Deep Learning Approaches for Auditory Perception in Robotics

This is the pre-preprint version.

Thèse n. 8000 (2021)
présentée le 23.03.2021
À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR LABORATOIRE DE L'IDIAP
PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE
École polytechnique fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Weipeng He

acceptée sur proposition du jury :

Prof. Pascal Frossard, président du jury
Dr. Jean-Marc Odobez, directeur de thèse
Dr. Petr Motlicek, directeur de thèse
Prof. Jean-Philippe Thiran, rapporteur
Prof. Joakim Gustafson, rapporteur
Dr. Antoine Deleforge, rapporteur

Lausanne, EPFL, 2021

Acknowledgements

I would like to thank all the people who supported me along this challenging and rewarding journey.

First and foremost, I am deeply grateful to my supervisors Jean-Marc Odobez and Petr Motlicek. I have been exceptionally fortunate to have two supervisors and receive a double amount of guidance. From them, I have learned more than I could write here, from how to approach research questions to how to present research work. This experience is invaluable for my career.

I would like to thank my thesis committee members, Prof. Pascal Frossard, Prof. Jean-Philippe Thiran, Prof. Joakim Gustafson, and Dr. Antoine Deleforge, for their time reviewing my thesis and the inspiring discussion we had during the oral exam.

I am truly grateful to all my past and present colleagues from the perception and speech groups at Idiap. I appreciate all the stimulating and fruitful discussions, which have inspired and motivated me to improve my research. I especially thank Olivier for helping me prepare the speaker location annotations, and Nam for providing me with the baseline method for speaker embedding. I also owe thanks to all the people and robots who participated in the data collection. My research would not be possible without these data.

I acknowledge the financial support as well as the scientific collaborations provided by the MuMMER project (funded by the European Commission).

Additionally, my sincere gratitude goes to all my friends at Idiap. I will always remember the wonderful moments we spend together for hiking, skiing & snowboarding, boardgames, and so on.

Last but not least, I would like to thank my parents and my brother, who have always been supportive. My brother deserves extra credit for proofreading the French abstract.

Martigny, March 26, 2021 Weipeng He

Abstract

Auditory perception is an essential part of a robotic system in Human-Robot Interaction (HRI), and creating an artificial auditory perception system that is on par with human has been a long-standing goal for researchers. In fact, this is a challenging research topic, because in typical HRI scenarios the audio signal is often corrupted by the robot ego noise, other background noise and overlapping voices. The traditional approaches based on signal processing seek analytical solutions according to the physical law of sound propagation as well as assumptions about the signal, noise and environments. However, such approaches either assume over-simplified conditions, or create sophisticated models that do not generalize well in real situations.

This thesis introduces an alternative methodology to auditory perception in robotics by using deep learning techniques. It includes a group of novel deep learning-based approaches addressing sound source localization, speech/non-speech classification, and speaker re-identification. The deep learning-based approaches rely on neural network models that learn directly from the data without making many assumptions. They are shown by experiments with real robots to outperform the traditional methods in complex environments, where there are multiple speakers, interfering noises and no a priori knowledge about the number of sources.

In addition, this thesis addresses the issue of high cost of data collection which arises with learning-based approaches. Domain adaptation and data augmentation methods are proposed to exploit simulated data and weakly-labeled real data, so that the effort for data collection is minimized. Overall, this thesis suggests a practical and robust solution for auditory perception in robotics in the wild.

Keywords: robotic auditory perception, deep learning, sound source localization, DOA estimation, domain adaptation, multi-task learning, human-robot interaction.

Résumé

La perception auditive est un élément essentiel d'un système robotique destiné à interagir avec des humains. La création d'un système de perception auditive artificielle aussi performant que celui de l'homme a été un objectif de longue date pour les chercheurs. En fait, il s'agit d'un sujet de recherche difficile, car dans les scénarios d'interaction typiques, le signal audio est souvent corrompu par le bruit produit par le robot lui-même, d'autres bruits de fond ou encore des voix qui se chevauchent. Les approches traditionnelles sont basées sur le traitement du signal et recherchent des solutions analytiques en fonction de la loi physique de propagation du son, et des hypothèses sur le signal, le bruit et l'environnement. Cependant, ces approches ou bien se reposent sur des hypothèses trop simplificatrices ou bien utilisent des modèles trop sophistiqués. Pour ces raisons, les approches traditionnelles sont difficiles à appliquer dans des situations réelles.

Cette thèse présente une méthodologie alternative s'appuyant sur des méthodes d'apprentissage à partir de données. Plus spécifiquement, elle introduit un ensemble de techniques basées sur de réseaux de neurones profonds pour accomplir des tâches telles que localisation des sources sonores, distinction entre parole ou non, et ré-identification des locuteurs. L'intérêt de ces approches est du faire peu d'hypothèses sur les modèles de propagation du son. Des expériences avec de vrais robots montrent qu'elles surpassent les méthodes traditionnelles dans des environnements complexes avec plusieurs locuteurs, avec des bruits interférents et sans connaissance a priori du nombre de sources.

En outre, cette thèse aborde la question du coût élevé de la collecte de données, question qui se pose inéluctablement avec les approches basées sur l'apprentissage. Des méthodes d'adaptation de domaine et d'augmentation des données sont proposées pour exploiter des données simulées et des données réelles faiblement étiquetées, de sorte que l'effort pour la collecte de données soit minimisé. Dans l'ensemble, cette thèse suggère une solution pratique et robuste pour la perception auditive en robotique dans des situations d'interaction naturelles.

Mots clés : perception auditive en robotique, apprentissage profond, localisation sonore, évaluation de lat direction d'arrivée du son, adaptation de domaine, apprentissage multi-tâches, interaction homme-robot.

Contents

List of Figures

List of Tables

List of Abbreviations

ACC	Accuracy.
ANN	Artificial Neural Network.
BN	Batch Normalization.
BSS	Blind Source Separation.
CCF	Cross-Correlation Function.
CNN	Convolutional Neural Network.
CPM	Convolutional Pose Machine.
CRNN	Convolutional Recurrent Neural Network.
DA	Domain Adaptation.
DNN	Deep Neural Network.
DOA	Direction-of-Arrival.
DUET	Degenerate Unmixing Estimation Technique.
EER	Equal Error Rate.
EM	Expectation-Maximization.
GCC	Generalized Cross-Correlation.
GCC-PHAT	Generalized Cross-Correlation with Phase Transform.
GCCFB	GCC-PHAT on filter bank.
GEV	Generalized Eigenvalue.
GMM	Gaussian Mixture Model.
GRL	Gradient Reversal Layer.
GSC	Generalized Sidelobe Canceller.
HRI	Human-Robot Interaction.
HRTF	Head Related Transfer Function.
ILD	Inter-channel Level Difference.
IPD	Inter-channel Phase Difference.
ITD	Interaural Time Difference.
JFA	Joint Factor Analysis.

LSTM	Long Short-Term Memory.
MAE	Mean Absolute Error.
MFCC	Mel-Frequency Cepstral Coefficient.
MLP	Multi-Layer Perceptron.
MMD	Maximum Mean Discrepancy.
MSE	Mean Squared Error.
MTL	Multi-Task Learning.
MUSIC	Multiple Signal Classification.
MVDR	Minimum Variance Distortionless Response.
NMF	Non-negative Matrix Factorization.
NN	Neural Network.
PLSOM	Parameter-Less Self-Organising Map.
ReLU	Rectified Linear Unit.
ResNet	Residual Network.
RNN	Recurrent Neural Network.
RT60	Reverberation Time.
SCM	Spatial Covariance Matrix.
SDA	Supervised Domain Adaptation.
SELD	Sound Event Localization and Detection.
SGD	Stochastic Gradient Descent.
SNR	Signal-to-Noise Ratio.
SNS	Speech/Non-Speech.
SRP-PHAT	Steered Response Power with PHAse Transform.
SSDA	Semi-Supervised Domain Adaptation.
SSL	Sound Source Localization.
SSS	Sound Source Separation.
STFT	Short-Time Fourier Transform.
SVM	Support Vector Machine.
TDOA	Time Difference of Arrival.
TF	Time-Frequency.
TSNN	Two-Stage Neural Network.
UBM	Universal Background Model.
UDA	Unsupervised Domain Adaptation.
VAD	Voice Activity Detection.
VAE	Variational Autoencoder.
WSDA	Weakly-Supervised Domain Adaptation.

1 Introduction

Robots have long been envisioned to be capable of interacting naturally and socially with humans (Breazeal, 2004; Dautenhahn, 2007). To realize this, robots are required to perceive the environments, make decision and act accordingly in a human-like way. Robotic perception is the first stage of this “perceive-decide-act” processing pipeline. Thus, precise and timely perception is the foundation for subsequent reasoning and decision making.

Like humans, robots perceive the environments using multiple modalities, including audio, visual and haptic. Among these modalities, audio signals carry unique and a substantial portion of the information conveyed during interactions. Audio signals are “omni-directional”, and complement the visual signal for object tracking when objects are outside the field of view or occluded. Speech, which includes both verbal and non-verbal information, is the major form of communication between humans and robots. Voices can be used to recognize people’s identities, which is important for maintaining long-term interactions. Moreover, various sound events characterize the environment where the interactions take place.

As an essential part of the robotic perception, *Robotic Auditory Perception* or *Robot Audition* (Okuno and Nakadai, 2015) interprets information about environments as well as the interacting persons from signals captured by audio sensors. This thesis presents several deep learning based approaches for robotic auditory perception. As an overview of this research, the rest of this chapter is organized as follows: First we introduce the specific components of robotic auditory perception and the challenges in this topic. Then we explain the drawback of traditional approaches, and that leads to our motivation of adopting deep learning. Based on the motivation and background of the MuMMER project, we set the specific objectives of thesis. Finally, our contributions are summarized.

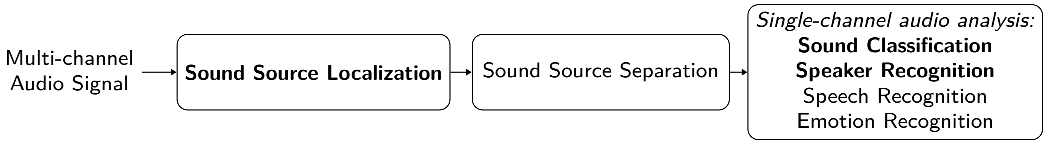


Figure 1.1 – The core functions of robotic auditory perception. Although most existing approaches apply these functions sequentially, there are exceptions (including our approaches) which solve two or more functions jointly. The focus of this thesis is highlighted with bold fonts.

1.1 Components of Robotic Auditory Perception

The human counterpart, human auditory perception, can pay attention to and extract information of a single voice in complex environments with multiple overlapping voices and noises, which is a situation known as the *Cocktail Party Effect* (Cherry, 1953; Haykin and Chen, 2005). The robotic auditory perception is expected to do the same, and more specifically, it should interpret information such as locations or *Direction-of-Arrival* (DOA) of the sound sources, type of the sound sources or events, speaker identity, speech content, and nonverbal cues in voices (such as emotion). Several specific technologies are involved for extracting these pieces of information, and they are applied sequentially in most auditory perception systems (Okuno and Nakadai, 2015; Argentieri et al., 2015) (Fig. 1.1):

- First, individual sound sources are localized with *Sound Source Localization* (SSL).
- Then, given the spatial information about the sound sources, *Sound Source Separation* (SSS) is applied to extract individual signals from different locations.
- Finally, the separated single-channel audio signals are processed for *Sound Classification*, *Speech Recognition*, *Speaker Recognition* and/or *Emotion Recognition*.

1.2 Challenges in Robotic Auditory Perception

Although substantial progress has been made recently in sound source localization, sound source separation, speech recognition, etc., their practical application in robotics is still limited. Especially, creating a human-level robotic auditory perception system is many years away. This is because several characteristics in the context of *Human-Robot Interaction* (HRI) make the auditory perception particularly challenging (Argentieri et al., 2015):

- Environments of HRI are dynamic and unpredictable. There are often multiple simultaneous speakers, of which the number is not known a priori. In addition, there are other unexpected background noises. All of these sounds reverberate in various room conditions.

- The choice of microphone array is limited in terms of size and budget of the robot. Because the microphone array is embedded on the robot, the possibility of using large microphone array and their corresponding technologies in array processing is excluded. Moreover, for robots designed with a tight budget, it is unlikely to use expensive microphones, such as high order ambisonic microphones. In fact, the number of microphones on the majority of the existing robots are between two and eight (Rascon and Meza, 2017).
- The physical embodiment of the robot generates noise and additional uncertainty. Some parts of the robot itself, such as fans, motors and speakers, may induce noise, which is termed *ego-noise*. Furthermore, the solid body also scatters the sound, making the approximation of the *transfer function* inaccurate.
- For practical applications, the perception response should be *real-time*. The computation complexity is limited so that prediction can be made within a guaranteed short delay.

1.3 Traditional Approaches

The traditional approaches for sound source localization and sound source separation are based on audio or array signal processing. These approaches seek analytical solutions according to the physical law of sound propagation without using labeled samples. The derivation of the solutions relies on assumptions about the acoustic environments, which may include known *Head Related Transfer Functions* (HRTFs), free-field anechoic sound propagation, high *Signal-to-Noise Ratio* (SNR), spatially white noise, and a known number of sources.

However, these assumptions may not hold well in real-world applications. For example, HRTF may not be precisely estimated due to the error in measurement of the microphone array geometry or obstacles (e.g. robot head) scattering the sound propagation in a way that is too complex to model. While direct measurement of the HRTF provides better estimation, it requires specialized equipment and significant amount of work. Moreover, there are often multiple simultaneous sound sources in the environments, and the number of sound sources is not known. The discrepancy between assumptions and reality may lead to significant performance degradation. Sophisticated modeling of a specific complex environment may mitigate the problem, but it is not clear how to generalize it as exhaustive modeling of all types of environments is unlikely.

1.4 Deep Learning for Auditory Perception

Alternative to traditional approaches, this thesis adopts a methodology of using *Deep Neural Networks* (DNNs) for auditory perception. Starting from automatic speech recognition (Bourlard and Morgan, 1994; Hinton et al., 2012), there have been many successful applications of DNNs in audio-related tasks (Purwins et al., 2019). With the radical increase of computational power and better optimization methods in recent years, it is now possible to train deeper and more complex neural network models which have been shown to outperform traditional signal processing approaches in challenging acoustic conditions. The deep learning approaches are advantageous in the following aspects:

- Instead of relying on physical laws of sound propagation, the deep learning approaches build models from *training samples*. Thus, they do not require many assumptions about the environment. In theory, DNNs are *universal function approximators* (Hornik, 1991; Zhou, 2020). That is, they can approximate any continuous function to an arbitrary accuracy, if the network size (in terms of depth or width) is large enough. In practice, as long as sufficient training data under the target conditions are available, it is possible to train neural network models for effective applications in dynamic environments.
- Neural networks are capable of maintaining implicit *prior models* of the target signals. In fact, humans use prior knowledge about sounds to help their auditory perception. Studies have shown that native listeners are better than non-native listeners at sentence processing in noisy environments (Florentine et al., 1984; Cooke et al., 2008). As for the neural networks, since they learn from examples, the features of the target audio signals in these examples are implicitly modeled by the neural networks. However, for signal processing approaches, direct modeling of audio signals, such as speech, is fundamentally difficult.
- Neural networks can be easily stacked and combined to create *Multi-Task Learning* (MTL) models or *end-to-end* models. MTL allows the task-specific knowledge to be shared among related tasks, so that they can help each other and create a synergy. The end-to-end design allows the indirect tasks (e.g. sound source localization with respect to speaker recognition) to be optimized directly for the target task at the end of the sequential processing pipeline, so that the error does not propagate through the pipeline. Examples are the increasing number of works in beamforming neural networks for far-field automatic speech recognition (Xiao et al., 2016; Braun et al., 2018; He et al., 2020)

Despite the success of deep learning approaches in many audio signal processing applications, some aspects of deep learning based auditory perception have not been studied in depth:

- *Simultaneous detection and localization of multiple sound sources in real HRI scenarios.* As previously mentioned, in real HRI scenarios there may be multiple simultaneous sound sources and no a priori knowledge about the number of them, so a practical SSL system is supposed to detect and localize all of them. However, most of the previous studies on deep learning based approaches (which we will summarize in more detail in Section 2.1.2) either only address localization of a single sound source, or overlook the fact that the number of sound sources is not known.

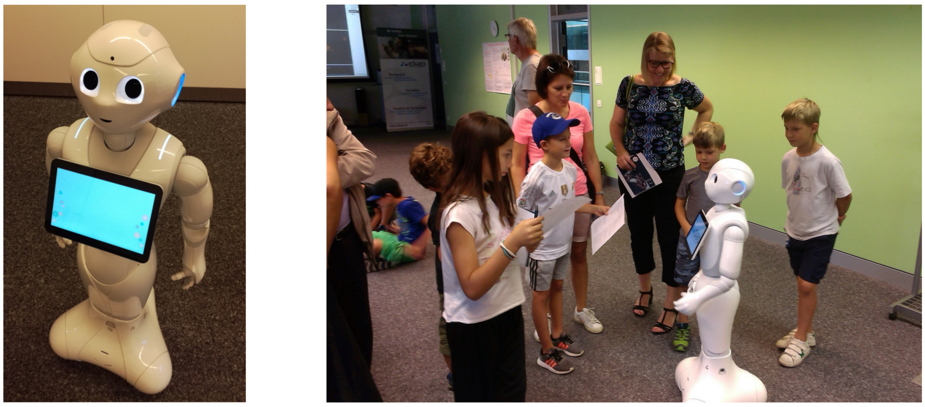
- *Learning with limited resource.* It is known that the deep learning based approaches rely heavily on a large number of labeled and unbiased training data. However, acquisition of labeled robotic audio data is especially costly. This is because audio data on distinct types of microphone arrays are radically different. Each type of microphone array requires individual data collection. In addition, annotation of sound source locations requires special procedure and external devices during the data collection, as the locations cannot be precisely labeled using the audio data alone. Instead of using real audio data, previous approaches have commonly used simulated data for training. However, simulation differs from the reality, therefore induces bias in the training data. Alternative learning techniques that train auditory perception models with limited data collection effort have not been well studied.
- *Multi-task learning for auditory perception in robotics.* As we have seen, an auditory perception system consists of many functions, but most of the previous studies consider them as separated modules in a sequential pipeline. There have been some works on joint sound source localization and separation (Mandel et al., 2010; Deleforge et al., 2013), joint localization and sound classification (May et al., 2011a; Taghizadeh et al., 2011; Crocco et al., 2017), as well as joint localization and speaker recognition (May et al., 2012). Nevertheless, the use of MTL neural networks for robotic auditory perception has not been well investigated.

1.5 Objectives

The vision of this thesis is to develop a practical deep learning based system for auditory perception in robotics with a focus on the following perception functions:

- Direction-of-arrival estimation;
- Speech/non-speech classification;
- Speaker re-identification.

These functions are the fundamental components of a full auditory perception system. The DOA estimation detects the sound sources and provides their spatial information for subsequent processing. Speech/non-speech classification enables the robot to distinguish speech from other sounds and pay attention to human actions. Understanding people's identity with speaker re-identification is the basis of long-term interactions. Moreover, the studies of these topics suggest a framework that can incorporate the other perception functions, such as speech and emotion recognition.



(a) Pepper.

(b) An example HRI scenario.

Figure 1.2 – The robot Pepper and a typical HRI scenario where a robot interacts with multiple persons.

Based on the previous discussion about the unsolved topics, the specific objectives of this thesis include:

- Investigate deep learning approaches for DOA estimation of multiple speakers, particularly under the condition that the number of speakers is not known a priori.
- Investigate domain adaptation methods for training DOA estimation DNNs with unlabeled or weakly-labeled real audio data.
- Investigate multi-task learning for auditory perception in robotics.

1.6 Background of Research

The research of this thesis was conducted within the *MuMMER project* (MultiModal Mall Entertainment Robot)[1]. The goal of this project is to build socially intelligent robots for entertainment in public spaces (Foster et al., 2016, 2019).

The robot platform *Pepper*[2] from Softbank Robotics has been used for this project (Fig. 1.2a). The robot is a 1.3 meter tall humanoid robot, equipped with four co-planar microphones as well as RGB and depth cameras. The cooling fans inside the robot head are very close to the microphones. The strong ego-noise produced by the fans, and the dynamic environments of HRI (Fig. 1.2b), make the auditory perception very challenging.

The approaches proposed in this thesis are all verified by experiments with the real robot. The result of the auditory perception is then combined with visual perception for long-term audio-visual person tracking and characterization (Foster et al., 2019).

[1] http://mummer-project.eu/

[2] https://www.softbankrobotics.com/emea/en/pepper

1.7 Contributions

This thesis reports the following contributions:

- a deep learning framework for joint sound source detection and DOA estimation in realistic HRI scenarios with the presence of short input, an unknown number of overlapping voices and strong ego-noise. For this, we proposed spatial spectrum output coding, that can handle an arbitrary number of sources. We study various input representations and network architectures. Our proposed approaches are shown to significantly outperforms traditional signal processing approaches. Part of this work is published in (He et al., 2018a).
- several domain adaptation approaches for training DOA estimation models with fully-labeled simulated data, and weakly-labeled or unlabeled real data. These approaches include a weakly-supervised domain adaptation method using the number of sound sources as weak label, and an unsupervised adaptation method with domain adversarial training (Ganin and Lempitsky, 2015). The weakly-supervised adaptation scheme is extended with data augmentation. With the extension, the weakly-supervised adaptation reaches a performance on par with the supervised approaches. This study suggests a practical deployment scheme for DNN based DOA estimation in real robotic application with minimal effort for data collection. Part of this work is published in (He et al., 2019).
- a novel multi-task neural network approach for joint DOA estimation and speech/non-speech classification. The proposed method achieves significantly better results in terms of speech/non-speech classification and speech source localization, compared to methods that separates localization and classification. Part of this work is published in (He et al., 2018b).
- another multi-task neural network for speaker embedding of overlapping voices using DOA estimation as an auxiliary task. This approach outperforms a speaker embedding DNN using beamformed signals as input in multi-source segments.
- more than 50 hours of original audio data with real robots, including sounds (speech and noise) played from loudspeakers as well as voices of human talkers. These data include as well the frame-level annotations of sound source locations, sound type, and speaker identities. We have released the data[3] for benchmarking of future studies on auditory perception in robotics.

1.8 Thesis Outline

The rest of this thesis is organized as follows:

- Chapter 2 summarizes the previous research on related topics of auditory perception.

[3] https://www.idiap.ch/dataset/sslr/

- Chapter 3 presents our research on deep neural networks for multi-speaker DOA estimation. A description of our data collection procedure is also included.
- Chapter 4 presents several domain adaptation approaches for DOA estimation models when training resource is limited.
- Chapter 5 studies a MTL neural network for joint DOA estimation and speech/non-speech classification.
- Chapter 6 introduces a MTL neural network for extracting embeddings of multiple speakers, which can be used for speaker re-identification.
- We conclude and suggest future research ideas in Chapter 7.

2 Literature Review

This chapter reviews the related research on auditory perception in robotics. As there is an extensive amount of literature on this topic, we limit the review to the studies of sound source localization, domain adaptation, and multi-task learning for auditory perception, as these are most relevant to the subsequent chapters of this thesis.

2.1 Sound Source Localization

Sound Source Localization (SSL) is the task of estimating *locations* of sound sources from audio signals captured by audio sensors (i.e. microphones). It can be separated into estimation of *Direction-of-Arrival* (DOA) and *distance* (Fig. 2.1). The DOA estimation is categorized into 1-dimensional, which is estimating the azimuth (horizontal direction) alone, and 2-dimensional, which is estimating both the azimuth and elevation of the sound sources.

In the context of robotics, most of the literature of SSL addresses DOA estimation because knowing the DOA is of greater interest and estimating distance is more challenging. DOA (or azimuth alone) allows robots to turn to speakers and respond to interactions. Associating sound sources to objects detected from vision system also requires DOA estimation. Many of the DOA estimation approaches rely on a grid search on the candidate DOAs, these approaches can be extended to location estimation under near-field conditions by replacing the DOA candidate set with locations. Therefore, it is common to see DOA estimation and SSL being used interchangeably in the literature.

The basis of SSL is that sound sources from various locations travel through different paths before arriving at the microphones, and such information is embedded in the captured audio signals. That is, in the signal processing language, the frequency-domain signals $\mathbf{X}(\omega) \in \mathbb{C}^M$ captured by M microphones are mixtures of N sound sources $S_i(\omega) \in \mathbb{C}$ filtered by different *transfer functions* $\mathbf{H}_i(\omega; \varphi_i, \theta_i, \rho_i) \in \mathbb{C}^M$, and an additional noise

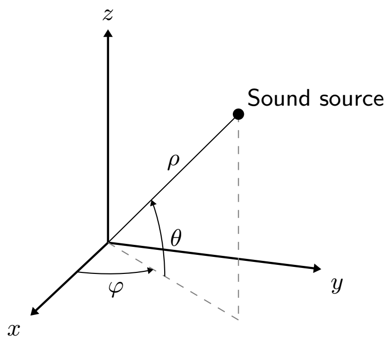


Figure 2.1 – Targets of sound source localization : azimuth φ, elevation θ and distance ρ. In typical reference coordinate systems, the origin is the center of the microphone array, and x-y plane is the horizontal plane.

$\mathbf{V}(\omega) \in \mathbb{C}^M$:

$$\mathbf{X}(\omega) = \sum_{i=1}^{N} \mathbf{H}_i(\omega; \varphi_i, \theta_i, \rho_i) S_i(\omega) + \mathbf{V}(\omega), \tag{2.1}$$

where ω is the frequency. Depending on the sound source location $(\varphi_i, \theta_i, \rho_i)$ (Fig. 2.1), the transfer functions filter the source signals for each channel by different delays, attenuation, diffraction and reverberation.

The information about the sound source locations, termed *sound localization cues*, are extracted explicitly or exploited implicitly for both robotic and human auditory perception. The most common cues are:

- *Time Difference of Arrival* (TDOA). Because the microphones are at different distances to the sound source, the sound wave arrives at the microphones with different delays (Fig. 2.2). The TDOA can be mapped back to space to get an estimation of the DOA. In the context of human ears or binaural microphones with artificial pinnae, the TDOA is also called *Interaural Time Difference* (ITD).
- *Inter-channel Phase Difference* (IPD). IPD for narrowband signals is the equivalent concept of TDOA in the frequency domain. While TDOA requires estimation, IPD can be directly computed. However, the mapping from IPD of high-frequency signals to TDOA or spatial locations is ambiguous due to *spatial aliasing*.
- *Inter-channel Level Difference* (ILD). The different distances of the sound propagation paths and the objects blocking sound propagation cause sound level differences between channels. For example, the human head or robot head can block the sound, so that when a sound is coming from the left or the right, the sound is more attenuated when it arrives at the ear on the opposite side. ILD is more prominent for high frequency signals as they are more affected by obstacles, while low frequency signals can travel around obstacles due to diffraction.
- *Spectral Cues*. Human or artificial pinnae (outer ears) amplify or attenuate different

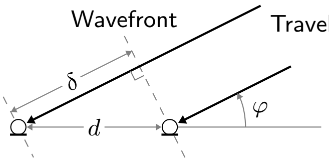


Figure 2.2 – TDOA caused by different sound travel distances. In this example, the sound wave is planar as a far-field condition is assumed. δ is the difference in travel distance, and the TDOA is δ/c, where c is the sound speed.

frequencies depending on the DOA of the signal. Such changes of response power in certain frequencies are the spectral cues. They can be used to distinguish sounds coming from the front or the back, which can be confusing for two microphones when relying on TDOA and ILD alone. Moreover, the spectral cues of a single microphone can be used for sound source localization if the source signal prior model and sound propagation are learned (Saxena and Ng, 2009; Georganti et al., 2011; El Badawy et al., 2017).

Once the localization cues are obtained, *cue-to-location mapping* is carried out to estimate the sound locations. The mapping depends on the type of localization cue and on a *propagation model*. This propagation model describes how sound is propagated to the microphones, or in other words, how the transfer functions $\mathbf{H}$ in Eq. (2.1) are like.

Depending on how the mapping procedure and propagation model are obtained, SSL methods can be categorized into *traditional signal processing approaches* and *learning-based approaches*. As we have mentioned in the introduction (Section 1.3), the traditional approaches rely on explicit modeling of sound propagation according to the acoustic environments and seek analytical solution based on the propagation model. In contrast, the learning-based approaches do not require explicit modeling of sound propagation. Instead, the propagation model as well as the mapping procedure are learned jointly from training samples, which are samples of the captured sound signals and corresponding sound locations.

In the following sections, we review the two groups of SSL approaches with an emphasis on those which are most related to our work. For more comprehensive reviews on sound source localization in robotics, readers are referred to (Argentieri et al., 2015) and (Rascon and Meza, 2017).

2.1.1 Traditional Signal Processing Approaches for SSL

In this section, we summarize the propagation models and mapping procedures of the traditional signal processing approaches.

Propagation Models

Propagation models are required for all traditional signal processing approaches. They are either estimated or measured. The estimation is normally based on simplified assumptions about the environments. For example, free-field propagation is commonly assumed, which means that there is no object in the space. Thus, the sound travels in a single direct path to the microphone, without reflection and scattering. In this case the transfer functions are simply time delays (phase shift in the frequency domain). For more accurate propagation modeling, the geometry of the robot head is simplified as a sphere, so that sound diffraction caused by the head can be taken into account (Nakadai et al., 2000, 2003; Kim et al., 2011, 2015). While most literature does not directly model the reverberation (which is considered as a part of the noise term $\mathbf{V}$ in Eq. (2.1)), there are a few approaches which include early reverberations in their propagation models, so that the DOA of the early reverberation can be exploited for SSL (An et al., 2018, 2020; Di Carlo et al., 2019).

Besides estimation, the propagation model can also be obtained through measurement. In binaural sound source localization research, measured *Head Related Transfer Functions* (HRTFs) are commonly used (Gardner and Martin, 1995; Algazi et al., 2001; Wierstorf et al., 2011). HRTF measurement requires an anechoic chamber and specific devices, and this is not practical for all research groups. Even if the HRTFs are obtained, an HRTF only constitutes the direct path part of the transfer function, and the reverberations need to be coped additionally.

Mapping Procedures

The mapping procedures from localization cue to location can be categorized into three types: *TDOA-based*, *grid search* and *clustering-based*.

TDOA-based. The TDOA-based methods first estimate the TDOA, and then infer the sound locations from the estimated TDOA. Among the various proposed methods, *Generalized Cross-Correlation with Phase Transform* (GCC-PHAT) (Knapp and Carter, 1976) is the most popular one due to its effectiveness and simplicity. According to GCC-PHAT, the TDOA is estimated by finding the delay that maximizes the generalized cross correlation between two channels:

$$\mathrm{TDOA} = \arg\max_{\tau} \int_{-\pi}^{\pi} \Psi(\omega) X_1(\omega) X_2(\omega)^* e^{j\omega\tau} d\omega, \tag{2.2}$$

where * denotes complex conjugation, and $\Psi(\omega)$ is the phase transform:

$$\Psi(\omega) = \frac{1}{|X_1(\omega) X_2(\omega)^*|}. \tag{2.3}$$

The intuitively-designed phase transform flattens the contribution of signals in all frequencies, and under ideal conditions (two channels are exactly the same except for a delay and scaling), the generalized cross correlation becomes a unit impulse shifted by the TDOA. It has been shown that GCC-PHAT is robust to reverberation (Brandstein and Silverman, 1997). However, the effectiveness of GCC-PHAT relies on the assumption that the target signal dominates all frequency bins. In the presence of interfering sound sources, frequency masking can be applied to select the frequency bins that correspond to the target sound source for computing the cross correlation (Valin et al., 2003; Kim et al., 2015; Grondin and Michaud, 2015). Although the GCC-PHAT is designed to estimate the TDOA of a single sound source, other peaks in the cross correlation can also be used to localize multiple sources with an optional modification of the filter function (Kwon et al., 2010).

Once the TDOAs are estimated, sound locations are solved according to the propagation model. For example, in the simple case of two microphones under the free-field and far-field assumptions (Fig. 2.2), the DOA can be solved from the TDOA and the geometry of the microphone array:

$$\varphi = \arccos \frac{\mathrm{TDOA} \cdot c}{d}, \tag{2.4}$$

where c is the speed of sound and d is the distance between the microphones. However, with only two microphones, sound sources from different directions in a cone ("cone of confusion") can produce exactly the same TDOA, therefore more microphone is required for unambiguous 2D DOA estimation. When using more than two microphones, the estimation of the DOA from pairwise TDOAs may be overdetermined, thus additional criteria such as least square (Smith and Abel, 1987; Brandstein et al., 1997; Apolinario et al., 2019) or maximum likelihood (Urruela and Riba, 2004; So et al., 2008) are needed.

Grid-search. The second type of mapping procedure relies on searching the solution on a grid of candidate DOAs or locations. These grid search based approaches compute a score for each DOA, and the DOAs with the maximum scores are the estimations. The scores constitute a function of DOA, which is known as a *spatial spectrum* or *angular spectrum* (Fig. 2.3). There are numerous methods to compute them. Each of these methods rely on different assumptions and objectives.

A popular group of these methods are based on *beamforming* or *steered response*. These approaches apply filters (linear transform) and summation over all channels (i.e. filter-and-sum), so that the signal assumed to come from a target direction is isolated. Then, criteria derived from the steered response are considered as the spatial spectrum. Examples of this type are *Steered Response Power with PHAse Transform* (SRP-PHAT) (DiBiase et al., 2001) and *Signal-to-Noise Ratio* (SNR) estimation with *Minimum Variance Distortionless Response* (MVDR) beamformer (Capon, 1969; Blandin et al., 2012).

Besides beamforming, there are many other ways to compute the spatial spectrum. For

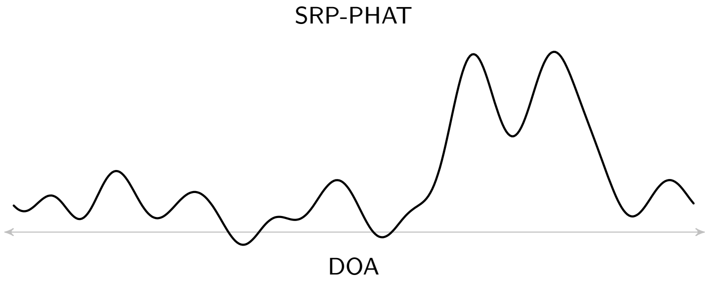


Figure 2.3 – Example spatial spectrum of SRP-PHAT. The abscissae of the two peaks in the spatial spectrum indicates the two predicted DOAs.

example, it can be obtained through the estimation of the subspace spanned by the signal from the target direction in the *Spatial Covariance Matrix* (SCM). This is the basis of the well-known *Multiple Signal Classification* (MUSIC) approach (Schmidt, 1986) and its variants (Dmochowski et al., 2007; Nakamura et al., 2009, 2011). Another example is the coherence between signals after inverting the HRTF (Keyrouz et al., 2006).

Clustering-based. The third type of mapping procedure is based on clustering (Mandel et al., 2006, 2010; Cobos et al., 2011; Blandin et al., 2012). These approaches assume there is one dominant sound source in each time-frequency bin. Therefore, each bin can be associated with a target sound source. This is accomplished by clustering the bins iteratively with the *Expectation-Maximization* (EM) algorithm (Dempster et al., 1977).

2.1.2 Learning-based Approaches for SSL

As we have seen from the previous section, the traditional signal processing approaches require explicit modeling of the sound propagation. However, sound propagation in real situation is very complex due to sound reflection, diffraction and scattering. Simplifying assumptions are made in order to obtain analytical solutions. Therefore, discrepancy between the assumptions and reality may severely impact their performance.

Recently, researchers have proposed learning-based approaches, which can learn the propagation model from sample data instead of manually specifying it. Specifically, these approaches create machine learning *models with adaptable parameters* that describe a relation between the features of the captured signals and sound locations. The parameters are searched through a *training process* to make the model fit the relation observed on the training samples. During *test* time, the model is expected to make predictions on unseen data based on the relation specified by the model parameters.

In the *supervised-learning* setting, the training samples are pairs of captured signals and corresponding location labels. Although most of the learning-based SSL approaches assume the supervised-learning setting with sufficient number of labeled samples, there are exceptions that apply learning without labels (*unsupervised*) which we will introduce in detail in Section 2.2.

Generative and Discriminative Approaches

There are two ways to model the relation between the features and labels: *generative* and *discriminative*. The generative approaches model the joint distributions of the features and the labels. For prediction, the posterior probability of the label conditioned on the features is derived from the joint distribution. Examples of these approaches are modeling binaural features and DOA with *Gaussian Mixture Model* (GMM) (May et al., 2011b, 2015; Ma et al., 2015b), and probabilistic piecewise-affine mapping (Deleforge and Horaud, 2012; Deleforge et al., 2015b).

The discriminative approaches approximate mappings from features to labels, without necessarily using a probabilistic framework. *Artificial Neural Networks* (ANNs) are well suited for such an approach, as they can in theory approximate any continuous function (Hornik, 1991; Zhou, 2020). In particular, with the advancement of computational power as well as theories in deep learning, there has been a surge of studies on using *Deep Neural Networks* (DNNs) for sound source localization in recent years. However, applying deep learning is not trivial, the understanding of sound signal and sound propagation is needed for designing effective input representation, output coding, neural network structure as well as training processes. These elements can have a high impact on the results and should be carefully chosen to avoid *underfitting* (the model does not describe well the training samples) or *overfitting* (the model does not generalize well to unseen data), both of which undermine the performance of the models. We summarize recent studies (Table 2.1) in term of these elements.

Input Representation

Both high-level hand-crafted features and low-level representations have been used as input of neural networks. The high-level hand-crafted features usually correspond to those which have been used in the traditional signal processing approaches, including ITD (Palmieri et al., 1991; Berglund and Sitte, 2005; Youssef et al., 2013), IPD (Datum et al., 1996; Berglund and Sitte, 2005; Pang et al., 2019), ILD (Palmieri et al., 1991; Datum et al., 1996; Berglund and Sitte, 2005; Youssef et al., 2013; Ma et al., 2015a, 2017; Pang et al., 2019), MUSIC eigenvectors (Takeda and Komatani, 2016b,a), *Cross-Correlation Function* (CCF) (Ma et al., 2015a, 2017), and GCC-PHAT coefficients (Xiao et al., 2015; He et al., 2018a; Ferguson et al., 2018; Vesperini et al., 2018; Vecchiotti et al., 2018).

Recently, instead of explicit feature extraction, some approaches use directly the low-level representations of the signal (i.e. raw signal) and expect the neural networks to learn by

Table 2.1 – Summary of neural network based SSL approaches. We include our approaches (will be presented in Chapter 3) for comparison. “1 (w noise)” in the second column means these approaches try to localize one target sound source from a mixture of directional interfering noises. The third column “a priori” indicates whether a priori knowledge about the number of sources is required for multi-source localization.

Approach	# of Sources	a priori	Input	Output	Architecture
Palmieri et al. (1991)	1	-	ITD, ILD	Gaussian coding	MLP
Neti et al. (1992)	1	-	Power spectrum	Gaussian coding	MLP
Datum et al. (1996)	1	-	IPD, ILD	Gaussian coding	MLP
Berglund and Sitte (2005)	1	-	ITD, IPD, ILD	Posterior probability	PLSOM
Murray and Erwin (2011)	1	-	Spectral analysis	Posterior probability	MLP
Youssef et al. (2013)	1	-	ITD, ILD	DOA	MLP
Xiao et al. (2015)	1	-	GCC-PHAT coefficients	Posterior probability	MLP
Ma et al. (2015a, 2017)	Multiple	required	CCF, ILD	Posterior probability	MLP
Ma and Brown (2016)	1 (w noise)	-	CCF, ILD	Posterior probability	MLP
Takeda and Komatani (2016b)	0 or 1	no	MUSIC eigenvectors	Posterior probability	modified MLP
Takeda and Komatani (2016a)	0, 1, 2	no	MUSIC eigenvectors	Marginal posterior probability	modified MLP
Yalta et al. (2017)	0 or 1	no	Power spectrogram	Posterior probability	CNN
Chakrabarty and Habets (2017)	1	-	Phase spectrogram	Posterior probability	CNN (ResNet)
Pertilä and Cakir (2017)	1	-	Magnitude spectrogram	TF-mask (for SRP-PHAT)	CNN
Adavanne et al. (2018)	Multiple	no	Magnitude and phase spect.	(DOA-wise) Posterior prob.	CRNN
Chakrabarty and Habets (2019)	Multiple	required	Phase spectrogram	(DOA-wise) Posterior prob.	CNN
Ferguson et al. (2018)	1	-	GCC, cepstrogram	DOA and distance	CNN
Perotin et al. (2018)	1	-	Ambisonic intensity vector	Soft-assigned (1 and 0.5)	CRNN
Vesperini et al. (2018)	1	-	GCC-PHAT	2D coordinate	CNN
Vecchiotti et al. (2018)	1	-	GCC-PHAT, logmel	2D coordinate	CNN
Pak and Shin (2019)	Multiple	required	sin and cos of IPD	Direct-path IPD	MLP
Pang et al. (2019)	1	-	IPD, ILD	Posterior probability	CNN
Pertila and Parviainen (2019)	1	-	GCC-PHAT, logmel	TDOA	LSTM
Tang et al. (2019)	1	-	Ambisonic intensity vector	DOA or posterior prob.	CRNN
Vecchiotti et al. (2019)	1	-	Raw waveform	Posterior probability	CNN
Mack et al. (2020)	1 (w noise)	-	Phase spectrogram	Posterior probability	CNN with LSTM attention
Our approaches					
He et al. (2018a)	Multiple	no	GCC-PHAT	Gaussian spatial spectrum	MLP
He et al. (2018a)	Multiple	no	GCC-PHAT on filter bank	Gaussian spatial spectrum	CNN, TSNN
He et al. (2018b)	Multiple	no	STFT (real & imaginary)	Gaussian spatial spectrum	CNN (ResNet)

themselves to extract the optimal localization features for SSL. These representations include raw waveform (Vecchiotti et al., 2019) in the time domain, power spectrum (Neti et al., 1992) in the frequency domain, ambisonic intensity vectors (Perotin et al., 2018; Tang et al., 2019) in the spherical harmonic domain, and several kinds of representations in the *Time-Frequency* (TF) domain. For instance, there are power/magnitude spectrogram (Yalta et al., 2017; Pertilä and Cakir, 2017; Adavanne et al., 2018) phase spectrogram (Chakrabarty and Habets, 2017, 2019; Adavanne et al., 2018; Mack et al., 2020), and raw *Short-Time Fourier Transform* (STFT) with its real and imaginary parts (He et al., 2018b).

Output Coding

Output coding defines how labels are encoded into desired network outputs, and how the network outputs are decoded into predictions. The type of output coding dictates the choice of the loss function used during training.

Single sound source localization is commonly considered as either a *regression* problem or a *classification* problem. In the regression setting, the network directly outputs the DOA or locations as a continuous value (Youssef et al., 2013; Vesperini et al., 2018; Vecchiotti et al., 2018; Tang et al., 2019). These approaches usually use *Mean Squared Error* (MSE) between the prediction and ground truth as the loss function.

In the classification setting (Xiao et al., 2015; Takeda and Komatani, 2016b; Vecchiotti et al., 2019; Tang et al., 2019; Mack et al., 2020), the network outputs a vector that is interpreted as the posterior distribution probability. Each element in the vector is associated with a location or DOA, and its value is the posterior probability of *the sound*[1] coming from that location (conditioned on the given input signal). The desired output is a one-hot vector, with value 1 at the ground truth location and 0 at other locations. During test time, the prediction is the direction with the highest output value. A label of "silence" can be added to the set of classes, so that the model can detect and localize a sound source without a priori knowledge about whether it is active (Takeda and Komatani, 2016b). A *softmax* function is often used at the output layer, so that the output values sum to one, and a *cross-entropy* loss is used as the training target.

In contrast to the posterior probability coding, the *soft-assigned spatial spectrum coding* relaxes the constraint of "one-hot" (hard assignment), so that the output values at the directions that are next to the ground truth are assigned between 0 and 1. For example, Gaussian coding uses a Gaussian function centered at the ground truth direction as the desired output (Palmieri et al., 1991; Neti et al., 1992; Datum et al., 1996). As another example, in (Perotin et al., 2018), the desired output values are 1 at the ground truth direction, and 0.5 at its neighboring directions. In these approaches, the network

[1] In single sound source localization, it is assumed that there is one and only one sound source.

output is no longer a distribution (because it does not sum to one), and it resembles the spatial spectrum used in traditional signal processing approaches[2]. Soft assignment takes account of the variance in DOA estimation and correlation among neighboring directions. It is shown to generate better models than the hard assignment (Perotin et al., 2018). In these approaches, a *sigmoid* function is often used at the output layer, so that the output values are bounded between 0 and 1, and an MSE loss defined on the spatial spectrum is often used together with it.

Posterior probability coding can be extended to multiple sound source localization, by aggregating single sound source predictions over time (Ma et al., 2015a, 2017). Such an approach first trains neural networks for localizing a single sound source, Then with the assumptions of each time frame is dominated by one sound source, the neural network predicts the posterior probability of the dominant sound source locations in each frame, and the outputs are averaged over time to get the final spatial spectrum. However, this method usually applies averaging at the utterance level, and is not applicable for detecting multiple sound sources in a short frame.

Other types of output coding have been proposed for multiple sound source location at the frame level. *Marginal posterior probability coding* is used for localization of maximum two sound sources (Takeda and Komatani, 2016a). In this approach, the network outputs two vectors, each of which encodes the posterior probability of the location of one sound source. The sound source to vector assignment, which can be ambiguous, is resolved by location-based ordering. Nevertheless, there might be some confusion for the network, as sound sources from the same direction may be assigned to the first source or the second source depending on the spatial location of other sound sources.

DOA-wise posterior probability coding is another approach (Adavanne et al., 2018; Chakrabarty and Habets, 2019), where network is used for multi-class multi-label classification. Binary classification is applied to each direction, in which an output value indicates the posterior probability of the presence of a sound in that direction. The desired output is a hard-assigned spatial spectrum, with multiple 1s at the ground truth directions and 0 for the rest. At test time, if there is no a priori knowledge about the number of sources, a threshold of 0.5 is applied to the output values to find the prediction. If the number of sources k is known[3], the directions with the k highest values are the prediction. A sigmoid function at the output layer and a *binary cross entropy* loss are often used in this type of output coding.

Different from other approaches, we have proposed using a Gaussian based spatial spectrum coding for multiple sound source localization, which we will introduce in Chapter 3.

[2] In fact, the posterior probability coding is the hard-assignment case of the spatial spectrum coding.

[3] Strictly speaking, if the number of sources k is known, the network output can no longer be interpreted as posterior probability.

Network Architecture

The choices of network architectures depends on the input representation, output coding and complexity of the problem. Early approaches, which normally relied on high-level feature extraction, were primarily based on *Multi-Layer Perceptron* (MLP) (Palmieri et al., 1991; Ma et al., 2015a; Xiao et al., 2015). As the high-level features already contain direct information regarding the sound source locations, exploiting simple neural networks may be sufficient.

However, for approaches with low-level input representation, deeper models are required, so that high-level localization cues can be extracted implicitly by the network. In addition, the low-level representations contain topological structures, that is information is distributed along time and frequency axes. Thus, *Convolutional Neural Networks* (CNNs) are commonly used for these inputs (Chakrabarty and Habets, 2017; He et al., 2018a; Vecchiotti et al., 2018). For example, 1D convolutions are typically applied to time-domain signals, and 2D convolutions are applied to time-frequency domain signals. The weight sharing of the convolution kernels significantly reduces the number of model parameters, making deep architectures possible. The usage of *Residual Networks* (ResNets) have also been studied (Yalta et al., 2017; He et al., 2018b). With residual connection added to the CNN, ResNet allows even deeper models to be trained without being affected by the vanishing gradient problem.

While CNNs exploit the temporal structure of the sound signal, their receptive fields (the part of the signal which the prediction is based on) are limited. Using *Recurrent Neural Networks* (RNNs) allows network to incorporate information from an unlimited context. For example, some methods use *Convolutional Recurrent Neural Networks* (CRNNs), which are the combinations of CNN with recurrent layers (Adavanne et al., 2018; Perotin et al., 2018).

Hybrid Approaches

Besides the learning-based approaches we have mentioned, there are *hybrid approaches* that also use machine learning models, but do not directly model the relation between signal and locations. Instead, these approaches use machine learning models to predict some statistics that can be combined with traditional signal processing approaches for location estimation. For example, DNN is applied for TDOA estimation (Pertila and Parviainen, 2019; Di Carlo et al., 2019), and then TDOA to DOA mapping can be carried out with a manually-specified propagation model. In other examples, DNNs are used for estimating the TF mask of the target signal, which then can be used for weighting the TF-bins for SRP-PHAT (Pertilä and Cakir, 2017) , computing the SCM for MUSIC (Xu et al., 2017) or refining the IPD for clustering-based DOA estimation (Pak and Shin, 2019).

Data Acquisition

Unlike traditional signal processing approaches, the performance of the learning-based approaches heavily relies on a large number of unbiased training samples. In other words, the training samples should cover various conditions in terms of sound locations, number of sources, room and noise conditions. These training conditions should match those in the target application scenarios. Otherwise, the training samples may not generalize well to unseen audio data.

There are two main ways to acquire training data. The first way is real data collection by audio recording. Sounds from loudspeakers or human talkers are recorded by microphone arrays and their location labels are extracted with external devices, such as cameras (Deleforge et al., 2015b; He et al., 2018a) and motion capture systems (Löllmann et al., 2018), or using robot motor sensor data (Deleforge et al., 2015a). Obtaining frame-wise labels (active or silence for each source in each frame) requires further voice activity annotation, either manually or automatic if clean source signal is available. The whole data collection process for real data is costly and the collected data is exclusive for only the type of microphone array that is used during the recording. Applying learning-based approaches to new types of microphone arrays requires individual data collection. In addition to manually controlled data collection, there has also been research on recording data automatically with mobile robots. With the images captured from the cameras, the robots can autonomously annotate or verify the sound source locations, either according to fixed data collection procedures (Le Roux et al., 2015) or via self-supervised learning (Liu et al., 2019).

Alternative to audio recording, training data can be obtained through acoustic simulation. While simulation generates a large number of data with various conditions at low cost, the simulated data is biased. The most commonly used room acoustic simulation methods (Allen and Berkley, 1979; Kulowski, 1985), only handle over-simplified room settings (Campbell et al., 2005; Habets, 2006). Recently, advanced simulation techniques considering sound reflection and scattering caused by solid objects (e.g. robot head or other objects in rooms) have been proposed (Nakadai et al., 2003; Schimmel et al., 2009; Jarrett et al., 2012; Schissler and Manocha, 2016; Tang et al., 2019). Nevertheless, in practice it is still difficult to measure and simulate the surfaces of complex solid objects, thus simulation cannot perfectly reproduce the directivity and frequency response patterns of the real microphone arrays. For binaural sound localization, simulation with measured HRTFs can generate more realistic data (Ma et al., 2015a).

2.2 Domain Adaptation

Ideal machine learning settings assume a sufficient number of training data. However, in reality, due to the high cost of data collection and annotation, labeled data of a *target* *domain* are not always available or sufficient. As a compromise for this case, data from another different but related domain, the *source domain*, can be used for training. While models can be well trained with the abundant data from the source domain, these models are not optimal for the target domain, because of the discrepancy between two domains. For example, SSL models trained with simulated data (source domain) may not perform well in real applications (target domain). *Domain Adaptation* (DA) aims to solve such an issue by exploiting data from both domains (Ben-David et al., 2010; Wang and Deng, 2018).

Domain adaptation is a branch of *transfer learning*, which studies how knowledge of a task can be exploited for another related task (Pan and Yang, 2010; Zhuang et al., 2020). In domain adaptation, source and target domains are *homogeneous*, that is they share an identical feature space, whereas transfer learning also studies situations where source and target domains are *heterogeneous*, that is they have distinct feature spaces. Domain adaptation is also related to *domain generalization* (Blanchard et al., 2011; Motiian et al., 2017), which is applied to situations where target domain data are not available. In contrast, domain adaptation assumes there are target domain data, which might be scarce in quantity or of which the exact labels are unavailable.

Depending on such label availability in the target domain, domain adaptation can be classified into four categories:

- *Supervised Domain Adaptation* (SDA): labeled but scarce target domain data are available;
- *Unsupervised Domain Adaptation* (UDA): labels are not available;
- *Semi-Supervised Domain Adaptation* (SSDA): some but not all of the target domain data are labeled;
- *Weakly-Supervised Domain Adaptation* (WSDA): the target domain data are annotated with related but not exact labels (*weak labels*).

Many domain adaptation approaches have been proposed. We will summarize their basic ideas in the following section, and present how they have been used in sound source localization.

2.2.1 Domain Adaptation Approaches

Common domain adaptation strategies include *instance weighting*, *feature transformation*, *parameter control* and *regularization*. Instance weighting tries to weight samples so that the loss function on the source samples is corrected to approximate that on the target domain (Huang et al., 2007; Sugiyama et al., 2008).

The feature transformation strategy seeks to construct a common representation space, so that a model can be used for inference on both domains. This can be achieved through aligning correlation on the data from both domains (Sun et al., 2016), or minimizing some metrics between the source domain and target domain data distributions, like in the *Maximum Mean Discrepancy* (MMD) approach (Borgwardt et al., 2006). Moreover, *domain-adversarial training* creates a common representation space such that features in this space cannot be distinguished for its domain by an additional network (domain classifier) (Ganin and Lempitsky, 2015).

Parameter control is another strategy which fixes some parameters of a model, such as the first n layers of a network, and fine-tuning the rest (Yosinski et al., 2014). In the supervised adaptation setting, the loss function is usually domain-agnostic and can be directly applied to the samples in the target domain to conduct the fine-tuning. The knowledge from the source domain is transferred via the frozen parameters.

Regularization strategy is used when labels of the target domain data are not available. It is based on heuristics about the unlabeled data. An example is the *entropy minimization principle* (Grandvalet and Bengio, 2004; Yves and Yoshua, 2006; Long et al., 2016). The entropy of the predicted labels is a measure of the class overlap. With the assumption that classes are well separated, models that generate minimum prediction entropy are favored. *Pseudo-labeling* is another type of regularization (Lee, 2013; Choi et al., 2019), which works by iteratively updating model parameters with its own prediction as ground truth. This is based on the *continuity assumption*, that is the samples with their features close to each other are likely to share the same label.

2.2.2 Domain Adaptation for Sound Source Localization

As we have previously mentioned, the cost of data collection for learning-based sound source localization methods is particularly high, while generating a large number of data with simulation is easy. Under such a context, domain adaptation is a critical technique for developing practical sound source localization systems. However, there have been so far only a few studies on this topic.

Takeda and Komatani (2017) have investigated unsupervised domain adaptation with entropy minimization for single sound source localization. In this study, the source domain data are generated from measured anechoic transfer functions, and target domain data are generated from measured reverberant transfer functions. A neural network is trained for classification of sound locations, thus entropy minimization is applicable. This study shows that unsupervised domain adaption is able to improve the localization performance. However, the improvement is sensitive to the stopping criterion. If adaptation is not stopped correctly, overfitting may occur and severely undermine the performance. This study is extended in (Takeda et al., 2018) by adding a eliminative constraint on top of entropy minimization. The eliminative constraint removes the least probable locations that are determined from the MUSIC spatial spectra, and reduces the risk of overfitting.

As we can see, domain adaption for multiple sound source localization neural networks, in particular with simulated source domain and real target domain, has not been well studied.

2.3 Multi-Task Learning for Auditory Perception

As mentioned in the introduction, robotic auditory perception consists of several components (Section 1.1). They are integrated into a whole system by two different strategies: *sequential* integration and *joint* integration. In the following sections, we will summarize specific research on robotic auditory perception under these two settings.

2.3.1 Sequential Approaches

Sequential integration combines the components in a sequential pipeline (Fig. 1.1). The processing of a component is independent of its subsequent ones. The components can be studied individually and combined freely. For example, any single-channel analysis component, such as automatic speech recognition, speaker recognition and sound classification, can be used directly with the result of any sound source separation approach. However, error and noise produced by each component can accumulate through the pipeline. Thus, to obtain an optimal solution, subsequent components should ideally be designed or trained to handle such noise. Moreover, the improvement in an intermediate component may not necessarily yield better performance in its following components. For example, a better speech enhancement system, that increases SNR of the audio signal, does not necessarily improve the word error rate of a subsequent automatic speech recognition module, because it can be trained to be robust to noise with noisy training data.

In the following sections, we will first introduce approaches for sound source separation as it is the key component that connects to all other components in a sequential auditory perception systems. Then, examples of sequential systems are presented.

Sound Source Separation

Sound source separation tries to reconstruct source signals from sound mixtures. *beamforming* is a class of techniques for separating spatially distributed sound sources when a priori knowledge about the sound source locations is available. Beamformers applies filter-and-sum to the multi-channel input signal to get the individual signals. The filter weights (i.e. beamforming weights) are chosen based on the target source direction and the input signal, so that the signals from other directions are suppressed. The most common beamformers are *delay-and-sum* beamformer, MVDR beamformer (Capon, 1969), and *Generalized Sidelobe Canceller* (GSC) beamformer (Griffiths and Jim, 1982). A special example is the *Generalized Eigenvalue* (GEV) beamformer (Warsitz and Haeb-Umbach, 2007), which does not require a priori knowledge about the sound source locations.

Time-frequency masking is another class of techniques. As speech signal is sparse in the time-frequency domain, it is possible to assign time-frequency bins to different sound sources. Based on the assignment, a binary mask (0 and 1 values for each TF bin) or soft mask (values between 0 and 1 for each TF bin) is element-wisely multiplied to the input signal, and the result is the separated signal. This is equivalent to applying a Wiener filter (Wiener, 1949) to the input signal. Example methods for estimating TF mask include the *Degenerate Unmixing Estimation Technique* (DUET) algorithm (Jourjine et al., 2000; Yilmaz and Rickard, 2004), *Non-negative Matrix Factorization* (NMF) (Smaragdis and Brown, 2003; Virtanen, 2007; Ozerov and Fevotte, 2010), and various recent approaches with DNN (Wang and Chen, 2018).

Besides beamforming and time-frequency masking, a group of approaches are based on *Independent Component Analysis* (Bell and Sejnowski, 1995; Smaragdis, 1998). By assuming signals are mutually independent, finding the independent components separates individual signals.

Examples of Sequential Systems

As previously mentioned, any single-channel audio analysis component can be used directly with separated signals. For example, sound classification is applied to separated signals to localize and track a specific type of sound sources (Lim et al., 2015; Crocco et al., 2017; Wakabayashi et al., 2020). Nevertheless, approaches that work well for clean signals may not perform the same for separated signals, which are unreliable due to imperfect separation. Many studies have attempted to solve such an issue.

Missing data classification is such an example. An estimated TF mask indicates which TF bins belong to the target signal and which bins belong to the noise. The noisy TF bins are then considered as missing data (with respect to the target signal), which may receive a distinctive processing. Three strategies have been proposed for missing data classification: *reconstruction*, *marginalization* and *direct masking* (Cooke et al., 2001; Hartmann et al., 2013). Reconstruction tries to reconstruct the complete observation according to a prior distribution of the data. Marginalization uses marginalized posterior on the reliable data in place of the complete posterior. Direct masking applies a constant attenuation to the noisy TF bins, and use them as if they were from the target signal. Missing data classification methods allow applications in noisy environments, such as localization and classification of multiple sound sources (May et al., 2011a), robust automatic speech recognition (Cooke et al., 2001; Hartmann et al., 2013), and robust speaker recognition (May et al., 2012; Zhao et al., 2012, 2014).

Besides missing data classification, there are other approaches dealing with unreliable input feature. For example, feature normalization with linear transformation or histogram equalization reduces the discrepancy between feature distributions of unreliable signal and clean signal (Squartini et al., 2012). Moreover, data augmentation can be used to train models robust to noise (Ming et al., 2007).

2.3.2 Joint Approaches

Joint integration, in contrast to sequential integration, solves multiple tasks at the same time. In this framework, the knowledge of individual tasks is shared among them to improve each other. From the view of *Multi-Task Learning* (MTL), jointly solving multiple tasks adds regularization on the model, and improves its generalization ability (Thrun, 1995; Caruana, 1997; Ruder, 2017).

Many of the clustering-based sound source localization approaches (Section 2.1.1) incorporate simultaneous sound source separation (Mandel et al., 2010; Deleforge et al., 2013). They usually jointly estimate the location of each TF bin and associate them with different sources to obtain TF masks. Jointly solving these two problems leads to more robust performance compared to sequential approaches.

A similar idea has been explored for joint sound source separation and speaker recognition (Zegers and Van hamme, 2016). In this work, speaker-specific basis vectors are estimated during sound source separation with NMF. The speaker-specific basis vectors provide prior knowledge about the spectral envelopes of the different voices, which is useful for better sound separation.

Recently, multi-task learning with deep neural networks has been studied for joint approaches. As an early attempt, (Hirvonen, 2015) uses CNN to estimate the joint posterior of location and type (speech/music) of a single sound source. Later, many MTL neural networks have been proposed for marginal estimations, that is the networks give multiple outputs, each of which is a prediction of a single task. These approaches have been applied to joint sound source localization and speech/non-speech classification (He et al., 2018b; Vecchiotti et al., 2018), joint *Sound Event Localization and Detection* (SELD) (Adavanne et al., 2019; Grondin et al., 2019; Kapka and Lewandowski, 2019; Xue et al., 2019), as well as joint sound source separation and speaker recognition (Drude et al., 2018; Shi et al., 2020).

3 Neural Network Models for Multi-Speaker DOA Estimation

This chapter, based on (He et al., 2018a) and (He et al., 2018b), discusses several *Deep Neural Network* (DNN) models for estimating *Direction-of-Arrival* (DOA) of multiple simultaneous sound sources. In particular, we consider the DOA estimation under the following conditions:

- Multiple simultaneous speakers;
- No a priori knowledge about the exact number of speakers;
- Short segments of speech;
- Presence of strong robot ego-noise.

While these are common conditions in *Human-Robot Interaction* (HRI), previous DNN-based methods (discussed in Section 2.1.2) have not addressed DOA estimation under all of these conditions.

In Section 3.1, we formalize the DOA estimation problem in a supervised learning setting, and present an overview of our deep learning-based approaches.

The absence of knowledge about the number of speakers invalidates the previous learning-based approaches based on regression or classification (Section 2.1.2). To solve this issue, we propose a spatial spectrum output coding scheme (Section 3.2). According to this output coding, neural networks are trained to predict spatial spectra, like those encountered in SRP-PHAT (DiBiase et al., 2001) and MUSIC (Schmidt, 1986), which can then be decoded into predictions of an arbitrary number of sound sources.

Then, three types of input representations are introduced in Section 3.3. These input representations include high-level hand-crafted features, such as *Generalized Cross-Correlation with Phase Transform* (GCC-PHAT) coefficients and *GCC-PHAT on filter bank* (GCCFB), as well as a low-level representation directly derived from the *Short-Time Fourier Transform* (STFT).

According to the characteristics of the different input representations, we design various network structures (Section 3.4), including a *Multi-Layer Perceptron* (MLP), a *Convolutional Neural Network* (CNN), a *Two-Stage Neural Network* (TSNN) that exploits the sub-band structures in the frequency domain, and a deep *Residual Network* (ResNet), which is able to implicitly extract high-level features from the STFT input.

We collected more than 24 hours of training and evaluation data with a real robot for experiments. These data includes both recordings of sound played from loudspeakers and recordings of human talkers. The data collection procedure is described in Section 3.5.

Section 3.6 discusses the localization and detection performance of the proposed deep learning-based approaches, which are compared to the traditional spatial spectrum based approaches. Unlike previous works, we emphasize the condition that there is no a priori knowledge about the number of sources, and include precision-recall curves in the performance criteria. The results show that our proposed approaches significantly outperform the traditional ones.

3.1 Overview

We seek algorithms that can detect sound sources and estimate their DOA from multichannel audio segments captured by microphone arrays on robots without any a priori knowledge about the number of speakers. Each audio segment s is a mixture of background noise and voices of some (may be zero) speakers from various directions. The set of speaker directions is denoted by the *label* $y \in Y$, and Y is the *label space* that include all finite subsets of possible directions:

$$Y = \{y \subset \Phi : |y| < \infty\},$$

where $|y|$ is the cardinality of y (i.e. number of sources), and Φ is the set of all possible directions. Depending on the application scenario, Φ can be all horizontal directions ($\Phi = [-\pi, \pi]$), all directions in the three-dimensional space ($\Phi = S^2$ is a sphere), or even a set of locations for sound source localization. The goal of the DOA estimation algorithms is to predict $\hat{y} \in Y$ that is the same or close to the label y.

We propose using neural network models to approximate the relationship between audio segments and DOA labels. This approximation is carried out in three steps (Fig. 3.1):

1. Input representation: conversion of the raw signal s to $x \in X$ (the *input space*), which can be used by the neural networks. This step may involve high-level feature extraction.
2. Neural network model: the neural network model parameterized by θ takes x as input and predicts $o = f_\theta(x)$, an element in the *output space* O. The output space of the neural network is different from the label space, because it is difficult for

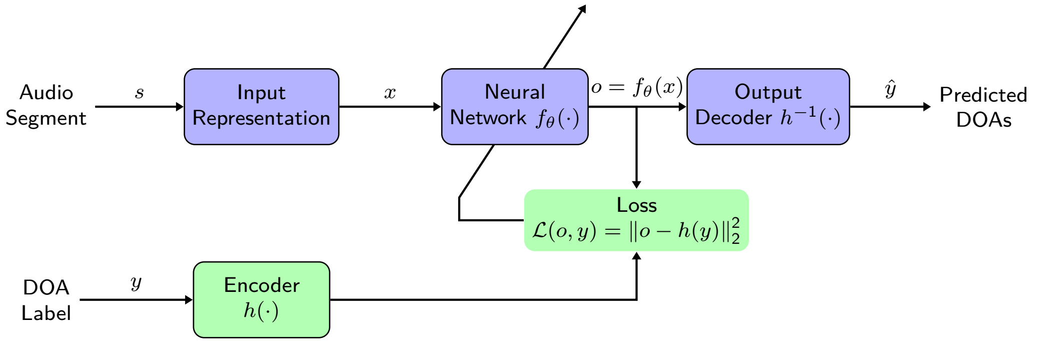


Figure 3.1 – Overview of neural network-based approach for multi-speaker direction of arrival estimation. The prediction process (depicted as blue blocks) involves extracting input representation from a raw audio segment, applying a neural network model, and decoding the network output into a set of predicted DOAs. The supervised learning process (depicted as green blocks) includes encoding the ground truth DOA label, computing the loss, which is the difference between the network output and the ideal output, and tuning the neural network parameters such that the loss is minimized.

neural network to directly predict DOA labels, which have variable sizes. Instead, the network output has a fixed number of dimensions.

3. Output decoder: finally, the network output is decoded into a DOA label $\hat{y} = h^{-1}(o)$.

Given the above procedure, we aim to find, with a set of training data, the neural network parameters θ such that the output can be decoded into the correct DOA prediction. Specifically, we are given a set of input-label pairs:

$$D = \{(x_i, y_i)\}_{i=1}^{N} \subset X \times Y,$$

and the parameters θ are trained to minimize a loss function $\mathcal{L}$ on these samples:

$$\theta^* = \arg\min_{\theta} \mathop{\mathbf{E}}_{(x,y)\in D} \mathcal{L}\left(f_\theta(x), y\right), \tag{3.1}$$

where the loss function $\mathcal{L}$ is defined based on a metric between the network output $f_\theta(x)$ and ideal output $h(y)$ that is encoded by the ground truth label y. The function h is the output encoder, which is coupled with the decoder h^{-1}. We use the *Mean Squared Error* (MSE) as the loss function, that is:

$$\mathcal{L}(f_\theta(x), y) = \|f_\theta(x) - h(y)\|_2^2. \tag{3.2}$$

Although the goal is to search for the global optimizer θ^*, in practice, achieving such a goal is not likely due to the large search space and its non-convex characteristic. Instead, the common approach is to compute the gradient of the loss with respect to the model parameters:

$$\nabla_\theta \mathcal{L}(f_\theta(x), y)$$

using *back propagation* (Rumelhart et al., 1986), and then apply *gradient descent* to modify the parameters iteratively. In our work, we use *Stochastic Gradient Descent* (SGD) that modifies network parameters based the mean gradient of the loss on a *mini-batch* of samples:

$$\theta \leftarrow \theta - \alpha g \left(\frac{1}{|B|} \sum_{(x,y) \in B} \nabla_\theta \mathcal{L} \left(f_\theta(x), y \right) \right), \tag{3.3}$$

where B is the mini-batch of samples, α is the *learning rate*, and $g(\cdot)$ is a function which modifies the gradient according to different optimization strategies. Examples of the common optimization strategies include using momentum (Sutskever et al., 2013), *Adam* (Kingma and Ba, 2015), *Adagrad* (Duchi et al., 2011), and *ADADELTA* (Zeiler, 2012).

Based on this framework, the following sections answer these remaining questions:

- What is the output encoding (h) and decoding (h^{-1}) scheme?
- What input representations are suitable for multi-speaker DOA estimation?
- What are the neural network structures and how do we train them?

3.2 Spatial Spectrum Coding

We use a spatial spectrum coding to handle an arbitrary number of sound sources. The spatial spectrum is a function of the DOA ($o : \Phi \to \mathbb{R}$), and its value indicates how likely there is a sound source at a given DOA. Unlike signal processing approaches, where the aim is to find the analytical solution for the spatial spectrum, our approach trains models to approximate an ideal spatial spectrum that we can arbitrarily define. Thus, the localization problem becomes a spatial spectrum regression problem.

3.2.1 Encoding

According to such a coding scheme, the network output vector $\mathbf{o} = \{o_l\}_{l=1}^{L}$ indicates values of the spatial spectrum on the sampled directions $\{\varphi_l\}_{l=1}^{L} \subset \Phi$, where l is the index of the DOA. In our experiments, $\{\varphi_l\}$ are 360 evenly-spaced azimuth directions. We define the ideal spatial spectrum of a label y as the maximum of Gaussian curves centered at the sound source directions (Fig. 3.2):

$$h(y) = \{o_l^*\}_{l=1}^{L}, \tag{3.4}$$

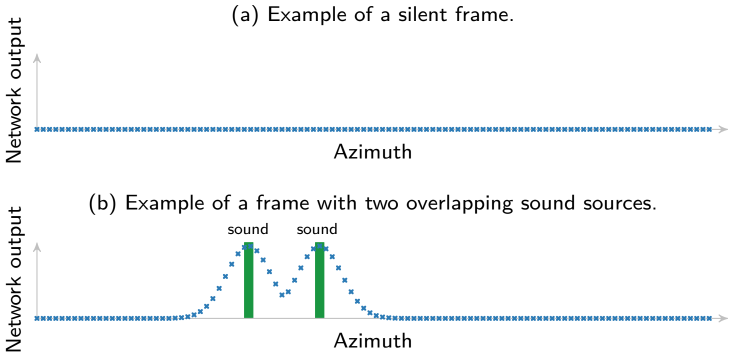


Figure 3.2 – Gaussian-based spatial spectrum output coding for multiple sources.

and

$$o_l^* = \begin{cases} \max\limits_{\varphi' \in y} \left\{ e^{-d(\varphi_l, \varphi')^2/\sigma^2} \right\} & \text{if } |y| > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{3.5}$$

where $d(\cdot, \cdot)$ is the angular distance, and σ is a constant that controls the width of the Gaussian curves. The ideal output values are close to zero at directions away from the sound sources. They peak at the ground truth directions with values of one, and gradually decrease to zero as the distance to the sound source increases. The values in each direction can be interpreted as a score or likelihood of the presence of a sound source in that direction.

Unlike previous posterior probability coding (Section 2.1.2), the spatial spectrum coding is not constrained as a probability distribution (the output layer is not normalized by a softmax function). It can be all zero when there is no sound source, or contains the same number of peaks as the number of sound sources. This coding scheme allows detection of an arbitrary number of sound sources. In addition, the soft assignment of the output values, in contrast to the 0/1 assignment in posterior probability coding, takes the correlation between adjacent directions into account, allowing better generalization of the neural networks.

3.2.2 Decoding

During inference, the network output is decoded to a prediction $\hat{y} \in Y$ by finding the local maxima in the predicted spatial spectrum. When the number of sources z is unknown, the peaks above a given threshold ξ are taken as predictions:

$$\hat{y} = h^{-1}(o; \xi) = \left\{ \varphi_l : o_l > \xi \quad \text{and} \quad o_l = \max_{d(\varphi_i, \varphi_l) < \sigma_n} o_i \right\}, \tag{3.6}$$

where $o = f_\theta(x)$ is the network output and σ_n is the neighborhood size for non-maxima suppression. When z is known, the z highest peaks are taken as predictions:

$$\hat{y} = h^{-1}(o; z) = \left\{ \varphi_l : \text{among the } z \text{ greatest } o_l = \max_{d(\varphi_i, \varphi_l) < \sigma_n} o_i \right\}. \tag{3.7}$$

3.3 Input Representations for DOA Estimation

We investigate three different input representations, that are extracted from the raw input signals captured by M microphones. The short time Fourier transform (STFT) of the input signal is denoted by $S_i(\omega), i = 1, \ldots, M$, where i is the microphone index and ω is the frequency in the discrete domain. The time index is omitted for clarity. The specific parameters of the input representations are chosen according to our experimental setting. In this setting, we are given four-channel audio signals ($M = 4$) sampled at 48 kHz. The input representations are extracted from 170 ms (8192 samples) segments with 50% overlap. Such a segment size provides a good balance between the amount of information and time resolution.

3.3.1 GCC-PHAT Coefficients

The GCC-PHAT (Knapp and Carter, 1976) is one of the most popular method for estimating the *Time Difference of Arrival* (TDOA) between microphones. The GCC-PHAT between channel i and j is formulated as:

$$g_{ij}(\tau) = \mathcal{R}\left(\sum_{\omega} \frac{S_i(\omega) S_j(\omega)^*}{|S_i(\omega) S_j(\omega)^*|} e^{j\omega\tau} \right) \tag{3.8}$$

where τ is the delay in the discrete domain, $(\cdot)^*$ denotes the complex conjugation, and $\mathcal{R}(\cdot)$ denotes the real part of a complex number. Usually, the peak in the GCC-PHAT is used for TDOA estimation. However, under real condition, the GCC-PHAT is corrupted by noise and reverberation. Therefore, we use the values of the GCC-PHAT coefficients as the input feature instead of a single estimation of the TDOA (Fig. 3.3). Specifically, the input representation is the concatenation of the center GCC-PHAT coefficients of all $M(M-1)/2$ microphone pairs, as it was proposed in (Xiao et al., 2015):

$$x = \{g_{i,j}(\tau)\}_{i \neq j; \tau = -D, -D+1, \ldots, D} \,. \tag{3.9}$$

Here, D is the center width, and it should be set according to the microphone array size, namely:

$$D \geq \frac{\max\{d_{ij}\}}{c} f_s, \tag{3.10}$$

where d_{ij} is the distance between the microphone pair (i, j), f_s is the sampling rate

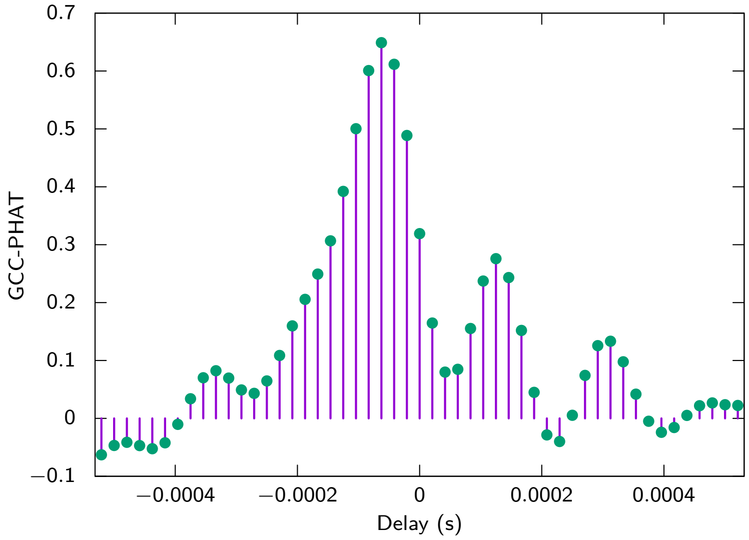


Figure 3.3 – Example of GCC-PHAT coefficients extracted from the audio signals of a pair of microphones. The input representation x is the concatenation of these coefficients extracted from all $M(M-1)/2$ pairs.

and c is the speed of sound. With the above formula satisfied, the coefficients include information of all possible time delay. In the experiments, we choose $D = 25$, therefore there are 51 coefficients for each pair of microphones.

3.3.2 GCC-PHAT on the Mel-scale Filter Bank

The GCC-PHAT is not optimal for TDOA estimation of multiple source signals since it equally sums over all frequency bins disregarding the "sparsity" of speech signals in the *Time-Frequency* (TF) domain and the randomly distributed noise which may be stronger than the signal in some TF bins.

To preserve delay information on each frequency band and to allow sub-band analysis, we propose the GCCFB. Hence, the second type of input feature is formulated as:

$$g_{ij}(f,\tau) = \mathcal{R}\left(\frac{\sum_{\omega\in\Omega_f} H_f(\omega)\frac{S_i(\omega)S_j(\omega)^*}{|S_i(\omega)S_j(\omega)^*|}e^{j\omega\tau}}{\sum_{\omega\in\Omega_f} H_f(\omega)}\right), \tag{3.11}$$

where f is the filter index, H_f is the transfer function of the f-th mel-scaled triangular filter, and Ω_f is the support of H_f. Under this setting, the network input is:

$$x = \{g_{i,j}(f,\tau)\}_{i\neq j; f=1,\dots,F; \tau=-D,-D+1,\dots,D}, \tag{3.12}$$

where F is the number of filters.

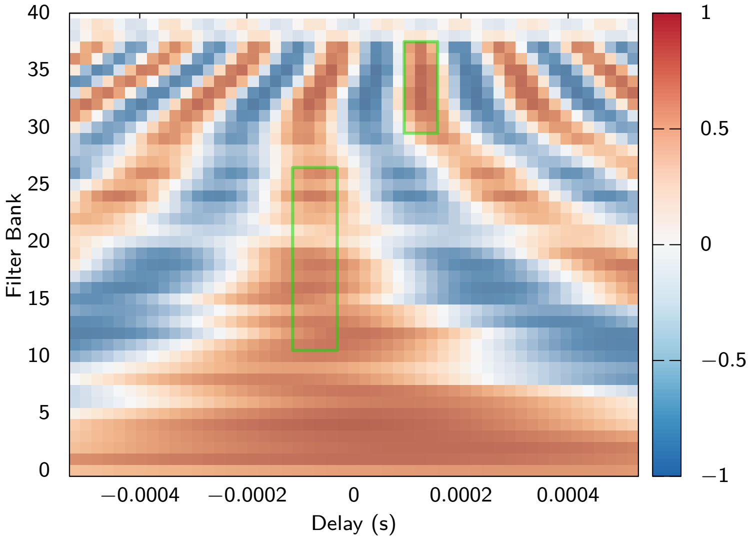


Figure 3.4 – Example of GCCFB extracted from a frame with two overlapping sound sources.

Fig. 3.4 shows an example of the GCCFB representation of a frame in which two speech signals overlap. Each row corresponds to the GCC-PHAT in an individual frequency band. The frequency-based decomposition allows the estimation of the TDOAs by looking into local areas rather than across all frequency bins. In the example, the areas marked by the green rectangles correspond to two separate sources with different delays and which produce high cross-correlation values in different frequency bands (and hence, for different filter indices). In the experiments, we use 40 mel-scale filters covering the frequencies from 100 to 8000 Hz.

3.3.3 Short Time Fourier Transform

The third type of input representation simply comprises the real and imaginary parts of the time-frequency domain signal. In contrast to high-level features extraction, such a representation retains all the information of the signals and allows the network to implicitly extract informative features for localization, which potentially include both inter-channel cues (i.e. level/phase difference) and intra-channel cues (i.e. spectral features). In addition, as speech is known to be sparse in the time-frequency domain, with such a representation the network can learn to separate overlapping sound sources in the mixed input signals.

Specifically, we compute the STFT of the segments with a frame size of 43 ms (2048 samples) and 50% overlap. Thus, there are seven frames in each segment (170ms). We only use the frequency bins between 100 and 8000 Hz, so that the number of frequency bins is reduced to 337. We take the real and imaginary part of the complex values instead of the phase and power, so that we avoid the discontinuity problem of the phase at π

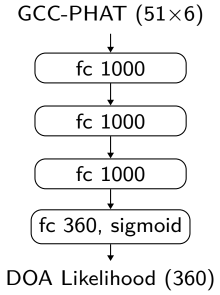


Figure 3.5 – Neural network architecture of the MLP-GCC model. Batch normalization and ReLU activation function after each hidden layer are omitted in the graph.

and $-\pi$. Eventually, the dimension of the input vector is $7 \times 337 \times 8$.

3.4 Neural Network Architectures

We investigate four different neural network architectures for multi-speaker DOA estimation.

3.4.1 Multilayer perceptron with GCC-PHAT: MLP-GCC

MLP-GCC uses GCC-PHAT as input and contains three hidden layers (Fig. 3.5), each of which is a fully-connected layer with a *Rectified Linear Unit* (ReLU) activation function (Nair and Hinton, 2010) and *Batch Normalization* (BN) (Ioffe and Szegedy, 2015). The last layer is a fully-connected layer with sigmoid activation function. The sigmoid function is bounded between 0 and 1, which is the range of the desired output. According to our experiments, this helps the network to converge to a better result.

3.4.2 Convolutional neural network with GCCFB: CNN-GCCFB

MLP with fully-connected layers is not suitable for high-dimensional input features (such as GCCFB) because the large input dimension requires a large number of parameters to be trained, making the network computationally expensive and prone to overfitting. Convolutional neural networks can learn local features with a reduced number of parameters by weight sharing. This leads to the idea of using CNN for the input feature of GCCFB.

We use the CNN structure shown in Fig. 3.6, which consists of four convolutional layers (with ReLU activation and BN) and a fully connected layer at the output (with sigmoid activation). The local features are not shift-invariant since the position of the feature (the delay and frequency) is the important cue for SSL. Therefore, we do not apply any pooling after the convolutions. Instead, we apply the filters with a stride of 2, expecting

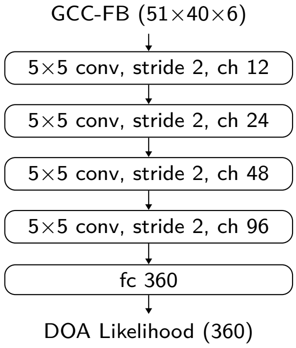

GCC-FB (51×40×6)
5×5 conv, stride 2, ch 12
5×5 conv, stride 2, ch 24
5×5 conv, stride 2, ch 48
5×5 conv, stride 2, ch 96
fc 360
DOA Likelihood (360)

Figure 3.6 – Neural network architecture of the CNN-GCCFB model. Batch normalization and ReLU activation function after each hidden layer are omitted in the graph.

that the network learns its own spatial downsampling.

3.4.3 Two-stage neural network with GCCFB: TSNN-GCCFB

The CNN-GCCFB considers the input features as images without taking their properties into account, which may not yield the optimal model. For the third architecture, we design the weight sharing in the network exploiting these properties of the GCCFB:

- In each TF bin, there is generally only one predominant speech source, thus we extract local features in each frequency band before such information is aggregated into a broadband prediction.
- Features with the same delay in different channels of the GCCFB (Fig. 3.4) do not correspond to each other locally, because a sound source has different delays among different microphone pairs. Instead of using local convolution kernels, feature extraction should take the whole delay axis into account.

Based on these considerations, we propose a two-stage neural network (Fig. 3.7). The first stage extracts latent DOA features from narrow-band signals, by repeatedly applying *Subnet 1* on individual frequency regions (five filters) that span all delays and all microphone pairs. The second stage aggregates information across all frequencies in a neighbor DOA area and outputs the spatial spectrum. Similarly, the *Subnet 2* is repeatedly used for all DOAs in the second stage.

To train such network, we adopt a two-step training scheme: First, we train the Subnet 1 in the first stage using the ideal spatial spectrum as the desired latent feature for all frequencies. In such way, we obtain DOA and frequency-related features that help the NN to converge to a better result in the next step. During the second step, both stages are trained in an end-to-end manner. In our experiments, Subnet 1 is a 2-hidden-layer MLP, and Subnet 2 is a 1-hidden-layer MLP. All the hidden layers are of size 500.

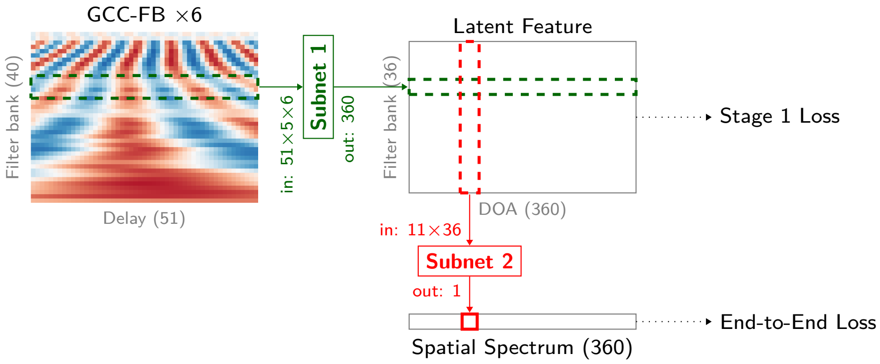


Figure 3.7 – NN architecture of our two-stage neural network with GCCFB as input (TSNN-GCCFB). The first and second stages are marked as green and red, respectively. The number of filters in the latent feature is less than that in the input feature, because there is no padding applied to the moving window of the input of *Subnet 1*.

3.4.4 Residual Network with STFT : ResNet-STFT

We design a fully-convolutional residual neural network (ResNet) for the STFT input (Fig. 3.8). In addition to weight sharing for reducing the number of parameters, the residual connection in the ResNet allows the construction of very deep neural network models, and therefore increases their capabilities at extracting high-level features (He et al., 2015).

Similar to TSNN-GCCFB, this network comprises two parts. Each part convolves along different axes. In the first part, the network convolves along the time and frequency axes. Specifically, it includes two layers of strided convolution in the frequency axis for downsampling as well as feature extraction, five residual blocks for the extraction of higher level features, and a layer projecting the features to the DOA space. The output of the first part of the network is time-frequency local, meaning that each output value is derived from a local time-frequency region in the input.

In the second part, the network convolves along the DOA axis. It aggregates features from all TF bins at the neighboring directions, and outputs the spatial spectrum.

Like the TSNN-GCCFB, the training of the ResNet-STFT consists of two stages. In the first stage, we train the first part of the network by considering its output as short-term narrow-band predictions of the spatial spectrum, and using a loss function that repeats the ultimate loss (Eq. (3.2)) across TF bins:

$$\mathcal{L}_I\left(f_{I,\theta}(x), y\right) = \sum_{t,k} \mathcal{L}\left(f_{I,\theta}(x)[t,k], y\right), \tag{3.13}$$

where $f_{I,\theta}(x)[t,k]$ is the output of the first part of the network at time t and frequency k. The pre-trained parameters are then used to initialize the network for the second stage,

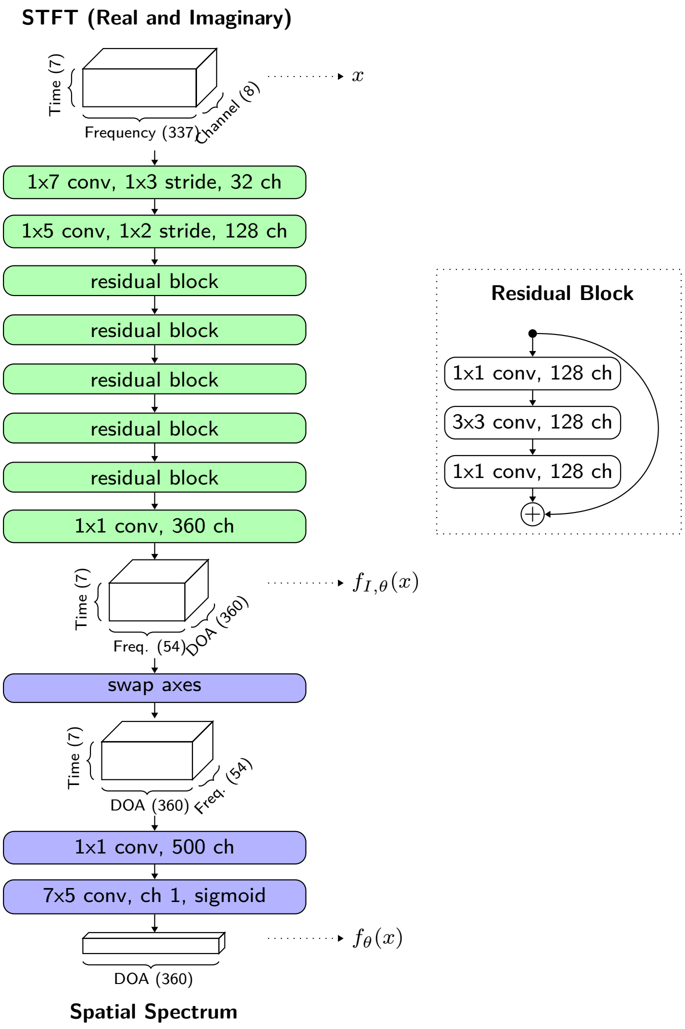
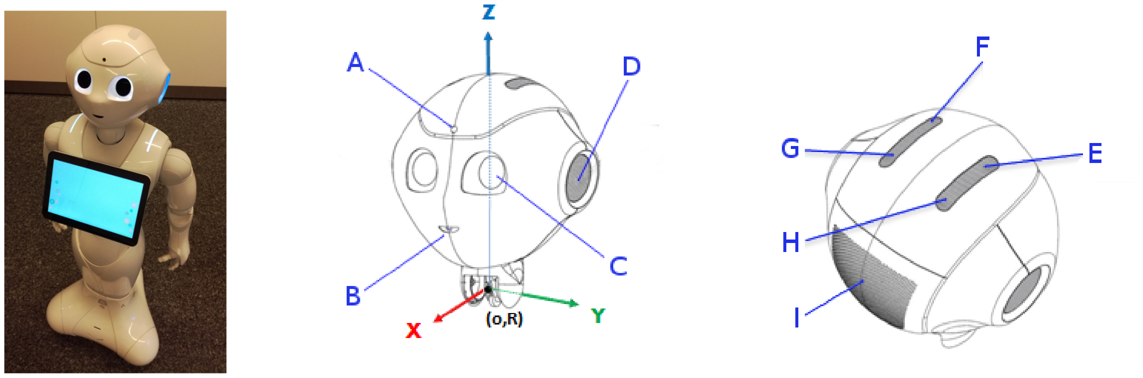


Figure 3.8 – Network architecture of ResNet-STFT. It uses STFT of the audio signals as the input and predicts the spatial spectrum of the sound sources. It consists of two parts: the first part (green) applies convolution along the time and frequency axes, and the second part (blue) applies convolution along the DOA axis. The intermediate TF-local output $f_{I,\theta}(x)$ is used for the first-stage training with Eq. (3.13), while the network output $f_\theta(x)$ is used with Eq. (3.2) for the end-to-end training. Batch normalization and ReLU activation functions after each hidden layer are omitted in the graph.



(a) Pepper.

(b) Front view of the head.

(c) Rear view of the head.

Figure 3.9 – Pepper and its sensors on the robot. The labeled parts are: (A) front camera; (B) bottom camera; (C) depth camera; (D) loudspeaker; (E) front right microphone; (F) front left microphone; (G) rear left microphone; (H) rear right microphone; (I) vent holes for the cooling fans.

where the whole network is trained with the loss function defined by Eq. (3.2).

3.5 Data Collection

We collected around 24 hours of real audio recordings with a robot. These data include sounds from both loudspeakers and human talkers (Table 3.1).

3.5.1 Robot Platform

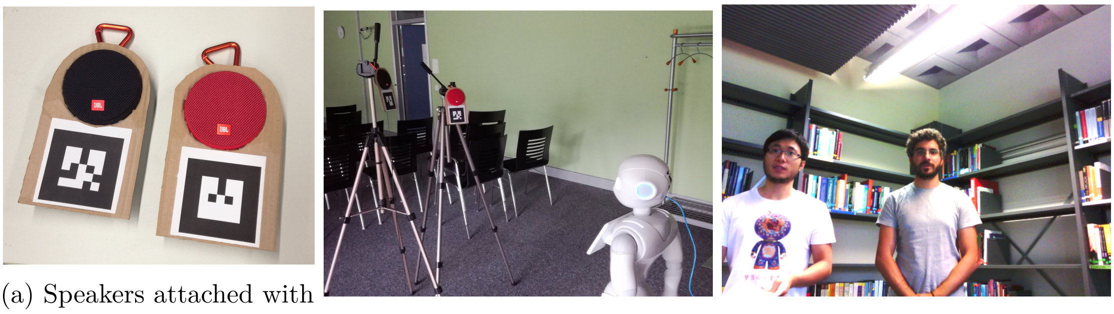
We used a Pepper robot[1] developed by Softbank Robotics. The robot is a 1.3 meter tall humanoid robot (Fig. 3.9a). There are two RGB cameras and a depth camera on the robot head (Fig. 3.9b). We used the front camera for extracting sound location labels.

There are four microphones on the top of the robot head (Fig. 3.9c). The four microphones are coplanar, forming a rectangle with 5.80 cm along the front-rear sides and 6.86 cm along the left-right sides. The microphones are directional with a forward look direction, and the sampling rate is 48 kHz. The captured audio signals are strongly affected by the robot ego noise. It mainly consists of the noise of the fans next to the microphone array (Fig. 3.9c).

[1] http://doc.aldebaran.com/2-5/home_pepper.html

Table 3.1 – Specifications of the recorded data

	Loudspeaker		Human talkers
	Training	Test	Test
# of frames	507k	262k	2098
- no source	106k	54k	1169
- single source	350k	179k	788
- two sources	51k	29k	144
# of files	4208	2393	21
- single source	2808	1597	–
- two sources	1400	796	21
# of male speakers	105	8	12
# of female speakers	43	8	2
Total duration	16 hours	8 hours	4 min
Azimuth (°)	$[-180, 180]$	$[-180, 180]$	$[-24, 23]$
Elevation (°)	$[-39, 56]$	$[-29, 45]$	$[-14, 13]$
Distance (m)	$[0.5, 1.8]$	$[0.5, 1.9]$	$[0.8, 2.1]$



(a) Speakers attached with markers.

(b) Loudspeakers.

(c) Human talkers.

Figure 3.10 – Data collection with Pepper.

3.5.2 Recording with Loudspeakers

We collected data by playing clean speech from loudspeakers and recording the sounds with the robot (Fig. 3.10b).

The clean speech data were taken from the AMI corpus (McCowan et al., 2005), which contains spontaneous speech of people interacting in meetings. We selected the non-overlapping segments recorded from the headset microphones. For the convenience of recording, segments with a minimum length of five seconds were used. The clean speech data are split into training and test sets in a way that speakers in the training set are not in the test set (Table 3.2). We applied phoneme forced alignment (with annotated transcripts) on the clean speech data to acquire the phoneme and voice activity labels analyzed at 100Hz.

Table 3.2 – Specifications of source clean speech segments used for loudspeaker data collection.

	Training	Test
# of segments	5409	472
# of speakers	148	16
# of male speakers	105	8
# of female speakers	43	8
Average duration (s)	11.6	10.6

For each recorded audio file, we played speech segments from one or two loudspeakers. When two loudspeakers were playing overlapping sounds, the offset between the two sounds was randomly chosen from a uniform distribution of $[-2, 2]$ seconds.

The data collection was conducted in three rooms at Idiap with different sizes and settings, which are:

- **Room 106** a large meeting room (used for both training and test data collection).
- **Room 301** a small meeting room (used exclusively for training data collection).
- **Library** a small room with book shelves (used exclusively for test data collection).

All of these rooms are ordinary office rooms without any reverberation control. We programmed the robot to move its head automatically to acquire a large diversity of relative locations of the sound sources. After every 40 different head poses, we moved the robot and loudspeakers a different set of locations, and continued the same procedure.

The sound location labels were automatically obtained through detecting the markers (Fig. 3.10a) attached on the loudspeakers with the camera on the robot. In addition, the frame-level annotation is derived from the voice activity labels of the clean signal.

3.5.3 Recording with Human Talkers

To evaluate SSL methods in real HRI, we collected a second test set with human talkers (Fig. 3.10c). During the recording, talkers spoke to the robot with phrases for interactions. This dataset includes recordings with both single utterances and overlapping ones. Another room, which is of the same size of Room 301 but with different objects, was used for the human talker recordings.

We manually annotated the voice activity labels and automatically acquired the mouth position by applying a multiple person tracker (Khalidov and Odobez, 2017) with detection from the *Convolutional Pose Machine* (CPM) (Wei et al., 2016; Cao et al., 2017).

3.6 Experiments

We implemented the proposed methods and compared them to the traditional SSL approaches. We consider frame-level azimuth estimation of sound sources with frames of 170ms long. For output encoding (Eq. (3.4)) and decoding (Eq. (3.6)), the parameters are chosen as: $\sigma = \sigma_n = 8°$.

3.6.1 Network Training

We trained the *Neural Networks* (NNs) with the loudspeaker training set, which includes a total of 506k frames. We used the Adam optimizer (Kingma and Ba, 2015) with a mini-batch size of 256 and an initial learning rate of 0.001. The learning rate was decreased by half after every two epochs. MLP-GCC and CNN-GCCFB were trained for ten epochs. We trained TSNN-GCCFB and ResNet-STFT for four epochs for the first stage and another ten epochs for end-to-end training.

3.6.2 Evaluation Protocol

We considered two evaluation settings:

- when the number of sound sources is known, or
- when it is not.

When the number of sound sources is known, we evaluate how close the predicted DOAs are from the ground truth. In this case, the predictions $\hat{y}_i = \{\hat{\varphi}_{ij} : j = 1, \dots, z_i\}$ are the DOAs of the z_i (number of sound sources) highest peaks in the output spatial spectrum (according to Eq. (3.7)). The indices js are selected such that the predicted DOA $\hat{\varphi}_{ij}$ is nearest to the ground truths DOA φ_{ij} in label $y_i = \{\varphi_{ij} : j = 1, \dots, z_i\}$. As performance measure, we compute the *Mean Absolute Error* (MAE) in terms of angular distance between the predictions and the ground truth:

$$\text{MAE} = \frac{\sum_i \sum_{j=1}^{z_i} d(\hat{\varphi}_{ij}, \varphi_{ij})}{\sum_i z_i}. \tag{3.14}$$

We also compute the *Accuracy* (ACC), that is the percentage of the predictions of which the error is less than a given admissible error E_a:

$$\text{ACC} = \frac{\sum_i \sum_{j=1}^{z_i} \mathbf{1}_{d(\hat{\varphi}_{ij}, \varphi_{ij}) < E_a}}{\sum_i z_i}, \tag{3.15}$$

where $\mathbf{1}$ is the indicator function.

When the number of sound sources is unknown, we evaluate the DOA estimation in terms of sound source detection. The predictions $\hat{y}_i = \{\hat{\varphi}_{ik} : k = 1, \dots, \hat{z}_i\}$ decoded from the network output by Eq. (3.6) are matched with the ground truth DOAs. We use $m(\hat{\varphi}_{ik}, \varphi_{ij})$ to denote a match. The number of predicted sound sources $\hat{z}_i$ may not be equal to the number of ground truth sources z_i, and each ground truth source is matched with at most one prediction (could be none), which is the nearest prediction with an error less than E_a:

$$m(\hat{\varphi}_{ik}, \varphi_{ij}) = \begin{cases} 1 & \text{if } d(\hat{\varphi}_{ik}, \varphi_{ij}) < E_a \text{ and } k = \arg\min_{l=1}^{\hat{z}_i} d(\hat{\varphi}_{il}, \varphi_{ij}), \\ 0 & \text{otherwise.} \end{cases} \tag{3.16}$$

We vary the prediction threshold ξ (Eq. (3.6)) and plot *precision-recall curves*. The *precision* is the percentage of correct predictions among all predictions:

$$\text{Precision} = \frac{\sum_i \sum_{j=1}^{z_i} \sum_{k=1}^{\hat{z}_i} m(\hat{\varphi}_{ik}, \varphi_{ij})}{\sum_i \hat{z}_i}. \tag{3.17}$$

The *recall* is the percentage of correct detection out of all ground truth sources:

$$\text{Recall} = \frac{\sum_i \sum_{j=1}^{z_i} \sum_{k=1}^{\hat{z}_i} m(\hat{\varphi}_{ik}, \varphi_{ij})}{\sum_i z_i}. \tag{3.18}$$

An admissible error $E_a = 5°$ is used for the evaluation.

3.6.3 Baseline Methods

We include the following popular spatial spectrum-based methods for comparison:

- **SRP-PHAT**: steered response power with phase transform (DiBiase et al., 2001);
- **SRP-NONLIN**: SRP-PHAT with a non-linear modification of the score. It is a multi-channel extension of GCC-NONLIN from (Blandin et al., 2012);
- **MVDR-SNR**: *Minimum Variance Distortionless Response* (MVDR) beamforming with *Signal-to-Noise Ratio* (SNR) scoring (Blandin et al., 2012);
- **SEVD-MUSIC**: *Multiple Signal Classification* (MUSIC) (Schmidt, 1986), assuming spatially white noise and one signal in each bin;
- **GEVD-MUSIC**: MUSIC with generalized eigenvector decomposition (Schmidt, 1986; Nakamura et al., 2009), assuming noise is pre-measured and one signal in each TF bin.

For all the above methods, the empirical spatial covariance matrices are computed using blocks containing 7 frames of 2048 samples with 50% overlap, so that each block is 170ms long (same as the network input of the proposed approaches).

Table 3.3 – DOA estimation performance on the loudspeaker dataset with a priori knowledge about the number of sources. $E_a = 5°$.

Dataset	Loudspeaker					
Subset (# of frames)	Overall (207k)		$N = 1$ (178k)		$N = 2$ (29k)	
	MAE (°)	ACC	MAE (°)	ACC	MAE (°)	ACC
MLP-GCC	4.9	0.92	4.2	0.94	9.2	0.77
CNN-GCCFB	4.8	0.90	4.1	0.93	9.1	0.73
TSNN-GCCFB	5.4	0.91	4.6	0.93	10.1	0.77
ResNet-STFT	**3.1**	**0.94**	**2.7**	**0.95**	**5.8**	**0.85**
SRP-PHAT	21.5	0.78	19.0	0.82	37.0	0.50
SRP-NONLIN	25.7	0.73	23.8	0.77	37.6	0.51
MVDR-SNR	23.2	0.76	21.2	0.79	35.2	0.55
SEVD-MUSIC	29.1	0.66	27.6	0.69	38.1	0.47
GEVD-MUSIC	25.4	0.64	23.2	0.67	39.3	0.44

Table 3.4 – DOA estimation performance on the human talkers dataset with a priori knowledge about the number of sources. $E_a = 5°$.

Dataset	Human Talkers					
Subset (# of frames)	Overall (929)		$N = 1$ (788)		$N = 2$ (141)	
	MAE (°)	ACC	MAE (°)	ACC	MAE (°)	ACC
MLP-GCC	5.0	0.93	4.4	0.94	8.1	0.84
CNN-GCCFB	4.8	0.93	4.2	0.96	8.3	0.77
TSNN-GCCFB	4.1	0.95	3.8	0.96	5.8	0.90
ResNet-STFT	**2.6**	**0.97**	**2.0**	**0.98**	**5.7**	**0.93**
SRP-PHAT	5.4	0.88	2.6	0.93	20.9	0.56
SRP-NONLIN	4.8	0.90	2.5	0.94	18.1	0.68
MVDR-SNR	4.4	0.90	2.5	0.94	15.2	0.68
SEVD-MUSIC	6.4	0.85	3.0	0.88	25.1	0.64
GEVD-MUSIC	6.5	0.81	3.6	0.85	22.2	0.63

3.6.4 DOA Estimation Performance

Table 3.3 and Table 3.4 show the results of localization with a priori knowledge about the number of sound sources. On the loudspeaker dataset (Table 3.3), all four proposed glsdnn models achieve on average less than 5° error and more than 90% accuracy. In particular, the best approach ResNet-STFT outperforms others in all criteria, with 3.1°error and 94% accuracy. In contrast, the best baseline method (SRP-PHAT) has 21.5° error and only 78% accuracy.

For the human talker dataset (Table 3.4), the baseline methods have slightly better MAE on single-source frames than three of the DNN approaches. However, their performance degrades radically on frames with overlapping sources. In contrast, all the proposed approaches outperform the baseline methods in terms of accuracy, and their performance is not much affected by the condition of overlapping sources. The loudspeaker dataset is in general more challenging because it contains samples with lower SNR and wider range of azimuth directions. The sources from the rear are difficult to detect due to the directivity of the microphones.

In terms of simultaneous detection and localization with an unknown number of sound sources (Fig. 3.11), all proposed methods outperform the baseline methods, achieving more than 90% precision and recall on both datasets. We also notice that, unlike signal processing approaches, our DNN-based methods are not affected by the condition of no a priori knowledge about the number of sound sources. This indicates that our output coding and data-driven approach are effective for detecting the number of sound sources. This is because all peaks in the desired spatial spectrum have a height of one, no matter what the SNR is or how many sources there are. In contrast, the peak height in the traditional spatial spectrum such as SRP-PHAT is affected by SNR and overlapping sources.

Among the proposed models, both TSNN-GCCFB and ResNet-STFT achieve better results on frames with overlapping sound. This justifies the usage of features decomposed in the frequency domain (GCCFB) or TF domain (raw STFT), that allows networks to separate the overlapping sounds. ResNet-STFT clearly outperforms the other DNN-based approaches, which shows that with a deep network structure, the network can learn by itself to extract the optimal localization cues from raw signals. Moreover, STFT with both phase and power spectral information is indeed useful for sound source localization.

A demo video of the results is available online[2].

[2] https://www.youtube.com/watch?v=_4EwuVlE_pU&t=153s

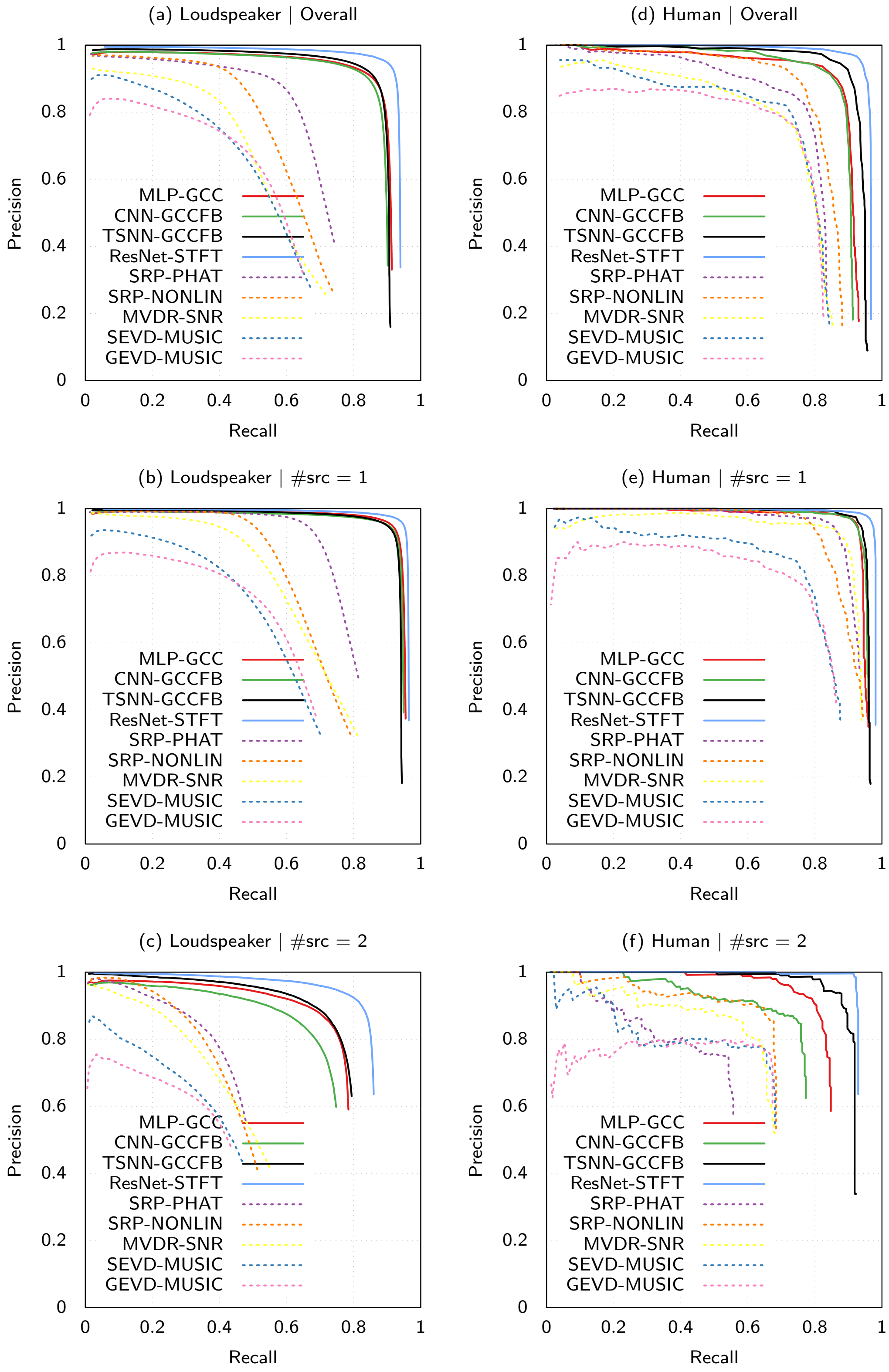


Figure 3.11 – Detection and localization performance. The figure titles indicates the test set (loudspeakers or human talkers) and selected frames (all, single-source frame, or two-source frame). $E_a = 5°$.

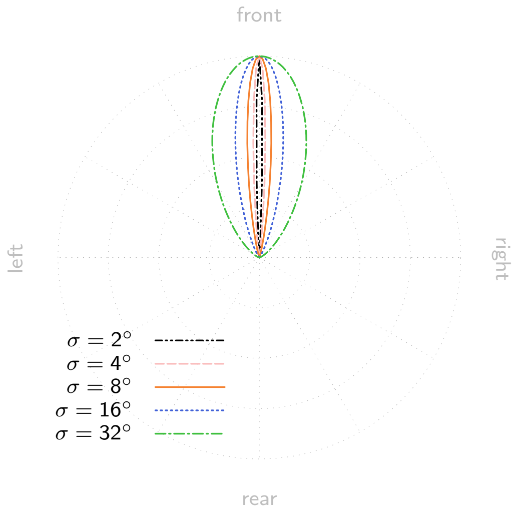


Figure 3.12 – Ideal spatial spectra of one active sound source with different σ values in a polar plot.

Table 3.5 – DOA estimation performance on the loudspeaker dataset with different σ values for output encoding.

Dataset	Loudspeaker					
Subset (# of frames)	Overall (207k)		$N = 1$ (178k)		$N = 2$ (29k)	
	MAE (°)	ACC	MAE (°)	ACC	MAE (°)	ACC
ResNet-STFT ($\sigma = 2°$)	4.1	0.91	3.4	0.94	8.3	0.77
ResNet-STFT ($\sigma = 4°$)	3.6	0.93	3.1	0.95	6.4	0.82
ResNet-STFT ($\sigma = 8°$)	**3.1**	**0.94**	**2.7**	**0.96**	**5.8**	**0.85**
ResNet-STFT ($\sigma = 16°$)	3.8	0.91	2.8	0.94	9.9	0.77
ResNet-STFT ($\sigma = 32°$)	6.5	0.85	3.3	0.88	26.1	0.63

3.6.5 Curve Width for Output Coding

In the results reported above, the desired output spatial spectrum are all encoded with a curve width of $\sigma = 8°$ (Eq. (3.4)). We modified this parameter and examined how it impacts the DOA estimation performance. We trained ResNet-STFT with different σ values ranging from 2° to 32° (Fig. 3.12). The DOA estimation performance (Tables 3.5 and 3.6, Fig. 3.13) shows that $\sigma = 8°$ is the optimal setting. The spatial resolution is much reduced when using large σ values, therefore we can notice a significant performance degradation on frames with overlapping sound sources. In contrast, small σ values may generate spatial spectra with better spatial resolution. However, such spatial spectra have smaller support, thus tend to be sensitive to condition changes.

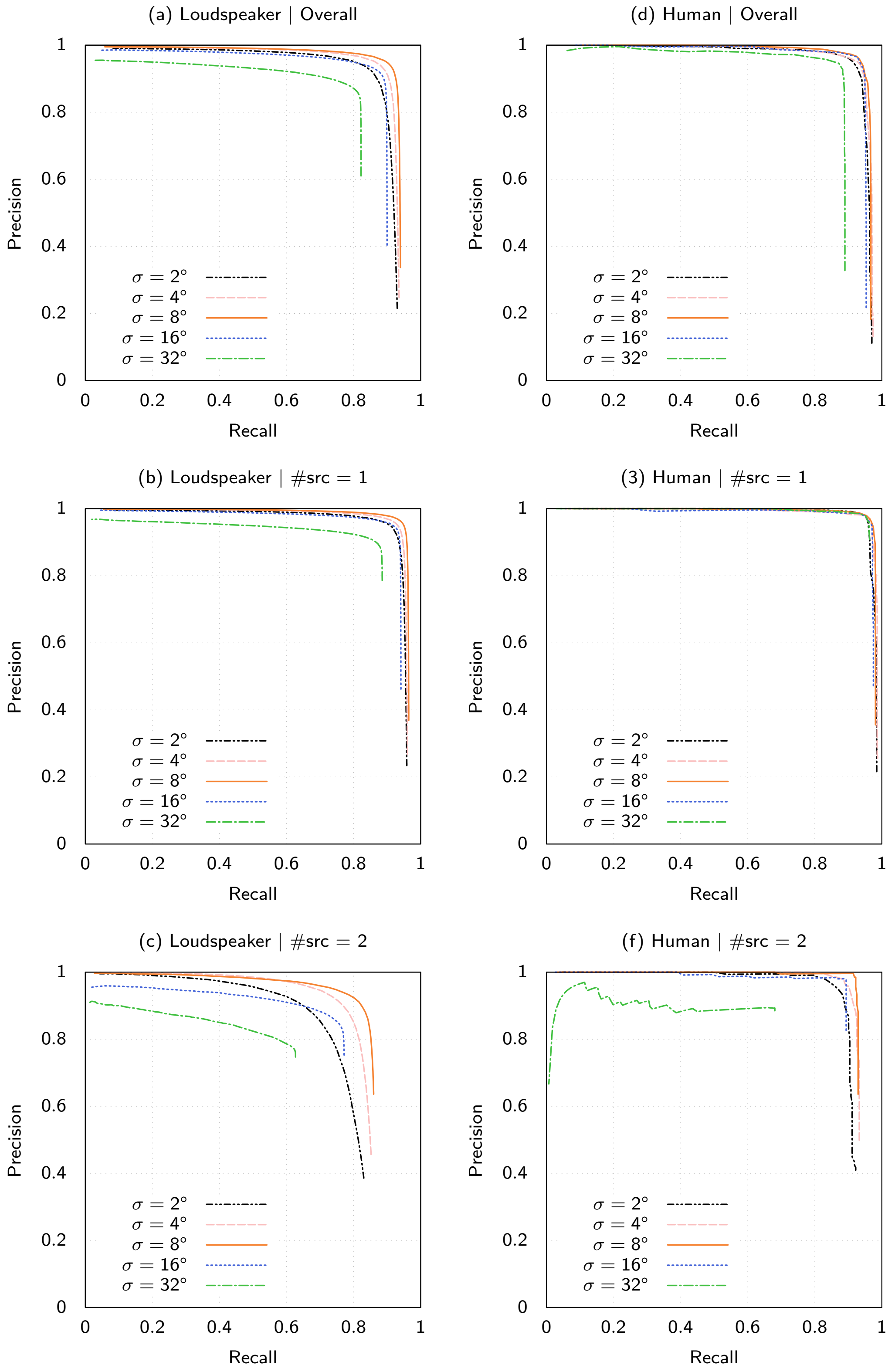


Figure 3.13 – Detection and localization performance of different σ values for output encoding.

Table 3.6 – DOA estimation performance on the human talkers dataset with different output coding curve widths.

Dataset	Human Talkers					
Subset (# of frames)	Overall (929)		$N = 1$ (788)		$N = 2$ (141)	
	MAE (°)	ACC	MAE (°)	ACC	MAE (°)	ACC
ResNet-STFT ($\sigma = 2°$)	2.8	0.96	2.4	0.97	4.7	0.90
ResNet-STFT ($\sigma = 4°$)	**2.4**	**0.97**	2.1	**0.98**	**4.2**	**0.93**
ResNet-STFT ($\sigma = 8°$)	2.6	**0.97**	**2.0**	**0.98**	5.7	**0.93**
ResNet-STFT ($\sigma = 16°$)	3.3	0.96	2.2	0.97	9.8	0.89
ResNet-STFT ($\sigma = 32°$)	7.2	0.92	2.5	0.96	33.1	0.68

3.7 Summary

This chapter has investigated four neural network models for simultaneous detection and localization of sound sources. We have proposed a Gaussian-based spatial spectrum output coding, making it possible to train DNNs to detect an arbitrary number of overlapping sound sources. This approach is among the first deep learning based approaches for multi-speaker DOA estimation with no a priori knowledge about the number of sources.

We have collected a large amount of real data, including recordings with loudspeakers and human talkers, for both training and evaluation. The results of the comprehensive evaluation show that our proposed methods significantly outperform the traditional spatial spectrum-based approaches.

In addition, ResNet-STFT, the network with deep structures and raw signal input, achieves the best performance. This shows that low-level input representations, as they retain all information about the signals, can be better than hand-crafted high-level features for deep neural networks.

4 Domain Adaptation for DOA Estimation Models

We have shown in the previous chapter that deep learning based approaches significantly outperform the traditional spatial spectrum based approaches. However, this is based on the assumption that a sufficient number of unbiased labeled data are available. In fact, in the learning-based approaches, the difficulties have been shifted from modeling the complex environments to the need of collecting a sufficient number of training data covering all variabilities in the target test environment. Such variabilities include various sound classes, samples per class, source locations, reverberation, noises, and objects in the scene.

In addition to making audio recordings, annotating them with the location labels is also particularly costly. As audio data do not intrinsically contain direct information for annotation of the sound source locations, external devices, such as camera or motion capture systems, are needed. Moreover, since multi-channel audio data are distinct among different types of microphone arrays, individual target-domain data collection is needed for each new type of microphone array.

One possible solution to costly data collection is to develop device-independent *Sound Source Localization* (SSL) models, allowing real audio data to be reused for multiple devices. This is a difficult problem and little research has been conducted in this direction. Models using uniform input representation, such as the ambisonics intensity vectors (Perotin et al., 2018; Tang et al., 2019), could potentially be applied to multiple devices. However, this idea has not been verified by experiments and the conversion of multi-channel audio signal to ambisonics intensity vectors is limited to non-coplanar microphone arrays.

As we have discussed in Section 2.1.2, acoustic simulation is a popular way to obtain training data for SSL. Although we can simulate a large number of audio samples under various conditions at low cost, the simulated samples are biased, as it is unlikely to perfectly reproduce real sound propagation. Therefore, the models trained with simulated data are usually suboptimal for real applications.

Domain Adaptation (DA) techniques (discussed in Section 2.2) aim to mitigate the effect of the condition mismatch between the source domain (simulated data) and the target domain (real situations). Although there is a substantial amount of research in DA (Ben-David et al., 2010; Wang and Deng, 2018), its application to SSL has not been studied in depth so far. DA can potentially exploit the wide variety of conditions from the simulated data and the unbiased samples from the available real data. It aims to train model with the best performance in real test scenarios. Previously, domain adaptation by entropy minimization has been studied for single-source localization (Takeda and Komatani, 2017; Takeda et al., 2018). However, this approach is only applicable to classification problems, which are not suitable for multi-source localization, which is our target.

This chapter, partially based on (He et al., 2019), discusses domain adaptation approaches for multi-speaker *Direction-of-Arrival* (DOA) estimation neural networks under three different scenarios: *supervised*, *weakly-supervised* and *unsupervised*, which are organized as follows:

- In Section 4.1, we present the supervised domain adaptation approach, which is simply combining the loss terms of the two domains.
- In Section 4.2, we introduce a novel weakly-supervised domain adaptation framework with the number of sound sources as the weak label. This framework includes the minimum distance adaptation criterion and pseudo-labeling relying on augmented data.
- Finally, we investigate the application of domain adversarial training for unsupervised domain adaptation for SSL in Section 4.3.

These approaches are applied to the ResNet-STFT model (Section 3.4.4), and evaluated with audio data collected from two different versions of Pepper (Section 4.4). We will show that the weakly-supervised approach achieves comparable performance as the supervised adaptation approach. This suggests a practical and effective framework for applying deep learning based approaches in real situations.

4.1 Supervised Adaptation

We first consider the supervised domain adaptation in which we are given a set of labeled simulated data

$$D_s \subset X \times Y,$$

together with a set of labeled real audio data

$$D_t \subset X \times Y,$$

where X and Y are respectively the *input space* and *label space* that we have defined in Section 3.1. To apply supervised domain adaptation, we first use the simulated data to pre-train a model, which is the initialization for subsequent optimization processes. Then, we train a model that minimizes the loss on both the source domain and the target domain:

$$\theta^* = \arg\min_{\theta} \mu_t \mathop{\mathbf{E}}_{(x,y)\in D_t} \mathcal{L}\left(f_\theta(x), y\right) + \mu_s \mathop{\mathbf{E}}_{(x,y)\in D_s} \mathcal{L}\left(f_\theta(x), y\right), \tag{4.1}$$

where $\mathcal{L}$ is the *Mean Squared Error* (MSE) loss (Eq. (3.2)), and μ_t and μ_s are the weighting parameters for the two domains. In practice, the weighting is implemented by changing the proportion of source and target domain samples in each mini-batch. The added loss term relying on the simulated data can reduce the bias caused by data insufficiency of the real data.

4.2 Weakly-Supervised Adaptation

Although with supervised domain adaptation we can reduce the requirement for a large number of real samples, annotation of these real samples still demands a heavy workload. Therefore, we propose a weakly-supervised adaptation framework to further reduce the effort for data collection. According to this framework (Fig. 4.1), we first generate a large number of simulated data from clean speech and background noise. These data serve as the source domain data and are used to pre-train the ResNet-STFT model (Section 3.4.4). Then, we collect a relatively small set of real audio data in which only the number of sound sources is manually labeled (as weak labels), hence the high cost of exact location annotation is avoided. These real audio data are also augmented by mixing single-source frames. Lastly, we adapt the pre-trained model to the real condition using a combination of the simulated data, the weakly-labeled real data and the augmented data in a weakly-supervised fashion.

The adaptation process involves two adaptation schemes: *minimum distance criterion* relying on the real data and *pseudo-labeling* relying on the augmented data, which we will introduce respectively in the following sections.

4.2.1 Minimum Distance Criterion

In the weakly-supervised adaptation setting, instead of fully-labeled data D_t, we are given a set of weakly-labeled real data:

$$D_w = \{(x_i, z_i)\}_{i=1}^{N_w} \subset X \times Z,$$

accompanied by a set of fully-labeled simulated (source domain) data D_s. Each value z_i from the weak label domain $Z = \{0, 1, 2, \ldots\}$ indicates the number of sources in the input frame x_i.

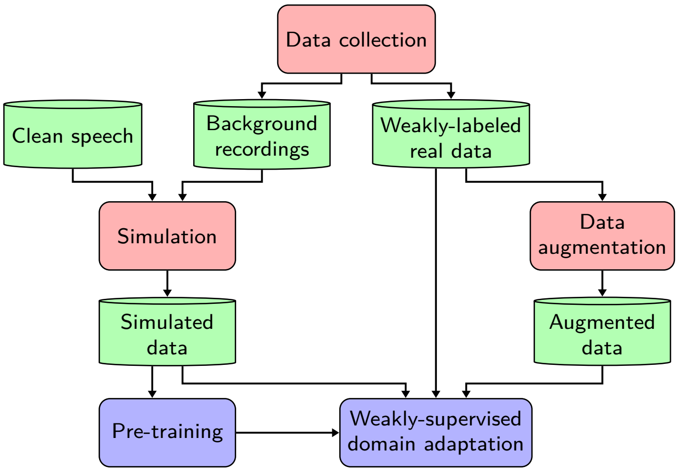


Figure 4.1 – Overview of our framework for neural network-based multi-speaker DOA estimation with weakly-supervised domain adaptation. The arrows indicate which datasets (green) are required or generated by data preparation procedures (red), and which datasets are used for the training processes (blue).

We apply adaptation by minimizing a weak supervision loss $\mathcal{L}_w$ on the target domain as well as the supervised loss (Eq. (3.2)) on the source domain:

$$\theta^* = \arg\min_{\theta} \ \mu_w \mathop{\mathbf{E}}_{(x,z)\in D_w} \mathcal{L}_w\left(f_\theta(x), z\right) + \mu_s \mathop{\mathbf{E}}_{(x,y)\in D_s} \mathcal{L}\left(f_\theta(x), y\right), \tag{4.2}$$

where μ_w and μ_s are weighting parameters. The minimum distance criterion defines the weak supervision loss as the minimum distance in the output space between the network output and all possible labels that satisfy the weak label:

$$\mathcal{L}_w(f_\theta(x), z) = \min_{y\in r(z)} \|f_\theta(x) - h(y)\|_2^2, \tag{4.3}$$

where $h(\cdot)$ is the output encoding defined by Eq. (3.4), and $r(z)$ is the set of all sound DOA labels that satisfy the weak label z, i.e. the number of sources in y is z:

$$r(z) = \{y \in Y : |y| = z\}.$$

The objective function and weakly-supervised loss are based on the assumptions that a good model should:

- predict well the DOAs in the source domain (supervised loss);
- output curves that are close to ideal spatial spectra (weakly-supervised loss);
- make correct detection of the number of sources in the target domain (information from weak labels).

The weak supervision can also be viewed as a pseudo-labeling approach, because the loss function L_w can be rewritten as:

$$\begin{aligned}\mathcal{L}_w\left(f_\theta(x), z\right) &= \left\| f_\theta(x) - h\left(\operatorname*{arg\,min}_{y \in r(z)} \|f_\theta(x) - h(y)\|_2^2\right)\right\|_2^2 \\ &= \mathcal{L}\left(f_\theta(x), \operatorname*{arg\,min}_{y \in r(z)} \|f_\theta(x) - h(y)\|_2^2\right) \\ &= \mathcal{L}\left(f_\theta(x), p_\theta(x, z)\right),\end{aligned} \tag{4.4}$$

with

$$p_\theta(x, z) = \operatorname*{arg\,min}_{y \in r(z)} \|f_\theta(x) - h(y)\|_2^2 \tag{4.5}$$

as the pseudo-labeling function. We can see that the weak supervision loss function is equivalent to the supervised loss using $p_\theta(x, z)$ as the ground-truth label.

Furthermore, we can visualize the pseudo-labels in the output space to see how the minimum distance criterion works (Fig. 4.2). When the number of sources is zero, the network is supervised to output zeros, thus reducing the false positives caused by unseen noise (Fig. 4.2a). When the number of sources is one or more, the network is supervised to give more certain prediction on the most prominent peaks, thus increasing the recall (Fig. 4.2c). At the same time, the other peaks that are caused by unseen conditions are suppressed (Fig.4.2b, c). However, the effectiveness of the weakly-supervised adaptation depends on the initial performance of the network model. If the network initial output is too far away from the ground truth, the weak supervision will lead to incorrect pseudo-labels (Fig.4.2d, e).

4.2.2 Pseudo-labeling with Data Augmentation

In practice, we observe that the network trained on simulated data initially performs worse on the multi-source audio segments as illustrated in Fig. 4.2e. Thus, in order to increase the correctness of the pseudo-labeling on multi-source audio frames, we augment the real data by generating mixture data with known single-source components, and extend the weak supervision method using a modified pseudo-labeling approach. The idea is that we apply the pseudo-labeling to the easier single-source components rather than to the multi-source mixtures, so that we can obtain more effective weak supervision.

Data augmentation. The augmented mixture dataset D_a consists of a set of mixture x_i and their single-source components $\mathbf{u}_i = \{u_{ij}\}_{j=1}^{z_i}$:

$$D_a = \{(x_i, \mathbf{u}_i)\}_{i=1}^{N_a} \subset X \times 2^X.$$

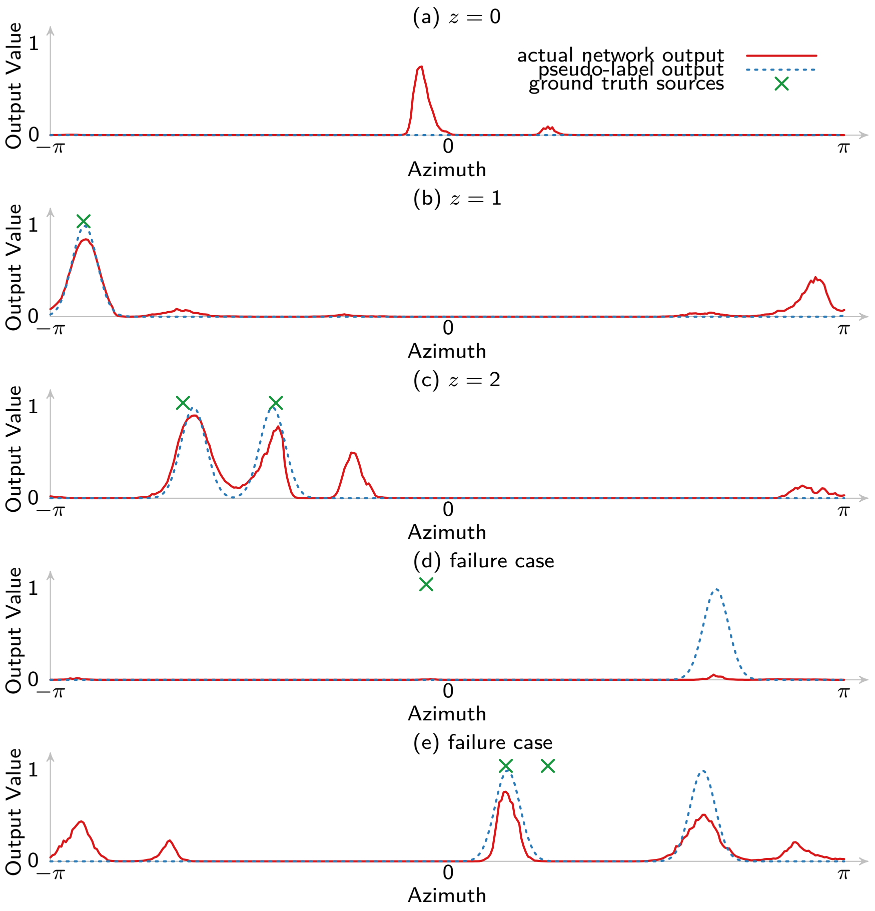


Figure 4.2 – Examples of weak supervision with a known number of sources on real audio segments. The ground truth locations are shown but are not used for weak supervision.

Here, the mixtures are generated by linear combination:

$$x_i = \sum_{j=1}^{z_i} \alpha_{ij} u_{ij}, \tag{4.6}$$

where $\{u_{ij}\}$ are single-source segments randomly sampled from the weakly-labeled dataset D_w, z_i is the number of components (sources), and $\{\alpha_{ij}\}$ are random scaling factors. Since all the real recordings include background noise, we scale each single source frame in a way that the power of the background noise is constant in the mixture[1], that is:

$$\sum_{j=1}^{z_i} \alpha_{ij}^2 = 1. \tag{4.7}$$

A benefit of such data augmentation is that it increases the number of realistic multi-source segments, which is difficult to obtain by recording. In addition, as the combinations

[1] We assume the background noise segments are pairwise independent.

of sound directions increases exponentially with the number of sources, we need a large number of multi-source training samples to cover such variabilities.

Pseudo-labeling on components. The other benefit is that the knowledge of the single-source components allows us to apply reliable pseudo-labeling on this dataset: we first apply pseudo-labeling (Eq. (4.5)) to its single-source components, that are $p_\theta(u_{ij}, 1), j = 1 \ldots z_i$ (Fig. 4.3a,b). Then, we use the union of these pseudo-labels for the multi-source frame (Fig. 4.3c). Thus, the loss function of the modified adaptation is:

$$\mathcal{L}_a(f_\theta(x_i), \mathbf{u}_i) = \mathcal{L}\left(f_\theta(x_i), \cup_{j=1}^{z_i} p_\theta(u_{ij}, 1)\right), \tag{4.8}$$

and the optimization target becomes:

$$\theta^* = \arg\min_\theta \mu_a \mathop{\mathbf{E}}_{(x,\mathbf{u})\in D_a} \mathcal{L}_a\left(f_\theta(x), \mathbf{u}\right) + \mu_w \mathop{\mathbf{E}}_{(x,z)\in D_w} \mathcal{L}_w\left(f_\theta(x), z\right) + \mu_s \mathop{\mathbf{E}}_{(x,y)\in D_s} \mathcal{L}\left(f_\theta(x), y\right), \tag{4.9}$$

where μ_a controls the weight of the modified weak-supervision loss on the augmented dataset.

4.3 Domain-Adversarial Training

We apply *domain adversarial training* (Ganin et al., 2016) for unsupervised adaptation. Under this setting, we are given labeled simulated data D_s and unlabeled real data

$$D_u = \{x_i\}_{i=1}^{N_u} \subset X.$$

The goal of domain adversarial training is to learn a domain-invariant feature representation, so that a model can estimate DOA in the target domain as equally well as in the source domain. To achieve this, we first separate the ResNet-STFT network into two parts: a *feature extractor* $g_f(\cdot; \theta_f)$, and a *DOA Estimator* $g_y(\cdot; \theta_y)$. Then, we add a *domain classifier* $g_d(\cdot; \theta_d)$ subnet after the feature extractor (Fig. 4.4). The output of the domain classifier is a scalar between 0 and 1, indicating the posterior probability of the input being from the target domain. We train these three parts in a way that the output of the feature extractor is suited for DOA estimation and indistinguishable by the domain classifier.

Specifically, depending on the part of the network, the following objective function is either minimized or maximized during adaptation:

$$E(\theta_f, \theta_y, \theta_d) = \mathop{\mathbf{E}}_{(x,y)\in D_s} \mathcal{L}\left(f(x), y\right) - \mu_d \left(\mathop{\mathbf{E}}_{x\in D_s} \mathcal{L}_d\left(f_d(x), 0\right) + \mathop{\mathbf{E}}_{x\in D_u} \mathcal{L}_d\left(f_d(x), 1\right)\right), \tag{4.10}$$

where $f(x) = g_y(g_f(x; \theta_f); \theta_y)$ is the output of the DOA estimator, $f_d(x) = g_d(g_f(x; \theta_f); \theta_d)$ is the output of the domain classifier, $\mathcal{L}$ is the supervised loss (Eq. (3.2)), μ_d is the

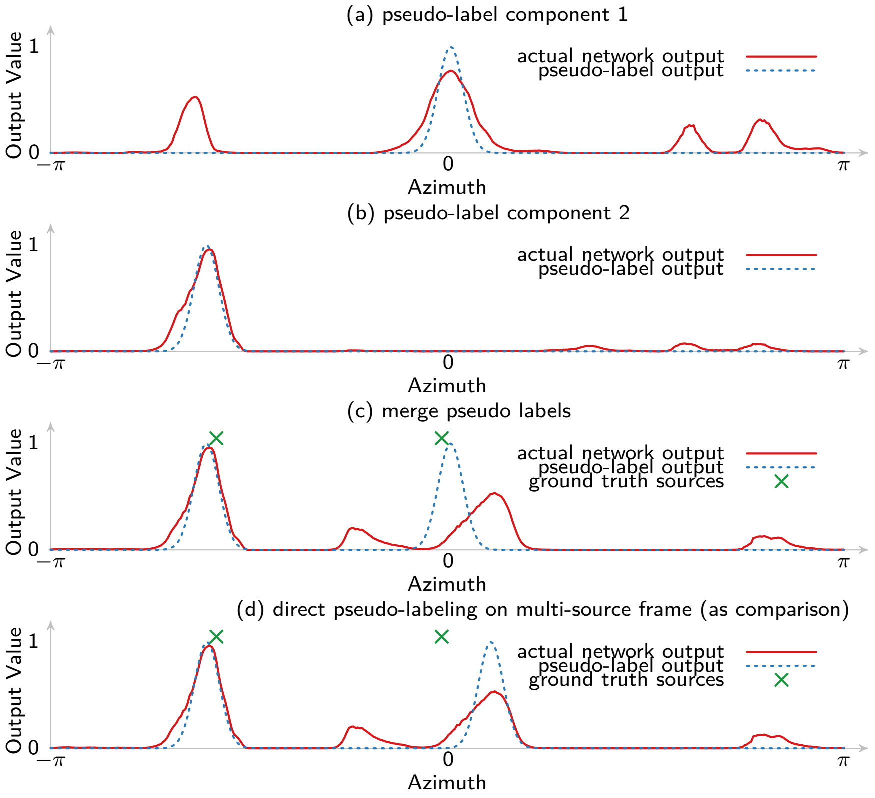


Figure 4.3 – Example of weak supervision by pseudo-labeling on mixture components. (a, b) The pseudo-labeling is applied first on the single-source components. (c) Then, the pseudo-label of the two-source mixture is obtained by merging the pseudo-labels of its components. This approach is more reliable than directly applying the pseudo-labeling to the mixture as shown in (d).

weight for domain classification loss, and $\mathcal{L}_d$ is the binomial cross entropy loss for domain classification:

$$\mathcal{L}_d(f_d(x), v) = -v\log f_d(x) - (1-v)\log(1 - f_d(x)), \tag{4.11}$$

where v indicates the domain (0: source domain, 1: target domain).

The target of adaptation is to make the feature extractor and the DOA estimator minimize the objective function (Eq. (4.10)), while the domain classifier maximizes that function. Specifically, it seeks to find the saddle point such that:

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg\min_{\theta_f, \theta_y} E(\theta_f, \theta_y, \hat{\theta}_d), \tag{4.12}$$

$$\hat{\theta}_d = \arg\max_{\theta_d} E(\hat{\theta}_f, \hat{\theta}_y, \theta_d). \tag{4.13}$$

Based on Eq. (4.12), the feature extractor parameters are modified to decrease the

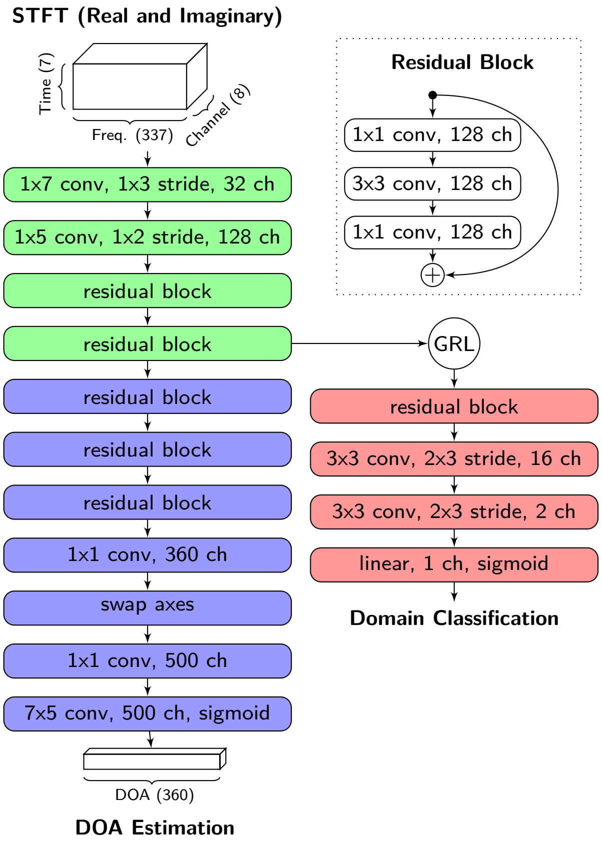


Figure 4.4 – Network architecture for domain adversarial training. It includes a feature extractor g_f (green), a DOA estimator g_y (blue), and a domain classifier g_d (red). GRL is a *Gradient Reversal Layer* (Ganin et al., 2016). The *Batch Normalization* (BN) and *Rectified Linear Unit* (ReLU) after each hidden layer are omitted in this graph.

DOA estimation loss and increase the domain classification loss. In contrast, based on Eq. (4.13), the domain classifier parameters are modified in the adversarial direction, that is decreasing the domain classification loss. Domain-invariant features are expected to be obtained on this saddle point.

Combination with weakly-supervised adaptation. Domain adversarial training can be combined with weakly-supervised adaptation, by adding a weakly-supervised loss term to the objective function:

$$\begin{aligned} E(\theta_f, \theta_y, \theta_d) = & \underset{(x,y)\in D_s}{\mathbf{E}} \mathcal{L}\left(f(x), y\right) + \underset{(x,z)\in D_w}{\mathbf{E}} \mathcal{L}_w\left(f(x), z\right) \\ & - \mu_d \left(\underset{x\in D_s}{\mathbf{E}} \mathcal{L}_d\left(f_d(x), 0\right) + \underset{x\in D_w}{\mathbf{E}} \mathcal{L}_d\left(f_d(x), 1\right) \right), \end{aligned} \tag{4.14}$$

where $\mathcal{L}_w$ is the weak supervision loss (Eq. (4.3) in Section 4.2.1). The target of adaptation remains the same as finding the saddle point which satisfies Eqs. (4.12) and (4.13).

4.4 Experiments

We applied the proposed approaches with simulated data and real data from two robots, and we verified its effectiveness in two ways — by analyzing the correctness of pseudo-labeling and evaluating the performance of DOA estimation.

4.4.1 Robots and Data

Robots. We used two versions of Pepper in our experiments. In addition to the one we have described in Section 3.5.1, we added data from a second Pepper that has omni-directional microphones, which are different from the directional microphones on the first robot. We use `P1` and `P2` to denote the two versions respectively.

Source-domain (simulated) data. We generated the source domain data by convolving clean speech audio with simulated room impulse responses (Table 4.1). The room impulse responses are simulated with the RIR-Generator[2] (Habets, 2006). The clean audio speech data were the close talking recordings randomly selected from the AMI corpus (McCowan et al., 2005) (Table 3.2). We first generated spatialized audio of speech signals from random locations in random rooms. The microphone array geometry was set according to that on the robot. We tried to simulate both omni-directional and cardioid directivity patterns of the microphones, and found out that the models trained with omni-directional simulation have in general better performance, even for the robot `P1`, whose actual microphones are directional. We hypothesize that this is because the simulation cannot replicate exactly the directivity patterns of the real microphones, and models trained with omni-directional simulation are easier to adapt to other situations. Therefore, we used the omni-directional simulation for both robots throughout our experiments. Then the single-source simulated audio frames are mixed randomly at runtime with other frames as well as the real robot background recordings.

In total, we generated one million mixture frames (47 hours). The number of sources varies from zero to four. This include a significant number of source locations and audio content for training. We experimented with both anechoic and reverberant room conditions. For the reverberant simulation, the *Reverberation Time* (RT60) is randomly selected between 200 and 800 ms.

Target-domain data. We collected real data with the robots. For `P1`, we used the same data that were used in the previous chapter (Table 3.1). For `P2`, we conducted the data collection in the same way and obtained a set of loudspeaker data (Table 4.2). As we have mentioned, the sound source locations are fixed during each piece of recording, therefore the coverage in terms of varoius sound source locations in the real recordings is considerably less than that of the simulated data.

[2] https://github.com/ehabets/RIR-Generator

Table 4.1 – Specifications of the simulated data.

	Simulation `P1` & `P2`
Total duration	47 hours
# of frames	1 million
- no source	200k
- single source	400k
- two sources	300k
- more	100k
# of male speakers	105
# of female speakers	43
Azimuth (°)	[-180, 180.0]
Elevation (°)	[-74, 75]
Distance (m)	[0.5, 10.8]
Room length (m)	[8.0, 12.0]
Room width (m)	[6.0, 9.0]
Room height (m)	[2.0, 5.0]
RT60 (ms)*	[200, 800]

*RT60 values only apply to the reverberant simulation.

Table 4.2 – Specifications of the target-domain data for `P2`

	`P2`-Loudspeaker	
	Training	Evaluation
Total duration	3.8 hours	2.1 hours
# of frames	122k	67k
- no source	26k	14k
- single source	90k	46k
- two sources	6k	7k
# of male speakers	101	8
# of female speakers	41	8
Azimuth (°)	[-180, 180]	[-178, 180]
Elevation (°)	[-39, 56]	[-29, 48]
Distance (m)	[0.5, 1.8]	[0.9, 2.0]

These data were annotated with exact sound source locations, from which we derived the weak labels (number of sound sources in each frame). Although only the weak labels are required for the weakly-supervised adaptation approach, the availability of the fully-labeled training data allowed us to analyze quantitatively the effectiveness of the weak supervision, and compare it to the fully-supervised methods.

4.4.2 Training Parameters

According to the proposed framework (Fig. 4.1), we first pre-trained a model on the fully-labeled simulated data using the two-stage training scheme. The model was trained for one epoch in the first stage (Eq. (3.13)) and four epochs in the second stage (Eq. (3.2)).

For the weakly-supervised domain adaptation step, the pre-trained model was used as initialization, and the weights of the components in the optimization target function Eq. (4.9) where $\mu_w = 0.9$ (for the weak supervision loss), $\mu_a = 0.1$ (for the modified weak supervision loss on augmented data), and $\mu_s = 1.0$ (for supervised loss on simulated data). This is equivalent to composing mini-batches using 45%, 5% and 50% of the samples from the weakly-labeled dataset, augmented dataset, and the simulated dataset, respectively. Models were adapted for ten epochs for `P1` and 40 epochs for `P2`. We used a learning rate of 0.001 and reduced it by half every two (`P1`) or eight (`P2`) epochs. During all the training processes, the models were optimized with the Adam optimizer (Kingma and Ba, 2015) and a mini-batch size of 100.

For domain adversarial training, the network was adapted for ten epochs. Domain classification weight μ_d varied from 0 to 10^{-3} during the adaptation process:

$$\mu_d(p) = \left(\frac{2}{1 + \exp(-10p)} - 1\right) \times 10^{-3}, \tag{4.15}$$

where $p \in [0, 1]$ is the adaptation progress. In each mini-batch, an equal number of source domain data and target domain data were sampled. The optimizer and mini-batch size were the same as those used for the weakly-supervised adaptation scheme.

4.4.3 Analysis of the Pseudo-Labeling

To better understand the minimum distance criterion, we analyzed how the effectiveness of the pseudo-labeling depends on the initial model performance and the number of overlapping sources. Our expectation is that good pseudo-labels will have a positive impact on the adaptation process if the target spatial spectra encoded from the pseudo-labels are on average closer to the ground truth spatial spectra than the actual network outputs. Therefore, we computed the loss reduction between the MSE loss (Eq. (3.2)) of

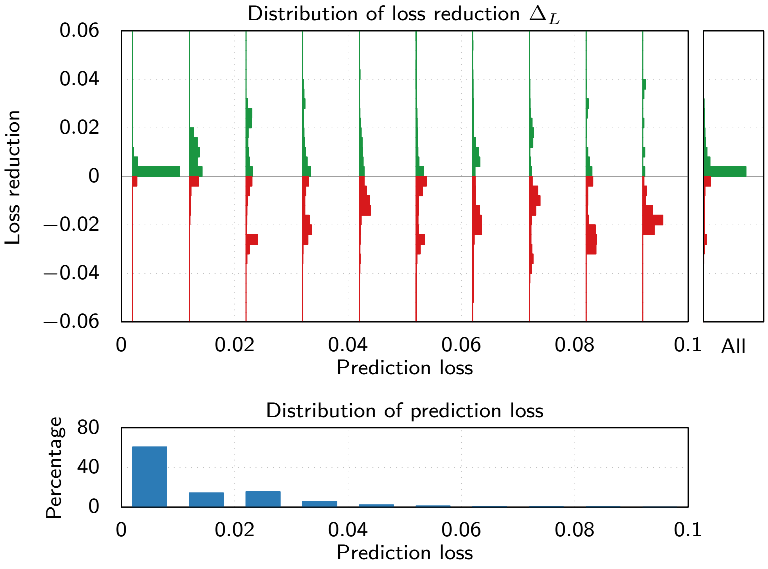


Figure 4.5 – The distribution of the pseudo-labeling loss reduction on all samples in the `P1` training data. *Top figure*: each histogram (plotted vertically) shows a distribution of the loss reduction (Eq. (4.16)) on the samples with the indicated prediction loss (left histograms) and on all samples (right-most histogram). The green bars indicate positive reduction (correct weak supervision), while the red bars indicate negative reduction (incorrect weak supervision). *Bottom figure*: the distribution of the initial prediction loss. The network is pre-trained with the anechoic simulation data.

the model prediction and that of the pseudo-label:

$$\Delta_L = \mathcal{L}\left(f_\theta(x), y\right) - \mathcal{L}\left(h(p_\theta(x,z)), y\right), \tag{4.16}$$

where y and z are, respectively, the location label and the weak label corresponding to the audio segment x. A positive loss reduction indicates that the pseudo-labeling should be beneficial for the model.

We extracted the pseudo-labels on the target-domain data using the minimum distance criterion (Eq. (4.3)) and the pre-trained model. Then, we computed the distributions of the loss reduction (Eq. (4.16)) on samples with different prediction accuracy, which is characterized here by the loss values of the predictions. The result (Fig. 4.5) shows that weak supervision is mostly correct when the prediction loss is small (below 0.02), and becomes unreliable as the prediction loss increases.

By comparing the loss reduction distributions on the single-source (Fig. 4.6) and the multi-source samples (Fig. 4.7), we can verify the assumption that weak supervision is more reliable on single-source frames, since the pre-trained model initially performs better on the single-source frames.

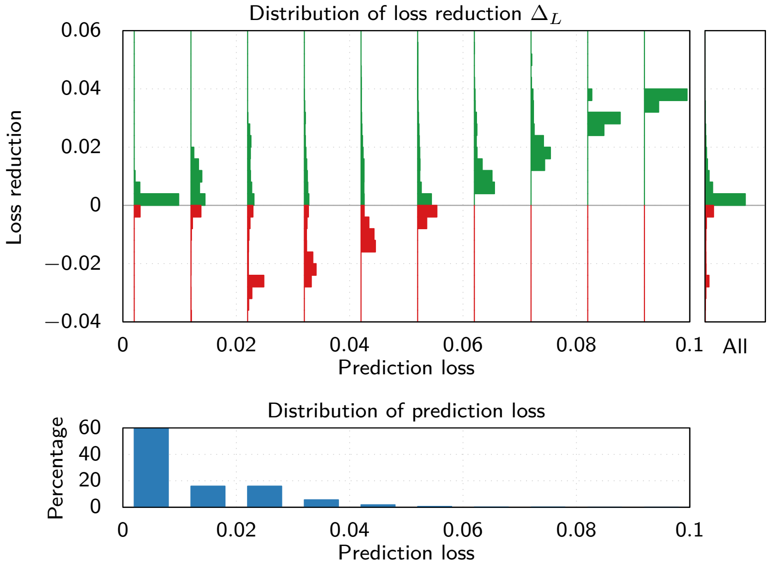


Figure 4.6 – The distribution of the pseudo-labeling loss reduction on the single-source samples from the `P1` training data.

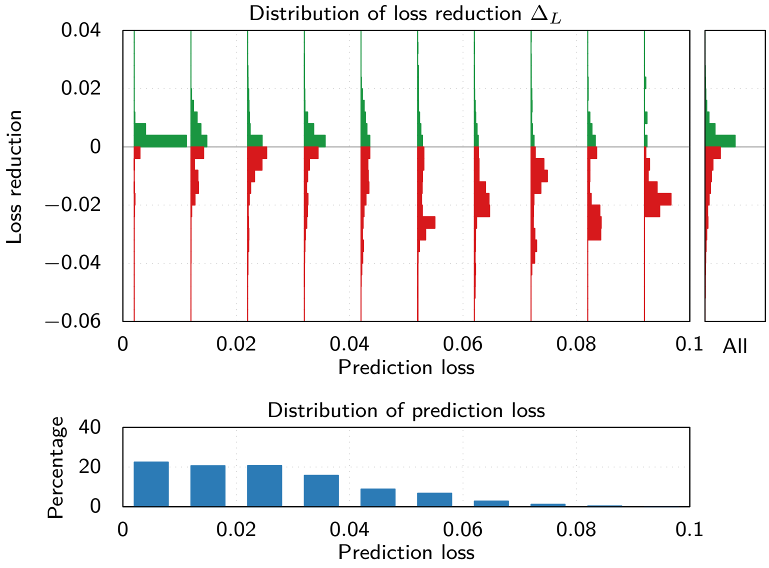


Figure 4.7 – The distribution of the pseudo-labeling loss reduction on the multi-source samples from the `P1` training data.

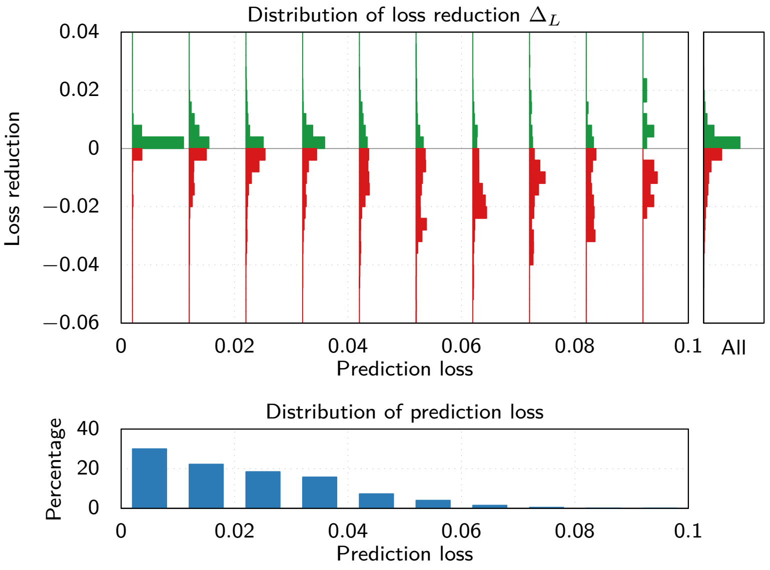


Figure 4.8 – The distribution of the pseudo-labeling loss reduction on the P1 augmented data using the minimum distance criteria.

We also computed the loss reduction distributions of two different strategies for pseudo-labeling on the multi-source augmented data. The first strategy is to directly apply the minimum distance criterion (Eq. (4.3)) on the mixed signal (Fig. 4.8), while the other strategy is, as we proposed, to extract pseudo-labels from the mixture components and then merge the pseudo-labels using Eq. (4.8) as the loss function (Fig. 4.9). Since the modified adaptation relies on pseudo-labels of the single-source components, it generates more reliable results than the direct application of pseudo-labeling on the multi-source frames. Even when the initial prediction loss is larger than 0.02, the pseudo-labels are more likely to have a positive reduction.

4.4.4 List of Methods

The following approaches were included for comparison:

- **SRP-PHAT**: steered response power with phase transform (DiBiase et al., 2001).
- **SUP.REAL**: the fully-supervised approach described in Chapter 3 using only fully-labeled real data for training (two-stage training with loss functions Eqs. (3.2) and (3.13)).
- **SUP.SIM**: a model trained with only the simulated data. This is also the initialization for all the following adaptation approaches.
- **SDA**: the supervised domain adaptation approach. Its objective function is Eq. (4.1).
- **WDA.MD**: the weakly-supervised domain adaptation with minimum distance crite-

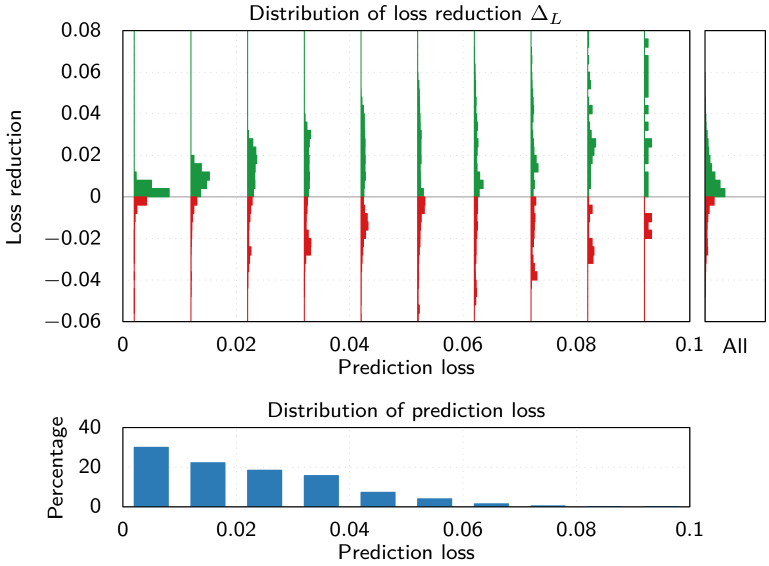


Figure 4.9 – The distribution of the pseudo-labeling loss reduction on the `P1` augmented data, using the modified pseudo-labeling that takes advantage of the knowledge about the mixing components.

rion. Its objective function is Eq. (4.2).

- **WDA.AUG**: the weakly-supervised domain adaptation approach with both minimum distance criterion and pseudo-labeling on augmented data. Its objective function is Eq. (4.9).
- **UDA.DAT**: the unsupervised domain adaptation approach with domain adversarial training. Its objective functions are Eqs. (4.10), (4.12) and (4.13).
- **WDA.DAT**: the combined approach of domain adversarial training and minimum distance weakly-supervised adaptation. Its objective functions are Eqs. (4.12) to (4.14).

We experimented with both *anechoic* and *reverberant* simulation for generating source domain data, therefore two models were obtained for each of the above methods (except for SRP-PHAT and SUP.REAL).

4.4.5 DOA Estimation Results

We applied these approaches to both the robots `P1` and `P2`, and evaluated them on their respective test sets (real data). We report their performance on single-source and two-source frames, as well as the overall performance on all test frames. The evaluation criteria are the same as those in the previous chapter (Section 3.6.2).

Table 4.3 – *Mean Absolute Error* (MAE) and *Accuracy* (ACC) on the P1-Loudspeaker dataset. Performance is evaluated on different subsets: all frames, single-source frames and two-source frames. The source-domain data are simulated with two different room conditions (anechoic and reverberant).

Dataset	P1-Loudspeaker					
Subset	All		$z = 1$		$z = 2$	
	MAE (°)	ACC (%)	MAE (°)	ACC (%)	MAE (°)	ACC (%)
SRP-PHAT	21.5	78	19.0	82	37.0	50
SUP.REAL	3.0	94	2.6	96	5.6	84
Anechoic Simulation						
SUP.SIM	13.1	80	11.6	82	22.6	66
SDA	3.3	94	2.7	95	7.1	86
WDA.MD	7.5	87	4.3	91	26.6	62
WDA.AUG	4.5	93	3.3	95	12.2	83
UDA.DAT	12.9	81	11.3	83	22.6	69
WDA.DAT	8.1	87	4.1	92	32.3	58
Reverberant Simulation						
SUP.SIM	11.7	86	10.0	88	22.7	69
SDA	3.8	94	3.1	95	7.9	84
WDA.MD	8.8	89	4.9	93	32.9	63
WDA.AUG	5.2	92	3.8	94	14.2	79
UDA.DAT	9.6	89	8.2	91	18.2	76
WDA.DAT	10.0	89	5.0	94	41.0	58

Learning-based vs SRP-PHAT. From the performance of the approaches on the P1-Loudspeaker data (Table 4.3 and Fig. 4.10), we see that all learning-based approaches outperform SRP-PHAT. Because there is a strong background noise in the robot audio data, SRP-PHAT, which assumes the target signal is dominant across all frequencies, is more affected. The learning-based approaches, on the other hand, can learn from the samples to implicitly suppress the noise.

Simulation vs Real Data. Comparing the models trained with simulated data (SUP.SIM) to those trained with real data (SUP.REAL), we see the expected performance degradation caused by the discrepancy between the acoustic simulation and real recordings.

Supervised Adaptation. The model first pre-trained with simulated data and then adapted with fully-labeled real data (SDA) achieves similar performance as that directly trained on real data (SUP.REAL) in the P1-Loudspeaker test set. Nevertheless, a noticeable difference (see Fig. 4.10c) is that the adapted model has better precision and recall in the two-source frames. This is probably because the simulated data provide a

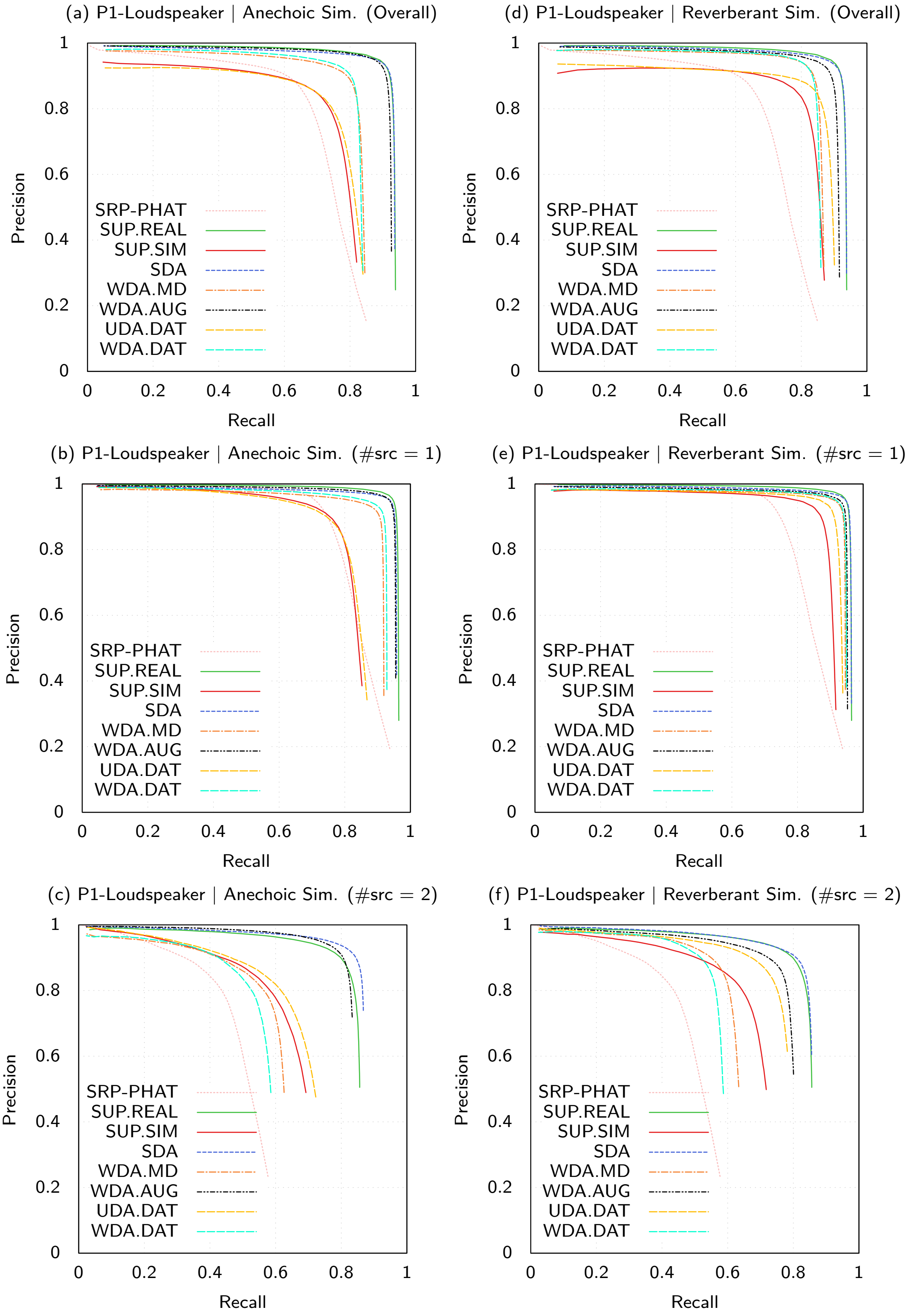


Figure 4.10 – Precision-recall curves as a sound source detection problem on the P1-Loudspeaker dataset. The curves are generated by varying the prediction threshold ξ in Eq. (3.6). DOA estimation with less than 5° error is considered correct. As indicated in the figure titles, left column shows results of model trained with anechoic simulation, and right column shows results of model trained with reverberant simulation.

broader coverage of sound source directions, especially in the multi-source case, than the real data.

Weakly-supervised Adaptation. Using the weakly-labeled real data, both the weakly-supervised domain adaptation approaches (WDA.MD and WDA.AUG) significantly outperforms the pre-trained model (SUP.SIM). The discrepancy between the simulation and real data is mitigated. Between both the approaches, the performance of the approach relying on augmented data (WDA.AUG) is significantly better, especially on the two-source frames, with an accuracy of 83% (vs 62% for WDA.MD) for instance. In fact, directly applying the minimum distance criterion (WDA.MD) on the multi-source frames is not reliable and generates wrong pseudo-labels. Therefore, its performance on the two-source frames is even worse than the pre-trained model. Applying the adaptation on the single-source components of the augmented data prevents unreliable pseudo-labeling and improves the adaptation result. As a result, our approach achieves comparable results, in terms of accuracy as well as precision and recall (Fig. 4.10(a,d)), as those using fully-labeled real data. This shows that we can substitute exact labels in the real data with weak labels, thus the workload of annotation can be significantly reduced.

Domain-Adversarial Training. The adaptation with domain-adversarial training only shows insignificant improvement (Fig. 4.10). Moreover, combining domain-adversarial training with minimum distance weakly-supervised adaptation does not improve the result with respect to using only weakly-supervised adaptation. We have explored various values of μ_d, architecture of the domain classifier, and layers of features extractor, however further improvements were not obtained. In practice, introducing domain-invariance suffers the risk of reducing the discriminative power of the features, because the feature extractor may produce irrelevant features in order to fool the domain classifier. Overall, finding the balance between domain-invariance and discriminative power is difficult.

Anechoic vs Reverberant Simulation. Comparing the different simulation conditions, we find that the pre-trained models with reverberant simulation in general outperform those with anechoic simulation, as they matches the evaluation data better, which are collected in reverberant environments. However, after domain adaptation, the models with the anechoic simulation achieve better performance in most conditions. This is probably because the models trained with simpler source-domain conditions (anechoic simulation) are more capable to adapt, while the models trained with reverberant simulation might be already locked in a local optimum that favors the specific complex conditions in the simulated training data.

`P1`-Human Data. We can draw similar conclusions from the evaluation on this test set (Table 4.4 and Fig. 4.11), except that the performance of SRP-PHAT is much better than that in the `P1`-Loudspeaker data. This is because in the human talker recordings all speakers are from the front (inside the field of view of the robot camera), and the SNR is higher. SPR-PHAT is less affected by the background noise. It achieves better

Table 4.4 – MAE and ACC on the P1-Human dataset. The source-domain data are simulated with anechoic condition.

Dataset	P1-Human					
Subset	All		$z = 1$		$z = 2$	
	MAE (°)	ACC (%)	MAE (°)	ACC (%)	MAE (°)	ACC (%)
SRP-PHAT	5.4	88	2.6	93	20.9	56
SUP.REAL	2.9	97	2.4	98	5.6	94
Anechoic Simulation						
SUP.SIM	4.9	92	3.9	94	10.4	83
SDA	3.9	97	3.2	98	8.0	92
WDA.MD	6.6	93	2.6	96	28.8	72
WDA.AUG	4.4	96	3.2	97	11.2	91
UDA.DAT	4.3	94	3.6	95	8.2	87
WDA.DAT	7.0	91	2.4	97	32.9	62

MAE than some learning-based approaches. However, in terms of detection ACC, the learning-based approaches are better, as their ideal spatial-spectrum is normalized to one (not dependent on the SNR or signal power), thus a certain prediction threshold may work uniformly well for all samples.

P2-Loudspeaker Data. We notice that this dataset is in general more challenging than the P1-Loudspeaker data, as indicated by the accuracy as well as precision and recall of SRP-PHAT, SUP.REAL and SUP.SIM in the results (Table 4.5 and Fig. 4.12). The proposed approach relies on initial performance of the pre-trained model, therefore it does not perform as well as that in the P1 data. In spite of this, the proposed approach (WDA.AUG) shows a significant improvement over the pre-trained model. We also find that the model trained with both simulated and real data (SDA) outperforms significantly the models using only real data. This is because there are less real training data for P2 (compared to P1), and adding the simulated data may help especially when the real training data are not sufficient.

4.4.6 Scalability with Data Size

We analyzed the scalability of the different approaches. Specifically, we examined on P1-Loudspeaker (Fig. 4.13) and P1-Human (Fig. 4.14) how their F1-scores evolve with the size of the target-domain training data. The F1-scores are computed with the precision and recall values that generate the best F1-scores. Both figures show that the performance of all approaches generally increases as more real data are used (except the pre-trained model, which does not use real data). The domain adaptation approaches, including weakly-supervised, outperform the supervised approach when the data size is

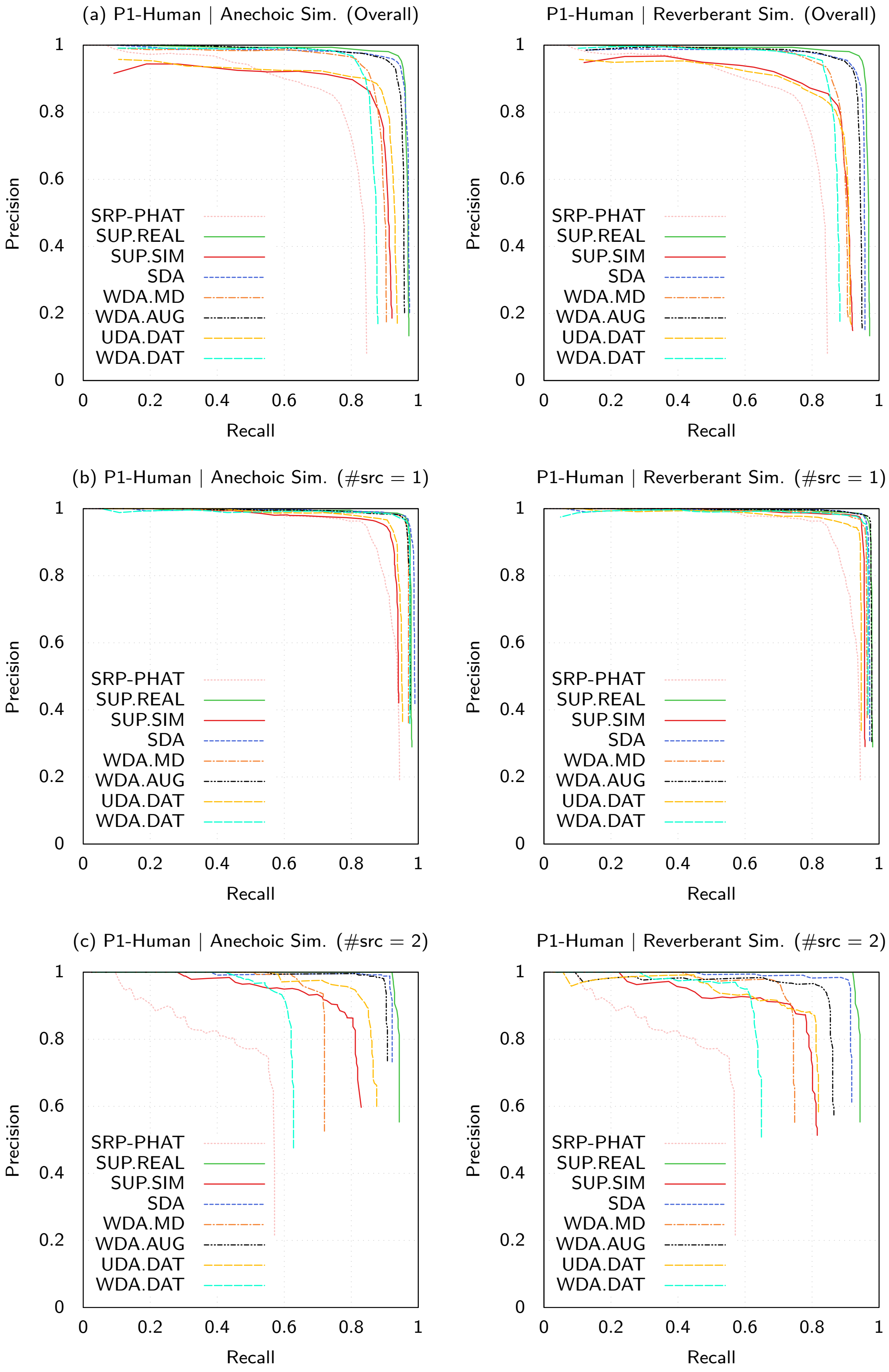


Figure 4.11 – Precision-recall curves as a sound source detection problem on the `P1`-Human dataset.

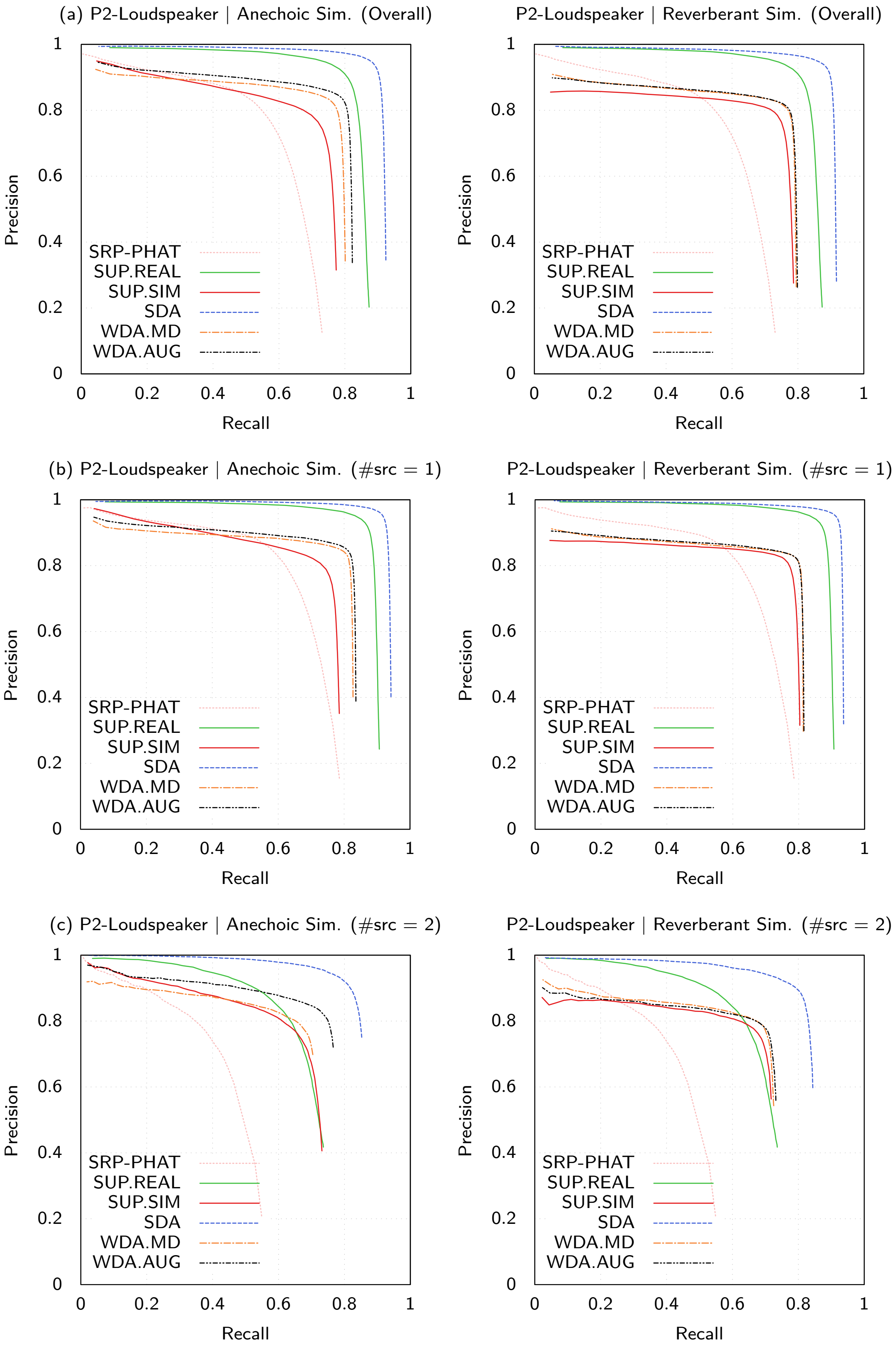


Figure 4.12 – Precision-recall curves as a sound source detection problem on the P2-Loudspeaker dataset.

Table 4.5 – MAE and ACC on the P2-Loudspeaker dataset. The source-domain data are simulated with anechoic condition.

Dataset	P2-Loudspeaker					
Subset	All		$z = 1$		$z = 2$	
	MAE (°)	ACC (%)	MAE (°)	ACC (%)	MAE (°)	ACC (%)
SRP-PHAT	13.5	72	11.1	75	29.3	51
SUP.REAL	5.5	87	4.5	89	12.0	70
Anechoic Simulation						
SUP.SIM	7.2	77	6.5	78	12.3	70
SDA	3.5	92	3.2	94	5.6	85
WDA.MD	5.1	81	4.4	82	9.7	71
WDA.AUG	4.7	82	4.4	83	7.1	77

small.

4.5 Summary

In this chapter, we have introduced a framework to train deep neural networks for multi-source DOA estimation. The framework uses simulated data together with weakly-labeled (number of sources) data under a domain adaptation setting. We have also proposed a data augmentation scheme combining our weakly-supervised adaptation approach with reliable pseudo-labeling of mixture components in the augmented data. This approach prevents incorrect adaptation caused by difficult multi-source samples. The proposed weakly-supervised method (WDA.AUG) achieves almost equal performance to the fully-labeled case on the data of the P1 robot. As for the P2 robot, the proposed method significantly improves the pre-trained model, but does not achieve the same level of performance as the supervised approaches due to the inaccurate initial prediction of the pre-trained model.

Moreover, we have explored the application of domain-adversarial training for unsupervised domain adaptation of multi-source DOA estimation models. Domain-adversarial training aims to construct a domain-invariant feature representation. However, this compromises the discriminant power of the features and no significant adaptation improvement has been observed with this approach.

Overall, the proposed weakly-supervised framework can be used for deploying learning-based DOA approaches to new microphone arrays with minimal effort for data collection.

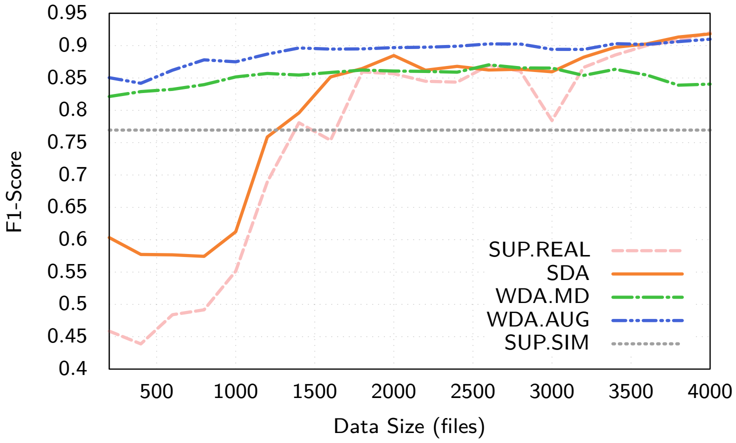


Figure 4.13 – Sound source detection F1-score evaluated on the P1-Loudspeaker test set versus the training data size (number of files). The number of files indicates the variabilities of sound source positions in the dataset, as sound source locations are fixed in each audio file. Source-domain data are simulated with anechoic condition. The pre-trained model (SUP.SIM), which does not use any real data, is presented for reference.

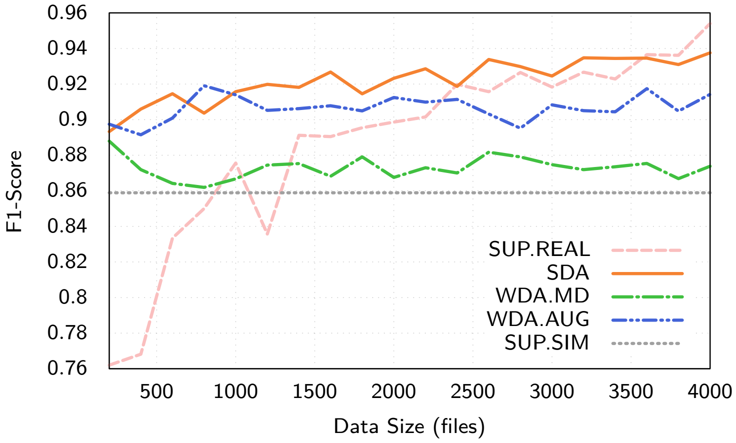


Figure 4.14 – Sound source detection F1-score evaluated on the P1-Human test set versus the training data size (number of files).

5 Joint DOA Estimation and Speech/Non-Speech Classification

The previous chapters have focused on studying *Deep Neural Network* (DNN) for *Direction-of-Arrival* (DOA) estimation of multiple speech sources. However, in real *Human-Robot Interaction* (HRI) environments, there are other types of sounds besides speech, which a robot should distinguish in order to respond accordingly. This chapter, partially based on (He et al., 2018b), aims to develop DNNs for DOA estimation under the following challenging conditions:

- Multiple simultaneous sound sources;
- No a priori knowledge about the number of sound sources;
- Presence of strong robot ego-noise;
- Presence of directional interfering non-speech sources besides speech sources.

As summarized in Section 2.3, previous research on localization and tracking of a specific type of sound sources in the presence of interfering noise sources can be categorized into *sequential approaches* and *joint approaches*. The sequential approaches first obtain signals from individual directions through *Sound Source Separation* (SSS), and then apply classification individually on the separated signals (May et al., 2012; Lim et al., 2015; Crocco et al., 2017; Wakabayashi et al., 2020). The methods to achieve SSS include beamforming (Lim et al., 2015; Crocco et al., 2017), time-frequency masking (May et al., 2012), and spatially-constrained *Blind Source Separation* (BSS) method (Wakabayashi et al., 2020). These methods apply disjoint source separation and classification. Specifically, the classification is either independent or subsequent of the sound source localization and separation.

Joint approaches, in contrast, solve localization and classification simultaneously, allowing knowledge to be shared between tasks. In fact, localization and classification of sources in sound mixtures are closely related. Localization can help classification by providing spatial information, which is useful for better separation or enhancement of source signals. Vice versa, knowing the types of the sources provides spectral information that can help the localization. Such an idea has been explored by applying *Multi-Task Learning* (MTL) neural networks for joint localization and classification of a single sound source (Hirvonen, 2015; Vecchiotti et al., 2018). However, there has been little discussion on joint localization and classification of multiple sound sources.

In this chapter, we present two deep multi-task neural networks for joint DOA estimation and *Speech/Non-Speech* (SNS) classification of multiple sound sources. These networks output two scores per direction: one for sound activity and the other for sound type. By combining these scores, the networks can detect and classifier an arbitrary number of sound sources. The two networks differ in their architectures. The first architecture is based on the idea of sharing features across tasks, while the second architecture explores the idea of merging initial estimation of the two tasks to get refined estimation (re-estimation). Experiments with real noisy recordings show that these two approaches outperform a sequential approach based on *Minimum Variance Distortionless Response* (MVDR) beamformer, and two single-task networks.

5.1 Approach

We describe the multi-task neural networks in term of network input/output, loss function, and network architectures.

5.1.1 Network Input

We adopt the raw *Short-Time Fourier Transform* (STFT) as our network input, as it contains all the required information for both tasks. If we would extract high-level localization features, such as cross correlation, *Inter-channel Phase Difference* (IPD), *Inter-channel Level Difference* (ILD) or subspace-based features, the signal power spectral information is lost. Therefore, such high-level features are not suitable for sound classification. Conversely, high-level sound classification features, such as *Mel-Frequency Cepstral Coefficients* (MFCCs) (Martin et al., 2001; Hughes and Mierle, 2013), do not contain phase information for sound localization. In contrast, the raw signal, which retains all information for both tasks, is preferable. In addition, we have shown in Chapter 3 that DNNs can learn by themselves to extract suitable high-level features from raw STFT. Therefore, we follow the approach of ResNet-STFT in Section 3.3.3 to extract the same input representation for joint DOA estimation and SNS classification.

Specifically, the raw data received by the robot are 4-channel audio signals sampled at 48 kHz. Their STFT is computed in frames of 2048 samples (43 ms) with 50% overlap. Then, a block of 7 consecutive frames (170 ms) are considered an input unit for analysis. The 337 frequency bins between 100 and 8000 Hz are used, and the real and imaginary parts of the STFT coefficients form two individual channels. Therefore, the input representation

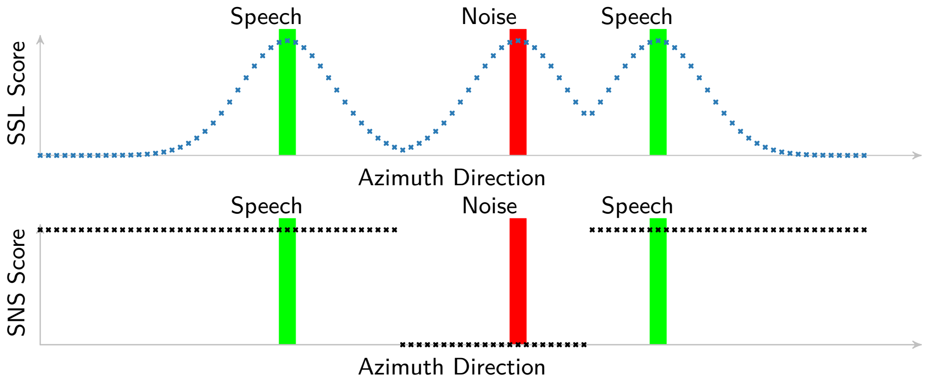


Figure 5.1 – Desired outputs for joint DOA estimation and SNS classification. In this example, there are two speech sources and one noise source indicated by green and red bars, respectively. The desired SSL scores form a spatial spectrum with peaks at the sound source directions. The desired SNS scores are either 1 (speech) or 0 (noise), depending on the type of the nearest sound source.

has a dimension of $7 \times 337 \times 8$ (temporal frames × frequency bins × channels).

5.1.2 Network Output and Loss Function

For each spatial direction, the multi-task networks predict a score of sound activity, $\mathbf{p} = \{p_i\}$ (SSL scores), and a score of sound type, $\mathbf{q} = \{q_i\}$ (SNS scores). The elements p_i and q_i are associated with one of the 360 azimuth directions φ_i.

Encoding. According to the Gaussian-based spatial spectrum coding described in Section 3.2, the desired SSL scores are the maximum of Gaussian functions centered at the DOAs of the ground truth sources (Fig. 5.1):

$$p_i = \begin{cases} \max_{\varphi \in y} \left\{ e^{-d(\varphi_i, \varphi)^2/\sigma^2} \right\} & \text{if } |y| > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{5.1}$$

where $y = y^{(s)} \cup y^{(n)}$ is the union of the ground truth speech source and interfering source DOAs, σ is a parameter controlling the width of the Gaussian curves, $d(\cdot, \cdot)$ denotes the azimuth angular distance, and $|\cdot|$ denotes the cardinality of a set.

The desired SNS scores are either 1 or 0 depending on the type of the nearest source[1] (Fig. 5.1):

$$q_i = \begin{cases} 1 & \text{if the nearest sound source is speech} \\ 0 & \text{otherwise} \end{cases}. \tag{5.2}$$

[1] It is assumed that sound sources are not from the exact same direction.

Loss function. The loss function is defined as the sum of the *Mean Squared Error* (MSE) of both predictions:

$$\text{Loss} = \|\hat{\mathbf{p}} - \mathbf{p}\|_2^2 + \mu \sum_i w_i \left|\hat{q}_i - q_i\right|^2 , \tag{5.3}$$

where $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ are the network outputs, $\mathbf{p}$ and $\mathbf{q}$ are the desired outputs, and μ is a constant. The SNS loss is weighted by $\{w_i\}$, which depends on its distance to the nearest source (w_i differs from p_i only in the parameter for curve width):

$$w_i = \begin{cases} \max_{\varphi \in y} \left\{ e^{-d(\varphi_i,\varphi)^2/\sigma_w^2} \right\} & \text{if } |y| > 0 \\ 0 & \text{otherwise} \end{cases} . \tag{5.4}$$

The weighting emphasizes training around the directions of the active sources, and the SNS scores can be arbitrary at the other directions.

Decoding. During test, the method localizes the sound sources by finding the local maxima in the SSL likelihood that are above a given threshold (this is the same as Eq. (3.6)):

$$\hat{y} = \left\{ \varphi_i : p_i > \xi \quad \text{and} \quad p_i = \max_{d(\varphi_j,\varphi_i)<\sigma_n} p_j \right\} , \tag{5.5}$$

where ξ is the prediction threshold and σ_n is the neighborhood distance for peak finding. Furthermore, to detect speech sources and estimate their DOAs, we combine the SSL and SNS likelihood to further refine the peaks in the SSL likelihood:

$$\hat{y}^{(s)} = \left\{ \varphi_i : p_i q_i > \xi \quad \text{and} \quad p_i = \max_{d(\varphi_j,\varphi_i)<\sigma_n} p_j \right\} . \tag{5.6}$$

We set $\sigma = \sigma_n = 8°$, $\mu = 1$ and $\sigma_w = 16°$ in the experiments.

5.1.3 Network Architectures and Training Procedure

We investigate two multi-task network architectures, one with shared features and one relying on re-estimation. For comparison, we also include a single-task approach with two separate networks for individual tasks.

Multi-Task Network with Shared Features. Inspired by ResNet-STFT, the multi-task network with shared features is a fully convolutional neural network consisting of a residual network common trunk and two task-specific branches (Fig. 5.2). The common trunk starts with the reduction of the size in the frequency dimension using two layers of strided convolution. These initial layers are followed by five residual blocks. The identity connections in the residual blocks allow a deeper network to be trained without being affected by the vanishing gradients problem. We have shown that the ResNet is effective for multi-speaker DOA estimation (Chapter 3). The trunk extracts features that are used for both tasks. The hard parameter sharing in such common trunk provides regularization and reduces the risk of overfitting (Ruder, 2017).

The task-specific branches are identical in structure. They both start with a 1 × 1 convolutional layer with 360 output channels (corresponding to 360 azimuth directions). The layers until this point represent *Stage 1*, in which all the convolutions are along the *Time-Frequency* (TF) domain, therefore the outputs have local receptive fields in the TF domain and can be considered as the initial estimation (of SSL and SNS) for individual TF points. In the rest of the network, *Stage 2*, the convolutions are local in time and DOA dimensions but global in the frequency dimension. In practice, this is achieved by swapping the DOA and the frequency axes. The final output of each branch is a 360-dimension vector indicating the likelihood of SSL and SNS respectively. In addition, *Batch Normalization* (BN) (Ioffe and Szegedy, 2015) and *Rectified Linear Unit* (ReLU) activation functions (Nair and Hinton, 2010) are applied after all convolutional layers except for the output layer.

Following the same training procedure of ResNet-STFT, we train the network with a two-stage training scheme. We first train *Stage 1* for four epochs by imposing supervision to its output. The loss function at this stage is defined as the sum of Eq. (5.3) applied to all the TF points[2]. Such supervision provides a better initialization of the *Stage 1* parameters for further training, as we will show in the experiments.

Then, the whole network is trained in an end-to-end fashion (using the loss function of Eq. (5.3) at the output) for ten epochs. We use the Adam optimizer (Kingma and Ba, 2015) with mini-batches of size 128 for training.

Separated Single-Task Networks. For comparison, we also implement a single-task approach with separated networks (Fig. 5.3). This approach simply uses two identical single-task networks, both of which are copies of ResNet-STFT. Each network outputs the prediction of one task. It can be viewed as separating the trunk in the shared-feature multi-task network. Without the hard feature sharing, these networks can theoretically approximate better the training data, but they are more prone to overfitting compared to the multi-task network. The separated networks are trained in the same way as the shared-feature multi-task network with the two-stage training.

Multi-Task Network with Re-estimation. We design another multi-task network by extending the separated single-task networks with a *re-estimation module* (Fig. 5.4). The idea of the re-estimation is that the independent preliminary TF local predictions for both task are combined to refine the final TF local predictions, so that the predictions are shared explicitly between tasks. The re-estimation module uses the concatenation of the *Stage 1* outputs of both tasks as input, predicts individually the "refinement" for each

[2] We don't use individual ground truth for each TF point, because it is impractical to acquire.

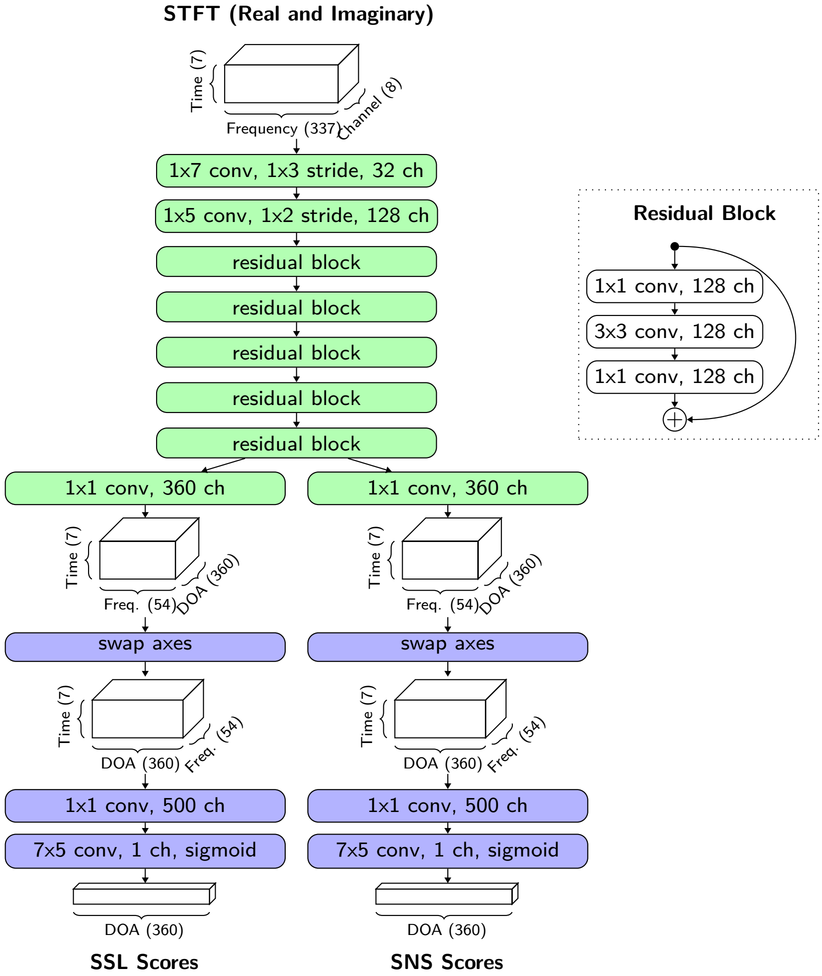


Figure 5.2 – Multi-task network with shared features (MT.SHARED). *Stage 1* consists of the green layers, and *Stage 2* consists of the blue layers. The BN and ReLU activation functions after each hidden layer are omitted in this figure.

task, and adds them element-wisely to the original *Stage 1* outputs. Each "refinement" computation consists of two convolutional layers for downsampling, and two transposed convolutional layers (Dumoulin and Visin, 2018) for upsampling back to the same size of the original *Stage 1* output.

To train the multi-task network with re-estimation, we first initialize it by the separated single-task networks, and then train it with the same end-to-end fashion as the other approaches.

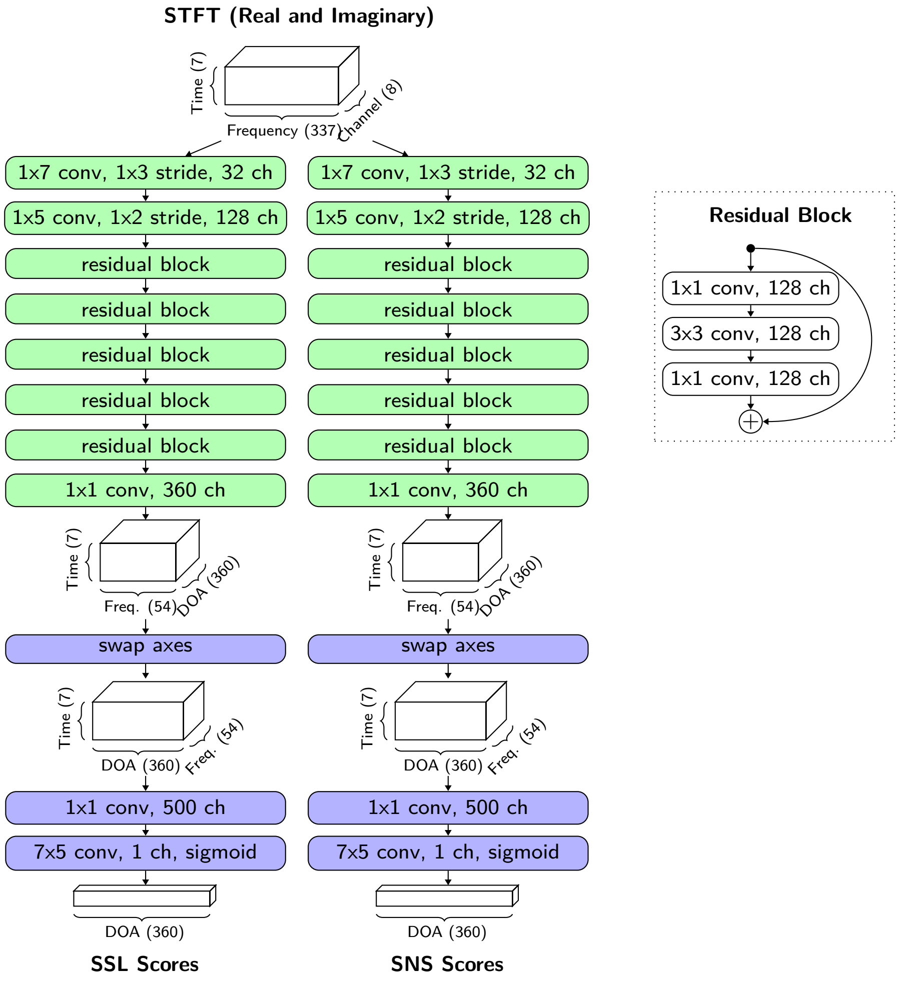


Figure 5.3 – Separated single-task networks (ST.SEP). *Stage 1* consists of the green layers, and *Stage 2* consists of the blue layers. The BN and ReLU activation functions after each hidden layer are omitted in this figure.

5.1.4 Adding Temporal Context

We extend the multi-task network with shared feature by simply adding more temporal context to the input. That is, in addition to the block of 7 frames to be analyzed (i.e. for which we want to make a prediction), we add 10 frames (210 ms) in the past and 10 frames (210 ms) in the future as input to the network, thus reaching an input duration of 600 ms. As the network is fully convolutional, its structure remains the same except for the last convolutional layer where the kernel shape is changed from 7×5 to 27×5 (temporal frames $\times$ DOA).

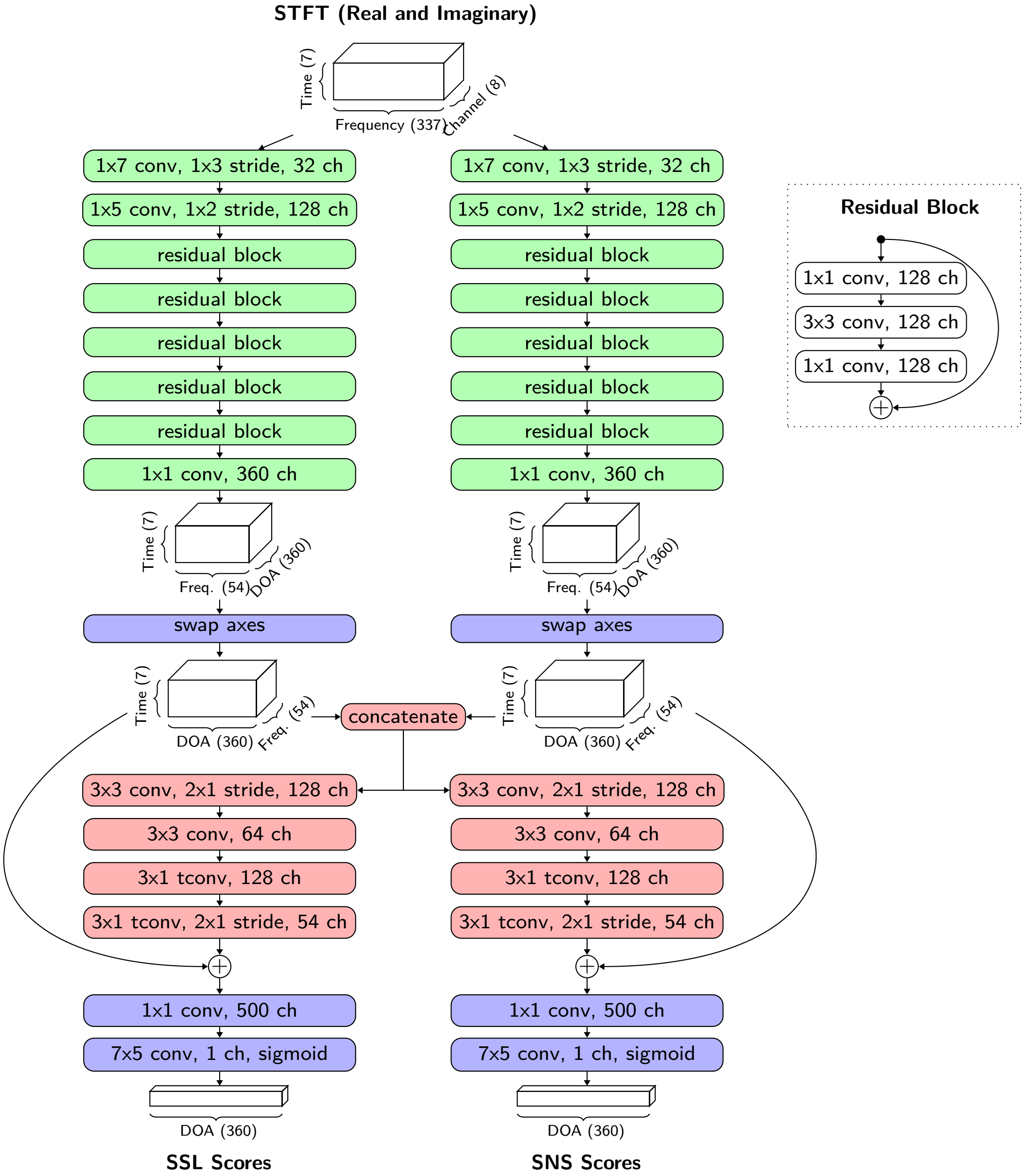


Figure 5.4 – Multi-task network with re-estimation (MT.RE-EST). *Stage 1* consists of the green layers, and *Stage 2* consists of the blue layers. Red layers constitute the *re-estimation module*. “tconv” indicates transposed convolution. The BN and ReLU activation functions after each hidden layer are omitted in this figure.

Table 5.1 – Specifications of the recorded data. 360° means the source can be from any azimuth direction. FoV is the camera's field of view.

	Loudspeaker		Human
	Training	Test	Test
Total duration	32 hours	17 hours	8 min
Max. # of speech	2	2	3
Max. # of noise	1	1	1
# of speakers	148	16	7
DOA range (speech)	360°	360°	in FoV
DOA range (noise)	360°	360°	360°

5.2 Experiments

We collected noisy recordings with our robot Pepper, and evaluated the performance of the methods in terms of sound localization, SNS classification, as well as speech localization.

5.2.1 Data

The collected recordings consist of two sets: the loudspeaker mixtures and human recordings (Table 5.1). The loudspeaker mixture recordings are an extension of the loudspeaker dataset from Section 3.5 by mixing additional non-speech recordings with the speech recordings. The non-speech recordings were collected by playing non-speech audio segments from loudspeakers in the same condition as the speech recordings. These segments are from the Audio Set (Gemmeke et al., 2017) and cover a wide range of audio classes, including a variety of noises, music, nature sounds, animal sounds and non-speech human sounds.

The human recordings involve people having natural conversation or reading with provided scripts while non-speech segments were played from loudspeakers. Ground truth source locations were automatically annotated, while the voice activity was manually labeled.

5.2.2 List of Methods

We include the following methods for comparison:

- **SRP-PHAT**: Steered response power with phase transform (DiBiase et al., 2001).
- **ST.SEP**: Separated single-task networks (Fig. 5.3).
- **ST.SSL**: A single-task network for sound source localization (the SSL part of ST.SEP).
- **ST.SPEECH**: A single-task network (same structure as ST.SSL) for speech localiza-

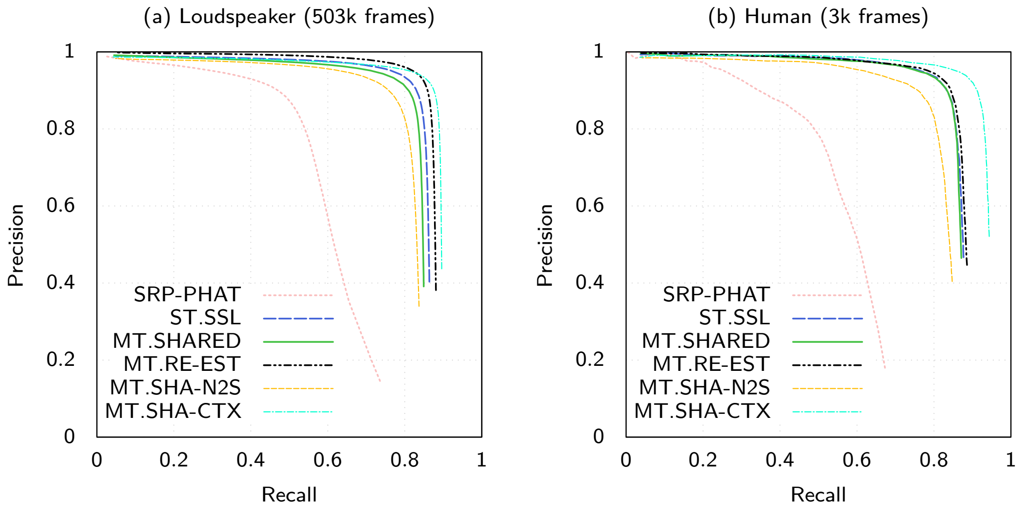


Figure 5.5 – Sound source localization performance.

tion (trained to localize speech sources and ignore noises).

- **SEQUENTIAL**: It first localizes sounds with ST.SSL, then extracts the signals from the estimated DOAs by the MVDR beamformer (Capon, 1969), and finally classifies their sound type with a SNS neural network (with similar ResNet structure).
- **MT.SHARED**: The proposed multi-task network with shared features (Fig. 5.2).
- **MT.RE-EST**: The proposed multi-task network with re-estimation (Fig. 5.4).
- **MT.SHA-N2S**: MT.SHARED trained without using the two-stage training scheme.
- **MT.SHA-CTX**: MT.SHARED with temporal context extension.

5.2.3 Sound Source Localization Results

We evaluate the sound source localization as a detection problem, where the number of sources is not known a priori. To do this, we compute the precision and recall with a varying prediction threshold ξ of Eq. (5.5). A prediction is considered to be correct if it is within 5° of error from a ground truth DOA. Then, we plot the precision vs. recall curves on the two datasets (a) loudspeaker mixtures (b) human recordings (Fig. 5.5). Both proposed multitask approaches (MT.SHARED and MT.RE-EST) achieve more than 90% precision and 80% recall on both datasets. MT.SHARED is slightly worse than the single task network (ST.SSL) on the loudspeaker dataset, it is because adding the regularization with hard feature sharing may compromise the approximation ability of one task. MT.RE-EST, in contrast, outperforms ST.SSL and shows re-estimation refines correctly the DOA estimation. Note that all neural network-based methods are significantly better than SRP-PHAT.

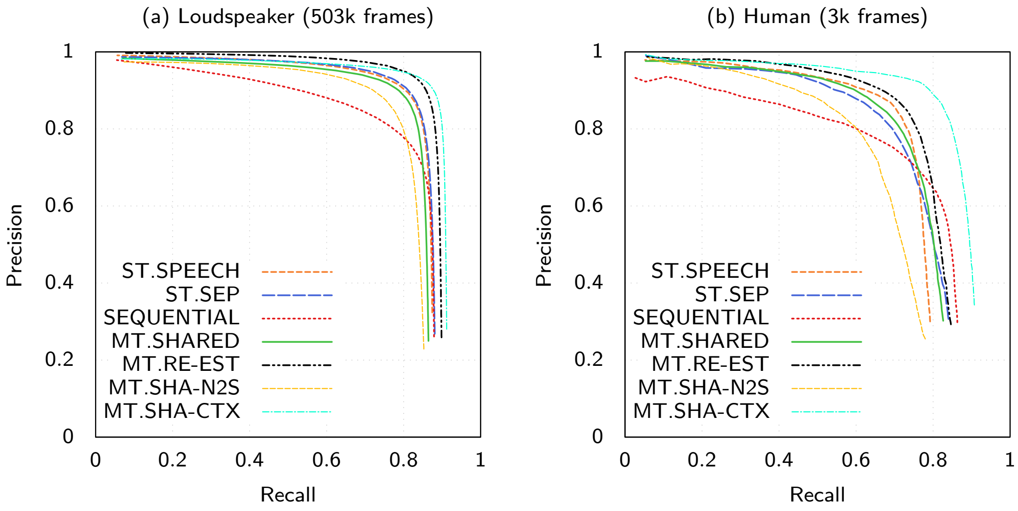


Figure 5.6 – Speech source localization performance.

Table 5.2 – Speech/non-speech classification accuracy. The classification is extracted on directions according to the ground truth (G.T.), or the DOA predictions (Pred.) that have less than 5° of error. The numbers in the parentheses indicate the recall (Rec.) of the DOA predictions.

Dataset	Loudspeaker		Human	
Directions	G.T.	Pred. (Rec.)	G.T.	Pred. (Rec.)
ST.SEP	0.94	0.97 (0.83)	0.81	0.82 (0.83)
SEQUENTIAL	0.80	0.81 (0.83)	0.68	0.73 (0.83)
MT.SHARED	**0.95**	0.97 (0.81)	**0.85**	0.86 (0.82)
MT.RE-EST	**0.95**	**0.98** (0.85)	**0.85**	**0.87** (0.83)
MT.SHA-N2S	0.93	0.96 (0.79)	0.82	0.83 (0.76)
MT.SHA-CTX	0.96	0.98 (0.85)	0.89	0.89 (0.86)

5.2.4 Speech/Non-Speech Classification Results

To evaluate the performance of speech/non-speech classification, we compute the classification accuracy under two conditions: considering the SNS predictions (a) in the ground truth directions, and (b) in the predicted directions (Table 5.2). Specifically, under condition (a), for each ground truth sound source, we check how accurate the method predict its type in the ground truth DOA. Such evaluation is independent of the localization method. Under condition (b), we first detect sound sources using Eq. (5.5) with $\xi = 0.5$, and select the sound sources of which the estimated DOAs are close to the ground truth (error $< 5°$). Then we evaluate the accuracy of SNS classification at these estimated directions. In this case, not all ground truth sources are matched to a prediction (recall is lower than 1 and indicated in the parentheses in Table 5.2). The result is thus dependent on the DOA estimation. Ideally, we prefer methods with high detection recall and high SNS classification accuracy. We can notice that the performance using the predicted DOAs is better than using the ground truth DOAs. This is because the sound sources which are correctly detected have on average higher *Signal-to-Noise Ratio* (SNR), thus it is easier to classify them.

Both proposed multitask approaches (MT.SHARED and MT.RE-EST) achieve more than 95% accuracy on the loudspeaker recordings and more than 85% accuracy on the human recordings. They are both better than ST.SEP and the difference is more prominent on the human recordings, which have more condition mismatch with the training data than the loudspeaker test set. This shows that single-task networks are not as good as the multi-task approaches at generalization. All the end-to-end neural network approaches are significantly better than SEQUENTIAL, which extracts signals by beamforming and then applies classification.

5.2.5 Speech Source Localization Results

We evaluated the speech source localization performance in the same way as that for sound source localization (Fig. 5.6). The results show that the multi-task approaches significantly outperform the sequential approach, due to their better performance in SNS classification. Although MT.SHARED is slightly worse than the single-task approaches (ST.SEP and ST.SPEECH) in the loudspeaker recordings, it achieves better performance than ST.SEP and similar performance as ST.SPEECH in the human recordings. This again indicates that the multi-task learning achieves better generalization under unmatched conditions. The other multi-task approach MT.RE-EST outperforms both single-task approaches as well as MT.SHARED in both datasets.

5.2.6 Two-stage Training and Temporal Context

Results of all three evaluation criteria show that the shared-feature multi-task network trained in two stages (MT.SHARED) is superior than training it with only the end-to-end stage (MT.SHA-N2S). This implies that the two-stage training scheme effectively helps the training process.

In addition, we see that adding temporal context (MT.SHA-CTX) improves both the sound source localization and classification performance, and as a result, greatly improves the speech localization performance. However, such an extension compromises the real-time response of the approach, and may overfit if there are only static sound sources present in the training data. Demonstration videos of MT.SHARED and MT.SHA-CTX are available online[3].

[3] https://www.youtube.com/watch?v=O7bQvg03RTc

5.3 Summary

In this chapter, we have described two novel multi-task neural network approaches for joint DOA estimation and speech/non-speech classification, including one with shared features (MT.SHARED) and one with re-estimation (MT.RE-EST). Both networks use raw STFT as input and predict SSL and SNS scores at each direction. The benefit of sharing features between tasks is adding regularization to the model so that it generalizes better under mismatched conditions, whereas the idea of re-estimation is to share information of initial predictions between task and obtain refined estimation.

Both multi-task approaches achieve better SNS classification and speech localization performance than separated single-task networks, and a sequential approach, which applies DOA estimation, beamforming and classification sequentially. The shared-feature approach (MT.SHARED) does not outperform single-task SSL network and single-task speech localization network in terms of sound localization and speech localization, respectively. This is because regularization with hard parameter sharing may compromise the performance of one task in order to improve the other. In contrast, the re-estimation approach (MT.RE-EST) achieves better performance in all criteria than the single-task approaches. This indicates that re-estimation can improve the performance for both tasks. In addition, we have shown that a simple extension to the shared-feature approach by adding temporal context to inputs can significantly improve its performance.

6 Speaker Embedding for Re-Identification

Person re-identification aims at identifying whether a detected person has been observed before. For person tracking, re-identification is required to establish consistent identity labeling and associate disconnected tracks of the same person. In the context of *Human-Robot Interaction* (HRI), person re-identification is a crucial part of a robot as it needs to "remember" the identities of people for long-term natural interactions.

6.1 Introduction

Re-identification systems are usually based on vision, audio, or both. Vision-based systems identify persons by their visual appearance, such as clothing and faces, and body features, such as gait. A number of visual re-identification approaches have been studied for multi-camera video surveillance as well as HRI (Bedagkar-Gala and Shah, 2014; Wang et al., 2019; Leng et al., 2020). Audio-based systems identify persons by their voices. Audio is used to compensate vision, when talking persons are occluded or outside of the field of view, or used alone when no visual sensor is available. Re-identification is particularly important for audio tracking, because of the transient nature of speech. In natural interactions, people move and do not talk continuously, audio tracking relies on re-identification to keep consistent identity labels for detected voices. Audio and visual person models are combined with early fusion or late fusion for robust multi-modal re-identification (Brutti and Cavallaro, 2017; Marras et al., 2020).

Techniques for audio re-identification are generally known as *speaker recognition*, which includes two different tasks: *speaker identification* and *speaker verification* (Hansen and Hasan, 2015). Speaker identification aims to identify an unknown speaker from a set of known speakers, whereas speaker verification aims to verify if a voice and some enrolled voices are from the same speaker. On top of these two tasks, speaker re-identification also has to actively manage the enrollment process. That is, when an unknown voice is detected, the re-identification system compares it with the models of known speakers, and decides whether it is from a known speaker or a new identity needs to be created. Although the goals of these tasks are different, the techniques they rely on are similar. In fact, most of the speaker recognition methods are based on mapping speech segments to a speaker embedding space where they can be compared using a metric for identification or verification. In an ideal embedding space, distances between voices of the same speaker are smaller than distances between voices of different speakers. Obtaining effective speaker embeddings is key for speaker re-identification.

Speaker recognition with clean and segmented single-channel audio has been extensively studied. Well-known approaches include *Gaussian Mixture Model* (GMM) (Reynolds and Rose, 1995), GMM with *Universal Background Model* (UBM) (Reynolds et al., 2000), *Joint Factor Analysis* (JFA) (Kenny, 2005), *Support Vector Machine* (SVM) for GMM supervector classification (Campbell et al., 2006), and the *i-vector* system (Dehak et al., 2011). Recently, many deep learning based approaches have been shown to outperform the traditional ones (Variani et al., 2014; Snyder et al., 2016, 2018; Bredin, 2017; Le and Odobez, 2018). These deep learning based approaches extract speaker embeddings in two ways. One way is to train a network for speaker identification and use the activation at one of the last hidden layers as speaker embeddings (Variani et al., 2014; Snyder et al., 2018). In contrast, the other way is to use directly the network output as the speaker embedding, and train the network with objective functions that are defined on the distances between same-speaker and different-speaker pairs. Examples of the objective functions include *contrastive loss* (Snyder et al., 2016), which separately minimizes distances between same-speaker pairs and maximizes those between different-speaker pairs, and *triplet loss* (Bredin, 2017; Le and Odobez, 2018), which maximizes the difference between different-speaker distances and same-speaker distances up to a given margin.

Besides speaker recognition using clean audio signals, a number of studies address speaker recognition in the presence of noise and simultaneous speakers. As we have summarized in Section 2.3, these approaches rely on separating the sound sources (from either single-channel or multi-channel audio signals), so that speaker recognition is applied on separated single-channel signals. Sound separation is applied prior and independently to the speaker recognition in the *sequential approaches* (May et al., 2013; Zhao et al., 2012, 2014). Alternatively, sound separation and speaker recognition are solved jointly in the *joint approaches* (Zegers and Van hamme, 2016; Drude et al., 2018; Shi et al., 2020). Nevertheless, using deep neural networks for speaker recognition in multi-speaker conditions is still an emerging topic. Specifically, joint *Direction-of-Arrival* (DOA) estimation and recognition of multiple speakers but has not been studied so far.

This chapter investigates deep neural networks for speaker recognition in multi-speaker conditions using DOA estimation as an auxiliary task. We study an idea that is similar to what is explored in the previous chapter: we use the neural works to extract features for each direction, which are shared for both DOA estimation and speaker embedding. In contrast to previous works, our approach does not rely on explicit separation of the signals. Instead, the network learns to implicitly separate the sound features through end-to-end training. Our proposed neural network shares similarities with the well-known X-vector network (Snyder et al., 2018), that both networks are trained using speaker identification loss and extract speaker embeddings from hidden layers. Moreover, temporal statistic pooling is used in both approaches to accommodate input sequences of variable lengths. The difference between our approach and the X-vector approach is that we address the speaker recognition of multiple overlapping speakers from multi-channel audio, while X-vector approach extract speaker embeddings from single-channel single-speaker audio.

6.2 Approach

We describe our multi-task neural network approach in terms of input representation, network output, loss function and network architecture.

6.2.1 Network Input

We use the raw *Short-Time Fourier Transform* (STFT) as the network input. As we have discussed in the previous chapter (Section 5.1.1), STFT includes both the spectral power information as well as the phase information of the input signal. For DOA estimation, *Inter-channel Level Difference* (ILD) and spectral cues can be extracted from the power information, and *Inter-channel Phase Difference* (IPD) can be extracted from the phase information. The power information, in addition, includes necessary features for speaker recognition.

The STFT is processed in the same way as in the previous chapter, except that the input segment can be arbitrarily long to incorporate more information for speaker recognition. Specifically, STFT is extracted from 4-channel input audio signals at 48 kHz sampling rate using frames of 2048 samples (43 ms) with 50% overlap. The 337 frequency bins between 100 and 8000 Hz are used. The real and imaginary parts of the STFT coefficients are split into two individual channels. Therefore, the input feature of each unit has a dimension of $T \times 337 \times 8$, where the number of frames T varies across different segments.

6.2.2 Network Output and Loss Function

The network output includes frame-wise prediction of the spatial spectrum $\mathbf{p}_t = \{p_{td}\}_{d=1}^{D} \in [0, 1]^D$ for DOA estimation, and segment-wise prediction of speaker posterior probability at each direction $\mathbf{q}_d = \{q_{ds}\}_{s=1}^{S} \in [0, 1]^D$ for speaker identification. The subscripts $t \in \{1, 2, \ldots, T_o\}$ is the frame index, $d \in \{1, 2, \ldots, D\}$ is the direction index, and $s \in \{1, 2, \ldots, S\}$ is the speaker ID. Due to downsampling, the frame rate of predicted spatial spectrum is different from that of the input, thus $T_o \neq T$.

Encoding. The desired output spatial spectrum is encoded by the Gaussian based spatial spectrum coding (Section 3.2), that is:

$$p_{td} = \begin{cases} \max_{\varphi \in y_t} \left\{ e^{-d(\varphi_d, \varphi)^2/\sigma^2} \right\} & \text{if } |y_t| > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{6.1}$$

where $y_t \subset \Phi$ is the set of ground truth directions at frame t, σ is the parameter to control the width of the Gaussian curves, $d(\cdot, \cdot)$ denotes the azimuth angular distance, and $|\cdot|$ denotes the cardinality of a set.

Similar to how we encode the sound type in Section 5.1.2, the speaker ID prediction at direction φ_d depends on the nearest sound source (speaker) to that direction, that is:

$$q_{ds} = \begin{cases} 1 & \text{if Speaker } s \text{ is the nearest speaker to } \varphi_d \\ 0 & \text{otherwise} \end{cases}. \tag{6.2}$$

Loss Functions. The target loss function is a linear combination of the individual task-specific loss functions:

$$\text{Loss} = \mu \, \text{Loss}_{DOA} + \lambda \, \text{Loss}_{ID}, \tag{6.3}$$

where μ and λ are weighting parameters. We use the *Mean Squared Error* (MSE) loss for DOA estimation:

$$\text{Loss}_{DOA} = \frac{1}{T_o} \sum_{t=1}^{T_o} \|\hat{\mathbf{p}}_t - \mathbf{p}_t\|_2^2 , \tag{6.4}$$

where $\hat{\mathbf{p}}_t$ and $\mathbf{p}_t$ are the actual and desired spatial spectrum outputs, respectively. The speaker identification loss is the weighted sum of cross entropy loss at individual directions:

$$\text{Loss}_{ID} = -\sum_{d=1}^{D} w_d \sum_{s=1}^{S} q_{ds} \log \hat{q}_{ds}, \tag{6.5}$$

where $\hat{q}_{ds}$ and q_{ds} are the actual and desired speaker identity outputs, respectively. The weighting $\{w_d\}$ are same as those in the previous chapter:

$$w_d = \begin{cases} \max_{\varphi \in y} \left\{ e^{-d(\varphi_d, \varphi)^2/\sigma_w^2} \right\} & \text{if } |y| > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{6.6}$$

where y contains the segment-level ground truth directions.

Decoding. During test time, the network outputs frame-wise spatial spectra $\mathbf{p}_t$ and speaker embedding $\mathbf{r}_d$ per direction (will be explained in Section 6.2.3). To get segment-level DOA prediction, we compute the average of frame-wise spatial spectra:

$$\mathbf{p} = \frac{1}{T_o} \sum_{t=1}^{T_o} \mathbf{p}_t \tag{6.7}$$

and apply peak finding according to Eq. (3.6). For any detected sound source, the speaker embedding output at its estimated direction is the predicted speaker embedding.

6.2.3 Network Architecture

We design a shared-feature multi-task network for speaker embedding using DOA estimation as an auxiliary task. Its architecture, depicted in Fig. 6.1, consists of a trunk for feature extraction, and two task-specific branches. The trunk (green blocks in the figure) applies 2D convolutions along time and frequency axes to extract *Time-Frequency* (TF) local features. It starts with two downsampling convolutions to reduce the computational cost. They are followed by five residual blocks, which are used for extracting high-level TF-local features, each of which is a 480-dimensional vector. Each of these feature is then separated into DOA-wise features at 120 directions (4-dimensional vector per direction). Then, they are re-organized by merging features across all frequencies (54 bins after down-sampling). As a result, the trunk extracts time-DOA local features, each of which is a 216-dimensional vector ($216 = 4 * 54$). These features are then used as input for the task-specific branches.

The DOA estimation branch (blue blocks in the figure) applies two layers of 2D convolutions along time and DOA axes. The borders are padded circularly along the DOA axis, preserving its actual topology. This branch outputs one value per direction per frame, which is bounded between 0 and 1 by the sigmoid function. This output is the frame-wise spatial spectrum $\mathbf{p}_t$.

The speaker recognition branch (red blocks in the figure) starts with two layers of 2D convolutions to extract frame-wise speaker features per direction $\mathbf{f}_{td} \in \mathbb{R}^{512}$, which is then pooled along the time axis using *weighted average and standard deviation*:

$$\mathbf{f}_d^{(avg)} = \frac{\sum_{t=1}^{T_o} p_{td} \mathbf{f}_{td}}{\sum_{t=1}^{T_o} p_{td}}, \tag{6.8}$$

$$\mathbf{f}_d^{(std)} = \sqrt{\frac{\sum_{t=1}^{T_o} p_{td} \left(\mathbf{f}_{td} - \mathbf{f}_d^{(avg)} \right)^2}{\sum_{t=1}^{T_o} p_{td}}}, \tag{6.9}$$

where $\sqrt{\cdot}$ and $\cdot^2$ are element-wise square root and square, respectively. As indicated in the formulas, we use the output of the DOA estimation branch $\{p_{td}\}$ as the weighting parameters, because the DOA estimation output (i.e. spatial spectrum) indicates whether there is an active sound at that frame and direction. This can be viewed as an attention

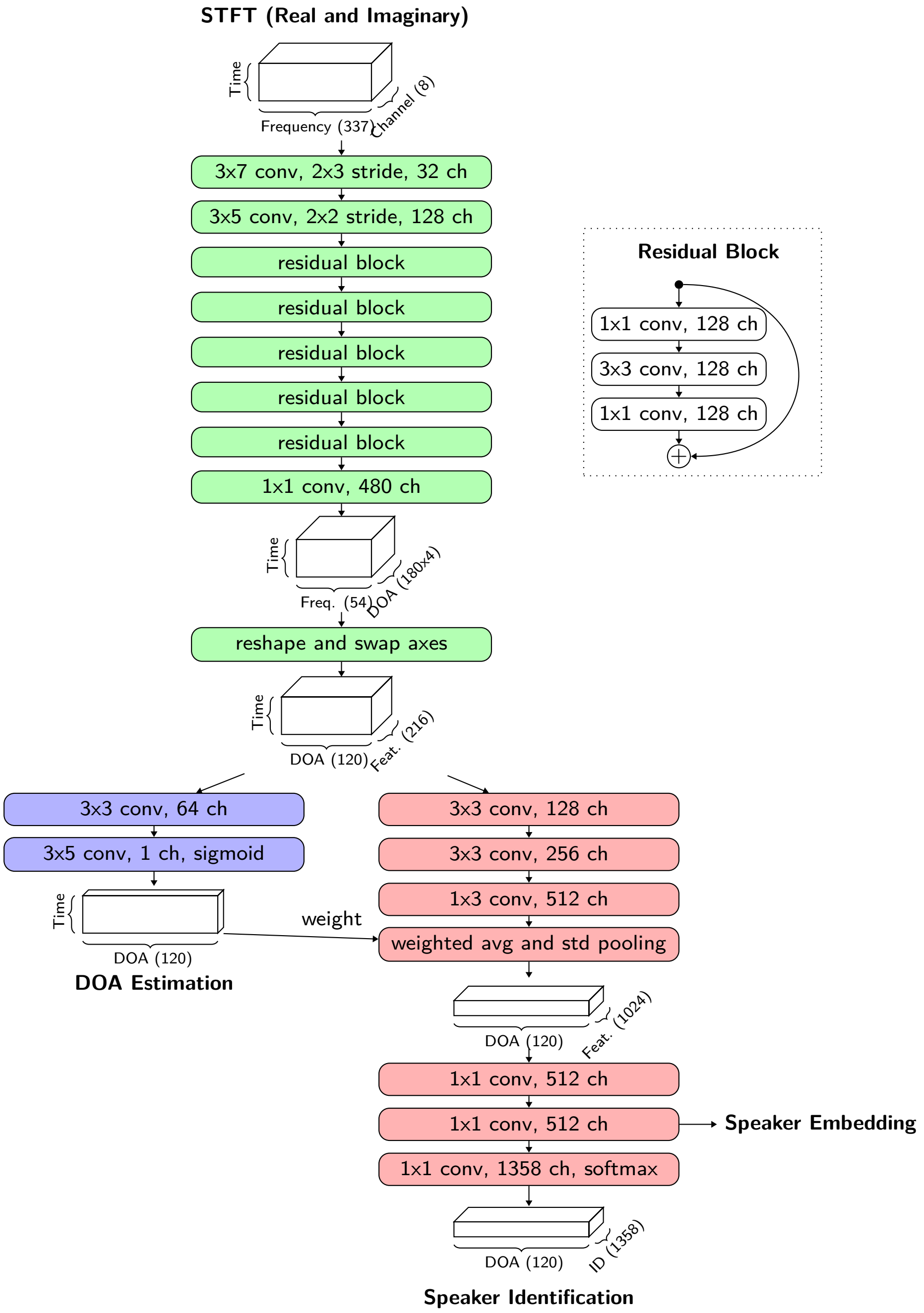


Figure 6.1 – The architecture of the multi-task network for speaker recognition. *Green blocks*: *feature extraction* trunk, which first uses convolutions along time and frequency axes and then assigns individual channels in the TF-local features to different directions to get time-DOA local features. *blue blocks*: *DOA estimation* branch, which consists of convolutions along time and DOA axes. It outputs the DOA estimation which is spatial spectra for each frame. *red blocks*: *speaker recognition* branch, which first extracts frame-level speaker features for each direction, which are then pooled into segment-level speaker features for each direction. The DOA estimation results are used as the weighting parameters for the temporal pooling. It outputs posterior probability of the speaker ID at each direction, and the activation of the last hidden layer is used as speaker embeddings at each direction.

mechanism, that is the network chooses by itself which frames to attend to.

Their concatenation $\mathbf{f}_d = [\mathbf{f}_d^{(avg)}\ \mathbf{f}_d^{(std)}] \in \mathbb{R}^{1024}$ is the segment-level speaker feature per direction. Then, the speaker identity posterior probability is computed from these features with fully-connected layers (1×1 convolutions) and a softmax layer.

At test time, the 512-dimensional activation of the last hidden layer after batch normalization is the speaker embedding $\mathbf{r}_d$ at the direction φ_d.

6.3 Experiments

We compared the proposed approach to a sequential approach on our loudspeaker dataset in a speaker verification setting.

6.3.1 Data

We used the loudspeaker recordings desribed in Section 3.5 for training and evaluation. The specifications of the loudspeaker data are listed in Table 3.1. Training models that identify speakers per direction requires more variability in DOAs of individual speaker as well and number of identities. Otherwise, the network may overfit to a wrong state where spatial locations are used as the clues for speakers' identity. Therefore, we added simulated data to complement the real loudspeaker recordings for training. The simulation process was the same reverberant room acoustic simulation described in Section 4.4.1, except that the source signals were selected from the VoxCeleb1 dataset (Nagrani et al., 2017), which includes more speaker identities than our previous simulated dataset. In total, there are 1358 speakers (147 from the loudspeaker data and 1211 for simulated data) for training and 16 different speakers (8 male and 8 female) for evaluation.

6.3.2 Details on Parameters and Training Process

The parameters are chosen as $D = 120$, $\sigma = \sigma_n = 8°$, and $\sigma_w = 16°$ in the experiments. Various settings of loss weighting parameters μ and λ are experimented, which we will explain in Section 6.3.4.

During training, the input segments in each mini-batch are randomly truncated to the same length between 3 to 10 seconds or 2 to 5 seconds, depending on the training stage. In each mini-batch, 10.0% of the sequences are sampled from the loudspeaker dataset, and the rest are sampled from the simulated dataset. We select the number of sequences in each mini-batch, such that they approximately fill up the memory of a GPU with 11 GB memory. Thus, depending on the sequence length, the number of sequences in each mini-batch varies between 10 to 90. The network is trained for 80 epochs with an Adam optimizer (Kingma and Ba, 2015). The learning rate is 0.001 for the first 40 epochs and reduced by half for the other 40 epochs.

6.3.3 Evaluation Protocol

We evaluate the *Equal Error Rate* (EER) of the speaker embedding methods under a speaker verification setting. First the loudspeaker test data are segmented into 2, 3, 5, and 10 second frames or used directly at utterance level. For each sound source in each segment, we extract the speaker embedding according to its ground truth direction, estimated direction, or direction estimated by an external model (i.e. the ResNet-STFT in Chapter 3). Then, we generate verification trials using 5 million randomly-sampled pairs of sound sources, or all possible pairs if the number of them are fewer than 5 million. We compute the cosine similarity scores between the speaker embeddings of all trial pairs. Comparing the scores to a threshold, we can make predictions on whether the speakers from a trial pair have the same identity. The EER is the rate when false acceptance rate and false rejection rate are equal while varying the threshold.

In addition to segment duration, speaker verification are also strongly affected by whether there are overlapping speech in the audio signal. Therefore, we additionally report the speaker verification performance on trail pairs of these different conditions:

- Both speaker embeddings are sampled from the single-source segments;
- One is from the single-source segments and one is from the multi-source segments;
- Both are from the multi-source segments.

6.3.4 List of Methods

We include the following methods for comparison:

- **SEQ (original)**: A sequential approach based on the *Minimum Variance Distortionless Response* (MVDR) beamformer (Capon, 1969) and a deep neural network for speaker embedding (Le and Odobez, 2018). The neural network directly output speaker embedding using single-channel audio input. It is trained on the VoxCeleb1 dataset (Nagrani et al., 2017) with a triplet loss and intra-class distance variance regularization. The DOA for beamforming is either based on the ground truth or an estimation from the ResNet-STFT model.
- **SEQ (fine-tuned)**: A variant of SEQ (original) with a fine-tuned speaker embedding neural network model. The model is fine-tuned using the beamformed signals extracted from the loudspeaker training data according to the ground truth DOAs.
- **PROP ($\mu = 1$, $\lambda = 0.1$)**: The proposed approach using 3-10 second training segments and the loss weighting that emphasizes more the DOA estimation loss.

Table 6.1 – EER (%) on trials with pairs of sound sources from any segments. The second column indicates how the DOA used for speaker embedding extraction is obtained: according to the ground truth (G.T.), prediction made by the external model (P. (E)) or prediction made the DOA estimation branch of the multi-task model (P.). The best EERs on predicted DOAs (i.e. excluding G.T.) are highlighted.

Method	DOA	Utterance	10s	5s	3s	2s
SEQ (original)	G.T.	13.32	9.68	12.83	15.33	18.29
-	P. (E)	13.34	9.64	12.81	15.33	18.29
SEQ (fine-tuned)	G.T.	11.49	8.72	11.07	13.48	18.90
-	P. (E)	11.51	8.67	11.04	**13.49**	18.93
PROP ($\mu = 1$, $\lambda = 0.1$)	G.T.	9.60	5.84	10.37	14.70	20.14
-	P. (E)	9.65	5.89	10.44	14.78	20.18
-	P.	9.53	**5.69**	10.22	14.57	20.03
PROP ($\mu = 0.1$, $\lambda = 1$)	G.T.	10.45	6.66	11.57	16.43	21.77
-	P. (E)	10.49	6.74	11.60	16.47	21.78
-	P.	14.79	9.98	14.12	18.23	23.07
PROP (with pre-tr.)	G.T.	9.05	5.91	9.76	13.42	17.80
-	P. (E)	**9.11**	5.97	**9.83**	13.50	**17.85**
-	P.	11.21	6.96	10.85	14.32	18.63

- **PROP ($\mu = 0.1$, $\lambda = 1$)**: The proposed approach using 3-10 second training segments and the loss weighting that emphasizes more the speaker identification loss.
- **PROP (with pre-tr.)**: The proposed approach with pre-training. This model is pre-trained with an emphasis on DOA estimation ($\mu = 1$, $\lambda = 0.1$) and 3-10 second training segments. Then, as the second step, it is trained with an emphasis on speaker identification ($\mu = 0.1$, $\lambda = 1$) and 2-5 second training segments.

6.3.5 Speaker Verification Performance

We compute the EER of the aforementioned methods under various conditions according to the evaluation protocol. Table 6.1 shows the results when the verification trial pairs are sampled from any audio segments, whereas Table 6.2 shows the case when the trial pairs are sampled from segments containing only one sound source, Table 6.3 shows the case when one sound source is sampled from single-source segments and the other is from multi-source segments, and Table 6.4 shows the case when both sound sources are from multi-source segments. In the rest of this section, we discuss how DOA estimation, loss-weighting and pre-training impact the speaker verification performance, and how our proposed multi-task learning approach is compared to sequential methods.

DOA estimation. Comparing the results of a method with different DOA estimation, the speaker embeddings extracted from the directions predicted by an external model is as good as using the ground truth. This indicates that the prediction of the external model is accurate and the speaker embedding approaches are robust to small error in DOA estimation. However, the performance is degraded when an inaccurate DOA estimation is used for speaker embedding extraction. This is the case when models are trained with an emphasis on speaker identification and the DOA estimation accuracy is compromised.

Loss weighting and pre-training. Among the three proposed approaches with different loss weighting, the one with an emphasis on speaker recognition after a pre-training step geared towards DOA estimation achieves the best overall performance using either ground truth directions or estimations of the external model. Direct training with the same weighting parameters for the DOA estimation and speaker recognition losses does not achieve the same performance, probably because without substantial supervision on the DOA estimation the network cannot learn a proper weighting parameters for the temporal statistic pooling in the speaker recognition branch. In contrast, pre-training with an emphasis on DOA estimation initializes the model with reasonable temporal weighting parameters. Better weighting parameters, which might not be the ideal spatial spectra, can then derived from the subsequent training setting with a higher weight on the speaker identification loss.

Proposed vs. sequential methods. The proposed methods, compared to the sequential approaches, achieve better overall performance in long segments (utterance level, 10-second segments and 5-second segments), while their EERs in short segments (2 and 3 second segments) are similar. Their performance under different trial conditions (Tables 6.2 to 6.4) indicates that while the proposed methods are not as good as extracting speaker embedding in single-source segments, they are better in general under the multi-source conditions. In the single-source case, the sequential approach is not influenced much by the beamformer, as sound source separation is not necessary and the single-channel speaker embedding network can be trained to handle noisy input (as what the fine-tuning is for). In contrast, our proposed multi-task network is trying to extract speaker embeddings on all directions, and is more complex than a single-channel single-speaker embedding network. Therefore, it is more difficult to train. In our experiments, we find that the single-channel speaker embedding approach is more suitable for single-source conditions. However, for segments containing multiple sound sources, the sequential approach relies on the beamformer to separate the signals. Its speaker embedding performance may degrade due to imperfect sound separation, whereas our proposed approach does not require explicit sound separation.

6.4 Summary

In this chapter, we have presented a multi-task network for extracting speaker embeddings of multiple simultaneous speakers using DOA estimation as an auxiliary task. The network learns to output a spatial spectrum score and a speaker embedding for each direction.

Table 6.2 – EER (%) on trials with pairs of sound sources from single-source segments.

Method	DOA	Utterance	10s	5s	3s	2s
SEQ (original)	P. (E)	8.89	5.73	8.85	11.75	**15.06**
SEQ (fine-tuned)	P. (E)	**6.77**	**3.80**	**7.07**	**10.19**	16.42
PROP ($\mu = 1$, $\lambda = 0.1$)	P. (E)	8.94	4.87	9.22	13.77	19.17
PROP ($\mu = 0.1$, $\lambda = 1$)	P. (E)	9.66	6.05	10.36	15.14	20.38
PROP (with pre-tr.)	P. (E)	8.44	5.24	8.83	12.50	16.88

Table 6.3 – EER (%) on trials with pairs of sound sources that one is from single-source segments and the other from multi-source segments.

Method	DOA	Utterance	10s	5s	3s	2s
SEQ (original)	P. (E)	14.03	11.32	14.44	17.16	20.10
SEQ (fine-tuned)	P. (E)	12.27	10.82	12.73	15.23	20.40
PROP ($\mu = 1$, $\lambda = 0.1$)	P. (E)	9.75	6.43	10.93	15.36	20.74
PROP ($\mu = 0.1$, $\lambda = 1$)	P. (E)	10.57	6.94	12.00	17.09	22.45
PROP (with pre-tr.)	P. (E)	**9.21**	**6.36**	**10.22**	**13.99**	**18.43**

Table 6.4 – EER (%) on trials with pairs of sound sources from multi-source segments.

Method	DOA	Utterance	10s	5s	3s	2s
SEQ (original)	P. (E)	16.12	12.44	15.87	18.87	22.09
SEQ (fine-tuned)	P. (E)	14.65	12.24	14.29	16.87	21.70
PROP ($\mu = 1$, $\lambda = 0.1$)	P. (E)	10.17	6.46	11.50	16.13	21.47
PROP ($\mu = 0.1$, $\lambda = 1$)	P. (E)	11.19	7.60	13.16	18.61	23.95
PROP (with pre-tr.)	P. (E)	**9.54**	**6.34**	**10.76**	**14.71**	**19.18**

The spatial spectrum is used as weighting parameters for weighted average and standard deviation pooling of the frame-wise speaker features along the time axis. Compared to a sequential approach that applies separately DOA estimation, beamforming and speaker embedding extraction, our proposed approaches achieves better overall performance for audio segments with overlapping sound sources.

7 Conclusion

This chapter summarizes the contributions of this thesis, and suggests possible directions for future research.

7.1 Summary of Contributions

This thesis has reported progress in three important aspects of robotic auditory perception:

- *Deep Neural Network* (DNN) architectures and input/output representations for multi-speaker *Direction-of-Arrival* (DOA) estimation;
- DNN training procedure with domain adaptation;
- Multi-task learning for robotic auditory perception.

We have proposed several DNN models, and deployed them on our Pepper robot in challenging *Human-Robot Interaction* (HRI) scenarios. Their effectiveness is verified by experiments with real data. Furthermore, some of the models are combined with visual tracking, and integrated into a real-time robotic system, which has been successfully demonstrated in public (Foster et al., 2019).

Although the earliest works on artificial neural networks for DOA estimation date back to 1990s, the neural network based approaches under challenging acoustic conditions were not studied in depth until recently. Our approaches presented in Chapter 3 are among the first deep learning based approaches for multi-speaker DOA estimation. We have proposed the Gaussian based spatial spectrum output coding, which can handle an arbitrary number of sources and does not rely on a priori knowledge about the number of them. Unlike "0-1" assignment in the posterior coding, our approach adopts soft-assignment, taking account of the variance in estimation and correlation among neighboring directions. This idea has been followed by other researchers as well (Nguyen et al., 2020). In terms of performance, our deep learning based approaches achieve significantly better results than the traditional spatial spectrum based approaches. This is because neural networks can implicitly learn the propagation and signal models from example data, while it is difficult to create these models analytically. Learning-based approaches in general are more adaptive to complex environments as long as a sufficient number of training data are available.

As the second topic, we have studied domain adaptation methods for DOA estimation neural networks. In fact, the success of deep learning based approaches rely on a sufficient number of data, thus the cost of data collection should not be overlooked. In particular, since audio data from distinct microphone arrays are radically different, the high cost of data collection actually limits the practical application of deep learning based approaches. In this view, we believe domain adaptation techniques can mitigate such issues related to costly data collection. Although it is an important topic, very few research has been done so far. Our research presented in Chapter 4 is the first to address domain adaptation of multi-speaker DOA estimation models. Specifically, we have studied domain adaptation methods under supervised, weakly-supervised and unsupervised settings. We have found that our proposed weakly-supervised approach based on minimum distance criterion and pseudo-labeling on augmented data is well suited for practical applications. This approach requires only the labeling of the number of sound sources instead of exact locations, and reduces significantly the annotation workload. It has shown very promising results, achieving similar performance as supervised approaches.

Finally, we have investigated multi-task learning neural networks for robotic auditory perception, including joint DOA estimation and *Speech/Non-Speech* (SNS) classification (Chapter 5) and speaker embedding in multi-speaker environments (Chapter 6). In contrast to sequential approaches, our approaches solve multiple tasks simultaneously, allowing knowledge of one task to be shared with the others. In fact, the estimated directions of the sound sources provide spatial information which is useful for the sound classifier or the voice embedding system to separate the features of the sound sources, and vice versa, knowledge about the sound type or speaker identity provides prior information of sound spectral distribution that can help the DOA estimation. Our experiments show that our joint approaches for DOA estimation and SNS classification outperform single-task approaches and sequential approaches that separately apply DOA estimation, beamforming, and sound classification. Furthermore, our multi-task speaker embedding approach achieves better speaker verification performance in multi-speaker environments than the sequential approaches.

7.2 Future Work

Besides the aforementioned progress, we have also observed the following limitations about our current research:

- More complex conditions are not yet addressed. We have studied DOA estimation of up to three overlapping static sound sources. However, we do not know whether our approaches can generalize well to other conditions such as moving sound sources, more than three overlapping sound sources, or presence of babble noise, robot motor noise when it is moving, and voice of the robot when it is speaking.

- Adaptive speaker model management for person re-identification is not implemented, and the speaker embedding performance of our proposed multi-task network is not as good as the sequential approaches under single-source conditions.
- Unified auditory perception neural network model is not studied. So far, the speaker embedding network is separate from the network that is for joint DOA estimation and SNS classification, and it requires an external model for DOA estimation. Although, these models can run in parallel on an external computer that receives streaming data from the robot, it is not possible to run them using the on-board chip. A unified auditory perception model may reduce the total computational requirement and potentially improve the overall performance.
- Integration with visual perception is not studied. We have integrated the speech detection with a visual tracker using a simple decision fusion rule. However, this thesis does not address how to fuse both modalities and adapt models taking account of their different reliability.

In future work, we suggest several directions that might solve these issues.

Collection of more data. Availability of a sufficient number of real data is the basis for evaluation and training of models. By collecting data under the aforementioned complex conditions, we can objectively and comprehensively study the generalization of our DNN approaches. Moreover, collecting more data for training may potentially solve the performance issues of our models, because bigger and deeper network models can be trained with more data without being affected by overfitting.

Temporal context. Following our approaches, more research could be done for incorporating temporal context. Besides the experiment in Section 5.1.4, we haven't explored the temporal context due to the lack of training data with moving sources. However, temporal context provides more information and is required for more robust auditory perception. Besides increasing the local receptive fields of the *Convolutional Neural Networks* (CNNs), as what we use, *Recurrent Neural Network* (RNN) models can be explored to incorporate long temporal context.

Integration of more functions. We have studied multi-task DNN models for at most two tasks. However, the same idea can be applied to create a unified system for more auditory perception functions including DOA estimation, SNS classification, speaker embedding as well as speech recognition. The benefit of integrating more functions is that it works as a regularization on the models, and can create models with better generalization ability. However, the degree of parameter sharing and loss weighting need to be carefully chosen, as the performance of one task may be compromised to improve the other tasks.

Online and autonomous model adaptation. A possible extension to our domain adaptation methods for DOA estimation DNN models is to develop online adaption that could be one autonomously by robots during their exploration and interactions. Visual feedback, such as visual *Voice Activity Detection* (VAD), could be used for self-supervised long-term learning. In addition, modeling turn-taking in conversations may provide prior knowledge of the speech activity, which can also be useful for the adaptation. Multiple robots could be used together to simulate multi-party interactions.

Explicit signal modeling. Another interesting topic is explicit modeling of the signal prior distribution, since it is known that prior knowledge about the target signals helps auditory perception. One advantages of DNNs over traditional approaches is that they can implicitly learn the signal prior. Recent studies have shown that it is possible to model speech prior with *Variational Autoencoder* (VAE) for speech enhancement (Bando et al., 2018; Leglaive et al., 2019; Sekiguchi et al., 2019). However, this idea has not been applied to sound source localization. Unlike multi-channel audio recordings which are scarce, clean single-channel audio data are abundant. Models of signal prior could lead to unsupervised (no labeled multi-channel audio recordings required) learning approaches for sound source localization, which model the sound sources in a probabilistic framework, and infer the sound signals as well as their locations with the maximum likelihood. In addition, the separated signals, as a byproduct, can be used for other perception functions, such as speaker and speech recognition.

Bibliography

Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. (2019). Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48.

Adavanne, S., Politis, A., and Virtanen, T. (2018). Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *Proceedings of 2018 26th European Signal Processing Conference (EUSIPCO)*, Rome.

Algazi, V., Duda, R., Thompson, D., and Avendano, C. (2001). The CIPIC HRTF database. In *2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 99–102.

Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950.

An, I., Jo, B., Kwon, Y., Choi, J.-w., and Yoon, S.-e. (2020). Robust sound source localization considering similarity of back-propagation signals. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*.

An, I., Son, M., Manocha, D., and Yoon, S.-E. (2018). Reflection-aware sound source localization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 66–73.

Apolinario, J. A., Yazdanpanah, H., Nascimento, A. S., and de Campos, M. (2019). A data-selective ls solution to TDOA-based source localization. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4400–4404, Brighton, United Kingdom. IEEE.

Argentieri, S., Danès, P., and Souères, P. (2015). A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech & Language*, 34(1):87–112.

Bando, Y., Mimura, M., Itoyama, K., Yoshii, K., and Kawahara, T. (2018). Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 716–720.

Bedagkar-Gala, A. and Shah, S. K. (2014). A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, 79(1):151–175.

Berglund, E. and Sitte, J. (2005). Sound source localisation through active audition. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 653–658.

Blanchard, G., Lee, G., and Scott, C. (2011). Generalizing from several related classification tasks to a new unlabeled sample. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2178–2186. Curran Associates, Inc.

Blandin, C., Ozerov, A., and Vincent, E. (2012). Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. *Journal of Signal Process.*, 92(8):1950–1960.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):49–57.

Bourlard, H. A. and Morgan, N. (1994). *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers.

Brandstein, M., Adcock, J., and Silverman, H. (1997). A closed-form location estimator for use with room environment microphone arrays. *IEEE Transactions on Speech and Audio Processing*, 5(1):45–50.

Brandstein, M. S. and Silverman, H. F. (1997). A robust method for speech signal time-delay estimation in reverberant rooms. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 375–378.

Braun, S., Neil, D., Anumula, J., Ceolini, E., and Liu, S.-C. (2018). Multi-channel attention for end-to-end speech recognition. In *Interspeech 2018*, pages 17–21.

Breazeal, C. (2004). Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 34(2):181–186.

Bredin, H. (2017). TristouNet: Triplet loss for speaker turn embedding. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5430–5434.

Brutti, A. and Cavallaro, A. (2017). Online cross-modal adaptation for audio-visual person identification with wearable cameras. *IEEE Transactions on Human-Machine Systems*, 47(1):40–51.

Campbell, D., Palomaki, K., and Brown, G. (2005). A matlab simulation of "shoebox" room acoustics for use in research and teaching. *Computing and Information Systems*, 9(3):48.

Campbell, W., Sturim, D., and Reynolds, D. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299.

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.

Chakrabarty, S. and Habets, E. A. P. (2017). Broadband DOA estimation using convolutional neural networks trained with noise signals. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 136–140.

Chakrabarty, S. and Habets, E. A. P. (2019). Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):8–21.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5):975–979.

Choi, J., Jeong, M., Kim, T., and Kim, C. (2019). Pseudo-labeling curriculum for unsupervised domain adaptation. *arXiv:1908.00262 [cs]*.

Cobos, M., Lopez, J. J., and Martinez, D. (2011). Two-microphone multi-speaker localization based on a laplacian mixture model. *Digital Signal Processing*, 21(1):66–76.

Cooke, M., Garcia Lecumberri, M. L., and Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, 123(1):414–427.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285.

Crocco, M., Martelli, S., Trucco, A., Zunino, A., and Murino, V. (2017). Audio tracking in noisy environments by acoustic map and spectral signature. *IEEE Transactions on Cybernetics*, PP(99):1–14.

Datum, M. S., Palmieri, F., and Moiseff, A. (1996). An artificial neural network for sound localization using binaural cues. *The Journal of the Acoustical Society of America*, 100(1):372–383.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Deleforge, A., Forbes, F., and Horaud, R. (2013). Variational EM for binaural sound-source separation and localization. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 76–80.

Deleforge, A., Forbes, F., and Horaud, R. (2015a). Acoustic space learning for sound-source separation and localization on binaural manifolds. *International Journal of Neural Systems*, 25(01).

Deleforge, A. and Horaud, R. (2012). 2d sound-source localization on the binaural manifold. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6.

Deleforge, A., Horaud, R., Schechner, Y., and Girin, L. (2015b). Co-localization of audio sources in images using binaural features and locally-linear regression. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):718–731.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Di Carlo, D., Deleforge, A., and Bertin, N. (2019). Mirage: 2d source localization using microphone pair augmentation with echoes. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 775–779, Brighton, United Kingdom.

DiBiase, J. H., Silverman, H. F., and Brandstein, M. S. (2001). Robust localization in reverberant rooms. In Brandstein, M. and Ward, D., editors, *Microphone Arrays: Signal Processing Techniques and Applications*, Digital Signal Processing, pages 157–180. Springer, Berlin, Heidelberg.

Dmochowski, J. P., Benesty, J., and Affes, S. (2007). Broadband music: Opportunities and challenges for multiple source localization. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 18–21.

Drude, L., von Neumann, T., and Haeb-Umbach, R. (2018). Deep attractor networks for speaker re-identification and blind source separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11–15.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Dumoulin, V. and Visin, F. (2018). A guide to convolution arithmetic for deep learning. *arXiv:1603.07285 [cs, stat]*.

El Badawy, D., Dokmanić, I., and Vetterli, M. (2017). Acoustic doa estimation by one unsophisticated sensor. In Tichavský, P., Babaie-Zadeh, M., Michel, O. J., and Thirion-Moreau, N., editors, *Latent Variable Analysis and Signal Separation*, Lecture Notes in Computer Science, pages 89–98, Cham. Springer International Publishing.

Ferguson, E. L., Williams, S. B., and Jin, C. T. (2018). Sound source localization in a multipath environment using convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Florentine, M., Buus, S., Scharf, B., and Canevet, G. (1984). Speech reception thresholds in noise for native and non-native listeners. *The Journal of the Acoustical Society of America*, 75(S1):S84–S84.

Foster, M. E., Alami, R., Gestranius, O., Lemon, O., Niemelä, M., Odobez, J.-M., and Pandey, A. K. (2016). The MuMMER project: Engaging human-robot interaction in real-world public spaces. In *Social Robotics*, pages 753–763. Springer, Cham.

Foster, M. E., Craenen, B., Deshmukh, A., Lemon, O., Bastianelli, E., Dondrup, C., Papaioannou, I., Vanzo, A., Odobez, J.-M., Canévet, O., Cao, Y., He, W., Martínez-González, A., Motlicek, P., Siegfried, R., Alami, R., Belhassein, K., Buisan, G., Clodic, A., Mayima, A., Sallami, Y., Sarthou, G., Singamaneni, P.-T., Waldhart, J., Mazel, A., Caniot, M., Niemelä, M., Heikkilä, P., Lammi, H., and Tammela, A. (2019). MuMMER: Socially intelligent human-robot interaction in public spaces. *arXiv:1909.06749 [cs]*.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Gardner, W. G. and Martin, K. D. (1995). HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America*, 97(6):3907–3908.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Georganti, E., May, T., van de Par, S., Harma, A., and Mourjopoulos, J. (2011). Speaker distance detection using a single microphone. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1949–1961.

Grandvalet, Y. and Bengio, Y. (2004). Semi-supervised learning by entropy minimization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, pages 529–536, Cambridge, MA, USA. MIT Press.

Griffiths, L. and Jim, C. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34.

Grondin, F., Glass, J., Sobieraj, I., and Plumbley, M. D. (2019). Sound event localization and detection using CRNN on pairs of microphones. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*.

Grondin, F. and Michaud, F. (2015). Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6149–6154.

Habets, E. A. (2006). Room impulse response generator. *Technische Universiteit Eindhoven, Tech. Rep*, 2(2.4):1.

Hansen, J. H. and Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, 32(6):74–99.

Hartmann, W., Narayanan, A., Fosler-Lussier, E., and Wang, D. (2013). A direct masking approach to robust ASR. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):1993–2005.

Haykin, S. and Chen, Z. (2005). The cocktail party problem. *Neural Computation*, 17(9):1875–1902.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv:1512.03385 [cs]*.

He, W., Lu, L., Zhang, B., Mahadeokar, J., Kalgaonkar, K., and Fuegen, C. (2020). Spatial attention for far-field speech recognition with deep beamforming neural networks. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7499–7503.

He, W., Motlicek, P., and Odobez, J.-M. (2018a). Deep neural networks for multiple speaker detection and localization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 74–79.

He, W., Motlicek, P., and Odobez, J.-M. (2018b). Joint localization and classification of multiple sound sources using a multi-task neural network. In *Interspeech 2018*, pages 312–316.

He, W., Motlicek, P., and Odobez, J.-M. (2019). Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 770–774.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hirvonen, T. (2015). Classification of spatial audio location and content using convolutional neural networks. In *Proc. Audio Eng. Soc. Conv. 138*, Warsaw, Poland.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.

Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., and Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems*, pages 601–608.

Hughes, T. and Mierle, K. (2013). Recurrent neural networks for voice activity detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7378–7382.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *PMLR*, pages 448–456.

Jarrett, D. P., Habets, E. a. P., Thomas, M. R. P., and Naylor, P. A. (2012). Rigid sphere room impulse response simulation: Algorithm and applications. *The Journal of the Acoustical Society of America*, 132(3):1462–1472.

Jourjine, A., Rickard, S., and Yilmaz, O. (2000). Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 2985–2988.

Kapka, S. and Lewandowski, M. (2019). Sound source detection, localization and classification using consecutive ensemble of CRNN models. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019).*

Kenny, P. (2005). Joint factor analysis of speaker and session variability: Theory and algorithms. Tech. Rep. CRIM-06/08-13, CRIM, Montreal, Quebec, Canada.

Keyrouz, F., Naous, Y., and Diepold, K. (2006). A new method for binaural 3-d localization based on hrtfs. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*.

Khalidov, V. and Odobez, J.-M. (2017). Real-time multiple head tracking using texture and colour cues. Technical Report Idiap-RR-02-2017, Idiap.

Kim, U.-H., Mizumoto, T., Ogata, T., and Okuno, H. G. (2011). Improvement of speaker localization by considering multipath interference of sound wave for binaural robot audition. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2910–2915.

Kim, U.-H., Nakadai, K., and Okuno, H. G. (2015). Improved sound source localization in horizontal plane for binaural robot audition. *Applied Intelligence*, 42(1):63–74.

Kingma, D. P. and Ba, J. (2015). Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diego.

Knapp, C. and Carter, G. (1976). The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327.

Kulowski, A. (1985). Algorithmic representation of the ray tracing technique. *Applied Acoustics*, 18(6):449–469.

Kwon, B., Park, Y., and Park, Y.-s. (2010). Analysis of the GCC-PHAT technique for multiple sources. In *ICCAS 2010*, pages 2070–2073.

Le, N. and Odobez, J.-M. (2018). Robust and discriminative speaker embedding via intra-class distance variance regularization. In *Interspeech 2018*, pages 2257–2261.

Le Roux, J., Vincent, E., Hershey, J. R., and Ellis, D. P. W. (2015). Micbots: Collecting large realistic datasets for speech and audio research using mobile robots. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5635–5639.

Lee, D.-H. (2013). Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on challenges in representation learning, ICML.*

Leglaive, S., Girin, L., and Horaud, R. (2019). Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 101–105.

Leng, Q., Ye, M., and Tian, Q. (2020). A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1092–1108.

Lim, H., Yoo, I.-C., Cho, Y., and Yook, D. (2015). Speaker localization in noisy environments using steered response voice power. *IEEE Transactions on Consumer Electronics*, 61(1):112–118.

Liu, H., Zhang, Z., Zhu, Y., and Zhu, S.-C. (2019). Self-supervised incremental learning for sound source localization in complex indoor environment. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2599–2605.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2016). Unsupervised domain adaptation with residual transfer networks. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 136–144. Curran Associates, Inc.

Löllmann, H. W., Evers, C., Schmidt, A., Mellmann, H., Barfuss, H., Naylor, P. A., and Kellermann, W. (2018). The LOCATA challenge data corpus for acoustic source localization and tracking. In *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 410–414.

Ma, N. and Brown, G. J. (2016). Speech localisation in a multitalker mixture by humans and machines. In *INTERSPEECH 2016*, pages 3359–3363.

Ma, N., Brown, G. J., and May, T. (2015a). Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions. *Proceedings of Interspeech 2015*, pages 3302–3306.

Ma, N., May, T., and Brown, G. J. (2017). Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2444–2453.

Ma, N., May, T., Wierstorf, H., and Brown, G. J. (2015b). A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2699–2703.

Mack, W., Bharadwaj, U., Chakrabarty, S., and Habets, E. A. P. (2020). Signal-aware broadband DOA estimation using attention mechanisms. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4930–4934.

Mandel, M. I., Ellis, D. P. W., and Jebara, T. (2006). An EM algorithm for localizing multiple sound sources in reverberant environments. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pages 953–960, Cambridge, MA, USA. MIT Press.

Mandel, M. I., Weiss, R. J., and Ellis, D. P. W. (2010). Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):382–394.

Marras, M., Marín-Reyes, P. A., Lorenzo-Navarro, J., Castrillón-Santana, M., and Fenu, G. (2020). Deep multi-biometric fusion for audio-visual user re-identification and verification. In De Marsico, M., Sanniti di Baja, G., and Fred, A., editors, *Pattern Recognition Applications and Methods*, Lecture Notes in Computer Science, pages 136–157, Cham. Springer International Publishing.

Martin, A., Charlet, D., and Mauuary, L. (2001). Robust speech/non-speech detection using LDA applied to MFCC. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, volume 1, pages 237–240.

May, T., Ma, N., and Brown, G. J. (2015). Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2679–2683.

May, T., Par, S. v. d., and Kohlrausch, A. (2011a). Binaural detection of speech sources in complex acoustic scenes. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 241–244.

May, T., Par, S. v. d., and Kohlrausch, A. (2011b). A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):1–13.

May, T., van de Par, S., and Kohlrausch, A. (2012). A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(7):2016–2030.

May, T., van de Par, S., and Kohlrausch, A. (2013). Binaural localization and detection of speakers in complex acoustic scenes. In Blauert, J., editor, *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, pages 397–425. Springer, Berlin, Heidelberg.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., and others (2005). The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88.

Ming, J., Hazen, T. J., Glass, J. R., and Reynolds, D. A. (2007). Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1711–1723.

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. (2017). Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5715–5725.

Murray, J. C. and Erwin, H. R. (2011). A neural network classifier for notch filter classification of sound-source elevation in a mobile robot. In *The 2011 International Joint Conference on Neural Networks*, pages 763–769.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620. ISCA.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 807–814.

Nakadai, K., Lourens, T., Okuno, H. G., and Kitano, H. (2000). Active audition for humanoid. In *AAAI/IAAI*, pages 832–839.

Nakadai, K., Matsuura, D., Okuno, H., and Kitano, H. (2003). Applying scattering theory to robot audition system: robust sound source localization and extraction. In *2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1147–1152, Las Vegas, NV, USA.

Nakamura, K., Nakadai, K., Asano, F., Hasegawa, Y., and Tsujino, H. (2009). Intelligent sound source localization for dynamic environments. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 664–669.

Nakamura, K., Nakadai, K., Asano, F., and Ince, G. (2011). Intelligent sound source localization and its application to multimodal human tracking. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 143–148.

Neti, C., Young, E. D., and Schneider, M. H. (1992). Neural network models of sound localization based on directional filtering by the pinna. *The Journal of the Acoustical Society of America*, 92(6):3140–3156.

Nguyen, T. T. N., Gan, W.-S., Ranjan, R., and Jones, D. L. (2020). Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1.

Okuno, H. G. and Nakadai, K. (2015). Robot audition: Its rise and perspectives. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5610–5614, South Brisbane, Queensland, Australia. IEEE.

Ozerov, A. and Fevotte, C. (2010). Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563.

Pak, J. and Shin, J. W. (2019). Sound localization based on phase difference enhancement using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1335–1345.

Palmieri, F., Datum, M., Shah, A., and Moiseff, A. (1991). Sound localization with a neural network trained with the multiple extended kalman algorithm. In *1991 International Joint Conference on Neural Networks (IJCNN)*, pages 125–131, Seattle.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Pang, C., Liu, H., and Li, X. (2019). Multitask learning of time-frequency CNN for sound source localization. *IEEE Access*, 7:40725–40737.

Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (2018). CRNN-based joint azimuth and elevation localization with the ambisonics intensity vector. In *IWAENC 2018 - 16th International Workshop on Acoustic Signal Enhancement*.

Pertila, P. and Parviainen, M. (2019). Time difference of arrival estimation of speech signals using deep neural networks with integrated time-frequency masking. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 436–440, Brighton, United Kingdom. IEEE.

Pertilä, P. and Cakir, E. (2017). Robust direction estimation with convolutional neural networks based steered response power. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6125–6129.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.

Rascon, C. and Meza, I. (2017). Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 96:184–210.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41.

Reynolds, D. A. and Rose, R. C. (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv:1706.05098 [cs, stat]*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Saxena, A. and Ng, A. Y. (2009). Learning sound location from a single microphone. In *2009 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1737–1742.

Schimmel, S. M., Muller, M. F., and Dillier, N. (2009). A fast and accurate “shoebox” room acoustics simulator. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 241–244.

Schissler, C. and Manocha, D. (2016). Interactive sound propagation and rendering for large multi-source scenes. *ACM Transactions on Graphics (TOG)*, 36(4).

Schmidt, R. (1986). Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280.

Sekiguchi, K., Bando, Y., Nugraha, A. A., Yoshii, K., and Kawahara, T. (2019). Semi-supervised multichannel speech enhancement with a deep speech prior. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2197–2212.

Shi, Y., Huang, Q., and Hain, T. (2020). Speaker re-identification with speaker dependent speech enhancement. *arXiv:2005.07818 [cs, eess]*.

Smaragdis, P. (1998). Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1):21–34.

Smaragdis, P. and Brown, J. (2003). Non-negative matrix factorization for polyphonic music transcription. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 177–180.

Smith, J. and Abel, J. (1987). Closed-form least-squares source location estimation from range-difference measurements. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(12):1661–1669.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., and Khudanpur, S. (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 165–170.

So, H. C., Chan, Y. T., and Chan, F. K. W. (2008). Closed-form formulae for time-difference-of-arrival estimation. *IEEE Transactions on Signal Processing*, 56(6):2614–2620.

Squartini, S., Principi, E., Rotili, R., and Piazza, F. (2012). Environmental robust speech and speaker recognition through multi-channel histogram equalization. *Neurocomputing*, 78(1):111–120.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., and Kawanabe, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1433–1440. Curran Associates, Inc.

Sun, B., Feng, J., and Saenko, K. (2016). Return of frustratingly easy domain adaptation. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147.

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

Takeda, R. and Komatani, K. (2016a). Discriminative multiple sound source localization based on deep neural networks using independent location model. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 603–609.

Takeda, R. and Komatani, K. (2016b). Sound source localization based on deep neural networks with directional activate function exploiting phase information. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 405–409.

Takeda, R. and Komatani, K. (2017). Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2217–2221.

Takeda, R., Kudo, Y., Takashima, K., Kitamura, Y., and Komatani, K. (2018). Unsupervised adaptation of neural networks for discriminative sound source localization with eliminative constraint. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3514–3518.

Tang, Z., Kanu, J. D., Hogan, K., and Manocha, D. (2019). Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks. In *Interspeech 2019*, pages 654–658. ISCA.

Thrun, S. (1995). Is learning the n-th thing any easier than learning the first? In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, pages 640–646, Cambridge, MA, USA. MIT Press.

Urruela, A. and Riba, J. (2004). Novel closed-form ML position estimator for hyperbolic location. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 149–152.

Valin, J. M., Michaud, F., Rouat, J., and Letourneau, D. (2003). Robust sound source localization using a microphone array on a mobile robot. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 2, pages 1228–1233.

Variani, E., Lei, X., McDermott, E., Moreno, I. L., and Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4052–4056.

Vecchiotti, P., Ma, N., Squartini, S., and Brown, G. J. (2019). End-to-end binaural sound localisation from the raw waveform. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 451–455, Brighton, United Kingdom.

Vecchiotti, P., Squartini, S., Principi, E., and Piazza, F. (2018). Deep neural networks for joint voice activity detection and speaker localization. In *2018 26th European Signal Processing Conference (EUSIPCO)*.

Vesperini, F., Vecchiotti, P., Principi, E., Squartini, S., and Piazza, F. (2018). Localizing speakers in multiple rooms by using deep neural networks. *Computer Speech & Language*, 49:83–106.

Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074.

Wakabayashi, M., Okuno, H. G., and Kumon, M. (2020). Multiple sound source position estimation by drone audition based on data association between sound source localization and identification. *IEEE Robotics and Automation Letters*, 5(2):782–789.

Wang, D. and Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726.

Wang, M. and Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Wang, Y., Shen, J., Petridis, S., and Pantic, M. (2019). A real-time and unsupervised face re-identification system for human-robot interaction. *Pattern Recognition Letters*, 128:559–568.

Warsitz, E. and Haeb-Umbach, R. (2007). Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1529–1539.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press.

Wierstorf, H., Geier, M., and Spors, S. (2011). A free database of head related impulse response measurements in the horizontal plane with multiple distances. In *Audio Engineering Society Convention 130*. Audio Engineering Society.

Xiao, X., Watanabe, S., Erdogan, H., Lu, L., Hershey, J., Seltzer, M. L., Chen, G., Zhang, Y., Mandel, M., and Yu, D. (2016). Deep beamforming networks for multi-channel speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5745–5749.

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

Xu, C., Xiao, X., Sun, S., Rao, W., Chng, E. S., and Li, H. (2017). Weighted spatial covariance matrix estimation for MUSIC based TDOA estimation of speech source. In *Interspeech 2017*, pages 1894–1898. ISCA.

Xue, W., Tong, Y., Zhang, C., and Ding, G. (2019). Multi-beam and multi-task learning for joint sound event detection and localization. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*.

Yalta, N., Nakadai, K., and Ogata, T. (2017). Sound source localization using deep learning models. *Journal of Robotics and Mechatronics*, 29(1):37–48.

Yilmaz, O. and Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc.

Youssef, K., Argentieri, S., and Zarader, J. L. (2013). A learning-based approach to robust binaural sound localization. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2927–2932.

Yves, G. and Yoshua, B. (2006). Entropy regularization. In Chapelle, O., Scholkopf, B., and Zien, A., editors, *Semi-Supervised Learning*, pages 151–168. The MIT Press.

Zegers, J. and Van hamme, H. (2016). Joint sound source separation and speaker recognition. In *Interspeech 2016*, pages 2228–2232.

Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. *arXiv:1212.5701 [cs]*.

Zhao, X., Shao, Y., and Wang, D. (2012). CASA-based robust speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1608–1616.

Zhao, X., Wang, Y., and Wang, D. (2014). Robust speaker identification in noisy and reverberant conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):836–845.

Zhou, D.-X. (2020). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*.

Weipeng He

Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland
https://idiap.ch/~whe · heweipeng@gmail.com

Research Interests

Audio and Speech Processing · Machine Learning · Robot Auditory Perception

Education

École Polytechnique Fédérale de Lausanne (EPFL) Switzerland
Ph.D. Candidate anticipated graduation in Dec. 2020
— Thesis title: *Deep Learning Approaches for Auditory Perception in Robotics*
— Supervisors: Dr. Jean-Marc Odobez and Dr. Petr Motlicek

University of Hamburg Germany
M.Sc. in *Intelligent Adaptive Systems*, GPA: 1.38/1.0 (excellent) September 2015

Tsinghua University China
B.E. in *Computer Science and Technology*, GPA: 90/100 (top 10%) July 2012
B.S. in *Pure and Applied Mathematics* (second major), GPA: 83/100

Experience

Idiap Research Institute Switzerland
Research Assistant June 2016 — Present

- Investigate deep learning-based approaches for sound source localization in real human robot interaction scenarios. Achieve more than 90% precision and recall for detecting and localizing multiple simultaneous sources in real-time.
- Investigate simultaneous localization and speech/non-speech classification of multiple sound sources. The joint multi-task approach significantly outperforms traditional methods those sequentially process localization and classification.
- Study data augmentation and domain adaptation for low-resource training of sound source localization neural networks.

Facebook USA
Research Intern July — October 2019

- Research on deep beamforming networks for far-field automatic speech recognition. Propose the spatial attention that improves the recognition rate by 9% on smart speakers.

6Estates Singapore
Software Engineer January — May 2016

- Develop back-end web interface of a deep learning based business insight information retrieval system.

University of Hamburg Germany
Research Assistant October 2012 — November 2015

- Research on interactive object recognition with audio-visual sensory fusion.

Hulu (Beijing) China
Research Intern July — September 2012

- Optimize search engine with fast approximate string matching.

Publications

[1] IEEE SLT 2021 Alpha-mini Speech Challenge: Open Datasets, Tracks, Rules and Baselines.
Yihui Fu, Weipeng He, et al.
in 2021 IEEE *Spoken Language Technology Workshop* (SLT)

[2] Spatial attention for far-field speech recognition with deep beamforming neural networks.
Weipeng He, Lu Lu, Biqiao Zhang, Jay Mahadeokar, Kaustubh Kalgaonkar, and Christian Fuegen,
in 2020 IEEE *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP)

[3] The MuMMER Data Set for Robot Perception in Multi-Party HRI Scenarios.
Olivier Canevet, Weipeng He, Petr Motlicek, and Jean-Marc Odobez,
in 2020 *29th IEEE International Conference on Robot and Human Interactive Communication* (RO-MAN)

[4] Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training.
Weipeng He, Petr Motlicek, and Jean-Marc Odobez,
in 2019 IEEE *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP)

[5] MuMMER: Socially intelligent human-robot interaction in public spaces.
Mary Ellen Foster, Weipeng He, et al.
in *Proc. AI-HRI 2019*

[6] Joint localization and classification of multiple sound sources using a multi-task neural network.
Weipeng He, Petr Motlicek, and Jean-Marc Odobez,
in *Proc. Interspeech* 2018 (best student paper award finalist)

[7] Deep neural networks for multiple speaker detection and localization.
Weipeng He, Petr Motlicek, and Jean-Marc Odobez,
in 2018 IEEE *International Conference on Robotics and Automation* (ICRA)

[8] Multimodal object recognition from visual and audio sequences.
Weipeng He, Haojun Guan, and Jianwei Zhang,
in 2015 IEEE *Int. Conf. on Multisensor Fusion and Information Integration for Intelligent Systems* (MFI)

[9] What to do first: the initial behavior in a multi-sensory household object recognition and categorization system.
Haojun Guan, Weipeng He, and Jianwei Zhang,
in 2014 IEEE *Int. Conf. on Multisensor Fusion and Information Integration for Intelligent Systems* (MFI)

[10] THUNLP at TAC KBP 2011 in entity linking.
Yu Zhao, Weipeng He, Zhiyuan Liu, and Maosong Sun,
in *Proceedings of TAC*, 2011

Peer Reviews

- Journal of Sound and Vibration (JSV), 2020
- European Conference on Computer Vision (ECCV), 2020
- IEEE International Conference on Robotics and Automation (ICRA), 2020
- IEEE Signal Processing Letters, 2019
- ACM International Conference on Multimodal Interaction (ICMI), 2018

Other Research Activities

Challenges:	Co-organizer of the IEEE SLT 2021 Alpha-mini Speech Challenge.
Datasets:	Creator of the SSLR dataset and MuMMER dataset.

Technical Skills

Programming:	Python · C · C++ · Java · MATLAB
Libraries:	PyTorch · Gstreamer · OpenCV · GSL · OpenMP · OpenMPI · Spark
Other Tools:	Vim · Bash · LaTeX · GDB · Git · Mercurial · Gnuplot

Languages

English:	Fluent
Chinese:	Native Speaker
German:	Basic
French:	Basic

Miscellaneous

GitHub:	github.com/hwp (highlight projects: apkit, notGHMM)