\log |\mathbf{X}|^{\text{Mel}}$ by the main network to output a time-frequency mask $\mathbf{M}^{\text{Mel}}$. This mask is converted to the STFT domain and used to estimate the STFT magnitudes of the foreground and background components via (3.4). Such estimates are then combined with the mixture phase to obtain the time-domain signals $\widehat{\mathbf{f}}(t)$ and $\widehat{\mathbf{b}}(t)$ by means of inverse STFT.

3.2 Experimental setup

3.2.1 Dataset

In order to assess the generalization ability of the proposed framework in controlled conditions, we generated synthetic mixtures with foreground and background sound events by randomly sampling isolated sounds from the development and evaluation sound banks of the Domestic Environment Sound Event Detection (DESED) dataset (Turpault et al., 2019) and Audioset (Gemmeke et al., 2017). We cut the duration of all sound events to two seconds, and mixed them at different foreground-to-background signal-to-noise ratio (FBSNR), i.e., the ratio between the

[1]We adopted this approach since directly outputting an STFT-domain mask did not make a significant difference.

loudness of the foreground stream and the loudness of the background stream. We considered mixing foreground and background sounds at equal FBSNR, i.e., 0 dB, as well as at various FBSNRs randomly chosen between -3 and 3 dB. Since the duration of foreground sound events is typically shorter than that of the background, they are repeated as many times as needed to generate the foreground component of the mixture. Repeating short sounds withing the mixture also avoids biasing the separation models towards the background sound classes which, unlike foreground sound events, span the duration of the mixture.

The data are split in three subsets for training, validation and evaluation. The corresponding sound classes are listed in Table 3.1. We considered a total of 25 sound classes: 10 classes of foreground events and 15 classes of background sounds. The isolated signals used to generate each subset are disjoint.

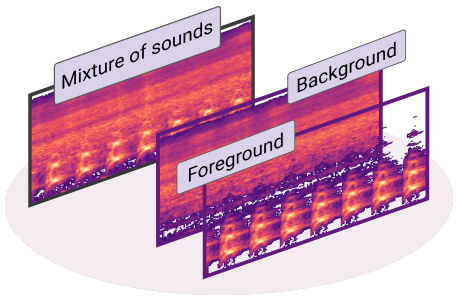
Table 3.1: Sound classes considered for the creation of the dataset.

<table>
<tr><th>Set</th><th>Foreground</th><th>Background</th></tr>
<tr><td>Training</td><td rowspan="3">DESED: <code>dog, speech, cat, dishes, alarm-bell-ringing</code></td><td rowspan="3">DESED: <code>vacuum cleaner, blender, frying, running water, electric shaver-toothbrush</code>, Audioset: <code>bathub, mechanical fan, microwave oven, hair dryer, drill</code></td></tr>
<tr><td>Validation</td></tr>
<tr><td>Evaluation C1</td></tr>
<tr><td>Evaluation C3</td><td>Audioset: <code>door, slam, squeak, coins, chopping food</code></td></tr>
<tr><td>Evaluation C2</td><td>DESED: <code>dog, speech, cat, dishes, alarm-bell-ringing</code></td><td rowspan="2">Audioset: <code>pink noise, white noise, noise, waterfall, vibration</code></td></tr>
<tr><td>Evaluation C4</td><td>Audioset: <code>door, slam, squeak, coins, chopping food</code></td></tr>
</table>

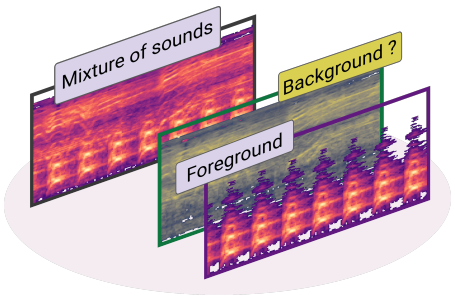
The training and validation sets comprise 5 foreground classes from DESED and 10 background classes from both Audioset and DESED datasets. The foreground and background*seen* classes relate to domestic environments. We further augment the training data using Scaper (Salamon et al., 2017).

On the other hand, the evaluation set comprises mixtures of all 15 seen classes and the remaining 10 *unseen* classes, including 5 foreground classes and 5 background classes from Audioset. Although the unseen background classes are unrelated to domestic environments, they share similar characteristics with domestic background sounds. The evaluation set is composed of four subsets: C1, C2, C3 and C4, which are illustrated and described in Figure 3.4. Each subset comprises mixtures of foreground and background sound events either seen or unseen during model training. This helps asses the generalization ability of the model in the foreground-background separation task.

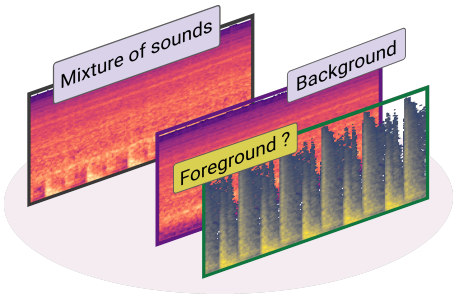
Overall, the training set consists of $15,000$ mixtures representing a total of 8.3 hours. Such mixtures were generated from 604 isolated foreground events and 786 background events. The validation set comprises $6,000$ mixtures spanning a total of 3.3 hours. The mixtures in this subset were generated from 131 foreground events and 175 background events. Lastly, each evaluation subset consists of $1,000$ mixtures representing 0.5 hour. Such mixtures were generated from 194 and 326 isolated foreground and background events, respectively. All data are single-channel signals sampled at 44.1 kHz.



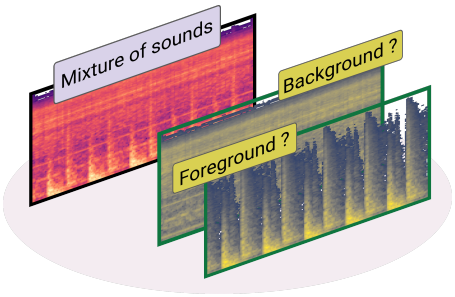
(a) C1: mixtures of seen foreground and background classes.



(b) C2: mixtures of seen foreground classes and unseen background classes.



(c) C3: mixtures of unseen foreground classes and seen background classes.



(d) C4: mixtures of fully unseen foreground and background classes.

Figure 3.4: Subsets of the evaluation dataset.

3.2.2 Input features

3.2.3 Mel-spectrograms

We compute the STFT with a window size of 2048 samples (46 ms) and a hop size of 512 samples (12 ms), leading to an overlap of 75% across frames. We then compute the log-Mel spectrogram using a Mel-filterbank of $F' = 128$ filters.

3.2.4 Per-channel energy normalization

In addition to computing log-Mel spectrograms as input features, we compute a time-frequency representation based on per-channel energy normalization (PCEN) (Lostanlen et al., 2018). In fields such as automatic speech recognition (Wang et al., 2017) and acoustic event detection (Lostanlen et al., 2018), PCEN spectrograms have recently shown to outperform Mel spectrograms as an acoustic frontend in classification scenarios involving frequency transposition of stationary and transient sound events. Since the spectro-temporal structure of the foreground and background events in the mixtures falls into a similar scenario, we explore the use of PCEN as a way to enhance the foreground sound events over the background. PCEN converts background noise into a spectro-temporal texture over which the foreground sounds stand out. Instead of logarithmic compression, PCEN uses a simple feed-forward automatic gain control to dynamically stabilize signal levels. It is defined as

$$\mathrm{PCEN}(n, f') = \left(\frac{|X|^{\mathrm{Mel}}(n, f')}{(\epsilon + \overline{|X|^{\mathrm{Mel}}}(n, f'))^{\alpha}} + \delta \right)^{r} - \delta^{r},$$

where f' denotes the Mel band index and $\overline{|X|^{\text{Mel}}}(n, f')$ is a smoothed version of $|X|^{\text{Mel}}(n, f')$, which is computed using a first-order infinite impulse response (IIR) filter as

$$\overline{|X|^{\text{Mel}}}(n, f') = (1 - s) \times \overline{|X|^{\text{Mel}}}(n - 1, f') + s \times |X|^{\text{Mel}}(n, f'),$$

with s the smoothing constant. This normalization scheme preserves frequency patterns which enhances the transients of foreground events, and filters out stationary background sounds, thus making it a suitable acoustic front-end for the separation task. We adopted the default PCEN parameters defined by Lostanlen et al. (2018), i.e., $s = 0.025$, $\epsilon = 10^{-6}, \alpha = 0.98, \delta = 2$, and $r = 0.5$, which are suited for indoor applications, e.g., domestic environments.

3.2.5 Main network

The neural network architecture is similar to those used by Žmolíková et al. (2019) for target speaker extraction and speaker separation, respectively. The main neural network, which is depicted in Figure 3.5, comprises a stack of 3 recurrent layers with bidirectional long short-term (BLSTM) memory cells, each followed by a dense layer with a hyperbolic tangent activation function. Each BLSTM layer has 300 units in each direction and 0.2 dropout is applied on the output, which is then projected by the dense layer to 256 dimensions. The last layer is a dense layer with an output dimension of 128 that matches the number of frequency bands, and a sigmoid nonlinearity that ensures the output values are between 0 and 1. The network is trained on 170-frame segments using the Adam optimizer with learning rate of 0.0001 for a total of 250 epochs.

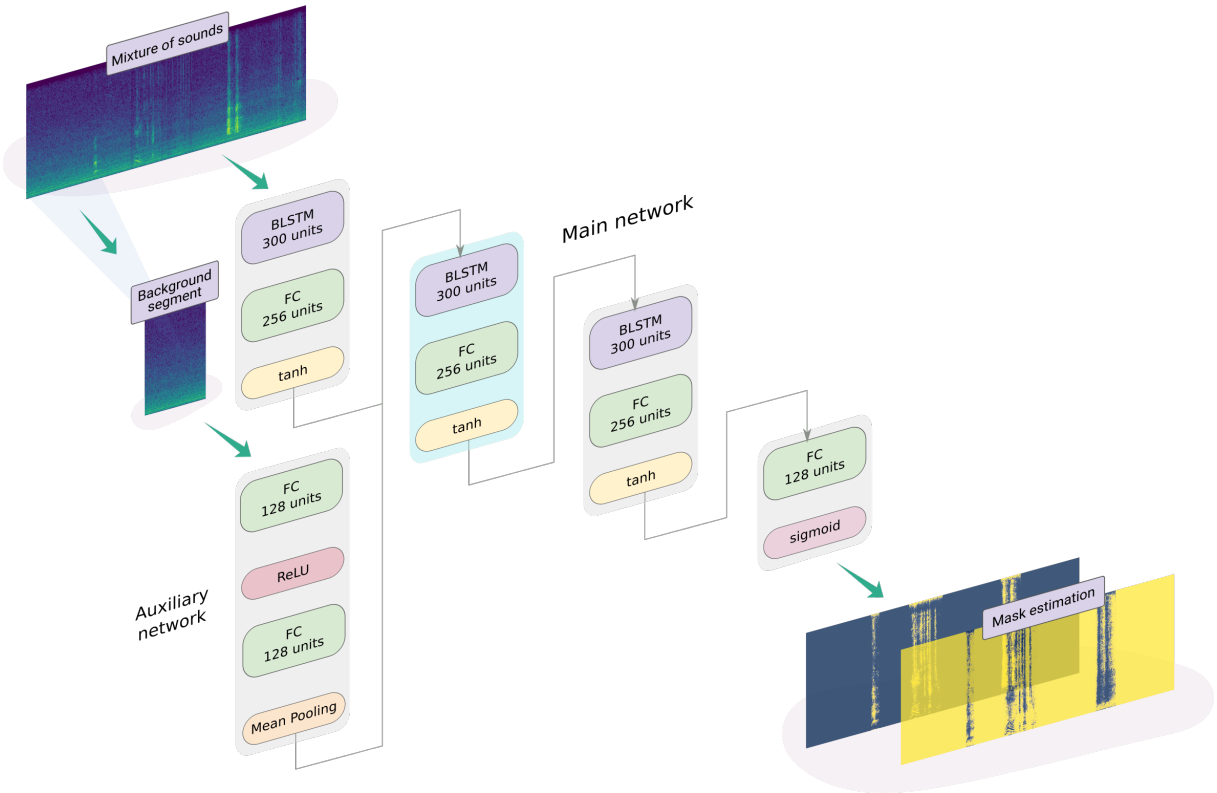


Figure 3.5: Block diagram of the proposed deep neural network architecture.

3.2.6 Auxiliary network

The auxiliary network is a sequence-summarizing network (Veselỳ et al., 2016) which compresses the adaptation segment into a fixed-sized vector

$$\boldsymbol{\lambda} = \frac{1}{N_a} \sum_{n=1}^{N_a} g(\log |\mathbf{a}|^{\text{Mel}}(n)), \tag{3.5}$$

where $|\mathbf{a}|^{\text{Mel}}(n)$ denotes the n-th frame of $|\mathbf{A}|^{\text{Mel}}$, and $g(\cdot)$ is the nonlinear transformation carried out by the network over each frame. This transformation comprises two dense layers with 128 neurons each. As shown in the lower block of Figure 3.5, following the first dense layer a rectified linear unit (ReLU) activation is used, while no activation function is used after the second dense layer.

The network outputs are averaged across all N_a frames to produce the adaptation vector $\boldsymbol{\lambda}$. The outputs of the first BLSTM layer of the main network are then multiplied element-wise by the entries of this vector. Both the auxiliary network and the the main network are trained jointly.

3.2.7 Training objective

Whether the auxiliary network is active or not, the model is trained by minimizing the squared Frobenius norm between the Mel spectrogram of the ground truth foreground event and the Mel spectrogram of the mixture multiplied element-wise by the estimated mask:

$$\min \|\mathbf{M}^{\text{Mel}} \odot |\mathbf{X}|^{\text{Mel}} - |\mathbf{F}|^{\text{Mel}}\|_F^2. \tag{3.6}$$

3.2.8 Model configurations

We built four models to investigate the separation task, which are depicted in Figure 3.6 and described in the following. The first model, referred to as *M1*, is trained with log-Mel spectrograms as inputs and does not rely on the auxiliary network for the mask estimation process. The second model, *M1+*, also takes log-Mel spectrograms as inputs but, unlike *M1*, it makes use of the auxiliary network. Models *M2* and *M2+* are defined similarly to *M1* and *M1+*, respectively, but use PCEN spectrograms as inputs instead of log-Mel spectrograms.

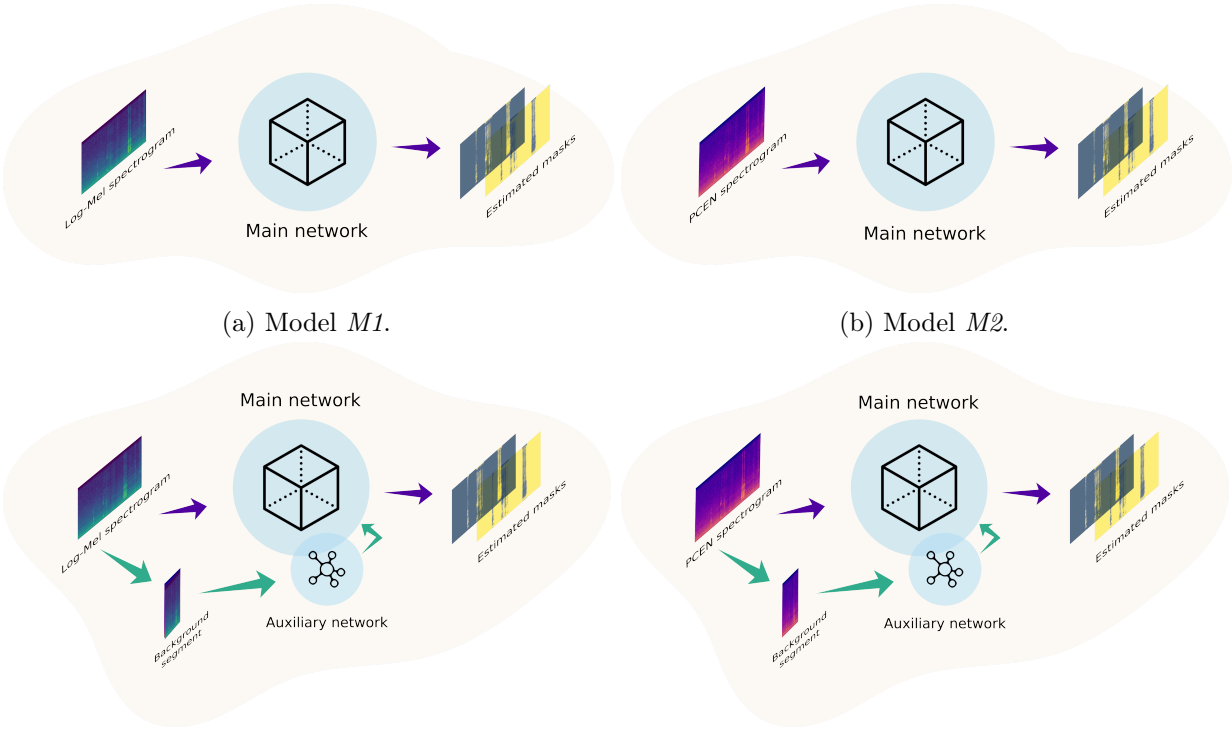
We used as a baseline an oracle separation method based on the ideal ratio mask (IRM). We compute the IRM as $\mathbf{M}^{\text{IRM}} = \mathbf{F}^{\text{Mel}}/(\mathbf{F}^{\text{Mel}} + \mathbf{B}^{\text{Mel}})$, where $\mathbf{F}^{\text{Mel}}$ and $\mathbf{B}^{\text{Mel}}$ are the Mel spectrograms of the ground truth foreground and background signals. The separation results provided by the IRM thus provide an upper bound on how well a masking-based separation model can perform on the foreground-background separation task.

3.3 Quantitative results

3.3.1 Metrics

In order to evaluate the separation quality achieved by the proposed models, we used as objective criteria the source separation metrics described in Section 2.2.4. The signal-to-distortion ratio (SDR), the source-to-interference ratio (SIR) and the signal-to-artifacts ratio (SAR). With the help of these metrics, we define

- SDR Imp. as the difference between the output SDR and the input SDR expressed in dB,



(a) Model *M1*.

(b) Model *M2*.

(c) Model *M1+*.

(d) Model *M2+*.

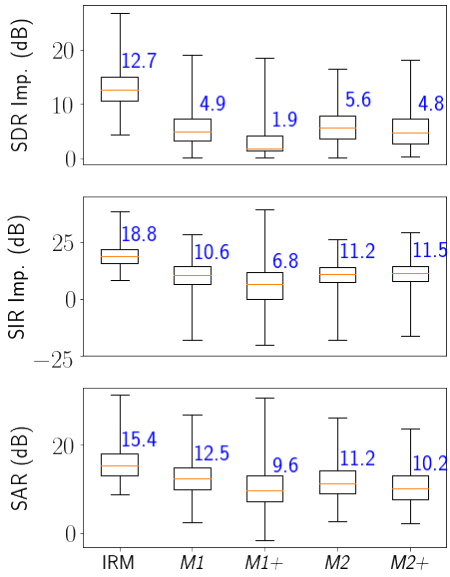
Figure 3.6: Model configurations.

- SIR Imp. as the difference between the output SIR and the input SIR expressed in dB.

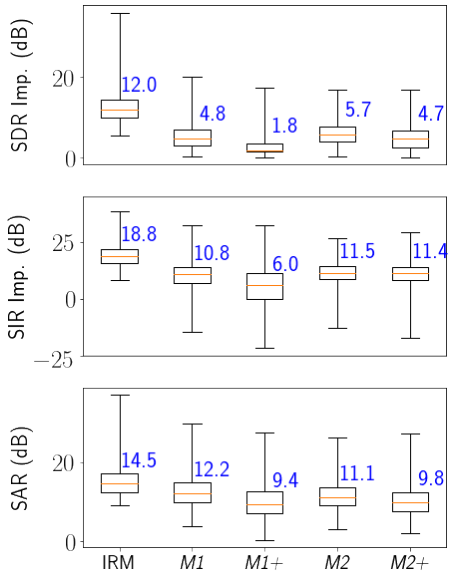
The box plots shown in Figure 3.7 reports the SDR and SIR improvements and the SAR achieved by the four model configurations and by the IRM on the four subsets of the evaluation set for mixtures mixed at 0 dB FBSNR, whereas Figure 3.8 shows the separation results for mixtures mixed at various FBSNRs between -3 and 3 dB.

3.3.2 Impact of input features

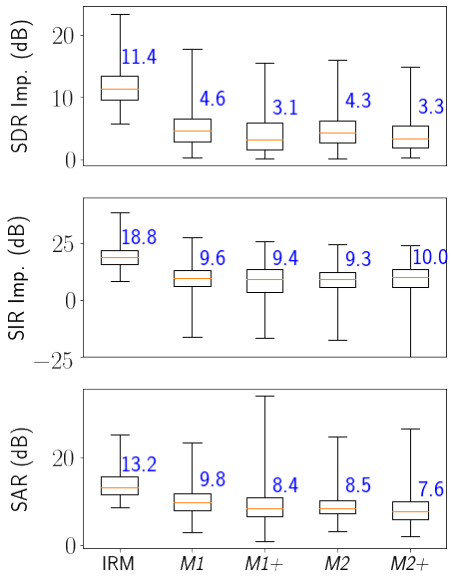
As shown in Figure 3.8, except for configuration *C3* (mixtures with unseen foreground and seen background) models using PCEN spectrograms (*M2* and *M2+*) achieved higher SDRs scores than the models (*M1* and *M1+*) using log-Mel spectrograms. Comparing the best two models for each input feature type (model *M1* vs. *M2*), model *M2* achieved a higher median SDR improvement than model *M1*, increasing the median SDR improvement by 0.9, 1.3 and 0.5 dB for *C1*, *C2* and *C4*, respectively. Model *M2* also achieved higher results than model *M1* in terms of SIR for the previous configurations, improving the SIR by 0.4, 0.1 and 0.5 dB for configurations *C1*, *C2* and *C3*, respectively. These results indicate that PCEN is beneficial for the separation task in terms of SDR and SIR. Even though the median SDR and SIR scores achieved by *M1* for configuration *C3* are higher than those achieved by model *M2*, the separation performance is comparable. Regarding SAR scores, the models using log-Mel spectrograms outperformed the models using PCEN spectrograms in all configurations where the auxiliary network is inactive, and configurations *C3* and *C4* when the auxiliary network is active. These results indicate that the models using PCEN spectrograms, despite estimating signals with less interference, produce more artifacts than the models using log-Mel spectrograms.



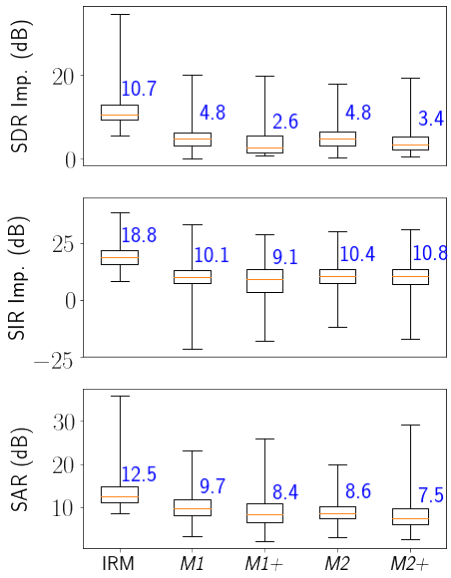
(a) C1: seen foreground and background classes.



(b) C2: seen foreground and unseen background classes.



(c) C3: unseen foreground and seen background classes.



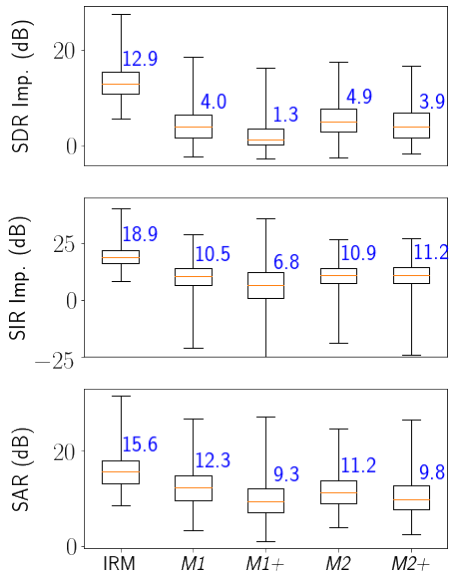
(d) C4: unseen foreground and background classes.

Figure 3.7: SDR and SIR improvements and SAR in decibels (dB) achieved by the four model configurations and by the IRM on the four subsets of the evaluation set. Mixtures mixed at 0 dB FBSNR.

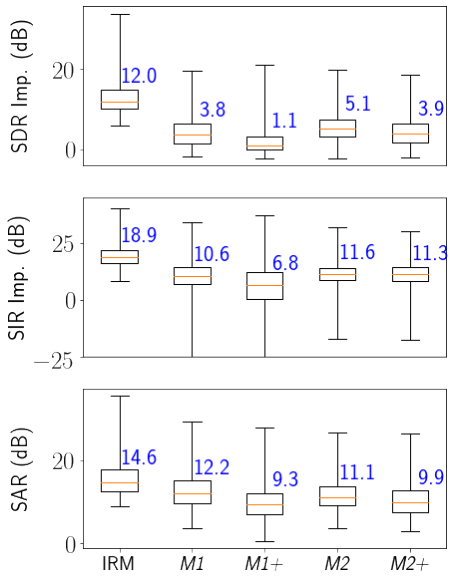
3.3.3 Impact of the auxiliary network

Regarding models using the auxiliary network in the mask estimation process, model *M1+* and *M2+* performed worse in terms of SDR than their counterparts *M1* and *M2*, respectively. In fact, model *M1+* trained on log-Mel spectrograms is the worst among all models.

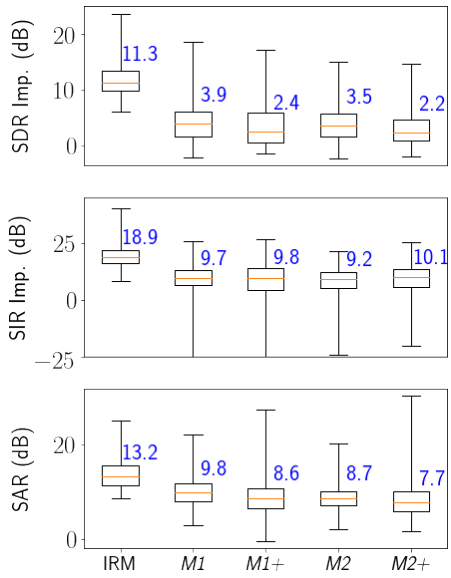
Model *M2+* showed however similar scores (conf. *C2* and *C4*) or higher scores (conf. *C1* and



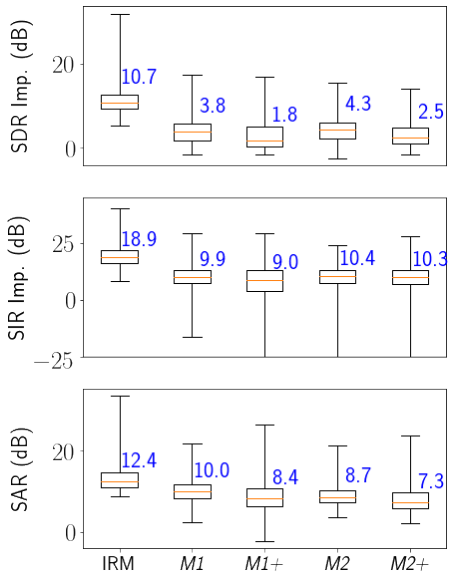
(a) C1: seen foreground and background classes.



(b) C2: seen foreground and unseen background classes.



(c) C3: unseen foreground and seen background classes.



(d) C4: unseen foreground and background classes.

Figure 3.8: SDR and SIR improvements and SAR in decibels (dB) achieved by the four model configurations and by the IRM on the four subsets of the evaluation set. Mixtures mixed at various FBSNRs randomly chosen between -3 and 3 dB.

C3) than *M2* in terms of SIR, but its lower SAR scores impacted negatively its SDR performance. Model *M2+* using PCEN spectrograms achieved comparable median SIR scores in configurations *C2* and *C4* than when not using the auxiliary network (model *M2*), or even higher scores in configurations *C1* and *C3*. However, model *M2+* achieved lower SAR scores than model *M2*, which impacted negatively its SDR performance.

We also observe that the performance of *M2+* is similar to that of model *M1* on subsets with seen foreground sound events (*C1* and *C2*), but its performance falls behind that of model *M1* when separating unseen foreground events (*C3* and *C4*).

These results indicate that use of the auxiliary network in the mask estimation process may not be necessary since the foreground and background streams are sources of substantially different nature (fast transient vs. stationary sounds). This stands in contrast with the target speaker extraction task in which the signals to be separated are from the same class, or with the target sound extraction task, in which the target sound may be any arbitrary sound class. In both tasks the use of an auxiliary network for the mask estimation process has shown to be beneficial.

3.3.4 Robustness to unseen events

We assess the generalization ability of the foreground-background separation models by comparing the SDR scores in Figure 3.8 for each evaluation configuration. The interest lies particularly in the performance of the models in configurations other than *C1*, where we desire to estimate the foreground or background streams of a mixture containing sounds that were not seen in the training stage, e.g., configurations *C2*, *C3*, *C4*. We observe that models not relying on the auxiliary network (*M1*, *M2*) are more robust to unseen foreground or background sounds than their counterparts (*M1+*, *M2+*). The SDR improvement achieved by models *M1* and *M2* is persistent over the four evaluation subsets, showing good generalization for any combination of either seen or unseen foreground-background sound classes. Model *M2* generalizes better than model *M1*, but it shows a slight degradation in performance when estimating unseen foreground classes (conf. *C3* and *C4*), in which scores can be compared to that of model *M1*.

3.4 Qualitative results

The series of Figures 3.9, 3.10, 3.11 and 3.12 shows the ground-truth (GT) and estimated foreground and background components from example mixtures in all evaluation configurations. The time-frequency representation of such signals serves to analyze the proposed models from a qualitative perspective.

3.4.1 Effects of the auxiliary network on overall signal distortion

Adding the adaptation segment to the main network through the auxiliary network causes the background sound to be more strongly reduced, leading to strong distortion of the foreground, notably for model *M1+* trained with log-Mel spectrograms. We observe that the separation quality of models when the auxiliary network is inactive (*M1* and *M2*) is closer to the ground-truth signals in all four configurations, which goes in line with the objective evaluation in terms of SDR. We invite the reader to visit the accompanying website[2] for more audio examples.

[2] http://molveraz.com/ambient-sound-scene-separation/

3.4.2 Effects of the auxiliary network on interferences and artifacts

In accordance with the objective evaluation, we observe qualitatively that the strong removal of the background sound, causes the foreground component to be estimated with less interferences from the background. However, this is counterproductive, as it leads to partial removal of the spectro-temporal structure of the foreground sound, particularly for higher frequencies, but also leads to artifacts. These effects are consistent in all four configurations.

3.4.3 Effects of PCEN

Differences in separation quality depending on the input features can be made by comparing models *M1* and *M2*, and models *M1+* and *M2+*. We observe through all evaluation configurations that the use of PCEN spectrograms leads to estimated foreground sounds with less distortions compared to log-Mel spectrograms. In fact, PCEN provides two advantages over log-Mel spectrograms in the foreground-background separation task. First, it enhances the spectral information of short sound events which improves the quality of the mask estimation. Second, the filtering operation attenuates the strong impact of the auxiliary network to remove the background sound. This explains why the SDR scores of model *M2+* did not suffer a degradation comparable to model *M1+* trained with log-Mel spectrograms.

3.5 Conclusion

We presented the foreground-background ambient sound separation task, in which short duration events occur on top of a background sound. We consider that this task is closer to the conditions faced by real-life ambient sound analysis systems than the mixtures of arbitrary sounds considered in some previous studies. We performed a series of experiments to assess the performance of a deep learning-based source separation model to discriminate the rapidly varying spectro-temporal features of foreground events against the slowly varying features of background sounds. We explored a scheme comprising a main mask estimation network and analyzed the effects of feeding an auxiliary network with an adaptation segment. Furthermore, we assessed the impact of log-Mel and PCEN spectrograms as input features. Under the proposed experimental protocol, we found that the use of an auxiliary network in the mask estimation process is detrimental, while PCEN is a more suitable acoustic front-end than log-Mel spectrograms in terms of SDR and SIR. The objective evaluation of the proposed models showed generalization capabilities over mixtures of seen and unseen foreground and/or background sound events. This capability is desirable for real-world applications that need to be robust to label shift problems.

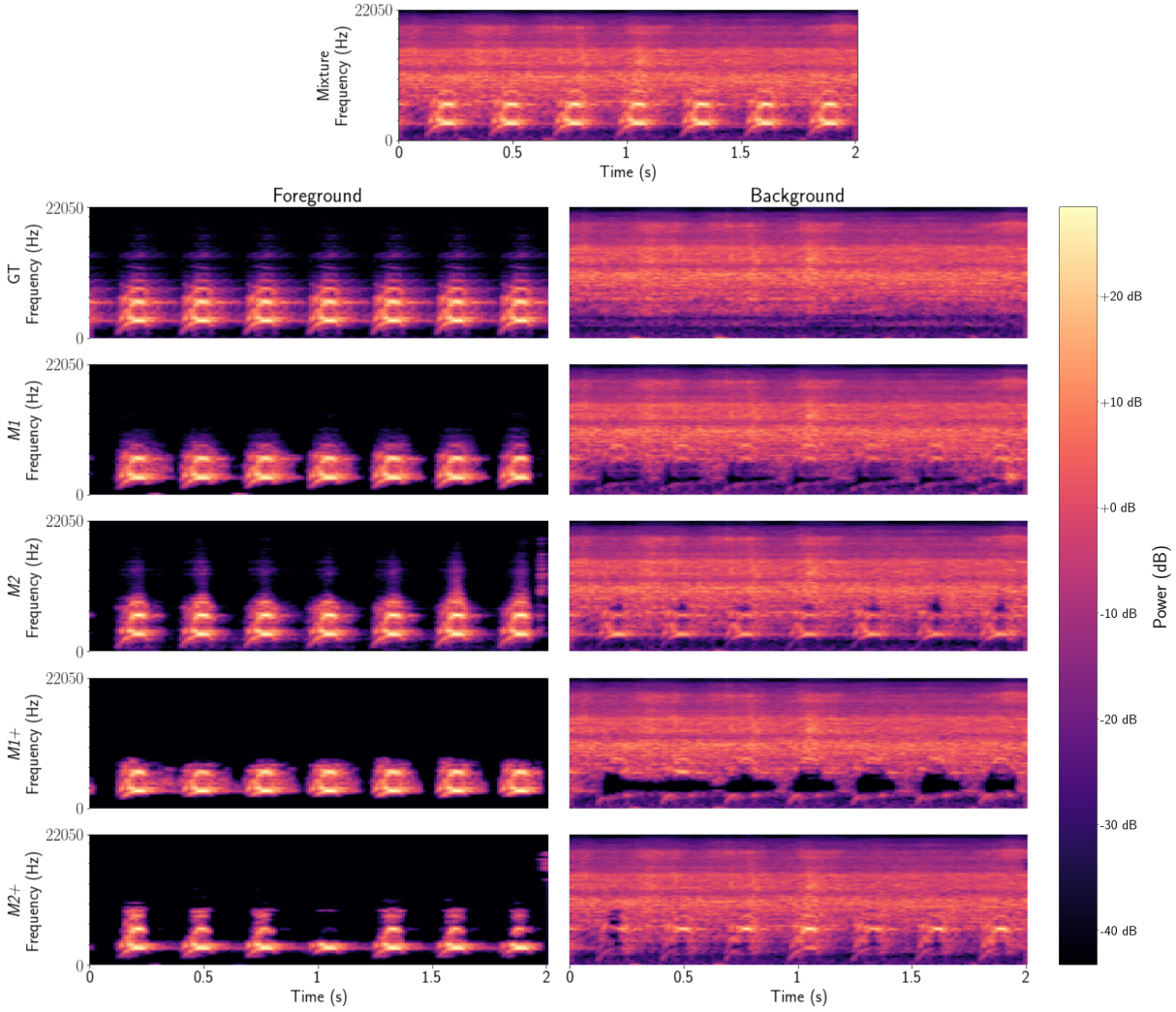


Figure 3.9: Example of foreground and background estimates from the four model configurations and their comparison with the ground-truth signals (GT) on the evaluation subset *C1*: mixtures of seen foreground and background classes. Foreground sound: *dog*; background sound: *electric shaver*.

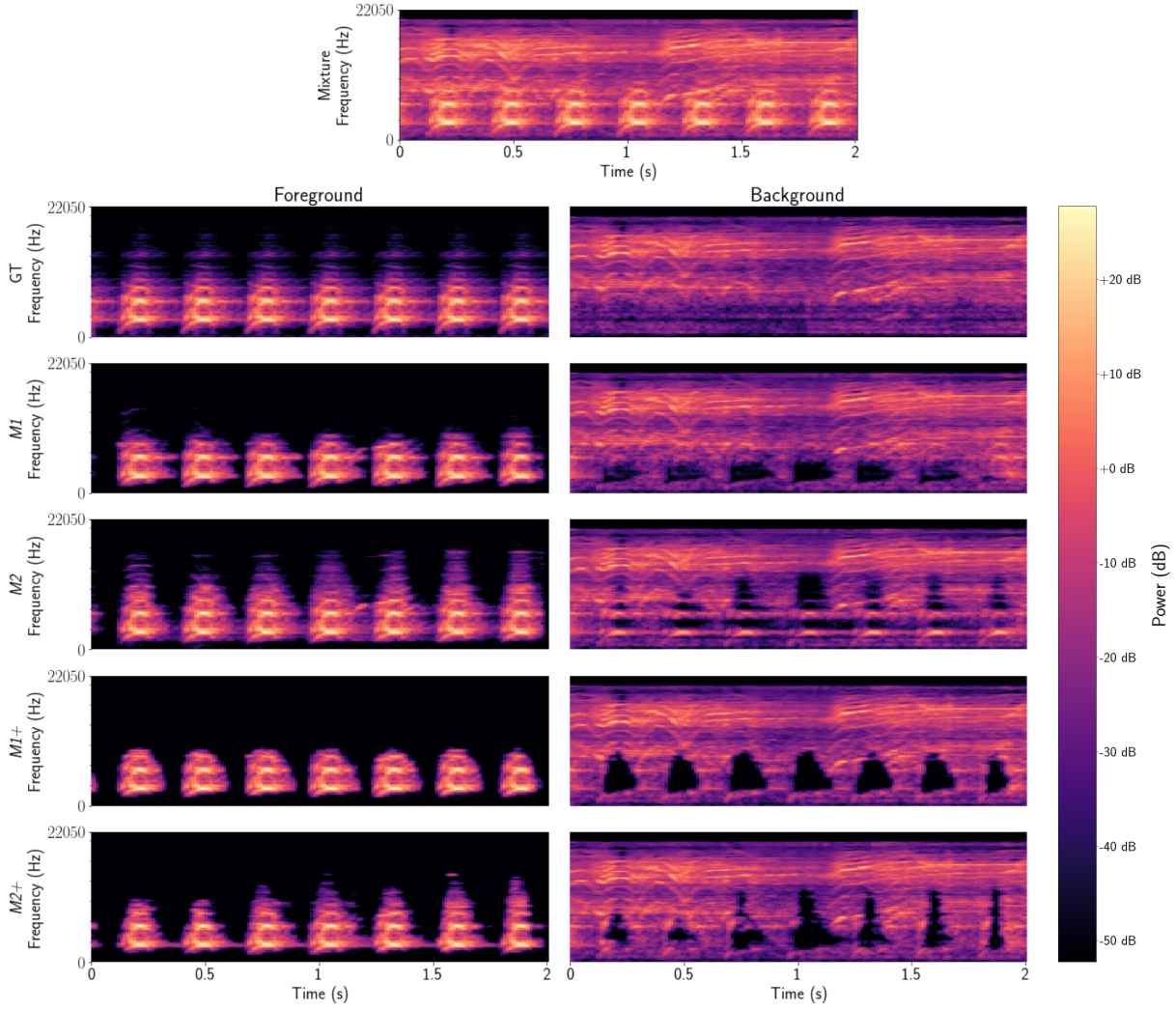


Figure 3.10: Example of foreground and background estimates from the four model configurations and their comparison with the ground-truth signals (GT) on the evaluation subset *C2*: mixtures of seen foreground classes and unseen background classes. Foreground sound: *dog*; background sound: *drill*.

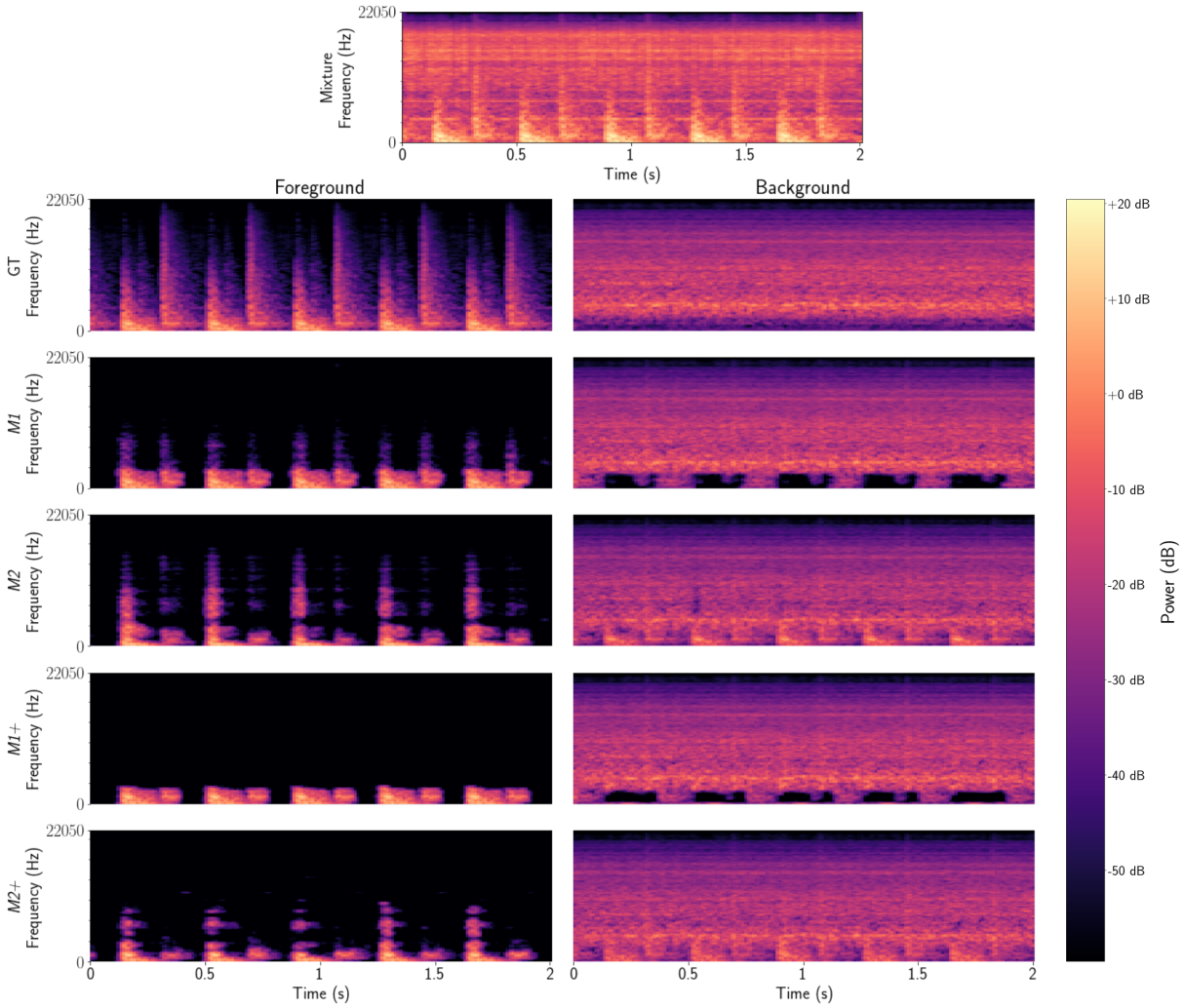


Figure 3.11: Example of foreground and background estimates from the four model configurations and their comparison with the ground-truth signals (GT) on the evaluation subset *C3*: mixtures of unseen foreground classes and seen background classes. Foreground sound: *door*; background sound: *vacuum cleaner*.

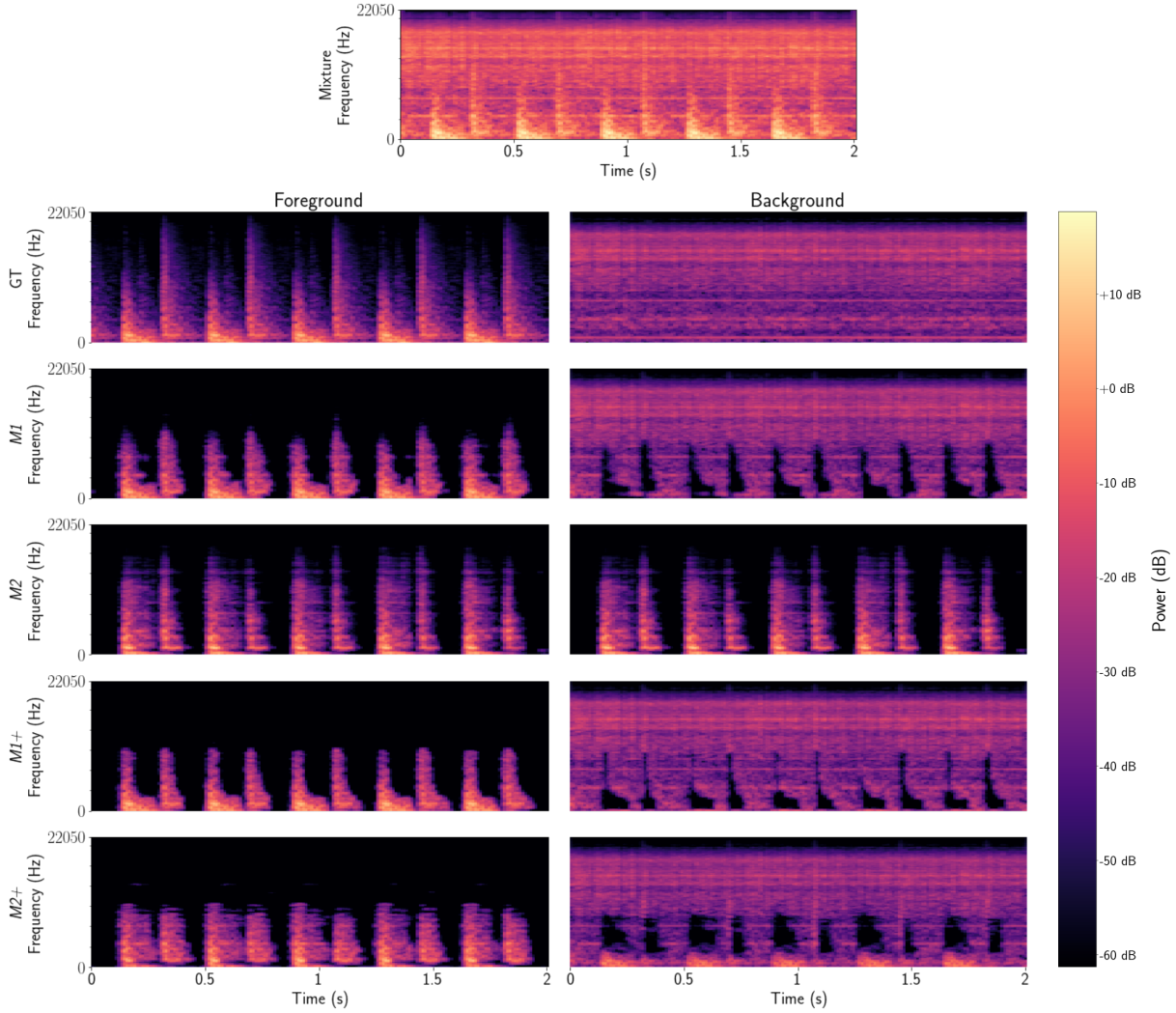


Figure 3.12: Example of foreground and background estimates from the four model configurations and their comparison with the ground-truth signals (GT) on the evaluation subset *C4*: mixtures of unseen foreground and background classes. Foreground sound: *door*; background sound: *vibration*.

4

Normalization strategies for unsupervised domain adaptation

In the previous chapter we introduced foreground-background ambient sound separation as a potential benefit to make audio classification tasks robust to overlapping sounds of different spectro-temporal characteristics. Moving forward, our focus will be on a methodology that enables audio analysis systems to achieve robustness in diverse acoustic conditions beyond those considered during development. In Section 2.3 we discussed that different sources of mismatches affect the performance of classification and detection systems. In this chapter, in the context of the Acoustic Scene Classification (ASC) task illustrated in Figure 4.1, we are particularly interested in the covariate shift that arises from the use of different recording devices at training and test time. Such mismatched recording conditions degrade the performance of ASC systems and prevent generalization to data captured with new recording devices. Figure 4.2 illustrates this scenario.

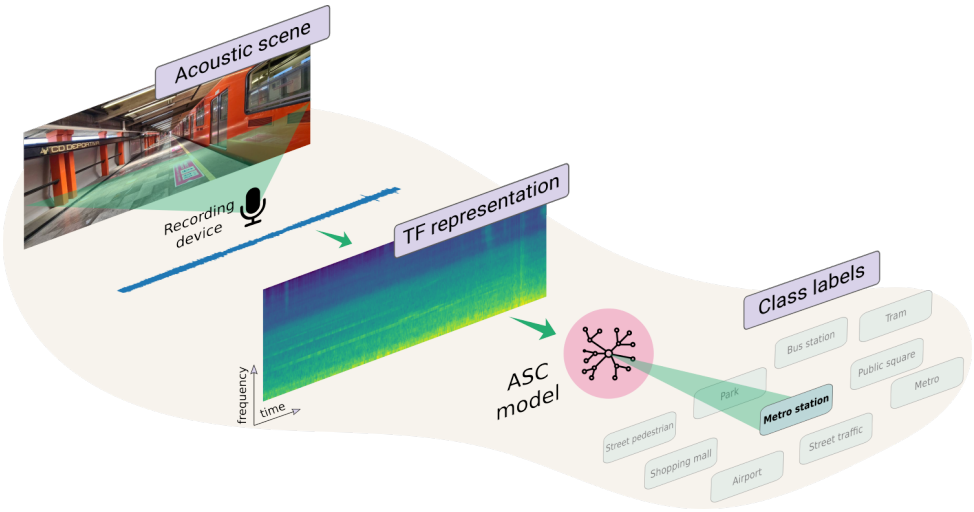
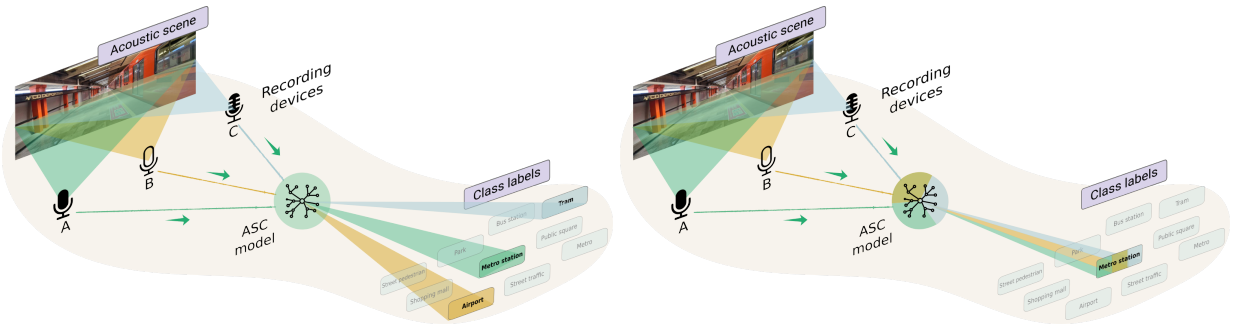


Figure 4.1: Acoustic scene classification.

This setup has been widely popularized by the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge Task 1 series (Mesaros et al., 2018a, 2019, 2018b), which provides a dataset with a large number of recordings from a source device but only a limited amount of data from target devices. The primary goal is to improve generalization to the underrepresented



(a) ASC in mismatched conditions. A model trained on device A only does not generalize to devices B and C.

(b) ASC in matched conditions. A model trained on device A and adapted on devices B and C generalizes to all devices.

Figure 4.2: Acoustic Scene Classification in mismatched vs. matched conditions.

devices. Supervised machine learning algorithms have been proposed to account for the data imbalance problem and are often combined with data augmentation, regularization and fine tuning approaches (Nguyen and Pernkopf, 2019; Hu et al., 2021a; Takeyama et al., 2021). As the dataset contains recordings captured simultaneously by the source and target devices, some methods leverage these parallel data to compensate for the effects of the frequency responses of the devices (Kośmider, 2020; Nguyen et al., 2020; Primus et al., 2019).

A few works have also used this dataset to investigate the more practical scenario where recordings of the source and target devices are available, but only the source recordings are labeled. To improve generalization on the target devices, they have applied unsupervised domain adaptation (UDA) methods (Ganin et al., 2016; Courty et al., 2017c). In particular, adversarial domain adaptation (Tzeng et al., 2017) has proven to be effective (Gharib et al., 2018; Drossos et al., 2019). Despite its effectiveness, it requires a large number of recordings from the target device to carry out the adaptation process. Furthermore, the adaptation process could implicitly benefit from the parallel data present in the dataset, or explicitly use these data to ease generalization (Mun and Shon, 2019; Yang et al., 2021). To overcome these limitations, band-wise statistical matching (BSWM) (Mezza et al., 2020, 2021) was introduced as a simple, linear feature transformation method that does not require any adaptation stage and that is equally effective as adversarial-based adaptation methods. The integration of this method with non-linear, learning-based UDA methods has not yet been explored. Given the effectiveness of stand-alone feature-based and adversarial-based UDA methods, we thoroughly analyze the impact of various feature normalization and moment matching strategies and their integration with adversarial domain adaptation strategies to further improve generalization in the target domain without assuming the availability of parallel source and target device data. Within the proposed adaptation strategies, we show that BWSM is a particular moment matching strategy and we further extend the adversarial domain adaptation framework to the conditional case. We show experimentally the individual scopes and limitations of such techniques on the development set of the DCASE Challenge 2018 Task 1B, that is the preferred choice for the development of unsupervised adaptation methods that address the domain shift due to mismatched recording devices.

The structure of this chapter is as follows. We first introduce in Section 4.1 the unsupervised domain adaptation strategies based on linear transformations related to moment normalization and moment matching, as well as the non-linear adversarial domain adaptation framework. We

present the experimental setup in Section 4.2. In particular, we discuss the problem of parallel recordings in unsupervised domain adaptation, as well as the evaluation configurations that integrate moment normalization and matching strategies to the adaptation framework. Section 4.3 shows the results and analysis and Section 4.4 concludes the chapter. The contents of this chapter have been published at ICASSP 2022 (Olvera et al., 2022).

4.1 Problem formulation

4.1.1 Linear distortion model

Let us denote by x_{nmk} the log-magnitude short-time Fourier transform (STFT) coefficients in time frame m and frequency bin k of the actual (undistorted) signal of some acoustic scene indexed by n, and by x_{dnmk} the log-magnitude STFT coefficients of the same signal captured by some recording device d with time-invariant linear magnitude frequency response h_{dk}. Assuming that there is no other distortion besides the effects of the frequency response of the recording device, the captured acoustic scene can be expressed as $x_{dnmk} = x_{nmk} + \log h_{dk}$. The undistorted signal x_{nmk} can therefore be recovered as

$$\hat{x}_{dnmk} = x_{dnmk} - \log h_{dk}. \tag{4.1}$$

We keep index d in the notation $\hat{x}_{dnmk}$ to emphasize the fact that this estimate was obtained from device d. In practice h_{dk} is unknown, hence $\hat{x}_{dnmk}$ must be obtained from x_{dnmk} only.

4.1.2 Moment normalization

Moment normalization consists of applying a domain-dependent linear transform to the data in the source and target domains so that their first- and second-order moments are fixed. Mean normalization is the simplest form of this technique. It consists of subtracting the average log-magnitude spectrum

$$\mu_{dk} = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} x_{dnmk} \tag{4.2}$$

of the data recorded by each device d from the original data:

$$\hat{x}^{\mathrm{MN}}_{dnmk} = x_{dnmk} - \mu_{dk}. \tag{4.3}$$

This is equivalent to (4.1), where μ_{dk} can be seen as an estimate of $\log h_{dk}$ up to an arbitrary frequency response which is common to all devices. Kośmider (2020) used mean normalization as the basis of their method *Spectrum correction* to compensate for the different microphone frequency responses prior to training a supervised model that generalizes to several recording devices.

Mean and variance normalization (a.k.a. standardization) additionally requires the computation of the sample standard deviation

$$\sigma_{dk} = \sqrt{\frac{1}{NM-1} \sum_{n=1}^{N} \sum_{m=1}^{M} (x_{dnmk} - \mu_{dk})^2} \tag{4.4}$$

of the data for each device d. Normalization is achieved by

$$\hat{x}^{\mathrm{MVN}}_{dnmk} = \frac{x_{dnmk} - \mu_{dk}}{\sigma_{dk}}. \tag{4.5}$$

Although not addressing any specific physical distortion, this is a common preprocessing step in machine learning (Friedman, 2017; Bishop, 1995).

4.1.3 Moment matching

Unlike moment normalization, moment matching does not eliminate the distortion due to the recording device d from the source data. Instead, it transforms the target data recorded by some other device d' so that its first- and second-order moments match those of the source data.

Matching the means of x_{dnmk} and $x_{d'nmk}$ can be achieved by removing from $x_{d'nmk}$ the distortion due to device d' as in (4.3), and then introducing the distortion due to device d:

$$\hat{x}^{\mathrm{MM}}_{d'nmk} = x_{d'nmk} - \mu_{d'k} + \mu_{dk}. \tag{4.6}$$

After moment matching, x_{dnmk} and $\hat{x}^{\mathrm{MM}}_{d'nmk}$ share the same sample mean, which according to (4.1) should suffice to transfer the distortion from one device to another. In practice, further robustness may be obtained by also matching the variances. This is achieved by first normalizing the mean and variance of $x_{d'nmk}$ as in (4.5), and then scaling and shifting the standardized frequency bands by σ_{dk} and μ_{dk}:

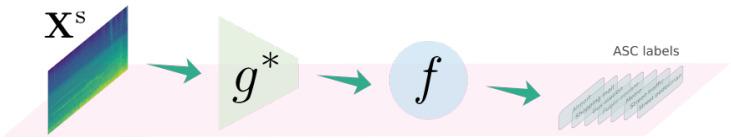
$$\hat{x}^{\mathrm{MVM}}_{d'nmk} = \frac{(x_{d'nmk} - \mu_{d'k})\sigma_{dk}}{\sigma_{d'k}} + \mu_{dk}. \tag{4.7}$$

The series of works carried out by Mezza et al. (2020, 2021, 2022) relies on this approach to correct the covariate shift due to mismatched recording devices. Similarly to Spectrum Correction, their method BWSM (Mezza et al., 2021) is devised to perform feature transformations on the data, but unlike Spectrum Correction, such transformations are intended to be applied to data from unseen devices before they are fed to a pre-trained classifier, thus avoiding learning phase.

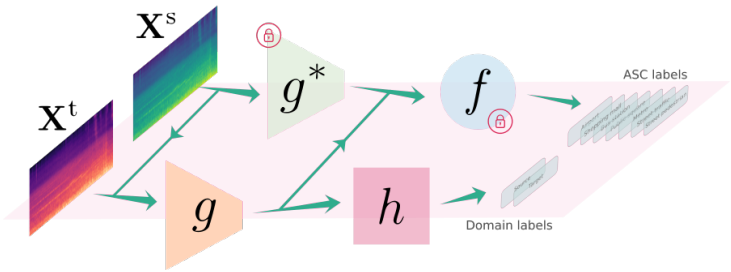
4.1.4 Adversarial domain adaptation

The moment normalization and moment matching techniques defined by equations (4.3), (4.5), (4.6) and (4.5) improve robustness to linear filtering of the acoustic scene, however they fail to compensate for non-linear mismatches, e.g., reverberation or phase distortion.

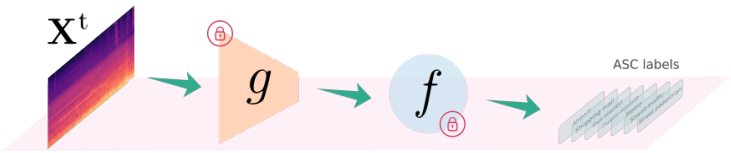
To mitigate these non-linear mismatches we follow the unsupervised domain adaptation method for ASC proposed by Drossos et al. (2019). Since this method does not rely on reference annotations for the classification task, it is suitable to scenarios in which no class labels are available in the target domain. The general framework is a two-step adversarial domain adaptation process based on the Wasserstein generative adversarial networks (WGAN) formulation (Arjovsky et al., 2017). It relies on three deep neural network-based models: a feature extractor g, a classifier f which outputs the vector of posterior probabilities of all ASC classes, and a discriminator h which outputs the posterior probability that the input data is from the target (as opposed to the source) domain. We regard as source domain data $\mathbf{X}^{\mathrm{s}} = \{x_{dnmk}\}_{n=1}^{N_{\mathrm{s}}}$ the acoustic scenes from device d, with one-hot labels $\mathbf{y}^{\mathrm{s}}$ of the considered classes. We regard as target domain data $\mathbf{X}^{\mathrm{t}} = \{x_{d'nmk}\}_{n=1}^{N_{\mathrm{t}}}$ the acoustic scenes recorded by some other device d', without class labels. Starting from a pre-trained feature extractor g^* and a classifier f trained on source data, the goal is to regularize the feature extractor g using the discriminator h so that it produces features $g(\mathbf{X}^{\mathrm{s}})$ and $g(\mathbf{X}^{\mathrm{t}})$ which exhibit the same distribution across domains. This adaptation procedure is carried out in three stages: *pretraining*, *adaptation* and *inference*, which are illustrated in Figure 4.3 and described below.



(a) Pretraining.



(b) Adaptation.



(c) Inference.

Figure 4.3: Adversarial domain adaptation framework.

Pretraining stage: This stage corresponds to the development of a model trained in a supervised way only on the source domain data. As such, we obtain the pre-trained feature extractor g^* and label classifier f using the source domain data $\mathbf{X}^{\mathrm{s}}$ together with their corresponding reference annotations $\mathbf{y}^{\mathrm{s}}$ and minimize the objective function

$$\mathcal{L}_s = -\sum_{n=1}^{N_{\mathrm{s}}} \mathbf{y}_n^{\mathrm{s}} \cdot \log(f(g^*(\mathbf{X}_n^{\mathrm{s}})), \tag{4.8}$$

where $\cdot$ denotes the dot product.

Adaptation stage: In this stage, two new models are introduced to perform adversarial domain adaptation: a feature extractor g meant to perform well on the target domain and a binary classifier h that discriminates between the source and target domain data. The feature extractor g is initialized as the pretrained model g^*, and g and h are jointly trained on source and target domain data by minimizing

$$\mathcal{L}_h = \sum_{n=1}^{N_{\mathrm{s}}} h(g^*(\mathbf{X}_n^{\mathrm{s}})) - \sum_{n=1}^{N_{\mathrm{t}}} h(g(\mathbf{X}_n^{\mathrm{t}})) \tag{4.9}$$

$$\mathcal{L}_g = \sum_{n=1}^{N_{\mathrm{t}}} h(g(\mathbf{X}_n^{\mathrm{t}})) - \sum_{n=1}^{N_{\mathrm{s}}} \mathbf{y}_n^{\mathrm{s}} \cdot \log(f(g(\mathbf{X}_n^{\mathrm{s}})) \tag{4.10}$$

to enforce domain-invariant distributions. The second term in (4.10) is a classification loss that prevents g from losing performance on the source domain data. Following Drossos et al. (2019), $\mathcal{L}_h$ and $\mathcal{L}_g$ are iteratively minimized by updating h according to the gradient of (4.9) w.r.t. h with g fixed, and updating g according to the gradient of (4.10) w.r.t. g with h fixed.

Inference stage: After adaptation, the feature extractor g and classifier f are used to classify acoustic scenes from both source and target devices.

4.1.5 Conditional adversarial domain adaptation

The above adversarial domain adaptation strategy ensures the alignment of the marginal distributions of the source- and target-domain features, however it does not guarantee the alignment of their class-conditional distributions. To address this issue, we extend the adversarial-based UDA framework with conditional adversarial domain adaptation (CADA) (Long et al., 2018): an alternative adversarial domain adaptation formulation that enforces the joint distribution alignment of features and class labels. As shown in Figure 4.4, the key idea is to condition the domain discriminator h on the class-posteriors from f with the joint variable $w(\mathbf{X}) = g(\mathbf{X}) \otimes f(g(\mathbf{X}))$ which aims to capture the multimodal information of g and f. Introducing the multilinear mapping through $w(\mathbf{X})$, the losses in (4.9) and (4.10) become

$$\mathcal{L}_h = \sum_{n=1}^{N_\mathrm{s}} h(w^*(\mathbf{X}_n^\mathrm{s})) - \sum_{n=1}^{N_\mathrm{t}} h(w(\mathbf{X}_n^\mathrm{t})) \tag{4.11}$$

$$\mathcal{L}_g = \sum_{n=1}^{N_\mathrm{t}} h(w(\mathbf{X}_n^\mathrm{t})) - \sum_{n=1}^{N_\mathrm{s}} \mathbf{y}_n^\mathrm{s} \cdot \log(f(g(\mathbf{X}_n^\mathrm{s}))). \tag{4.12}$$

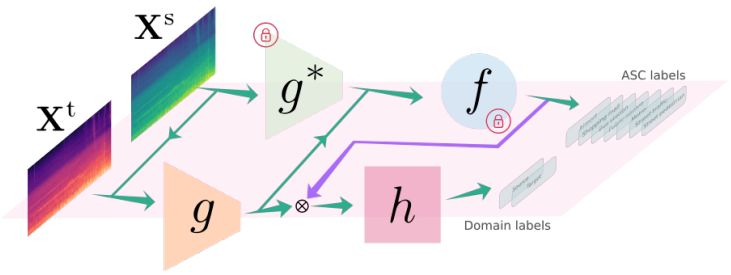


Figure 4.4: Conditional adversarial domain adaptation.

4.2 Experimental setup

4.2.1 Dataset

In order to assess the impact of moment normalization and moment matching, as well as their integration with adversarial domain adaptation, we perform experiments on the development dataset of the DCASE Challenge 2018 Task 1B.

Description

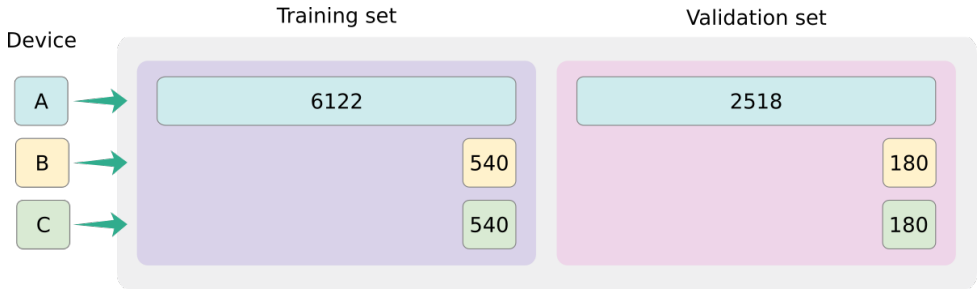
As shown in Table 4.1 the dataset is composed of 10 different acoustic scenes grouped into three different environments. The acoustic scene recordings span 10 s and were captured in six European cities using three different recording devices, namely devices A, B and C. Device A is a Soundman OKM II Klassik/studio A3, electret binaural microphone and a Zoom F8 audio recorder. Smartphones such as the Samsung Galaxy 7 and IPhone SE correspond to device B and C, respectively. The number of available recordings are 8,640, 720, and 720 from device A, B, and C, respectively. This is equivalent to 28 hours of audio out of which 24 hours are from device A, 2 hours from device B, and 2 hours from device C.

Table 4.1: Labels considered in the TUT urban acoustic scenes 2018 mobile development dataset.

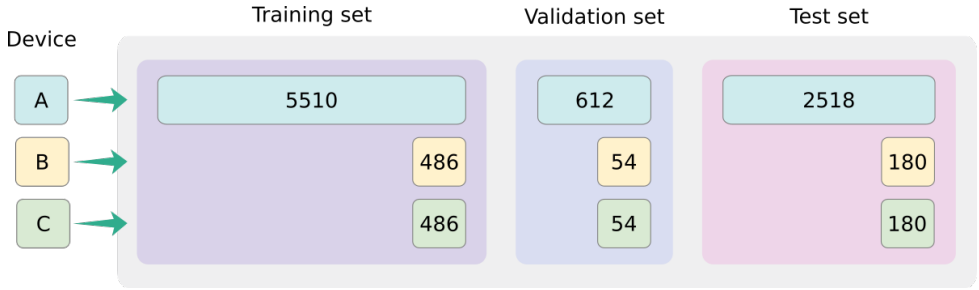
Categories	Indoor	Outdoor	Vehicles
Acoustic scenes	Airport Metro station Shopping mall	Park Public square Pedestrian street Street traffic	Bus Tram Metro

Original dataset setup for unsupervised domain adaptation

In the context of the DCASE Challenge 2018, the available recordings constituted a development set split into a training and validation sets. [3] Hence, Gharib et al. (2018) proposed to use the validation data as the test data for evaluation, and selected randomly 10% of the original training split as the validation data. Figure 4.5 shows the number of segments in the original dataset setup and in the split proposed by Gharib et al. (2018).



(a) Number of samples in the original setup.



(b) Number of samples per split (after the new validation split).

Figure 4.5: Dataset setup proposed by Gharib et al. (2018).

Problem of simultaneous recordings

Following the same setup as in Gharib et al. (2018), we found that 8.8% of the data that constitute the training and validation splits comprise recordings that were captured simultaneously by all recording devices, i.e., *parallel data* (see Figure 4.6). Since the data that composes the validation

[3] While the DCASE Challenge 2018 remained open, an evaluation set comprising test recordings was not publicly available. The evaluation set was only released after the challenge ended and served to ranked participants according to their performance on the given tasks.

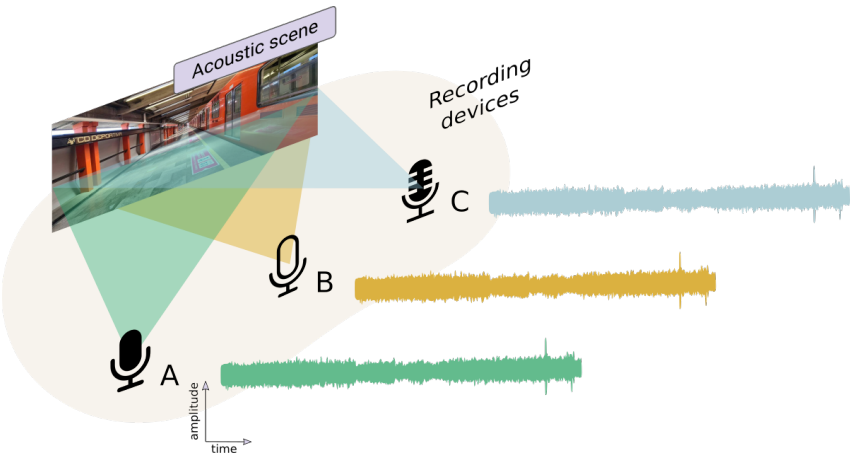
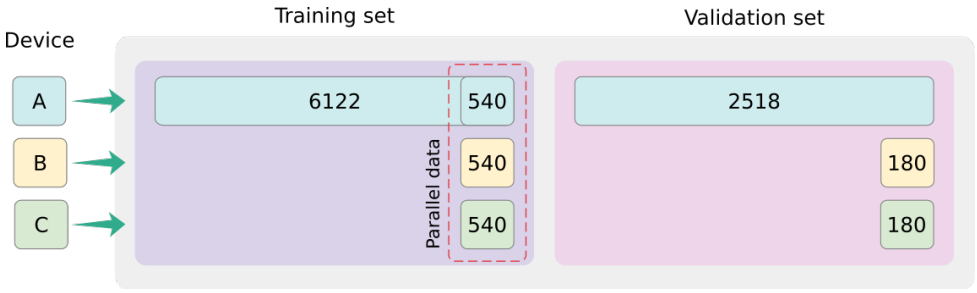
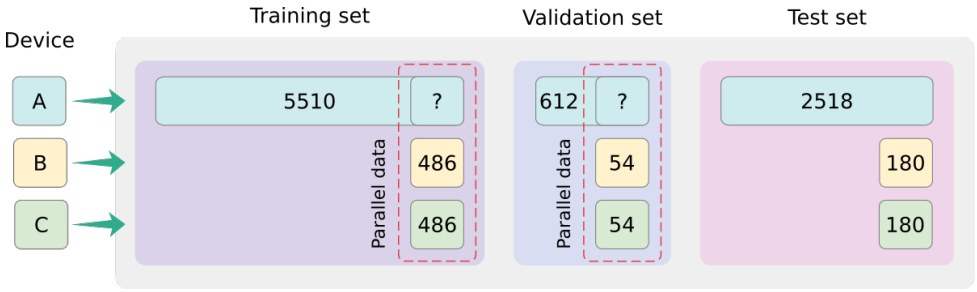


Figure 4.6: Capturing an acoustic scene simultaneously by different recording devices leads to *parallel* recordings.

set was randomly sampled from the training set, these parallel data are also randomly distributed in both training and validation sets. Figure 4.7 illustrates this matter.



(a) Number of parallel samples in the original setup.



(b) Parallel samples randomly distributed in the training and validation sets.

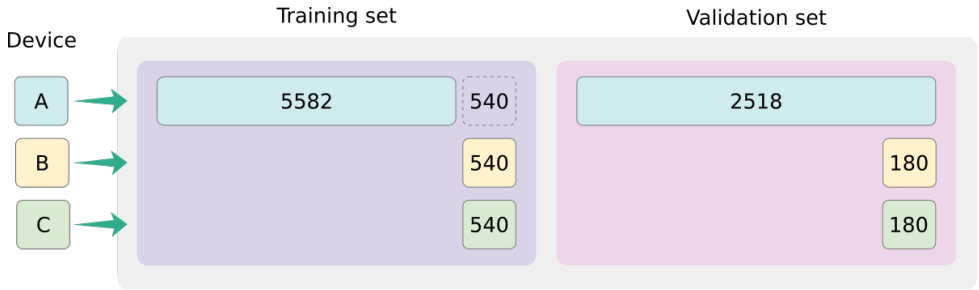
Figure 4.7: Issue with parallel data in the dataset.

Having parallel recordings raises two issues in the context of unsupervised domain adaptation. First, in the adaptation step, the model may discriminate distortions between parallel recordings more easily; second, in the inference step, the parallel recordings from the target devices could be considered by the model as transformed examples of acoustic scenes already seen during training, thus making their classification easier. The work of Gharib et al. (2018) as well as following works on UDA for ASC (Drossos et al., 2019; Mezza et al., 2021) using this setup suffer implicitly from these issues. We note however that Yang et al. (2021) exploited explicitly the subset of parallel

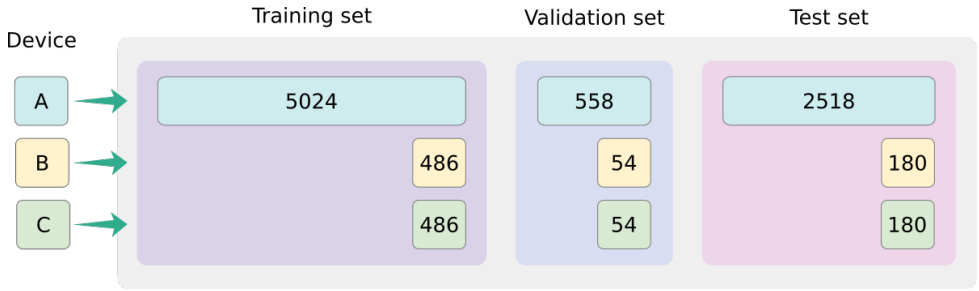
recordings to ease the adaptation process. [4]

Refined dataset setup

In order to address unsupervised domain adaptation in the fully blind setting where a pretrained ASC system must be deployed on devices with unknown microphone responses, the assumption of the availability of parallel recordings is not realistic. Accordingly, we discarded the subset of recordings from device A which are parallel to recordings from devices B and C prior to generating the validation set comprising 10% of the data from the original training set. Our final setup without parallel recordings, which is illustrated in Figure 4.8, comprises 5,024 training audio scenes from the source device A, and 486 for each target device B and C. The validation set comprises 558 acoustic scenes from the source device A, and 54 acoustic scenes from each target device B and C. The test set is composed of 2,518 acoustic scenes from device A, and 180 acoustic scenes from each target device B and C.



(a) Number of samples in the refined setup without the subset of parallel recordings from device A (in the dash-lined square).



(b) Number of samples per split.

Figure 4.8: Refined dataset to perform fully unsupervised domain adaptation.

4.2.2 Input features

From each acoustic scene recording, we extracted 64-dimensional log-Mel spectra, using a Hamming window of 2,048 samples (equivalent to 46 ms) and a hop size of 1,024 samples (23 ms), leading to an overlap of 50% across frames. Since the amount of available data from device A is significantly higher than the amount of data from devices B and C), we over sampled the generated log-Mel spectrograms from these last devices such that we approximate the same amount of data as device A only for the pretraining and adaptation stages.

[4] In practice, experiments with parallel data (not shown in this thesis) resulted in a higher mean accuracy for 72 out of the 90 results reported in Table 4.2, out of which 8 were statistically significant.

4.2.3 Model and training

We employ the model architecture referred to as *Kaggle* by Gharib et al. (2018); Drossos et al. (2019); Mezza et al. (2021). The network architecture of the Kaggle model is shown in Figure 4.9. The feature extractor g consists of five convolutional neural network (CNN) layers, with square kernel shapes of widths 11, 5, 3, 3, 3, and 48, 128, 192, 192, 128 channels. The stride is (2, 3) for the first two layers and (1, 1) for the rest. All layers are followed by rectified linear unit (ReLU) activation, and the first two and last layers use batch normalization and max pooling, with square kernels of width 2 and a stride of (1,2), (2,2),(1,2). The label classifier f consists of two linear layers with ReLU activations and dropout of 25% followed by a linear layer with softmax activation. The domain discriminator h consists of a linear layer with ReLU activation followed by a linear layer without activation. The RMSProp optimizer is used with a learning rate of 5×10^{-5}. We use a batch size of 16 and the feature classifier g was trained for 300 epochs.

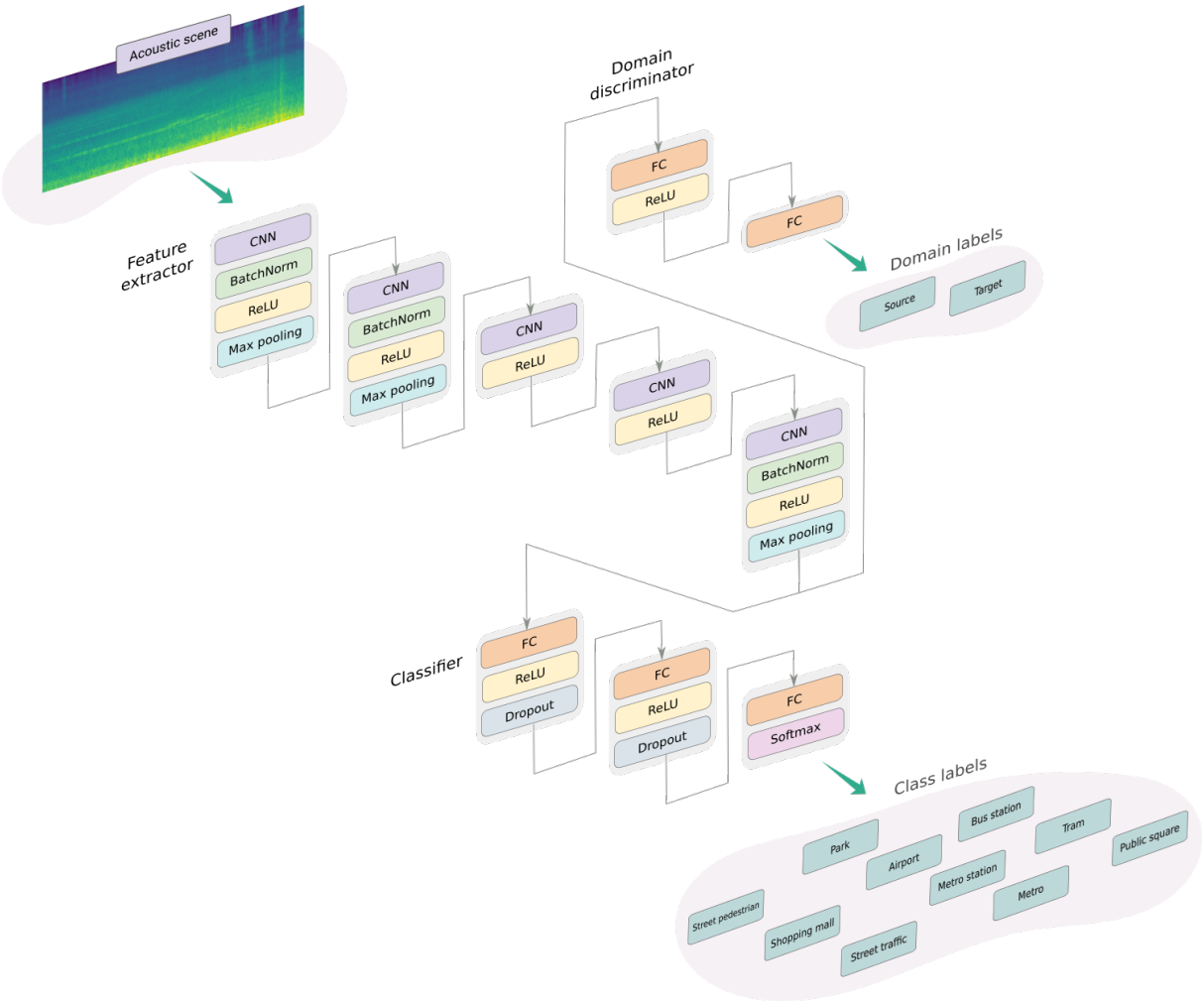
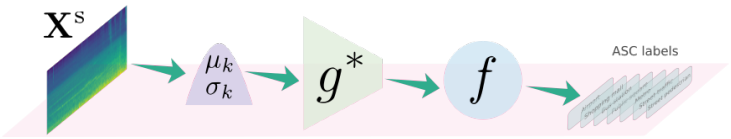


Figure 4.9: Network architecture of the *Kaggle* model.

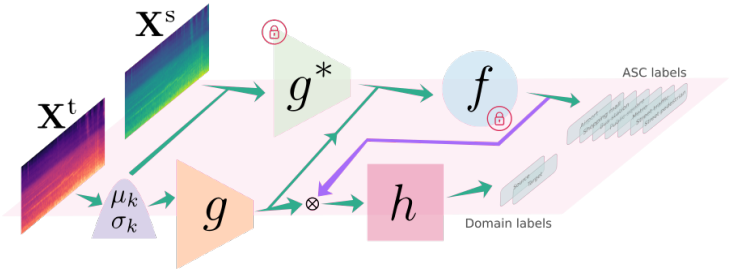
4.2.4 Experiments

We carry out experiments to analyze the impact of moment normalization and moment matching used alone or in combination with adversarial domain adaptation (ADA) or conditional adversarial domain adaptation (CADA). Figure 4.10 shows a general overview of the proposed method.

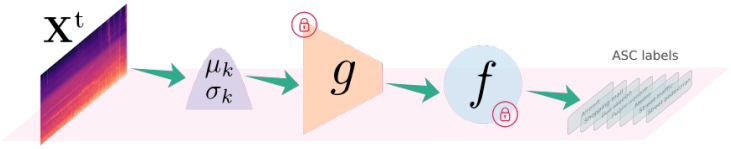
More specifically, we transform the source device data by mean normalization (MN) or mean and variance normalization (MVN) in the pretraining step. At inference time, we transform the target device data by the same MN or MVN strategy when it was applied in the pretraining step, or by mean matching (MM) or mean and variance matching (MVM) when no normalization was applied in the pretraining step. In addition, we also test a hybrid mean normalization and variance matching (MNVM) strategy, where mean normalization is applied to the source and target data and the variances are subsequently matched. In the adaptation step, the above strategies are used to transform the target data prior to ADA or CADA. We also evaluate ADA and CADA alone for comparison.[5] Figure 4.10 illustrates the proposed integration of moment normalization and/or matching strategies (as a pre-processing module) within the adversarial domain adaptation framework in all its three stages.



(a) Pretraining.



(b) Adaptation.



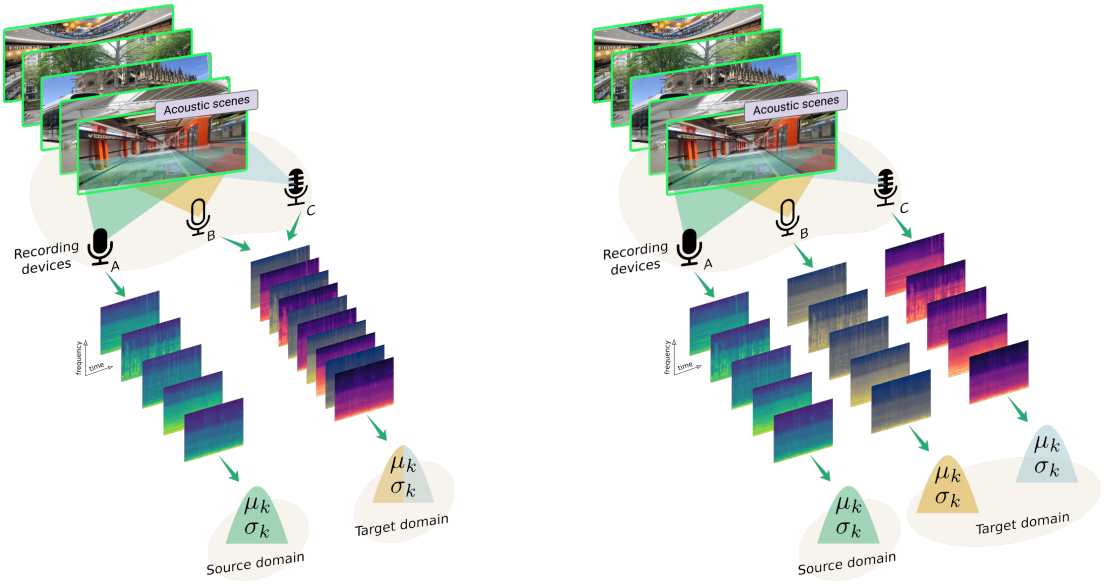
(c) Inference.

Figure 4.10: Proposed integration of moment normalization and/or matching strategies to the adversarial domain adaptation framework.

For ADA/CADA, we consider devices B and C as one domain, because the amount of training data from each device is small (Drossos et al., 2019). By contrast, we perform moment normalization and matching in two settings: *device-independent*, in which we regard devices B and C as one domain and transform the data using the average sample statistics of the two devices, and *device-dependent*, in which we regard B and C as two distinct domains and transform the

[5] Applying MVM alone at inference time is equivalent to BWSM (Mezza et al., 2021). The results differ from those of Drossos et al. (2019) and Mezza et al. (2021) due to discarding parallel data and not standardizing the data using the average statistics of devices A, B and C.

data using their respective statistics. We illustrate such configuration in Figure 4.11. When no moment normalization/matching is performed, the system is categorized as device-independent.



(a) Device independent setting. The statistics of the target domain are computed from devices B and C altogether.

(b) Device dependent setting.The statistics are computed for each device in the source and target domains.

Figure 4.11: Device settings.

4.3 Results

Initial domain mismatch

We start by noting from Table 4.2 that the accuracy of the system pretrained on unnormalized data reaches 59.3% on unnormalized source domain data, but drops to 13.6% on unnormalized target domain data (1st row). An absolute difference of 46% between source and target domain performance exposes how severe mismatched recording conditions can be for an ASC model. In fact, an analysis of the mean magnitude values of the input features, which is illustrated in Figure 4.12, shows abrupt differences among all recording devices, being more noticeable the difference between the source and target domain devices. According to Figure 4.13 the differences in the frequency-wise distribution of the input features causes a shift in the learned audio embeddings, which limits the model to be optimal only for the source domain data and bias the model towards a couple of classes (e.g., metro and metro station) in the target domain.

Effects of moment normalization

Pretraining a model on normalized data (MN or MVN) and applying the same normalization during inference is sufficient to largely correct the mismatch between the source and target distributions. Figure 4.14 shows the effects of MN in the learned embedding space and classification performance in the target domain. As discussed in Section 4.1.2, first- and second-order moment normalization has the effect of removing the linear distortions introduced by the recording device. The system's performance in the target domain increases by up to 30% absolute in the

Table 4.2: Average ASC accuracy (%) and standard deviation (in parentheses) on the test set achieved over 20 training runs. Bold numbers show the best statistically significant (p-value $<$ 0.05) results in the target domain (with respect to the best performing model).

Method			Device independent		Device dependent	
Pretraining	Adaptation	Inference	source	target	source	target
-	-	-	59.3(5.1)	13.6(2.3)	N/S	N/S
-	ADA (Drossos et al., 2019)	-	62.3(2.0)	36.2(2.6)	N/S	N/S
-	CADA	-	61.0(2.7)	39.4(2.7)	N/S	N/S
MN	-	MN	62.5(2.0)	43.5(2.3)	62.1(2.1)	51.1(2.4)
MN	MN-ADA	MN	64.7(1.0)	50.8(1.6)	64.0(2.8)	58.5(1.2)
MN	MN-CADA	MN	64.5(1.1)	51.6(1.3)	64.2(1.3)	59.0(1.6)
MVN	-	MVN	62.3(1.9)	41.8(1.0)	62.5(2.1)	51.0(2.0)
MVN	MVN-ADA	MVN	64.8(1.4)	**52.0(1.2)**	64.7(1.3)	**59.3(1.4)**
MVN	MVN-CADA	MVN	64.3(1.3)	**52.7(1.6)**	65.0(1.4)	**60.0(1.2)**
-	-	MM	59.3(5.1)	37.3(5.5)	59.3(5.1)	50.0(3.5)
-	ADA	MM	62.3(2.0)	41.3(2.0)	62.3(2.0)	52.0(2.7)
-	MM-ADA	MM	62.1(2.8)	47.0(2.3)	62.2(2.0)	56.2(1.8)
-	CADA	MM	61.0(6.5)	40.3(2.0)	61.0(6.5)	51.9(2.6)
-	MM-CADA	MM	62.5(1.5)	48.9(2.2)	62.7(1.2)	57.7(1.7)
-	-	MVM (Mezza et al., 2021)	59.3(5.1)	38.8(2.9)	59.3(5.1)	50.1(3.5)
-	ADA	MVM	62.3(2.0)	38.2(2.2)	62.3(2.0)	51.0(2.3)
-	MVM-ADA	MVM	63.7(1.1)	47.4(1.9)	62.6(1.5)	56.5(1.7)
-	CADA	MVM	62.4(2.8)	38.1(1.8)	62.4(2.8)	51.2(2.1)
-	MVM-CADA	MVM	63.0(1.3)	50.4(1.2)	63.2(1.6)	58.5(2.0)
MN	-	MNVM	64.2(2.4)	42.3(4.3)	63.1(1.8)	51.8(2.2)
MN	MN-ADA	MNVM	64.7(1.0)	48.6(1.9)	64.0(2.22)	56.0(1.7)
MN	MN-CADA	MNVM	64.0(1.6)	49.4(1.2)	64.0(1.6)	57.2(1.3)
MN	MNVM-ADA	MNVM	64.2(1.8)	50.6(1.1)	64.2(1.0)	**59.9(1.6)**
MN	MNVM-CADA	MNVM	64.0(1.9)	51.7(1.1)	64.6(1.8)	**60.1(1.8)**

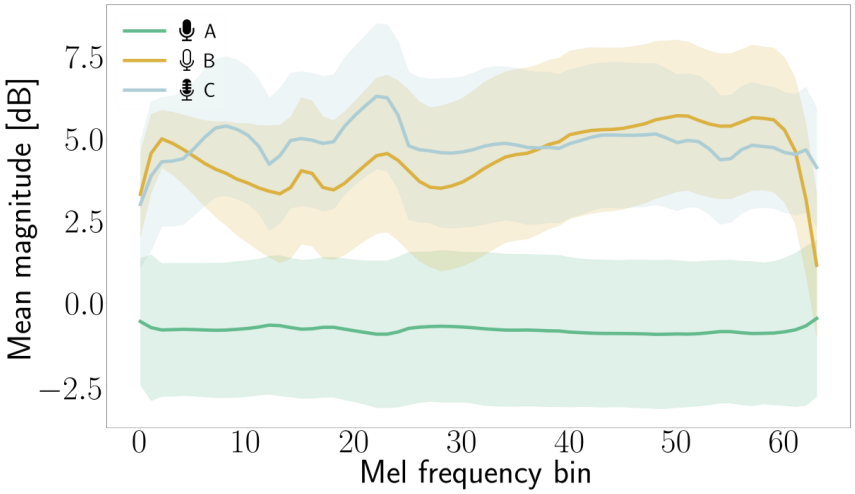
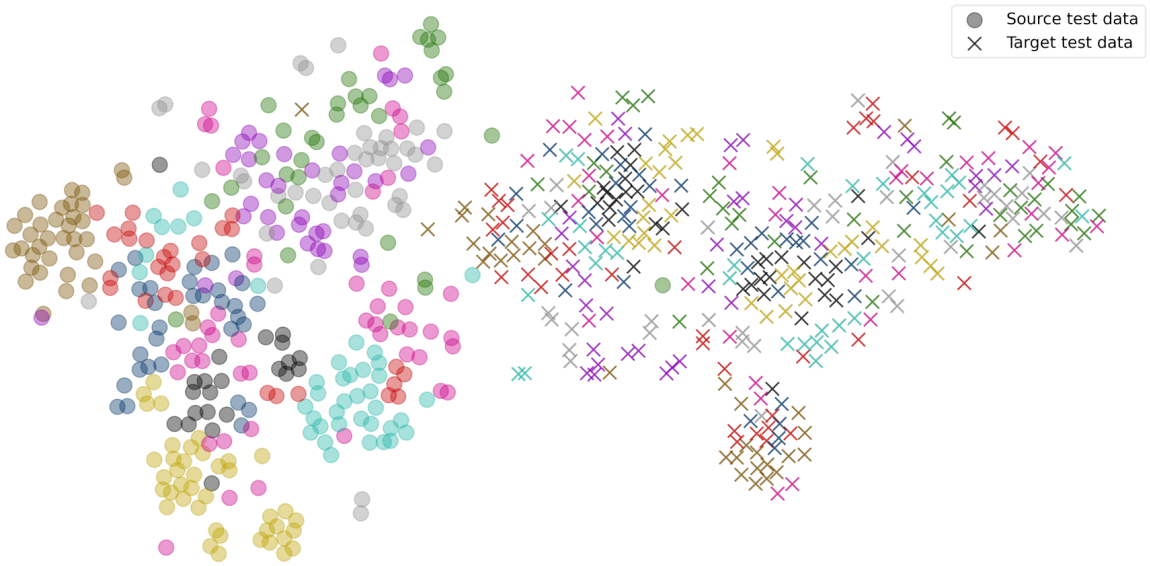
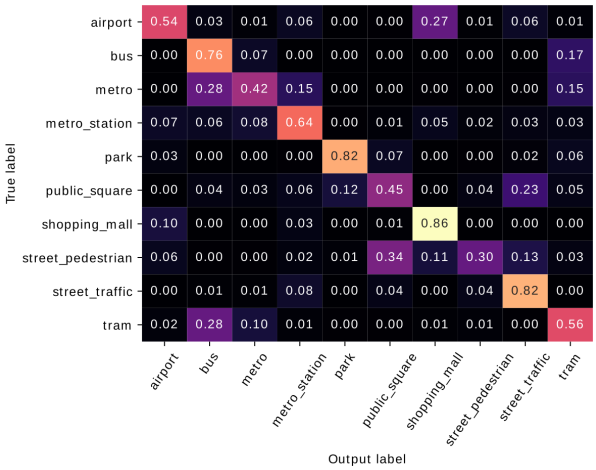


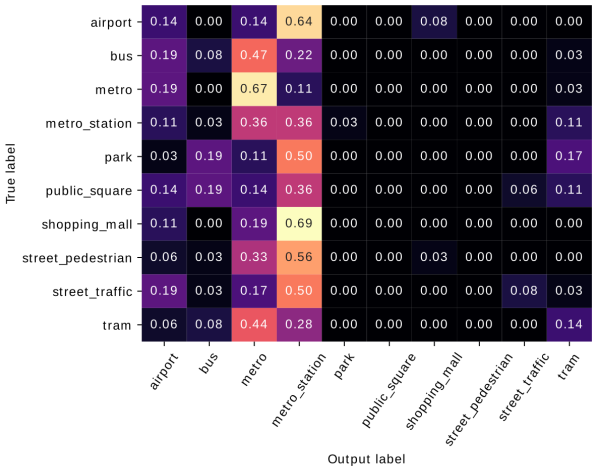
Figure 4.12: Mean and standard deviation per frequency bin for all recording devices.



(a) Audio embeddings from devices A (leftmost cloud of embeddings), B and C (rightmost cloud of embeddings) extracted from the source (non-adapted) model. A gap between the source (device A) and the target (devices B and C) embeddings shows the initial distribution mismatch.



(b) Confusion matrix for the source (non-adapted) model on the *source* domain data.



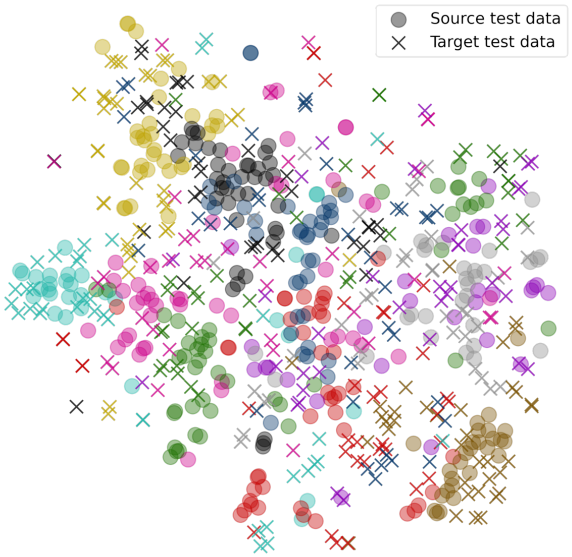
(c) Confusion matrix for the source (non-adapted) model on the *target* domain data.

Figure 4.13: Audio embeddings and confusion matrices for the source (non-adapted) model on the target domain data (see 1st row of Table 4.2).

device-independent setting and by 37% absolute in the device-dependent setting (4th and 7th rows). In the former setting MN outperforms MVN, while in the latter normalizing by MN or MVN is similarly effective.

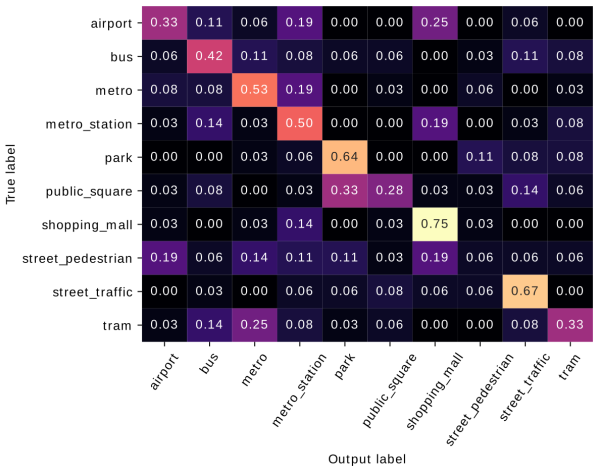
Effects of moment matching

Adapting an unnormalized pretrained system through moment matching (MM or MVM) increases the accuracy in the target domain by up to 25% absolute in the device-independent setting and by 36% absolute in the device-dependent setting (10th and 15th rows). Figure 4.15

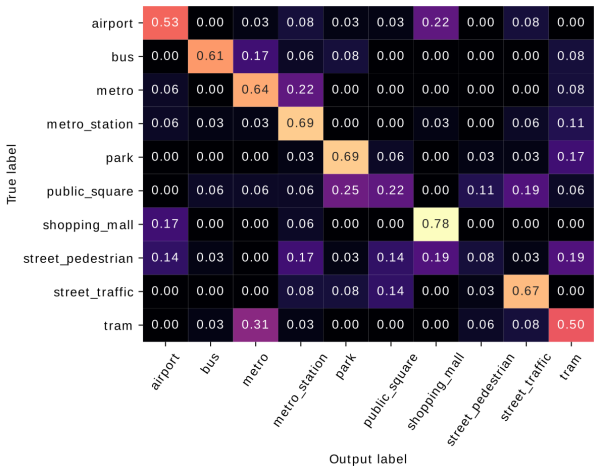


(a) Embedding space (device indep. setting).

(b) Embedding space (device dep. setting).



(c) Confusion matrix (device indep. setting).



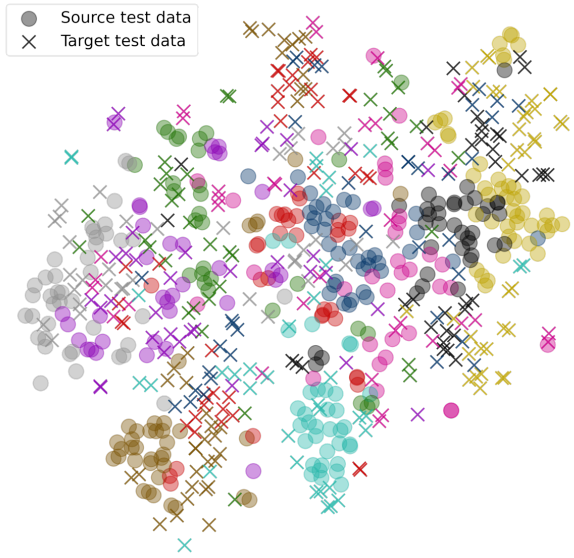
(d) Confusion matrix (device dep. setting).

Figure 4.14: Audio embeddings and confusion matrices from a model performing MN: *mean normalization* (see 4th row of Table 4.2) for the two device settings.

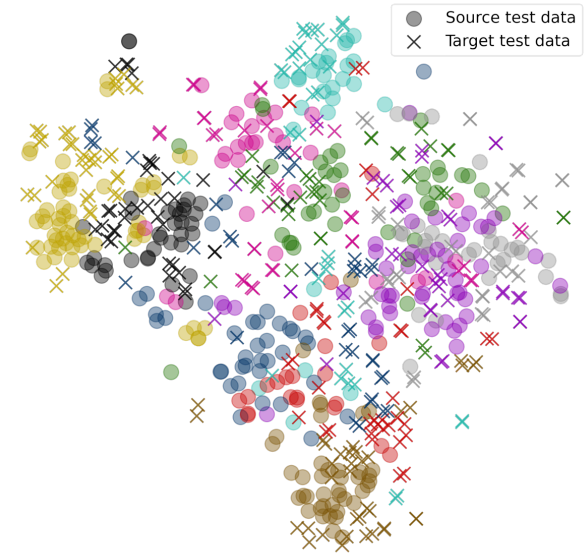
shows the effects of MVM in the learned embedding space and classification performance in the target domain. As discussed in Section 4.1.3, these results show that moment matching effectively transfers distortions from one recording device to another. In both device settings, MM and MVM are not statistically different.

Effects of adversarial domain adaptation

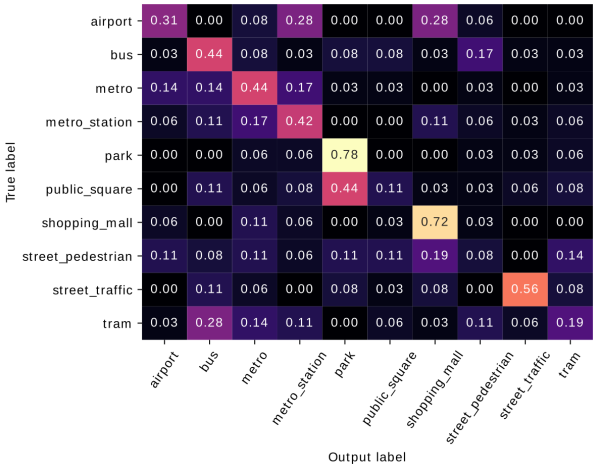
Adversarial domain adaptation methods with unnormalized data (2nd and 3rd rows) boost accuracy by up to 26% absolute in the target domain and by up to 3% absolute in the source domain. CADA obtains significantly higher average accuracy than ADA (p-value < 0.05) for the conditions in the 3rd, 14th, 19th and 24th rows in the device independent setting, and for



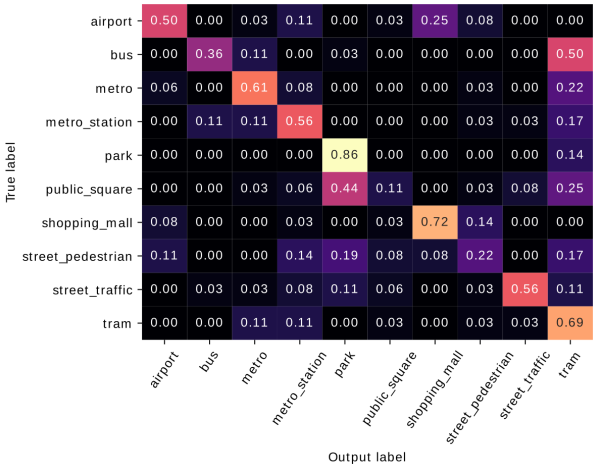
(a) Embedding space (device indep. setting).



(b) Embedding space (device dep. setting).



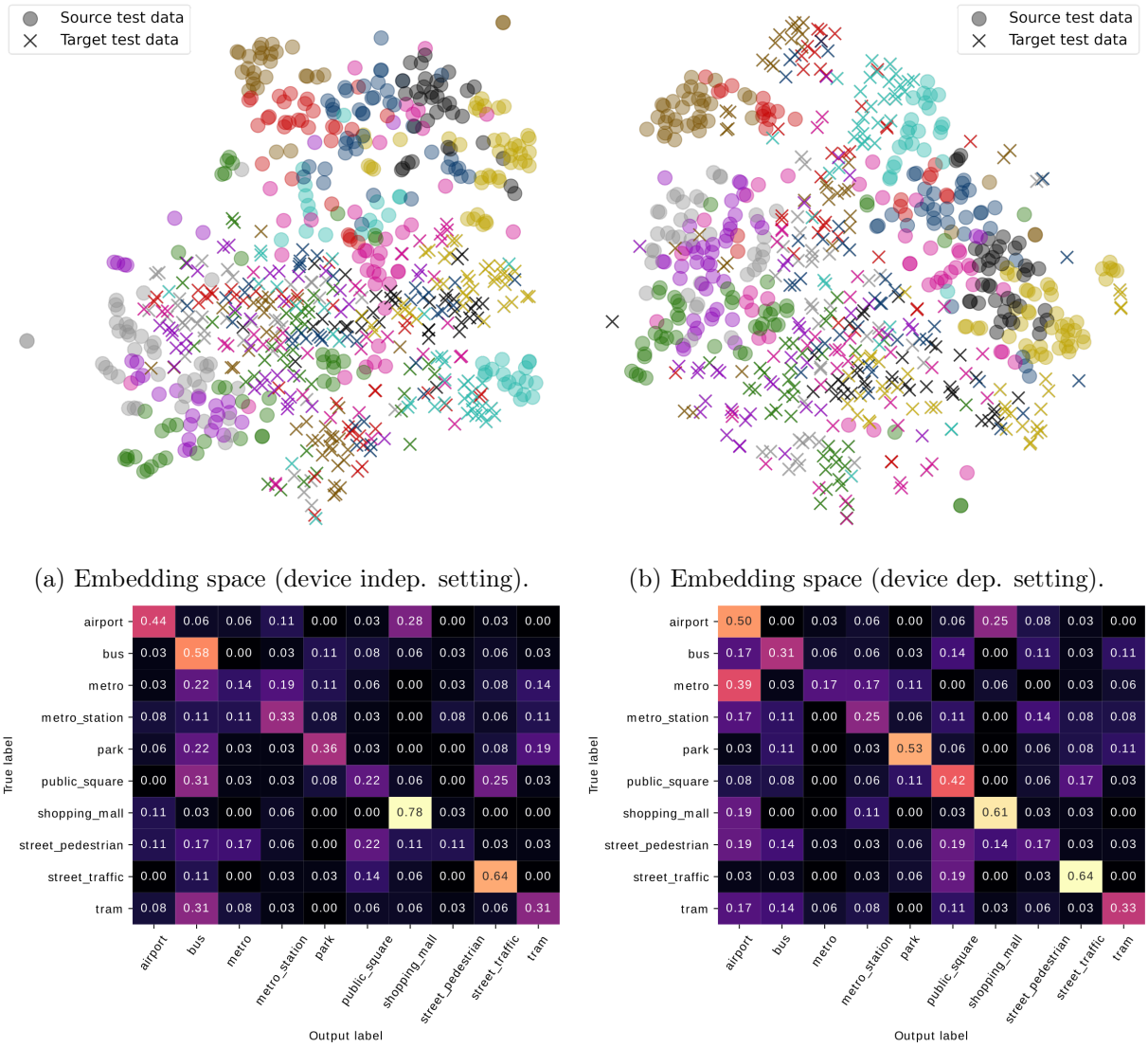
(c) Confusion matrix (device indep. setting).



(d) Confusion matrix (device dep. setting).

Figure 4.15: Audio embeddings and confusion matrices from a model performing MVM: *mean and variance matching* (see 15th row of Table 4.2) for the two device settings.

those in the 14th, 19th and 22nd rows in the device-dependent setting. In all conditions in which ADA and CADA are compared, CADA obtains a higher average accuracy than ADA, which is statistically significant when the difference in performance exceeds 1.1% (p-value < 0.05). Figure 4.16 compares ADA and CADA from a qualitative perspective by visualizing the learned embedding space after adaptation. CADA achieves its goal of correcting the class-conditional shift between the source and target data distributions, and as shown by the confusion matrices, the average accuracy is better distributed among all classes than with ADA, thus avoiding bias towards certain classes.



(a) Embedding space (device indep. setting).

(b) Embedding space (device dep. setting).

(c) Confusion matrix (device indep. setting).

(d) Confusion matrix (device dep. setting).

Figure 4.16: Comparison between ADA: *adversarial domain adaptation* (left column) and CADA: *conditional adversarial domain adaptation* (ADA). See 2nd and 3rd row of Table 4.2) for the performances in the two device settings.

Integration of moment normalization and/or matching with adversarial domain adaptation

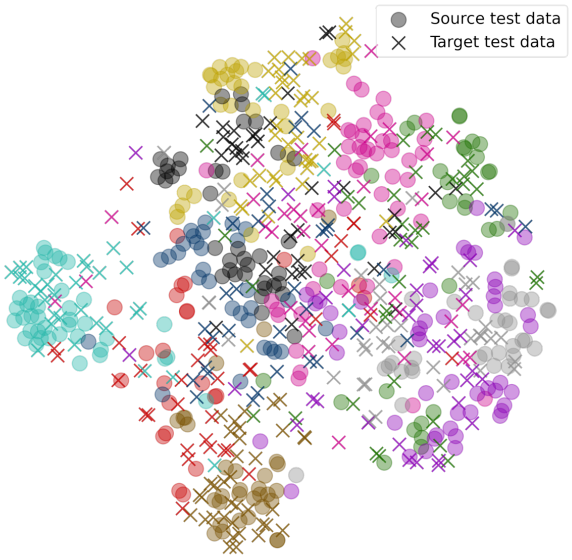
By integrating moment normalization or moment matching with adversarial methods, the gap between the source and target domains is further reduced. The best accuracy achieved in the target domain is 53% in the device-independent setting and 60% in the device-dependent one. Such large mismatch correction is obtained by standardizing the source and target data during adaptation regardless of the adaptation method (MVN-ADA or MVN-CADA, 8th and 9th rows). An equally good performance is obtained by normalizing the means and matching the variances of the data during adaptation (MNVM-ADA or MNVM-CADA, 23rd and 24th rows) in the device-dependent setting. This shows that second-order moment normalization or matching help further improve the performance in the target domain compared to methods that use first-order statistics only. Figure 4.17 shows the audio embeddings and confusion matrices from a MVN-CADA model. Compared to moment normalization, moment matching or adversarial methods alone, the integration of these methods results in invariant audio embeddings to both source and target domains, which results in a classification accuracy in the target domain similar to that obtained in the source domain, effectively reducing the domain discrepancy.

Achieving source domain performance on the target domain

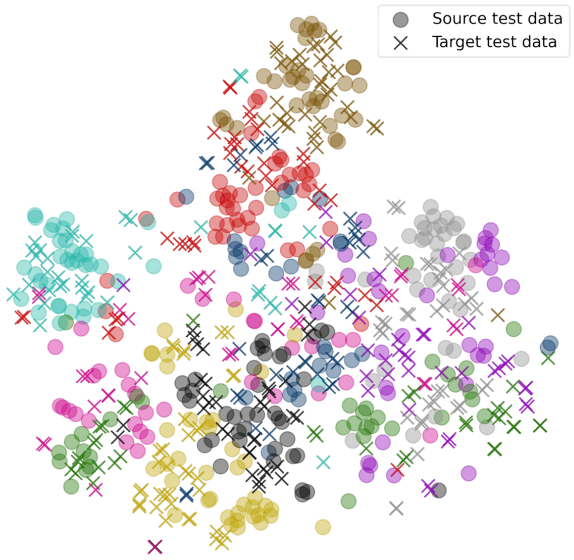
Figure 4.18 shows the average accuracy achieved by all proposed adaptation strategies in both device settings. From the plot, we identify a big gap between two groups of adaptation strategies, which suggest that those combining moment normalization and/or matching with adversarial domain adaptation are better choices than those lacking adversarial domain adaptation or that do not combine moment normalization and/or matching with adversarial domain adaptation. Regarding adversarial domain adaptation methods, CADA tends to outperform ADA by a small margin. Among the best performing adaptation strategies, those that apply moment normalization and/or matching in all steps tend to outperform moment matching strategies applied in the adaptation and inference steps. Moreover, as discussed in the previous section, these strategies are able to match the performance on the source domain. Figure 4.19 shows audio embeddings extracted from a non-adapted model to show the initial mismatch problem as well as audio embeddings extracted from an adapted model to show the mismatch correction.

4.4 Conclusion

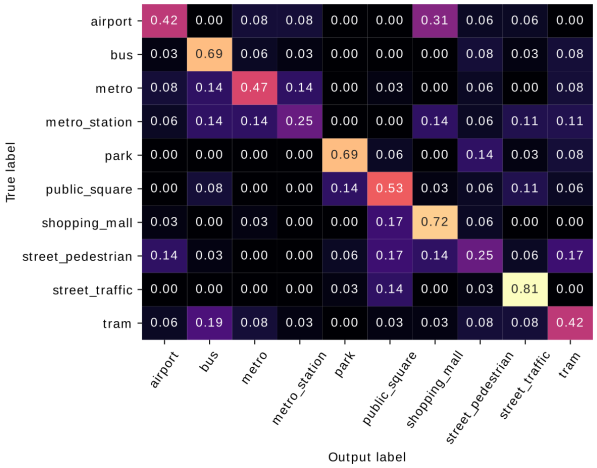
We experimentally assessed the impact of moment normalization and moment matching strategies as well as their integration with adversarial domain adaptation methods for acoustic scene classification with mismatched recording devices. We showed that normalization strategies are particular instances of a linear distortion model that improve robustness to mismatched recording devices. The combination of moment normalization and/or matching strategies with adversarial domain adaptation methods reduces remaining mismatch due to non-linear effects. Results indicate that such integration achieves a performance in the target domain close to that obtained in the source domain.



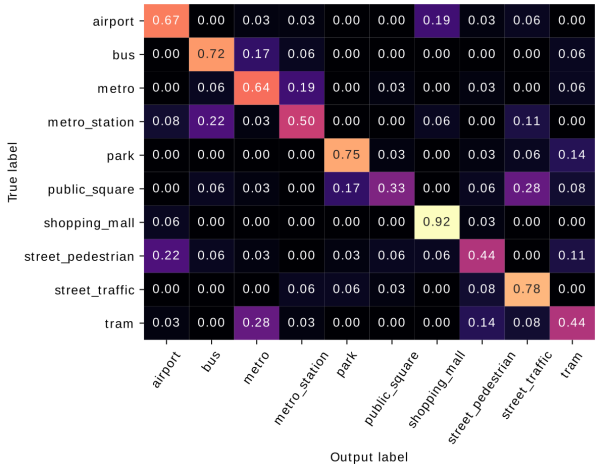
(a) Embedding space (device indep. setting).



(b) Embedding space (device dep. setting).



(c) Confusion matrix (device indep. setting).



(d) Confusion matrix (device dep. setting).

Figure 4.17: Audio embeddings and confusion matrices from a model performing MVN-CADA: *mean and variance normalization and conditional adversarial domain adaptation* (see 9th row of Table 4.2) for the two device settings.

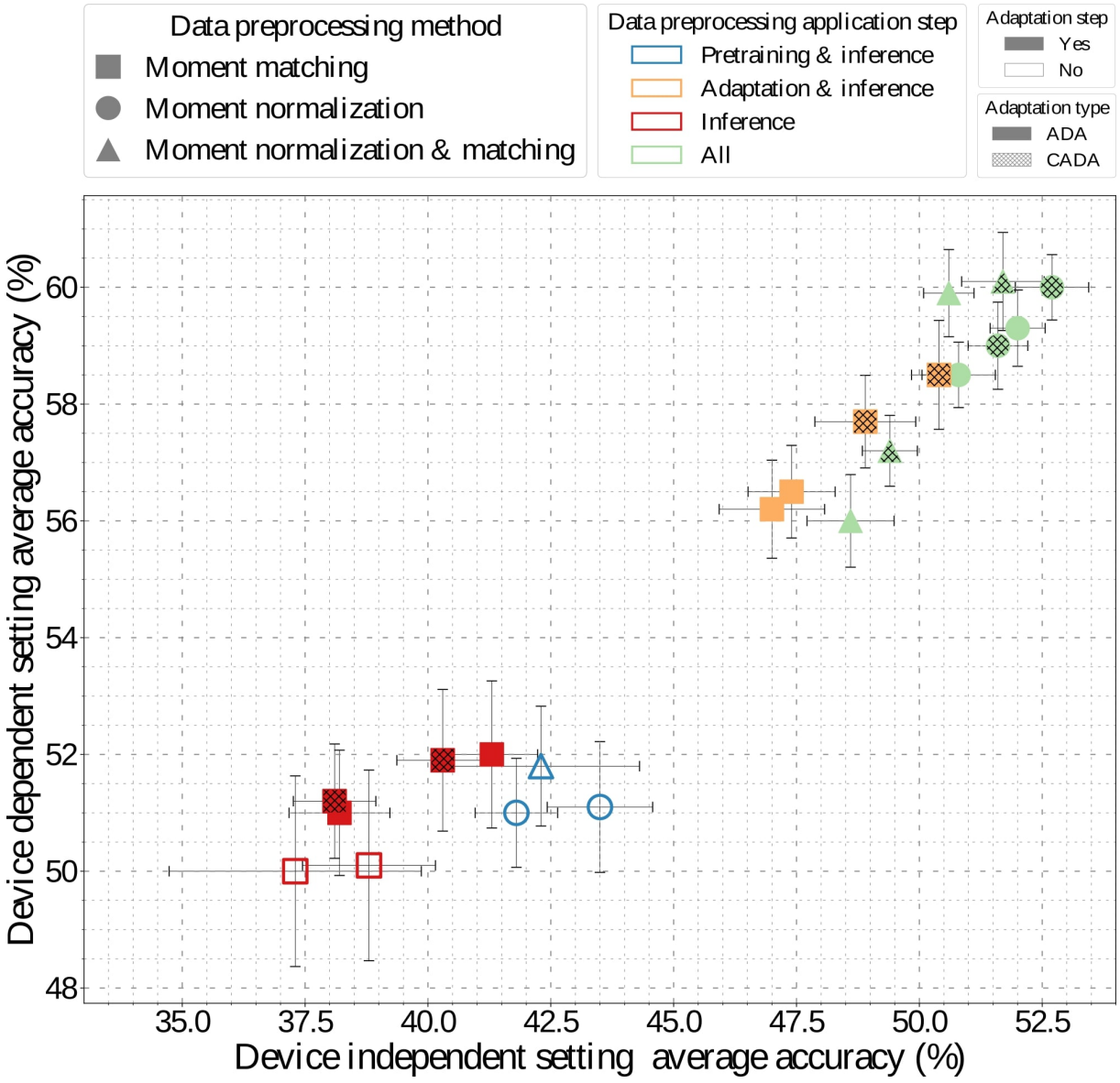
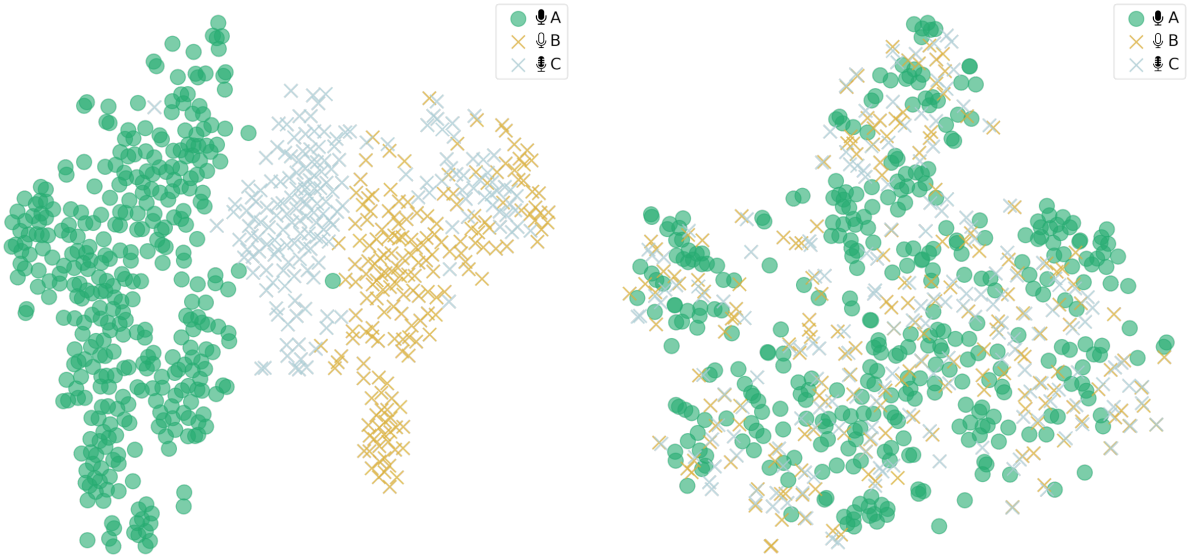


Figure 4.18: Target domain average accuracy in the device-independent setting and corresponding average accuracy in the device-dependent setting for the adaptation strategies in Table 4.2. Error bars indicate 95% confidence intervals.



(a) Embedding space of non-adapted model.

(b) Embedding space of adapted model.

Figure 4.19: Audio embeddings from devices A, B and C showing the initial mismatch between the source and target domains (left) as well as the mismatch correction by model MVN-CADA (right).

5

Improving sound event detection

Unsupervised domain adaptation strategies based on correction of the target test data or that involve a retraining stage such as the ones presented in the previous chapter are effective in counteracting degradation due to mismatched training and test recording devices that often hinders the performance of acoustic scene classification systems. Indeed, this source of mismatch can potentially affect any other audio analysis task. In this chapter, we are interested in another type of mismatch that similarly hampers the optimal performance of audio analysis tasks, namely, the mismatch between synthetic and real data. In particular, we focus on tackling this mismatch for the sound event detection task in domestic scenarios.

Over the past five years, Task 4 of the DCASE Challenge has encouraged the development of methods that improve the SED task. Commonly, SED models are trained in a semi-supervised way on a heterogeneous dataset (Turpault and Serizel, 2020) that includes a set of synthetic soundscapes with annotations indicating class labels and timestamps (strong labels), as well as a set of real recordings, mostly unlabeled and with only a small subset of them containing at most information about the active sound classes in the recordings (weak labels). The aim of SED, as illustrated in Figure 5.1, is to identify both the class label and time boundaries of all active sound events composing an acoustic scene. While this training approach aims to learn invariant representations for both types of data, there is a gap in performance between synthetic and real recordings. In order to reduce such a mismatch, many improvements to the SED task comprise augmentation schemes to ease generalization (Li, 2021; Miyazaki et al., 2020), changes in the acoustic front-end with alternative time-frequency representations to log-Mel spectrograms (Copiaco et al., 2021) or time-frequency resolutions for each sound event class (de Benito-Gorron et al., 2020), and class-dependent approaches and post-processing techniques to refine the detection scores (Huang et al., 2020a; Cances et al., 2019). Only few works have relied on adversarial domain adaptation to improve performance on real data (Cornell et al., 2020a; Yang et al., 2020; Zheng et al., 2021).

In this chapter we introduce several complementary methods to improve SED in domestic environments. First, we explore an acoustic front-end suited for domestic sounds based on multiple trainable PCEN transformations as an alternative to log-Mel spectrograms. Secondly, we propose to classify domestic sounds by their spectro-temporal content as foreground or background events as an auxiliary task for SED. In Chapter 3 we saw that a deep neural network-based source separation model is able to discriminate between the rapidly and slowly varying spectro-temporal features of foreground and background sounds. We aim to investigate whether this categorization is beneficial for SED too. The proposed foreground-background classifier is jointly trained with the SED branch in a multi-task fashion and the combination of both branches is also explored.

Third, we propose a domain adaptation strategy based on optimal transport (Courty et al., 2017a; Damodaran et al., 2018) to explicitly reduce the mismatch between synthetic and real recordings. Our proposed strategy aligns the empirical distributions of the feature representations of active and inactive frames of synthetic and real data. Lastly, we extend the proposed domain adaptation framework and combine it with an active learning strategy to further improve detection scores in real environments. We evaluate the performance of the proposed methods in terms of the event-based macro F1-score on the Domestic Environment Sound Event Detection Dataset (DESED) validation and public evaluation sets (Turpault et al., 2019; Serizel et al., 2020).

The structure of this chapter is as follows. We describe in Section 5.1 the trainable PCEN acoustic front-end for SED, and we present the proposed integration of foreground-background classification to the training recipe of SED. Then, we describe the optimal transport-based domain adaptation strategy to reduce the mismatch between training data and its extension to an active learning setup. Section 5.2 describes the experimental setup and 5.3 shows the results and analysis of the aforementioned improvements to the SED task. Section 5.4 concludes the chapter. The contents of this chapter are part of the work carried out for our participation in Task 4 of the 2020 and 2021 DCASE Challenges and have been published in subsequent DCASE Workshops (Cornell et al., 2020b; Olvera et al., 2021a).

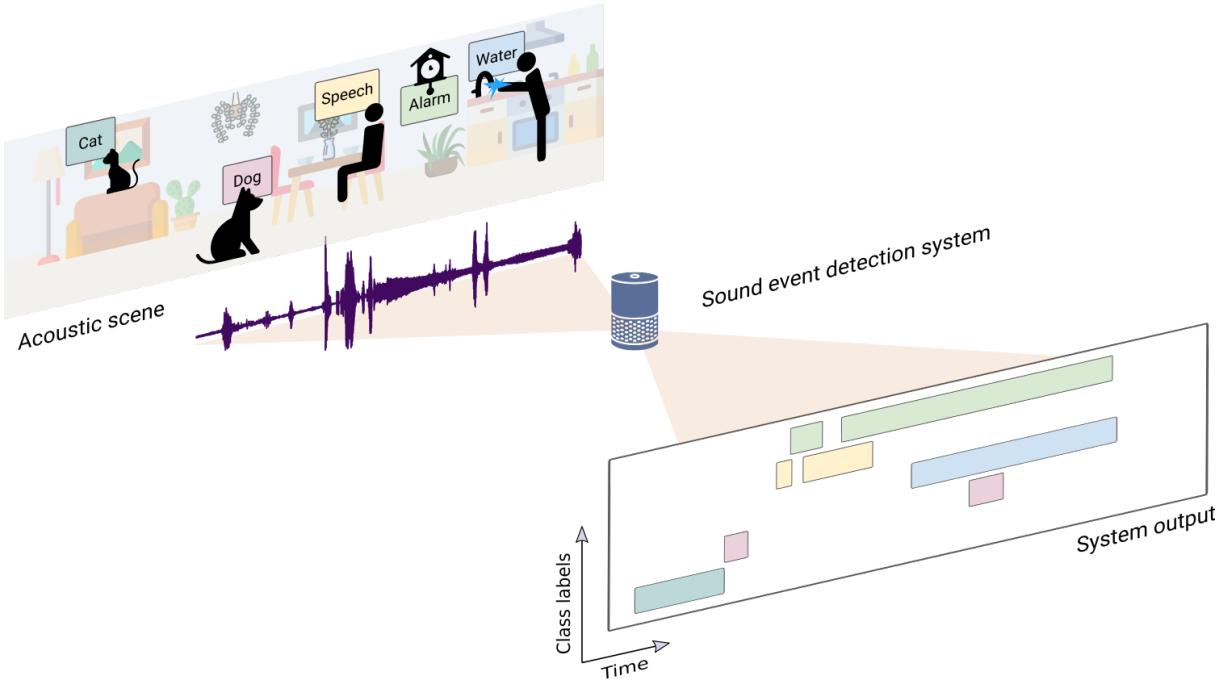


Figure 5.1: Overview of the sound event detection task. The sound event detection model analyzes an input recording and outputs class labels of active sound events with their respective onset and offset times.

5.1 Proposed Methods

In this section we describe the proposed methods that improve sound event detection. These methods are applied in different stages of the detection pipeline.

5.1.1 Per-channel energy normalization

Similarly to Chapter 3, we explore per-channel energy normalization (PCEN) as an alternative acoustic front-end to log-Mel spectrograms. PCEN was originally defined by Wang et al. (2017) as

$$\mathrm{PCEN}(n, f) = \left(\frac{x(n, f)}{(\epsilon + \bar{x}(n, f))^{\alpha}} + \delta \right)^{r} - \delta^{r}, \tag{5.1}$$

where n and f denote time and frequency band index of a time-frequency representation $x(n, f)$. The parameters α, ϵ, r and δ are positive constants, while $\bar{x}(n, f)$ is a smoothed version of $x(n, f)$, which is computed with a first-order IIR filter as $\bar{x}(n, f) = (1 - s)\bar{x}(n - 1, f) + sx(n, f)$, with s the smoothing coefficient.

As discussed in Chapter 3, PCEN is beneficial to enhance short sound events in the presence of long duration stationary sounds. However, in domestic settings in which it is equally important to detect sounds with fast transients and sounds of stationary characteristics, the filtering operation may have a negative impact on the sound classes of interest with the latter attributes (e.g., blender, vacuum cleaner, electric shaver) thus affecting their timely detection.

Therefore, for domestic sounds, it is necessary for a PCEN transformation to have parameters that achieve a trade-off between highlighting short sounds and avoiding degradation of long duration stationary-like sounds. This may however be sub-optimal for sound event detection. Thus, instead of using a PCEN transformation with fixed parameters, we propose as a better trade-off the use of multiple PCEN transformations as neural network layers, whose parameters are learned automatically during the training of a SED model. In this trainable acoustic front-end we seek to optimize parameters α, δ and r. We thus investigate whether PCEN transformations with fixed or learned parameters are better suited than log-Mel spectrograms as the acoustic front-end for sound event detection in domestic environments.

5.1.2 Foreground-background classification

Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be the input, output and latent spaces. For the SED task we denote the soundscape time-frequency representation by $x \in \mathcal{X}$ with corresponding annotations $y \in \mathcal{Y}$. We have access to a synthetic dataset with strong labels $\mathcal{D}^{\mathcal{S}} = \{(x_i^{\mathrm{s}}, y_i^{\mathrm{s}})\}_{i=1}^{n_{\mathrm{s}}}$ and two datasets of real recordings: a weakly labeled dataset $\mathcal{D}^{\mathcal{W}} = \{(x_i^{\mathrm{w}}, y_i^{\mathrm{w}})\}_{i=1}^{n_{\mathrm{w}}}$ and an unlabeled dataset $\mathcal{D}^{\mathcal{U}} = \{x_i^{\mathrm{u}}\}_{i=1}^{n_{\mathrm{u}}}$.

The SED model, illustrated in Figure 5.2 is a Mean Teacher model (Tarvainen and Valpola, 2017) in which both the student and the teacher models have the same convolutional-recurrent neural network (CRNN) architecture. We use the CRNN from the student model as a representation mapping $g : \mathcal{X} \to \mathcal{Z}$, where the log-Mel spectrograms are mapped to the latent space. The SED model is represented by the function $f : \mathcal{Z} \to \mathcal{Y}$ that maps the latent vectors to the output space.

FB branch

Motivated by the broad categorization of sound events into foreground and background according to their spectro-temporal structure, we propose a foreground-background (FB) auxiliary classifier $f_{\mathrm{FB}} : \mathcal{Z} \to \mathcal{Y}^{\mathrm{FB}}$ that maps the latent space to foreground-background labels. We learn this classifier jointly with the SED model in a multi-task fashion, hypothesizing that these different yet related classification tasks will help improve the network's generalization capability. Analogously, for the teacher model we denote by g', f' and f'_{FB} the CRNN embedding function, the SED

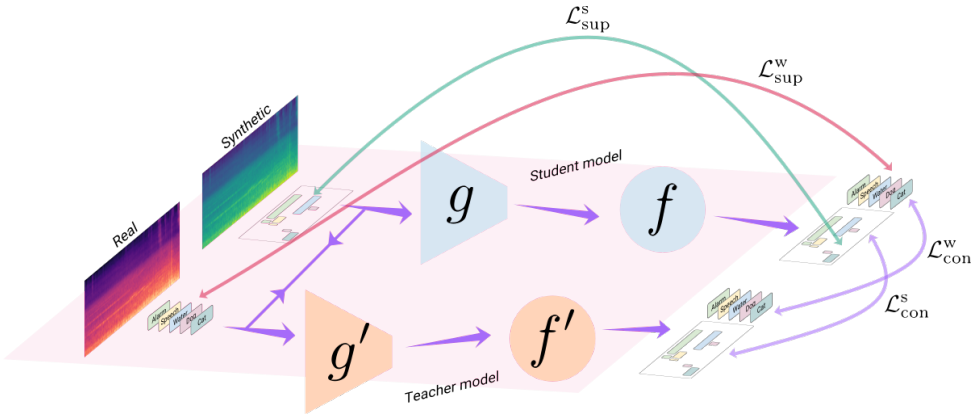


Figure 5.2: Baseline model based on the Mean Teacher framework for semi-supervised training. The training scheme comprises two pairs of cost functions: $\mathcal{L}^{s}_{sup}, \mathcal{L}^{w}_{sup}$ that train the system in a supervised way with ground-truth strong and weak labels, respectively; and $\mathcal{L}^{s}_{con}, \mathcal{L}^{w}_{con}$ that force consistency between the the teacher and student model outputs.

branch and the FB branch, respectively. Figure 5.3 shows the proposed system depicting the foreground-background classification branch.

In order to train the FB classifier in a multi-task fashion, as illustrated by Figure 5.4 we derived foreground-background ground-truth annotations y^{fb}_i from the strong labels y^{s}_i of the synthetic data by combining the sound event labels in two categories: foreground (*alarm - bell ringing, speech, cat, dog, dishes*) and background (*blender, vacuum cleaner, frying, electric shaver - toothbrush, running water*). The SED model and the FB classifier are optimized by minimizing

$$\begin{aligned}\mathcal{L}_{\text{SED}} = L(y^{s}_i, f(g(x^{s}_i))) + \lambda L_{\text{strong}}(f(g(x_i)), f'(g'(x_i))) + \\ L(y^{w}_i, f(g(x^{w}_i))) + \lambda L_{\text{weak}}(f(g(x_i)), f'(g'(x_i))) + \\ L(y^{fb}_i, f_{\text{FB}}(g(x^{s}_i))) + \lambda L_{\text{strong}}(f_{\text{FB}}(g(x_i)), f'_{\text{FB}}(g'(x_i)))\end{aligned} \quad (5.2)$$

where $L(\cdot,\cdot)$ is a binary cross-entropy classification loss, and $L_{\text{strong}}(\cdot,\cdot)$ and $L_{\text{weak}}(\cdot,\cdot)$ are mean-

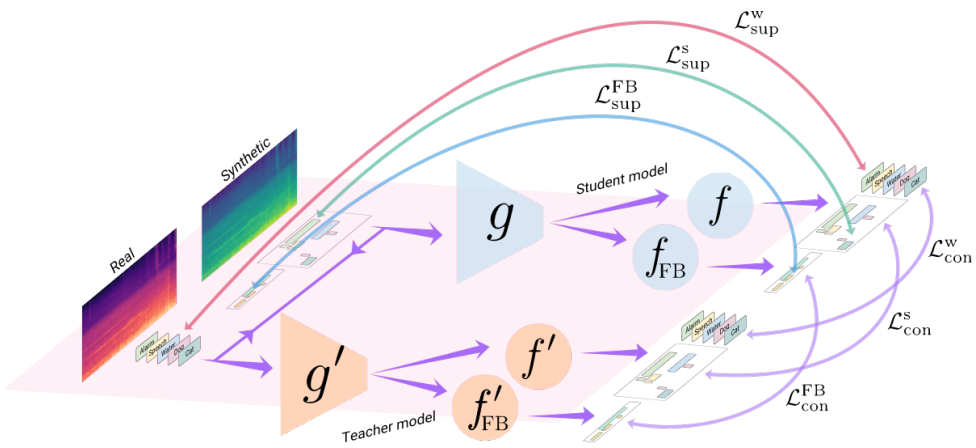


Figure 5.3: Baseline model with added foreground-background classification (FB) branch. $\mathcal{L}^{FB}_{sup}$ corresponds to the FB classification objective, while $\mathcal{L}^{FB}_{con}$ accounts for the consistency of foreground-background classification outputs between the teacher and the student model.

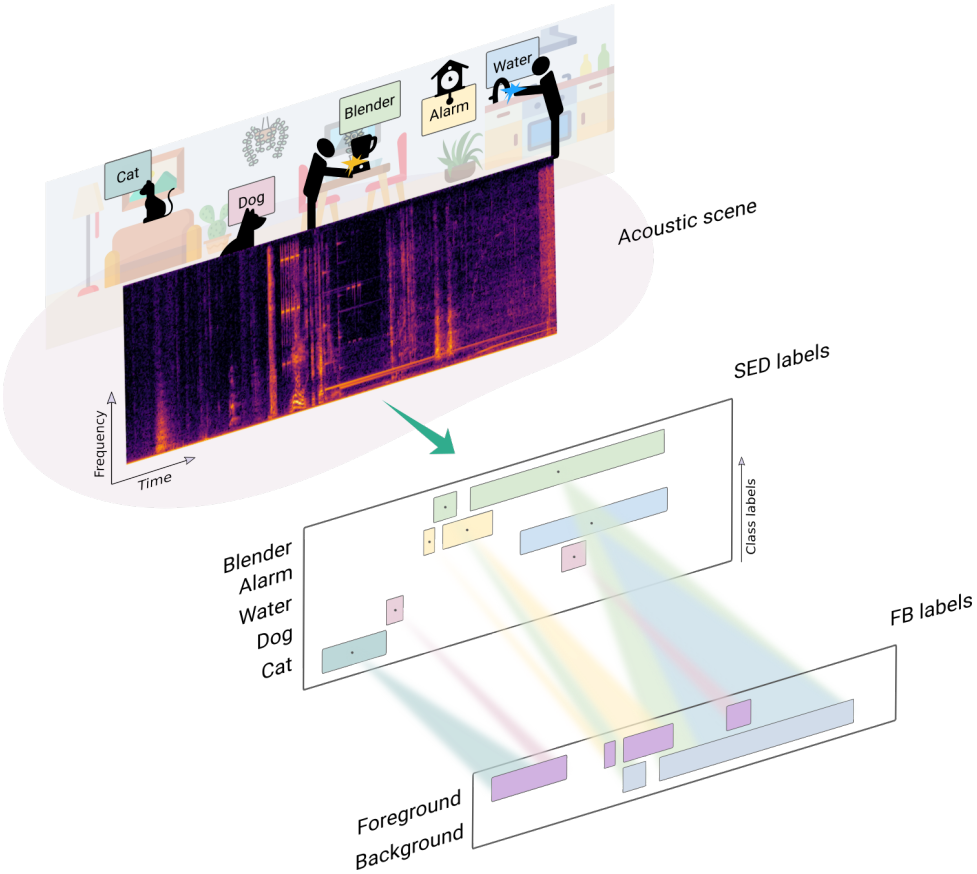


Figure 5.4: Obtaining ground-truth foreground-background labels from ground-truth strong labels. Sound events with rapidly-varying spectro-temporal characteristics (e.g., alarm, speech, dog, cat) are merged into the *foreground* class. Conversely, stationary-like sound events (e.g., blender, water) are merged into the *background* class.

square error consistency costs which are differentiable on their second parameter over strong (frame-level) and weak (clip-level) scores, respectively. The consistency weight λ is tied to all consistency costs.

SEDFB branch

Going beyond the proposed FB classification branch, we explored its fusion with the SED branch into a detection branch (SEDFB) to refine outputs. This model is illustrated in Figure 5.5. The SEDFB branch is represented by a function $f_{\text{SEDFB}} : \mathcal{Y} \times \mathcal{Y}^{\text{FB}} \to \mathcal{Y}$ (f'_{SEDFB} for the teacher model). The input of the SEDFB branch is the outer product of the outputs from the SED and FB branches $w_i = f(g(x_i)) \otimes f_{\text{FB}}(g(x_i))$, as this fusion creates a representation containing information from the joint interaction of the SED and FB classifiers. The following classification-consistency cost pair is added to the training objective in (5.2):

$$\mathcal{L}_{\text{SEDFB}} = L(y_i^{\text{s}}, f_{\text{SEDFB}}(w_i^{\text{s}})) + \lambda L_{\text{strong}}(f_{\text{SEDFB}}(w_i), f'_{\text{SEDFB}}(w_i)). \tag{5.3}$$

The overall cost involving the SEDFB branch is given by

$$\mathcal{L} = \mathcal{L}_{\text{SED}} + \mathcal{L}_{\text{SEDFB}}. \tag{5.4}$$

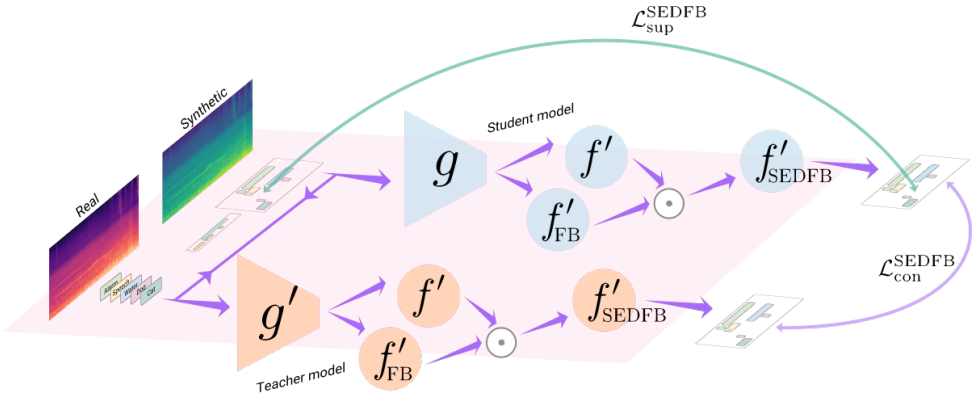


Figure 5.5: Baseline model with the improved sound event detection (SEDFB) branch. It combines the output scores of the FB branch and the frame-level scores of the baseline SED model.

5.1.3 Domain adaptation for sound event detection

From the unsupervised domain adaptation perspective, we regard the synthetic dataset with strong labels as the source domain $\mathcal{S} = \mathcal{D}^{\mathcal{S}}$, and the combination of real recordings from the weakly and unlabeled datasets as the target domain $\mathcal{T} = \mathcal{D}^{\mathcal{W}} \cup \mathcal{D}^{\mathcal{U}}$. We denote as x^{s} and x^{t} the soundscapes from $\mathcal{S}$ and $\mathcal{T}$, respectively.

In contrast to adversarial adaptation approaches that introduce a domain discriminator to reduce the distribution discrepancy between domains (Cornell et al., 2020a; Yang et al., 2020; Park et al., 2019b), our proposed strategy relies on optimal transport for its ability to find correspondences between samples by exploiting the geometry of the underlying space. We adopt the DeepJDOT framework (Damodaran et al., 2018), to correct the mismatch between the distributions of learned feature representations in the two domains.

Joint distribution optimal transport

We recall from Section 2.3.6 the formulation of joint distribution optimal transport. Let $\mu_{\text{s}} = \sum_{i=1}^{n_{\text{s}}} a_i \delta_{g(x_i^{\text{s}}),y_i^{\text{s}}}$ and $\mu_{\text{t}} = \sum_{i=1}^{n_{\text{t}}} b_i \delta_{g(x_i^{\text{t}}),y_i^{\text{t}}}$ be two empirical distributions on the product space $\mathcal{Z} \times \mathcal{Y}$, where $\delta_{g(x_i),y_i}$ is the Dirac function at position $(g(x_i), y_i) \in \mathcal{Z} \times \mathcal{Y}$, and a_i and b_i are (uniform) probability weights, i.e. $\sum_{i=1}^{n_{\text{s}}} a_i = \sum_{i=1}^{n_{\text{t}}} b_i = 1$. The associated cost for moving the i-th source and j-th target element can be expressed as a weighted combination of costs in the latent and label spaces

$$d(g(x_i^{\text{s}}), y_i^{\text{s}}; g(x_j^{\text{t}}), y_j^{\text{t}}) = \alpha c(g(x_i^{\text{s}}), g(x_j^{\text{t}})) + \beta \mathcal{L}(y_i^{\text{s}}, y_j^{\text{t}}) \tag{5.5}$$

where $c(\cdot,\cdot)$ is the squared ℓ_2 distance, $\mathcal{L}(\cdot,\cdot)$ is a cross-entropy loss that enforces regularity between the source and target domain labels, and α and β are two scalar non-negative values. Since no labels y_j^{t} are available in the target domain they are replaced with pseudo-labels $f(g(x_j^{\text{t}}))$ obtained from the classifier $f : \mathcal{Z} \to \mathcal{Y}$. We seek for a transportation coupling $\gamma \in \mathbb{R}^{n_{\text{s}} \times n_{\text{t}}}$ in the space $\Gamma(\mu_{\text{s}}, \mu_{\text{t}})$ of joint probability distributions with marginals $\gamma \mathbf{1}_{n_{\text{t}}} = \mu_{\text{s}}$ and $\gamma^T \mathbf{1}_{n_{\text{s}}} = \mu_{\text{t}}$, where $\mathbf{1}_d$ is a d-dimensional vector of ones, and a pair of mapping functions g and f that minimize

$$\min_{\gamma \in \Gamma(\mu_{\text{s}},\mu_{\text{t}}),g,f} \sum_{i,j} \gamma_{i,j} d(g(x_i^{\text{s}}), y_i^{\text{s}}; g(x_j^{\text{t}}), f(g(x_j^{\text{t}}))). \tag{5.6}$$

We follow a two-step procedure to solve this optimization problem. In the first step, we compute the optimal coupling matrix γ with fixed model parameters f and g,

$$\min_{\gamma \in \Gamma(\mu_s, \mu_t)} \sum_{i,j} \gamma_{ij} (\alpha ||g(x_i^s) - g(x_j^t)||^2 + \beta \mathcal{L}(y_i^s, f(g(x_j^t)))). \tag{5.7}$$

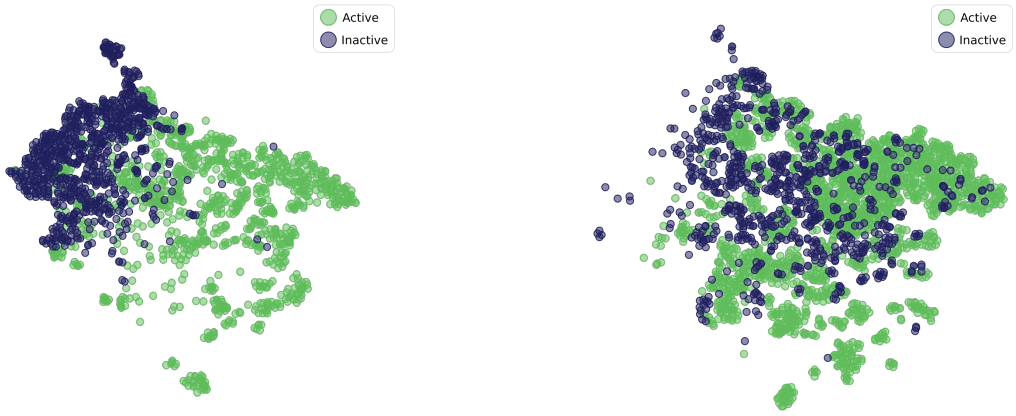
In the second step, with fixed γ, we update the models g and f as

$$\min_{g,f} \mathcal{L}_s + \sum_{i,j} \gamma_{ij} (\alpha ||g(x_i^s) - g(x_j^t)||^2 + \beta \mathcal{L}(y_i^s, f(g(x_j^t)))) \tag{5.8}$$

where $\mathcal{L}_s$ is the classification cost on the source domain to avoid losing performance on synthetic data.

Sampling strategy

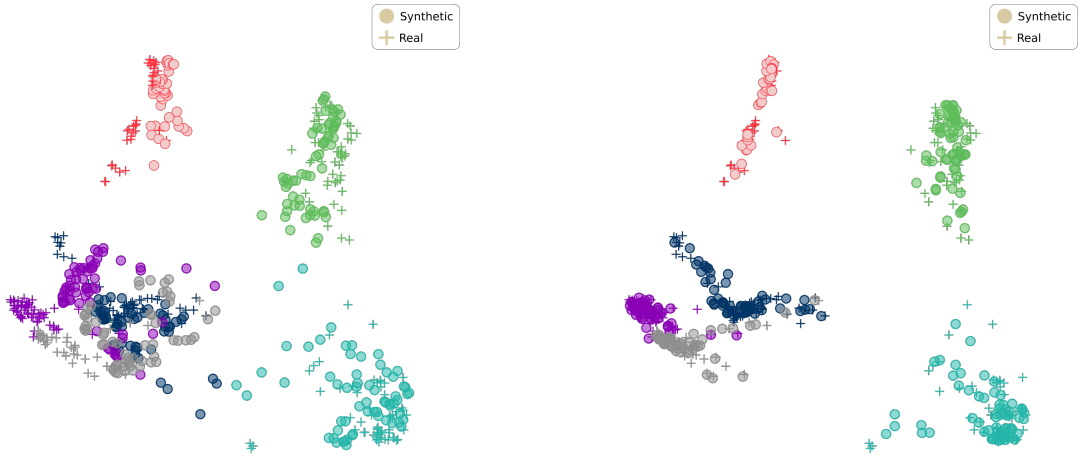
A visual inspection of the audio embeddings learned by the baseline model led us to realize that the semi-supervised framework to train a SED model does not fully account for the mismatch between synthetic and real data. As illustrated in Figure 5.6, the empirical distributions of active and inactive frames in both types of data show a clear discrepancy between them. Therefore, for each data batch we sample all active and inactive frames from the source and target domains as indicated by the strong labels and the pseudo-labels. For both domains we only keep active frames where no sound event overlap occurs, so that the optimal transport takes place between the empirical distributions of the sound classes of both domains. The number of sampled active frames per class can vary considerably from batch to batch for synthetic and real data, which can lead to the absence of certain classes in one type of data for some batches. To account for this imbalance problem we only keep active frames from those classes which are common to both domains, and then balance all classes by resampling them randomly, taking as many elements as there are in the class with fewer elements. This is illustrated in Figure 5.7. Similarly, the number of inactive frames in the source and target domains varies from batch to batch, so we sample randomly each set by taking as many inactive frames as there are in the set with fewer elements.



(a) Distribution of synthetic training data.

(b) Distribution of real evaluation data.

Figure 5.6: t-SNE plot showing the distribution of sound embeddings extracted from a baseline model trained on synthetic and real data with semi-supervised learning.



(a) Sampled synthetic and real sound embeddings. (b) Matched embeddings with optimal transport.

Figure 5.7: t-SNE plot showing the distribution of sound embeddings sampled from a mini-batch of synthetic and real training data (left) and matching of the data with optimal transport (right). This example shows a balanced sampling of active frames from six sound event classes which are common to both domains.

Pseudo-label refinement

To improve the reliability of the pseudo-labels assigned to real data, we leverage the provided annotations of the weakly labeled set to refine pseudo-labels on this subset. The refinement process consists of fusing the frame-level outputs of the SED branch f on soundscapes from $\mathcal{D}^{\mathcal{W}}$ with their clip-level annotations by an element-wise multiplication. The target domain pseudo-labels are thus updated for weakly labeled data as $\hat{y}_j^{\mathrm{t}} = f(g(x_j^{\mathrm{w}})) \odot y_j^{\mathrm{w}}$, $j = 1, \ldots, n_{\mathrm{w}}$. This operation constrains the estimated labels to contain at most the same classes present in the weakly labeled soundscapes. Filtering out all extra classes helps reduce false positives and allows more reliable pseudo-labels to be obtained for the proposed sampling strategy and domain adaptation process.

Training objectives

We denote as $\hat{z}^{\mathrm{s}}$ and $\hat{z}^{\mathrm{t}}$ the sampled active frames and as $\bar{z}^{\mathrm{s}}$ and $\bar{z}^{\mathrm{t}}$ the sampled inactive frames from the source and target domain latent representations z^{s} and z^{t}, respectively. After an initial pre-training stage using (5.4), we construct the following objective function to account for the mismatch between the empirical distributions of active and inactive learned feature representations

$$\mathcal{L}_{\mathrm{s}} + \mathcal{L}_{\mathrm{active}} + \mathcal{L}_{\mathrm{inactive}} \tag{5.9}$$

where

$$\mathcal{L}_{\mathrm{s}} = \frac{1}{n_{\mathrm{s}}} \sum_{i=1}^{n_{\mathrm{s}}} L(y_i^{\mathrm{s}}, f(g(x_i^{\mathrm{s}}))) + \frac{1}{n_{\mathrm{s}}} \sum_{i=1}^{n_{\mathrm{s}}} L(y_i^{\mathrm{fb}}, f_{\mathrm{FB}}(g(x_i^{\mathrm{s}}))) \tag{5.10}$$

corresponds to the first and third classification cost terms of the training classification cost in (5.2). As only the student model undergoes adaptation, no consistency losses are included in the

above objective to train the source domain classifier. The cost function $\mathcal{L}_{\text{active}}$ is the distribution alignment loss for active frames

$$\mathcal{L}_{\text{active}} = \frac{1}{|C_{\text{active}}|} \sum_{i,j}^{N_{\text{active}}} \gamma_{ij}^{\text{active}} (\alpha ||\hat{z}_i^{\text{s}} - \hat{z}_i^{\text{t}}||^2 + \beta \mathcal{L}(y_i^{\text{s}}, \hat{y}_j^{\text{t}})) \tag{5.11}$$

where $|C_{\text{active}}|$ is the cardinality of the subset of labels $C_{\text{active}} \in C$ representing the total number of active classes in the batch. The second term in (5.11) enforces consistency between the target domain pseudo-labels and source domain labels. The cost function $\mathcal{L}_{\text{inactive}}$ accounts for the alignment of the marginal distributions of the learned representations of inactive frames in both domains:

$$\mathcal{L}_{\text{inactive}} = \sum_{i=1}^{N_{\text{inactive}}} \gamma_{ij}^{\text{inactive}} ||\bar{z}_i^{\text{s}} - \bar{z}_i^{\text{t}}||^2. \tag{5.12}$$

Figure 5.8 depicts the proposed frame-level domain adaptation strategy based on optimal transport for the SED task.

5.1.4 Domain adaptation and active learning for sound event detection

The use of pseudo-labels in the domain adaptation approach in Section 5.1.3 limits the correct alignment of class-conditional distributions of synthetic and real data. While we use a refinement method to improve their reliability, they can still cause negative transfer that misguides the adaptation process. To mitigate this issue, we explore an active learning strategy that replaces pseudo-labels with annotations prompted to the user (simulated by an oracle). The goal is to further reduce the mismatch between synthetic and real data. At each iteration of the process, a set of audio segments are selected for labeling and the SED model is adapted on ground-truth data. We note that current works on active learning consider semi-supervised learning or domain adaptation as a separate problem and their optimal combination hasn't been explored.

The active learning strategy is based on the *mismatch-first farthest-traversal* method proposed by Shuyang et al. (2020). Mismatch-first looks for samples with mismatched labels according to two different labeling systems and farthest-traversal maximizes the diversity of the selected samples. The three main steps involved in this active learning strategy are: soundscape segmentation, segment selection for annotation and weakly supervised learning. In our setting, rather than weakly supervised learning we perform segment-level domain adaptation.

Soundscape segmentation

Audio segmentation is carried out with change point detection. The likelihood of a change point is measured at each time frame by the cosine distance between the means of the past and the future M frames. Empirically, we set $M = 8$ frames corresponding to 0.128 ms. Peaks in the likelihood (between 0 and 1) are selected as change points. We selected peaks above 0.5. Figure 5.9 illustrates the change point selection method applied to the input log-Mel spectrogram, to the output of the CNN model and to the output of the RNN model. Recent works in the speech domain have relied on learned representations for audio similarity analysis tasks (Zhang and Duan, 2017, 2016). In the same way, a visual inspection of the segmentation method on CNN and RNN embeddings led us to choose RNN embeddings to segment the soundscapes.

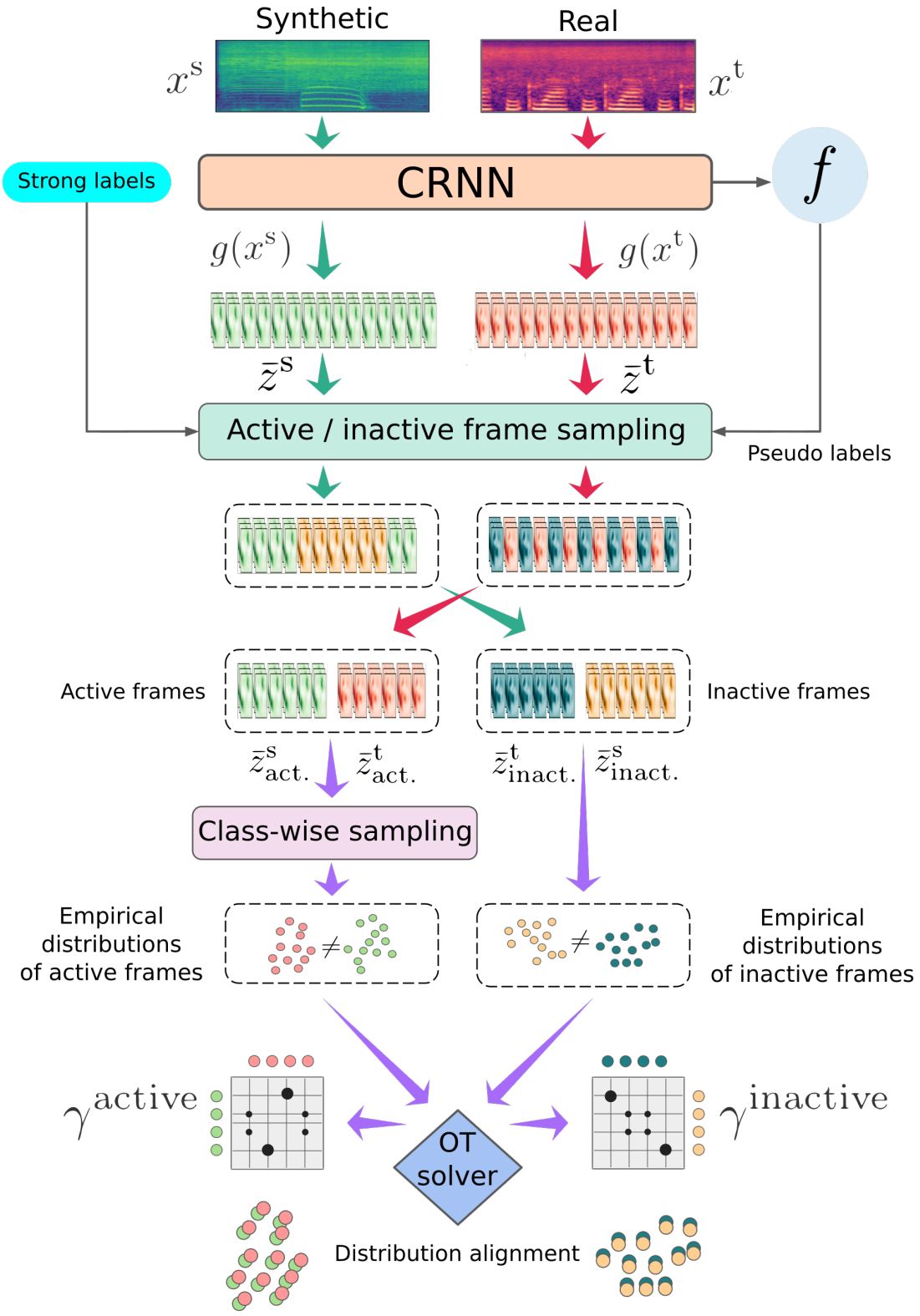


Figure 5.8: Proposed domain adaptation strategy based on optimal transport to correct the mismatch between synthetic and real data.

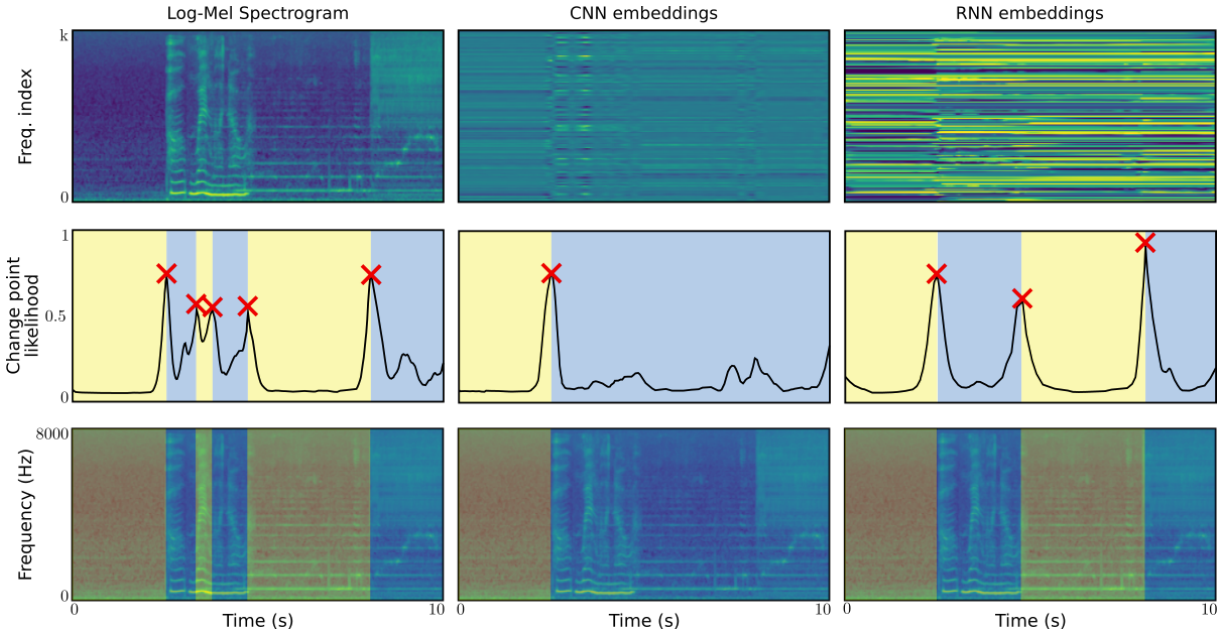


Figure 5.9: Soundscape segmentation based on change point detection on the input log-Mel spectrogram (left column), CNN embeddings (middle column) and RNN embeddings (right column). Peaks in the likelihood above 0.5 correspond to change points as marked by the red crosses.

Sample selection strategy

The sample selection strategy comprises two criteria illustrated in Figure 5.10 and it is described as follows.

Mismatch-first criterion All sampled segments for domain adaptation are assigned two labels: pseudo-labels generated by the pretrained SED model and a propagated label generated with nearest neighbor prediction. The model pseudo-labels are regarded as $\mathcal{A}_x$ and the propagated labels as $\mathcal{B}_x$. The mismatch between the two sets of labels is measured with the Jaccard Index as

$$J(\mathcal{A}_x, \mathcal{B}_x) = \begin{cases} \frac{|\mathcal{A}_x \cap \mathcal{B}_x|}{|\mathcal{A}_x \cup \mathcal{B}_x|} & \text{if } \mathcal{A}_x \cup \mathcal{B}_x \neq 0 \\ 1 & \text{if } \mathcal{A}_x \cup \mathcal{B}_x = 0. \end{cases} \tag{5.13}$$

The samples are ranked by the lowest index, mismatch first. Samples with similar scores are further ranked by the second selection criterion.

Farthest-first criterion It aims at selecting diverse samples after the mismatch-first primary criterion. A farthest-traversal search is performed on the audio segments with the lowest Jaccard index. Samples are ranked by the cosine distance to the previously selected samples, farthest first. The selected samples are then presented to the annotator which provides strong labels.

Segment-level domain adaptation

We extend the proposed domain adaptation method for sound event detection based on the DeepJDOT framework to operate on the sound segments from the segmentation step. We thus extend frame-level domain adaptation to segment-level domain adaption. Figure 5.11 shows the modification to the original setup. A sound segment is represented by the mean of its frame embeddings and the domain adaptation method proceeds with the training objective in Equation

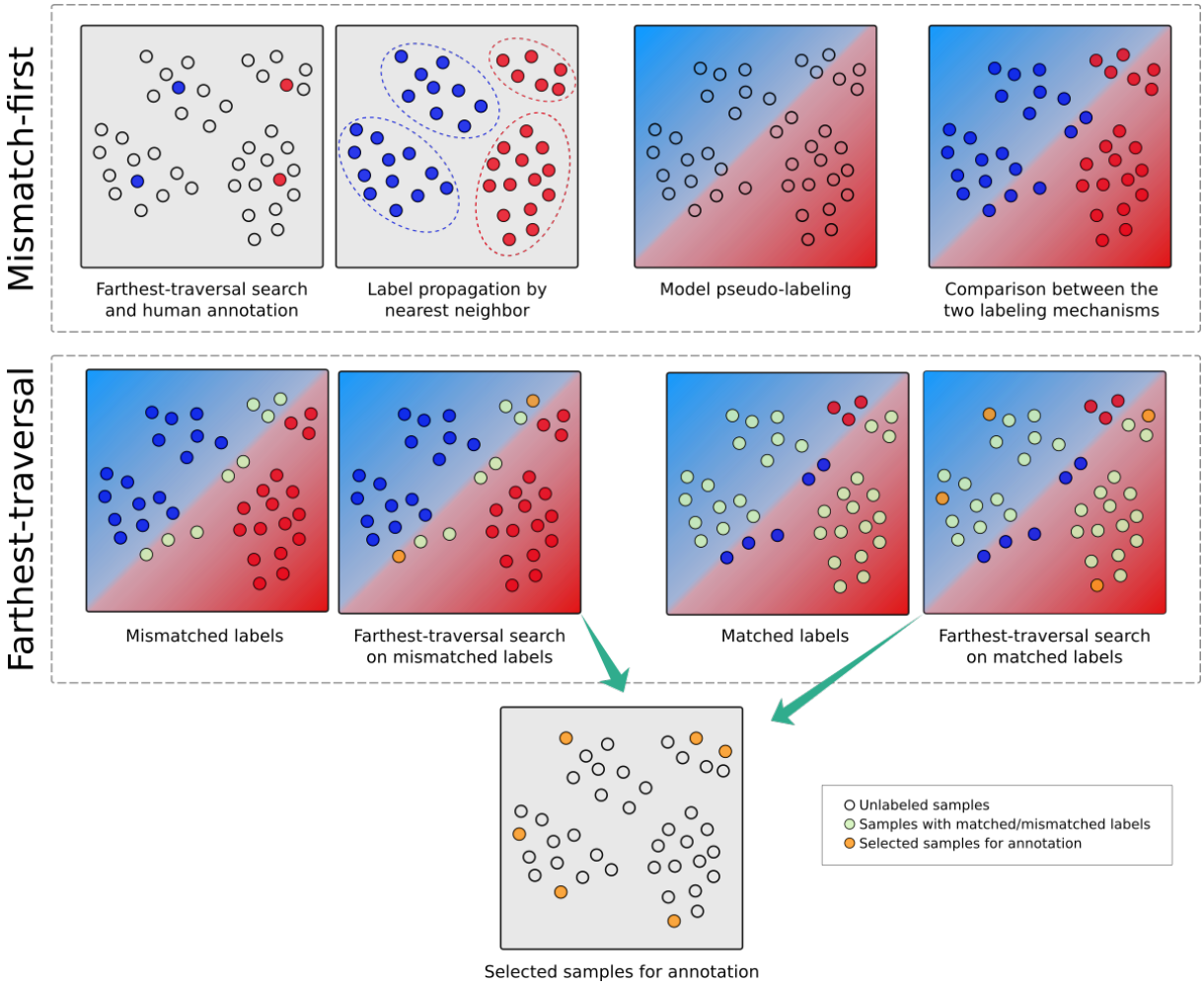


Figure 5.10: Mismatch-first farthest-traversal strategy for sample selection.

(5.9). Instead of performing active/inactive frame-wise sampling, we thus perform active/inactive segment-level sampling for domain adaptation. Active segments correspond to audio segments in which a sound event occurs (as indicated by the ground-truth or pseudo-label), while inactive segments correspond to audio segments lacking annotations. The alignment between the active and inactive empirical distributions of sound segments follows the same procedure as for frame-level adaptation.

5.1.5 Model

The selected model architecture is the same as the baseline system of DCASE 2020. It is illustrated in Figure 5.12. The CNN part is composed of 7 layers with 16, 32, 64, 128, 28, 128, and 128 filters, respectively. A kernel of size 3x3 was used with max-pooling [2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2] and [1, 2], respectively. A gated linear unit activation is applied to the convolution operations. The RNN part is composed of 2 layers of 128 bidirectional gated recurrent units. The outputs of the overall CRNN feed two branches. The first branch is composed of a dense layer with sigmoid activation and outputs frame-level (strong) class-wise posteriors. The second branch comprises a dense layer with softmax activation. The outputs of this branch are combined

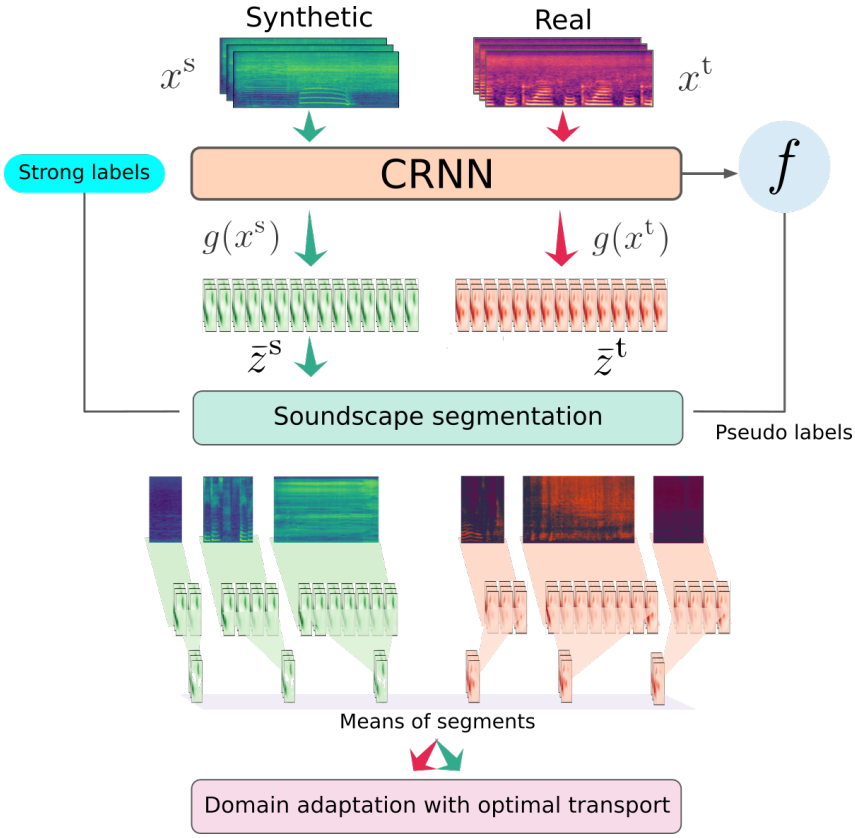


Figure 5.11: Segment-level sampling for domain adaptation.

with those of the first branch and then passed to an attention pooling block performing mean temporal aggregation to produce clip-level (weak) scores. The FB branch consists of a dense layer with sigmoid activation, which acts upon the outputs of the RNN block. The SEDFB branch is composed of a bidirectional RNN with 128 gated recurrent units and a dense layer with sigmoid activation.

5.2 Experimental setup

5.2.1 Dataset

We conducted experiments on the Domestic Environment Sound Event Detection Dataset (DESED) dataset (Turpault et al., 2019; Serizel et al., 2020), composed of a training set of 2,584 synthetic audio clips generated by Scaper (Salamon et al., 2017), 1,578 real soundscapes with clip-level annotations and 14,412 unlabeled real recordings. We evaluate the performance of the proposed methods on the validation and public evaluation sets, comprising 1,168 and 692 clips, respectively.

5.2.2 Training hyper-parameters

In the training stage, the model was trained for 200 epochs with the Adam optimizer, a dropout value of 0.5, and a gradually increasing learning rate with a max value of 10^{-3}. The consistency weight λ was set to 1. In the adaptation stage, the student model was adapted for 300 epochs. We used cost weights $\alpha = 0.2$, $\beta = 5.0$, and the contribution of the source domain classifier

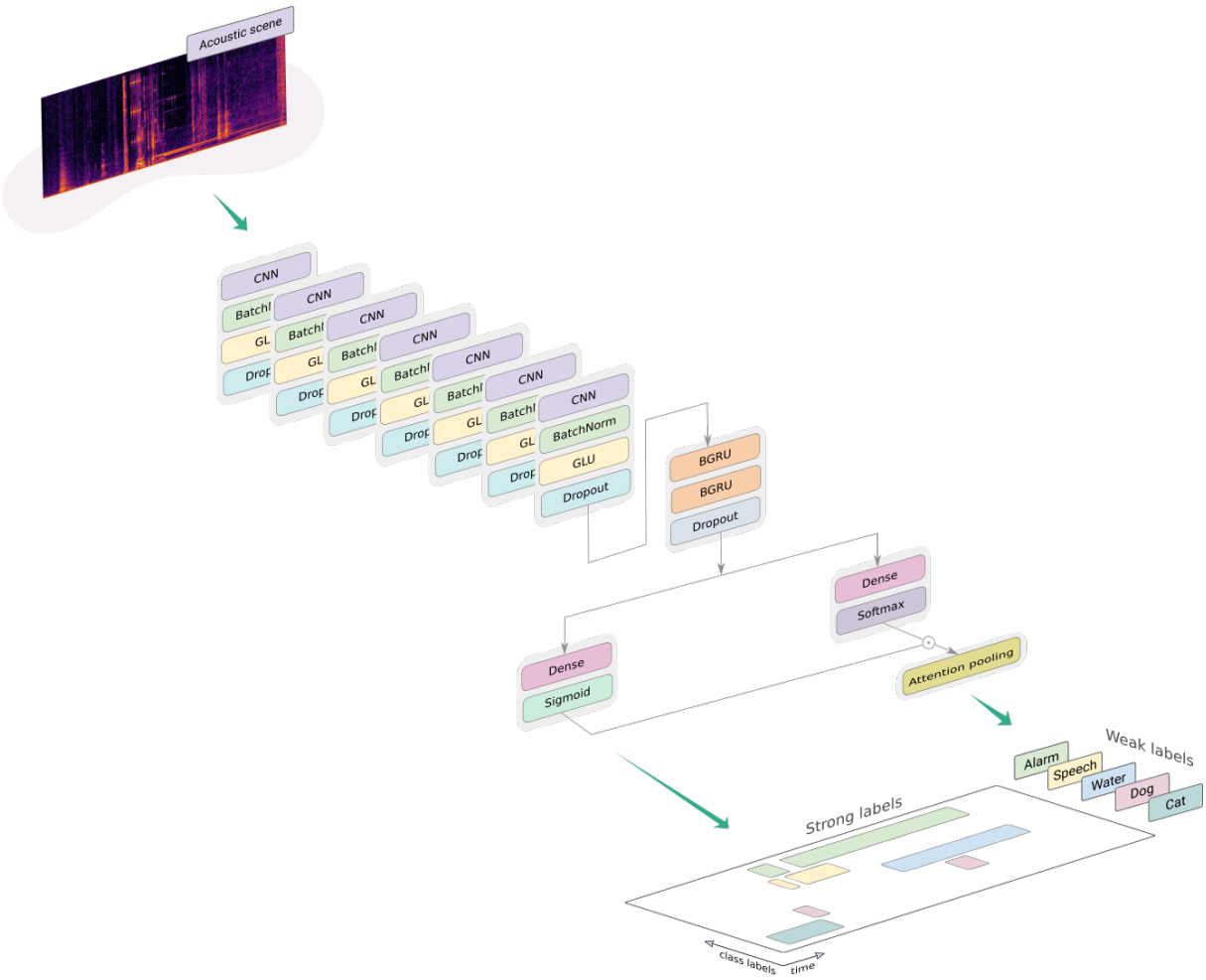


Figure 5.12: Network architecture of the baseline model.

cost $\mathcal{L}_s$ to the total adaptation cost was multiplied by 100. The learning rate was fixed to 10^{-4}. Experiments with optimal transport were performed using the Python Optimal Transport package (Flamary et al., 2021).

Output post-processing

We used two methods to post-process the SED frame-level scores. The first method corresponds to smoothing the frame-level scores with a median filter of 0.45 s. The second approach consists of Hidden Markov Model (HMM) decoding. Following the same procedure as Cornell et al. (2020a), we determined the optimal transition probabilities for each sound event class using the validation set. We contrast the contribution of both post-processing schemes in Section 5.3.

5.3 Results

5.3.1 Improving SED with learnable PCEN as acoustic front-end

In Table 5.1 we report the performance achieved on the development set by a baseline model trained on log-Mel spectrograms as well as on different PCEN-based acoustic front-ends. We

measure the performance in terms of the event-based macro F1 score. We report the best performance achieved on the validation set. We first analyze the results of PCEN transformations with fixed parameters. *Fixed PCEN 1* uses parameters $\epsilon = 10^{-6}$, $\alpha = 0.98$, $\delta = 2$ and $T = 400$ ms. This setup filters stationary sound events yielding a sub-optimal performance. *Fixed PCEN 2* uses the same parameters except for α, which is set to 0.8. This setup no longer suppresses sounds with slowly-changing spectral features and its performance is comparable to the log-Mel acoustic front-end. *Fixed PCEN 3* uses the same configuration, except that the time constant is set to 60 ms. This configuration favors all sound classes in the task and brings a modest improvement over log-Mel. Regarding the trainable version of PCEN, we observe that learning a single PCEN transformation (*Trainable PCEN*) jointly with the sound event detection objective brings a modest improvement over *Fixed PCEN 3*. Training a second layer in parallel (*P-PCEN on Mel* spectrograms) brings a little more than 1% absolute improvement over learning the parameters of a single PCEN transformation. Lastly, a similar setup with log-Mel spectrograms as input features (*P-PCEN on log-Mel*) outperforms the log-Mel front-end by 3.41% absolute. We found that learning the parameters of two PCEN transformations is sufficient to give significant performance improvement over the common log-Mel acoustic front-end and that adding more PCEN transformations does not bring further improvement to the detection scores. Figure 5.13 shows the output of the proposed 2-layer P-PCEN front-end for an input soundscape. We observe that the first PCEN transformation filters stationary sounds to stand out fast transient sounds, while the second layer limits the filtering operation to avoid degradation of slowly-changing background sounds.

Table 5.1: Performance on the validation set per acoustic front-end.

Acoustic front-end	F1 score (%)
Log-Mel	35.93
Fixed PCEN 1	28.71
Fixed PCEN 2	35.42
Fixed PCEN 3	36.84
Trainable PCEN	36.92
P-PCEN on Mel	**38.20**
P-PCEN on log-Mel	**39.34**

5.3.2 Improving SED with domain adaptation

In Table 5.2 we show the best results achieved on the evaluation sets by our proposed optimal transport-based domain adaptation strategy. Adapting the SED model by aligning the empirical distributions of active frames of synthetic and real data brings only a modest absolute improvement of 2.1% over the the Baseline performance. The opposite effect is achieved when only adapting the empirical distributions of inactive frames. Given the large imbalance between active and inactive frames, the adaptation process drags the distribution of active frames towards the distribution of inactive frames. This causes the SED system to experience severe performance degradation and lose around 10 percentage points. The best result is achieved by aligning the empirical distributions of both active and inactive frames. In this setting a performance boost of 7.3% absolute was achieved over the Baseline performance, showing that is equally important to align both the empirical distributions of active and inactive frames. The last three rows of Table 5.2 correspond to oracle adaptation, i.e., adaptation with active and inactive frames sampled

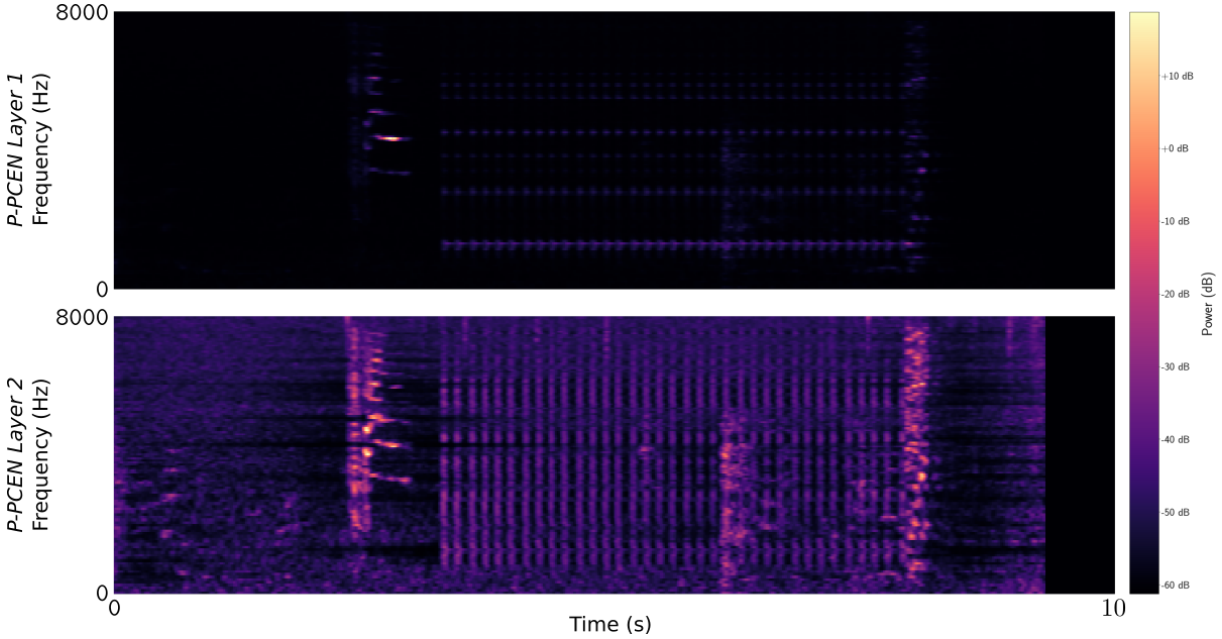
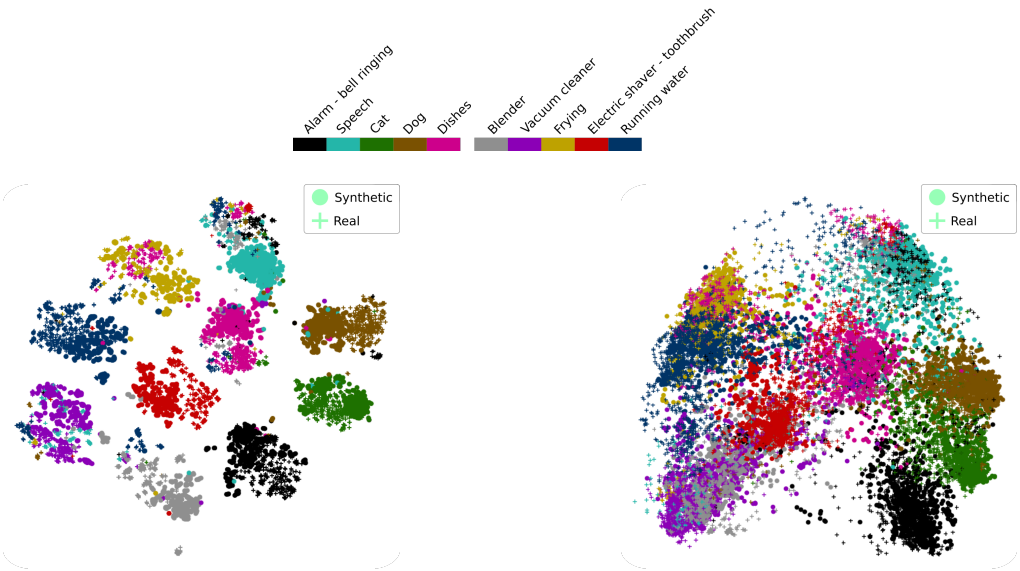


Figure 5.13: Output PCEN spectrograms from the proposed trainable *P-PCEN* transformation. The input soundscape is composed of an alarm co-occurring with the sound of a vacuum cleaner. The first PCEN transformation filters the vacuum cleaner and isolates to some extent the sound of the alarm, while the second transformation limits the filtering operation on the vacuum cleaner to avoid its degradation.

with ground-truth labels. Results confirm the need of active and inactive frame alignment to reduce the mismatch between synthetic and real data. Figures 5.14, 5.15 and 5.16 show the learned embedding space of a SED model adapted using active, inactive and active and inactive frames, respectively. These figures serve to compare the mismatched embedding space of the Baseline model with the matched embedding space of the proposed domain adaptation strategies. The figures also show a visualization of an embedding space invariant to synthetic and real data as a reference for oracle adaptation.

Table 5.2: Performance comparison per adaptation method in the validation set.

Adaptation method	F1 score (%)
Baseline	34.80
Active frames only	36.93
Inactive frames only	24.17
Active + inactive frames	**42.41**
Active frames only (oracle)	50.66
Inactive frames only (oracle)	16.34
Active + inactive frames (oracle)	70.23



(a) t-SNE visualization of the embedding space learned by the Baseline model.

(b) PCA visualization of the embedding space learned by the Baseline model.

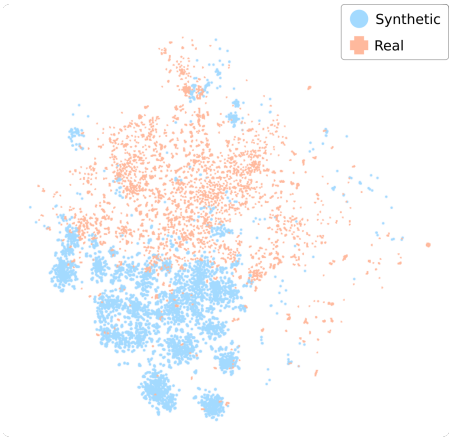
(c) t-SNE visualization of the embedding space learned by the best adapted model.

(d) PCA visualization of the embedding space learned by the best adapted model.

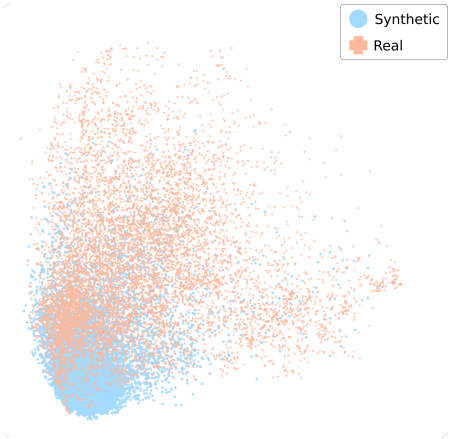
(e) t-SNE visualization of the embedding space learned by the oracle adapted model.

(f) PCA visualization of the embedding space learned by the oracle adapted model.

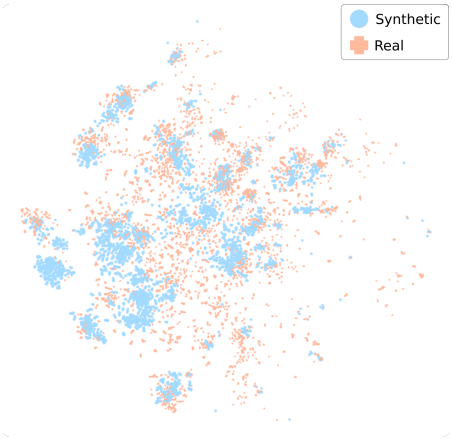
Figure 5.14: t-SNE (left column) and PCA (right column) visualizations of the distribution of embeddings computed by the Baseline model (top row), by the best model adapted using **active** frames only (middle row) and by the oracle model using **active** frames only and performing adaptation with ground-truth data (bottom row).



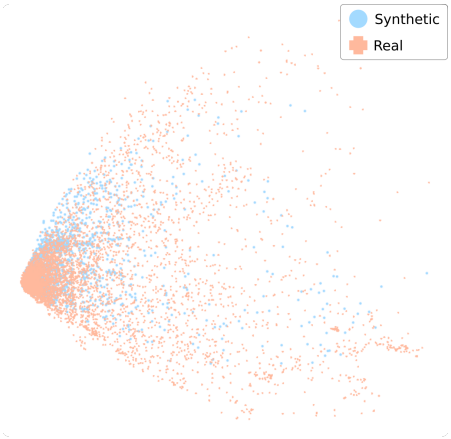
(a) t-SNE visualization of the embedding space learned by the Baseline model.



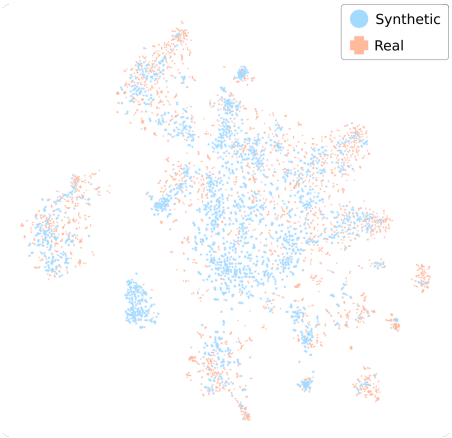
(b) PCA visualization of the embedding space learned by the Baseline model.



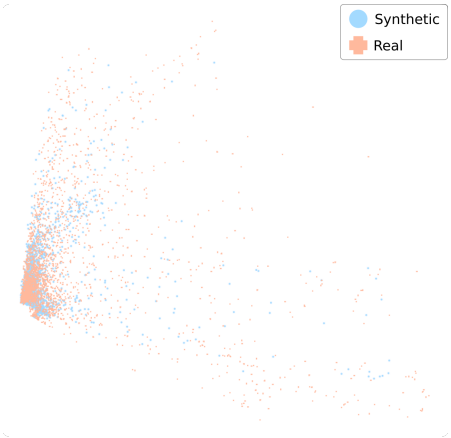
(c) t-SNE visualization of the embedding space learned by the best adapted model.



(d) PCA visualization of the embedding space learned by the best adapted model.

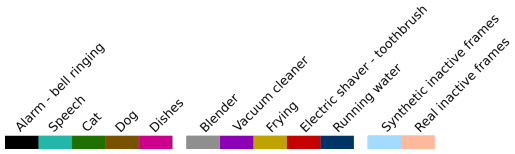


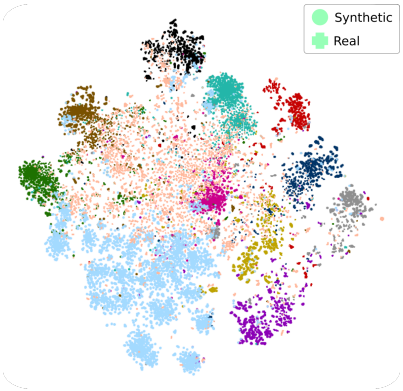
(e) t-SNE visualization of the embedding space learned by oracle adapted model.



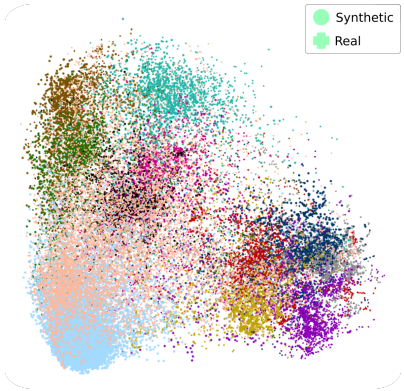
(f) PCA visualization of the embedding space learned by oracle adapted model.

Figure 5.15: t-SNE (left column) and PCA (right column) visualizations of the distribution of learned embeddings by the Baseline model (top row), by the best model adapted using **inactive** frames only (middle row) and by the oracle model using **inactive** frames only and performing adaptation with ground-truth data (bottom row). In this setting, the performance of the best model is below the baseline model.

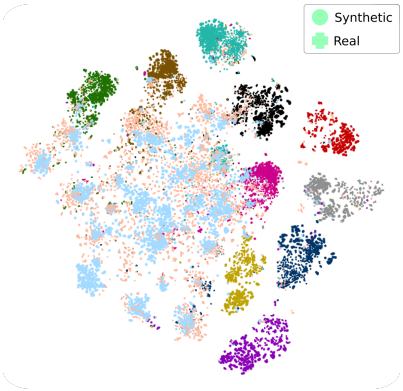




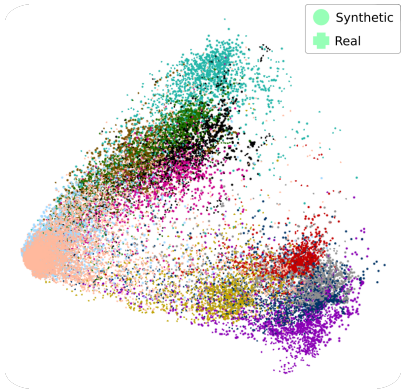
(a) t-SNE visualization of the embedding space learned by the Baseline model.



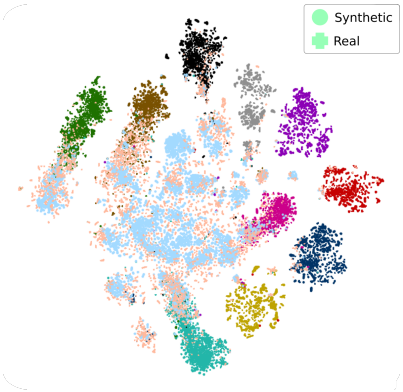
(b) PCA visualization of the embedding space learned by the Baseline model.



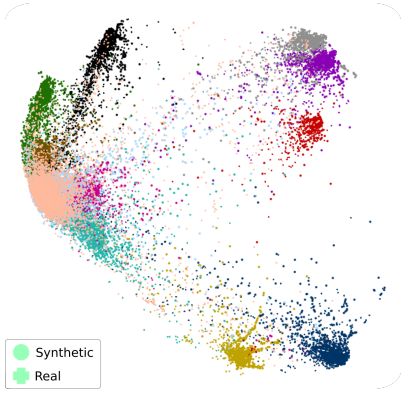
(c) t-SNE visualization of the embedding space learned by the best adapted model.



(d) PCA visualization of the embedding space learned by the best adapted model.



(e) t-SNE visualization of the embedding space learned by oracle adapted model.



(f) PCA visualization of the embedding space learned by oracle adapted model.

Figure 5.16: t-SNE (left column) and PCA (right column) visualizations of the distribution of embeddings computed by the Baseline model (top row), by the best model adapted using **active** and **inactive** frames (middle row) and by the oracle model using **active** and **inactive** frames and performing adaptation with ground-truth data (bottom row).

5.3.3 Improving SED with foreground-background classification and domain adaptation

In Table 5.3 we compare the best results obtained by the proposed methods on the validation and public evaluation sets of the DESED dataset in terms of the event-based macro F1 score. The model labeled as Baseline corresponds to the baseline system of the DCASE 2020 Challenge Task 4. For each evaluation set we show the performance with median filtering (+MF) or HMM smoothing (+HMM) based post-processing.

Table 5.3: Performance on the validation (val) and public evaluation (eval) sets.

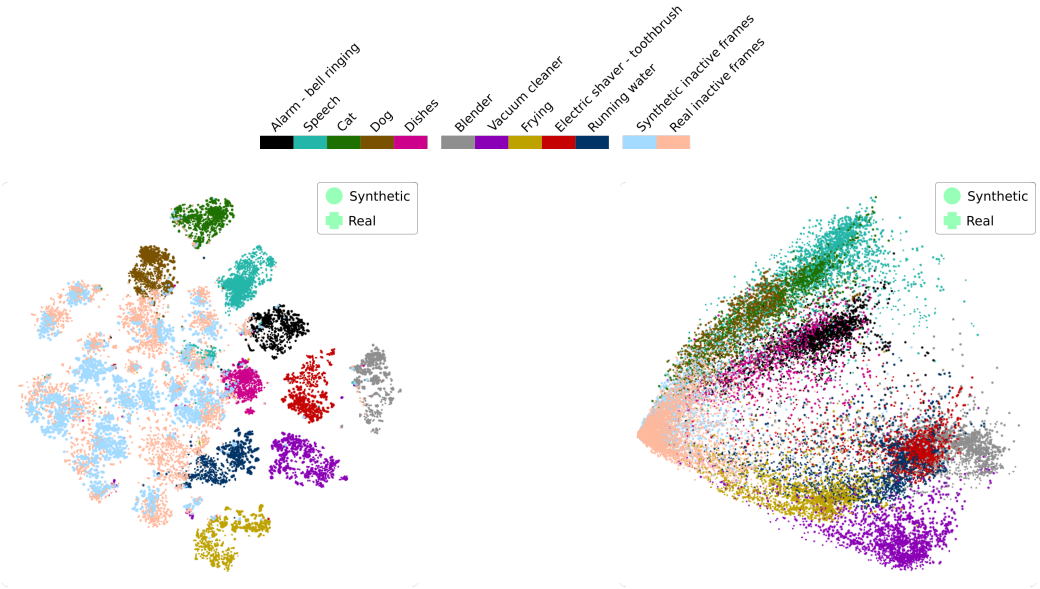
Method	F1 score (%)			
	val +MF	val +HMMs	eval +MF	eval +HMMs
Baseline	34.80	37.60	38.1	41.1
+ DA	42.41	43.89	44.8	47.12
+ FB	43.12	45.42	46.06	49.38
+ FB + DA	45.68	47.77	50.79	53.10
+ SEDFB	46.15	46.20	48.40	49.79
+ SEDFB + DA	**47.61**	47.75	**52.12**	53.30
DCASE 1	45.13	**48.07**	50.58	**53.35**
DCASE 2	45.15	47.08	50.28	52.23

Effects of the auxiliary FB branch and improved SEDFB branch

Adding the FB branch is beneficial to the SED task as it improved results on the validation set compared to the Baseline by 8.3% absolute. This large performance improvement validates the use of auxiliary information about the spectro-temporal nature of the sound events considered in domestic scenarios. Further improvement was achieved by refining the output scores with HMM smoothing, which boosted performance by 10.6% absolute. Moreover, the combination of the FB branch with the SED branch into the SEDFB branch brought an additional gain of 3% absolute with median filtering and 0.8% absolute with HMM smoothing. A similar improvement is observed in the public evaluation set.

Combining foreground-background auxiliary information and domain adaptation

As discussed in the previous section, model adaptation with the proposed optimal transport strategy (DA) improved the Baseline performance by 7.6% absolute as a standalone method. This setting corresponds to the distribution alignment of active and inactive frames of synthetic and real data. The PCA plots of the audio embeddings extracted from an adapted model (Figure 5.16) reveal a gap between foreground and background sounds, which increases when DA is combined with the FB and SEDFB branches, favoring sound event detection. This effect is illustrated in the PCA plot in Figure 5.17. The overall improvement over the Baseline performance was 10.8% and 12.8% absolute, respectively. These results prove the effectiveness of the system in further reducing the mismatch between synthetic and real data. We note that the performance achieved on the public set was about 2% absolute larger compared to the validation set, which might be due to the fact that the empirical distributions of the active and inactive frames in this set are more similar to those of the provided real training data on which adaptation was carried out.



(a) t-SNE plot of learned embeddings.

(b) PCA plot of learned embeddings.

Figure 5.17: t-SNE (left column) and PCA (right column) visualizations of the distribution of the active and inactive frame embeddings learned by the *SEDFB+DA* model.

Effects of output post-processing

HMM smoothing as a post-processing method yielded greater improvement to the detection scores over median filtering for all proposed models. For the SEDFB-based methods, +MF and +HMM provided similar scores, which implies that the SEDFB branch plays a similar role as HMM decoding in the modeling of time-varying spectra of sound events. Since in this setting the SEDFB branch produces more robust detection scores, a less complex post-processing method such as median filtering suffices to refine the outputs.

Participation in the DCASE 2021 Challenge

DCASE 1 and DCASE 2 are model ensembles comprising three and two + FB + DA systems from different training runs, respectively. Model ensembling is achieved by averaging the model outputs. These models correspond to the submissions made to the DCASE 2021 Challenge Task 4 and are labeled as *Olvera_INRIA_task4_SED_1* and *Olvera_INRIA_task4_SED_2*, respectively. Both systems showed competitive performance (ranked 11th and 14th, respectively) in terms of the event-based macro F1-score on the evaluation and public evaluation sets among 65 systems.

5.3.4 Improving SED with active learning

Figure 5.18 shows the performance of the proposed segment-level domain adaptation and active learning framework as a function of the labeling budget. Results show performance on the validation and public evaluation sets.

The recent availability of strong labels for a subset of unlabeled or weakly labeled recordings from Audioset (Hershey et al., 2021) helped us assessed the impact of incorporating ground-truth labels for real data in the training recipe of SED models. The second row of Table 5.4

Table 5.4: Performance of baseline models and domain adaptation approaches on the validation and public evaluation sets.

Method	F1 score (%)	
	validation set	public evaluation set
Baseline	34.80	37.6
Baseline + audioset	40.0	42.6
+ DA frame-level	42.41	44.8
+ DA segment-level	**43.26**	**45.08**

shows the performance of the Baseline model trained with such labels (Baseline + audioset). While the use of strong labels has a positive impact on the SED model, the overall improvement is limited compared to the methods presented in the previous section, since its performance is below domain adaptation with pseudo labels or the use of auxiliary classification branches. We however benefit from the availability of strong labels for real training data to explore the proposed domain adaptation framework with an active learning strategy. We aim to further reduce the mismatch between synthetic and real data by prompting the user (oracle labeling) for a fixed labeling budget at each iteration of the adaptation process. Figure 5.18 shows the performance of the extension of the domain adaptation framework to perform segment-level adaptation with the mismatch-first farthest-traversal active learning strategy.

We first observe that the extension of domain adaptation from frame-level to segment-level leads to marginally better detection scores (3rd and 4th rows of Table 5.4), achieving an extra 0.85% absolute on the validation set and only 0.28% absolute in the public evaluation set.

We now discuss the performance of domain adaptation at the segment-level with active learning (DA + AL), compared with the same adaptation strategy and a random-sampling active learning strategy (DA + RS). In the latter setting, random audio segments are prompted to the user to label, such that the selection of the least reliable and sufficiently diverse samples for the SED model is not guaranteed.

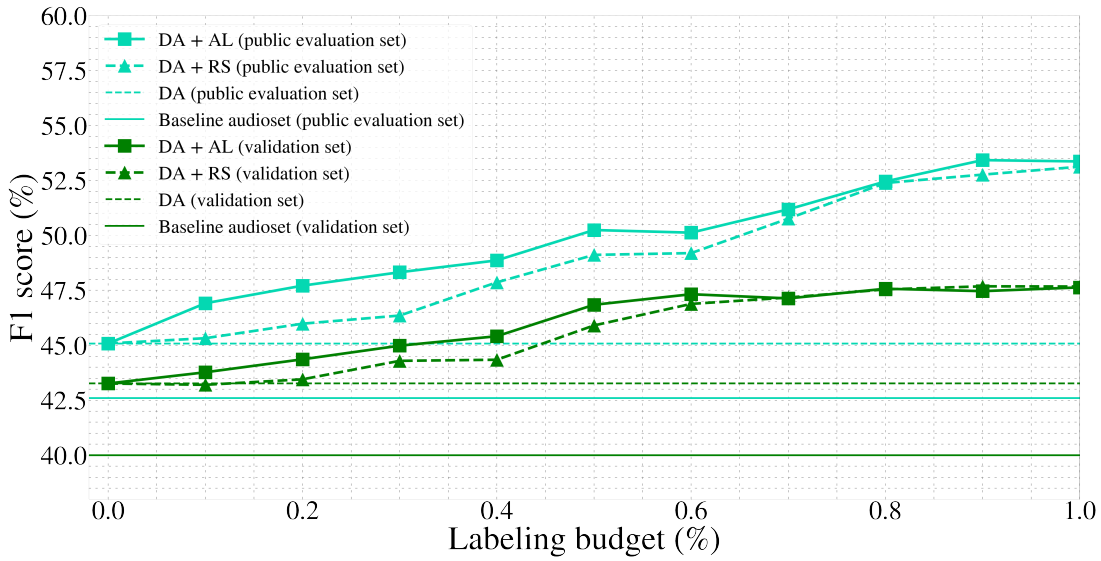


Figure 5.18: Performance of segment-level domain adaptation and active learning as a function of the labeling budget. The labeling budget is relative to the amount of ground-truth annotations available for real training data (around 10% of the data).

From Figure 5.18, DA + AL outperforms DA + RA for small labeling budgets by a small margin, while performance is comparable for larger labeling budgets. The performance of active learning strategies with limited labeling budgets is particularly interesting since the goal is to minimize the labeling effort. We note however that little improvement is achieved in such a setting. Only a 3.58% absolute improvement (compared to +DA segment-level) was achieved on the validation set and a 5.16% absolute improvement was achieved on the public evaluation set when using 50% of the available ground-truth labels for real data. Nevertheless, these results are promising considering that only around 10% of the real training data of the DESED dataset benefited from the release of strong annotations in Audioset. Once again, we point out that the optimal transport-based domain adaptation strategy has a larger benefit on the public evaluation set, implying that the distribution of real training data better fits this distribution.

5.4 Conclusion

In this chapter we presented several complementary methods that improve sound event detection models trained on synthetic and real data. We showed that learning the parameters of multiple PCEN transformations outperforms log-Mel spectrograms as input features. In addition, categorizing domestic sounds according to their spectro-temporal characteristics as foreground or background events is helpful auxiliary information that improves generalization. Moreover, we investigated an explicit domain adaptation strategy based on optimal transport to further reduce the mismatch between synthetic and real data. Results indicate that matching the empirical distributions of active and inactive frames and audio segments of both types of data during training is beneficial for SED. We showed that foreground-background classification and active learning strategies combined with domain adaptation substantially improves detection robustness.

6

Conclusions

This thesis addressed problems that audio analysis systems often encounter when deployed in real environments, in particular, robustness to overlapping sounds and robustness to sources of mismatch such as the use of different recording devices and the use of synthetic and real data. This led us to touch on several tasks, including audio source separation and domain adaptation for acoustic scene classification and sound event detection. In this chapter, we conclude the thesis with a summary of the proposed methods and contributions in Section 6.1, and present future research directions in Section 6.2.

6.1 Summary

In **Chapter** 3 we introduced the foreground-background ambient sound separation task. In this task of practical interest, we performed a series of experiments to investigate whether a deep learning-based source separation model can discriminate between short sound events with rapidly evolving spectral characteristics (i.e., foreground sounds) and long duration sounds with slowly varying spectro-temporal structure (i.e., background sounds). We conducted experiments on simulated foreground-background mixtures of isolated sounds from the DESED and Audioset datasets. We aimed to assess the ability of the proposed separation model to generalize to unseen sound classes whose frequency content exhibit this categorization, which is particularly important in scenarios where label shift occurs. We relied on a scheme comprising a main mask estimation network and analyzed the effects of feeding an auxiliary network that summarizes the statistics of the background sound to ease the task. Within our separation framework, we investigated the impact of log-Mel and PCEN spectrograms as input features. Under the proposed experimental protocol, we found that the use of an auxiliary network in the mask estimation process is detrimental, while the use of PCEN as an acoustic front-end for the task is more suitable than log-Mel spectrograms. This suggests that PCEN's properties can serve as a local strategy for domain adaptation that provides robustness to label shift. The objective evaluation of the proposed models in terms of the signal-to-distortion ratio showed the generalization ability of the proposed foreground-background separation model on mixtures of seen and unseen foreground and/or background sound events.

In **Chapter** 4, we addressed unsupervised domain adaptation for acoustic scene classification with mismatched recording devices. We investigated linear feature-based adaptation and its integration with non-linear learning-based domain adaptation to compensate for the bias induced by the microphone's frequency response. To address domain adaptation in a blind setting with

respect to the target domain, we did not assume the availability of simultaneous recordings captured by training and test microphones, since practical settings demand pretrained models to be deployed on devices with unknown microphone responses. We thoroughly analyzed the impact of first- and second- order moment normalization and moment matching strategies and their combination in all stages of the classification pipeline of an adversarial domain adaptation framework. This framework, however, only guarantees the marginal alignment of the source and target data distributions, thus failing to entirely capture the relationship between feature representations and class posteriors, which is crucial to domain adaptation. To overcome this limitation we extended the adversarial adaptation formulation to a conditional formulation to fully match the source and target data distributions. Our experimental results showed that adaptation strategies following this improvement achieved higher classification rates. Moreover, our results showed that first- and second-order moment normalization has the effect of removing the linear distortions introduced by the microphone, while moment matching effectively transfers these distortions from one dataset to another. Moreover, when combined with adversarial domain adaptation, the remaining mismatch due non-linear effects is reduced. This integration was observed to largely reduce the impact of mismatch when moment normalization and/or matching strategies are applied at all stages of the classification pipeline, i.e., during pretraining, adaptation and inference. We found that in such a setting, the performance achieved in the target domain is comparable to that in matched conditions. Our comprehensive study shows the individual scope and limitations of the proposed adaptation strategies and serves as design choices to counteract the performance degradation experienced by ASC with mismatched recording devices.

In **Chapter** 5, we proposed several complementary methods that improve sound event detection models trained on both synthetic and real data. In a first stage, we explored a trainable acoustic front-end based on PCEN to boost detection scores. We showed that learning the parameters of multiple PCEN transformations outperforms common log-Mel spectrograms and PCEN transformations with fixed parameters. The flexible dynamic compression of parallel PCEN transformations was shown to favor domestic sounds compared to fixed transformations that cannot highlight both transient sounds and stationary sounds of interest. In a second stage, we categorized domestic sounds as foreground or background events according to their spectro-temporal characteristics and investigated the use of this categorization as auxiliary information to improve detection scores. We derived foreground-background ground-truth annotations from the strong labels of the synthetic data and jointly optimized a foreground-background classification branch with the sound event detection model in a multi-task fashion. Moreover, we investigated a conditional approach that combines the proposed foreground-background classification branch with the main sound event detection branch into an enhanced detection branch that further improves robust sound event detection. Results showed the added benefit of incorporating auxiliary information regarding the spectro-temporal characteristics of domestic sounds, which can be thought of as hierarchical sound event classification. Following the previous studies, we addressed domain adaptation for sound event detection. In particular, we focused on reducing the mismatch between synthetic and real data within the customary semi-supervised learning framework for training sound event detection systems. We first designed a frame-level domain adaptation strategy based on optimal transport that aligns the empirical distributions of the feature representations of active and inactive frames of synthetic and real data. Then, we extended this framework to perform segment-level domain adaptation and we combined it with the mismatch-first farthest traversal active learning strategy to optimize adaptation. We found that the proposed methods lead to enhanced performance in terms of the event-based macro F1-score on the DESED validation and public evaluation sets, and that their combination leads

to further enhanced detection scores on real test data.

In summary, in this thesis, we studied how to strengthen sound analysis systems in the face of the intrinsic complexity of everyday acoustic scenes and the issues that arise due to label and covariate shift, which can degrade the accuracy of these systems in real-world settings. We tackled problems of practical interest touching on audio source separation, acoustic scene classification and sound event detection. In a first line of research, we investigated the separation of short duration sounds with fast evolving spectral characteristics from long duration stationary sounds. This task naturally arises as a potential benefit for the classification or detection of overlapping sounds. In a second line of research, we designed methods to make audio classification and detection systems robust to mismatched training and test conditions. In a nutshell, we addressed two scenarios: acoustic scene classification with mismatched recording devices and training of sound event detection (SED) systems with synthetic and real data. In the first scenario we showed that feature correction of the test data combined with learning-based adaptation mitigates the effects of mismatched microphones. In the second scenario, we showed that the integration of domain adaptation to the semi-supervised training recipe of sound event detection leads to models whose learned representations are less invariant to synthetic and real data. Our investigation in the aforementioned scenarios led us to explore other ways to achieve robustness, such as suitable acoustic front-ends and the use of auxiliary information that complement the task at hand. This thesis aimed to solve difficulties that audio analysis systems often encounter in adverse scenarios. We hope that our contributions will spark new ideas and set the base for novel methods that further improve machine listening applications in real environments.

6.2 Perspectives

The problems tackled in this thesis can give rise to various possible modifications of the systems that we have proposed. We have identified five areas of improvement that we present in this section.

6.2.1 Selection and refinement of pseudo-labels for domain adaptation

In Chapter 5 we presented a domain adaptation strategy for sound event detection to further reduce the mismatch between synthetic and real training data that is partially solved by the semi-supervised training recipe of SED. Our proposed method seeks for the joint distribution alignment of features and labels of synthetic and real audio representations. We validated in Chapter 4 the benefits of conditional distribution alignment over marginal distribution alignment. Unlike the studied adaptation method for ASC with mismatched recording devices, SED conditional alignment is more challenging, since adaptation is carried out either at the frame or segment level rather than globally at the scene level. Furthermore, the DeepJDOT framework explored for SED relies on sampling unlabeled real instances according to pseudo-labels generated by the SED model. Herein lies the major limitation of our proposed domain adaptation method and overcoming it would help us achieve a bigger fraction of the oracle adaptation performance (see Figure 5.16). The procedure is prone to large amounts of negative transfer as errors in the detection scores may gradually accumulate and lead to undesired performance degradation. Moreover, deep learning-based methods often suffer from miscalibrated scores (Guo et al., 2017) which is further accentuated under distribution shifts (Chen et al., 2019a). Our pseudo-label refinement method described in Section 5.1.3 simply filters out pseudo-labels from weakly labeled data according to their clip-level references. This subset of data represents however around 10%

of the available real training data. We believe that a more principled refinement strategy to improve the quality of the pseudo-labels should be tailored for SED and address explicitly the mismatch between synthetic and real data. This opens up the possibility not only to improve unsupervised domain adaptation, but also to improve current semi-supervised learning and self-training strategies for SED.

In a first step, primary refinement criteria such as confidence-based or re-weighting strategies to select trustworthy pseudo-labels for domain adaptation can be interesting to explore, but we caution that the performance gain can be inconsistent in severe mismatched settings (Liu et al., 2021). We argue that pseudo-labeling mechanisms with external models may be a better choice, especially if more complex models trained on large-scale datasets such as Audioset are used. Along the same line, state-of-the-art cluster-based methods (Mahon and Lukasiewicz, 2021) or neighbor-based methods (Zhong et al., 2020) can be used to divide target domain samples into different clusters and propagate pseudo-labels. We highly encourage the exploration of more complex learning-based methods such as *SimPLE* (Hu et al., 2021b) or *Fixmatch* (Grollmisch and Cano, 2021) for reliable pseudo-label generation and mismatch correction. The latter method has already been recently explored for audio classification (Grollmisch and Cano, 2021).

6.2.2 Robustness to the data imbalance of active and inactive frames in SED

Impact of inactive frames in domain adaptation

Section 5.3.2 points out that the adaptation of SED models with active frames leads only to a small improvement, while adaptation with both active and inactive frames produces the best result. Clearly, adaptation benefits more from correcting the mismatch experienced by "silent" frames (see Figure 5.6). A recent study by Imoto et al. (2022) on the impact of data imbalance caused by inactive frames showed that inactive frames tend to overwhelm the model training. This implies that detecting frames accurately leads to ignoring active frames. In our experiments we considered equal contributions for the losses accounting for the distribution matching of active and inactive frames. Similarly to the work of Imoto et al. (2022), it would be interesting to down-weight or up-weight the contribution of the loss accounting for the mismatch of inactive frames. We observed that adapting only inactive frames is detrimental for SED. Thus, would controlling its impact over active frames help improve detection scores? If beneficial, the proposed approach can be extended to a background-only adaptation process, which is of practical interest for sound analysis tasks under noise/background mismatch.

Leveraging polyphonic sounds for domain adaptation

Regarding adaptation with active frames, we recall that sampling was carried out over frames or segments exhibiting information from a single sound event, thus we discarded frames where sound events overlap. While the proposed method over frames or segments with class information of a single sound induces the alignment of the synthetic and real data distributions, this approach is sub-optimal. In fact, this multi-class setting may worsen the effects of data imbalance between active and inactive frames or segments. An extension of the proposed adaptation method based on optimal transport is worth pursuing for the intrinsic multi-label multi-class nature of SED. The adaptation method, with a suitable multi-label training objective (Frogner et al., 2015), could potentially achieve enhanced detection scores by leveraging polyphonic frames and sound segments.

6.2.3 Extension of the proposed domain adaptation methods to other practical scenarios

The proposed domain adaptation methods presented in this work addressed two sources of mismatch that hinder the performance of audio analysis systems: mismatch between training and test recording devices in Chapter 4 and mismatch between synthetic and real data in Chapter 5. Modeling the linear distortion caused by mismatched recording devices is simple and relatively easy to correct. In contrast, modeling the complexity of real soundscapes is surely a difficult endeavor. To address this mismatch we proposed a domain adaptation strategy that leverages a heterogeneous dataset (Turpault and Serizel, 2020) aiming to produce audio representations invariant to a wide range of recordings conditions. While accounting for the differences between simulated and real data is important to achieve robustness in real-world settings, except for the mismatch due to differences in the acquisition hardware, this study has not touched on isolated sources of mismatch such as those intrinsic to sound events (e.g., duration, frequency, co-occurrence) or to the acoustic scene (e.g., background noise, reverberation, FBSNR) that make recognition challenging in particular cases. The formulation of our proposed domain adaptation methods can be used to investigate more scenarios of practical interest. We list below a couple of them.

- Noise/background mismatch. This setting corresponds to scenarios in which audio analysis systems are required to recognize target sounds of interest in the presence of unknown noise types or varying FBSNR levels, e.g., an urban monitoring system trained with the sounds of a loud city and deployed in a quieter village. Another example corresponds to a wildlife audio detection system aiming to recognize species in different geographical zones.
- Room acoustics mismatch. In this setting a model deployed in a fixed location needs to cope with the acoustic conditions of its physical surroundings, for instance a sound surveillance system deployed in the living room of a family house. This scenario of practical interest would require adapting the system once installed and whenever acoustic conditions change, for example, when the system is moved to the kitchen or when new pieces of furniture badly influence the performance of the model.

6.2.4 Evaluation and optimization with contrastive metrics

In Section 5.3.3 we discussed our participation in the DCASE Challenge 2021. The proposed systems achieved competitive results in terms of the F1 score, while their performance in terms of the PSDS (Bilen et al., 2020) metric was below the baseline model. Indeed, since metrics influence the analysis and comparisons between systems, it would be interesting to optimize the proposed methods for SED with metrics such as the PSDS with parameters that reflect more realistic scenarios. These include scenarios when the system needs to react quickly upon an event detection (PSDS 1 in the DCASE Challenge) or when the system trades-off between-class confusion and time reaction (PSDS 2 in the DCASE Challenge).

6.2.5 Further improving ASC with mismatched recording devices

In Chapter 4 we showed that among the best adaptation strategies for ASC with mismatched recording devices, those that apply moment normalization and/or matching in combination with adversarial domain adaptation at all steps of the classification pipeline achieve a performance in the target domain close to that obtained in the source domain. While these strategies correct to a greater extent the distribution shift, there is room for improvement in the adversarial domain

adaptation framework. A simple extension lies in the domain classification model h. Instead of a binary classifier that discriminates between the source and the target domain data, it could be modified to perform multi-class classification among the devices A, B, C. This setting would correspond to a source to an unsupervised multi-target domain adaptation problem (Yang et al., 2021) and would benefit from the similar approach carried out by the proposed moment matching and moment normalization strategies when using device dependent statistics.

7

Résumé étendu

7.1 Introduction

Le sens de l'ouïe chez l'homme est remarquable pour sa capacité à nous connecter au monde par le son. Du murmure le plus subtil au bruit le plus fort, nous recueillons, traitons et donnons un sens aux sons que nous rencontrons continuellement dans notre vie quotidienne. Ils sont essentiels à notre bien-être et à notre survie. Ils nous procurent de la joie à travers le son de la musique, et nous pouvons exprimer du mécontentement à travers un son rauque. Qu'ils soient agréables ou ennuyeux, les sons transmettent des informations significatives. Ils ouvrent la voie à la communication par la parole, nous alertent du danger par des sirènes et des coups de feu, et nous appellent à l'action par des cris de bébé et des alarmes.

Plus que jamais, la recherche constante de méthodes informatiques pour comprendre notre environnement sonore a porté ses fruits dans de nombreux domaines, notamment la robotique, la surveillance urbaine, la bio-acoustique et la surveillance. De l'industrie aux applications d'intérêt général, l'analyse automatique des scènes et événements sonores permet d'interpréter le flux continu de sons quotidiens.

De nos jours, les produits électroniques grand public impliquant la détection sonore et l'analyse automatique du son sont commercialisés à grande échelle, malgré le fait qu'ils ont tendance à être peu performants dans des environnements réels. Une des principales dégradations rencontrées lors du passage des conditions de laboratoire au monde réel est due au fait que les scènes sonores ne sont pas composées d'événements isolés mais de plusieurs événements simultanés. Des différences entre les conditions d'apprentissage et de test des modèles de détection surviennent aussi souvent en raison de facteurs extrinsèques, tels que le choix du matériel d'enregistrement et des positions des microphones, et de facteurs intrinsèques aux événements sonores, tels que leur fréquence d'occurrence, leur durée et leur variabilité. Augmenter la robustesse des systèmes d'analyse automatique du son face aux adversités des environnements du monde réel permettra aux applications d'aller au-delà de l'identification de scènes sonores ou de sons isolés avec une grande précision. Donner du sens aux relations complexes entre les sons, le lieu et le temps dans lesquels ils se produisent pourrait fournir une meilleure description des environnements sonores, ce qui pourrait potentiellement conduire à l'émergence de nouvelles applications et améliorer les applications existantes.

7.2 Contexte

La recherche sur l'analyse des sons ambiants a bénéficié des progrès réalisés dans des domaines plus établis tels que l'étude des signaux vocaux et musicaux. Par exemple, les méthodes de séparation des sons ambiants se sont inspirées des méthodes de séparation supervisée de la parole afin de séparer des événements sonores d'une grande variété de classes et d'attributs beaucoup moins structurés. Des études récentes menées par Kavalerov et al. (2019) et Wisdom et al. (2020) ont montré des résultats prometteurs vers la séparation des sons arbitraires. La séparation des sources audio appliquée aux événements sonores au-delà de la parole ou de la musique peut devenir un outil potentiel pour les tâches d'analyse audio impliquant la reconnaissance du son. En principe, cela peut améliorer la robustesse des systèmes de classification et de détection pour cibler les sons d'intérêt se produisant simultanément ou en présence de bruit ou d'autres interférences non ciblées. Cependant, il est tout aussi important de fournir des systèmes d'analyse solides et robustes face à d'autres problèmes du monde réel. De fait, lors du développement de systèmes d'analyse audio, il existe une hypothèse sous-jacente selon laquelle les échantillons d'apprentissage sont des échantillons i.i.d. tirés de la même distribution que les échantillons de test. En pratique, cette hypothèse est fausse : les systèmes d'analyse audio sont sujets à des changements de domaine, qui résultent de l'évolution des conditions du développement au déploiement. La solution à ce problème est connue sous le nom d'*adaptation de domaine* (Ben-David et al., 2006). L'objectif est d'adapter un modèle pour qu'il fonctionne bien sur des échantillons de test d'un domaine *cible* dont la distribution est différente du domaine *source* duquel les données d'apprentissage ont été tirées. L'adaptation de domaine nécessite l'accès à un jeu de données annoté tiré de la distribution source et à un jeu de données tiré de la distribution cible. En fonction de la disponibilité des annotations pour les données du domaine cible, le processus d'adaptation peut être effectué de différentes manières. Si aucune annotation n'est fournie, une *adaptation de domaine non supervisée* est effectuée. Lorsque seules quelques annotations sont disponibles pour un sous-ensemble des données du domaine cible, le processus d'adaptation est appelé *adaptation de domaine semi-supervisée*. Dans les deux cas, l'objectif est de réduire l'écart existant entre les distributions d'apprentissage et de test.

Kull and Flach (2014) catégorisent le changement de distributions des données en *label shift*, *covariate shift* ou *concept shift*. Les scénarios du monde réel combinent un ou plusieurs de ces changements. La plupart des travaux sur l'adaptation de domaine pour les tâches d'analyse sonore abordent le *covariate shift*. Par exemple, la Tâche 1 du défi DCASE motive le développement de méthodes qui généralisent à des dispositifs d'enregistrement différents pour la tâche de classification de scène sonore. Beaucoup d'entre eux s'appuient sur des techniques d'augmentation de données, et seule une minorité réduit explicitement la différence de distribution avec des méthodes d'adaptation de domaine. Dans des conditions d'enregistrement identiques, les distorsions dues au bruit d'arrière-plan, à la réverbération et à la présence de sons non-ciblés peuvent également dégrader la qualité du signal de test. Ces sources d'erreur ont été particulièrement étudiées pour la tâche de détection d'événements sonores (Turpault and Serizel, 2020; Ronchini et al., 2021). Malgré l'effort de simulation des conditions du monde réel, un écart de performance subsiste entre les données de test synthétiques et les données réelles.

7.3 Séparation de sources sonores ambiantes

Les progrès récents sur la séparation des sources audio font de ce domaine de recherche une base potentielle pour les tâches d'analyse audio qui nécessitent une identification robuste des événe-

ments sonores en présence d'interférences. Sur la base de l'observation que les systèmes d'analyse audio fonctionnent sur des enregistrements continus impliquant plusieurs sons de courte durée sur un arrière-plan stationnaire, nous acons introduit la tâche de *séparation avant-plan/arrière-plan de sources sonores ambiantes*. Nous visons trois objectifs. Tout d'abord, nous examinons si un système de séparation basé sur l'apprentissage profond est capable de différencier les caractéristiques spectro-temporelles à variation rapide des événements audio courts par rapport aux caractéristiques à variation plus lente des sons d'arrière-plan rencontrés dans des environnements réels. Deuxièmement, nous explorons l'application de la normalisation d'énergie par canal (PCEN) aux représentations temps-fréquence d'entrée pour améliorer les performances de séparation. Troisièmement, nous étudions si un modèle de séparation de sources de cette nature est capable de généraliser à des classes de sons d'avant-plan et/ou d'arrière-plan non vues à l'apprentissage.

7.3.1 Formulation du problème

Pour un enregistrement audio monocanal, nous définissons la séparation avant-plan/arrière-plan comme la tâche d'estimer les composantes d'avant-plan et d'arrière-plan du mélange. Le mélange d'entrée $x(t)$ est modélisé comme

$$x(t) = f(t) + b(t), \tag{7.1}$$

où

$$f(t) = \sum_{i=1}^{I} f_i(t). \tag{7.2}$$

Les $\{f_i(t)\}_{i=1..I}$ sont les événements d'avant-plan au nombre total de I, et $b(t)$ est un bruit d'arrière-plan dont les caractéristiques spectro-temporelles sont supposées stationnaires. Étant donné le mélange $x(t)$, notre objectif est d'estimer les signaux d'avant-plan et d'arrière-plan $f(t)$ et $b(t)$.

Nous nous concentrons uniquement sur la récupération d'un signal d'avant-plan $f(t)$ contenant des événements sonores d'une seule classe, c'est-à-dire $I = 1$, superposé à un son d'arrière-plan de durée longue.

La tâche de séparation avant-plan/arrière-plan est effectuée dans le domaine de la transformée de Fourier à court terme (TFCT). Les coefficients TFCT $X(n, f)$ du mélange $x(t)$ dans l'intervalle de temps n et à la fréquence f satisfont

$$X(n, f) = F(n, f) + B(n, f), \tag{7.3}$$

où $F(n, f)$ et $B(n, f)$ sont les coefficients de TFCT de $f(t)$ et $b(t)$, respectivement. Nous utiliserons la notation $\mathbf{X}$, $\mathbf{F}$, $\mathbf{B}$ pour les matrices de taille $N_x \times F$ comprenant tous les coefficients à valeurs complexes $X(n, f)$, $F(n, f)$ et $B(n, f)$, N_x étant le nombre de trames temporelles et F le nombre de canaux de fréquence. Les composantes $\widehat{\mathbf{F}}$ et $\widehat{\mathbf{B}}$ estimées par le processus de séparation sont retransformées en signaux du domaine temporel $\widehat{\mathbf{f}}(t)$ et $\widehat{\mathbf{b}}(t)$ en calculant la TFCT inverse.

7.3.2 Méthode de séparation

La méthode de séparation avant-plan/arrière-plan proposée est décrite dans la Figure 7.1. Inspirée de la tâche d'extraction du locuteur cible (Žmolíková et al., 2019), elle repose sur un réseau de neurones profond principal et un réseau auxiliaire optionnel pour localiser les composantes d'arrière-plan et d'avant-plan dans le plan temps-fréquence. Les deux réseaux fonctionnent sur l'échelle de fréquence Mel. On note F' le nombre de bandes Mel et $|\mathbf{X}|^{\text{Mel}}$ le spectrogramme

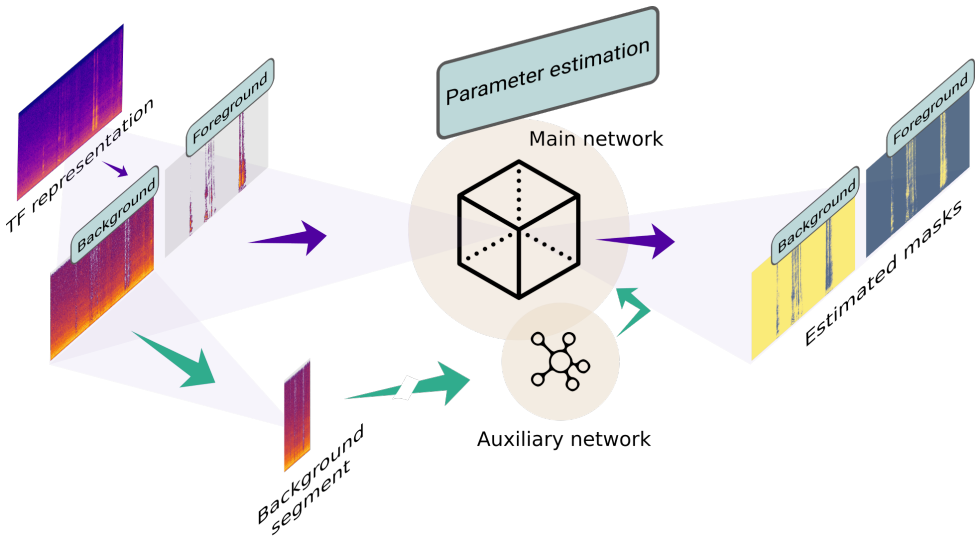


Figure 7.1: Schéma général de la méthode de séparation avant-plan/arrière-plan proposée.

Mel du mélange qui s'obtient en multipliant l'amplitude de la TFCT $|\mathbf{X}|$ par la matrice de taille $F' \times F$ représentant le banc de filtres Mel. Le réseau auxiliaire peut être actif ou inactif dans le processus de séparation. Lorsqu'il est *inactif*, le réseau principal prend en entrée le spectrogramme log-Mel du mélange $\log|\mathbf{X}|^{\text{Mel}}$ et sort un masque temps-fréquence $\mathbf{M}^{\text{Mel}}$, c'est-à-dire une matrice de taille $N_x \times F'$ avec des entrées à valeurs réelles dans $[0, 1]$ qui quantifient la proportion de son d'avant-plan dans chaque case temps-fréquence. Cette matrice est ensuite projetée sur le domaine TFCT pour obtenir un masque $\mathbf{M}$ de dimensions $N_x \times F$[6]. À l'aide du masque, les amplitudes TFCT estimées des composantes d'avant-plan et d'arrière-plan sont obtenues comme

$$|\widehat{\mathbf{F}}| = \mathbf{M} \odot |\mathbf{X}| \quad \text{et} \quad |\widehat{\mathbf{B}}| = (\mathbf{1} - \mathbf{M}) \odot |\mathbf{X}|, \tag{7.4}$$

où $\odot$ désigne la multiplication élément par élément et $\mathbf{1}$ est une matrice dont les coefficients sont égaux à 1.

Lorsque le réseau auxiliaire est *actif*, l'information préalable sur le bruit d'arrière-plan est supposée disponible sous la forme d'un segment d'adaptation $a(t)$, qui peut être un court échantillon du mélange à traiter ou un intervalle de temps précédent qui a été classé avec une grande confiance comme étant uniquement en arrière-plan, c'est-à-dire sans aucun son d'avant-plan superposé. Le réseau auxiliaire compresse le spectrogramme log-Mel $\log|\mathbf{A}|^{\text{Mel}}$ du segment d'adaptation, de taille $N_a \times F'$ avec N_a le nombre de trames correspondantes, en un vecteur de taille fixe $\boldsymbol{\lambda}$ qui est utilisé avec le spectrogramme log-Mel du mélange $\log|\mathbf{X}|^{\text{Mel}}$ par le réseau principal pour produire un masque temps-fréquence $\mathbf{M}^{\text{Mel}}$. Ce masque est converti dans le domaine TFCT et utilisé pour estimer les amplitudes des TFCT des composantes d'avant-plan et d'arrière-plan via (7.4). Ces estimations sont ensuite combinées avec la phase du mélange pour obtenir les signaux dans le domaine temporel $\widehat{\mathbf{f}}(t)$ et $\widehat{\mathbf{b}}(t)$ au moyen de la TFCT inverse.

[6]Nous avons adopté cette approche car l'estimation d'un masque directement dans le domaine TFCT ne faisait pas de différence significative.

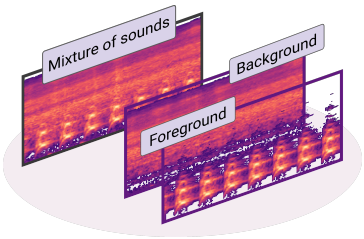
7.3.3 Protocole expérimental

Jeu de données

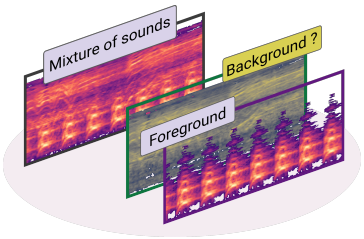
Nous avons généré des mélanges synthétiques avec des événements sonores d'avant-plan et d'arrière-plan en échantillonnant de manière aléatoire des sons isolés à partir des ensembles de développement et d'évaluation du corpus de détection d'événements sonores en environnement domestique DESED (Turpault et al., 2019) et d'Audioset (Gemmeke et al., 2017). Les classes sonores correspondantes sont répertoriées dans le Tableau 7.1 et forment quatre configurations illustrées dans la Figure 7.2.

Table 7.1: Classes sonores considérées pour la création du jeu de données.

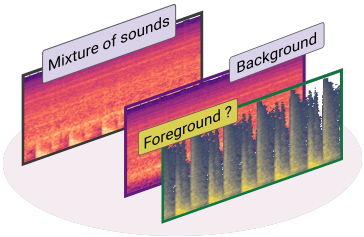
Set	Avant-plan	Arrière-plan
Apprentissage	DESED: `chien, parole, chat, vaisselle, sonnette d'alarme`	DESED: `aspirateur, mixeur, friture, eau courante, rasoir-brosse à dents électrique`, Audioset: `baignoire, ventilateur mécanique, four micro-ondes, séchoir à cheveux, perceuse`
Validation		
Évaluation C1		
Évaluation C3	Audioset: `porte, claquer, grincer, pièces de monnaie, hacher la nourriture`	
Évaluation C2	DESED: `chien, parole, chat, vaisselle, sonnette d'alarme`	Audioset: `bruit rose, bruit blanc, bruit, cascade, vibration`
Évaluation C4	Audioset: `porte, claquer, grincer, pièces de monnaie, hacher la nourriture`	



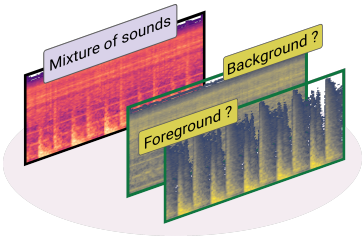
(a) C1: mélange de classes d'avant-plan et d'arrière-plan vues à l'apprentissage.



(b) C2: mélange de classes d'avant-plan vues et d'arrière-plan non vues à l'apprentissage.



(c) C3: mélange de classes d'avant-plan non vues et d'arrière-plan vues à l'apprentissage.



(d) C4: mélange de classes d'avant-plan et d'arrière-plan non vues à l'apprentissage.

Figure 7.2: Sous-ensembles de l'ensemble de données d'évaluation.

Caractéristiques d'entrée

En plus de calculer les spectrogrammes log-Mel comme caractéristiques d'entrée, nous calculons une représentation temps-fréquence basée sur la normalisation de l'énergie par canal (PCEN) (Lostanlen et al., 2018). Cette méthode de normalisation améliore les transitoires des événements d'avant-plan tout en filtrant les sons stationnaires. Elle est définie comme

$$\mathrm{PCEN}(n, f') = \left(\frac{|X|^{\mathrm{Mel}}(n, f')}{(\epsilon + \overline{|X|^{\mathrm{Mel}}}(n, f'))^{\alpha}} + \delta \right)^{r} - \delta^{r},$$

où f' désigne l'indice de bande Mel et $\overline{|X|^{\mathrm{Mel}}}(n, f')$ est une version lissée de $|X|^{\mathrm{Mel}}(n, f')$, qui est calculée à l'aide d'un filtre à réponse impulsionnelle infinie (IIR) du premier ordre comme

$$\overline{|X|^{\mathrm{Mel}}}(n, f') = (1 - s) \times \overline{|X|^{\mathrm{Mel}}}(n - 1, f') + s \times |X|^{\mathrm{Mel}}(n, f'),$$

avec s la constante de lissage. Nous avons adopté les paramètres PCEN par défaut définis par Lostanlen et al. (2018), c'est-à-dire $s = 0,025$, $\epsilon = 10^{-6}, \alpha = 0,98, \delta = 2$ et $r = 0,5$.

Architecture et apprentissage du modèle

Le réseau principal comprend un empilement de 3 couches BLSTM avec 300 unités. La dernière couche est une couche dense avec 128 unités et une activation sigmoïde. Le réseau auxiliaire suit l'architecture de Veselỳ et al. (2016) et compresse le segment d'adaptation en un vecteur de taille fixe

$$\boldsymbol{\lambda} = \frac{1}{N_a} \sum_{n=1}^{N_a} g(\log |\mathbf{a}|^{\mathrm{Mel}}(n)), \tag{7.5}$$

où $|\mathbf{a}|^{\mathrm{Mel}}(n)$ désigne la n-ième trame de $|\mathbf{A}|^{\mathrm{Mel}}$, et $g(\cdot)$ est une transformation non-linéaire par trame. Ce réseau comprend deux couches denses avec 128 neurones et activation ReLU et linéaire, respectivement.

Que le réseau auxiliaire soit actif ou non, le modèle est entraîné en minimisant la norme de Frobenius au carré entre le spectrogramme Mel de l'événement d'avant-plan cible et le spectrogramme Mel du mélange multiplié élément par élément par le masque estimé :

$$\min \|\mathbf{M}^{\mathrm{Mel}} \odot |\mathbf{X}|^{\mathrm{Mel}} - |\mathbf{F}|^{\mathrm{Mel}}\|_F^2. \tag{7.6}$$

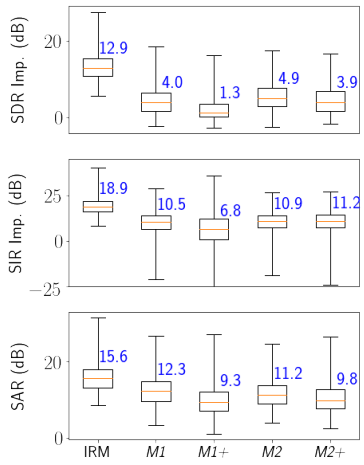
Nous avons construit quatre modèles pour étudier la tâche de séparation. Le modèle *M1* est entraîné avec des spectrogrammes log-Mel en entrée et ne repose pas sur le réseau auxiliaire pour l'estimation du masque. Le modèle *M1+* prend aussi des spectrogrammes log-Mel comme entrées et utilise le réseau auxiliaire. Les modèles *M2* et *M2+* sont définis de manière similaire à *M1* et *M1+*, respectivement, mais utilisent des spectrogrammes PCEN comme entrées au lieu de spectrogrammes log-Mel.

Nous avons utilisé comme méthode de base une méthode de séparation oracle basée sur le masque de rapport idéal (IRM). Nous calculons l'IRM comme $\mathbf{M}^{\mathrm{IRM}} = \mathbf{F}^{\mathrm{Mel}}/(\mathbf{F}^{\mathrm{Mel}} + \mathbf{B}^{\mathrm{Mel}})$, où $\mathbf{F}^{\mathrm{Mel}}$ et $\mathbf{B}^{\mathrm{Mel}}$ sont les spectrogrammes Mel des signaux d'avant-plan et d'arrière-plan cibles.

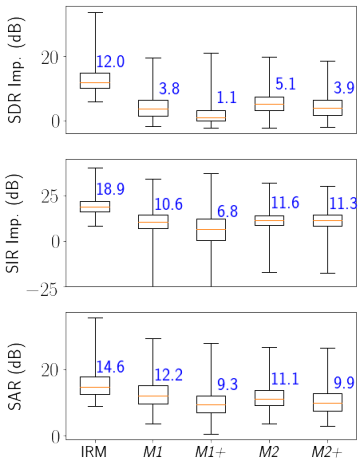
7.3.4 Résultats

Résultats qualitatifs

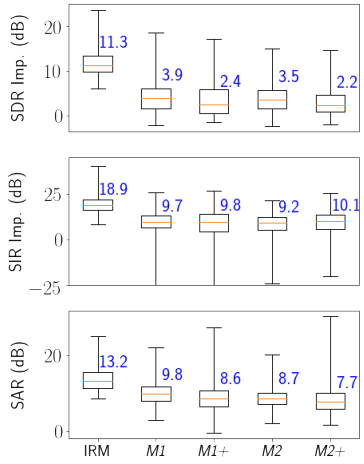
Nous évaluons les modèles de séparation proposés en terme des rapports signal-à-distorsion (SDR), signal-à-interférence (SIR) et signal-à-artefacts (SAR). Les boîtes à moustaches présentées dans la Figure 7.3 montrent les améliorations du SDR et du SIR et le SAR obtenus



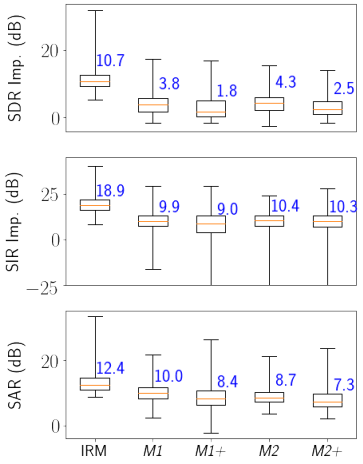
(a) C1: classes d'avant-plan et d'arrière-plan vues.



(b) C2: classes d'avant-plan vues et d'arrière-plan non vues.



(c) C3: classes d'avant-plan non vues et d'arrière-plan vues.



(d) C4: classes d'avant-plan et d'arrière-plan non vues.

Figure 7.3: Améliorations du SDR et du SIR et SAR en décibels (dB) obtenus par les quatre configurations de modèle et l'IRM sur les quatre sous-ensembles de l'ensemble d'évaluation.

par les quatre configurations de modèle et l'IRM sur les quatre sous-ensembles de l'ensemble d'évaluation.

Impact des caractéristiques d'entrée. À l'exception de la configuration *C3* (mélanges avec classes d'avant-plan non vues et arrière-plan vues à l'apprentissage), les modèles utilisant des spectrogrammes PCEN (*M2* et *M2+*) ont obtenu des scores SDR et SIR plus élevés que les modèles entraînés avec des spectrogrammes log-Mel. Ces résultats indiquent que PCEN est bénéfique pour la tâche de séparation en termes de SDR et de SIR, malgré la production de plus d'artefacts dans les signaux estimés que les spectrogrammes log-Mel.

Impact du réseau auxiliaire. En général, les modèles utilisant le réseau auxiliaire ont obtenu de moins bons résultats en terme de SDR. L'utilisation d'un réseau auxiliaire dans le processus d'estimation de masque peut ne pas être nécessaire car les sources à séparer sont de natures sensiblement différentes.

Robustesse aux événements invisibles. L'intérêt de la capacité de généralisation des

modèles de séparation avant-plan/arrière-plan réside notamment dans la performance des modèles dans des configurations où les mélanges contiennent des classes de sons non vues à l'apprentissage. L'amélioration de SDR des modèles appris sur des spectrogrammes log-Mel ou PCEN montre une bonne généralisation pour toute combinaison de classes de sons d'avant-plan et d'arrière-plan vues ou non vues.

Résultats qualitatifs

Impact du réseau auxiliaire dans la distorsion du signal. L'ajout du segment d'adaptation au réseau principal via le réseau auxiliaire entraîne une réduction plus forte du bruit d'arrière-plan, entraînant une forte distorsion de l'avant-plan, notamment pour le modèle *M1+* entraîné sur des spectrogrammes log-Mel. La Figure 7.4 montre la vérité-terrain (GT) et les composantes d'avant-plan et d'arrière-plan estimées à partir d'exemples de mélanges pour la configuration *C4*. **Impact du réseau auxiliaire sur les interférences et les artefacts**. La forte suppression du bruit d'arrière-plan entraîne l'estimation de la composante d'avant-plan avec moins

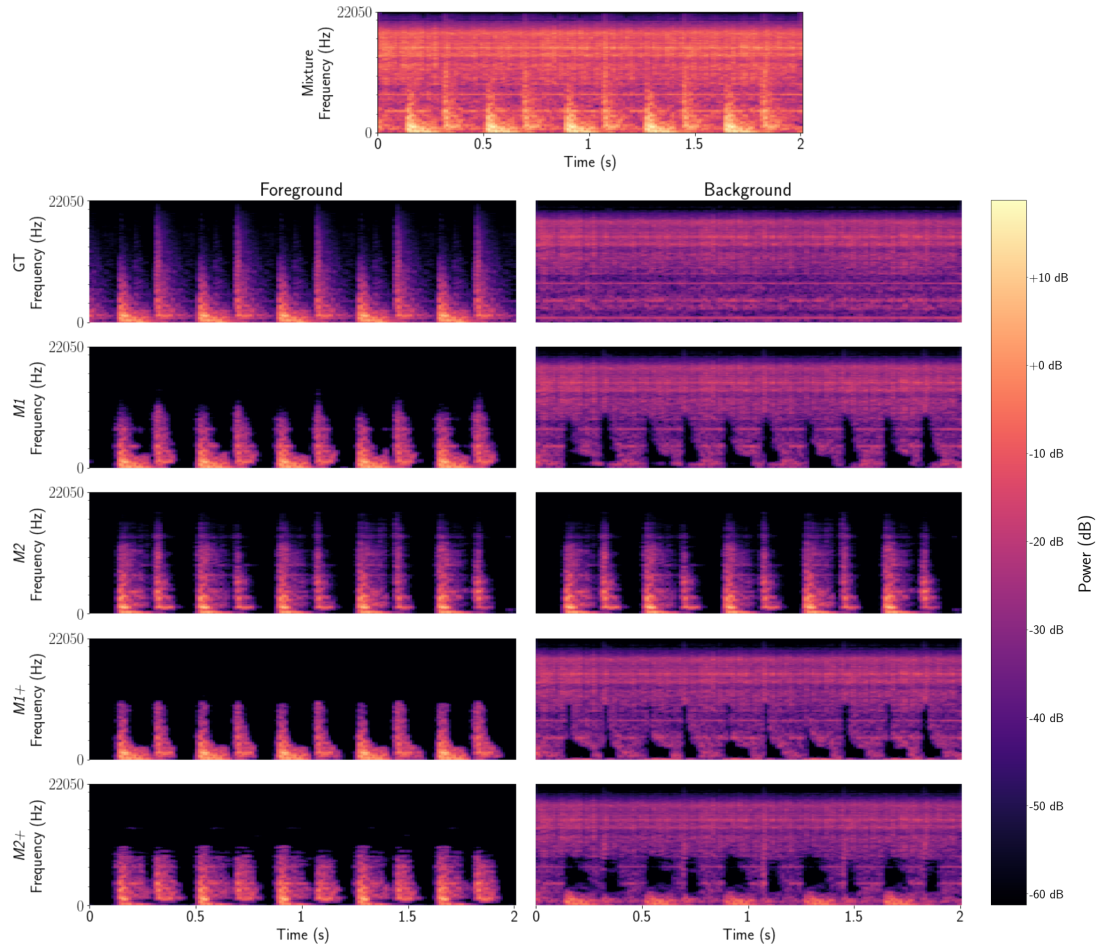


Figure 7.4: Exemple de composantes d'avant-plan et d'arrière-plan estimées pour les quatre configurations de modèles et comparaison avec les signaux de vérité terrain (GT) sur le sous-ensemble d'évaluation *C4* : mélanges de classes d'avant-plan et d'arrière-plan non vues. Avant-plan : *porte* ; arrière-plan : *vibration*.

d'interférences de l'arrière-plan. Cela conduit à une suppression partielle de la structure spectro-temporelle du son d'avant-plan, en particulier pour les fréquences plus élevées et conduit également à des artefacts. Cet effet est récurrent dans les quatre configurations.

Impact de PCEN comme pré-traitement. L'utilisation de spectrogrammes PCEN conduit à des sons d'avant-plan estimés avec moins de distorsions par rapport aux spectrogrammes log-Mel. PCEN améliore les informations spectrales des événements sonores courts, ce qui améliore la qualité de l'estimation du masque et atténue le fort impact du réseau auxiliaire lors de la suppression du bruit d'arrière-plan.

7.3.5 Conclusion

Nous avons introduit la tâche de séparation des sons ambiants d'avant-plan et d'arrière-plan, dans laquelle des événements de courte durée se produisent simultanément à un bruit d'arrière-plan. Nous considérons que cette tâche est plus proche des conditions rencontrées par les systèmes d'analyse du son ambiant dans des conditions réelles que les mélanges de sons arbitraires considérés dans les études précédentes. Nous avons effectué une série d'expériences pour évaluer les performances d'un modèle de séparation de sources basé sur l'apprentissage profond pour discriminer les caractéristiques spectro-temporelles à variation rapide des événements d'avant-plan par rapport aux caractéristiques à variation lente des bruits d'arrière-plan. Nous explorons une architecture comprenant un réseau d'estimation de masque principal et analysons l'apport d'un réseau auxiliaire avec un segment d'adaptation. De plus, nous avons évalué l'impact des spectrogrammes log-Mel et PCEN en tant que caractéristiques d'entrée. Dans le cadre du protocole expérimental proposé, nous avons constaté que l'utilisation d'un réseau auxiliaire dans le processus d'estimation du masque est préjudiciable, tandis que les spectrogrammes PCEN sont plus appropriés que les spectrogrammes log-Mel. L'évaluation objective des modèles proposés a montré des capacités de généralisation sur des mélanges de classes d'événements sonores d'avant-plan et/ou d'arrière-plan vues ou non-vues à l'apprentissage en terme de SDR.

7.4 Stratégies de normalisation pour l'adaptation de domaine non supervisée

Dans le cadre de la tâche de classification de scène sonore (ASC), nous nous intéressons particulièrement au *covariate shift* qui découle de l'utilisation de différents appareils d'enregistrement pour l'apprentissage et le test. Des conditions d'enregistrement différentes dégradent les performances des systèmes ASC et empêchent la généralisation à des données capturées avec de nouveaux appareils d'enregistrement. Nous abordons ce problème avec une adaptation de domaine non supervisée, dans laquelle les enregistrements des appareils source et cible sont disponibles, mais seuls les enregistrements source sont annotés. Les travaux antérieurs traitant de cette tâche se sont appuyés sur l'adaptation de domaine antagoniste (Tzeng et al., 2017; Gharib et al., 2018; Drossos et al., 2019), ou sur des méthodes de transformation de caractéristiques linéaires qui, bien qu'elles ne nécessitent aucune étape d'adaptation, sont tout aussi efficaces (Mezza et al., 2020, 2021). L'intégration de ces diverses méthodes n'a pas encore été explorée. Nous comblons cette lacune en réalisant un ensemble complet d'expériences qui montrent l'impact sur la généralisation au domaine cible de diverses stratégies de normalisation et d'appariement de caractéristiques et de leur intégration avec des stratégies d'adaptation de domaine antagonistes.

7.4.1 Formulation du problème

Modèle de distorsion linéaire

Désignons par x_{nmk} les coefficients de log-amplitude de la TFCT dans la trame m et la bande de fréquence k du signal réel (non déformé) d'une scène sonore indexée par n, et par x_{dnmk} les coefficients de log-amplitude de la TFCT du même signal capturé par un dispositif d'enregistrement d avec une réponse en fréquence d'amplitude linéaire et invariante dans le temps h_{dk}. La scène sonore capturée peut être exprimée comme $x_{dnmk} = x_{nmk} + \log h_{dk}$. Le signal non distordu x_{nmk} peut donc être récupéré comme

$$\hat{x}_{dnmk} = x_{dnmk} - \log h_{dk}. \tag{7.7}$$

Nous gardons l'indice d dans la notation $\hat{x}_{dnmk}$ pour l'estimation obtenue à partir de l'appareil d. En pratique, h_{dk} est inconnu, donc $\hat{x}_{dnmk}$ doit être obtenu uniquement à partir de x_{dnmk}.

Normalisation des moments

La normalisation des moments consiste à appliquer une transformation linéaire dépendante du domaine aux données dans les domaines source et cible afin que leurs moments de premier et de second ordre soient fixes. La normalisation de la moyenne consiste à soustraire le spectre de log-amplitude moyen

$$\mu_{dk} = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} x_{dnmk} \tag{7.8}$$

des données enregistrées par chaque appareil d des données d'origine :

$$\hat{x}^{\mathrm{MN}}_{dnmk} = x_{dnmk} - \mu_{dk}. \tag{7.9}$$

Ceci équivaut à (7.7), où μ_{dk} peut être vu comme une estimation de $\log h_{dk}$.

La normalisation de la moyenne et de la variance (appelée standardisation) nécessite en outre le calcul de l'écart-type

$$\sigma_{dk} = \sqrt{\frac{1}{NM-1} \sum_{n=1}^{N} \sum_{m=1}^{M} (x_{dnmk} - \mu_{nsp})^2} \tag{7.10}$$

des données pour chaque appareil d. La normalisation est obtenue par

$$\hat{x}^{\mathrm{MVN}}_{dnmk} = \frac{x_{dnmk} - \mu_{dk}}{\sigma_{dk}}. \tag{7.11}$$

Appariement des moments

Contrairement à la normalisation des moments, l'appariement des moments n'élimine pas la distorsion due au dispositif d'enregistrement d des données source. Elle transforme les données cibles enregistrées par un autre appareil d' afin que ses moments de premier et de second ordre correspondent à ceux des données source. Faire correspondre les moyennes de x_{dnmk} et $x_{d'nmk}$ peut être obtenu en supprimant de $x_{d'nmk}$ la distorsion due à l'appareil d' comme dans (7.9), puis en introduisant la distorsion due à l'appareil d :

$$\hat{x}^{\mathrm{MM}}_{d'nmk} = x_{d'nmk} - \mu_{d'k} + \mu_{dk}. \tag{7.12}$$

Après l'appariement des moments, x_{dnmk} et $\hat{x}^{\mathrm{MM}}_{d'nmk}$ partagent la même moyenne, qui selon (7.7) suffit à transférer la distorsion d'un appareil à l'autre. En pratique, une plus grande robustesse peut être obtenue en normalisant d'abord la moyenne et la variance de $x_{d'nmk}$ comme dans (7.11), puis en mettant à l'échelle les bandes de fréquences normalisées de σ_{dk} et μ_{dk} :

$$\hat{x}^{\mathrm{MVM}}_{d'nmk} = \frac{(x_{d'nmk} - \mu_{d'k})\sigma_{dk}}{\sigma_{d'k}} + \mu_{dk}. \tag{7.13}$$

Adaptation de domaine antagoniste

La méthode adoptée pour atténuer les distorsions non-linéaires restantes est une méthode d'adaptation de domaine antagoniste en deux étapes basée sur le principe des réseaux génératifs antagonistes de Wasserstein (WGAN) (Arjovsky et al., 2017; Drossos et al., 2019). Elle s'appuie sur trois modèles par réseaux de neurones profonds : un extracteur de caractéristiques g, un classifieur f et un discriminateur h. Nous considérons comme données du domaine source $\mathbf{X}^{\mathrm{s}} = \{x_{dnmk}\}_{n=1}^{N_{\mathrm{s}}}$ les scènes sonores enregistrées par l'appareil d, avec des annotations *one-hot* $\mathbf{y}^{\mathrm{s}}$ des classes considérées. Nous considérons comme données du domaine cible $\mathbf{X}^{\mathrm{t}} = \{x_{d'nmk}\}_{n=1}^{N_{\mathrm{t}}}$ les scènes sonores enregistrées par un autre appareil d', sans annotations. La procédure d'adaptation se déroule en trois étapes décrites ci-dessous : *pré-apprentissage*, *adaptation* et *inférence*.

Étape de pré-apprentissage : Dans cette étape, l'extracteur de caractéristiques pré-appris g^* et le classifieur f sont obtenus à l'aide des données de domaine source $\mathbf{X}^{\mathrm{s}}$ avec les annotations de référence correspondantes $\mathbf{y}^{\mathrm{s}}$ en minimisant la fonction de coût

$$\mathcal{L}_s = -\sum_{n=1}^{N_{\mathrm{s}}} \mathbf{y}^{\mathrm{s}}_n \cdot \log(f(g^*(\mathbf{X}^{\mathrm{s}}_n)), \tag{7.14}$$

où $\cdot$ désigne le produit scalaire.

Étape d'adaptation : Dans cette étape, un nouvel extracteur de caractéristiques g à adapter et un classifieur binaire h qui discrimine les données du domaine source et cible sont introduits. L'extracteur de caractéristiques g est initialisé en tant que modèle pré-appris g^*, et g et h sont appris conjointement sur les données des domaines source et cible en minimisant

$$\mathcal{L}_h = \sum_{n=1}^{N_{\mathrm{s}}} h(g^*(\mathbf{X}^{\mathrm{s}}_n)) - \sum_{n=1}^{N_{\mathrm{t}}} h(g(\mathbf{X}^{\mathrm{t}}_n)) \tag{7.15}$$

$$\mathcal{L}_g = \sum_{n=1}^{N_{\mathrm{t}}} h(g(\mathbf{X}^{\mathrm{t}}_n)) - \sum_{n=1}^{N_{\mathrm{s}}} \mathbf{y}^{\mathrm{s}}_n \cdot \log(f(g(\mathbf{X}^{\mathrm{s}}_n) \tag{7.16}$$

pour favoriser des distributions invariantes au domaine. Le deuxième terme de (7.16) est un coût de classification qui empêche g de perdre en performances sur les données du domaine source. Suivant Drossos et al. (2019), $\mathcal{L}_h$ et $\mathcal{L}_g$ sont itérativement minimisés en mettant à jour h selon le gradient de (7.15) par rapport à h avec g fixe, et mise à jour de g selon le gradient de (7.16) par rapport à g avec h fixe.

Étape d'inférence : Après adaptation, l'extracteur de caractéristiques g et le classifieur f sont utilisés pour classer les scènes sonores des appareils source et cible.

7.4.2 Adaptation de domaine contradictoire conditionnelle

La stratégie d'adaptation de domaine antagoniste ci-dessus est étendue à l'adaptation de domaine antagoniste conditionnelle (CADA) (Long et al., 2018) pour aligner les distributions conjointes

des caractéristiques et des annotations. L'idée clé est de conditionner le discriminateur de domaine h sur les classes issues de f avec la variable jointe $w(\mathbf{X}) = g(\mathbf{X}) \otimes f(g(\mathbf{X}))$ pour capturer les informations multimodales de g et f. En introduisant la transformation multilinéaire via $w(\mathbf{X})$, les coûts (7.15) et (7.16) deviennent

$$\mathcal{L}_h = \sum_{n=1}^{N_\mathrm{s}} h(w^*(\mathbf{X}_n^\mathrm{s})) - \sum_{n=1}^{N_\mathrm{t}} h(w(\mathbf{X}_n^\mathrm{t})) \tag{7.17}$$

$$\mathcal{L}_g = \sum_{n=1}^{N_\mathrm{t}} h(w(\mathbf{X}_n^\mathrm{t})) - \sum_{n=1}^{N_\mathrm{s}} \mathbf{y}_n^\mathrm{s} \cdot \log(f(g(\mathbf{X}_n^\mathrm{s})). \tag{7.18}$$

7.4.3 Protocole expérimental

Jeu de données

Afin d'évaluer l'impact de la normalisation ou de l'appariement des moments, ainsi que leur intégration avec l'adaptation de domaine antagoniste, nous effectuons des expériences sur l'ensemble de données de développement de la tâche 1B du Challenge DCASE 2018.

Le jeu de données est composé de 10 scènes sonores différentes : *aéroport, station de métro, shopping, parc, place publique, rue piétonne, trafic routier, bus, tram, métro*. Les enregistrements de scènes sonores durent 10 s et ont été capturés dans six villes européennes à l'aide de trois appareils d'enregistrement différents appelés A, B et C. Afin de traiter l'adaptation de domaine non supervisée dans un cadre totalement aveugle par rapport au domaine cible, nous avons rejeté 10% des données de l'ensemble de données d'origine. Notre configuration finale comprend 5 024 scènes d'apprentissage à partir de l'appareil source A et 486 pour chaque appareil cible B et C. L'ensemble de validation comprend 558 scènes à partir de l'appareil source A et 54 scènes à partir de chaque appareil cible B et C. L'ensemble de test est composé de 2 518 scènes de l'appareil A et de 180 scènes de chaque appareil cible B et C. De chaque enregistrement, nous avons extrait des spectres log-Mel à 64 dimensions.

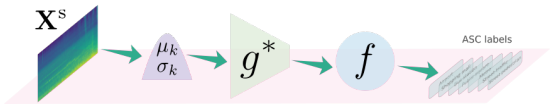
Architecture et apprentissage du modèle

Nous utilisons l'architecture de modèle appelée *Kaggle* par Gharib et al. (2018), Drossos et al. (2019) et Mezza et al. (2021). L'extracteur de caractéristiques g se compose de cinq couches convolutives (CNN), tandis que le classifieur f et le discriminateur de domaine h se composent de deux couches denses. L'optimiseur RMSProp est utilisé avec un pas d'apprentissage de 5×10^{-5}. Nous utilisons une taille de lot de 16 et le classifieur g a été appris pour 300 époques.

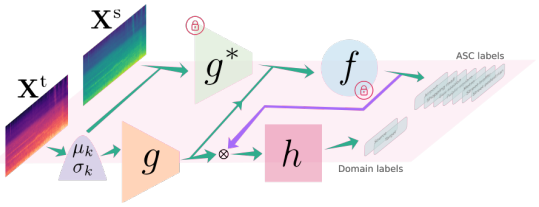
7.4.4 Expériences

Nous analysons l'impact de la normalisation et de l'appariement des moments utilisés seuls ou en combinaison avec l'adaptation de domaine contradictoire (ADA) ou l'adaptation de domaine contradictoire conditionnelle (CADA). Nous transformons les données de l'appareil source par normalisation de la moyenne (MN) ou normalisation de la moyenne et la variance (MVN) dans l'étape de préapprentissage. Au moment de l'inférence, nous transformons les données de l'appareil cible par la même stratégie MN ou MVN lorsqu'elle a été appliquée à l'étape de pré-apprentissage, ou par appariement de la moyenne (MM) ou appariement de la moyenne et la variance (MVM) lorsqu'aucune normalisation n'a été appliquée à l'étape de pré-apprentissage. Nous testons également une stratégie hybride de normalisation de la moyenne et d'appariement

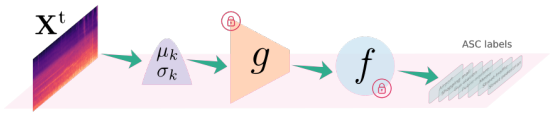
de la variance (MNVM). Dans l'étape d'adaptation, les stratégies ci-dessus transforment les données cibles avant ADA ou CADA. Nous évaluons également ADA et CADA seuls à des fins de comparaison. La Figure 7.5 illustre l'intégration proposée de la normalisation des moments et/ou des stratégies d'appariement (en tant que module de prétraitement) dans le cadre d'adaptation du domaine contradictoire dans ses trois étapes.



(a) Pré-apprentissage.



(b) Adaptation.



(c) Inférence.

Figure 7.5: Intégration proposée de stratégies de normalisation et/ou d'appariement des moments et de l'adaptation de domaine antagoniste.

Nous effectuons la normalisation et l'appariement des moments dans deux scénarios : *indépendant de l'appareil*, dans lequel nous considérons les appareils B et C comme un domaine et transformons les données en utilisant les statistiques moyennes des deux appareils, et *dépendant de l'appareil*, dans lequel nous considérons B et C comme deux domaines distincts et transformons les données en utilisant leurs statistiques respectives.

7.4.5 Résultats

Les résultats de notre étude sont présentés dans le Tableau 7.2. Nous rapportons la précision moyenne de l'ASC et l'écart-type obtenus par toutes les stratégies dans les domaines source et cible.

Impact de la différence de domaine initiale

Nous commençons par noter que la précision du système dans des conditions de test différentes de celles de l'apprentissage chute considérablement à 13,6% sur des données du domaine cible non normalisées (1ère ligne du Tableau 7.2). Une différence absolue de 46% entre les performances des domaines source et cible montre à quel point des conditions d'enregistrement différentes peuvent être graves pour un modèle d'ASC.

Impact de la normalisation et l'appariement des moments

Le pré-apprentissage d'un modèle sur des données normalisées (MN ou MVN) et l'application de la même normalisation lors de l'inférence sont suffisants pour réduire en grande partie le

Table 7.2: Précision d'ASC moyenne (%) et écart-type (entre parenthèses) sur l'ensemble de test obtenus sur 20 apprentissages. Les nombres en gras indiquent les meilleurs résultats statistiquement significatifs ($p < 0.05$) dans le domaine cible.

Méthode			Ind. de l'appareil		Dép. de l'appareil	
Préapprentissage	Adaptation	Inférence	source	cible	source	cible
-	-	-	59.3(5.1)	13.6(2.3)	N/S	N/S
-	ADA (Drossos et al., 2019)	-	62.3(2.0)	36.2(2.6)	N/S	N/S
-	CADA	-	61.0(2.7)	39.4(2.7)	N/S	N/S
MN	-	MN	62.5(2.0)	43.5(2.3)	62.1(2.1)	51.1(2.4)
MN	MN-ADA	MN	64.7(1.0)	50.8(1.6)	64.0(2.8)	58.5(1.2)
MN	MN-CADA	MN	64.5(1.1)	51.6(1.3)	64.2(1.3)	59.0(1.6)
MVN	-	MVN	62.3(1.9)	41.8(1.0)	62.5(2.1)	51.0(2.0)
MVN	MVN-ADA	MVN	64.8(1.4)	**52.0(1.2)**	64.7(1.3)	**59.3(1.4)**
MVN	MVN-CADA	MVN	64.3(1.3)	**52.7(1.6)**	65.0(1.4)	**60.0(1.2)**
-	-	MM	59.3(5.1)	37.3(5.5)	59.3(5.1)	50.0(3.5)
-	ADA	MM	62.3(2.0)	41.3(2.0)	62.3(2.0)	52.0(2.7)
-	MM-ADA	MM	62.1(2.8)	47.0(2.3)	62.2(2.0)	56.2(1.8)
-	CADA	MM	61.0(6.5)	40.3(2.0)	61.0(6.5)	51.9(2.6)
-	MM-CADA	MM	62.5(1.5)	48.9(2.2)	62.7(1.2)	57.7(1.7)
-	-	MVM (Mezza et al., 2021)	59.3(5.1)	38.8(2.9)	59.3(5.1)	50.1(3.5)
-	ADA	MVM	62.3(2.0)	38.2(2.2)	62.3(2.0)	51.0(2.3)
-	MVM-ADA	MVM	63.7(1.1)	47.4(1.9)	62.6(1.5)	56.5(1.7)
-	CADA	MVM	62.4(2.8)	38.1(1.8)	62.4(2.8)	51.2(2.1)
-	MVM-CADA	MVM	63.0(1.3)	50.4(1.2)	63.2(1.6)	58.5(2.0)
MN	-	MNVM	64.2(2.4)	42.3(4.3)	63.1(1.8)	51.8(2.2)
MN	MN-ADA	MNVM	64.7(1.0)	48.6(1.9)	64.0(2.22)	56.0(1.7)
MN	MN-CADA	MNVM	64.0(1.6)	49.4(1.2)	64.0(1.6)	57.2(1.3)
MN	MNVM-ADA	MNVM	64.2(1.8)	50.6(1.1)	64.2(1.0)	**59.9(1.6)**
MN	MNVM-CADA	MNVM	64.0(1.9)	51.7(1.1)	64.6(1.8)	**60.1(1.8)**

fossé entre les distributions source et cible. De même, l'adaptation d'un système pré-entraîné non normalisé par l'appariement des moments (MM ou MVM) augmente la précision dans le domaine cible.

Impact de l'adaptation de domaine antagoniste

Les méthodes d'adaptation de domaine antagonistes avec des données non-normalisées améliorent la précision jusqu'à 26% en valeur absolue dans le domaine cible. En général, CADA obtient une précision moyenne supérieure à ADA, qui est statistiquement significative lorsque la différence de performances dépasse 1,1% ($p < 0,05$).

Intégration de la normalisation et/ou l'appariement des moments avec ADA/CADA

En intégrant la normalisation ou l'appariement des moments avec des méthodes antagonistes, l'écart entre les domaines source et cible est encore réduit. La meilleure précision obtenue dans le domaine cible est de 53% dans le scénario indépendant de l'appareil et de 60% dans celui dépendant de l'appareil. Une telle précision est obtenue en normalisant les données source et cible pendant l'adaptation, quel que soit le procédé d'adaptation.

La Figure 7.6 montre la précision moyenne obtenue par toutes les stratégies d'adaptation

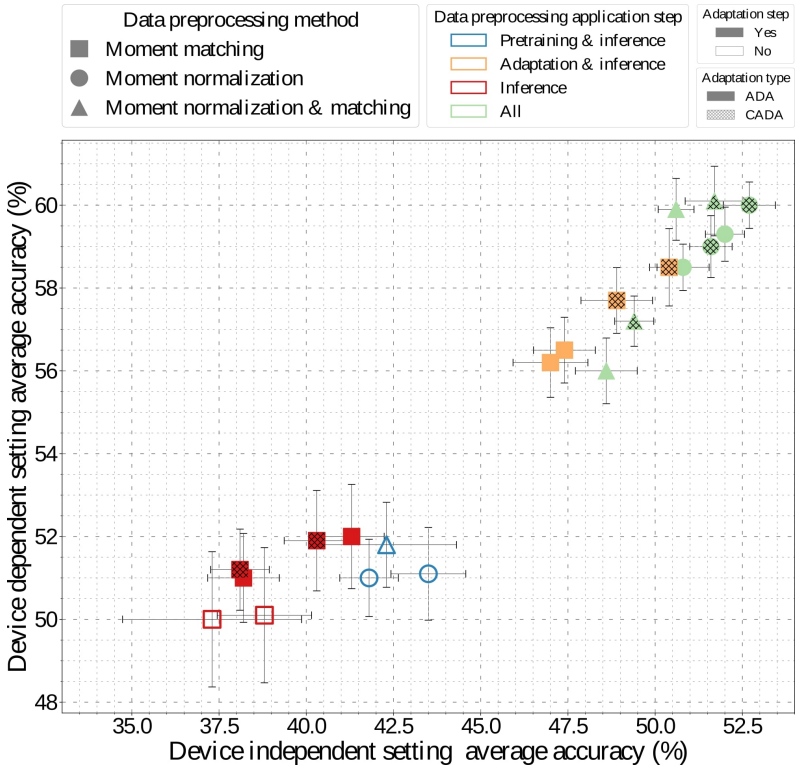


Figure 7.6: Précision moyenne sur le domaine cible dans le scénario indépendant de l'appareil et précision moyenne correspondante dans le scénario dépendant de l'appareil pour les stratégies d'adaptation dans le Tableau 7.2. Les barres d'erreur indiquent des intervalles de confiance de 95%.

proposées dans les deux scénarios. Combiner la normalisation et/ou l'appariement des moments avec l'adaptation au domaine de l'adversaire est la meilleure stratégie. Cette stratégie est capable d'atteindre les performances sur le domaine source.

7.4.6 Conclusion

Nous avons évalué expérimentalement l'impact des stratégies de normalisation et d'appariement des moments ainsi que leur intégration avec des méthodes d'adaptation de domaine antagonistes pour la classification de scènes sonores enregistrées par des dispositifs d'enregistrement différents. Nous avons montré que les stratégies de normalisation sont des exemples particuliers d'un modèle de distorsion linéaire qui améliore la robustesse aux dispositifs d'enregistrement différents. La combinaison de la normalisation des moments et/ou des stratégies d'appariement avec des méthodes d'adaptation de domaine contradictoires réduit l'écart restant dû aux effets non-linéaires. Les résultats indiquent qu'une telle intégration atteint une performance dans le domaine cible proche de celle obtenue dans le domaine source.

7.5 Amélioration de la détection des événements sonores

Nous proposons une stratégie d'adaptation de domaine basée sur le transport optimal (Courty et al., 2017a; Damodaran et al., 2018) pour réduire explicitement le fossé entre les enregistrements

synthétiques et réels. Notre stratégie aligne les distributions empiriques des représentations des caractéristiques des trames actives et inactives des données synthétiques et réelles. Nous étendons et combinons le cadre d'adaptation de domaine proposé avec une stratégie d'apprentissage actif pour améliorer encore les scores de détection. Ce travail nous a amenés à explorer d'autres moyens d'augmenter la robustesse, tels qu'un prétraitement approprié et l'utilisation d'informations auxiliaires.

7.5.1 Normalisation d'énergie par canal parallèle apprenable

De manière similaire à ci-dessus, nous explorons PCEN comme un prétraitement optionnel des spectrogrammes log-Mel. Nous proposons l'utilisation de plusieurs transformations PCEN en tant que couches de réseaux de neurones, dont les paramètres sont appris lors de l'apprentissage du modèle de SED. Nous comparons cette approche avec des transformations PCEN fixes et des spectrogrammes log-Mel.

7.5.2 Classification avant-plan/arrière-plan

Soient $\mathcal{X}$, $\mathcal{Y}$ et $\mathcal{Z}$ les espaces d'entrée, de sortie et l'espace latent. Nous notons la représentation temps-fréquence de la scène sonore par $x \in \mathcal{X}$ avec les annotations correspondantes $y \in \mathcal{Y}$. Nous avons accès à un jeu de données synthétique avec des annotations fortes $\mathcal{D}^{\mathcal{S}} = \{(x_i^{\mathrm{s}}, y_i^{\mathrm{s}})\}_{i=1}^{n_{\mathrm{s}}}$ et deux jeux d'enregistrements réels : un jeu de données faiblement annoté $\mathcal{D}^{\mathcal{W}} = \{(x_i^{\mathrm{w}}, y_i^{\mathrm{w}})\}_{i=1}^{n_{\mathrm{w}}}$ et un jeu de données non annoté $\mathcal{D}^{\mathcal{U}} = \{x_{je}^{\mathrm{u}}\}_{i=1}^{n_{\mathrm{u}}}$.

Le modèle de SED est un modèle Mean Teacher (Tarvainen and Valpola, 2017). Nous utilisons le CRNN du modèle étudiant comme une transformation $g : \mathcal{X} \to \mathcal{Z}$ des spectrogrammes log-Mel vers l'espace latent. Le modèle de SED est représenté par la fonction $f : \mathcal{Z} \to \mathcal{Y}$ qui associe les vecteurs latents à l'espace de sortie.

Branche FB

Nous proposons un classifieur auxiliaire avant-plan/arrière-plan (FB) $f_{\mathrm{FB}} : \mathcal{Z} \to \mathcal{Y}^{\mathrm{FB}}$ qui transforme l'espace latent en classes avant-plan/arrière-plan. Nous apprenons ce classifieur conjointement avec le modèle SED de manière multitâche. De manière analogue, pour le modèle enseignant, nous désignons par g', f' et f'_{FB} la fonction d'intégration CRNN, la branche SED et la branche FB, respectivement.

Nous avons dérivé des annotations de vérité terrain y_i^{fb} à partir des annotations fortes y_i^{s} des données synthétiques en combinant les classes d'événements sonores en deux catégories : avant-plan (*alarme - sonnerie, parole, chat, chien, plats*) et arrière-plan (*mixeur, aspirateur, friture, rasoir électrique - brosse à dents, eau courante*) . Le modèle SED et le classifieur FB sont optimisés en minimisant

$$\begin{aligned}\mathcal{L}_{\mathrm{SED}} = L(y_i^{\mathrm{s}}, f(g(x_i^{\mathrm{s}}))) + \lambda L_{\mathrm{strong}}(f(g(x_i)), f'(g'(x_i))) + \\ L(y_i^{\mathrm{w}}, f(g(x_i^{\mathrm{w}}))) + \lambda L_{\mathrm{weak}}(f(g(x_i)), f'(g'(x_i))) + \\ L(y_i^{\mathrm{fb}}, f_{\mathrm{FB}}(g(x_i^{\mathrm{s}}))) + \lambda L_{\mathrm{strong}}(f_{\mathrm{FB}}(g(x_i)), f'_{\mathrm{FB}}(g'(x_i)))\end{aligned} \quad (7.19)$$

où $L(\cdot, \cdot)$ est un coût de classification d'entropie croisée binaire, et $L_{\mathrm{strong}}(\cdot, \cdot)$ et $L_{\mathrm{weak}}(\cdot, \cdot)$ sont des coûts de cohérence d'erreur quadratique moyenne qui sont différenciables sur leur deuxième paramètre sur les scores forts (au niveau de l'image) et faibles (au niveau du clip), respectivement. Le même poids de cohérence λ est lié à tous les coûts de cohérence.

Branche SEDFB

Nous explorons la fusion de la branche de classification FB avec la branche SED en une branche de détection (SEDFB) pour affiner les sorties. La branche SEDFB est représentée par une fonction $f_{\text{SEDFB}} : \mathcal{Y} \times \mathcal{Y}^{\text{FB}} \to \mathcal{Y}$ (f'_{SEDFB} pour le modèle enseignant). L'entrée de la branche SEDFB est le produit extérieur des sorties des branches SED et FB $w_i = f(g(x_i)) \otimes f_{\text{FB}}(g(x_i))$, car cette fusion crée une représentation contenant des informations issues de l'interaction conjointe des classifieurs SED et FB. Les deux coûts de classification et de cohérence suivants sont ajoutés à l'objectif d'apprentissage (7.19) :

$$\mathcal{L}_{\text{SEDFB}} = L(y_i^{\text{s}}, f_{\text{SEDFB}}(w_i^{\text{s}})) + \lambda L_{\text{strong}}(f_{\text{SEDFB}}(w_i), f'_{\text{SEDFB}}(w_i)). \tag{7.20}$$

Le coût global avec la branche SEDFB est donné par

$$\mathcal{L} = \mathcal{L}_{\text{SED}} + \mathcal{L}_{\text{SEDFB}}. \tag{7.21}$$

7.5.3 Adaptation de domaine pour la détection d'événements sonores

Du point de vue de l'adaptation de domaine non supervisée, nous considérons le jeu de données synthétique avec des annotations fortes comme le domaine source $\mathcal{S} = \mathcal{D}^{\mathcal{S}}$, et la combinaison d'enregistrements réels des ensembles de données faiblement annoté et non annoté comme le domaine cible $\mathcal{T} = \mathcal{D}^{\mathcal{W}} \cup \mathcal{D}^{\mathcal{U}}$. Nous notons x^{s} et x^{t} les scènes sonores de $\mathcal{S}$ et $\mathcal{T}$, respectivement. Nous adoptons le cadre de la méthode DeepJDOT (Damodaran et al., 2018).

Transport optimal de distribution conjointe

Soient $\mu_{\text{s}} = \sum_{i=1}^{n_{\text{s}}} a_i \delta_{g(x_i^{\text{s}}), y_i^{\text{s}}}$ et $\mu_{\text{t}} = \sum_{i=1}^{n_{\text{t}}} b_i \delta_{g(x_i^{\text{t}}), y_i^{\text{t}}}$ deux distributions empiriques sur l'espace produit $\mathcal{Z} \times \mathcal{Y}$, où $\delta_{g(x_i), y_i}$ est la fonction de Dirac à la position $(g(x_i), y_i) \in \mathcal{Z} \times \mathcal{Y}$, et a_i et b_i sont des poids de probabilité (uniformes), c'est-à-dire $\sum_{i=1}^{n_{\text{s}}} a_{je} = \sum_{i=1}^{n_{\text{t}}} b_i = 1$. Le coût associé pour déplacer le i-ième élément source vers le j-ième élément cible peut être exprimé comme la combinaison pondérée de coûts dans l'espace latent et sur les annotations

$$d(g(x_i^{\text{s}}), y_i^{\text{s}}; g(x_j^{\text{t}}), y_j^{\text{t}}) = \alpha c(g(x_i^{\text{s}}), g(x_j^{\text{t}})) + \beta \mathcal{L}(y_i^{\text{s}}, y_j^{\text{t}}) \tag{7.22}$$

où $c(\cdot, \cdot)$ est la distance ℓ_2 au carré, $\mathcal{L}(\cdot, \cdot)$ est un coût d'entropie croisée qui impose le lien entre la source et les annotations du domaine cible, et α et β sont deux valeurs scalaires positives. Comme aucune annotation y_j^{t} n'est disponible dans le domaine cible, elle est remplacée par des pseudo-annotations $f(g(x_j^{\text{t}}))$ obtenues à partir du classifieur $f : \mathcal{Z} \to \mathcal{Y}$. On cherche un couplage de transport $\gamma \in \mathbb{R}^{n_{\text{s}} \times n_{\text{t}}}$ dans l'espace $\Gamma(\mu_{\text{s}}, \mu_{\text{t}})$ des distributions de probabilité conjointes avec des distributions marginales $\gamma \mathbf{1}_{n_{\text{t}}} = \mu_{\text{s}}$ et $\gamma^T \mathbf{1}_{n_{\text{s}}} = \mu_{\text{t}}$, où $\mathbf{1}_d$ est un vecteur de valeurs égales à 1 de dimension d et les fonctions g et f minimisent

$$\min_{\gamma \in \Gamma(\mu_{\text{s}}, \mu_{\text{t}}), g, f} \sum_{i,j} \gamma_{i,j} d(g(x_i^{\text{s}}), y_i^{\text{s}}; g(x_j^{\text{t}}), f(g(x_j^{\text{t}}))). \tag{7.23}$$

Nous suivons une procédure en deux étapes pour résoudre ce problème d'optimisation. Dans la première étape, nous calculons la matrice de couplage optimale γ avec des paramètres de modèle fixes f et g:

$$\min_{\gamma \in \Gamma(\mu_{\text{s}}, \mu_{\text{t}})} \sum_{i,j} \gamma_{ij} (\alpha ||g(x_i^{\text{s}}) - g(x_j^{\text{t}})||^2 + \beta \mathcal{L}(y_i^{\text{s}}, f(g(x_j^{\text{t}})))). \tag{7.24}$$

Dans la deuxième étape, avec γ fixe, nous mettons à jour les modèles g et f comme

$$\min_{g,f} \mathcal{L}_s + \sum_{i,j} \gamma_{ij} (\alpha ||g(x_i^{\mathrm{s}}) - g(x_j^{\mathrm{t}})||^2 + \beta \mathcal{L}(y_i^{\mathrm{s}}, f(g(x_j^{\mathrm{t}})))) \tag{7.25}$$

où $\mathcal{L}_s$ est le coût de la classification sur le domaine source pour éviter de perdre en performances sur les données synthétiques.

Coût d'apprentissage

Nous notons $\hat{z}^{\mathrm{s}}$ et $\hat{z}^{\mathrm{t}}$ les trames actives échantillonnées et $\bar{z}^{\mathrm{s}}$ et $\bar{z}^{\mathrm{t}}$ les trames inactives échantillonnées à partir des représentations latentes des domaines source et cible z^{s} et z^{t}, respectivement. Après une première étape de préapprentissage utilisant (7.21), nous construisons la fonction de coût suivante pour tenir compte de la différence entre les distributions empiriques des représentations de caractéristiques apprises actives et inactives:

$$\mathcal{L}_{\mathrm{s}} + \mathcal{L}_{\mathrm{active}} + \mathcal{L}_{\mathrm{inactive}} \tag{7.26}$$

où

$$\mathcal{L}_{\mathrm{s}} = \frac{1}{n_{\mathrm{s}}} \sum_{i=1}^{n_{\mathrm{s}}} L(y_i^{\mathrm{s}}, f(g(x_i^{\mathrm{s}}))) + \frac{1}{n_{\mathrm{s}}} \sum_{i=1}^{n_{\mathrm{s}}} L(y_i^{\mathrm{fb}}, f_{\mathrm{FB}}(g(x_i^{\mathrm{s}}))) \tag{7.27}$$

correspond aux premier et troisième termes du coût de classification (7.19). Comme seul le modèle étudiant subit une adaptation, aucun coût de cohérence n'est inclus dans $\mathcal{L}_{\mathrm{s}}$. La fonction de coût $\mathcal{L}_{\mathrm{active}}$ est le coût d'alignement de distribution pour les trames actives

$$\mathcal{L}_{\mathrm{active}} = \frac{1}{|C_{\mathrm{active}}|} \sum_{i,j}^{N_{\mathrm{active}}} \gamma_{ij}^{\mathrm{active}} (\alpha ||\hat{z}_i^{\mathrm{s}} - \hat{z}_i^{\mathrm{t}}||^2 + \beta \mathcal{L}(y_i^{\mathrm{s}}, \hat{y}_j^{\mathrm{t}})) \tag{7.28}$$

où $|C_{\mathrm{active}}|$ est la cardinalité du sous-ensemble d'annotations $C_{\mathrm{active}} \in C$ représentant le nombre total de classes actives dans le lot. Le deuxième terme dans (7.28) renforce la cohérence entre les pseudo-annotations du domaine cible et les annotations du domaine source. La fonction de coût $\mathcal{L}_{\mathrm{inactive}}$ tient compte de l'alignement des distributions marginales des représentations apprises des trames inactives dans les deux domaines :

$$\mathcal{L}_{\mathrm{inactive}} = \sum_{i=1}^{N_{\mathrm{inactive}}} \gamma_{ij}^{\mathrm{inactive}} ||\bar{z}_i^{\mathrm{s}} - \bar{z}_i^{\mathrm{t}}||^2. \tag{7.29}$$

La Figure 7.7 décrit la stratégie proposée d'adaptation de domaine au niveau de la trame basée sur le transport optimal.

7.5.4 Adaptation de domaine et apprentissage actif pour la détection d'événements sonores

L'utilisation de pseudo-annotations pour l'adaptation de domaine limite l'alignement correct des distributions conditionnelles de classe des données synthétiques et réelles. Pour résoudre ce problème, nous explorons une stratégie d'apprentissage actif qui remplace les pseudo-annotations par des annotations demandées à l'utilisateur. Cette stratégie est basée sur la méthode *mismatch-first farthest-traversal* proposée par Shuyang et al. (2020). *Mismatch-first* recherche des échantillons avec des annotations différentes selon deux mécanismes d'annotation différents et *farthest-traversal* maximise la diversité des échantillons sélectionnés.

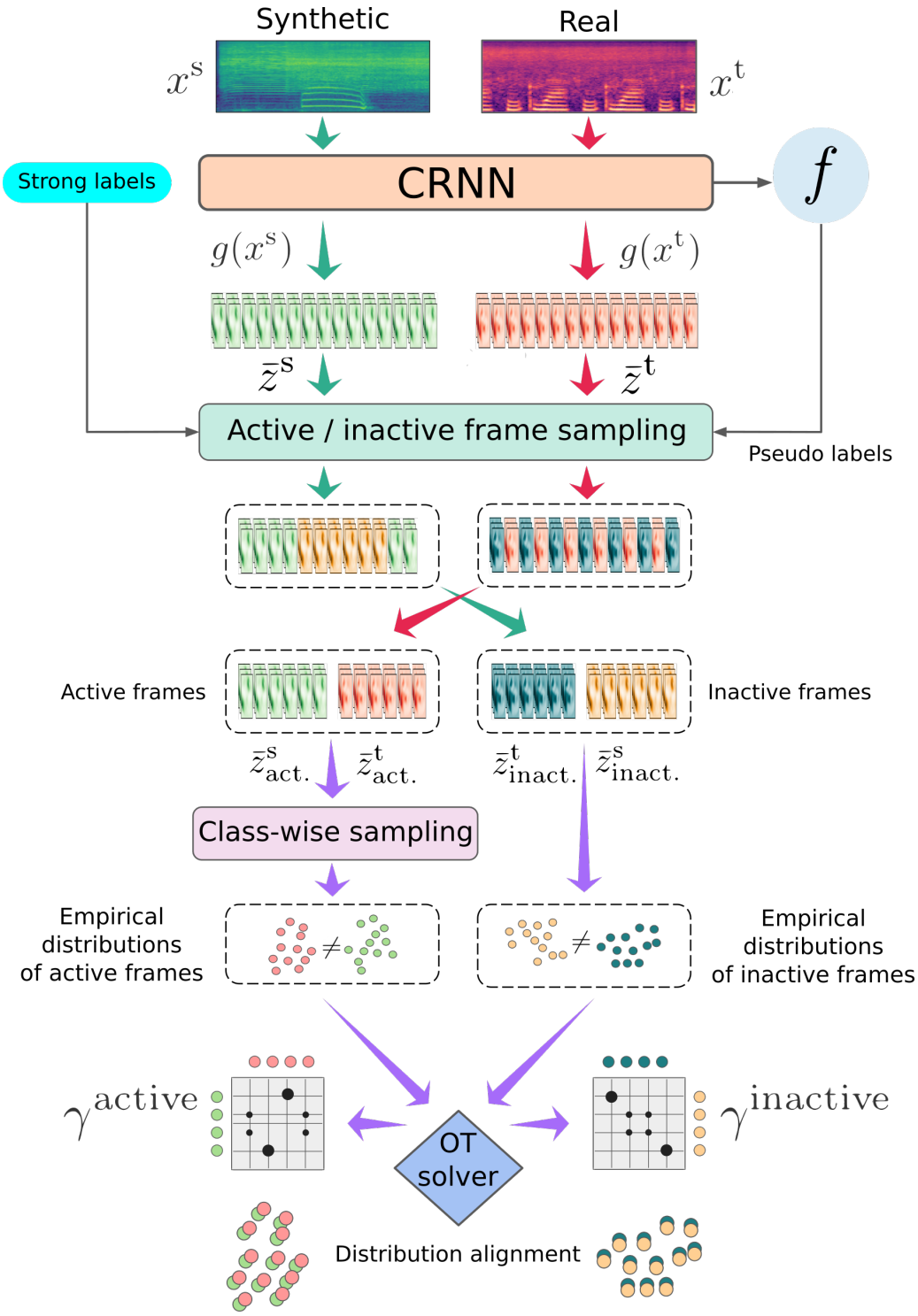


Figure 7.7: Stratégie proposée d'adaptation de domaine basée sur le transport optimal pour corriger la différence entre les données synthétiques et réelles au niveau de la trame.

Adaptation de domaine au niveau du segment

La méthode proposée d'adaptation de domaine au niveau de la trame est étendue au niveau du segment. Les segments actifs sont ceux dans lesquels un événement sonore se produit (comme

indiqué par la vérité terrain ou la pseudo-annotation), tandis que les segments inactifs sont ceux dépourvus d'annotations. L'alignement entre les distributions empiriques des segments actifs et inactifs suit la même procédure que pour l'adaptation au niveau de la trame.

7.5.5 Protocole expérimental

Jeu de données

Nous avons mené des expériences sur le jeu de données DESED (Turpault et al., 2019; Serizel et al., 2020). Nous évaluons les performances des méthodes proposées sur les ensembles de validation et d'évaluation public (Turpault et al., 2019; Serizel et al., 2020), en termes de score F1 macro.

Architecture et apprentissage du modèle

L'architecture du modèle sélectionné correspond au système de référence du Challenge DCASE 2020. Les extensions incluent une branche de classification avant-plan/arrière-plan (FB) formée d'une couche dense avec activation sigmoïde, et une branche de détection d'événements sonores modifiée (SEDFB) composée d'une couche récurrente bidirectionnelle. Le modèle a été appris de façon semi-supervisée avec la méthode Mean Teacher habituelle. Seul le modèle étudiant a été adapté.

Post-traitement de la sortie

Nous comparons deux méthodes pour post-traiter les sorties. La première méthode est un filtre médian (MF) de 0.45 s. La deuxième approche est un décodage par modèle de Markov caché (HMM). En suivant la même procédure que Cornell et al. (2020a), nous avons déterminé les probabilités de transition optimales pour chaque classe d'événements sonores à l'aide de l'ensemble de validation.

7.5.6 Résultats

Impact du prétraitement PCEN apprenable

Dans le Tableau 7.3, nous rapportons les performances obtenues sur l'ensemble de développement par un modèle de base entraîné sur des spectrogrammes log-Mel et avec les différents prétraitements basés sur PCEN. Nous observons que plusieurs transformations PCEN apprenables (P-PCEN) appliquées à des spectrogrammes Mel ou log-Mel surpassent les transformations PCEN avec des paramètres fixes et les spectrogrammes log-Mel seuls. L'apprentissage des paramètres de deux transformations PCEN s'avère suffisant.

Impact de l'adaptation de domaine

Les résultats de l'adaptation de domaine sont présentés dans le Tableau 7.4. La meilleure adaptation est obtenue en alignant simultanément les distributions empiriques des trames actives et inactives des données réelles et synthétiques. Aligner uniquement les distributions des trames actives ou inactives est sous-optimal.

Impact de la classification avant-plan/arrière-plan et du post-traitement de la sortie

Le Tableau 7.5 montre les résultats obtenus par les méthodes proposées sur les ensembles de validation et d'évaluation public du jeu de données DESED.

L'ajout de la branche FB est bénéfique car il améliore les résultats sur l'ensemble de validation de 8,3% dans l'absolu. Le même avantage est démontré avec la fusion de la branche FB avec la branche SED dans une branche de détection améliorée (SEDFB). Cette grande amélioration des performances valide l'utilisation d'informations auxiliaires sur la nature spectro-temporelle des événements sonores considérés dans les scénarios domestiques. La combinaison des branches de classification proposées avec l'adaptation de domaine apporte des avantages supplémentaires à la tâche.

Table 7.3: Comparaison des différentes transformations PCEN fixes ou apprenables.

Représentation d'entrée	score F1 (%)
Log-Mel	35.93
PCEN fixée 1	28.71
PCEN fixée 2	35.42
PCEN fixée 3	36.84
PCEN apprenable	36.92
P-PCEN sur Mel	**38.20**
P-PCEN sur log-Mel	**39.34**

Table 7.4: Comparaison des méthodes d'adaptation sur l'ensemble de validation.

Méthode d'adaptation	score F1 (%)
Pas d'adaptation	34.8
Trames actives seules	36.93
Trames inactives seules	24.17
Trames actives + inactives	**42.41**
Trames actives seules (oracle)	50.66
Trames inactives seules (oracle)	16.34
Trames actives + inactives (oracle)	70.23

Table 7.5: Comparaison des méthodes de classification avant-plan/arrière-plan et de post-traitement de la sortie sur les ensembles de validation (val) et d'évaluation public (eval).

Méthode	score F1 (%)			
	val +MF	val +HMMs	eval +MF	eval +HMMs
Baseline	34.8	37.6	38.1	41.1
+ DA	42.41	43.89	44.8	47.12
+ FB	43.12	45.42	46.06	49.38
+ FB + DA	45.68	47.77	50.79	53.10
+ SEDFB	46.15	46.20	48.40	49.79
+ SEDFB + DA	**47.61**	47.75	**52.12**	53.30
DCASE 1	45.13	**48.07**	50.58	**53.35**
DCASE 2	45.15	47.08	50.28	52.23

Impact de l'adaptation de domaine au niveau du segment et de l'apprentissage actif

La Figure 7.8 montre les performances de la méthode proposée d'adaptation de domaine et d'apprentissage actif au niveau du segment en fonction du budget d'annotation. Les résultats montrent les performances sur les ensembles de validation et d'évaluation public.

Table 7.6: Performance des modèles de référence et des approches d'adaptation de domaine sur les ensembles de validation et d'évaluation public.

Méthode	score F1 (%)	
	validation	évaluation public
Modèle de référence	34.8	37.6
Modèle de référence + Audioset	40	42.6
+ DA au niveau de la trame	42.41	44.8
+ DA au niveau du segment	**43.26**	**45.08**

L'extension proposée de l'adaptation de domaine du niveau de la trame au niveau du segment conduit à des scores de détection légèrement meilleurs comme le montre le Tableau 7.6. La Figure 7.8 montre que la combinaison de l'adaptation de domaine au niveau du segment avec l'apprentissage actif facilite la généralisation à mesure que le budget d'annotation augmente. Le scénario avec des budgets d'annotation limités est particulièrement intéressant puisque le but est de minimiser l'effort d'annotation.

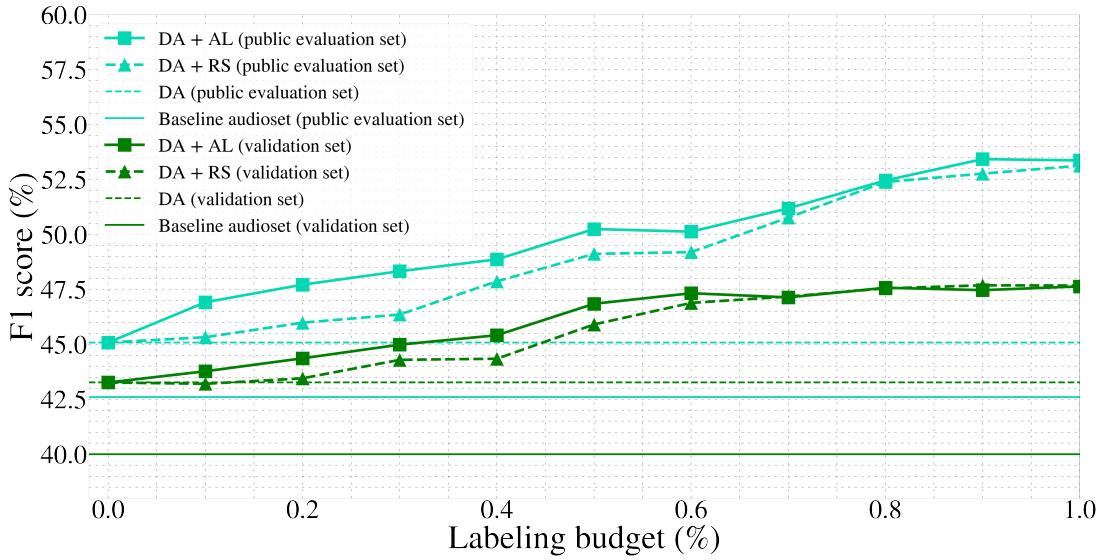


Figure 7.8: Performance de l'adaptation du domaine au niveau du segment et de l'apprentissage actif en fonction du budget d'annotation. Le budget d'annotation est relatif à la quantité d'annotations de vérité terrain disponibles pour les données d'apprentissage réelles (environ 10% des données).

Conclusion

Nous avons présenté plusieurs méthodes complémentaires qui améliorent les modèles de détection d'événements sonores appris sur des données synthétiques et réelles. Nous avons montré que l'apprentissage des paramètres de plusieurs transformations PCEN surpasse les spectrogrammes

log-Mel en tant que caractéristiques d'entrée. De plus, la catégorisation des sons domestiques en fonction de leurs caractéristiques spectro-temporelles en tant qu'événements d'avant-plan ou d'arrière-plan est une information auxiliaire utile qui améliore la généralisation. Nous avons aussi étudié une stratégie d'adaptation de domaine explicite basée sur le transport optimal pour réduire davantage le fossé entre les données synthétiques et réelles. Les résultats indiquent que la correspondance des distributions empiriques des trames actives et inactives et des segments audio des deux types de données pendant l'apprentissage est bénéfique. La classification avant-plan/arrière-plan et les stratégies d'apprentissage actif combinées à l'adaptation de domaine améliorent considérablement la robustesse de la détection.

7.6 Conclusions générales

Dans cette thèse, nous avons abordé des problèmes d'intérêt pratique concernant la séparation de sources audio, la classification de scènes sonores et la détection d'événements sonores. Dans un premier axe de recherche, nous avons étudié la séparation des sons de courte durée avec des caractéristiques spectrales à évolution rapide des sons stationnaires de longue durée. Dans un deuxième axe de recherche, nous avons conçu des méthodes pour rendre les systèmes de classification et de détection audio robustes aux conditions d'apprentissage et de test différentes. Nous avons abordé deux scénarios : la classification de scènes sonores avec des dispositifs d'enregistrement différents et l'apprentissage de systèmes de détection d'événements sonores avec des données synthétiques et réelles. Dans le premier scénario, nous avons montré que la correction des caractéristiques des données de test combinée à une adaptation basée sur l'apprentissage atténue les effets des microphones différents. Dans le deuxième scénario, nous avons montré que l'intégration de l'adaptation de domaine à une stratégie d'apprentissage semi-supervisé pour la détection d'événements sonores conduit à des représentations plus invariantes aux données synthétiques et réelles. Notre exploration de ces scénarios nous a amenés à rechercher d'autres moyens d'atteindre la robustesse, tels qu'un prétraitement approprié et l'utilisation d'informations auxiliaires. Cette thèse visait à résoudre les difficultés que les systèmes d'analyse audio rencontrent souvent dans des scénarios défavorables.

Bibliography

A. Alaoui-Belghiti, S. Chevallier, and E. Monacelli. Unsupervised anomaly detection using optimal transport for predictive maintenance. In *International Conference on Artificial Neural Networks*, pages 686–697, 2019.

R. M. Alsina-Pagès, J. Navarro, F. Alías, and M. Hervás. homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring. *Sensors*, 17(4):854, 2017.

D. Alvarez-Melis, S. Jegelka, and T. S. Jaakkola. Towards optimal transport with global invariances. In *22nd International Conference on Artificial Intelligence and Statistics*, pages 1870–1879, 2019.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

P. Arora and R. Haeb-Umbach. A study on transfer learning for acoustic event detection in a real life scenario. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2017.

D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*, 32(3):16–34, 2015.

B. Bayram and G. İnce. An incremental class-learning approach with acoustic novelty detection for acoustic event recognition. *Sensors*, 21(19):6622, 2021.

J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy. Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. *Communications of the ACM*, 62(2):68–77, 2019.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19, page 137–144, 2006.

Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović. A framework for the robust evaluation of sound event detection. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 61–65, 2020.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

A. L. Borker, M. W. McKown, J. T. Ackerman, C. A. Eagles-Smith, B. R. Tershy, and D. A. Croll. Vocal activity as a low cost and scalable index of seabird colony size. *Conservation Biology*, 28(4):1100–1108, 2014.

K. Brinker. Incorporating diversity in active learning with support vector machines. In *20th International Conference on Machine Learning (ICML)*, pages 59–66, 2003.

E. Cakır, 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*, 25(6):1291–1303, 2017.

L. Cances, T. Pellegrini, and P. Guyot. Multi task learning and post processing optimization for sound event detection. *Tech. Rep., DCASE2019 Challenge*, 2019.

E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.-R. Stöter. Musical source separation: An introduction. *IEEE Signal Processing Magazine*, 36(1):31–40, 2018.

M. Cantarini, L. Gabrielli, and S. Squartini. Few-shot emergency siren detection. *Sensors*, 22 (12):4338, 2022.

C.-F. Chan and W. Eric. An abnormal sound detection and classification system for surveillance applications. In *2010 18th European Signal Processing Conference*, pages 1851–1855, 2010.

C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu, and J. Huang. Progressive feature alignment for unsupervised domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019a.

L. Chen, Y. Zhang, R. Zhang, C. Tao, Z. Gan, H. Zhang, B. Li, D. Shen, C. Chen, and L. Carin. Improving sequence-to-sequence learning via optimal transport. *arXiv preprint arXiv:1901.06283*, 2019b.

K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *8th Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014.

G. Ciaburro and G. Iannace. Improving smart cities safety using sound events detection based on deep neural network algorithms. *Informatics*, 7(3), 2020.

A. Copiaco, C. Ritz, N. Abdulaziz, and S. Fasciani. A study of features and deep neural network architectures and hyper-parameters for domestic audio classification. *Applied Sciences*, 11(11): 4880, 2021.

S. Cornell, M. Olvera, M. Pariente, G. Pepe, E. Principi, L. Gabrielli, and S. Squartini. Domain-adversarial training and trainable parallel front-end for the DCASE 2020 Task 4 sound event detection challenge. In *5th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 26–30, 2020a.

S. Cornell, M. Olvera, M. Pariente, G. Pepe, E. Principi, L. Gabrielli, and S. Squartini. Task-aware separation for the DCASE 2020 Task 4 sound event detection and separation challenge. In *5th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 31–35, 2020b.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289, 2014.

N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 30, page 3733–3742, 2017a.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017b. doi: 10.1109/TPAMI.2016.2615921.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017c.

J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856, 2019.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, page 2292–2300, 2013.

B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *European Conference on Computer Vision (ECCV)*, pages 447–463, 2018.

DCASE. DCASE community, 2022. URL https://dcase.community/.

D. de Benito-Gorron, D. Ramos, and D. T. Toledano. A multi-resolution approach to sound event detection in DCASE 2020 Task 4. In *5th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 36–40, 2020.

M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, and S. Araki. Soundbeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning. *arXiv preprint arXiv:2204.03895*, 2022.

L. Delphin-Poulat, R. Nicol, C. Plapous, and K. Peron. Comparative assessment of data augmentation for semi-supervised polyphonic sound event detection. In *2020 27th Conference of Open Innovations Association (FRUCT)*, pages 46–53, 2020.

A. Diment and T. Virtanen. Transfer learning of weakly labelled audio. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 6–10, 2017.

C. N. Doukas and I. Maglogiannis. Emergency fall incidents detection in assisted living environments utilizing motion, sound, and visual perceptual components. *IEEE Transactions on Information Technology in Biomedicine*, 15(2):277–289, 2010.

K. Drossos, P. Magron, and T. Virtanen. Unsupervised adversarial domain adaptation based on the Wasserstein distance for acoustic scene classification. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 259–263, 2019.

H. Dubey, D. Emmanouilidou, and I. J. Tashev. In *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, pages 1–6, 2019.

T. B. Duman, B. Bayram, and G. İnce. Acoustic anomaly detection using convolutional autoencoders in industrial processes. In *International Workshop on Soft Computing Models in Industrial and Environmental Applications*, pages 432–442, 2019.

J. Ebbers, R. Haeb-Umbach, and R. Serizel. Threshold independent evaluation of sound event detection scores. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1021–1025, 2022.

B. Elizalde, A. Shah, S. Dalmia, M. H. Lee, R. Badlani, A. Kumar, B. Raj, and I. Lane. An approach for self-training audio event detectors using web data. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1863–1867, 2017.

A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia. A brief review of domain adaptation. In *Advances in Data Science and Information Engineering*, pages 877–894. Springer, 2021.

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

N. H. Fletcher. Animal bioacoustics. In *Springer Handbook of Acoustics*, pages 821–841. Springer, 2014.

A. Fleury, N. Noury, M. Vacher, H. Glasson, and J.-F. Seri. Sound and speech detection and classification in a health smart home. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4644–4647, 2008.

P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento. Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE Transactions on Intelligent Transportation Systems*, 17(1):279–288, 2015.

E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2017.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, volume 28, page 2053–2061, 2015.

D. Gabor. Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers — Part III: Radio and Communication Engineering*, 93 (26):429–441, 1946.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun. Ota: Optimal transport assignment for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 303–312, 2021.

J. F. Gemmeke, D. P. 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 *2017*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 776–780, 2017.

G. Gendron, R. Tremblay, A. Jolivet, F. Olivier, L. Chauvaud, G. Winkler, and C. Audet. Anthropogenic boat noise reduces feeding success in winter flounder larvae (pseudopleuronectes americanus). *Environmental Biology of Fishes*, 103(9):1079–1090, 2020.

B. Gfeller, D. Roblek, and M. Tagliasacchi. One-shot conditional audio filtering of arbitrary sounds. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 501–505, 2021.

S. Gharib, K. Drossos, E. Cakir, D. Serdyuk, and T. Virtanen. Unsupervised adversarial domain adaptation for acoustic scene classification. In *3rd Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 138–142, 2018.

P. Giannoulis, G. Potamianos, and P. Maragos. Overlapped sound event classification via multichannel sound separation network. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 571–575, 2021.

G. Gravier, J. Kharroubi, and G. Chollet. On the use of prior knowledge in normalization schemes for speaker verification. *Digital Signal Processing*, 10(1-3):213–225, 2000.

S. Grollmisch and E. Cano. Improving semi-supervised learning for audio classification with fixmatch. *Electronics*, 10(15):1807, 2021.

X. Gu, R. Li, M. Kang, F. Lu, D. Tang, and J. Peng. Unsupervised adversarial domain adaptation abnormal sound detection for machine condition monitoring under domain shift conditions. In *2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 139–146, 2021.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.

M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, pages 1–38, 2022.

Y. Guo, N. C. Codella, L. Karlinsky, J. V. Codella, J. R. Smith, K. Saenko, T. Rosing, and R. Feris. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision*, pages 124–141, 2020.

W. Han, E. Coutinho, H. Ruan, H. Li, B. Schuller, X. Yu, and X. Zhu. Semi-supervised active learning for sound classification in hybrid learning environments. *PloS one*, 11(9):e0162075, 2016.

N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito. ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions. In *6th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 1–5, 2021.

G. Heinzel, A. Rüdiger, and R. Schilling. Spectrum and spectral density estimation by the discrete fourier transform (DFT), including a comprehensive list of window functions and some new at-top windows. Technical report, Max-Planck-Institut für Gravitationsphysik, 2002.

T. Heittola, A. Mesaros, and T. Virtanen. Acoustic scene classification in DCASE 2020 Challenge: generalization across devices and low complexity solutions. In *5th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 56–60, 2020.

S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal. The benefit of temporally-strong labels in audio event classification. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370, 2021.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

H. Hu, S. M. Siniscalchi, Y. Wang, and C.-H. Lee. Relational teacher student learning with neural label embedding for device adaptation in acoustic scene classification. In *Interspeech*, pages 1196–1200, 2020.

H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu, F. Bao, Y. Zhao, S. M. Siniscalchi, Y. Wang, J. Du, and C.-H. Lee. A two-stage approach to device-robust acoustic scene classification. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 845–849, 2021a.

Z. Hu, Z. Yang, X. Hu, and R. Nevatia. Simple: similar pseudo label exploitation for semi-supervised classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15099–15108, 2021b.

P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Deep learning for monaural speech separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1562–1566, 2014.

W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas. Scream detection for home applications. In *2010 5th IEEE Conference on Industrial Electronics and Applications*, pages 2115–2120, 2010.

Y. Huang, L. Lin, S. Ma, X. Wang, H. Liu, Y. Qian, M. Liu, and K. Ouchi. Guided multi-branch learning systems for sound event detection with sound separation. In *5th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 61–65, 2020a.

Y. Huang, L. Lin, X. Wang, H. Liu, Y. Qian, M. Liu, and K. Ouchi. Learning generic feature representation with synthetic data for weakly-supervised sound event detection by inter-frame distance loss. *arXiv preprint arXiv:2011.00695*, 2020b.

C. Hummersone, T. Stokes, and T. Brookes. On the ideal ratio mask as the goal of computational auditory scene analysis. In *Blind Source Separation*, pages 349–368. Springer, 2014.

K. Imoto, S. Mishima, Y. Arai, and R. Kondo. Impact of data imbalance caused by inactive frames and difference in sound duration on sound event detection performance. *Applied Acoustics*, 196:108882, 2022.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

A. Jansen, M. Plakal, R. Pandya, D. Ellis, S. Hershey, J. Liu, C. Moore, and R. A. Saurous. Towards learning semantic audio representations from unlabeled data. In *NIPS Workshop on Machine Learning for Audio Signal Processing (ML4Audio)*, 2017.

S. Kacprzak and K. Kowalczyk. Adversarial domain adaptation with paired examples for acoustic scene classification on different recording devices. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 1030–1034, 2021.

E. Kanjo. Noisespy: A real-time mobile phone platform for urban noise monitoring and mapping. *Mobile Networks and Applications*, 15(4):562–574, 2010.

I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey. Universal sound separation. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 175–179, 2019.

K. Kilgour, B. Gfeller, Q. Huang, A. Jansen, S. Wisdom, and M. Tagliasacchi. Text-driven separation of arbitrary sounds. In *Interspeech*, pages 5403–5407, 2022.

B. Kim and B. Pardo. A human-in-the-loop system for sound event detection and annotation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–23, 2018.

B. Kim and B. Pardo. Sound event detection using point-labeled data. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, 2019.

J. Kim, K. Min, M. Jung, and S. Chi. Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition. *Building and Environment*, 181:107092, 2020.

Y. Kim, J. Sa, Y. Chung, D. Park, and S. Lee. Resource-efficient pet dog sound events classification using LSTM-FCN based on time-series data. *Sensors*, 18(11):4019, 2018.

L. N. Kloepper and A. M. Simmons. Bioacoustic monitoring contributes to an understanding of climate change. *Acoustic Today*, pages 8–15, 2014.

P. A. Knight. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

C.-Y. Koh, Y.-S. Chen, Y.-W. Liu, and M. R. Bai. Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 376–380, 2021.

E. Koh, F. Saki, Y. Guo, C.-Y. Hung, and E. Visser. Incremental learning algorithm for sound event detection. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020.

Y. Koizumi, M. Yasuda, S. Murata, S. Saito, H. Uematsu, and N. Harada. Spidernet: Attention network for one-shot anomaly detection in sounds. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 281–285, 2020.

M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913, 2017.

Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley. Sound event detection of weakly labelled data with CNN-Transformer and automatic threshold optimization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2450–2460, 2020.

M. Kośmider. Spectrum correction: acoustic scene classification with mismatched recording devices. In *Interspeech*, pages 4641–4645, 2020.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

M. Kull and P. Flach. Patterns of dataset shift. In *First International Workshop on Learning over Multiple Contexts*, 2014.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

A. Kumar, M. Khadkevich, and C. Fügen. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 326–330, 2018.

A. Kumar, Y. Wang, V. K. Ithapu, and C. Fuegen. Do sound event representations generalize to other audio tasks? A case study in audio transfer learning. In *Interspeech*, pages 1214–1218, 2021.

G. Lafay, M. Lagrange, M. Rossignol, E. Benetos, and A. Roebel. A morphological model for simulating acoustic scenes and its application to sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1854–1864, 2016.

J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey. SDR–half-baked or well done? In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630, 2019.

Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural Networks: Tricks of the Trade*, pages 9–48. Springer, 2012.

J. Lee, T. Kim, J. Park, and J. Nam. Raw waveform-based audio classification using sample-level CNN architectures. In *NIPS Machine Learning for Audio Signal Processing Workshop (ML4Audio)*, 2017.

D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

X. Li. Semi-supervised sound event detection using random augmentation and consistency regularization. *arXiv preprint arXiv:2102.00154*, 2021.

H. Liu, J. Wang, and M. Long. Cycle self-training for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 34, pages 22968–22981, 2021.

P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC press, 2007.

M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1647–1657, 2018.

V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello. Per-channel energy normalization: Why and how. *IEEE Signal Processing Letters*, 26(1): 39–43, 2018.

Y. Luo and N. Mesgarani. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700, 2018.

L. Mahon and T. Lukasiewicz. Selective pseudo-label clustering. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 158–178, 2021.

I. Martín-Morató, M. Harju, and A. Mesaros. Crowdsourcing strong labels for sound event detection. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 246–250, 2021.

A. Mesaros, T. Heittola, and T. Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162, 2016.

A. Mesaros, T. Heittola, and T. Virtanen. Acoustic scene classification: an overview of DCASE 2017 Challenge entries. In *16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 411–415, 2018a.

A. Mesaros, T. Heittola, and T. Virtanen. A multi-device dataset for urban acoustic scene classification. In *3rd Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 9–13, 2018b.

A. Mesaros, T. Heittola, and T. Virtanen. Acoustic scene classification in DCASE 2019 Challenge: Closed and open set classification and data mismatch setups. 2019.

A. I. Mezza, E. A. Habets, M. Müller, and A. Sarti. Feature projection-based unsupervised domain adaptation for acoustic scene classification. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2020.

A. I. Mezza, E. A. Habets, M. Müller, and A. Sarti. Unsupervised domain adaptation for acoustic scene classification using band-wise statistics matching. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 11–15, 2021.

A. I. Mezza, E. A. Habets, M. Müller, and A. Sarti. Unsupervised domain adaptation via principal subspace projection for acoustic scene classification. *Journal of Signal Processing Systems*, 94(2):197–213, 2022.

K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda. Convolution augmented transformer for semi-supervised sound event detection. In *5th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 100–104, 2020.

Z. Mnasri, S. Rovetta, and F. Masulli. Anomalous sound event detection: A survey of machine learning based methods and applications. *Multimedia Tools and Applications*, 81(4):5537–5586, 2022.

R. M. Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI.* Simon and Schuster, 2021.

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

S. Mun and S. Shon. Domain mismatch robust acoustic scene classification using channel information conversion. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 845–849, 2019.

J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, F. Antonacci, and M. Cobos. Open set audio classification using autoencoders trained on few data. *Sensors*, 20(13):3741, 2020.

T. Nguyen and F. Pernkopf. Acoustic scene classification with mismatched devices using cliquenets and mixup data augmentation. In *Interspeech*, pages 2330–2334, 2019.

T. Nguyen, F. Pernkopf, and M. Kosmider. Acoustic scene classification for mismatched recording devices using heated-up softmax and spectrum correction. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 126–130, 2020.

S. Ntalampiras. Automatic analysis of audiostreams in the concept drift environment. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2016.

T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki. Listen to what you want: Neural network-based universal sound selector. In *Interspeech*, pages 1441–1445, 2020.

Y. Okamoto, S. Horiguchi, M. Yamamoto, K. Imoto, and Y. Kawaguchi. Environmental sound extraction using onomatopoeic words. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 221–225, 2022.

M. Olvera, E. Vincent, and G. Gasso. Improving sound event detection with auxiliary foreground-background classification and domain adaptation. In *6th Workshop on Detection and Classification of Acoustic Scenes and Events*, 2021a.

M. Olvera, E. Vincent, R. Serizel, and G. Gasso. Foreground-background ambient sound scene separation. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 281–285, 2021b.

M. Olvera, E. Vincent, and G. Gasso. On the impact of normalization strategies in unsupervised adversarial domain adaptation for acoustic scene classification. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 631–635, 2022.

V. M. Panaretos and Y. Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6:405–431.

G. Parascandolo, H. Huttunen, and T. Virtanen. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6440–6444, 2016.

M. Pariente, S. Cornell, A. Deleforge, and E. Vincent. Filterbank design for end-to-end speech separation. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6364–6368, 2020.

D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, pages 2613–2617, 2019a.

H. Park, S. Yun, J. Eum, J. Cho, and K. Hwang. Weakly labeled sound event detection using tri-training and adversarial learning. In *4th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 184–188, 2019b.

S. E. Parks, M. Johnson, D. Nowacek, and P. L. Tyack. Individual right whales call louder in increased environmental noise. *Biology Letters*, 7(1):33–35, 2011.

J. Picaut, A. Can, N. Fortin, J. Ardouin, and M. Lagrange. Low-cost sensors for urban noise monitoring networks—a literature review. *Sensors*, 20(8):2256, 2020.

K. J. Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2015.

F. Pishdadian, G. Wichern, and J. Le Roux. Finding strength in weakness: Learning to separate sounds with weak supervision. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2386–2399, 2020.

J. Pons, J. Serrà, and X. Serra. Training neural audio classifiers with few data. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16–20, 2019.

A. N. Popper and A. D. Hawkins. An overview of fish bioacoustics and the impacts of anthropogenic sounds on fishes. *Journal of Fish Biology*, 94(5):692–713, 2019.

P. Primus, H. Eghbal-zadeh, D. Eitelsebner, K. Koutini, A. Arzt, and G. Widmer. Exploiting parallel audio recordings to enforce device invariance in CNN-based acoustic scene classification. In *4th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 204–208, 2019.

H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi. Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In *4th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 209–213, 2019.

K. Qian, Z. Zhang, A. Baird, and B. Schuller. Active learning for bird sound classification via a kernel-based extreme learning machine. *The Journal of the Acoustical Society of America*, 142(4):1796–1804, 2017a.

K. Qian, Z. Zhang, A. Baird, and B. Schuller. Active learning for bird sounds classification. *Acta Acustica united with Acustica*, 103(3):361–364, 2017b.

B. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, et al. A respiratory sound database for the development of automated classification. In *International Conference on Biomedical and Health Informatics*, pages 33–37, 2017.

F. Ronchini, R. Serizel, N. Turpault, and S. Cornell. The impact of non-target events in synthetic soundscapes for sound event detection. In *6th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 115–119, 2021.

L. Ruge, B. Altakrouri, and A. Schrader. Soundofthecity-continuous noise monitoring for a healthy city. In *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 670–675, 2013.

J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello. Scaper: A library for soundscape synthesis and augmentation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 344–348, 2017.

J. Sang, S. Park, and J. Lee. Convolutional recurrent neural networks for urban sound classification using raw waveforms. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2444–2448, 2018.

M. Schmitt and B. Schuller. End-to-end audio classification with small datasets–making it work. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019.

A. Seilacher. *Trace Fossil Analysis*. Springer Science & Business Media, 2007.

R. Serizel, N. Turpault, A. Shah, and J. Salamon. Sound event detection in synthetic domestic environments. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 86–90, 2020.

B. Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

S. Seyoum, L. Alfonso, S. J. Van Andel, W. Koole, A. Groenewegen, and N. Van De Giesen. A Shazam-like household water leakage detection method. *Procedia Engineering*, 186:452–459, 2017.

R. Sharma, S. Tiwari, R. Singla, S. Goyal, V. V. Patage, and R. T. Shankarappa. Sound event separation and classification in domestic environment using mean teacher. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–6, 2020.

B. She. Framework of footstep detection in in-door environment. In *International Congress on Acoustics*, pages 715–718, 2004.

Z. Shuyang, T. Heittola, and T. Virtanen. Active learning for sound event classification by clustering unlabeled data. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 751–755, 2017.

Z. Shuyang, T. Heittola, and T. Virtanen. An active learning method using clustering and committee-based sample selection for sound event classification. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 116–120, 2018.

Z. Shuyang, T. Heittola, and T. Virtanen. Active learning for sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2895–2905, 2020.

S. Singh, H. L. Bear, and E. Benetos. Prototypical networks for domain adaptation in acoustic scene classification. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 346–350, 2021.

O. Slizovskaia, L. Kim, G. Haro, and E. Gomez. End-to-end sound source separation conditioned on instrument labels. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 306–310, 2019.

O. Slizovskaia, G. Haro, and E. Gómez. Conditioned source separation for musical instrument performances. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2083–2095, 2021.

J. O. Smith. *Spectral Audio Signal Processing*. W3K, 2011.

J. C. Socoró, F. Alías, and R. M. Alsina-Pagès. An anomalous noise events detector for dynamic road traffic noise mapping in real-life urban and suburban environments. *Sensors*, 17(10):2323, 2017.

D. Stowell, M. Wood, Y. Stylianou, and H. Glotin. Bird detection in audio: a survey and a challenge. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2016.

Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai. Sound event aware environmental sound segmentation with Mask U-Net. *Advanced Robotics*, 34(20):1280–1290, 2020.

Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai. Multichannel environmental sound segmentation. *Applied Intelligence*, 51(11):8245–8259, 2021.

L. S. M. Sugai and D. Llusia. Bioacoustic time capsules: Using acoustic monitoring to document biodiversity. *Ecological Indicators*, 99:149–152, 2019.

B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450, 2016.

S. Takeyama, T. Komatsu, K. Miyazaki, M. Togami, and S. Ono. Robust acoustic scene classification to multiple devices using maximum classifier discrepancy and knowledge distillation. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 36–40, 2021.

E.-L. Tan, F. A. Karnapi, L. J. Ng, K. Ooi, and W.-S. Gan. Extracting urban sound information for residential areas in smart cities using an end-to-end iot system. *IEEE Internet of Things Journal*, 8(18):14308–14321, 2021.

T. Tang, X. Zhou, Y. Long, Y. Li, and J. Liang. CNN-based discriminative training for domain compensation in acoustic event detection with frame-wise classifier. In *13th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 939–944, 2021.

A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, page 1195–1204, 2017.

R. Torres, D. Battaglino, and L. Lepauloux. Baby cry sound detection: A comparison of hand crafted features and deep learning approach. In *International Conference on Engineering Applications of Neural Networks*, pages 168–179, 2017.

N. Turpault and R. Serizel. Training sound event detection on a heterogeneous dataset. In *5th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 200–204, 2020.

N. Turpault, R. Serizel, J. Salamon, and A. P. Shah. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *4th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 253–257, 2019.

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. In *5th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 205–209, 2020.

N. Turpault, R. Serizel, S. Wisdom, H. Erdogan, J. R. Hershey, E. Fonseca, P. Seetharaman, and J. Salamon. Sound event detection and separation: A benchmark on DESED synthetic soundscapes. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 840–844, 2021.

E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *IEEE / CVF Computer Vision and Pattern Recognition Conference*, pages 7167–7176, 2017.

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

E. Tzinis, Z. Wang, X. Jiang, and P. Smaragdis. Compute and memory efficient universal sound source separation. *Journal of Signal Processing Systems*, 94(2):245–259, 2022.

M. Vacher, F. Portet, A. Fleury, and N. Noury. Development of audio sensing technology for ambient assisted living: Applications and challenges. *International Journal of E-Health and Medical Communications (IJEHMC)*, 2(1):35–54, 2011.

P. van Hengel and J. Anemüller. Audio event detection for in-home care. In *International Conference on Acoustics (NAG/DAGA)*, pages 618–620, 2009.

D. D. Varma, R. Padmanabhan, and A. D. Dileep. Learning to separate: Soundscape classification using foreground and background. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 21–25, 2021.

R. Ventura, V. Mallet, and V. Issarny. Assimilation of mobile phone measurements for noise mapping of a neighborhood. *The Journal of the Acoustical Society of America*, 144(3):1279–1292, 2018.

K. Veselỳ, S. Watanabe, K. Žmolíková, M. Karafiát, L. Burget, and J. H. Černockỳ. Sequence summarizing neural network for speaker adaptation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5315–5319, 2016.

E. Vidaña-Vila, J. Navarro, C. Borda-Fortuny, D. Stowell, and R. M. Alsina-Pagès. Low-cost distributed acoustic sensor network for real-time urban sound monitoring. *Electronics*, 9(12): 2119, 2020.

E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46:535–557, 2017.

E. Vincent, T. Virtanen, and S. Gannot. *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

T. Virtanen, M. D. Plumbley, and D. Ellis. *Computational Analysis of Sound Scenes and Events*. Springer, 2018.

D. Wang. On ideal binary mask as the computational goal of auditory scene analysis. In *Speech Separation by Humans and Machines*, pages 181–197. Springer, 2005.

H. Wang, D. Yang, C. Weng, J. Yu, and Y. Zou. Improving target sound extraction with timestamp information. In *Interspeech*, pages 1526–1530, 2022a.

Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. In *Interspeech*, pages 2728–2732, 2019a.

Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous. Trainable frontend for robust and far-field keyword spotting. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5670–5674, 2017.

Y. Wang, A. E. M. Mendez, M. Cartwright, and J. P. Bello. Active learning for efficient audio annotation and classification with a large amount of unlabeled data. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 880–884, 2019b.

Y. Wang, N. J. Bryan, M. Cartwright, J. P. Bello, and J. Salamon. Few-shot continual learning for audio classification. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 321–325, 2021a.

Y. Wang, N. J. Bryan, J. Salamon, M. Cartwright, and J. P. Bello. Who calls the shots? Rethinking few-shot learning for audio. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 36–40, 2021b.

Z. Wang, C. Subakan, E. Tzinis, P. Smaragdis, and L. Charlin. Continual learning of new sound classes using generative replay. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 308–312, 2019c.

Z. Wang, C. Subakan, X. Jiang, J. Wu, E. Tzinis, M. Ravanelli, and P. Smaragdis. Learning representations for new sound classes with continual self-supervised learning. *arXiv preprint arXiv:2205.07390*, 2022b.

F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 91–99, 2015.

S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey. Unsupervised sound separation using mixture invariant training. In *Advances in Neural Information Processing Systems*, volume 33, pages 3846–3857, 2020.

Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley. Convolutional gated recurrent neural network incorporating spatial features for audio tagging. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3461–3466, 2017.

H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2272–2281, 2017.

D. Yang, H. Wang, and Y. Zou. Unsupervised multi-target domain adaptation for acoustic scene classification. *arXiv preprint arXiv:2105.10340*, 2021.

L. Yang, J. Hao, Z. Hou, and W. Peng. Two-stage domain adaptation for sound event detection. In *5th Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 41–45, 2020.

Z. Ye, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan, and K. Ouchi. Sound event detection transformer: An event-based end-to-end model for sound event detection. *arXiv preprint arXiv:2110.02011*, 2021.

M. Zabihi, A. B. Rad, S. Kiranyaz, M. Gabbouj, and A. K. Katsaggelos. Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In *2016 Computing in Cardiology Conference (CinC)*, pages 613–616, 2016.

M. Zappatore, A. Longo, and M. A. Bochicchio. Crowd-sensing our smart cities: A platform for noise monitoring and acoustic urban planning. *Journal of Communications Software and Systems*, 13(2):53–67, 2017.

N. Zeghidour and D. Grangier. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2840–2849, 2021.

Y. Zhang and Z. Duan. IMISOUND: An unsupervised system for sound query by vocal imitation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2269–2273, 2016.

Y. Zhang and Z. Duan. IMINET: Convolutional semi-Siamese networks for sound search by vocal imitation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 304–308, 2017.

C. Zhao, G. Liu, W. Shen, and L. Gao. A multi-representation-based domain adaptation network for fault diagnosis. *Measurement*, 182:109650, 2021.

J. Zhao, Q. Kong, X. Song, Z. Feng, and X. Wu. Feature alignment for robust acoustic scene classification across devices. *IEEE Signal Processing Letters*, 29:578–582, 2022.

X. Zheng, Y. Song, L.-R. Dai, I. McLoughlin, and L. Liu. An effective mutual mean teaching based domain adaptation method for sound event detection. In *Interspeech*, pages 556–560, 2021.

Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang. Learning to adapt invariance in memory for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (8):2723–2738, 2020.

Y. Zigel, D. Litvak, and I. Gannot. A method for automatic fall detection of elderly people using floor vibrations and sound—proof of concept on human mimicking doll falls. *IEEE Transactions on Biomedical Engineering*, 56(12):2858–2867, 2009.

K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ. Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):800–814, 2019.

Résumé

De l'industrie aux applications d'intérêt général, l'analyse automatique des scènes et événements sonores permet d'interpréter le flux continu de sons quotidiens. Une des principales dégradations rencontrées lors du passage des conditions de laboratoire au monde réel est due au fait que les scènes sonores ne sont pas composées d'événements isolés mais de plusieurs événements simultanés. Des différences entre les conditions d'apprentissage et de test surviennent aussi souvent en raison de facteurs extrinsèques, tels que le choix du matériel d'enregistrement et des positions des microphones, et de facteurs intrinsèques aux événements sonores, tels que leur fréquence d'occurrence, leur durée et leur variabilité. Dans cette thèse, nous étudions des problèmes d'intérêt pratique pour les tâches d'analyse sonore afin d'atteindre la robustesse dans des scénarios réels.

Premièrement, nous explorons la séparation des sons ambiants dans un scénario pratique dans lequel plusieurs événements sonores de courte durée avec des caractéristiques spectrales à variation rapide (c'est-à-dire des sons d'avant-plan) se produisent simultanément à des sons stationnaires d'arrière-plan. Nous introduisons la tâche de séparation du son d'avant-plan et d'arrière-plan et examinons si un réseau de neurones profond avec des informations auxiliaires sur les statistiques du son d'arrière-plan peut différencier les caractéristiques spectro-temporelles à variation rapide et lente. De plus, nous explorons l'usage de la normalisation de l'énergie par canal (PCEN) comme prétraitement et la capacité du modèle de séparation à généraliser à des classes sonores non vues à l'apprentissage. Les résultats sur les mélanges de sons isolés à partir des jeux de données DESED et Audioset démontrent la capacité de généralisation du système de séparation proposé, qui est principalement due à PCEN.

Deuxièmement, nous étudions comment améliorer la robustesse des systèmes d'analyse sonore dans des conditions d'apprentissage et de test différentes. Nous explorons deux tâches distinctes : la classification de scène sonore (ASC) avec des matériels d'enregistrement différents et l'apprentissage de systèmes de détection d'événements sonores (SED) avec des données synthétiques et réelles. Dans le contexte de l'ASC, sans présumer de la disponibilité d'enregistrements capturés simultanément par les matériels d'enregistrement d'apprentissage et de test, nous évaluons l'impact des stratégies de normalisation et d'appariement des moments et leur intégration avec l'adaptation de domaine antagoniste non supervisée. Nos résultats montrent les avantages et les limites de ces stratégies d'adaptation appliquées à différentes étapes du pipeline de classification. La meilleure stratégie atteint les performances du domaine source dans le domaine cible. Dans le cadre de la SED, nous proposons un prétraitement basé sur PCEN avec des paramètres appris. Ensuite, nous étudions l'apprentissage conjoint du système de SED et de branches de classification auxiliaires qui catégorisent les sons en avant-plan ou arrière-plan selon leurs propriétés spectrales. Nous évaluons également l'impact de l'alignement des distributions des données synthétiques et réelles au niveau de la trame ou du segment par transport optimal. Enfin, nous intégrons une stratégie d'apprentissage actif dans la procédure d'adaptation. Les résultats sur le jeu de données DESED indiquent que ces méthodes sont bénéfiques pour la tâche de SED et que leur combinaison améliore encore les performances sur les scènes sonores réelles.

Mots-clés: séparation de sources audio, classification de scènes sonores, détection d'événements sonores, adaptation de domaine, apprentissage profond

Abstract

From industry to general interest applications, computational analysis of sound scenes and events allows us to interpret the continuous flow of everyday sounds. One of the main degradations encountered when moving from lab conditions to the real world is due to the fact that sound scenes are not composed of isolated events but of multiple simultaneous events. Differences between training and test conditions also often arise due to extrinsic factors such as the choice of recording hardware and microphone positions, as well as intrinsic factors of sound events, such as their frequency of occurrence, duration and variability. In this thesis, we investigate problems of practical interest for audio analysis tasks to achieve robustness in real scenarios.

Firstly, we explore the separation of ambient sounds in a practical scenario in which multiple short duration sound events with fast varying spectral characteristics (i.e., foreground sounds) occur simultaneously with background stationary sounds. We introduce the foreground-background ambient sound separation task and investigate whether a deep neural network with auxiliary information about the statistics of the background sound can differentiate between rapidly- and slowly-varying spectro-temporal characteristics. Moreover, we explore the use of per-channel energy normalization (PCEN) as a suitable pre-processing and the ability of the separation model to generalize to unseen sound classes. Results on mixtures of isolated sounds from the DESED and Audioset datasets demonstrate the generalization capability of the proposed separation system, which is mainly due to PCEN.

Secondly, we investigate how to improve the robustness of audio analysis systems under mismatched training and test conditions. We explore two distinct tasks: acoustic scene classification (ASC) with mismatched recording devices and training of sound event detection (SED) systems with synthetic and real data. In the context of ASC, without assuming the availability of recordings captured simultaneously by mismatched training and test recording devices, we assess the impact of moment normalization and matching strategies and their integration with unsupervised adversarial domain adaptation. Our results show the benefits and limitations of these adaptation strategies applied at different stages of the classification pipeline. The best strategy matches source domain performance in the target domain. In the context of SED, we propose a PCEN based acoustic front-end with learned parameters. Then, we study the joint training of SED with auxiliary classification branches that categorize sounds as foreground or background according to their spectral properties. We also assess the impact of aligning the distributions of synthetic and real data at the frame or segment level based on optimal transport. Finally, we integrate an active learning strategy in the adaptation procedure. Results on the DESED dataset indicate that these methods are beneficial for the SED task and that their combination further improves performance on real sound scenes.

Keywords: audio source separation, acoustic scene classification, sound event detection, domain adaptation, deep learning