DEEP LEARNING APPROACHES FOR BINAURAL SPEAKER LOCALIZATION INCORPORATING AUDITORY-INSPIRED PERIODICITY FEATURES

Von der Fakultät für Medizin und Gesundheitswissenschaften
der Carl-von-Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
angenommene Dissertation

von
Reza Varzandeh
geboren am 13. August 1985
in Ahvaz, Iran

Reza Varzandeh: *Deep Learning Approaches for Binaural Speaker Localization Incorporating Auditory-Inspired Periodicity Features*

Erstgutachter:
Prof. Dr. Volker Hohmann, University of Oldenburg, Germany

Weitere Gutachter:
Prof. Dr. ir. Simon Doclo, University of Oldenburg, Germany
Dr. Ning Ma, University of Sheffield, UK

Tag der Disputation:
27. Februar 2025

Acknowledgments

This thesis, written in the Auditory Signal Processing and Signal Processing groups at the Department of Medical Physics and Acoustics of the University of Oldenburg, would not have been possible without the invaluable support of many individuals. I extend my genuine gratitude to every one of them.

First and foremost, I express my sincere gratitude to my supervisors, Prof. Volker Hohmann and Prof. Simon Doclo. Their invaluable guidance and insights have been fundamental in shaping the direction of this work. Working with both of them has been a privilege and an inspiration. I am profoundly grateful for their thoughtful feedback and detailed suggestions, which have consistently raised the quality of my work. Their belief in my potential and support throughout this journey has been instrumental. I am immensely thankful to Prof. Hohmann for trusting me and allowing me to work on my PhD at the University of Oldenburg.

My research journey in Oldenburg began with the Research and Development Group at HörTech (Hörzentrum Oldenburg), where I had the pleasure of working alongside talented colleagues under the leadership of Jörg-Hendrik Bach. I am incredibly grateful to Kamil Adiloglu and Hendrik Kayser for their unreserved support and insightful discussions throughout our fruitful projects.

I am deeply grateful to all my past and present colleagues in the Auditory Signal Processing and Signal Processing groups for their support and encouragement. My sincere thanks go to Hendrik Kayser, Ali Aroudi, Joanna Luberadzka, Gerard Llorach, Marvin Tammen, Maartje Hendrikse, Jürgen Otten, Giso Grimm, Nico Gößling, Piero Rivera Benois, Klaus Brümann, Henri Gode and Wiebke Middelberg for their constructive discussions and feedback, making my time both productive and memorable. I would also like to sincerely thank all the academic members and administrative staff of the Department of Medical Physics and Acoustics and the Collaborative Research Centre SFB 1330, especially Nicole Kulbach and Karin Klink.

Finally, I am deeply indebted to my family for their everlasting support and love. To my parents, who instilled in me the value of perseverance and the importance of education, thank you for your sacrifices and constant encouragement. To my brother, whose support has been a source of strength, and to my wonderful wife, whose presence, patience, and boundless support have carried me through the most challenging moments.

Oldenburg, February 2025
Reza Varzandeh

Summary

Individuals with hearing loss struggle with speech understanding and sound localization, which significantly impacts their ability to communicate and navigate noisy environments. Hearing aids can incorporate sound source localization techniques that often estimate the sound source direction of arrival (DOA). Accurate DOA estimation allows hearing aids to better separate target sounds from noise, improving speech intelligibility and listening comfort through techniques such as spatial filtering. However, existing methods, particularly those relying merely on binaural spatial cues, face challenges in real-world acoustic scenarios. These challenges arise from noise, reverberation, and multiple competing sound sources, which degrade the accuracy of DOA estimation algorithms. Speech inactivity also poses a significant challenge to DOA estimation systems, leading to unreliable estimates during silent periods. Deep neural networks (DNNs) can implicitly capture the complexities of sound propagation and diverse acoustic environments, leading to more robust and accurate DOA estimates. This enhanced accuracy can ultimately result in better hearing aid performance and an improved listening experience for individuals with hearing loss.

The human auditory system effectively analyzes complex acoustic scenes, segregating sound sources based on features such as periodicity and harmonicity, in addition to spatial cues. Inspired by the intricate mechanism of the human auditory system and potential of DNNs, the primary objective of this thesis is to improve DOA estimation performance in noisy and reverberant environments by leveraging auditory-inspired periodicity features in combination with binaural spatial features through DNN-based approaches. This thesis investigates DOA estimation in binaural hearing aids, focusing on both single-talker and multi-talker scenarios.

First, this thesis proposes a novel DNN-based approach for robust DOA estimation of a single talker in noisy and reverberant environments. The proposed method, referred to as speech-aware DOA estimation, eliminates the need for a separate voice activity detector (VAD) by directly estimating DOA upon speech detection that is merely conditioned on the DNN output. Novel feature combinations, consisting of spatial and periodicity features, are utilized within dual-path convolutional neural network (CNN) architectures for speech-aware DOA estimation. In these architectures, each feature is processed separately through parallel convolutional pathways. These feature combinations include conventional spatial features, such as broadband GCC-PHAT or narrowband cross-power spectrum (CPS), alongside a novel periodicity feature termed periodicity degree (PD). These PD features derived from subband-averaged periodicity analysis offer varying frequency selectivity. Evaluations are conducted for static and dynamic single-talker scenarios in various reverberant environments with different signal-to-noise ratios (SNRs) and in the presence of simulated and recorded background noise. Baseline systems consist of a CNN that utilizes a single spatial feature type, and cascaded with a pitch-based VAD. The results show that the proposed systems, using periodicity features in different feature combinations, consistently outperform baseline systems, regardless of the spatial feature type. Evaluation results demonstrate that incorporating PD features significantly enhances DOA estimation accuracy and reduces angular error, especially in adverse SNR conditions. The benefit of the speech-aware approach is highlighted by its reliable speech detection for the DOA estima-

tion task. The proposed approach effectively establishes the benefit of incorporating periodicity features for speech-aware DOA estimation.

Second, building upon the success of the speech-aware approach by integrating spatial and periodicity features, this thesis proposes to directly combine real and imaginary components of narrowband CPS features and PD features within the input of a CNN, eliminating the need for separate branches. In particular, a novel and computationally efficient two-stage CNN architecture for robust single-talker DOA estimation is proposed. A key innovation is the introduction of a feature reduction stage employing 1×1 convolutions to address the sparsity and high dimensionality of the narrowband PD features. This stage extracts compact PD saliency features, facilitating spectro-temporal processing within the CNN and leading to a more efficient model. Evaluation results for static scenarios across various SNR conditions, in the presence of recorded background noise or non-speech directional interference within reverberant environments, demonstrate significant improvements. The two-stage CNN outperforms a baseline system using only CPS features and a pitch-based VAD, with a comparable number of trainable parameters. Additionally, it surpasses a system using a direct feature combination strategy without feature reduction, while requiring significantly fewer parameters. This highlights the computational efficiency and effectiveness of the proposed two-stage architecture for practical applications.

Finally, the thesis extends the benefits of integrating periodicity and spatial features to multi-talker DOA estimation. A two-stage CNN, specifically adapted for the multi-talker task, is proposed, leveraging a combination of narrowband PD features and CPS phase as input. Various architectural choices are investigated for capturing temporal and spectro-temporal dependencies within the two-stage CNN. The findings reveal that processing each frequency band independently while capturing temporal dependencies within convolutional blocks yields the most effective configuration for multi-talker DOA estimation. This configuration offers the best performance with the lowest computational complexity. An evaluation in static multi-talker scenarios with two and three simultaneous speakers, in the presence of recorded background noise across different reverberant environments and SNRs, demonstrates the effectiveness of the proposed approach. The proposed two-stage CNN consistently outperforms baseline systems relying solely on the CPS phase or combined with magnitude spectrograms. This shows the benefit of integrating periodicity and spatial features in multi-talker scenarios, highlighting the adaptability of this approach to increasingly complex acoustic environments. The results of this thesis provide the basis for more effective DOA estimation in complex auditory scenarios. They highlight the value of periodicity features and optimized network design in binaural signal processing and provide new directions for future research in speech-related applications.

Zusammenfassung

Individuen mit Hörverlust haben Schwierigkeiten beim Sprachverständnis und bei der Schallokalisierung, was ihre Fähigkeit zur Kommunikation und Orientierung in lauten Umgebungen erheblich beeinträchtigt. Hörgeräte können Techniken zur Lokalisierung von Schallquellen einbeziehen, die häufig die Einfallsrichtung (DOA) der Schallquelle schätzen. Eine genaue DOA-Schätzung ermöglicht es Hörgeräten, Zielgeräusche von Hintergrundgeräuschen besser zu trennen und so durch Techniken wie räumliche Filterung die Sprachverständlichkeit und den Hörkomfort zu verbessern. Bestehende Methoden, insbesondere solche, die lediglich auf binauralen räumlichen Merkmalen beruhen, stoßen jedoch in realen akustischen Szenarien auf Herausforderungen. Diese Herausforderungen entstehen durch Rauschen, Nachhall und mehrere konkurrierende Schallquellen, die die Genauigkeit der DOA-Schätzalgorithmen beeinträchtigen. Auch Sprachpausen stellen eine große Herausforderung für DOA-Schätzsysteme dar, da sie während stiller Perioden zu unzuverlässigen Schätzungen führen. Tiefe neuronale Netze (DNNs) können implizit die Komplexität der Schallausbreitung und verschiedene akustische Umgebungen erfassen, was zu robusteren und genaueren DOA-Schätzungen führt. Diese verbesserte Genauigkeit kann letztendlich zu einer besseren Hörgeräteleistung und einem besseren Hörerlebnis für Menschen mit Hörverlust führen.

Das menschliche Gehör analysiert effektiv komplexe Hörszenen und trennt Schallquellen anhand von Merkmalen wie Periodizität und Harmonizität sowie räumlichen Hinweisen. Inspiriert durch den komplexen Mechanismus des menschlichen Gehörs und dem Potenzial von DNNs, besteht das Hauptziel dieser Arbeit darin, die Leistung der DOA-Schätzung in lauten und halligen Umgebungen zu verbessern, indem gehörinspirierte Periodizitätsmerkmale in Kombination mit binauralen räumlichen Merkmalen durch DNN-basierte Ansätze genutzt werden. In dieser Arbeit wird die DOA-Schätzung in binauralen Hörgeräten untersucht, wobei sowohl Szenarien mit einem Sprecher als auch mit mehreren Sprechern im Mittelpunkt stehen.

Zunächst wird in dieser Dissertation ein neuartiger DNN-basierter Ansatz für die robuste DOA-Schätzung eines einzelnen Sprechers in verrauschten und halligen Umgebungen vorgeschlagen. Die vorgeschlagene Methode, die als sprach-fokussierte DOA-Schätzung bezeichnet wird, macht einen separaten Sprachaktivitätsdetektor (VAD) überflüssig, indem die Einfallsrichtung direkt bei der Sprachdetektion geschätzt wird, die lediglich von der Ausgabe des DNN abhängig ist. Neue Merkmalskombinationen, bestehend aus räumlichen und Periodizitätsmerkmalen, werden in Dual-Pfad-Convolutional-Neural-Network (CNN)-Architekturen für die sprach-fokussierte DOA-Schätzung verwendet. In diesen Architekturen wird jedes Merkmal separat durch parallele Faltungswege verarbeitet. Diese Merkmalskombinationen umfassen konventionelle räumliche Merkmale wie Breitband-GCC-PHAT oder schmalbandiges Cross-Power-Spectrum (CPS) sowie ein neuartiges Periodizitätsmerkmal, genannt Periodizitätsgrad (PD). Diese PD-Merkmale, abgeleitet aus einer Frequenzband-gemittelten Periodizitätsanalyse, bieten eine variable Frequenzselektivität. Evaluationen wurden für statische und dynamische Einzelsprecherszenarien in verschiedenen halligen Umgebungen mit variierenden Signal-Rausch-Verhältnissen (SNRs) in Anwesenheit von simuliertem und aufgezeichnetem Hintergrundrauschen durchgeführt. Basissysteme bestehen aus einem CNN, das einen räumlichen Merkmalstyp verwendet und mit einer ton-

höhenbasierten VAD kaskadiert ist. Die Ergebnisse zeigen, dass die vorgeschlagenen Systeme, die Periodizitätsmerkmale in verschiedenen Merkmalskombinationen verwenden, die Basissysteme konsequent übertreffen, unabhängig vom räumlichen Merkmalstyp. Die Evaluationsergebnisse zeigen, dass die Einbeziehung von PD-Merkmalen die Genauigkeit der DOA-Schätzung erheblich verbessert und den Winkelabweichungsfehler verringert, insbesondere bei schwierigen SNR-Bedingungen. Der Vorteil des sprach-fokussierten Ansatzes wird durch seine zuverlässige Sprachdetektion für die DOA-Schätzung hervorgehoben. Der vorgeschlagene Ansatz verdeutlicht die Vorteile der Integration von Periodizitätsmerkmalen für die sprach-fokussierte DOA-Schätzung.

Zweitens, aufbauend auf dem Erfolg des sprach-fokussierten Ansatzes durch Integration von räumlichen und Periodizitätsmerkmalen, schlägt die Dissertation vor, die realen und imaginären Komponenten der schmalbandigen CPS-Merkmale und PD-Merkmale direkt im Eingabebereich eines CNNs zu kombinieren und separate Pfade zu vermeiden. Insbesondere wird eine neuartige und recheneffiziente zweistufige CNN-Architektur für die robuste DOA-Schätzung eines einzelnen Sprechers vorgeschlagen. Eine zentrale Innovation ist die Einführung einer Merkmalsreduktionsebene mittels 1×1 Faltungen, um die Spärlichkeit und hohe Dimensionalität der schmalbandigen PD-Merkmale zu bewältigen. Diese Stufe extrahiert kompakte PD-Salienzmerkmale, die die spektral-temporale Verarbeitung innerhalb des CNN erleichtern und zu einem effizienteren Modell führen. Bewertungsergebnisse für statische Szenarien unter verschiedenen SNR-Bedingungen, in Anwesenheit von aufgezeichnetem Hintergrundrauschen oder nichtsprachlicher, direktionaler Störung in halligen Umgebungen, zeigen erhebliche Verbesserungen. Das zweistufige CNN übertrifft ein Basissystem, das nur CPS-Merkmale und eine tonhöhenbasierte VAD verwendet, bei einer vergleichbaren Anzahl von trainierbaren Parametern. Zusätzlich übertrifft es ein System, das eine direkte Merkmalskombinationsstrategie ohne Merkmalsreduktion nutzt, während es deutlich weniger Parameter benötigt. Dies unterstreicht die Recheneffizienz und Wirksamkeit der vorgeschlagenen zweistufigen Architektur für praktische Anwendungen.

Schließlich erweitert die Dissertation die Vorteile der Integration von Periodizitäts- und räumlichen Merkmalen auf die DOA-Schätzung bei mehreren Sprechern. Es wird ein zweistufiges CNN vorgeschlagen, speziell angepasst für die Mehrsprecher-Aufgabe, das eine Kombination aus Schmalband-PD-Merkmalen und CPS-Phase als Eingabe nutzt. Es werden verschiedene architektonische Möglichkeiten untersucht, um zeitliche und spektral-temporale Abhängigkeiten innerhalb des zweistufigen CNN zu erfassen. Die Ergebnisse zeigen, dass die Verarbeitung jedes Frequenzbands unabhängig voneinander, während zeitliche Abhängigkeiten innerhalb von konvolutionalen Blöcken erfasst werden, die effektivste Konfiguration für die DOA-Schätzung bei mehreren Sprechern darstellt. Diese Konfiguration bietet die beste Leistung bei geringster Rechenkomplexität. Eine Evaluation in statischen Mehrsprecherszenarien mit zwei und drei gleichzeitigen Sprechern, in Anwesenheit von aufgezeichnetem Hintergrundrauschen in verschiedenen halligen Umgebungen und bei unterschiedlichen SNRs, zeigt die Wirksamkeit des vorgeschlagenen Ansatzes. Das vorgeschlagene zweistufige CNN übertrifft konsistent Basissysteme, die sich ausschließlich auf die CPS-Phase oder in Kombination mit Magnitudenspektren stützen. Dies zeigt den Vorteil der Integration von Periodizität und räumlichen Merkmalen in Szenarien mit mehreren Sprechern und unterstreicht die Anpassungsfähigkeit dieses Ansatzes an zunehmend komplexe akustische Umgebungen. Die Ergebnisse dieser Dissertation bilden die Grundlage für eine effektivere DOA-Schätzung in komplexen Hörszenarien. Sie unterstreichen den Vorteil von Periodizitätsmerkmalen und optimiertem Netzwerkdesign in der binauralen Signalverarbeitung und geben neue Richtungen für zukünftige Forschung in sprachbezogenen Anwendungen vor.

Acronyms

2D two-dimensional

ACF auto-correlation function

ASA auditory scene analysis

ATF acoustic transfer function

BRIR binaural room impulse response

CASA computational auditory scene analysis

CCF cross-correlation function

CNN convolutional neural network

CPS cross-power spectrum

DNN deep neural network

DOA direction of arrival

ESPRIT estimation of signal parameters via rotational-invariance techniques

GCC generalized cross-correlation

GCC-PHAT generalized cross-correlation with phase transform

GTFB gammatone filter bank

HRIR head-related impulse response

HRTF head-related transfer function

IIR infinite impulse response

ILD interaural level difference

IPD interaural phase difference

ITD interaural time difference

MAE mean absolute error

ML maximum likelihood

MUSIC multiple signal classification

PD periodicity degree

ReLU rectified linear unit

RNN recurrent neural network

RTF relative transfer function

SACF summary auto-correlation function

SNR signal-to-noise ratio

SRP-PHAT steered response power with phase transform

STFT short-time Fourier transform

TDOA time difference of arrival

VAD voice activity detector

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The human auditory system excels in analyzing complex auditory scenes, including multiple talkers, interference, and background noise, often referred to as cocktail party scenarios [2, 3]. This includes accurately localizing sound sources, a process crucial for communication, spatial awareness, and navigating dynamic environments. However, individuals with hearing loss often experience significant difficulties in sound localization, which impacts their quality of life, leading to challenges in social interaction and environmental awareness. Hearing aid technology enhanced by computational auditory scene analysis (CASA) [4] can mitigate this effect [5]. CASA aims to replicate the human ability to identify and segregate individual sound sources within complex acoustic environments using computational models. Sound source localization plays a fundamental role in the CASA systems, allowing binaural hearing aids to pinpoint the location from which individual sound sources originate relative to the listener. Instead of locating sound sources in three-dimensional space, binaural hearing aid algorithms often rely on direction of arrival (DOA) information [5, 6]. By accurate DOA estimation, hearing aids can implement advanced beamforming techniques to enhance speech intelligibility in noisy environments [7, 8]. This allows the devices to focus on desired speakers while suppressing background noise and competing talkers. Additionally, accurate DOA estimation facilitates better spatial awareness for hearing aid users, helping them localize sounds and navigate their auditory environment more effectively.

Most traditional binaural DOA estimation methods leverage the natural sound processing capabilities of the human auditory system by analyzing the spatial features that capture the differences in sound signals arriving at the left and right ears, namely the interaural time difference (ITD) and interaural level difference (ILD) [9, 10]. These methods include approaches comparing estimated binaural features with pre-computed templates obtained from head-related transfer functions (HRTFs), which describe how the listener's head and torso filter sound. Another approach employs relative transfer function (RTF) vectors, which capture the relationship between sound arriving at different microphones on a listener's head [8, 11]. Databases of either HRTFs or RTFs, covering a range of potential source directions, are used for this comparison. However, these methods face significant challenges in real-world acoustic environments. Background noise, interference, and reverberation introduce uncertainties into binaural cues. These uncertainties cause the extracted features from microphone signals to deviate from the pre-computed templates, leading to errors in DOA estimation.

Artificial intelligence, particularly deep neural network (DNN), has driven significant advancements in various domains, including DOA estimation [12]. Common input features for deep learning-based binaural DOA estimation include the ILD [13, 14], ITD [13], interaural phase difference (IPD) [15], the cross-correlation function (CCF) [14, 16], and the generalized cross-correlation with phase transform (GCC-PHAT) [17]. DNN-based techniques often directly map these features to the sound source DOA [14, 16, 18], while others employ a two-step approach, first enhancing signal features [19, 20]. In real-life scenarios, speech inactivity, particularly in the presence of background noise or non-speech interference, challenges DOA estimation systems, leading to unreliable results [21]. A common solution is to use a voice activity detector

(VAD) with the DOA estimator [21–24]. However, separate VADs often require manual tuning and can introduce errors, limiting overall system performance.

The auditory scene analysis (ASA) framework posits that the human auditory system groups signal components based on features like harmonicity and periodicity, and segregates them using spatial cues such as ITDs [4, 25]. However, while binaural spatial features are widely used in sound localization and DOA estimation, especially within the DNN-based approaches, the integration of periodicity and harmonicity remains largely unexplored in this context. This gap highlights the potential for innovative approaches that combine these features to enhance DOA estimation performance.

The primary objective of this thesis is to leverage the power of deep learning to tackle the challenge of binaural DOA estimation for speech signals in noisy and reverberant environments. Inspired by the intricate mechanisms of the human auditory system, this thesis **explores innovative DNN-based approaches for integrating auditory-inspired periodicity features with binaural spatial features, aiming to enhance DOA estimation performance in binaural hearing aids for both single- and multi-talker scenarios, without the need for a separate VAD**.

The remainder of this chapter presents a foundation for understanding auditory-inspired periodicity features and binaural localization. These concepts set the stage for understanding the contributions of this thesis. The remainder of this chapter is organized as follows. Section 1.2 presents a concise overview of the human auditory system, auditory scene analysis, and binaural sound localization. Section 1.3 explores the roles of periodicity and harmonicity in human auditory scene analysis and CASA systems and investigates various periodicity analysis methods. It also discusses classical CASA systems incorporating periodicity and harmonicity in conjunction with other cues, such as spatial cues, for sound separation and localization. Section 1.4 delves into the fundamentals of binaural DOA estimation, encompassing various aspects such as acoustic scenarios, HRTFs, and traditional DOA estimation approaches. Section 1.5 presents insights into deep learning techniques for DOA estimation, covering most important design choices made in this thesis, including learning strategies, output coding, input features, and network architectures. Section 1.6 highlights issues for DOA estimation, particularly those arising from speech inactivity. It discusses the limitations of common DNN-based approaches and explores the potential benefits of integrating auditory-inspired periodicity features for improved robustness. Finally, Section 1.7 outlines the challenges addressed and the main contributions of this thesis.

1.2 Human auditory system and sound source localization

Understanding the mechanisms of the auditory system and sound source localization is essential for advancements in areas such as hearing aids, audio signal processing, and artificial intelligence systems that mimic human hearing. This section provides a concise overview of the human auditory system's structure and function, the principles of auditory scene analysis, and the mechanisms underlying binaural sound localization.

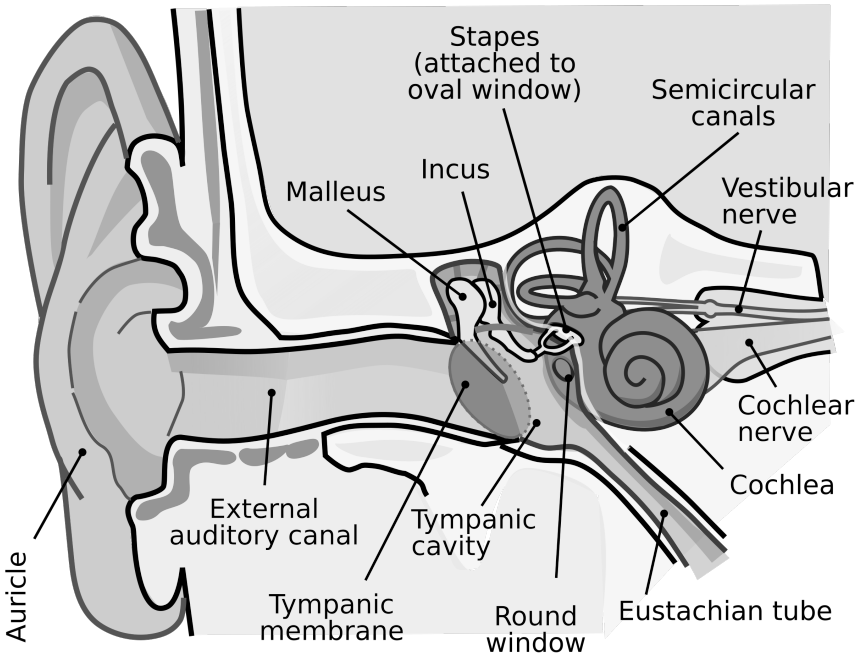


Figure 1·1: The peripheral auditory system. (Adapted from Lars Chittka; Axel Brockmann, Perception Space - The Final Frontier, A PLoS Biology Vol. 3, No. 4, e137, https://commons.wikimedia.org/w/index.php?curid=5957984. Modified from the original.)

1.2.1 Introduction to the human auditory system and auditory scene analysis

The human auditory system can be broadly divided into two main components: mechanical processing, also known as the peripheral auditory system, and neural processing [26]. Mechanical processing converts sound waves into electrical signals that the brain can interpret. Neural processing involves the network of neural pathways and brain regions responsible for analyzing, interpreting, and responding to auditory information.

Fig. 1·1 depicts the peripheral auditory system of the human ear [27]. In the peripheral auditory system, sound waves enter through the outer ear, where the unique shape of the pinna (or auricle) helps filter frequencies based on their direction, providing cues for sound localization. These sound waves travel through the outer ear canal and reach the eardrum (tympanic membrane), causing it to vibrate. These vibrations are transmitted through the three tiny middle ear bones, i.e., the malleus, incus, and stapes. This chain of bones acts as an impedance-matching system, transferring sound energy from the air-filled outer ear to the fluid-filled inner ear. Vibrations then reach the cochlea, a spiral-shaped organ in the inner ear containing the basilar membrane, whose stiffness varies along its length. As a result, different frequencies cause specific areas of the basilar membrane to vibrate, high frequencies near the base and low frequencies near the cochlea apex [26]. Hair cells on the basilar membrane convert the mechanical vibrations into electrical signals, which are sent to the brain via the auditory nerve.

Through neural processing, the auditory nerve transmits electrical signals from the cochlea

to the brainstem [26]. In the brainstem, these signals are processed in specific areas that extract different spatial and spectral features of sound. Certain regions process ITDs, allowing for the determination of timing differences between the ears. Others process ILDs, allowing to detect differences in sound intensity. Additional structures analyze spectral cues contributing to sound localization and segregation [28]. The midbrain integrates information from the brainstem, combining ITD, ILD, and spectral cues to refine the representation of auditory space. Higher-level auditory processing occurs in the auditory cortex, where features are integrated to create a unified spatial representation of sound sources. This includes stream segregation, which separates sounds from different sources based on acoustic features like pitch, timbre, and spatial location, as well as object recognition, which involves identifying and classifying sound sources using learned patterns [4].

ASA is a fundamental mechanism by which the human auditory system interprets acoustic signals to distinguish between different sound sources in the environment. Based on Bregman's conceptual framework for ASA [25], the human auditory system interprets acoustic signals using auditory features such as temporal continuity of harmonics and formants, periodicity and harmonicity of voiced speech, onset and offset synchrony, and spatial location. CASA [4] extends this concept into computational models, aiming to mimic the human auditory system's ability in machine listening systems such as hearing aids to analyze complex auditory scenes.

A key aspect of both the human auditory system and CASA systems is binaural sound source localization [29]. This process utilizes the difference in signals received at the two ears, known as binaural cues, to determine the direction and distance of sound sources.

1.2.2 Binaural sound localization

The human auditory system primarily employs two physical cues, ITD and ILD, to determine the location of sound sources [30]. These cues work in a complementary fashion, with ITD being most effective at lower frequencies and ILD taking precedence at higher frequencies [30, 31]. ITD arises from the time difference of arrival (TDOA) of a sound between the left and right ears, caused by the varying distance the sound must travel to each ear. This difference depends on the listener's head size. ITDs are assumed to be the predominant cue for binaural localization and exist across all frequencies, but they are most helpful in localizing sound components below approximately 1.5 kHz [32–34]. ITD is represented spectrally as the IPD.

ITD cues alone can be ambiguous, resulting in multiple potential source locations known as the *cone of confusion* [33]. The human auditory system leverages the complex shape of the external ears (pinnae) to resolve this ambiguity [33, 35]. Pinnae introduces spectral modifications to sound signals, mainly aiding in vertical localization (elevation) and resolving front-back ambiguities.

ILD occurs due to the *shadowing* effect of the head, which reduces the sound energy reaching the ear more distant from the sound source. This effect results from sound wave diffraction and reflection, dependent on the relationship between the sound's wavelength and the head's size [31]. Diffraction occurs when the sound's wavelength is larger than the head's diameter, causing the sound waves to bend around the head. Reflection happens when the sound's wavelength is

smaller than the head's diameter, leading to sound waves being reflected or backscattered. This reflection creates a distinct shadow zone behind the head, leading to ILD cues. ILDs are most pronounced at frequencies above approximately 1.5 kHz, where the reflection of sound waves at these frequencies is substantial.

1.3 Periodicity and harmonicity

Periodicity and harmonicity are crucial cues in human auditory scene analysis. These cues help the auditory system to segregate concurrent sounds and organize them into meaningful auditory objects or streams [4, 26, 36]. This section introduces the fundamental concepts of periodicity and harmonicity and their importance in human auditory scene analysis. We will detail the computational approaches CASA systems use to analyze periodicity and harmonicity. We also explain how CASA systems integrate them with other cues, particularly spatial information like ITD, to perform sound source separation and localization tasks.

1.3.1 Periodicity and harmonicity in auditory scene analysis

Periodicity is the quality of a signal that repeats itself over time. The repetition rate is called the fundamental frequency (or pitch), and its inverse is the fundamental period. While pitch and fundamental frequency are often used interchangeably, and pitch detection algorithms commonly refer to fundamental frequency estimation methods, they are not strictly equivalent [37]. For simple harmonic sounds, like pure tones, pitch, a perceptual attribute, is directly related to the fundamental frequency. However, for more complex sounds, such as those produced by musical instruments or the human voice, this relationship becomes less straightforward. In the context of speech, pitch perception helps listeners identify individual voices by grouping harmonics that are integer multiples of the fundamental frequency [26].

Harmonicity, closely related to periodicity, describes the harmonic structure of sounds where frequencies are integer multiples of a fundamental frequency. The regular spacing of these harmonics facilitates perceptual fusion, allowing us to perceive them as a single sound. Conversely, components that do not align with the harmonic series of the fundamental frequency are more likely to be perceived as separate sounds [26]. The human auditory system leverages this harmonic regularity to segregate concurrent sounds by identifying sets of harmonically related components within mixtures of frequencies [4]. For instance, if most frequency components can be grouped into two subsets with different fundamental frequencies, the auditory system interprets these as two distinct sound sources.

1.3.2 Periodicity and harmonicity analysis in CASA systems

CASA systems often employ a peripheral analysis model that simulates the frequency selectivity of the cochlea in the human ear using filter banks, such as the gammatone filter bank (GTFB) [38, 39]. This model incorporates auditory filters that are spaced based on the equivalent rectangular bandwidth (ERB), simulating how sounds are split into different frequency subbands for processing. This bank of filters mimics the frequency response of the basilar membrane, de-

composing the incoming audio signal into separate frequency subbands [38, 40–42]. The output of each filter is then processed to simulate the activity of inner hair cells, including rectification, compression, and low-pass filtering [10, 29, 36]. They then extract features e.g., related to the periodicity and harmonicity [43–48].

Various algorithms detect and analyze pitch, periodicity, and harmonicity [39]. A common method involves computing the auto-correlation function (ACF) or normalized ACF in each frequency subband [49, 50]. The resulting correlogram (or the normalized correlogram), a two-dimensional representation of ACFs in all subbands [51], facilitates periodicity analysis. Aggregating this correlogram across subbands yields a summary correlogram or summary auto-correlation function (SACF), revealing multiple periodicities in the signal [50–53].

The bandwidth of gammatone filters increases with frequency, mirroring the human auditory system but posing challenges for processing high-frequency harmonics in CASA systems [4, 39]. At low frequencies (below 1 kHz), narrow filters resolve individual harmonics. In contrast, at higher frequencies (above 1.5 kHz), broader filters capture multiple harmonics together, resulting in an amplitude-modulated signal that fluctuates at the fundamental frequency [54]. To address this, algorithms often extract the envelope of high-frequency signals to capture these amplitude fluctuations [43, 48, 53, 55–57].

To improve multi-pitch and periodicity analysis, methods like Tolonen and Karjalainen's model compute generalized ACFs for low-frequency signals and high-frequency envelopes, combining them into an enhanced SACF [43]. Such multi-band approaches are more noise-robust, adapting to periodicity in different frequencies and weighting or discarding those dominated by noise before combining the remaining ACFs [53, 55]. Advanced periodicity analysis methods, such as Tan and Alwan's multi-band summary correlogram (MBSC) algorithm, enhance pitch detection by emphasizing harmonic structures using comb filters and weighting ACFs based on the Harmonic-to-Subharmonic Ratio (HSR) [57]. Similarly, Chen and Hohmann introduced the periodicity degree (PD) by combining normalized auto-correlation (NAC), similar to the normalized ACF, and comb filter ratio (CFR) to robustly measure signal periodicity for speech enhancement [48]. The PD captures the dominance of harmonic components, distinguishing voiced speech from aperiodic components even in noisy conditions.

Some periodicity analysis methods [39, 45–47] leverage the *glimpsing* model of human speech perception, which suggests that perception relies on time-frequency segments with high local signal-to-noise ratio (SNR). By identifying *periodicity glimpses*, spectro-temporal bins with high periodicity, these methods can isolate segments where a single sound source dominates, providing a footprint of speech within complex mixtures [58, 59]. Josupeit and Hohmann utilized the normalized synchrogram, representing the ratio of harmonic energy to total energy in each time-frequency bin, to analyze the periodic structure of signals [46, 47]. This technique identifies prominent periodicities corresponding to fundamental periods, aiding in detecting periodicity glimpses in multi-talker environments for tasks like speech localization and recognition. Building on this, Luberadzka et al. integrated sparse periodicity auditory features within a predictive model to simulate human selective attention, effectively separating target speech from background noise in complex acoustic scenes [45].

1.3.3 Integrating periodicity and harmonicity with spatial cues

CASA systems leverage periodicity and harmonicity as primary features for sound source separation and segregation [36, 43, 44, 60–62]. Similar to the human auditory system, CASA systems benefit from combining periodicity and harmonicity with other auditory cues, such as spatial cues from multiple microphones. This integration of multiple cues leads to more robust and accurate sound source separation.

Darwin and Hukin [63] explored the role of ITD and frequency proximity in segregating a harmonic from a vowel. Their findings suggest that while ITD is a weak cue on its own, it can become more effective when combined with frequency-based cues. In contrast, frequency proximity plays a dominant role, as harmonically related components close in frequency tend to be perceptually fused into a single sound. The study highlights how harmonicity and regular frequency spacing help the auditory system group components, enhancing the segregation of concurrent sounds. Similarly, [64] investigated whether listeners could group sound components from different frequency ranges using ITD to help segregate concurrent speech sounds. Their results also demonstrate that listeners cannot use ITD alone to group frequency components across different subbands.

Both studies show that ITD works more effectively when combined with other auditory cues. For localization tasks, this means that CASA systems may benefit from a combination of spatial and spectral cues, e.g., when time-delay information such as ITD and spectral cues such as periodicity or harmonicity are available together. While the periodicity and harmonicity of speech have been extensively studied for source segregation and separation, a few studies have explored their benefits for sound localization [46, 65–67].

Brandstein introduced a method that leverages the harmonic structure in voiced speech segments [65]. This approach uses a modified generalized cross-correlation (GCC) weighting scheme based on the periodicity of speech harmonics. Christensen et al. [66] and Woodruff et al. [67] emphasized pitch-based grouping as a critical step for improving localization by isolating spectro-temporal regions or fragments dominated by single sources. While these method shows improved robustness to noise and reverberation compared to traditional approaches, their performance is inherently tied to the accuracy of the underlying pitch estimation. Josupeit and Hohmann [46] showed that combining ITD with periodicity-based glimpses improves speech localization in multi-talker environments, highlighting periodicity as a critical feature for segregating sounds in complex auditory scenes.

Despite the advancements in utilizing periodicity and harmonicity cues for different tasks, including sound localization, the existing approaches face limitations in complex acoustic environments. These methods often rely heavily on manually tuned thresholds and heuristics to identify and extract periodicity glimpses and accurately estimate pitch information. This dependence can limit the adaptability and robustness of these methods when dealing with noise, reverberation, or multiple overlapping sound sources. Fixed thresholds may not generalize well across diverse conditions, leading to limited performance. These limitations highlight the need for more flexible, data-driven approaches that can automatically learn to extract and integrate auditory cues without relying on preset heuristics.

1.4 Binaural DOA estimation

Sound source localization typically encompasses estimating both the direction and distance of sound sources [68]. However, the localization task in binaural hearing aids is often narrowed down to binaural DOA estimation. The binaural DOA estimation focuses on determining the angular position of sound sources in the azimuthal and elevation planes without explicitly calculating distance information. This simplification is justified by the far-field assumption, which is typically valid in most hearing aid scenarios where the listener's distance to sound sources is 5-10 times greater than the size of the listener's head (or the distance between two ears) [69]. Moreover, due to the physical constraints of hearing aids, namely their limited size and computational resources [5], focusing on DOA estimation, hearing aids can efficiently allocate their processing power to provide sufficient spatial information that can be utilized for beamforming, noise reduction, and overall improvement of the listening experience [5, 70].

A core component of traditional binaural DOA estimation algorithms involves the estimation of ITDs and ILDs through HRTFs, which encapsulate how sound waves are affected by the anatomy of an individual's head, torso, and pinnae. HRTFs are defined as the transfer functions conveying sound from a point source to a listener's eardrum in anechoic conditions, instrumental in characterizing binaural cues across frequencies. These functions are unique to each person and can be approximated using either a database or an acoustical model of the head [71, 72]. Their time-domain counterparts, head-related impulse responses (HRIRs), offer a temporal representation of these transfer functions. Historically, HRTFs have been primarily measured in far-field conditions, utilizing both human and artificial heads for various spatial directions as a function of azimuth and elevation. They are often provided as HRIR measurements for discrete angles [1, 73–76].

1.4.1 Acoustic scenarios

Fig. 1·2 depicts the most general acoustic scenario considered in this thesis for different tasks associated with binaural DOA estimation in reverberant environments. Considered acoustic scenarios may include single or multiple talkers, non-speech interference, or background noise. In this thesis, it is assumed that the number of talkers is known. Chapter 2 and Chapter 3 of this thesis consider the task of single-talker binaural DOA estimation in the presence of background noise, while Chapter 4 extends this to multi-talker scenarios. Typical background noises considered include diffuse, ambient, and babble noise. Appendix A explores single-talker binaural DOA estimation in the presence of directional non-speech interference, such as keyboard typing, vacuum cleaner noise, engine sounds, and other interior/domestic sounds.

The listener in Fig. 1·2 is wearing bilateral hearing aids, one on each ear, equipped with dual-microphone arrays with microphones near the ears. In addition to background noise or non-speech interference, each microphone captures the direct sound from each talker and the surrounding environment's reflections. These reflected sounds, known as reverberation, are caused by the interaction of sound waves with surfaces in the environment, such as walls, ceilings, and objects, and can complicate the process of DOA estimation [68, 77, 78].

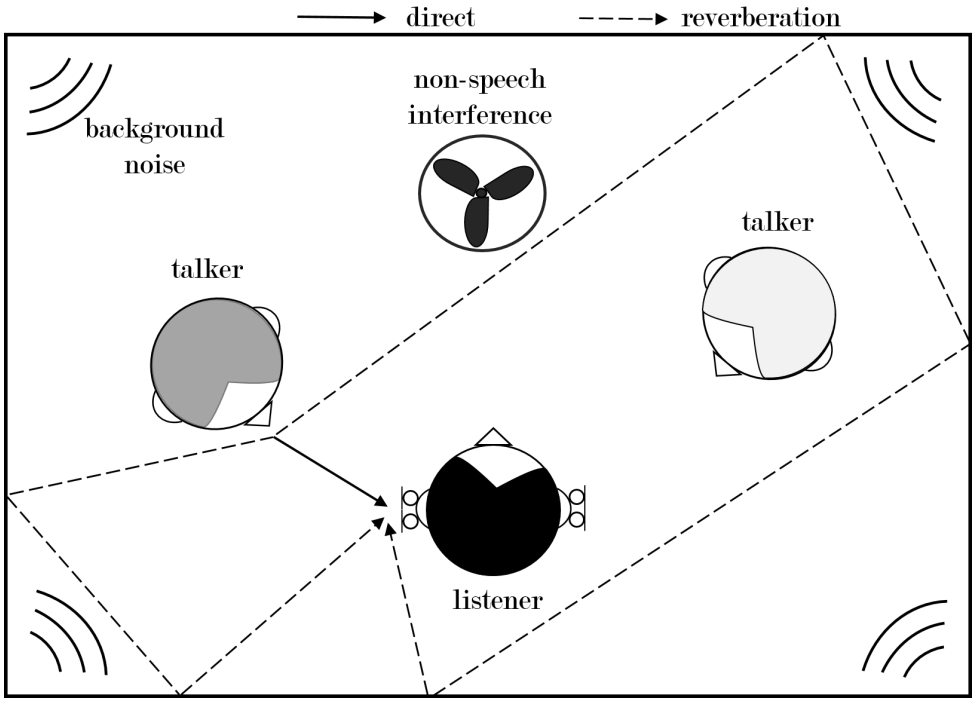


Figure 1·2: General acoustic scenario with multiple talkers, a single non-speech interference, and background noise in a reverberant environment. The listener is wearing two hearing aids on the left and right ear, each consisting of two microphones, where the microphones are located close to the ears on both sides.

1.4.2 Overview of binaural DOA estimation approaches

DOA estimation algorithms often determine the DOA in terms of the azimuth angle or both azimuth and elevation. Azimuth-only DOA estimation describes the DOA of the sound source relative to the listener in the horizontal plane. DOA estimation incorporating both azimuth and elevation provides a complete representation of the sound source's position in three-dimensional space. However, real-world listening scenarios where hearing aids are crucial, e.g., conversations in multi-talker scenarios, usually involve sound sources at roughly the same elevation as the listener. Hence, focusing on DOA estimation in the horizontal plane addresses the most important needs of hearing aid users. This thesis considers the DOA estimation in the azimuthal plane (0° elevation). Fig. 1·3 depicts the binaural hearing aid setup and the coordinate system in the azimuthal plane considered in this thesis. The DOA θ in Fig. 1·3 denotes the azimuth angle between the reference direction (look direction of the listener) and the talker in the far field.

Among the binaural cues commonly utilized for DOA estimation, ITD is the most prevalent. ITD extraction can be performed via auditory-inspired approaches using GTFBs [10, 79] or the broadband GCC-PHAT [79–81]. Research demonstrates that combining ITD (or IPD) and ILD cues improves DOA estimation accuracy, particularly in challenging multi-talker environments [9, 10, 82]. These studies often adopt a common approach in binaural DOA estimation to match the estimated binaural features with pre-computed feature templates derived from HRTF databases. Another class of binaural DOA estimation methods leverages RTF vectors [8, 11, 83]. These approaches use prototype RTF vectors pre-computed for various directions

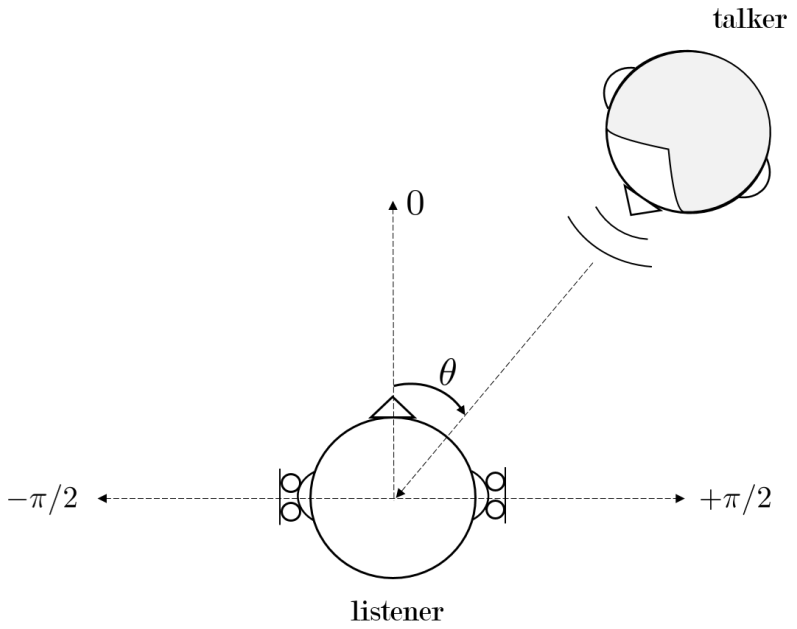


Figure 1·3: Hearing aid setup and the coordinate system in the azimuthal plane.

using measured HRTF databases. The RTF is commonly defined as the ratio of the HRTFs or source-to-microphone acoustic transfer functions (ATFs) for a given microphone pair and source position. It captures the difference in sound propagation paths between two microphones for a particular source location.

Alongside these cue-based methods, classical DOA estimation techniques [77, 78] are widely employed. TDOA-based approaches typically involve a two-step process, first estimating the time delays between microphone signals and then mapping these delays to the source's direction [77]. The GCC-PHAT [80] is a well-known TDOA-based method, utilizing cross-correlation between microphones to find the time delay corresponding to the DOA. It calculates the cross-correlation in the frequency domain after normalizing the cross-spectrum magnitude, making it less sensitive to room reverberation and noise. A class of DOA estimation methods, the so-called beamforming-based methods, involve steering a beamformer towards various candidate directions, usually searching for the direction that maximizes the output power [7, 84]. A well-known example is the steered response power with phase transform (SRP-PHAT) method [85], which weights the frequency components of the received signals using the PHAT weighting and calculates the power for each potential source direction. Model-based approaches like maximum likelihood (ML) estimation [7, 11, 86] make assumptions on statistical representations of the underlying signals. Subspace-based techniques such as multiple signal classification (MUSIC) [87, 88] and estimation of signal parameters via rotational-invariance techniques (ESPRIT) [89] offer alternative strategies for DOA estimation.

A significant challenge in binaural DOA estimation is the degradation caused by real-world acoustic conditions. Background noise, interference, and reverberation introduce uncertainties

into the binaural cues or violate assumptions on the signal models. These uncertainties can distort the extracted features, leading to mismatches with pre-computed templates and ultimately reducing DOA accuracy [10].

1.5 Deep learning for DOA estimation

In recent years, DOA estimation has increasingly transitioned towards DNN-based approaches, which offer several advantages over traditional methods [12]. Unlike traditional approaches that often rely on assumptions on signal distribution or spatial characteristics, DNN-based approaches can capture complex patterns and relationships in acoustic data, leading to more accurate and robust DOA estimation in challenging real-world conditions. DNN-based techniques have demonstrated strong performance in challenging acoustic environments, particularly in the presence of noise, reverberation, and interference.

1.5.1 Learning strategy

Learning strategies for DOA estimation are typically divided into unsupervised (or semi-supervised) [90–95] and supervised approaches [14–20, 23, 96–106]. Unsupervised approaches do not require labeled data and instead rely on identifying patterns or structures within the data to estimate the DOA. These methods can be beneficial when labeled data is scarce or difficult to obtain, but they often face challenges in complex environments where clear patterns are less evident, potentially impacting accuracy. In contrast, most DNN-based DOA estimation techniques utilize supervised learning, which involves training models on a large amount of labeled data. This supervised approach is particularly advantageous as it allows the model to capture complex patterns from the data, leading to more precise DOA estimation. In this thesis, we adopt the supervised approach due to the availability of large labeled datasets and the feasibility of generating additional training data as required. This approach enables us to fully exploit the capabilities of DNNs to achieve accurate and reliable DOA estimation.

1.5.2 Output coding

When designing a DNN for DOA estimation or, more generally, sound source localization, choosing between classification and regression as the output strategy is crucial. Classification involves dividing the localization space into distinct directions or zones, representing different classes. The DNN is trained to output the probability of a source being present in each direction or zone [14, 16, 17, 20, 23, 95–99, 102, 103, 105, 107–117]. This method is versatile, accommodating both single [16, 20, 99, 103, 107–109, 112, 114, 116–118] and multi-source [14, 17, 23, 95–98, 102, 105, 113, 115] localization, as the network learns to predict source activity for each zone regardless of the total number of sources. Conversely, regression aims to directly estimate the source's continuous spatial location, represented in Cartesian or spherical coordinates, or through generating intermediate features for traditional DOA estimation [15, 93, 100, 101, 104, 106].

While regression avoids the quantization errors associated with classification, it requires prior knowledge of the number of sources and often encounters the source permutation problem in

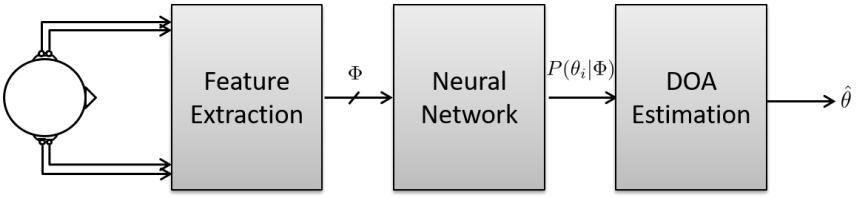


Figure 1·4: A block diagram of a typical classification-based DOA estimation system for binaural hearing aids. Binaural signals are initially processed to extract input features Φ. These features are then mapped to a posterior probability distribution for each DOA class i, representing the discretized DOA θ_i. Finally, the probability distribution is used for DOA estimation.

multi-source scenarios [109, 119–122]. This problem arises due to the ambiguity in matching estimated outputs with their corresponding targets during training and evaluation, making it difficult to determine which estimate corresponds to which source. This is a significant drawback of regression-based methods in multi-source scenarios, which require multiple regressors and network architectures. While the choice between regression and classification depends on the specific application and requirements, several studies suggest that classification might be a more suitable option for DOA estimation, especially in multi-source scenarios and challenging environments like low SNR conditions and when sources are closely spaced [109, 121, 123]. This thesis adopts the classification approach for DOA estimation, which requires a single network architecture, simplifying the system design and potentially improving performance in real-world hearing aid scenarios.

Considering the DOA estimation as a classification problem, the DOA range is typically discretized into a set of distinct DOA classes. For instance, the entire 360° azimuth range can be divided into 72 directions, resulting in discretized DOAs with a 5° resolution. Fig. 1·4 depicts the block diagram of a typical classification-based system for a binaural hearing aid setup. For single-talker scenarios, the DNN's output layer typically consists of neurons, each representing a DOA class. Taking the one-hot encoding scheme, the network is trained to activate the neuron associated with the sound source's DOA while suppressing others. A softmax activation function is commonly used to convert the network's outputs into a probability distribution across DOA classes [99, 107, 111, 112, 117]. The DOA with the highest probability is selected as the estimated DOA during inference. For multi-talker scenarios, the output layer typically has one neuron for each DOA class, similar to single-talker DOA estimation. However, instead of using the softmax activation function, with an assumption of independent output classes, i.e., source locations are independent of each other, the sigmoid activation function is often used as it allows all neuron outputs to be independent and range from 0 to 1 [98, 102, 113]. During inference, when the number of active sources is unknown, the approach analyzes the output probability distribution and identifies peaks exceeding a predefined threshold. Each peak indicates an active sound source, with the corresponding DOA class taken as the estimated DOA. This strategy is particularly useful in dynamic environments where the number of sources can vary over time. On the other hand, when the number of sources is known beforehand, the process involves selecting the DOA classes with the highest probability outputs corresponding to the known number of

sources. The DOAs associated with these classes are then taken as the estimated DOAs.

1.5.3 Input feature

Most DNN-based approaches rely on hand-crafted, pre-processed input features derived from established signal processing techniques rather than learning directly from raw waveforms. These features are designed to emphasize the spatial and time-frequency characteristics inherent in multichannel audio signals. Commonly used input features include those employed in classical DOA estimation approaches, such as CCF [14, 16, 101], GCC-PHAT [17], ITD [13, 105], ILD [13, 14, 16, 18, 101, 105], IPD [15, 18], and RTF [23, 93]. To move towards more generalized representations, time-frequency representations like spectrograms (based on magnitude, phase, real, or imaginary components) are often employed [20, 96–98, 102, 103]. Additionally, mel-scale and GTFB representations are utilized to approximate the auditory system's frequency response, aligning with human auditory perception [17, 104–106]. The cross-power spectrum (CPS), closely related to cross-correlation, provides another time-frequency representation used as input features [110]. In fact, the cross-correlation function can be obtained by taking the inverse Fourier transform of the CPS. Similarly, the GCC-PHAT is obtained by normalizing the CPS to isolate its phase information and then taking the inverse Fourier transform of the CPS phase. Finally, some models use the raw multichannel waveforms as input, allowing the network to learn its own representations [99, 100], though this approach can be computationally demanding. The choice of input features influences the complexity and performance of the model, and the optimal choice depends on factors like the application, acoustic environment, and available resources.

While deep learning approaches can learn directly from raw data or spectrograms, incorporating hand-crafted spatial features as input can enhance performance and reduce training requirements. These features can provide a more compact and informative representation of the spatial audio information, allowing neural networks to focus on learning higher-level patterns relevant to localization rather than having to learn basic acoustic principles.

1.5.4 Network architecture

A wide variety of DNN architectures [124–126] have been studied for sound source localization, ranging from more simple structures like feedforward neural networks to more complex models such as convolutional neural networks (CNNs), residual networks, recurrent neural networks (RNNs), attention-based networks, and encoder-decoder architectures. Initial attempts at using neural networks for sound source localization employed relatively simple architectures, such as multilayer perceptrons (MLPs), a particular type of feedforward neural networks (FFNNs) with fully-connected layers [13, 14, 108, 109]. However, these architectures generally have limited capacity to capture audio signals' spatial and temporal intricacies. Since the introduction of CNN architectures for sound localization application [107], they have been widely used. Researchers explored various CNN architectures, ranging from basic convolutional layers [17, 18, 22, 98, 99, 102, 110, 111, 117, 118, 127–130] to more sophisticated designs like Residual Networks (ResNets) [92, 95, 97, 100, 112]. RNNs, known for capturing long-term temporal de-

pendencies [124], are frequently integrated into convolutional recurrent neural network (CRNN) architectures, leveraging the combined strengths of RNNs and CNNs [96, 113, 131]. Due to their ability to model the temporal dynamics of sound signals, CRNNs have been predominantly used for sound event localization and detection, particularly in the DCASE challenge [132]. Beyond CNNs, attention-based and encoder-decoder architectures have also been applied to sound source localization. Encoder-decoder networks, including autoencoders and U-Nets, offer powerful ways to learn compressed representations and accurately estimate sound source DOA [23, 94, 106, 133]. Attention mechanisms, like multi-head self-attention [134], enable neural networks to selectively concentrate on specific segments of an input sequence. This capability is particularly useful to handle the temporal complexities of sound events [135, 136]. While effective, these methods often come with higher computational demands and complexity.

While each architecture has its strengths, CNNs have emerged as the most commonly used architecture for sound localization and DOA estimation systems. Although attention-based and encoder-decoder networks are powerful for tasks involving complex sequence modeling, they introduce additional complexity that may not be necessary for the relatively more straightforward task of DOA estimation in hearing aids. Additionally, CNNs use shared weights in their convolutional layers, significantly reducing the number of parameters compared to fully-connected networks. This design makes CNNs more computationally efficient and easier to train, especially when dealing with high-dimensional audio input. For example, in [22], a real-time CNN-based DOA estimation system was implemented on an Android smartphone using its built-in two microphones, tailored explicitly for hearing aid applications. CNNs excel at processing data with grid-like structures, making them well-suited for handling the spatial, spectral, and temporal patterns present in audio features like spectrograms, GCC-PHAT, and RTF representations. The key advantage of CNNs lies in their convolutional layers, which employ learnable filters that slide across the input data [124]. These layers extract local features, allowing the network to identify patterns and relationships between different time-frequency bins and microphone channels. This is particularly beneficial in sound localization as it enables the network to learn spatial patterns and relationships across frequency bands and microphone inputs.

1.6 Open issues

In real-life applications, speech inactivity poses a substantial challenge to DOA estimation systems, often leading to unreliable results. These challenges persist despite significant advancements in DOA estimation techniques, particularly those leveraging DNNs. A common approach to address speech inactivity is the integration of a VAD with the DOA estimator, which is considered a VAD-informed approach. For instance, research has demonstrated that incorporating a VAD can enhance DOA estimation systems utilizing CNNs for hearing aid applications by avoiding DOA estimation during noise-only frames [22]. Fig. 1·5 illustrates the block diagram of a classification-based system that leverages spatial features in a binaural hearing aid setup, combined with a VAD. This configuration allows the DOA estimation process to be conditioned on speech presence detected by the VAD.

However, this approach imposes some limitations. Separate VADs often require manual

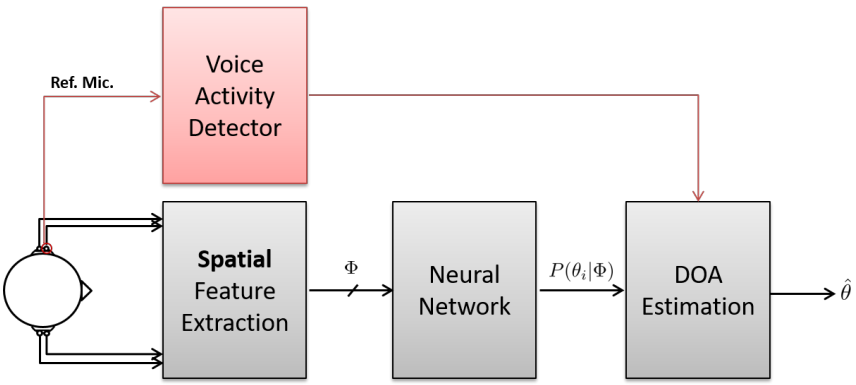


Figure 1·5: A block diagram of a classification-based DOA estimation system for binaural hearing aids, combined with a voice activity detector to mitigate the effect of speech inactivity on DOA estimation performance. A reference microphone signal is used for the voice activity detector, commonly employed to prevent DOA estimation when the signal is dominated by noise.

tuning, which can be time-consuming and may not generalize well across different acoustic environments. It should be noted that a VAD is primarily designed for speech detection, and its inclusion does not automatically translate to improved DOA estimation. The effectiveness of using a VAD for these purposes heavily relies on its accuracy. While a VAD can contribute positively, errors in VAD output, such as false positives, where background noise is misclassified as speech, can impair the DOA estimation process. Many VAD methods use features related to the pitch and periodicity of speech signals [137–140]. Real-life recordings for hearing aid applications are often contaminated by various noise sources, such as traffic, wind, or background conversations [141], which can degrade the performance of pitch and periodicity estimation methods [142].

Instead of relying on a VAD-informed approach for DOA estimation, a single DNN model could handle both speech detection and DOA estimation. This streamlined approach can potentially improve DOA estimation performance, particularly in challenging acoustic environments, while reducing computational overhead.

Moreover, while periodicity and harmonicity, especially when combined with spatial cues, have proven helpful in traditional CASA systems, their application to DNN-based DOA estimation remains unexplored. Current methods, often relying on manually tuned parameters, struggle in complex acoustic environments. This highlights the need for data-driven approaches that potentially leverage these auditory cues for more robust DOA estimation.

1.7 Thesis challenges and main contributions

This thesis introduces novel approaches, previously unexplored, for binaural DOA estimation leveraging the auditory-inspired periodicity features as the distinctive footprint of speech signals. By integrating these features with spatial features within DNN frameworks, the thesis improves DOA estimation in binaural hearing aids across single and multi-talker environments. This thesis highlights the potential of combining periodicity and spatial features for DOA estimation,

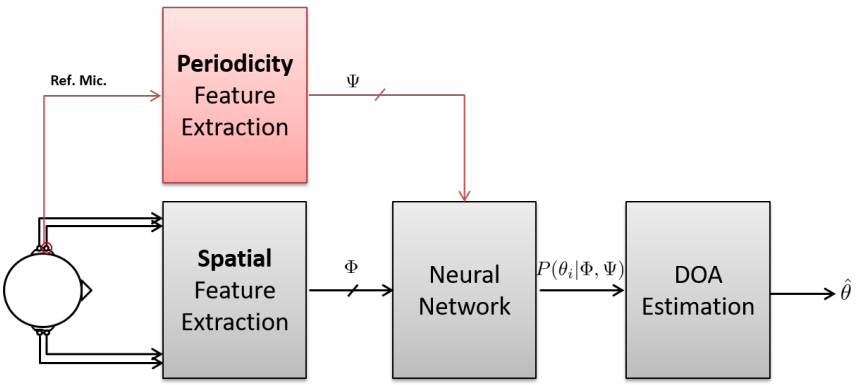


Figure 1·6: A general block diagram of a classification-based DOA estimation system for binaural hearing aids, utilizing auditory-inspired periodicity features (Ψ) as a monaural feature combined with spatial features (Φ).

paving the way for future advancements in sound source localization and speech enhancement for the hearing aid technology and speech processing fields.

Fig. 1·6 illustrates a general block diagram of a classification-based DOA estimation approach that leverages the proposed feature combination. Integrating spatial and periodicity features in a neural network is challenging due to the different nature of these features, requiring innovative approaches and network designs to efficiently and effectively combine these heterogeneous data types. This integration must be optimized to enhance DOA estimation accuracy while remaining computationally efficient. This could involve concatenation, attention mechanisms, or other fusion techniques that allow the network to learn the most relevant features for accurate DOA estimation. Moreover, periodicity features often have a sparse structure because they are derived from harmonic content that might not be present throughout the entire signal. The challenge is to design a network architecture that can effectively exploit these sparse features without being impaired by irrelevant or less informative parts of the signal. The neural network must be able to focus on and extract the most salient periodicity features that contribute to DOA estimation. By comparing the performance of the developed models against baseline systems in both single- and multi-talker scenarios, this thesis endeavors to demonstrate the advantages of the proposed feature integration strategies.

As the first contribution of this thesis, in Chapter 2, **we propose a novel DNN-based approach that reliably estimates the DOA of a single talker upon speech detection merely relying on the DNN output, without requiring a separate VAD. We propose different DOA estimation systems using dual-path CNNs with parallel convolutional pathways, exploiting novel feature combinations consisting of the periodicity and spatial features as separate inputs.** This contribution also presents a novel subband-averaged formulation of a periodicity feature, known as the PD, with varying frequency selectivity. We show that, regardless of the spatial feature type (subband-averaged or broadband), the proposed systems consistently demonstrate a clear benefit from the feature combination compared to baseline systems utilizing the same spatial feature in conjunction with a cascaded pitch-based VAD, achieving higher DOA estimation accuracy and reduced angular error.

As the second contribution, Chapter 3 **introduces an efficient DOA estimation system**

within a two-stage CNN architecture for single-talker binaural DOA estimation. Unlike the methodology detailed in Chapter 2, which processes PD and spatial features separately, this approach merges and jointly processes the input features in a single convolutional pathway. This two-stage CNN employs 1×1 convolutions to transform the sparse narrowband PD features into compact PD saliency features, allowing for a more computationally efficient system with improved performance rather than directly combining the spatial and PD features. The proposed system also outperforms the baseline system, which consists of a CNN using the same spatial feature and a cascaded pitch-based VAD.

As the third contribution, built upon the previous chapter, Chapter 4 **proposes to use the combination of narrowband spatial and PD features in a computationally efficient two-stage CNN architecture specifically adapted for multi-talker scenarios.** This system effectively captures the spectro-temporal dependencies in the input features across a few CNN layers, demonstrating the effectiveness of feature combination for multi-talker DOA estimation in binaural hearing aids.

1.8 Thesis outline

In the remainder of this section, we provide a clear chapter-by-chapter overview of this thesis, summarizing the content and highlighting the contributions of each chapter. A structured thesis overview is given in Fig. 1·7.

In **Chapter 2**, built upon our preliminary study in [143], we explore enhancing single-talker binaural DOA estimation by considering the speech-aware DOA estimation approach. This DNN-based approach formulates the DOA estimation task as a classification problem where, in addition to the DOA classes, a detection class in the output serves as an uncertainty class for reliable DOA estimation. This chapter pioneers combining spatial features with an auditory-inspired periodicity feature, known as the PD, as input to a CNN. In particular, we propose dual-path CNN architectures using parallel branches of convolutional layers, each receiving the spatial and PD features separately. The outputs of both branches are combined and used as input to a fully-connected path. This integration facilitates a speech-aware DOA estimation system that effectively operates without a separate VAD, distinguishing it from standard DNN-based systems. By training with both speech and non-speech signals, the CNN is enabled to capture the harmonic structure encoded in PD features, distinguishing speech from non-speech portions and accurately mapping spatial features to the sound source DOA upon speech detection. A significant contribution of this chapter is the evaluation of speech-aware DOA estimation systems utilizing both broadband and narrowband feature combinations compared to baseline systems. For narrowband features, we investigate using real/imaginary and magnitude/phase components of the CPS alongside a new subband-averaged PD representation. For broadband features, we combine GCC-PHAT with broadband PD features. This is a critical advancement over previous studies, providing insights into the benefits of employing PD features in both feature combinations for speech-aware binaural DOA estimation across different static and dynamic source scenarios. This chapter also presents the formulation of PD that incorporates auditory pre-processing with adjustable frequency resolution, generating a subband-averaged representa-

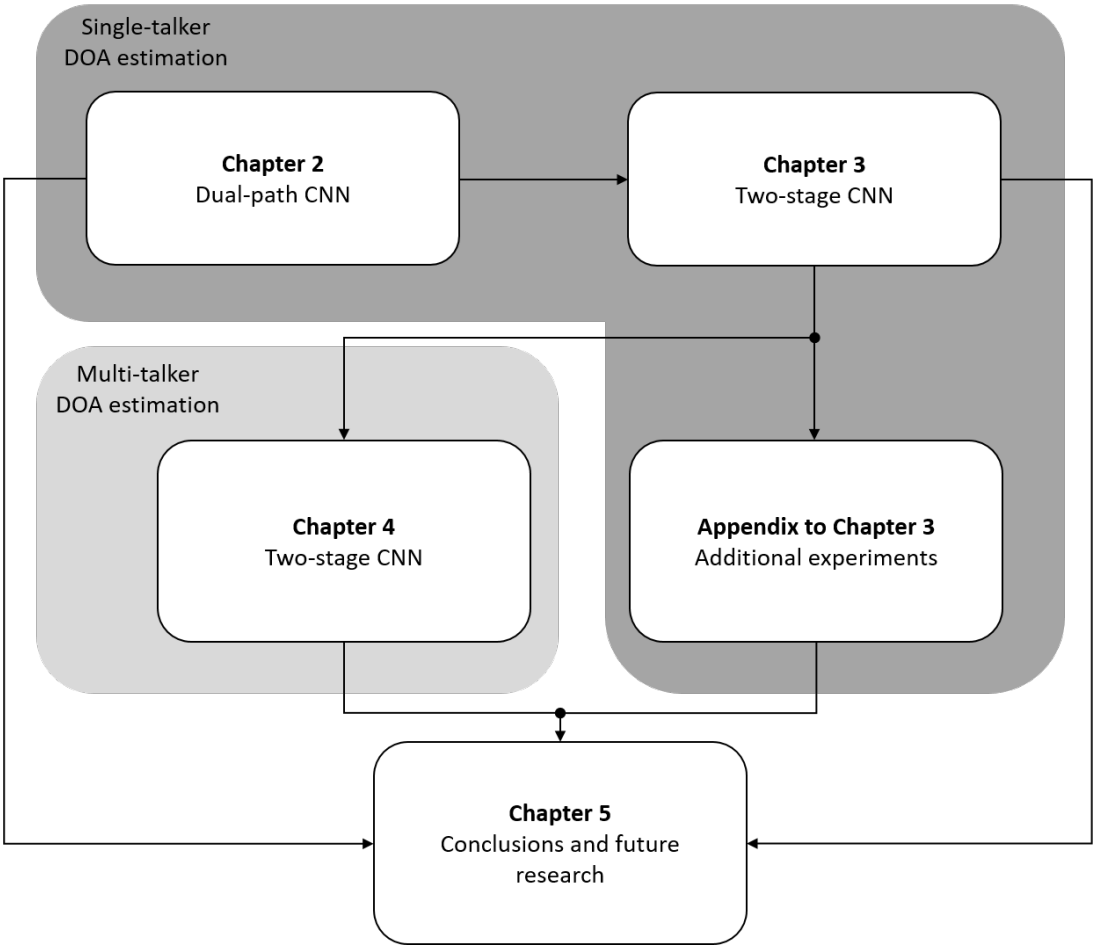


Figure 1·7: Structure of the thesis.

tion of PD. Extensive evaluations compare the performance of proposed narrowband systems with baseline systems in static-source scenarios within reverberant environments and dynamic scenarios with a single moving speech source. These evaluations, conducted under matched and unmatched background noise conditions and different SNR levels, demonstrate the benefit of incorporating PD features with any type of spatial feature for DOA estimation. The proposed speech-aware systems outperform baseline systems that rely solely on spatial features and a pitch-based VAD. The evaluations reveal that the proposed method, employing PD features, effectively estimates DOA in adverse SNR conditions and with higher degrees of spectro-temporal sparseness.

In **Chapter 3**, we refine the speech-aware binaural DOA estimation system from Chapter 2 by proposing a novel and computationally efficient two-stage CNN architecture. This new architecture provides a novel technique for integrating CPS and PD features while maintaining a similar training and testing paradigm for speech-aware DOA estimation. Firstly, instead of subband-averaged PD features, we use a narrowband representation of PD and the real and imaginary components of CPS. Secondly, instead of a dual-path CNN architecture used in Chapter 2, the two-stage CNN aims at directly combining the PD and CPS features on the same spectro-temporal regions. A pivotal contribution of this work is introducing a feature reduction stage based on 1×1 convolutions for the narrowband PD features before their joint processing with CPS features. This strategy is designed to exploit the sparsity property of speech signals more effectively by reducing the dimensionality of the PD features, which leads to a more efficient model that can guide the speech-aware DOA estimation process more effectively by focusing on compact, PD saliency features derived from the sparse structure of the PD features. This chapter delves into the benefits of PD dimensionality reduction, exploring its impact on both DOA estimation accuracy and computational efficiency. Evaluation results in terms of DOA estimation accuracy and angular error for static single-talker scenarios in two reverberant environments (with varying background noises and SNRs) demonstrate several key findings. Firstly, the proposed two-stage CNN performs better than a baseline system consisting of a CNN using only CPS features and a pitch-based VAD, even with a comparable number of trainable parameters. Secondly, it outperforms a speech-aware system that lacks PD feature reduction and requires a significantly lower number of trainable parameters. This highlights the computational efficiency of the proposed two-stage system, making it a practical and attractive solution for real-world applications.

In **Chapter 4**, we extend the research into more complex acoustic scenarios, focusing on multi-talker DOA estimation. Leveraging the insights and advancements from Chapter 3, this study addresses the gap in previous research by effectively utilizing narrowband PD features as monaural spectral features alongside the CPS phase as binaural spatial features in a two-stage CNN architecture adapted to the multi-talker DOA estimation task. The multi-talker DOA estimation approach contrasts with the speech-aware approach explored in Chapters 2 and 3. While the speech-aware approach included DOA classes alongside a detection class as the output classes, our multi-talker DOA estimation model only includes the DOA classes. Various architectural choices are investigated for the two-stage CNN, including different approaches to capture temporal and spectro-temporal dependencies of the PD and CPS features. Among the

different design choices investigated, the proposed system that captures temporal dependencies merely within convolutional blocks while independently processing each frequency emerges as the most effective system. This configuration offers the best DOA estimation performance with the lowest computational complexity, highlighting the importance of tailored architectural decisions. Conducting evaluations in static source scenarios with multiple talkers across different reverberant environments and SNRs with different background noises demonstrates that the proposed system consistently outperforms baseline systems that utilize either CPS features alone or in combination with magnitude spectrograms, highlighting the efficacy of integrating PD and CPS phase features.

In **Appendix A**, we consider the DOA estimation systems proposed in Chapter 3 and conduct additional experiments to investigate the benefit of the proposed systems for speech-aware DOA estimation in challenging scenarios. In particular, we evaluate the performance of the proposed systems and the baseline system in single-talker scenarios with non-speech interference in two reverberant environments. The proposed two-stage CNN, leveraging narrowband PD and CPS features, significantly enhances single-talker DOA estimation accuracy in reverberant environments with non-speech interference, outperforming the baseline system that utilizes a VAD. The improvement is notable across all tested SNRs, particularly under low SNR conditions. Additionally, the two-stage CNN with feature reduction consistently outperforms the proposed system without feature reduction. Although non-speech interference poses more significant challenges than the background noise used in previous experiments (Chapter 3), the proposed systems exhibit greater robustness to such interference than the baseline system.

In **Chapter 5**, we summarize the main findings of the thesis and provide an outlook on potential further research for future works.

Chapter 2

Speech-aware Binaural DOA Estimation Utilizing Periodicity and Spatial Features in Convolutional Neural Networks

---

This chapter is a reformatted reprint of the following publication:

R. Varzandeh, S. Doclo, V. Hohmann, "Speech-aware binaural DOA estimation utilizing periodicity and spatial features in convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1198-1213, 2024.

This study introduces a reliable DNN-based method to estimate a single talker's DOA during speech segments without needing a VAD. The proposed method uses a unique dual-path CNN that combines auditory-inspired periodicity features with spatial features. This feature combination within the DNN framework outperforms the standard approach employing a CNN that uses the same spatial features with a separate VAD, resulting in improved DOA estimation accuracy and lower angular error.

Abstract

In recent years, several supervised learning-based approaches have been proposed for estimating the direction of arrival (DOA) of a single talker in noisy and reverberant environments. In the absence of auxiliary information, such as a voice activity detector (VAD), the estimated DOA may be erroneous due to speech pauses or noise dominance. In this paper, we consider a speech-aware DOA estimation system for binaural hearing aids, which does not require a separate VAD. This system utilizes a combination of spatial features with an auditory-inspired periodicity feature called periodicity degree (PD) as input features of a convolutional neural network (CNN). Using speech and non-speech signals during the training, the CNN can capture the harmonic structure encoded in the PD features, thereby distinguishing speech from non-speech portions and simultaneously mapping spatial features to sound source DOA upon speech detection. To investigate the benefit of using PD features for speech-aware DOA estimation, we evaluated the performance of speech-aware systems that utilized either broadband or narrowband feature combinations compared to baseline systems. We propose to use a novel narrowband feature combination consisting of the narrowband cross-power spectrum (CPS) as the spatial feature and a new subband-averaged representation of PD features. The broadband feature combination consisted of the generalized cross-correlation with phase transform (GCC-PHAT) and the broadband PD features. The baseline systems considered in this work consisted of a CNN that exploits only a spatial feature, cascaded with a VAD. Evaluations in reverberant environments with different background noises for both static and dynamic single-talker scenarios demonstrate that incorporating the PD feature in conjunction with any type of spatial feature provides an advantage for binaural DOA estimation in terms of accuracy and angular error.

2.1 Introduction

Reliably estimating the DOA of a target speech source is a crucial task in applications such as binaural hearing aids. Several DOA estimation approaches have addressed this task. The model-based approaches [4, 8, 11, 21, 79, 144] typically rely on specific assumptions about the signal, noise, or reverberation model, which can be violated in adverse noisy and reverberant conditions, leading to a degraded DOA estimation performance. In addition to model-based DOA estimation approaches, in recent years several supervised learning-based DOA estimation approaches based on DNNs have been proposed [12, 14, 16, 18–20, 98, 99, 101], which can provide more robust performance in adverse scenarios when trained in different acoustic conditions [14, 99].

Most DNN-based binaural DOA estimation methods directly map features extracted from the signal to the sound source DOA [14, 16, 18], while some methods follow a two-step approach by first transforming signal features into enhanced features [19, 20]. The most frequently-used (spatial) features for binaural DOA estimation are the ILD, the ITD, the CCF, and the GCC-PHAT [80]. The complete CCF or the GCC-PHAT are typically used as the input feature for the DNN [14, 16], as this was shown to outperform using the ITD as the input feature [14]. Whereas most methods estimate the DOA in the azimuthal plane [14, 16, 20, 99], a few methods use multi-task learning approaches to jointly estimate the sound source azimuth together with elevation [18, 101]. In this work, we only consider binaural DOA estimation in the azimuthal

plane.

As a common DNN-based approach, the binaural DOA estimation task is often formulated as a classification problem, aiming at determining a mapping from input to a spatial probability map for a discretized azimuth range [14, 16, 99]. For instance, a binaural sound localization system was proposed in [99], which employs a CNN to find a mapping from the raw binaural signal waveforms to a posterior probability map. Although this system has been successfully able to outperform the baseline system with the GCC-PHAT input feature, it has only been trained and evaluated for noiseless scenarios, which is unrealistic in practical situations. Another category of DNN-based approaches involves the task of sound event localization and detection, which aims to identify and localize specific sound events in audio recordings, including both speech and non-speech events [97, 132]. In this paper, we focus on classification-based binaural DOA estimation, specifically aiming at DOA estimation of a single speech source.

A challenge when applying DOA estimation systems in real-life scenarios arises from speech inactivity, which can result in unreliable DOA estimates [21]. A general approach to deal with estimation errors due to speech inactivity in both model-based and DNN-based systems is to utilize a VAD [138] in parallel or cascaded with a DOA estimation system [21–24]. It should be realized that a separate VAD nonetheless usually requires manual and time-consuming parameter tuning which may entail readjustments when the system is used in different acoustic conditions. Moreover, a separate VAD itself can introduce errors that can restrict the overall performance of the system. In [22], a VAD was integrated into a CNN-based DOA estimation system for hearing aid applications, ensuring that the system avoids DOA estimation during noise-only frames. We will adopt a similar VAD-informed approach in the baseline systems considered in this paper. To address the speech inactivity problem in single-talker binaural DOA estimation, we will consider an alternative approach in this study. We treat it as a DOA estimation task without the need for a separate VAD, which we refer to as speech-aware DOA estimation.

To mitigate estimation errors caused by speech inactivity, classification-based systems commonly employ temporal averaging of the posterior probability map in the output over a relatively long duration [14, 16, 98, 99, 102]. Although this approach helps to smooth out unreliable estimates and improve the overall accuracy, it can compromise the reliability of the DOA estimation system when a new speech source emerges or becomes inactive. It also prevents the system from quickly detecting a change in the trajectory of a moving sound source.

A limited number of systems detect periods of silence within the output of a neural network [110, 112]. However, these approaches, primarily utilized in robot audition scenarios, have not been evaluated against the conventional classification-based approach, leaving their benefits unclear. Furthermore, some of these approaches have only been evaluated under unrealistic background noise conditions [112], while others are tailored for specific source distance and heights and have shown limited performance when tested in conditions that were not included in the training data [110].

It is assumed that the human auditory system groups signal components according to information such as periodicity of voiced speech and continuity of harmonics, and then ITD information is used to segregate the grouped components [4]. It is also known that about 75% of speech in spoken English is voiced and periodic [145]. This motivates the usage of an auditory-inspired

periodicity feature in combination with spatial features as input features of a neural network for DOA estimation of a single speech source.

In [48], an auditory-inspired feature called PD was proposed for fundamental period detection and estimation and was shown to be useful for VAD in low-SNR conditions. In [143], we proposed a classification-based speech-aware binaural DOA estimation system based on CNNs, which does not require a separate VAD. The proposed speech-aware system was compared to a baseline system that used a conventional classification-based approach. This study showed the benefit of using broadband PD features in combination with GCC-PHAT features as input features of the CNN for speech-aware binaural DOA estimation in static source scenarios.

In this paper, we extend our earlier study [143] by incorporating novel narrowband feature combinations. Our objective is to investigate the advantages of employing PD features in both narrowband and broadband feature combinations for speech-aware binaural DOA estimation across different static and dynamic source scenarios. We propose the novel narrowband feature combinations as follows: First, we introduce a formulation of the PD that incorporates an auditory pre-processing with an adjustable frequency resolution. This formulation generates a subband-averaged representation of the PD, allowing us to take advantage of the frequency selectivity of the human auditory system. Second, we propose to use narrowband CPS features (as spatial features) in combination with the subband-averaged PD feature as input features for the CNN. For the CPS feature, we consider either using the real and imaginary or the magnitude and phase components of the CPS. In summary, this study aims to investigate the benefits of PD features in the context of novel narrowband feature combinations, as established for the broadband feature combination in [143].

We conduct evaluations to compare the performance of the proposed narrowband systems with narrowband baseline systems consisting of a CNN utilizing only the CPS feature, cascaded with a state-of-the-art pitch-based VAD [138]. Additionally, We evaluate the performance of speech-aware and baseline systems that use broadband features as input features. All systems have been evaluated for static-source scenarios in reverberant environments with matched and unmatched background noise conditions. Furthermore, experiments were conducted for dynamic scenarios with a single moving speech source at different velocities for different SNR conditions. Our experimental results demonstrate the advantage of using the auditory-inspired PD feature in combination with any type of spatial feature (including the GCC-PHAT, real and imaginary parts, or magnitude and phase components of the CPS) for binaural DOA estimation.

The remainder of this paper is organized as follows. In Section 2.2, the single-talker DOA estimation problem is formulated as a classification problem and different approaches are discussed. In Section 2.3, we introduce the input features employed in this study. Section 2.4 provides a comprehensive description of the proposed and baseline systems. The details of the experimental setup for training and evaluation of all systems including datasets, data generation, training and network hyperparameters, and evaluation metrics appear in Section 2.5. The proposed and baseline systems are evaluated, and the results are discussed in Section 2.6. Section 2.7 summarizes the results and presents the conclusion.

2.2 DOA estimation as a classification problem

In this work, we consider the problem of single-talker DOA estimation in the azimuthal plane using a binaural hearing aid setup with M microphones, where the microphones are located close to the ears on both sides. The acoustic scenario consists of a (possibly moving) sound source at DOA θ in the azimuthal plane and background noise. The m-th microphone signal in the time domain at time t is given by

$$y_m(t) = x_m(t) + v_m(t), \tag{2.1}$$

where x_m and v_m denote the desired speech and noise signal components in the m-th microphone signal, respectively, which are assumed to be uncorrelated. In the short-time Fourier transform (STFT) domain, the m-th microphone signal at time frame n and frequency bin k (with K and D the STFT length and hop size, respectively) can be written as

$$Y_m(n,k) = X_m(n,k) + V_m(n,k). \tag{2.2}$$

By dividing the azimuth range into a set of C discrete DOAs $\{\theta_1, \cdots, \theta_C\}$, DOA estimation can be considered as a classification problem, where the DOA of a sound source should be assigned to one of the DOA classes. In this work, we consider $C = 72$ classes for the full 360° azimuth range, corresponding to a DOA map with 5° resolution. In the next subsections, two different classification-based approaches for binaural DOA estimation will be discussed.

2.2.1 Conventional DOA estimation

Conventionally, DOA estimation is formulated as a C-class classification task, where each output class corresponds to a DOA [16, 99]. During training, each training example belongs to only one output class that has been labeled using oracle DOA information. During testing, the neural network predicts a posterior probability map in the output. Under the single-source assumption, the DOA is then estimated by finding the DOA class with the highest posterior probability. To deal with erroneous DOA estimates (e.g., during speech pauses), a VAD can be cascaded to this system [22, 23], where a DOA is only estimated from the probability map if the VAD detects the signal as speech. In this work, we adopt the VAD-informed DOA estimation approach to design the baseline systems depicted in Fig. 2·3.

2.2.2 Speech-aware DOA estimation

In contrast to the VAD-informed classification-based approach, in [143] we proposed a classification-based approach referred to as speech-aware DOA estimation, which can estimate the DOA of a single talker, without needing a separate VAD. This problem is formulated as a $C + 1$-class classification task, where the first C classes represent the DOA classes and the last class represents the non-speech activity, regarded as the detection class. During training, via a one-hot encoding scheme, if a training example belongs to a speech source from a given direction, the DOA class corresponding to that direction is labeled by one, whereas all other classes (including

the detection class) are labeled by zero. On the other hand, if a training example belongs to a non-speech source, regardless of its direction, all DOA classes are labeled by zero, whereas the detection class is labeled by one. During testing, if the class with the highest posterior probability is a DOA class, the direction corresponding to that class indicates the sound source DOA. Otherwise, no reliable DOA could be estimated. In this work, we adopt the speech-aware DOA estimation approach in our proposed systems depicted in Fig. 2·4.

2.3 Input features

This section provides an overview of the spatial and periodicity features utilized as input features for various classification-based DOA estimation methods in this study. In Section 2.3.1, we present the broadband GCC-PHAT feature, which was also employed in [143], in addition to the newly introduced narrowband CPS features, as spatial features. In Section 2.3.2, we introduce the novel subband-averaged representation of the PD, along with the broadband PD used in [143]. Furthermore, we present the rationale for the incorporation of PD through exemplary visualizations that demonstrate different PD representations.

2.3.1 Spatial features

The GCC-PHAT has been successfully used as a feature for several data-driven DOA estimation methods[17, 109, 146, 147]. In this work, the broadband GCC-PHAT between the i-th pair of microphones is defined as the inverse Fourier transform of the phase of the instantaneous narrowband CPS which is given by

$$G_i(n,k) = Y_r(n,k)Y_q^*(n,k), \tag{2.3}$$

where microphones r and q constitute the i-th microphone pair and $(\cdot)^*$ denotes complex conjugate. We note that there are $M(M-1)/2$ microphone pairs, i.e., $i \in [1, M(M-1)/2]$. The GCC-PHAT for the i-th microphone pair at time frame n is computed as

$$\tau_i(n,d) = \mathcal{IFFT}\left(\frac{G_i(n,k)}{|G_i(n,k)|}\right), \tag{2.4}$$

where $|\cdot|$ denotes absolute value, and d represents the index of the time delay. In order to resolve fractional signal delays occurring for microphone pairs with a small distance (e.g., microphones on a hearing aid), it is useful to interpolate the GCC-PHAT function by using an oversampled inverse Fourier transform [79]. With an upsampling factor of κ, the relevant discrete time delays lie in the range $[-\kappa\tau_i^{max}, \kappa\tau_i^{max}]$, where τ_i^{max} denotes the maximum delay in samples, considered for the i-th microphone pair. The GCC-PHAT vector of the i-th microphone pair is defined as

$$\boldsymbol{\tau}_i(n) = [\tau_i(n,1), \cdots, \tau_i(n,\mathcal{T}_i)]^{\mathrm{T}}, \tag{2.5}$$

where $(\cdot)^{\mathrm{T}}$ denotes the vector transpose. The first and last elements in (5) correspond to $-\kappa\tau_i^{max}$ and $+\kappa\tau_i^{max}$, respectively. Therefore, the length of the GCC-PHAT vector is obtained by $\mathcal{T}_i = 2\kappa\tau_i^{max} + 1$. By concatenating the GCC-PHAT vectors $\boldsymbol{\tau}_i(n)$ for all possible microphone

pairs, and considering L consecutive time frames (including the current frame n and the previous $L-1$ frames), we obtain the two-dimensional (2D) GCC-PHAT input feature with dimensions $\mathcal{T} \times L$, where $\mathcal{T} = \sum_{i=1}^{M(M-1)/2} \mathcal{T}_i$. This 2D feature will be used as a spatial input feature for broadband systems in Section 2.4.

As can be seen in (2.4), the PHAT weighting eliminates the effect of spectral magnitude, such that phases contribute equally for all frequencies. Hence, as an alternative to the broadband GCC-PHAT, in this work, we will also consider the narrowband CPS [12], encoding both spectral magnitudes and phase differences, as an input feature.

As the CPS input feature, we consider either the magnitude and phase (denoted as MagPhase) or the real and imaginary parts (denoted as ReIm) of the complex-valued CPS $G_i(n,k)$ for all $M(M-1)/2$ unique microphone pairs, for $K/2+1$ frequencies (up to the Nyquist frequency, i.e., $k = 0, 1, \cdots, K/2$), and for L consecutive time frames. This means that the shape of the CPS input feature is equal to $(K/2+1) \times L \times 2M(M-1)/2$. We note here that the first, second, and third dimensions represent the height, width, and depth of the input feature, respectively, where the depth corresponds to the number of input channels. For the CPS input feature, $2M(M-1)/2$ input channels are constructed by stacking either the MagPhase or the ReIm for all microphone pairs. The CPS features will be used as spatial input features for narrowband systems in Section 2.4.

2.3.2 Periodicity features

Periodicity is an important cue to segregate and localize different talkers [47, 148]. Periodicity features are often estimated through an auditory pre-processing step followed by a feature extraction step [47] where they are estimated independently for each pre-processed subband signal. In [48] a periodicity feature called PD was introduced, which captures the salience of the periodic components in the input signal. In this work, we propose to use a subband-averaged representation of PD features, estimated for a set of N fundamental period candidates. Similar to spatial features, we will consider PD features from L consecutive time frames as input PD features. This section focuses on PD computation for time samples (t) spanning a block of L consecutive time frames.

To compute PD features, we use one of the M microphones, referred to as the reference microphone in this paper. It is important to note that the choice of the reference microphone is arbitrary, and determining the optimal microphone for PD estimation is not within the scope of this study. In the following, we present signal processing steps to compute the subband-averaged PD. In the pre-processing step, the reference microphone signal in the hearing aid setup is first decomposed into a set of subband signals using a complex-valued GTFB [48]. The real part of each subband signal is then passed through half-wave rectification, yielding the half-wave rectified signal $y(t,f)$ in the f-th gammatone subband. Although the PD is usually computed for each subband [47], in this paper we introduce a subband averaging step, enabling us to estimate the PD for frequency regions with adjustable bandwidths. The subband-averaged

signal is computed as

$$y_{avg}(t,\bar{f}) = \frac{1}{\sigma S} \sum_{f=(\bar{f}-1)S+1}^{\bar{f}S} y(t,f), \tag{2.6}$$

where $\bar{f}$ denotes the (averaged) frequency band index and S denotes the number of averaged subbands. The normalization parameter σ represents the standard deviation computed over all subbands and times (time samples of L consecutive STFT frames) for the signal $y(t,f)$. This subband averaging step results in F frequency bands. Subsequently, a fifth-order low-pass filter with 770 Hz cutoff frequency and a second-order high-pass filter with 40 Hz cutoff frequency are applied to $y_{avg}(t,\bar{f})$, resulting in bandpass-filtered signal envelopes $y_{env}(t,\bar{f})$.

In the feature extraction step, a set of N parallel infinite impulse response (IIR) comb filters designed for a given set of N fundamental period candidates $p_j, j = 1, \cdots, N$, filter the signal envelopes as

$$s(j,t,\bar{f}) = (1-\alpha) y_{env}(t,\bar{f}) + \alpha s(j, t-p_j, \bar{f}), \tag{2.7}$$

where α denotes the filter gain. The periodicity degree is defined as the mean amplitude of the comb-filtered signal, given by

$$PD(j,t,\bar{f}) = (1-\beta_j)|s(j,t,\bar{f})| + \beta_j PD(j,t-1,\bar{f}), \tag{2.8}$$

where the averaging parameter β_j for each fundamental period candidate is defined as $\beta_j = e^{-1/p_j}$.

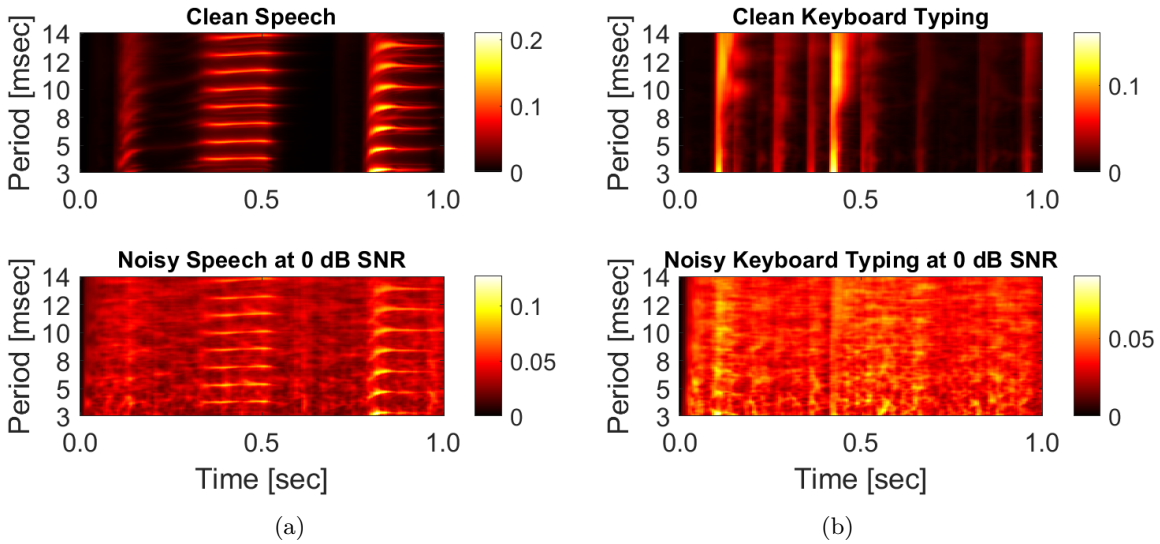
The PD features in (2.8) have the same temporal resolution as the time-domain signal. Since we aim at processing of the PD and CPS features by the neural network, it is desirable to represent both features at the same time resolution, which is the frame resolution of the STFT framework. Hence, the high-resolution PD features are temporally averaged as

$$\overline{PD}(j,n,\bar{f}) = \frac{1}{K} \sum_{t=(n-1)D+1}^{(n-1)D+K} PD(j,t,\bar{f}). \tag{2.9}$$

As the subband-averaged input PD feature, we consider PD features in (2.9) for all N fundamental period candidates, for L consecutive time frames, and for all F frequency bands. This means that the shape of the subband-averaged input PD is equal to $N \times L \times F$, where the first, second, and third dimensions represent the height, width, and depth of the input feature, respectively. This three-dimensional feature will be used as the periodicity input feature of the proposed narrowband speech-aware DOA estimation systems in Section 2.4.2.

As mentioned earlier, in this work, we also use the broadband PD feature from [143] for the broadband speech-aware DOA estimation system in Section 2.4.2. To obtain the broadband PD feature, signals of all gammatone subbands are averaged in (2.6), i.e., $F = 1$. The resulting broadband signal in (2.6) is utilized for PD feature extraction. Consequently, the broadband input PD has a shape of $N \times L$.

Fig. 2·1 depicts an exemplary representation of the broadband PD feature computed for $N = 180$ fundamental period candidates over a 1-second duration ($L = 199$) of both clean and



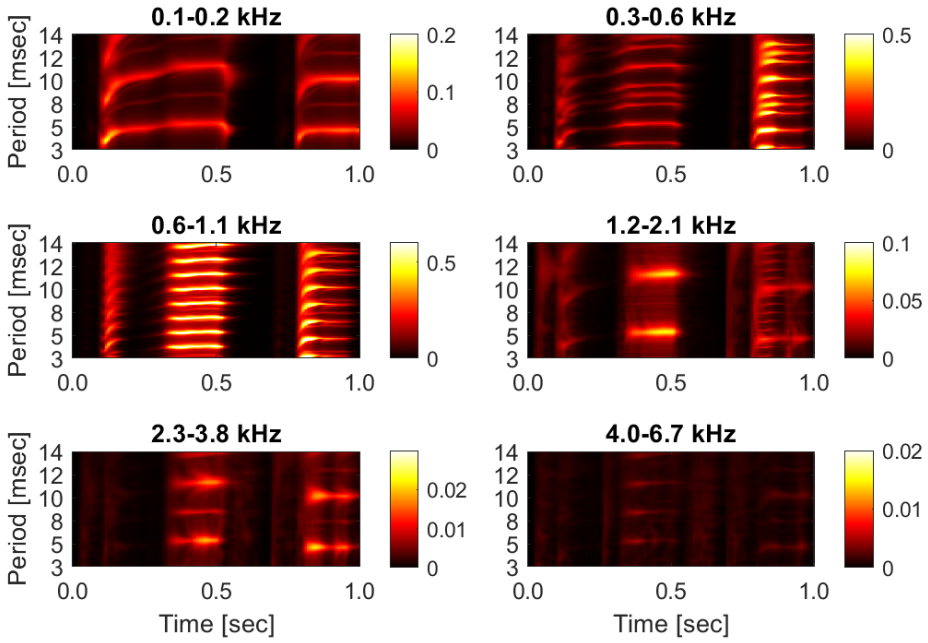
(a) (b)

Figure 2·1: An exemplary visualization of broadband PD computed for $N = 180$ fundamental period candidates and $L = 199$ consecutive time frames for clean and noisy (a) female speech and (b) keyboard typing signals in an anechoic environment with simulated diffuse noise at 0 dB SNR condition.

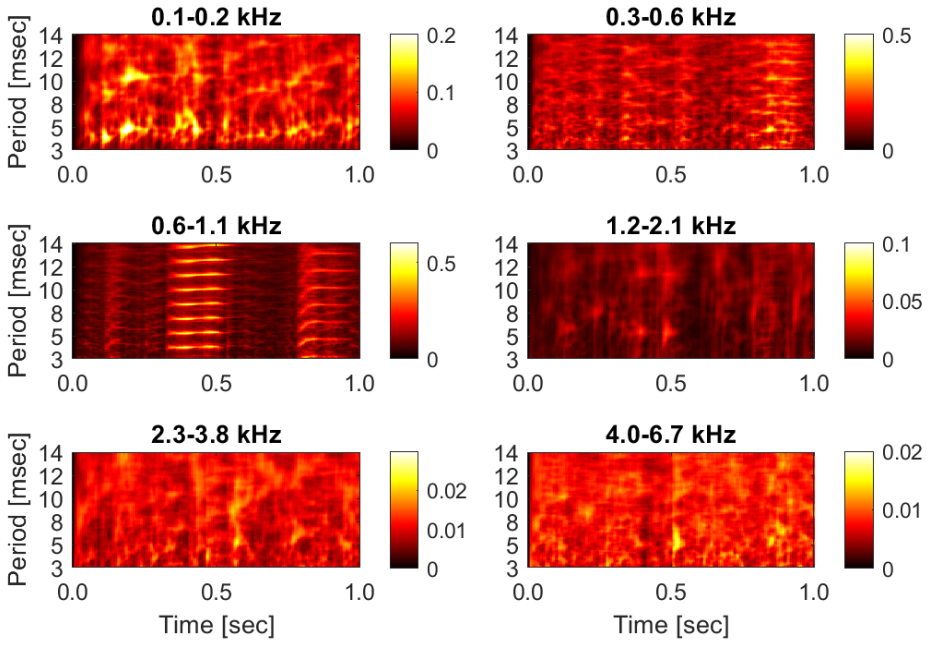
noisy speech, as well as non-speech (keyboard typing) signals. While for the clean and noisy speech signals, the fundamental period variation, its multiple harmonics and their temporal continuity are identifiable as a 2D structure over time, no such structure exists for the keyboard signal, and in general for non-speech signals. Although speech signals are not perfectly harmonic, we hypothesize that utilizing the fundamental period information encoded in the harmonic structure of the PD feature could facilitate a neural network's ability to differentiate between signal portions that are predominantly speech (and periodic) versus non-speech, particularly when trained with a combination of speech and non-speech signals.

The 2D structure of the pitch modulations and harmonics can also be identified in the subband-averaged PD features. Fig. 2·2 illustrates the subband-averaged PD features computed for $N = 180$ fundamental period candidates across $F = 6$ frequency bands ($S = 10$), using a 1-second duration ($L = 199$) of clean and noisy speech signals. As can be seen in Fig. 2·2a, the harmonic structure of the pitch information is captured in most frequency bands for the clean signal, particularly in frequency bands with a high degree of periodicity, e.g., in $0.3 - 0.6$ kHz and $0.6 - 1.1$ kHz frequency bands (with a maximum PD value of 0.5). However, as can be seen in Fig. 2·2b, for the noisy signal this information is substantially masked by the noise except for the frequency regions mainly in the $0.6 - 1.1$ kHz frequency band. By using subband-averaged PD features as input features, the neural network is expected to be able to select the most robust and salient periodicity information, particularly for those frequency bands in which speech signals have more energy, and hence are less susceptible to noise.

The primary rationale for employing PD features in conjunction with spatial features is to leverage the salient periodicity features as a footprint of speech signals in a noisy mixture [45, 46]. This approach enables the neural network to detect voiced speech portions of a signal while simultaneously mapping the CPS features of these portions to the talker's DOA.

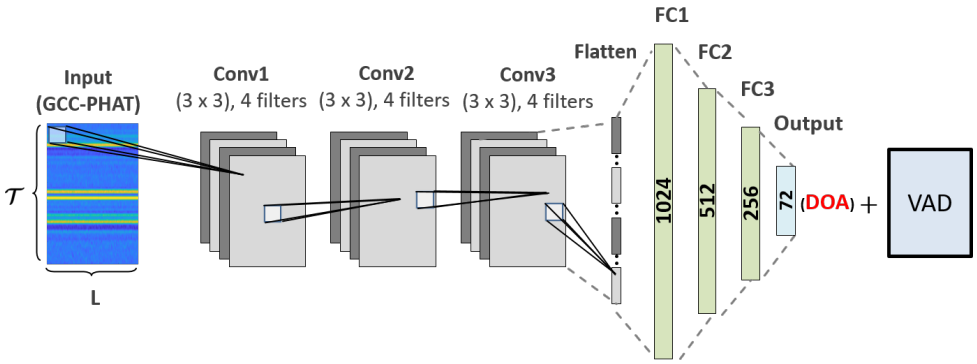


(a) Clean subband-averaged PD features

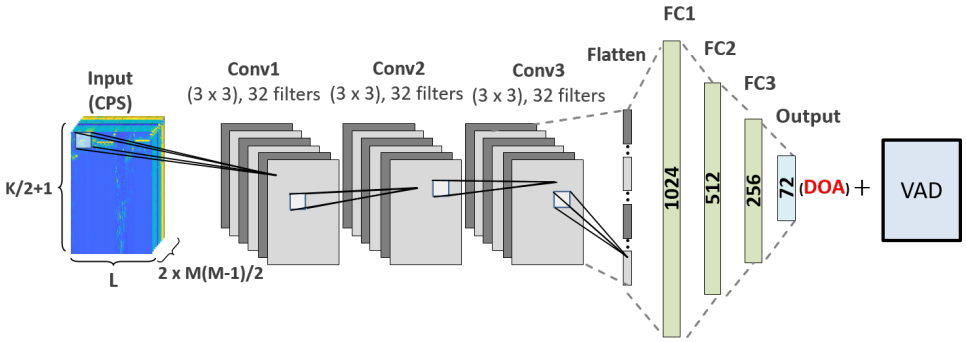


(b) Noisy subband-averaged PD features at 0 dB SNR

Figure 2·2: An exemplary visualization of the subband-averaged PD feature shown for (a) clean female speech, and (b) noisy female speech at 0 dB SNR, computed for $N = 180$ fundamental period candidates, $L = 199$ consecutive time frames, and $F = 6$ frequency bands in an anechoic environment with simulated diffuse noise as background noise. The frequency range corresponding to each frequency band is specified at the top of the images.



(a) With only broadband spatial feature (GCC-PHAT)



(b) With only narrowband spatial feature (CPS)

Figure 2·3: Baseline VAD-informed DOA estimation systems using only spatial features: (a) broadband spatial feature (GCC-PHAT), and (b) narrowband spatial feature (CPS).

2.4 CNN-based DOA estimation systems

This section outlines the CNN-based DOA estimation systems. The baseline systems are discussed in Section 2.4.1, which adopt a VAD-informed DOA estimation approach, utilizing only spatial features. The proposed systems are presented in Section 2.4.2, which adopt a speech-aware DOA estimation approach, utilizing a combination of spatial and PD features as input features. Finally, we discuss the computational complexity of the proposed and baseline systems in Section 2.4.3.

2.4.1 Baseline VAD-informed systems

Neural network architectures based on CNNs have been widely and successfully used for DOA estimation and sound source localization [12]. Fig. 2·3 depicts the baseline systems consisting of a CNN using only spatial features (cf. Section 2.3.1) as input, cascaded with a pitch-based binary VAD [138]. We consider three baseline systems:

- Broadband (Fig. 2·3a) using GCC-PHAT features,
- Narrowband-ReIm (Fig. 2·3b) using the real and imaginary parts of the CPS,

- Narrowband-MagPhase (Fig. 2·3b) using the magnitude and phase of the CPS.

The CNN architecture in all considered baseline systems starts with a cascade of three convolutional blocks, with each block (*Conv1* to *Conv3*) comprising a sequence of 2D convolutional, batch normalization, rectified linear unit (ReLU) activation, and 2D max-pooling layer. The outputs of the last pooling layer in *Conv3* are concatenated and then used as an input for a cascade of three fully-connected blocks (*FC1* to *FC3*), each representing a fully-connected dense layer followed by batch normalization, ReLU activation, and dropout layers. In the output layer, a softmax activation function predicts the posterior probability map for the C DOA classes.

In addition to the batch normalization layer implemented in the convolutional and fully-connected blocks of the CNNs, we applied a normalization scheme only in the input layer and directly on the input features before the first convolutional block to improve the performance of the CNNs. We applied layer normalization [149] on the GCC-PHAT features. Concerning the CPS input features, layer normalization was applied on all $2M(M-1)/2$ channels of the real and imaginary parts of the CPS to preserve phase information encoded by these features. As for the magnitude and phase parts of the CPS, group normalization [150] was applied to the two groups of magnitude and phase features each including $M(M-1)/2$ channels. This means that within each group, features are normalized separately. The reason for this is that the magnitude and phase have different statistical properties, and hence, joint normalization of the magnitude and phase may not be optimal. We note that all layer normalizations and group normalizations have been implemented without an affine transformation.

Each training example consists of a block of L consecutive time frames, i.e., we employ block-level labeling and the CNN generates its output for each block. We adopt a one-hot encoding scheme during the training, i.e., each training example belongs to only one output class that has been labeled using oracle DOA information. It is important to note that we assume a constant DOA when assigning a ground truth DOA label to a training example of a speech signal, which implies that the DOA remains consistent throughout the block of L consecutive time frames. During the testing phase, the CNN in the baseline system generates a posterior probability map $\boldsymbol{P} = [P_1, \cdots, P_C]$, which represents the likelihood of the sound source being located at each of the C possible DOA classes. It should be noted that we obtain consecutive input features with an overlap of $L-1$ frames for all systems. As input features consist of L consecutive time frames, this approach results in the generation of a new posterior probability map for each new frame.

To mitigate the effects of erroneous DOA estimates that can arise during periods of speech pauses, the system is augmented with a cascaded VAD. This configuration enables the DOA estimation process to be conditioned on the presence of speech, as determined by the VAD. Specifically, the DOA is estimated solely from the probability map when the VAD indicates the presence of speech, which is expected to lead to more robust and accurate DOA estimates. We note that the VAD decision is made using the same reference microphone signal that is used to compute PD features. As a common approach, the sound source DOA can be estimated as θ_I

for the DOA class I with the highest posterior probability, i.e.

$$I = \arg\max_i \ P_i. \tag{2.10}$$

In this work, to obtain continuous DOA estimates from discrete DOA classes, we estimate the sound source DOA by employing parabolic interpolation [151] on three DOA classes centered around θ_I, i.e., θ_{I-1}, θ_I and θ_{I+1}. As a result, this approach allows for a more precise estimation of the DOA with a higher spatial resolution.

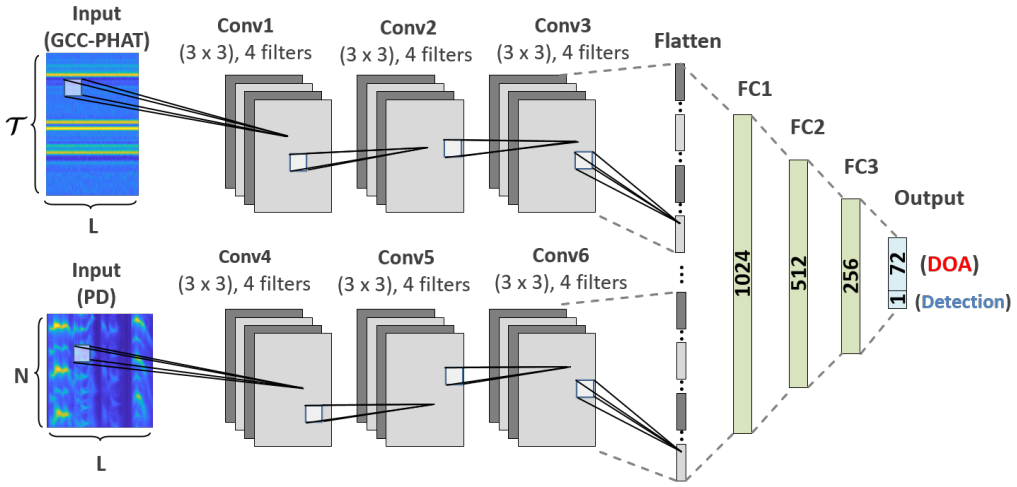
2.4.2 Proposed speech-aware systems

Instead of using a separate pitch-based VAD in combination with a CNN, we adopt the speech-aware approach described in Section 2.2.2. Fig. 2·4 depicts the proposed speech-aware DOA estimation systems, which use PD features (cf. Section 2.3.2) in combination with spatial features (cf. Section 2.3.1) as input features of the CNN. In this work, we consider three speech-aware systems:

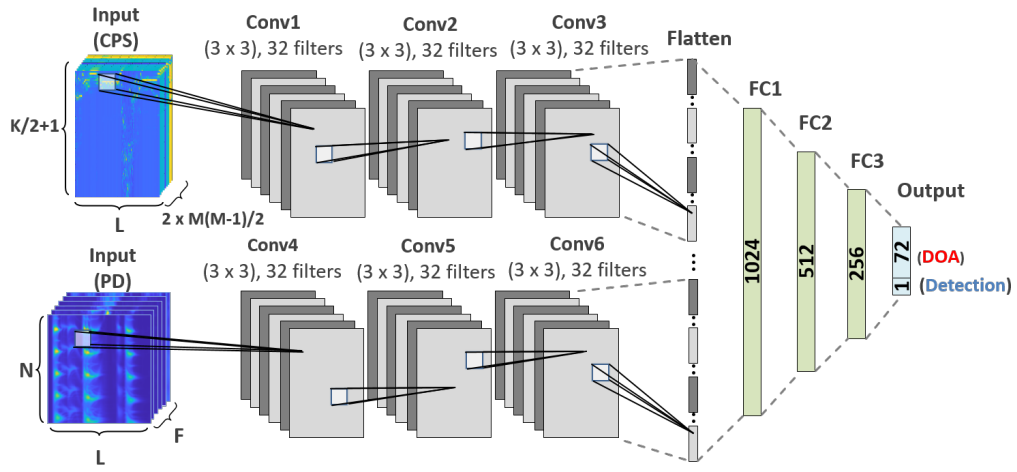
- Broadband (Fig. 2·4a) using GCC-PHAT features and the broadband PD features as input features,
- Narrowband-ReIm (Fig. 2·4b) using the real and imaginary parts of the CPS and the subband-averaged PD features as input features,
- Narrowband-MagPhase (Fig. 2·4b) using the magnitude and phase of the CPS and the subband-averaged PD features as input features.

Each proposed system in Fig. 2·4 consists of two parallel independent branches of three cascaded convolutional blocks. The top branch receives the spatial features as input features, whereas the bottom branch receives the PD features as input features. The outputs of both branches are then concatenated, which serves as a hybrid intermediate feature vector used by a cascade of three fully-connected blocks. Similar to the baseline systems described in Section 2.4.1, for the proposed systems each convolutional block consists of a 2D convolutional, batch normalization, ReLU activation, and 2D max-pooling layer. Each fully-connected block is comprised of a fully-connected dense layer followed by batch normalization, ReLU activation, and dropout layers. In the output layer, a softmax activation function predicts the posterior probability map for the $C+1$ classes. We applied layer normalization without an affine transformation on the PD input features of each proposed system. As for the spatial input features, we used the same normalization scheme that was applied to spatial features of the baseline systems (cf. Section 2.4.1). Please note that for each proposed system, the PD and spatial features have been normalized separately.

We note here that for each spatial feature, GCC-PHAT or CPS, the architecture of the convolutional branch with spatial input features in the proposed system and the architecture of the convolutional path in the baseline system are the same. This can be seen, for instance, by comparing Fig. 2·3a with the top branch in Fig. 2·4a, and also by comparing Fig. 2·3b with the top branch in Fig. 2·4b. As a result, the convolutional path in the baseline system and the



(a) With broadband feature combination (GCC-PHAT and PD)



(b) With narrowband feature combination (CPS and subband-averaged PD)

Figure 2·4: Proposed systems with (a) broadband feature combination (GCC-PHAT and PD), and (b) narrowband feature combination (CPS and subband-averaged PD). The architecture of the convolutional branch with a spatial input feature (top branch) in each proposed system is identical to the architecture of the convolutional branch in a baseline system using the same spatial feature depicted in Fig. 2·3.

top branch of a proposed system that use the same spatial features will learn the same number of parameters and filters, and ultimately contribute to the input of the fully-connected path by the same amount of (intermediate) features. Consequently, we can consider the contribution of the spatial features in the fully connected path of the proposed system to be equivalent to that of the baseline system. This allows us to analyze the benefits of using PD features and compare the two systems using the same spatial features.

We expect that by training the proposed systems with speech and non-speech signals, the network is able to capture the harmonic structure of the signal encoded in the PD features over consecutive frames. This allows the proposed system to discern between speech and non-speech

portions, while simultaneously mapping the spatial features to a sound source DOA when speech portions in the signal are detected.

The proposed systems were trained using oracle DOA and detection labels for speech and non-speech signals. All $C+1$ output classes were labeled as a single label, meaning each training example belonged to only one output class. This was achieved as follows: During the training phase, training examples of input features are provided for both speech and non-speech sources. For a given training example of a speech source, the direction of the speech source is associated with a particular DOA class, which is labeled by one. The remaining DOA classes, along with the detection class, are labeled by zero. In contrast, for a training example of a non-speech sound source, regardless of its direction, all DOA classes are labeled by zero, except for the detection class which is labeled by one.

During the testing phase, the proposed system generates a posterior probability map given by $\boldsymbol{P} = [P_1, \cdots, P_C, P_{C+1}]$ for a given number of directions C. We note here again that input features (each including L time frames) are consecutively obtained with an overlap of $L-1$ frames, i.e., a new posterior probability map is generated for every new frame. The process of speech-aware DOA estimation can be formulated by first introducing two hypotheses

$$\mathcal{H}_s: \quad \text{speech DOA detected}, \tag{2.11}$$

$$\mathcal{H}_{ns}: \quad \text{no speech DOA detected}, \tag{2.12}$$

and then defining the decision rule as

$$\begin{aligned} &\text{decide} \quad \mathcal{H}_{ns} \quad \text{if} \quad \arg\max_i \; P_i = C+1 \\ &\text{decide} \quad \mathcal{H}_s \quad \text{otherwise.} \end{aligned} \tag{2.13}$$

For the DOA estimation, we first consider the direction θ_I corresponding to the DOA class I with the highest posterior probability when speech DOA is detected, i.e.

$$I = \arg\max_i \; P_i | \mathcal{H}_s. \tag{2.14}$$

Then, we estimate the sound source DOA by employing parabolic interpolation [151] on three DOA classes centered around θ_I, i.e., θ_{I-1}, θ_I and θ_{I+1}. The process of speech-aware DOA estimation can be described as follows: the output class with the highest probability in the predicted probability map is selected. If the highest probability corresponds to the last class, which represents the detection class, it indicates that no reliable DOA estimation is possible. On the other hand, if the highest probability corresponds to a DOA class, the sound source DOA is estimated as the parabolic approximation of the direction associated with that DOA class.

It should be noted that the proposed speech-aware DOA estimation systems integrate both DOA estimation and VAD into a unified framework, with speech detection regarded as an implicit result of the proposed systems. Whereas for the VAD-informed systems, the DOA estimation is conditioned on the VAD decision (i.e., an *explicit* speech detection), for the speech-aware systems, the DOA estimation is merely conditioned on the joint probability distribution in

Table 2.1: Number of trainable parameters and multiply-accumulate operations (MACs) of the baseline and proposed systems.

CNN	Baseline broadband	Proposed broadband	Baseline narrowband	Proposed narrowband
Parameters/M	1.37	1.55	1.36	2.82
MACs/M	2.54	3.03	11.61	29.79

the CNN output, including the detection class (i.e., an *implicit* speech detection which we refer to as speech DOA detection in (2.11)). Therefore, in addition to the performance evaluation of all systems for DOA estimation, our study will assess their speech detection capabilities. This could offer a more comprehensive insight into these systems.

2.4.3 Computational complexity

Table 2.1 shows the number of trainable parameters and multiply-accumulate operations (MACs), both in millions for the baseline and proposed DOA estimation systems. The number of parameters, i.e. the model size, influences the memory required to store the model, while MACs provide an estimate of the arithmetic computations, which inherently affects energy consumption. By analyzing Table 2.1, we can observe that the size of the two CNNs employed for the baseline systems using the broadband (1.37 M) and narrowband (1.36 M) spatial features are comparable. It is important to note that the number of convolutional filters in each baseline system (cf. Fig. 2·3) was chosen to ensure that both systems have a comparable number of parameters. However, the proposed narrowband system exhibits a higher number of trainable parameters (2.82 M) in comparison to the broadband counterpart (1.55 M). Table 2.1 also shows that while the proposed broadband and narrowband systems exhibit a larger number of trainable parameters compared to their respective baseline counterparts, the difference in the number of trainable parameters is especially noticeable for the narrowband systems. To the best of our knowledge, it is not possible to directly implement the considered systems in current hearing devices. This may be possible after model size optimization, model quantization and pruning, which is however not the main topic of this study.

2.5 Experimental setup

In this section, we conduct experiments to assess the performance of the speech-aware systems proposed in Section 2.4.2 in comparison to the baseline systems described in Section 2.4.1. Furthermore, we provide details of the datasets utilized in this study in Section 2.5.1, and describe the procedures for generating training and evaluation data in Sections 2.5.2 and 2.5.3, respectively. Additionally, we present implementation details of the input features and the VAD in Section 2.5.4, and describe the training procedure and hyperparameters of the CNNs used in this study in Section 2.5.5. Evaluation metrics employed to assess the performance of all systems are described in Section 2.5.6.

Table 2.2: Summary of the training data

Source signals	Speech (TIMIT) and non-speech (ESC50)
Environment	Anechoic [1]
Background noise	Simulated diffuse noise
SNR	-5 dB to $+20$ dB in 5 dB steps
Source-to-head distance	3 m
Source positions	72 positions in the horizontal plane

Table 2.3: Summary of the evaluation data

	Static Scenario	**Dynamic Scenario**
Source signals	Speech (TIMIT)	Speech (TIMIT)
Environment	Cafeteria ($T_{60} \approx 1.3$ s) and Courtyard ($T_{60} \approx 0.9$ s) [1]	Office ($T_{60} \approx 0.3$ s) [1]
Background noise	Simulated diffuse noise and Recorded noise	Simulated diffuse noise
SNR	-5 dB to $+10$ dB in 5 dB steps	-5 dB to $+10$ dB in 5 dB steps
Source-to-head distance	1–1.6 m	1 m
Source positions	4 source positions with 2 head orientations in each environment	Trajectories from $-90°$ to $+90°$

2.5.1 Datasets

Signals from speech and non-speech datasets were used as sound source signals to generate the training and validation data required during the training of all systems. In particular, speech signals of 462 and 168 speakers from the TIMIT dataset [152] (including both male and female speakers) were used for training and validation, respectively. In addition, three categories (natural soundscapes and water sounds, interior and domestic sounds, exterior and urban noises) of the ESC50 dataset [153] were used as non-speech signals, where we used 960 and 240 distinct sound files for training and validation, respectively. For evaluation, only speech signals from the validation TIMIT dataset were used as source signals.

We used a database of multichannel binaural room impulse responses (BRIRs) [1] to generate data for training and evaluation. The considered binaural hearing aid setup consists of $M = 4$ microphones, where the front and rear microphones (approximate microphone distance of 15 mm) in both left and right hearing aids were used. The database in [1] contains BRIRs measured in anechoic conditions for different source-to-head distances, and for $C = 72$ directions in the azimuthal plane, i.e., with a resolution of 5°. This dataset also contains BRIRs in three reverberant environments (cafeteria with $T_{60} \approx 1.3$ s, courtyard with $T_{60} \approx 0.9$ s, office with $T_{60} \approx 0.3$ s). We generated the noisy binaural microphone signals by convolving the source signals with BRIRs and mixing the resulting clean binaural microphone signals with background noise. All systems were trained in noisy anechoic conditions and evaluated in noisy reverberant environments.

2.5.2 Training data

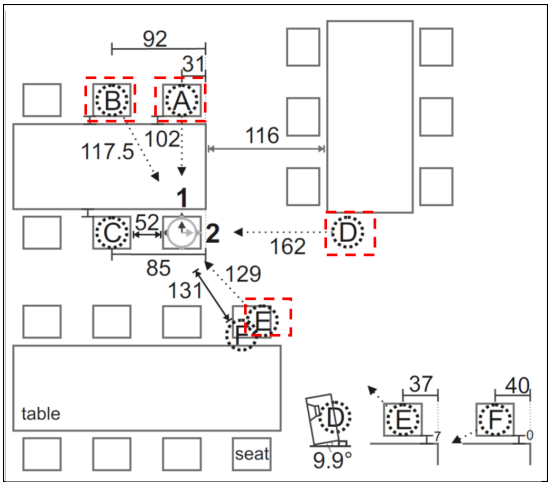
For training, the clean binaural microphone signals were generated by convolving both speech and non-speech source signals with anechoic BRIRs for each of the 72 directions with a source-to-head distance of 3 m. The noisy binaural microphone signals were generated by mixing the clean binaural microphone signals with simulated binaural diffuse noise at SNRs ranging from -5 dB to $+20$ dB in 5 dB steps. The noise at the microphones was generated by convolving uncorrelated speech-shaped noise from the ICRA noise database [154] with anechoic BRIRs, and summing all resulting binaural signals from 72 directions. Training examples were constructed for both speech and non-speech signals for all 72 directions at six different SNRs. It is important to note that in a data pre-processing step, a simple oracle broadband energy-based VAD was employed to identify segments containing enough speech content. This step ensures that for training examples associated with a speech source, only those containing meaningful speech content contribute to the loss function. Each training example consists of a block of $L = 20$ consecutive time frames (corresponding to 105 ms). In total, we obtained 5.9 million examples (about 172 hours) as *training set* and 2.4 million examples (about 70 hours) as *validation set*. A summary of the training data is presented in Table 2.2.

2.5.3 Evaluation data

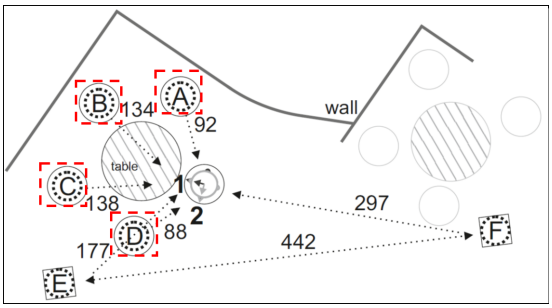
The performance of the baseline and proposed systems was evaluated for static and dynamic source scenarios in reverberant environments. As already mentioned, only speech signals from the validation TIMIT dataset were used as source signals. It should be noted that the source and background noise signals, acoustic conditions and source positions used during evaluations were different from those used during training and validation. A summary of the evaluation setup and data generation is presented in Table 2.3.

2.5.3.1 Static source scenario

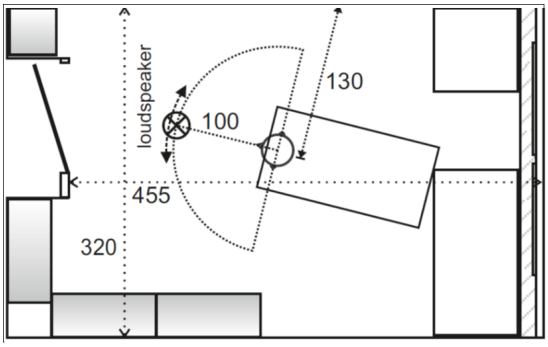
for the static source scenario, we considered two real environments (cafeteria and courtyard) with a reverberation time of approximately 1300 ms and 900 ms, respectively. The clean binaural microphone signals were generated by convolving the speech source signals with reverberant BRIRs[1]. The room configurations of both environments are depicted in Figs. 2·5a and 2·5b. In each environment, we considered four source positions (specified with dashed boxes), with two head orientations for each source position. All systems were evaluated at SNRs ranging from -5 dB to $+10$ dB either with matched or unmatched background noise. The same binaural diffuse noise as that used during training was utilized for the matched noise condition, whereas recorded cafeteria babble noise and courtyard ambient noise [1] were used for the unmatched noise condition. A total number of 150 speech segments randomly chosen from 30 unique male and female speakers (each with a length of 1 s) were selected from the validation TIMIT dataset.



(a) Cafeteria (static scenario)



(b) Courtyard (static scenario)



(c) Office (dynamic scenario)

Figure 2·5: Evaluation setups for static scenarios (a and b), and dynamic scenarios (c), adapted from [1]. In the cafeteria, static source positions A, B, D, E were considered, while in the courtyard, static source positions A, B, C, and D were considered. In both environments, the cafeteria and courtyard, the head position is indicated by dashed arrows extending from the source positions to the head. The two head orientations are denoted by the numbers 1 and 2 positioned in proximity to the head. For the dynamic scenarios in the office, the source position traveled from left to right with respect to the look direction of the head, either between $-45°$ to $+45°$ ($6\ °/s$ angular velocity) or between $-60°$ to $+60°$ ($8\ °/s$ angular velocity). The head position in the office environment is depicted in the middle of the room. All distances are specified in centimeters.

2.5.3.2 Dynamic source scenario

In [1], BRIRs of a reverberant office environment with a reverberation time of approximately 300 ms (specified in [1] as *Office I*) are provided, which cover the frontal azimuth range from $-90°$ to $+90°$ with a $5°$ resolution. Since only the BRIRs measured in the office environment allow to simulate moving sources, simulations for the dynamic source scenario were only performed for the office environment. To simulate a moving source, a time-aligned interpolation method [155] with shape-preserving piecewise cubic interpolation was used to interpolate the original

BRIRs from a 5° resolution to a 0.5° resolution. A total number of 10 speech segments were randomly chosen from 30 unique male and female speakers (each with a length of 15 s) from the validation TIMIT corpus. The clean binaural microphone signals were simulated for two source velocities (6 $^\circ/s$ and 8 $^\circ/s$ angular velocity) by partial convolution of the interpolated BRIRs with the clean speech signal using a frame length of 10 ms and 50% overlap. The office room configuration and the source movement trajectory are depicted in Fig. 2·5c. Simulated binaural diffuse noise was used to generate noisy binaural microphone signals at SNRs ranging from −5 dB to +10 dB.

2.5.4 Implementation details

All signals were sampled at 16 kHz. To compute the GCC-PHAT and CPS features, the microphone signals were transformed to the STFT domain using a Hann window of length $K = 160$ (corresponding to 10 ms), and a hop size of length $D = 80$ (corresponding to 5 ms), resulting in 81 STFT frequency bins. To compute GCC-PHAT features, we used an upsampling factor of $\kappa = 4$. In the case of a pair of microphones located on the same side of the head (left or right), the corresponding maximum delay τ_i^{max} is considered as 2, which translates to a time delay of 125 μs and $\mathcal{T}_i = 17$. Conversely, for a pair of microphones located on opposite sides, the maximum delay τ_i^{max} is considered as 20, corresponding to a time delay of 1.25 ms and $\mathcal{T}_i = 161$. We note that the chosen maximum delays are deliberately set to be greater than the maximum possible delay that can occur for each microphone pair by approximately a factor of two. In the considered binaural hearing aid setup, there are a total of four microphone pairs on opposite sides and two microphone pairs on the same side. As a result, GCC-PHAT feature vectors of size $\mathcal{T} = 678$ are obtained. For feature extraction, a block of $L = 20$ consecutive time frames is employed, leading to a GCC-PHAT input feature of size 678×20. For each pair of CPS features (real and imaginary parts, magnitude and phase components), the size of the input features is equal to $81 \times 20 \times 12$.

In this paper, we consider the front microphone of the left hearing aid as the reference microphone for the PD feature extraction, and also for the binary VAD decision employed in the baseline systems. To obtain a binary VAD decision on a block of L consecutive time frames, a majority vote rule is applied, where the block is classified as speech if at least 50% of the time frames are detected as such. In the baseline systems, we used the pitch-based binary VAD [138] (rVAD) with its original frame length but adjusted the frame hop size to 5 ms, aligning it with the proposed systems while keeping its spectral resolution unchanged.

PD features were computed using a 4-th order GTFB implementation [48] with 61 subbands, a group delay of 256, and minimum and maximum center frequencies of 60 Hz and 7200 Hz, respectively. By choosing the maximum and minimum fundamental frequencies as 320 Hz and 70 Hz, respectively, the range of fundamental period candidates for PD feature extraction lies between 3.1 ms and 14.3 ms for $N = 180$ period candidates. To compute the subband-averaged PD features, $F = 6$ frequency bands are obtained by averaging every $S = 10$ subband signals. The comb filter gain was chosen to be $\alpha = 0.7$. The size of the broadband and subband-averaged input PD features is equal to 180×20 and $180 \times 20 \times 6$, respectively.

2.5.5 Training and network hyperparameters

All systems were implemented using PyTorch [156]. For all CNNs, we used a 2D convolutional filter size of 3×3 with a stride size of 1×1. In each convolutional layer of the CNNs with broadband (GCC-PHAT) and narrowband (CPS) input, 4 and 32 filters were used, respectively. The max-pooling size was 2×2 with strides of the same size. The CNNs were trained using the Adam optimizer [157], a cross-entropy loss function, an initial learning rate of 10^{-5}, a mini-batch size of 128 and a dropout rate of 0.5. We used an early stopping regularization method which stopped the training if no improvement in validation loss was observed for 4 epochs, and a variable learning rate scheduler to halve the learning rate if the validation loss did not improve for 2 epochs.

The maximum epoch number for training all CNNs was set to 100. In each epoch, 1.63 million examples were randomly selected from the training set such that the network did not see the same example twice. Each mini-batch included 128 examples that were randomly chosen from different SNR conditions, DOA classes, and speech and non-speech signals. To calculate the validation loss at the end of each epoch, 200000 examples were randomly selected from the validation set and kept fixed throughout the training. The validation data were not seen by the network during the training.

2.5.6 Evaluation metrics

We evaluated the DOA estimation performance of the proposed and baseline systems in terms of mean absolute error (MAE) and accuracy (Acc.) [14, 98]. A DOA estimate in block l is considered accurate if the absolute error between the estimated DOA $\hat{\theta}_l$ and the oracle DOA θ_l is smaller than 5°, i.e., the minimum angular resolution of the database in [1]. The MAE (in degrees) and accuracy are defined as

$$\text{MAE} = \frac{1}{\mathcal{L}} \sum_{l=1}^{\mathcal{L}} \left| \hat{\theta}_l - \theta_l \right|, \tag{2.15}$$

$$\text{Acc} = \frac{\mathcal{L}_{\text{acc}}}{\mathcal{L}} \times 100, \tag{2.16}$$

where $\mathcal{L}$ denotes the total number of estimates, i.e., the number of signal blocks with positive speech detections, and $\mathcal{L}_{\text{acc}}$ denotes the total number of accurate estimates.

We evaluated the speech detection performance of the VAD used in the baseline systems and the performance of the speech DOA detection in the proposed system using the precision (P) and recall (R) metrics defined as

$$\text{P} = \frac{TP}{(TP + FP)}, \tag{2.17}$$

$$\text{R} = \frac{TP}{(TP + FN)}, \tag{2.18}$$

where for each evaluated system, the number of true positives (TP) represents the total number of signal blocks detected as speech by both the system and the oracle VAD, while the number of

false positives (FP) represents the total number of signal blocks detected as speech by the system but detected as non-speech by the oracle VAD. Conversely, the number of false negatives (FN) denotes the total number of signal blocks detected as non-speech by the system but detected as speech by the oracle VAD. While precision indicates the proportion of detected speech blocks that are actually correct, recall represents the proportion of actual speech blocks that are detected by the system. Both metrics range from 0 to 1.

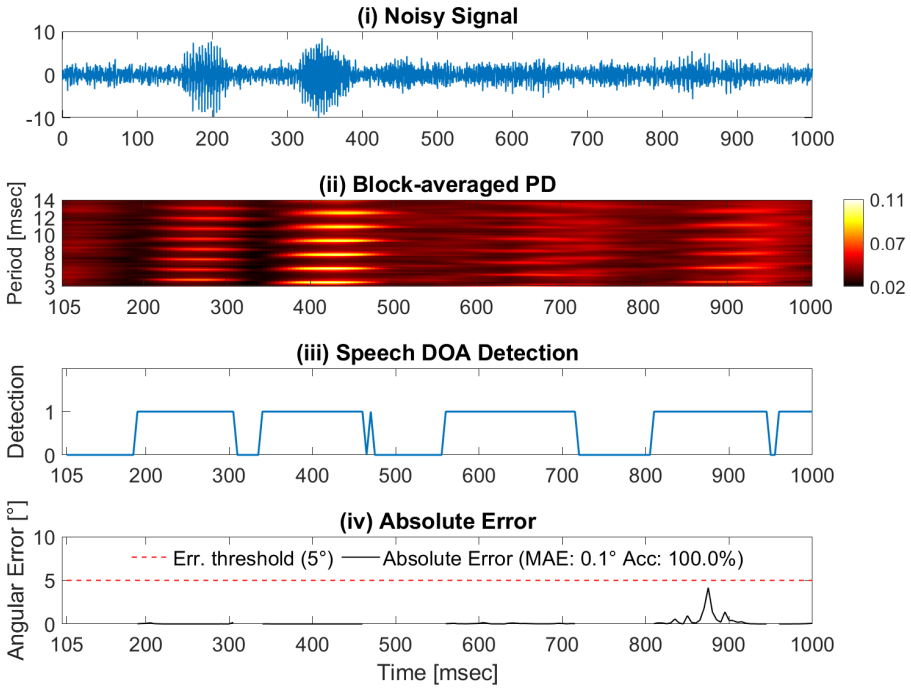
2.6 Results and discussion

In this section, we will present and analyze the performance evaluation results of speech-aware systems employing either broadband or narrowband feature combinations, in comparison to baseline systems. The baseline systems consist of a CNN that uses only spatial features, combined with a pitch-based VAD. We assessed the performance of all systems in various reverberant environments with different background noises for both static and dynamic single-talker scenarios in terms of accuracy and mean absolute error for DOA estimation, as well as precision and recall for speech detection. Section 2.6.1 serves as an exemplary demonstration of the proposed speech-aware DOA estimation using broadband input features. The evaluation results for static source scenarios in both matched and unmatched noise conditions are discussed in Section 2.6.2. The evaluation results for dynamic source scenarios are presented in Section 2.6.3. Finally, we discuss the limitations of this study and suggest potential future works in Section 2.6.4.

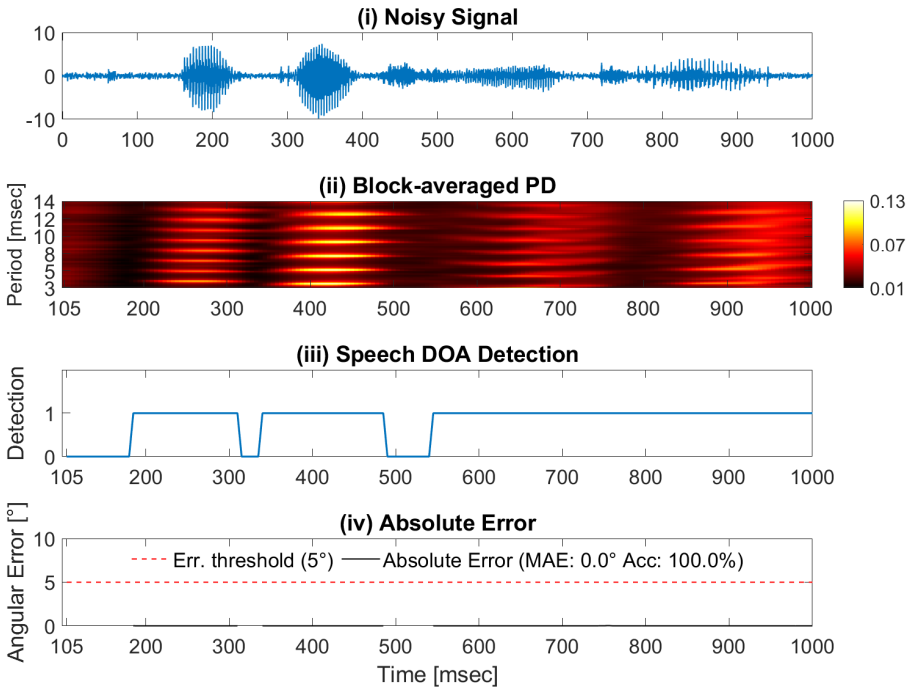
2.6.1 Speech-aware DOA estimation

To illustrate speech-aware DOA estimation, we consider an exemplary static source scenario in the courtyard (cf. Fig. 2·5b) for a female speech source at position C and head orientation 1 (corresponding to a DOA of $-20°$) with simulated diffuse noise at 0 dB and 10 dB SNR conditions. The proposed system with broadband input features (Fig. 2·4a) is chosen for DOA estimation in this scenario. Fig. 2·6b depicts the noisy reference microphone signal with a duration of 1 s, the corresponding block-averaged representation of the PD feature, the speech DOA detection (cf. (2.11) and (2.13)), and the DOA estimation error. Please note that the difference in the starting times between the reference microphone signal in subfigure (i) and the subsequent subfigures (ii-iv) is due to the design of the proposed system, which requires input features from consecutive time frames over a period of 105 ms before generating the first prediction. To aid in visualization, we obtained the block-averaged PD by averaging the PD values over consecutive time frames used by the CNN for each prediction.

When analyzing Fig. 2·6b, several key observations emerge. First, comparing speech DOA detection results in the two SNR conditions (Fig. 2.6a.iii and Fig. 2.6b.iii) shows that the speech-aware DOA estimation results in fewer signal blocks with DOA detections in the low SNR condition compared to the high SNR condition. Second, comparing the DOA detection results with the absolute error in either of the SNR conditions, e.g., in the low SNR condition (Fig. 2.6a.iii and Fig. 2.6a.iv), demonstrates that for this example, all estimated DOAs result in absolute errors below 5°, i.e., 100% accuracy. These findings illustrate the primary objective in designing the speech-aware DOA estimation systems, which is to reliably detect speech DOAs



(a) 0 dB SNR condition



(b) 10 dB SNR condition

Figure 2·6: An exemplary illustration of DOA estimation of the proposed system using the broadband feature combination in (a) SNR = 0 dB and (b) SNR = 10 dB: (i) Noisy reference microphone signal of a source at $-20°$ with simulated diffuse noise, (ii) Estimated block-averaged broadband PD, (iii) Speech DOA detection, (iv) Absolute angular error of the estimated DOAs over time specified by black lines, and the 5° error threshold specified by a red dashed line.

while excluding signal blocks prone to poor DOA estimation performance, without needing a separate VAD. As expected, such a system detects fewer signal blocks with reliable speech DOA in the low SNR condition. Moreover, when comparing block-averaged PD with DOA detection results, especially in the low SNR condition (Fig. 2.6a.ii and Fig. 2.6a.iii), it becomes evident that the proposed system predominantly estimates the DOA for blocks with a high degree of periodicity. These observations are noteworthy because they demonstrate that the proposed system automatically selects the most reliable signal blocks for DOA estimation, primarily those with a high degree of periodicity, which are less susceptible to noise.

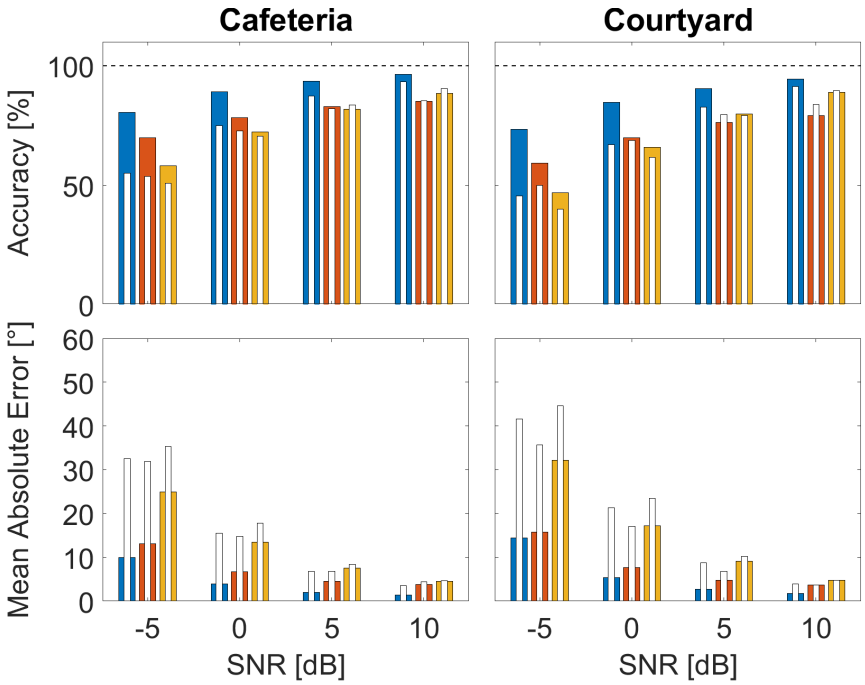
2.6.2 Evaluation results for static source scenarios

For the static source scenarios in two reverberant environments (cafeteria and courtyard), Fig. 2·7 shows the accuracy and mean absolute error at different SNRs for three proposed systems (Section 2.4.2) and three baseline systems (Section 2.4.1) using either broadband or narrowband features. Performance measures of three proposed systems, i.e., the proposed system with broadband PD and GCC-PHAT (Prop. broadband), the proposed system with subband-averaged PD and real and imaginary parts of CPS (Prop. narrowband ReIm), and the proposed system with subband-averaged PD and magnitude and phase parts of CPS (Prop. narrowband MagPhase) are depicted by colored bars. To facilitate the direct comparison between each proposed system and the corresponding baseline system using the same spatial feature, white narrow bars in front of the colored bars show the performance measures of the corresponding baseline system. A dashed line in the top plots of each figure shows the maximum accuracy of 100% that each system can achieve. In addition to the DOA estimation metrics depicted in Fig. 2·7, Fig. 2·8 shows the speech detection evaluation results in terms of precision and recall at different SNR conditions for three proposed systems (Section 2.4.2) and the rVAD [138] used in the baseline systems (Section 2.4.1).

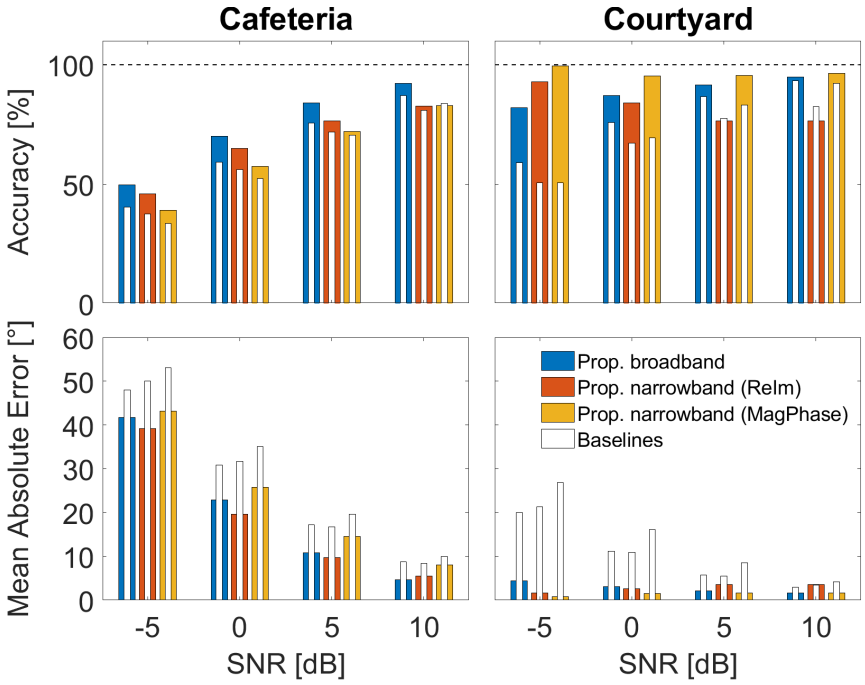
2.6.2.1 Matched noise condition

Fig. 2·7a depicts the performance measures for the matched noise condition with simulated diffuse background noise (also used during training). Comparing the performance of the proposed systems (colored bars) with the corresponding baseline systems (white bars), we can clearly observe the benefit of using PD in combination with a spatial feature in both environments for all systems at low SNRs (-5 dB and 0 dB), whereas the benefit also persists for the broadband system at high SNRs (5 dB and 10 dB). For example, for an SNR of 0 dB in the cafeteria environment, the benefit of using PD features in terms of accuracy is approximately 14% points for the broadband system, 6% points for the narrowband system (ReIm), and 1% points for the narrowband system (MagPhase), whereas the benefit in terms of MAE is 11.7° for the broadband system, 8° for the narrowband system (ReIm), and 3.8° for the narrowband system (MagPhase).

Fig. 2·8a depicts the speech detection performance measures for the matched noise condition. It can be observed that all proposed systems exhibit nearly perfect precision, approaching 1. This suggests a low likelihood of falsely detecting a signal portion for DOA estimation (i.e. low false positive). It can be clearly seen for all conditions that the proposed systems demonstrate

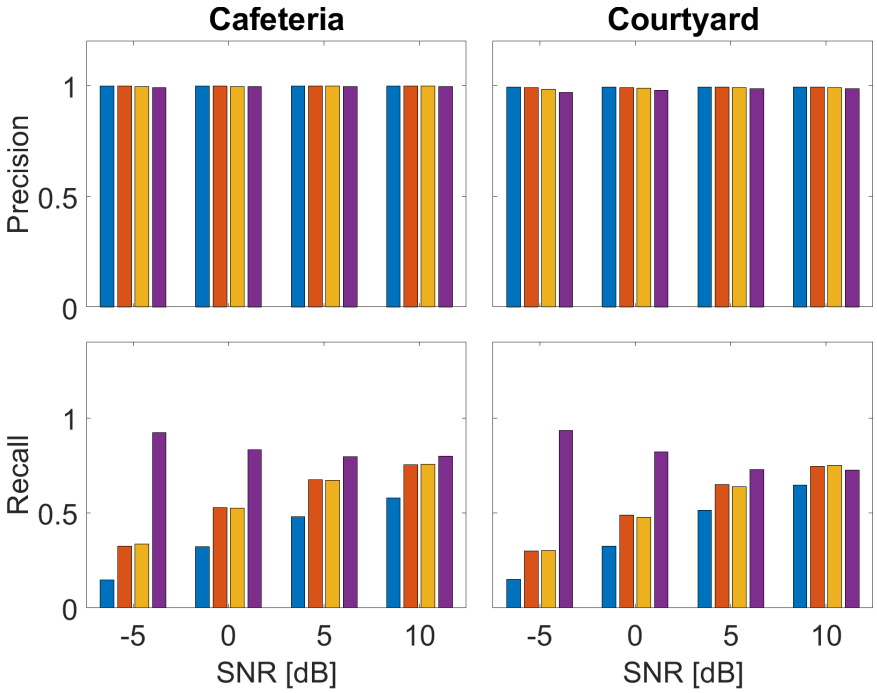


(a) simulated diffuse background noise (matched condition)

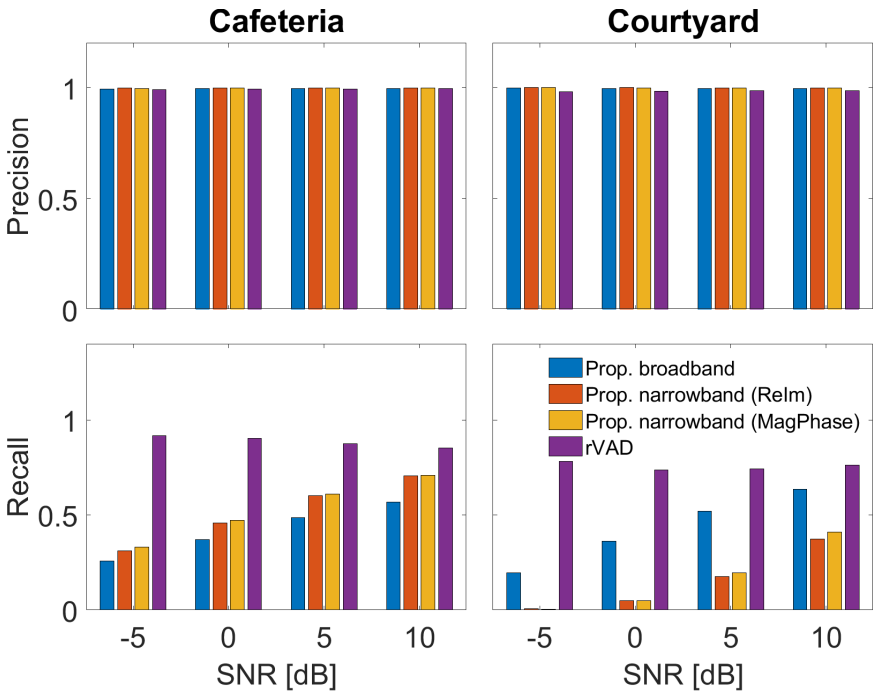


(b) recorded background noise (unmatched condition)

Figure 2·7: Accuracy and mean absolute error of the proposed and baseline systems for the static source scenarios in two reverberant environments (cafeteria and courtyard) for different SNRs: (a) matched noise condition with simulated diffuse background noise and (b) unmatched noise condition with recorded background noise. Colored bars show the performance of the proposed systems using broadband feature combination (GCC-PHAT and broadband PD) or narrowband feature combination (either CPS ReIm or MagPhase and subband-averaged PD), whereas white bars show the performance of the baseline systems using only broadband (GCC-PHAT) or narrowband (either CPS ReIm or MagPhase) spatial features.



(a) simulated diffuse background noise (matched condition)



(b) recorded background noise (unmatched condition)

Figure 2·8: The speech detection performance of the proposed systems and rVAD in terms of the precision and recall for the static source scenarios in two reverberant environments (cafeteria and courtyard) for different SNRs: (a) matched noise condition with simulated diffuse background noise and (b) unmatched noise condition with recorded background noise.

either better or comparable precision compared to the rVAD, but a lower recall. It is important to emphasize once more that the rVAD is specifically designed for speech detection, whereas the proposed systems are designed for speech DOA detection. This distinction is crucial, as the proposed systems leverage an output class which indeed serves as an uncertainty measure for DOA estimation. Although this class is regarded as a detection class, it has not been merely trained for the speech detection task.

2.6.2.2 Unmatched noise condition

Fig. 2·7b depicts the performance measures for the unmatched noise condition with recorded background noise (not seen during training). Except for the narrowband system (MagPhase) in the cafeteria environment at 10 dB SNR and the narrowband system (ReIm) in the courtyard environment at 10 dB SNR, the proposed systems using PD in combination with a spatial feature outperform the corresponding baseline systems for all SNRs in both environments. For example, for an SNR of 0 dB in the cafeteria environment, the benefit of using PD features in terms of accuracy is approximately 10% points for the broadband system, 9% points for the narrowband system (ReIm), and 5% points for the narrowband system (MagPhase), whereas the benefit in terms of MAE is 8° for the broadband system, 12.1° for the narrowband system (ReIm), and 9.2° for the narrowband system (MagPhase).

Fig. 2·8b depicts the speech detection performance measures for the unmatched noise condition. It can be observed that in the courtyard environment, the proposed narrowband systems result in notably low recall, particularly in low SNR conditions. The very low recall in this condition corresponds to a high number of missed detections (i.e. high false negative). However, as observed in Section 2.6.1, it's essential to emphasize that the primary objective of speech-aware systems is to detect the speech DOA for reliable localization, rather than solely focusing on speech activity detection. The good results for the proposed narrowband systems at low SNRs in the courtyard in the unmatched condition (Fig. 2·7b) can be attributed to the fact that these systems use only a small fraction of speech signal blocks for DOA estimation.

When comparing the performance measures between the matched and unmatched noise conditions (Fig. 2·7a and Fig. 2·7b), it can be clearly observed that in the cafeteria environment the performance for the recorded babble noise is worse than that for the simulated diffuse noise, whereas (somewhat surprisingly) in the courtyard environment the performance for the recorded ambient noise is better than that for the simulated diffuse noise. This can be explained by investigating the spectro-temporal sparsity of the signals for the different conditions. For the sparsity analysis, we use the Gini index [158], where a large Gini index (close to 1) corresponds to high sparsity, and a small Gini index (close to 0) corresponds to low sparsity. More in particular, we consider the joint spectro-temporal Gini index according to [159], computed on the STFT spectrogram of the noisy reference microphone signal. For each environment and background noise type and for different SNRs, Fig. 2·9 depicts the spectro-temporal Gini index averaged over all 150 speech segments. On the one hand, in the cafeteria environment, it can be observed for all SNRs that the spectro-temporal sparsity of the microphone signals with recorded babble noise is less than the spectro-temporal sparsity with simulated diffuse noise. On the other hand,

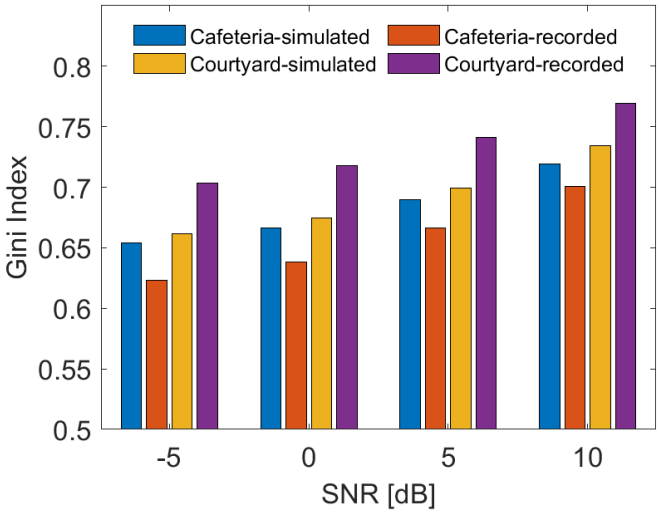


Figure 2·9: The average spectro-temporal Gini index for two environments (cafeteria, courtyard), two background noise types (simulated, recorded) and different SNRs.

in the courtyard environment, it can be observed for all SNRs that the microphone signals with recorded ambient noise exhibit a sparser spectro-temporal structure than with the simulated diffuse noise. Hence, in conjunction with the DOA estimation performance in Fig. 2·7, we can deduce that signals with sparser spectro-temporal structure appear to lead to better speech-aware DOA estimation.

Taking a closer look at Fig. 2·7b, it becomes evident that, in the courtyard environment with recorded background noise, the two proposed narrowband systems perform best under the lowest SNR condition (0 dB SNR). The Gini index, however, does not provide a comprehensive explanation for this particular case. Unlike the simulated diffuse noise and cafeteria babble noise, the courtyard ambient noise energy predominantly falls within the first frequency band of PD features. This means that at low SNRs, especially at -5 dB, the noise can mask the harmonic structure of speech signals in this frequency band. This masking potentially aids the CNN in almost perfectly identifying segments with prevalent noise, enhancing DOA estimation accuracy. As SNR increases, enhanced harmonics in low frequencies may introduce uncertainties, potentially compromising DOA estimation accuracy. However, this does not affect our main findings and conclusions for speech-aware DOA estimation.

2.6.3 Evaluation results for dynamic source scenarios

For the moving source scenario in the office environment, Fig. 2·10 depicts the DOA estimation performance measures of the proposed systems (colored bars) and the corresponding baseline systems using the same spatial feature (white bars) for different SNRs and two angular velocities. Similarly as for the static source scenario (Fig. 2·7), a clear benefit of using PD features can be observed, especially for the broadband system at all SNRs and for the narrowband system (ReIm) at low SNRs. For the narrowband system (MagPhase), whose performance is anyway

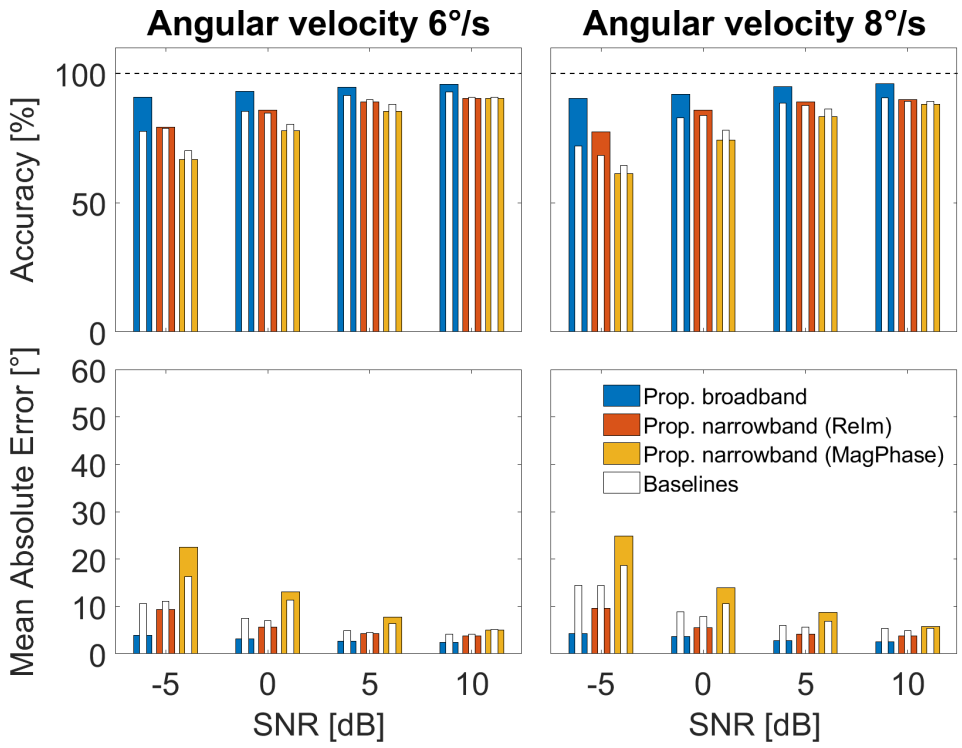


Figure 2·10: Accuracy and mean absolute error of the proposed and baseline systems using broadband and narrowband features for the dynamic source scenarios in the office environment for different SNRs and angular velocities.

lower than the narrowband system (ReIm), the baseline system (using MagPhase) exhibits comparable or better performance. These results further reveal the benefit of using PD features in the proposed speech-aware DOA estimation systems compared to the baseline systems using merely spatial features. This benefit even increases with angular velocity, particularly at low SNRs.

For the moving source scenario, the evaluation results of speech detection performance for all considered systems are illustrated in Fig. 2·11. It becomes evident that in dynamic scenarios across all conditions, the proposed narrowband systems yield a higher recall when compared to all other systems (including rVAD), while maintaining a high level of precision. This is particularly noteworthy, as the higher recall facilitates speech source tracking by generating more observations of the dynamic scene.

Evaluation results in Fig. 2·7 and Fig. 2·10 show that, except for the matched condition in the static source scenario, the proposed broadband system outperforms the proposed narrowband systems, while indicating a larger benefit from the inclusion of PD features. The results also demonstrate that the broadband baseline system using GCC-PHAT features typically outperforms narrowband baseline systems using CPS features. Despite a similar number of trainable parameters (cf. Table 2.1), the narrowband baseline systems must learn more intricate patterns from CPS features, whereas GCC-PHAT directly provides time delay information. This suggests that the narrowband baseline systems may need more capacity (trainable parameters) to match the performance of the broadband one. As our main goal was to study the benefit

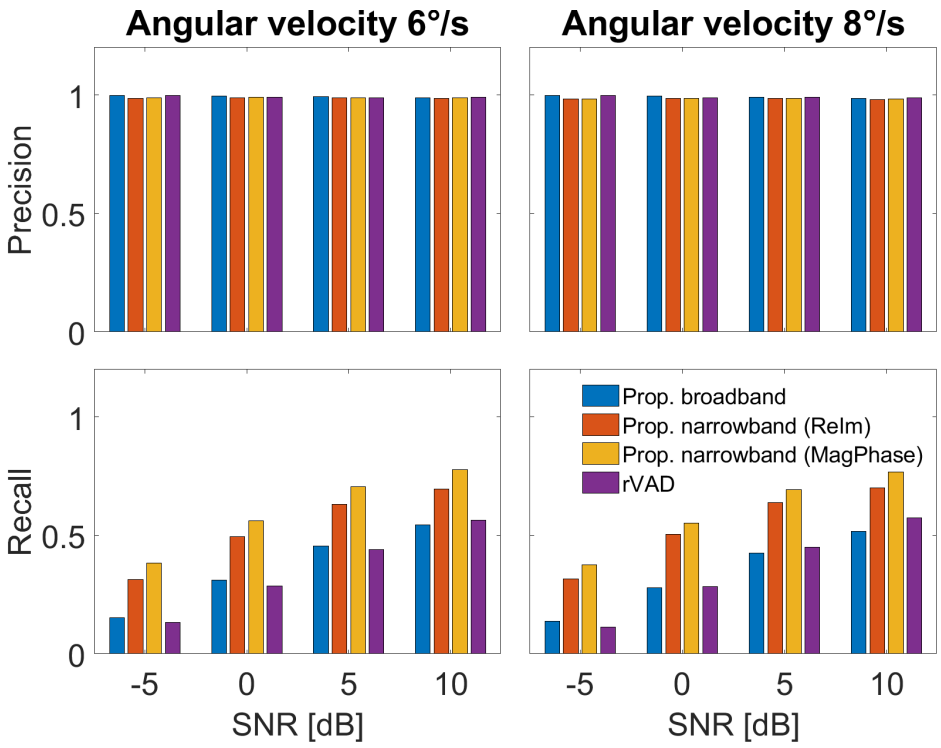


Figure 2·11: The speech detection performance of the proposed systems and rVAD in terms of the precision and recall for the dynamic source scenarios in the office environment for different SNRs and angular velocities.

of using PD features in the proposed narrowband and broadband systems, we didn't optimize the narrowband systems for performance parity with the broadband system, potentially causing performance limitations when combining CPS and PD features.

2.6.4 Limitations and future works

This study only considered binaural DOA estimation of a far-field speech source. For a speech source in the near field of a microphone array, accurate estimation of the time delay (and phase) involves considering both the range and the DOA of the sound source. The normalization inherent in the PHAT weighting (see (2.4)) eliminates the effect of the signal level (and hence range information) due to the source-microphone distance. Consequently, a model trained solely on the GCC-PHAT may have limited capability to leverage range-dependent information in the near-field scenarios. Although this study only considered binaural DOA estimation in the azimuthal plane, the proposed systems, in principle, can be extended for DOA estimation in terms of both azimuth and elevation as azimuth and elevation information are encoded by the spatial input features [12, 18].

In this study, we examined single-talker speech-aware DOA estimation in the presence of background noise. Future research may explore the potential benefits of using PD features for speech-aware DOA estimation in the presence of non-speech interference and binaural DOA estimation in multi-talker scenarios.

2.7 Conclusion

In this study, we proposed novel feature combinations for speech-aware DOA estimation in the context of binaural hearing aids. The proposed systems utilize CNNs and receive a spatial feature and an auditory-inspired periodicity feature as inputs to two parallel branches of convolutional layers. In particular, we introduced a subband-averaged PD feature as the periodicity feature, and combined it with either the real and imaginary or the magnitude and phase components of the narrowband CPS as the spatial feature. The performance of speech-aware systems was evaluated against CNN-based baseline systems which only use spatial features and a pitch-based VAD.

Comprehensive evaluations in static single-talker scenarios with different background noise types and SNRs demonstrate that for any type of spatial feature, the proposed method outperforms baseline systems in terms of DOA estimation accuracy and mean absolute error, particularly in adverse SNR conditions and in conditions with higher degrees of spectro-temporal sparseness. This study also shows that the proposed method using PD features is effective for speech-aware DOA estimation of a moving talker, and is robust to changes in talker velocity. Our proposed speech-aware system is able to estimate the sound source DOA when a high degree of periodicity is captured by the CNN, without any need for a separate VAD or pitch period estimation.

The primary finding of this study was that the usage of PD features in both narrowband and broadband feature combinations benefits the speech-aware binaural DOA estimation in different static and dynamic scenarios. It was also found that the proposed system employing the broadband feature combination typically demonstrated better performance than the proposed systems using the narrowband feature combinations in the specific system configuration employed in this study.

Overall, this study demonstrates the potential benefits of utilizing periodicity-based features in conjunction with spatial features for speech-related applications such as DOA estimation. The results also suggest that these features may have wider applications in other speech-related tasks. The findings of this study can contribute to the development of improved methods for sound source localization and speech enhancement in binaural hearing aids.

Chapter 3

A Two-stage CNN with Feature Reduction for Speech-aware Binaural DOA Estimation

---

This chapter is a reformatted reprint of the following publication:

R. Varzandeh, S. Doclo, V. Hohmann, "A two-stage CNN with feature reduction for speech-aware binaural DOA estimation," in Proc. *European Signal Processing Conference (EUSIPCO)*, Helsinki, Finland, 2023, pp. 241-245.

This chapter proposes an efficient two-stage CNN system for single-talker DOA estimation. This approach merges and processes the periodicity and spatial features within a single convolutional pathway. To achieve this efficiently, the system utilizes 1×1 convolutions to transform the periodicity features into compact features before combining them with spatial features. This leads to an improved performance compared to a system without feature reduction. This approach also performs better than a baseline system consisting of a CNN using only spatial features cascaded with a VAD. In this chapter, we focus on single-talker DOA estimation in the presence of background noise. In Appendix A, we evaluate the proposed and baseline systems under single-talker scenarios with directional non-speech interference.

Abstract

In recent years, several supervised learning-based approaches have been proposed to estimate the direction of arrival (DOA) of a single talker in noisy and reverberant environments. In this paper, we consider a speech-aware DOA estimation system for binaural hearing aids, which does not require a separate voice activity detector (VAD). We propose the combination of two narrowband features as the input features of a convolutional neural network (CNN), namely the cross-power spectrum as spatial features and narrowband auditory-inspired periodicity features. Prior to the joint processing of both features, we propose to reduce the dimensionality of the narrowband periodicity features using a feature reduction stage based on 1×1 convolutions. Simulation results for two reverberant environments with different background noises demonstrate the benefit of the feature reduction stage in terms of DOA estimation accuracy while significantly reducing the number of trainable parameters. In addition, simulation results show that the proposed system outperforms a baseline system consisting of a CNN using only spatial features and a pitch-based VAD.

3.1 Introduction

Reliably estimating the DOA of a talker is a crucial task in applications such as binaural hearing aids [160, 161]. In addition to model-based DOA estimation approaches [8, 80, 85, 162], in recent years several supervised learning-based DOA estimation approaches based on DNNs have been proposed [12, 14, 17, 23, 98, 99, 102]. In these approaches, the DOA estimation task is often formulated as a classification problem, aiming at determining a mapping from input features to a spatial probability map for a discretized DOA range. Without auxiliary information, e.g., a VAD, such approaches also provide a DOA estimate during speech pauses or when the signal is dominated by noise, which typically results in erroneous DOA estimates. Hence, a VAD is often cascaded with a DOA estimation system [21, 23]. However, a separate VAD usually requires manual and time-consuming parameter tuning, and may introduce errors that propagate through the DOA estimation system.

In [143], we proposed a speech-aware binaural DOA estimation system based on CNNs, which does not require a separate VAD. Simulation results showed the benefit of using broadband PD features in combination with GCC-PHAT features as input features for the CNN. However, the frequency integration of the CPS phase employed in the calculation of the GCC-PHAT feature [80, 143] does not allow the CNN to effectively exploit the sparsity property of speech signals in the time-frequency domain [163]. In addition, broadband PD only offers a coarse representation of the harmonic structure of a signal.

In this paper, we extend the speech-aware binaural DOA estimation system of [143] in two ways. First, aiming at exploiting the sparsity property of speech signals, we propose to use a narrowband representation of PD features in combination with narrowband CPS features (as spatial features) as input features for the CNN. Second, the key contribution of this paper is introducing a PD feature reduction stage before the joint processing of both narrowband features, resulting in a two-stage CNN architecture. We postulate here that the feature reduction stage better guides the DOA estimation by reducing the sparse structure of narrowband PD features to a set of more compact spectro-temporal features, referred to as PD saliency features. Evaluation

results for a single talker in two reverberant environments for different SNRs show the benefit of using the proposed PD feature reduction stage compared to a system without feature reduction. Evaluation results also show that the proposed systems combining narrowband CPS and PD features outperform a baseline system, consisting of a cascade of a CNN using only narrowband CPS features and a pitch-based VAD.

3.2 DOA estimation as a classification problem

In this work, we consider the problem of single-talker DOA estimation in the azimuthal plane using a binaural hearing aid setup with M microphones. In the STFT domain, the m-th microphone signal at time frame n and frequency bin k (with K the STFT length) can be written as

$$Y_m(n,k) = X_m(n,k) + V_m(n,k), \tag{3.1}$$

where X and V denote the sound source (at direction θ) and the uncorrelated background noise, respectively. By dividing the azimuth range into a set of C discrete DOAs $\{\theta_1, \cdots, \theta_C\}$, DOA estimation can be considered as a classification problem, where the DOA of a sound source should be assigned to one of the DOA classes. In this work, we consider $C = 72$ classes for the full 360° azimuth range, corresponding to a DOA map with 5° resolution. In the next subsections two different classification-based approaches for DOA estimation will be discussed.

3.2.1 Conventional DOA estimation

Conventionally, DOA estimation is formulated as a C-class classification task, where each output class corresponds to a DOA [14, 23, 98, 99, 102]. During training, each training example belongs to only one output class that has been labeled using oracle DOA information. During testing, the neural network predicts a posterior probability map in the output. The DOA is usually estimated as the DOA class with the highest posterior probability. In this work, to obtain continuous DOA estimates from discrete DOA classes, we estimate the sound source DOA by employing parabolic interpolation [151] on three DOA classes centered around the DOA class with the highest posterior probability.

To deal with erroneous DOA estimates (e.g., during speech pauses), a VAD can be cascaded to this system [21, 23], where a DOA is only estimated from the probability map, if the VAD detects the signal as speech. In this work, we adopt the VAD-informed DOA estimation approach to design the baseline system in Section 3.4.1.

3.2.2 Speech-aware DOA estimation

In contrast to the VAD-informed classification-based approach, in [143] we proposed a classification-based approach referred to as speech-aware DOA estimation. The purpose of speech-aware DOA estimation is to estimate the DOA of a sound source only for speech sources, without needing a separate VAD. This problem is formulated as a $C+1$-class classification task, where the first C classes represent the DOA classes and the last class represents the non-speech activity, regarded as the detection class. During training, via a one-hot encoding scheme, if a training example

belongs to a speech source from a given direction, the DOA class corresponding to that direction is labeled by one, whereas all other classes (including the detection class) are labeled by zero. On the other hand, if a training example belongs to a non-speech source, regardless of its direction, all DOA classes are labeled by zero, whereas the detection class is labeled by one. During testing, we consider the class with the highest posterior probability. If this class is a DOA class, we estimate the sound source DOA by employing parabolic interpolation [151] on three DOA classes centered around this class. Otherwise, no reliable DOA could be estimated. In this work, we adopt the speech-aware DOA estimation approach in our proposed systems in Section 3.4.2.

3.3 Narrowband input features

Aiming at exploiting speech sparsity in the STFT domain, in this section we describe the narrowband features that are used as input features for the DOA estimation, namely the cross-power spectrum (Section 3.3.1) and the periodicity degree (Section 3.3.2).

3.3.1 Cross-power spectrum (CPS)

In [143] the broadband GCC-PHAT, defined as the inverse Fourier transform of the CPS phase, was used as the spatial input feature. In this work, we propose to directly use the narrowband CPS. The instantaneous CPS between the r-th and q-th microphone is defined as

$$G_i(n,k) = Y_r(n,k)Y_q^*(n,k), \tag{3.2}$$

where $(\cdot)^*$ denotes complex conjugate and i denotes a microphone pair combination. From (3.2) it can be seen that the CPS encodes both the phase difference and the levels of a microphone pair. As CPS input feature, we consider the real and imaginary parts of $G_i(n,k)$ for all $M(M-1)/2$ unique microphone pairs for frequencies up to the Nyquist frequency, i.e., $k = 0, 1, \cdots, K/2$, for L consecutive time frames. This means that the shape of the CPS input feature is equal to $L \times (K/2+1) \times 2M(M-1)/2$. We note here that the first, second, and third dimension represent the height, width, and depth of the input feature, respectively, where the depth corresponds to the number of input channels. For the CPS input feature, $2M(M-1)/2$ input channels are constructed by stacking the real and imaginary parts for all microphone pairs.

3.3.2 Periodicity degree (PD)

In [143] broadband PD features, which only offer a coarse representation of the harmonic structure of a signal, were used as input features. In this work, we propose to use a narrowband formulation of the PD features, estimated for a set of N fundamental period candidates. The PD features are computed by first decomposing a reference microphone signal into a set of bandpass-filtered time signals using a GTFB [48]. The real part of each bandpass-filtered signal is then passed through a half-wave rectification, followed by a fifth-order low-pass filter with 770 Hz cutoff frequency and a second-order high-pass filter with 40 Hz cutoff frequency, resulting in bandpass-filtered signal envelopes $y_{env}(t,f)$ in time t and subband f. In the next step, a set of

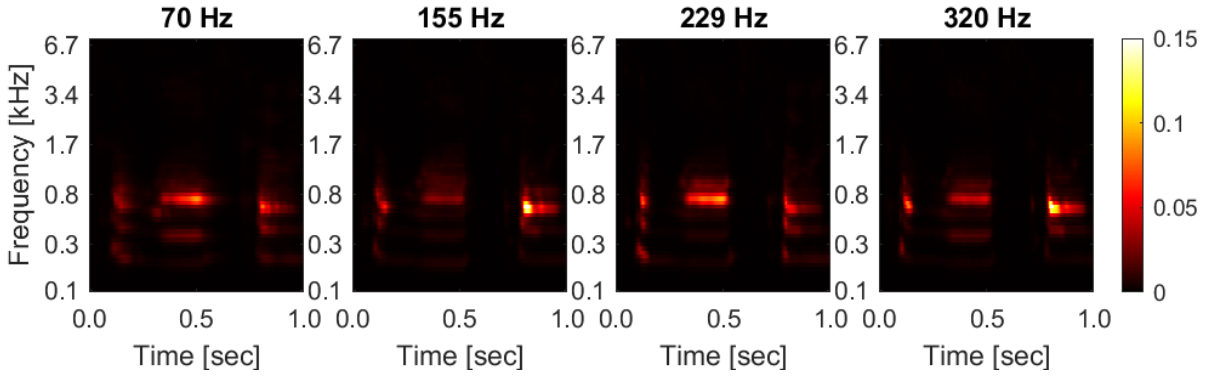


Figure 3·1: Illustrative visualization of narrowband PD features for a set of fundamental frequency candidates. The sparse spectro-temporal structure of these features motivates using a feature reduction stage prior to the joint processing of the CPS and PD features by the CNN.

N parallel IIR comb filters designed for a set of fundamental period candidates p_j,$j = 1, \cdots, N$, filter the signal envelopes as

$$s(t, f, j) = (1 - \alpha) y_{env}(t, f) + \alpha s(t - p_j, f, j), \tag{3.3}$$

where α denotes the filter gain. The periodicity degree is defined as the mean amplitude of the comb-filtered signal, computed as

$$PD(t, f, j) = (1 - \beta_j)|s(t, f, j)| + \beta_j PD(t - 1, f, j), \tag{3.4}$$

where $|\cdot|$ denotes the absolute value and the parameter β_j for each fundamental period candidate is defined as $\beta_j = e^{-1/p_j}$.

Since we aim at joint spectro-temporal processing of the PD and CPS features, it is required to represent both features at the same time-frequency resolution. To obtain the same time resolution as the CPS features, the PD features are averaged in each STFT frame. Unlike the linearly-spaced STFT frequency bands, the gammatone bands have a non-uniform frequency resolution that decreases with frequency. To obtain the same frequency resolution for the PD features as for the CPS features, for low STFT frequencies we average the PD features in gammatone bands associated with one STFT frequency band. In contrast, for high STFT frequencies we replicate the PD features of each gammatone band and assign them to those STFT frequency bands associated with one gammatone band. Similarly as for the CPS features, we consider L consecutive frames, such that the shape of the PD input feature is equal to $L \times (K/2 + 1) \times N$.

For a 1s clean signal of a female talker, Fig. 3·1 depicts exemplary 2D narrowband PD features, corresponding to a subset of fundamental frequency candidates (each representing an input channel). For a perfectly periodic signal with a certain fundamental frequency, a high PD value will be captured in each time-frequency bin across the N input channels associated with the harmonics and sub-harmonics of this fundamental frequency. Even though speech signals are not perfectly harmonic, their fundamental frequency variations and multiple harmonics exhibit a

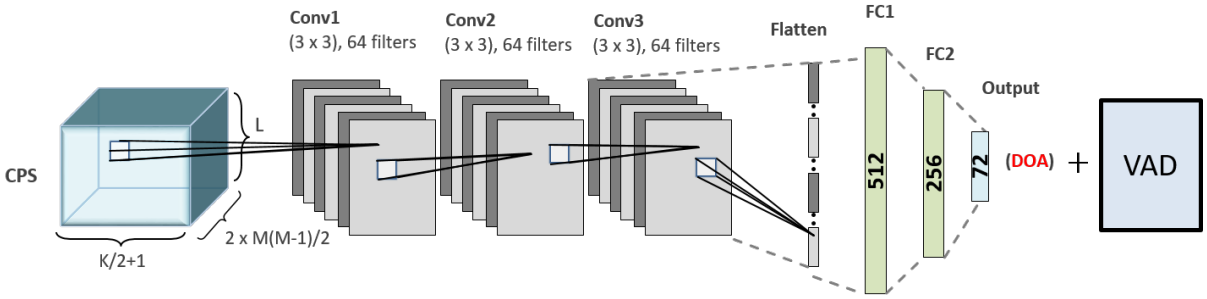


Figure 3·2: Baseline VAD-informed DOA estimation system using only CPS features.

spectro-temporal structure that can be identified in the input channels of the PD features. The main idea of using PD features in combination with CPS features is to use the salient periodicity features as a footprint of speech signals in a noisy mixture [45, 46]. This enables the CNN to detect voiced speech portions of a signal, at the same time mapping the CPS features of these portions to the DOA of the talker.

3.4 CNN-based DOA estimation systems

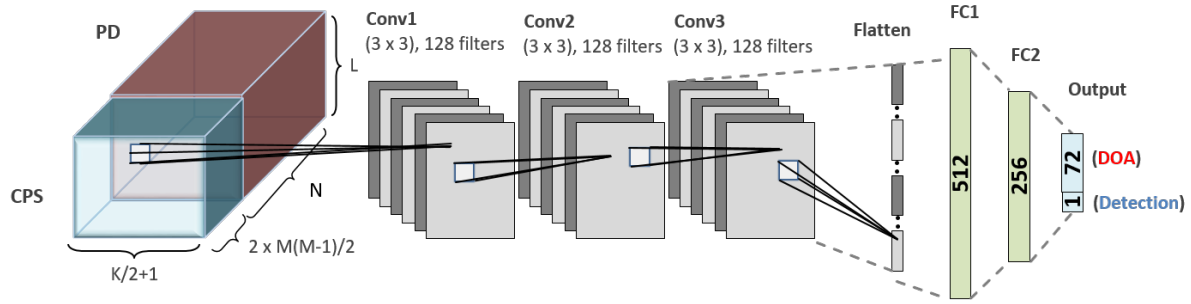
In this section, we describe the CNN-based DOA estimation systems. Section 3.4.1 discusses the baseline system, which adopts a VAD-informed DOA estimation approach and uses only the CPS features. Section 3.4.2 presents the proposed systems, which adopt a speech-aware DOA estimation approach and use a combination of the CPS features and the narrowband PD features as input features.

3.4.1 Baseline VAD-informed system

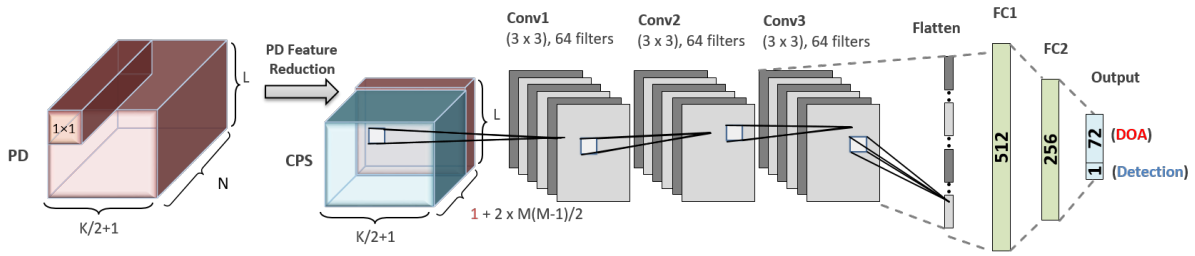
Fig. 3·2 depicts the baseline system consisting of a CNN using only spatial CPS features as input cascaded with a pitch-based binary VAD [138]. In the baseline CNN architecture, each convolutional block (*Conv1* to *Conv3*) consists of a cascade of 2D convolutional, batch normalization, ReLU activation, and 2D max-pooling layers. The outputs of the last pooling layers in *Conv3* are concatenated and then used as an input for a cascade of two fully-connected blocks (*FC1* to *FC2*), each representing a fully-connected dense layer followed by batch normalization, ReLU activation, and dropout layers. A softmax activation function predicts the posterior probability map for the C DOA classes.

3.4.2 Proposed speech-aware systems

Fig. 3·3 depicts the proposed speech-aware DOA estimation systems, which use narrowband PD features in combination with spatial CPS features as input features of the CNN. We expect that by training these systems with speech and non-speech signals, the CNN can capture the spectro-temporal structure of the signal encoded in the PD features, thereby distinguishing between speech and non-speech portions while simultaneously mapping the CPS features to a sound source DOA when speech portions in the signal are detected.



(a) without PD feature reduction



(b) with PD feature reduction stage (two-stage CNN)

Figure 3·3: Proposed speech-aware DOA estimation systems: (a) CPS and PD features are jointly processed by the CNN, (b) PD features are reduced to PD saliency features using 1×1 convolutions before being jointly processed with CPS features by the CNN.

The system in Fig. 3·3a directly employs 2D convolutional filters to the time-frequency regions of each input channel, i.e., PD and CPS features belonging to the same time-frequency bins are jointly processed, ensuring a proper association of both features. However, the spectro-temporal sparsity of the PD features (as visualized in Fig. 3·1) may complicate this task when a relatively large number of PD channels are correlated to the CPS features by the CNN. This motivates the usage of a PD feature reduction stage prior to the joint feature processing by the CNN.

Fig. 3·3b depicts the proposed two-stage CNN architecture including a PD feature reduction stage. The PD feature reduction stage aims at reducing the PD input depth, i.e., the number of channels, while keeping its width and height, i.e., the time-frequency resolution fixed. We propose to use 1×1 convolutions [164] to reduce the N-channel PD features to a 1-channel PD feature, which can be interpreted as a PD saliency feature for each time-frequency bin. In the next stage, the PD saliency features are jointly processed with the CPS features using 2D convolutional filters. It should be noted that both stages are jointly trained.

The CNN architecture of the proposed systems in Fig. 3·3 is very similar to the CNN architecture of the baseline system in Fig. 3·2. However, since the input features of the first convolutional block (*Conv1*) in the considered systems are different (CPS only, CPS and PD, CPS and PD saliency), the number of input channels is obviously different. In addition, the VAD-informed baseline system has C nodes in the output layer, whereas the speech-aware systems have $C+1$ nodes in the output layer. Finally, after hyperparameter optimization the best

performance was obtained when using 64 convolutional filters in the baseline system and the two-stage CNN (both corresponding to about 5.5 million trainable parameters), and using 128 convolutional filters in the proposed system without feature reduction (corresponding to about 11.2 million trainable parameters).

3.5 Experimental evaluation

In this section, we conduct experiments to evaluate the performance of the baseline system and the proposed speech-aware systems described in Section 3.4.1 and Section 3.4.2, respectively.

3.5.1 Datasets and data generation for training and evaluation

We used a database of multichannel BRIRs [1] to generate data for training and evaluation. The considered binaural hearing aid setup consists of $M = 4$ microphones, where the front and rear microphones in both left and right hearing aids were used. We used sound source signals from speech [152] and non-speech [153] datasets to generate the training and validation data required during the training of all CNNs. For evaluation, only speech signals from the validation TIMIT [152] data were used as source signals. Source signals were randomly chosen from unique speakers (both male and female) and from three categories [143] of non-speech signals. We generated the noisy binaural microphone signals by convolving the source signals with BRIRs and mixing the resulting clean binaural signals with a background noise at different SNRs. All systems were trained in noisy anechoic conditions and evaluated in noisy reverberant environments.

During training, we used a simulated binaural diffuse noise to generate noisy binaural microphone signals at SNRs ranging from -5 dB to $+20$ dB in 5 dB steps. This diffuse noise was generated by convolving uncorrelated speech-shaped noise taken from the ICRA noise database [154] with anechoic BRIRs and summing all resulting binaural signals from 72 directions. In total, we obtain 3.85 million training examples. To calculate the validation loss at the end of each epoch, 200000 examples were randomly selected from the validation data and kept fixed throughout training.

We generated the evaluation data for static-source scenarios in two real environments [1] (cafeteria and courtyard) with a reverberation time of approximately 1300 ms and 900 ms, respectively. The recorded cafeteria babble noise and courtyard ambient noise [1] were used to generate noisy binaural microphone signals. All systems were evaluated at SNRs ranging from -5 dB to $+10$ dB in 5 dB steps. A total of 150 speech segments randomly chosen from 30 unique male and female speakers (each with a length of 1 s) were used as source signals. In each environment, we considered BRIRs of two head orientations for four source positions [1]. It should be noted that the source and background noise signals, acoustic conditions, and source positions used during evaluation were different from those used during training and validation.

3.5.2 Implementation details

In our simulations, we used a sampling frequency $f_s = 16$ kHz and an STFT framework with a Hann window of length $K = 160$ (corresponding to 10 ms) and 50% overlap, resulting in 81 STFT

frequency bins. Each training example includes a block of $L = 20$ consecutive time frames. For the PD feature computation, we used a 4-th order GTFB implementation [48] with 61 frequency subbands, a group delay of 256, and minimum and maximum center frequency of 60 Hz and 7200 Hz. For PD features, we chose $N = 180$ fundamental period candidates corresponding to minimum and maximum fundamental frequencies of 70 Hz and 320 Hz, respectively. The comb filter gain in (3.3) was chosen to be $\alpha = 0.7$. We considered the front microphone of the left hearing aid as the reference microphone for the PD feature extraction.

All systems were implemented using PyTorch [156]. For all CNNs, we used a 2D convolutional filter size of 3×3 with a stride size of 1×1. The max-pooling size was 2×1, i.e., no pooling is applied across frequencies. In addition to the batch normalization used in the convolutional and fully-connected blocks of the CNNs, the layer normalization [149] was applied on the CPS and PD features separately at the input. The CNNs were trained using the Adam optimizer [157], a cross-entropy loss function, an initial learning rate of 10^{-5}, a mini-batch size of 128 and a dropout rate of 0.5. An early stopping regularization method on the validation loss and a variable learning rate scheduler with a factor of 0.5 were also employed. A softmax activation function is used at the output layer of all systems.

3.5.3 Performance measures

To evaluate the DOA estimation performance, we used MAE and accuracy (Acc). A DOA estimate in frame l is considered accurate if the absolute error between the estimated DOA $\hat{\theta}_l$ and the oracle DOA θ_l is smaller than 5°. The MAE (in degrees) and accuracy are defined as

$$\text{MAE} = \frac{1}{\mathcal{L}} \sum_{l=1}^{\mathcal{L}} \left| \hat{\theta}_l - \theta_l \right|, \tag{3.5}$$

$$\text{Acc} = \frac{\mathcal{L}_{\text{acc}}}{\mathcal{L}} \times 100, \tag{3.6}$$

where $\mathcal{L}$ and $\mathcal{L}_{\text{acc}}$ denote the total number of estimates and the total number of accurate estimates, respectively.

3.5.4 Results and discussion

Fig. 3·4 shows the performance of all considered systems in terms of accuracy and MAE. By comparing the proposed two-stage CNN with PD feature reduction to the proposed system without PD feature reduction, it can be observed that the two-stage CNN generally results in a better or similar performance. The benefit of using PD feature reduction is especially clear in challenging acoustic conditions, i.e., in the highly-reverberant cafeteria environment and in adverse SNR conditions in both environments. Although in terms of accuracy this benefit decreases with increasing SNR in the courtyard environment in favor of the proposed system without feature reduction, the proposed two-stage CNN maintains a lower MAE in all conditions.

The results in Fig. 3·4 clearly show that both proposed systems consistently outperform the baseline system in both environments and for all SNR conditions. This benefit decreases towards high SNR conditions, which is expected as there are fewer signal portions dominated by noise,

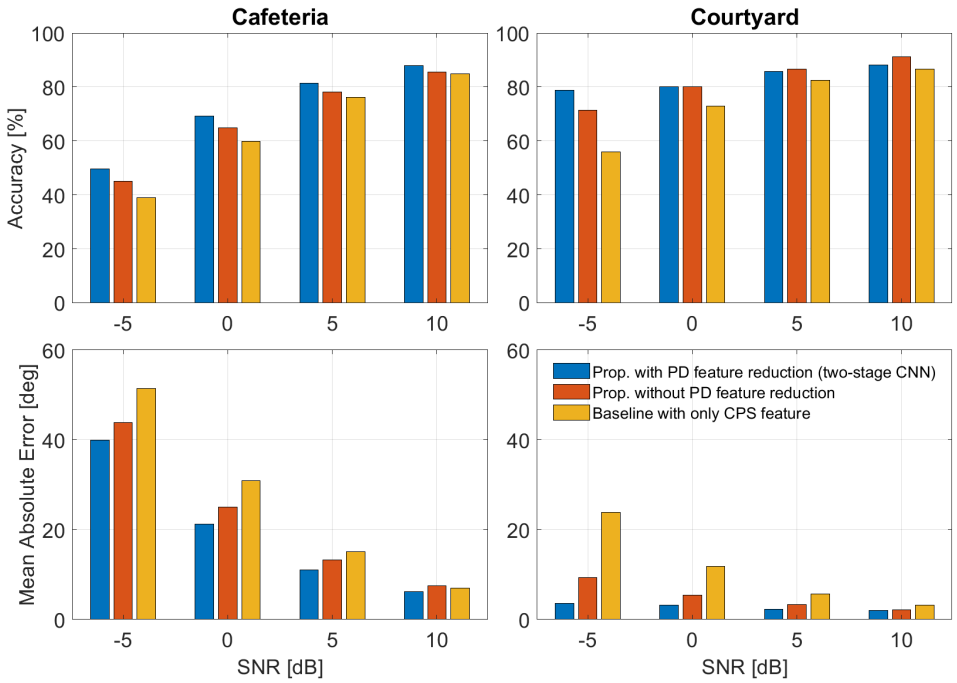


Figure 3·4: Accuracy and MAE of the proposed systems with narrowband feature combination evaluated against the baseline system using only CPS features in static-source scenarios for different SNR conditions in the cafeteria and courtyard environments.

which PD features can detect.

Considering the number of trainable parameters (cf. Section 3.4.2), compared to the baseline system the proposed two-stage CNN requires a comparable number of parameters while achieving a better performance. Moreover, the proposed two-stage CNN outperforms the proposed system without feature reduction while requiring significantly fewer parameters. This further highlights the benefit of employing the proposed feature reduction stage before the joint processing of the proposed narrowband feature combination.

3.6 Conclusion

In this paper, we proposed two speech-aware DOA estimation systems that use a combination of narrowband periodicity features and spatial CPS features as inputs of a CNN. In particular, we introduced a two-stage CNN with a periodicity feature reduction stage employing 1×1 convolutions. Evaluation results showed that the proposed systems yield a better DOA estimation performance than a baseline system using CPS features and a pitch-based VAD. While offering a lower computational complexity, the proposed two-stage CNN with feature reduction outperforms a system that jointly processes the feature combination without feature reduction. This study suggests that a feature reduction stage can effectively map the sparse periodicity features into more compact salient periodicity features, which combined with spatial features, provide robust features to guide speech-aware DOA estimation.

Chapter 4

Improving Multi-talker Binaural DOA Estimation by Combining Periodicity and Spatial Features in Convolutional Neural Networks

---

This chapter is a reformatted reprint of the following publication:

R. Varzandeh, S. Doclo, V. Hohmann, "Improving multi-talker binaural DOA estimation by combining periodicity and spatial features in convolutional neural networks," Submitted to *EURASIP Journal on Audio, Speech, and Music Processing*, 2024.

The chapter proposes a method to enhance the accuracy of multi-talker DOA estimation in binaural hearing aids by combining periodicity and spatial features. Using a two-stage CNN, the system integrates CPS phase as spatial features and PD as spectral features. This combination improves the robustness and accuracy of DOA estimation in noisy, multi-talker environments over baseline systems that only use the CPS phase or a combination of the CPS phase and magnitude spectrogram as spectral features.

Abstract

Deep neural network-based direction of arrival (DOA) estimation systems often rely on spatial features as input to learn a mapping for estimating the DOA of multiple talkers. Aiming to improve the accuracy of multi-talker DOA estimation for binaural hearing aids, we investigate the usage of periodicity features as a footprint of speech signals in combination with spatial features as input to a convolutional neural network (CNN). In particular, we propose a multi-talker DOA estimation system employing a two-stage CNN architecture that utilizes cross-power spectrum (CPS) phase as spatial features and an auditory-inspired periodicity feature called periodicity degree (PD) as spectral features. The two-stage CNN incorporates a PD feature reduction stage prior to the joint processing of PD and CPS phase features. We investigate different design choices for the CNN architecture, including varying temporal reduction strategies and spectro-temporal filtering approaches. The performance of the proposed system is evaluated in static source scenarios with 2-3 talkers in two reverberant environments under varying signal-to-noise ratios using recorded background noises. We consider baseline systems that utilize either CPS features or a combination of CPS and magnitude spectrogram as a spectral feature. Experimental results demonstrate the clear benefits of combining PD and CPS phase features within our proposed system. It consistently improves DOA estimation performance across all conditions and environments, outperforming both baseline systems in terms of accuracy.

4.1 Introduction

Multi-talker DOA estimation is integral to acoustic signal processing and plays a pivotal role in many applications, from enhancing auditory experiences in assisted listening devices to improving voice command detection in smart devices [165, 166]. In hearing aids, accurate DOA information facilitates improved speech intelligibility through beamforming, enables the suppression of competing noise sources, and can increase environmental awareness. This ultimately helps wearers navigate conversations in complex social environments. While the human auditory system is uniquely able to localize speech sources in noisy and reverberant environments, this remains a challenging task for machine listening systems such as hearing aids [4, 68]. This study addresses multi-talker DOA estimation in the context of binaural hearing aids.

Binaural DOA estimation, the process of determining the direction of sound sources using signals received by two microphones (e.g., in a binaural hearing aid setup), primarily leverages binaural cues inspired by the human auditory system, namely ITD, IPD, and ILD [4, 79]. While ITD and IPD pertain to the differences in time and phase of a sound arriving at two microphones, respectively, ILD concerns the difference in sound intensity levels captured by the microphone pair. ITD information can be defined either through an auditory-inspired approach in auditory gammatone filterbank channels [10, 79] or using the broadband GCC function [79, 80]. Integrating ITD with ILD cues has been shown to enhance the accuracy of binaural DOA estimation compared to using ITD alone [9, 10]. Studies show that combining IPD and ILD information improves DOA estimation accuracy in multi-talker scenarios [82]. Most of these approaches usually match the estimated binaural features with pre-computed feature templates from HRTF databases to obtain the DOA. Another class of conventional binaural DOA estima-

tion approaches utilizes RTF vectors [8, 11, 83]. These methods employ a database of prototype RTF vectors, pre-computed for each direction using a measured HRTF database.

A major challenge for these approaches in real-world acoustic conditions is that background noise, interference, and reverberation introduce uncertainties into binaural cues. These uncertainties distort the extracted binaural features from microphone signals, leading to mismatches with the pre-computed templates and subsequently degrading DOA accuracy [10]. To address the limitations of these approaches, researchers have developed supervised learning techniques using DNNs [12, 14, 16–20, 23, 96, 98–106, 121]. When trained on diverse acoustic conditions, these techniques demonstrate more robust performance in adverse scenarios [14, 99]. In this paper, we address the multi-talker binaural DOA estimation using supervised DNN-based techniques.

Supervised DNN-based approaches commonly formulate the multi-talker DOA estimation task as a classification or regression problem [12, 121]. In classification-based approaches, the neural network predicts a spatial probability map for a discretized DOA range. Peaks within this map indicate the probable locations of active sound sources. A common assumption here is that the number of sources is known in advance, and their DOAs are found by peak detection [14, 17, 98, 102, 103]. In regression-based approaches, the neural network provides continuous estimates directly in the output [100, 121]. This offers potential for improved performance compared to the classification-based approaches, as the latter limits the DOA resolution depending on the discretized DOA range. However, in regression-based approaches, the number of simultaneously active sources must be known before training, which may not align with the number of active sources during testing. Another drawback of regression-based approaches is the source permutation problem in multi-talker scenarios, where it becomes ambiguous which predicted output corresponds to which target speaker. In this paper, we focus on classification-based techniques for binaural DOA estimation.

DNN-based methods for binaural DOA estimation typically utilize spatial features extracted from binaural signals. Commonly used features include the ILD, ITD (or IPD), CCF, and GCC-PHAT [12, 108]. Most of these methods directly utilize the input features for DOA estimation utilizing the network output [14, 16, 18], while some methods adopt a two-step approach, first refining the input features into enhanced representations using the DNN and then estimate the DOA from the enhanced features [15, 19, 20]. Previous research indicates that using the complete CCF or GCC-PHAT can be more effective than relying solely on ITD [14]. While most existing methods focus on DOA estimation in the azimuthal plane [14, 16, 20, 99], some employ multi-task learning to simultaneously estimate both azimuth and elevation [18, 101], or azimuth alongside distance [167]. In this work, we focus specifically on DOA estimation within the azimuthal plane.

It is assumed that the human auditory system groups signal components according to information such as periodicity of voiced speech and continuity of harmonics, and then ITD (or IPD) information is used to segregate the grouped components [4]. Motivated by that, a learning-based method for multi-talker DOA estimation [67] proposed to incorporate a monaural pathway including pitch-based analysis to group time-frequency units dominated by the same talker. The grouping provided constraints for the integration of binaural cues, improving azimuth estimation

accuracy. It has also been shown in [47] that periodicity-based salient features yield a sparse auditory time-frequency representation capable of decoding complex auditory scenes.

While binaural features are widely used for DOA estimation, the benefit of their combination with monaural spectral features, such as salient periodicity features as input features for DNN-based multi-talker DOA estimation has not been investigated. In [168], we proposed a classification-based system based on CNNs for single-talker binaural DOA estimation. This study showed the benefit of using a periodicity feature called PD in combination with spatial features as input to a two-stage CNN.

In this paper, we propose a DOA estimation system that builds upon our earlier work [168] by incorporating a unique feature combination within a computationally efficient two-stage CNN adapted for multi-talker DOA estimation. Our main objective is to explore the potential benefits of incorporating periodicity features, alongside spatial features for DNN-based multi-talker DOA estimation, as established for the single-talker scenarios in [168, 169]. As the spatial feature, we use the phase component of CPS, which is closely related to the IPD for a pair of microphones. We hypothesize that combining the CPS phase as the spatial feature with a compact representation of PD, obtained through a feature reduction stage inspired by [168], will outperform using the CPS phase alone in multi-talker scenarios.

To optimize performance and computational efficiency, we investigate different two-stage CNN architectural choices, including different temporal reduction strategies (e.g., dilation, max pooling) and different approaches to spectro-temporal filtering using convolutional kernels. We show that the proposed system which captures the temporal dependencies for each frequency independently via convolutional kernels leads to the best performance for DOA estimation with the lowest computational complexity. We conducted evaluations in different static-source scenarios, including different SNRs using recorded background noises, reverberant environments, and different number of talkers. The proposed system was compared to baseline systems that utilized either the CPS phase alone or a combination of the CPS phase and magnitude spectrogram as input features. Experimental results demonstrated that the proposed system outperforms baseline systems in all evaluated scenarios.

The remainder of this paper is structured as follows. First, in Section 4.2, the multi-talker DOA estimation is formulated and discussed as a classification problem. Section 4.3 introduces the input features employed in this study. In Section 4.4, comprehensive descriptions of the proposed and baseline systems are presented. The details of the experimental setup for training and evaluation of all systems including datasets, data generation, training and network hyperparameters, and evaluation metrics appear in Section 4.5. The proposed and baseline systems are evaluated, and the results are discussed in Section 4.6. Section 4.7 summarizes the results and presents the conclusion.

4.2 DOA estimation as a classification problem

In this work, we consider the problem of multi-talker DOA estimation in the azimuthal plane using a binaural hearing aid setup with M microphones, where the microphones are located close to the ears on both sides. The acoustic scenario consists of multiple speech sources and

background noise, which are assumed to be mutually uncorrelated. The m-th microphone signal in the time domain at time t is given by

$$y_m(t) = \sum_{i=1}^{\mathcal{I}} x_m^i(t) + v_m(t), \tag{4.1}$$

where x_m^i and v_m denote the desired i-th speech source at DOA θ_i in the azimuthal plane, and noise signal components in the m-th microphone signal, respectively. In the STFT domain, the m-th microphone signal at time frame n and frequency bin k (with K and D the STFT length and hop size, respectively) can be written as

$$Y_m(n,k) = \sum_{i=1}^{\mathcal{I}} X_m^i(n,k) + V_m(n,k). \tag{4.2}$$

Conventionally, by discretizing the azimuth range into C DOAs $\{\phi_1, \cdots, \phi_C\}$, multi-talker DOA estimation is formulated as a C-class classification task, where output classes correspond to independent DOAs, i.e., sound source locations are mutually independent [98, 102, 103]. The goal is to assign the DOAs of multiple incoming sound sources to corresponding DOA classes. In this study, we use $C = 72$ classes spanning the full 360° azimuth range, yielding a DOA map with 5° resolution.

By taking a supervised approach, during training, each training example may belong to one or more output classes that are labeled using ground truth DOA information. In other words, each training example can represent situations where multiple speakers are active simultaneously. However, this approach can complicate the training data generation, as the differences in signal levels for these scenarios can significantly impact the performance of the DOA estimation system. In this work, we generate training examples involving only a single active speaker and evaluate the system's ability to generalize to multi-talker scenarios where each speaker contributes equally to the microphone signal.

During testing, the neural network predicts a posterior probability for each DOA class in the output. The generated posterior probability map $\boldsymbol{P} = [P_1, \cdots, P_C]$ represents the likelihood of the sound source being located at each of the C possible DOAs. As a common approach, with $\mathcal{I}$ active sources, the $\mathcal{I}$ DOA classes with the highest probability values in $\boldsymbol{P}$ are selected as the estimated DOAs. In this study, we will take a slightly different approach for DOA estimation in Section 4.4.

4.3 Input features

This section outlines the spatial and spectral features used as inputs for the classification-based DOA estimation methods in this study. Section 4.3.1 introduces narrowband CPS features as spatial features. Section 4.3.2 presents narrowband PD features (as introduced in [168]), alongside the magnitude spectrogram as an alternative spectral feature.

4.3.1 Spatial features

In [168], the real and imaginary components of the CPS was used as the spatial feature, which encodes both spectral magnitudes and phase differences. In this work, we consider the phase component of the CPS as spatial feature used in the baseline and proposed systems in Section 4.4. The instantaneous CPS between the r-th and q-th microphone is defined as

$$G_d(n,k) = Y_r(n,k)Y_q^*(n,k), \tag{4.3}$$

where $(\cdot)^*$ denotes complex conjugate and d denotes a microphone pair combination. As CPS input, we consider the phase components of $G_d(n,k)$ for all $M(M-1)/2$ unique microphone pairs for frequencies up to the Nyquist frequency, i.e., $k = 0, 1, \cdots, K/2$, for L consecutive time frames. This means that the shape of the CPS input is equal to $L \times (K/2+1) \times M(M-1)/2$. We note here that the first, second, and third dimensions represent the height, width, and depth of the input feature, respectively, with the depth corresponding to the number of input channels.

4.3.2 Spectral features

Periodicity is an important cue to segregate and localize different talkers [47, 148]. Periodicity features typically require an auditory pre-processing stage followed by feature extraction [47]. In [143, 168, 169] a periodicity feature called PD was used, which captures the salience of the periodic components in the input signal. In this work, we propose to use narrowband PD features [168] for multi-talker DOA estimation, computed for a set of N fundamental period candidates.

To compute PD features, we select one of the M microphones as the reference. Please note that this microphone is selected arbitrarily, and optimal microphone selection for PD estimation is beyond the scope of this study. In the pre-processing step, the reference microphone signal in the hearing aid setup is first decomposed into signals in different gammatone frequency bands using a complex-valued GTFB [48]. The real part of each signal then undergoes half-wave rectification, yielding signal $y(t,f)$ in the f-th gammatone frequency band. In each frequency band, $y(t,f)$ is processed with a fifth-order low-pass filter (770 Hz cutoff) and a second-order high-pass filter (40 Hz cutoff), resulting in bandpass-filtered signal envelopes $y_{env}(t,f)$. These envelopes serve as the basis for our PD feature extraction.

In the feature extraction step, we filter the signal envelopes using a set of N parallel IIR comb filters. These filters are designed for a set of N fundamental period candidates $p_j, j = 1, \cdots, N$. The comb-filtered signals are computed by

$$s(j,t,f) = (1-\alpha)y_{env}(t,f) + \alpha s(j,t-p_j,f), \tag{4.4}$$

where α denotes the filter gain. The periodicity degree is defined as the mean amplitude of the comb-filtered signal, given by

$$PD(j,t,f) = (1-\beta_j)|s(j,t,f)| + \beta_j PD(j,t-1,f), \tag{4.5}$$

where the averaging parameter β_j for each fundamental period candidate is defined as $\beta_j =$

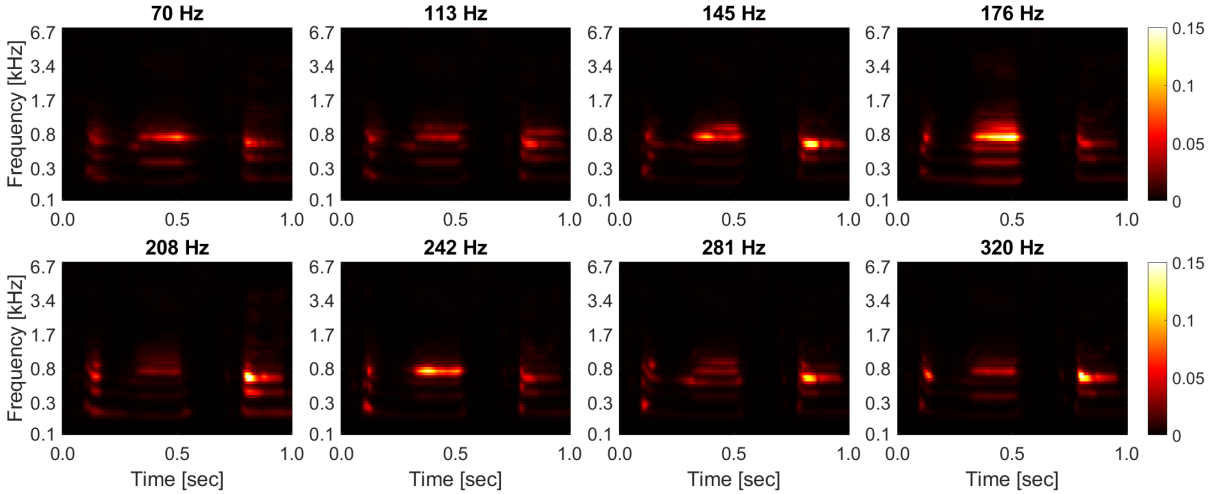


Figure 4·1: PD features computed over $L = 199$ consecutive time frames and 61 gammatone bands for a clean female speech in an anechoic environment. A small set of fundamental frequency candidates (specified above each image) is shown for visualization. The sparse spectro-temporal structure of PD features contains sufficient information to decode complex auditory scenes, and motivates using a feature reduction stage to learn the salient PD features prior to joint processing with the CPS phase.

e^{-1/p_j}.

To enable joint spectro-temporal processing of PD and CPS features, their time-frequency resolutions must be aligned. Since PD features in (4.5) initially have the temporal resolution of the time-domain signal, we achieve the necessary alignment with CPS features by temporally averaging them within each STFT frame as

$$\overline{PD}(j,n,f) = \frac{1}{K} \sum_{t=(n-1)D+1}^{(n-1)D+K} PD(j,t,f). \tag{4.6}$$

The non-uniform frequency resolution of gammatone bands (decreasing with frequency) contrasts with the linear spacing of STFT frequency bands. To align the frequency resolution of PD features with CPS features, we employ different strategies based on STFT frequency. For low frequencies, PD features from multiple gammatone bands corresponding to a single STFT band are averaged, while for high frequencies, PD features from each gammatone band are replicated and assigned to the associated STFT frequency bands.

AS input PD feature used for the proposed system (cf. Fig. 4·3), we consider PD features in (4.6) for all N fundamental period candidates, for L consecutive time frames, and for all $K/2+1$ STFT frequency bands. This leads to an input PD feature of size $L \times (K/2+1) \times N$, which will be used as the input PD feature of the proposed DOA estimation system in Section 4.4.2.

For a 1s clean signal of a female talker, Fig. 4·1 depicts exemplary 2D images of PD features, corresponding to a subset of fundamental frequency candidates. For a perfectly periodic signal characterized by a specific fundamental frequency, a high PD value will be captured for candidates associated with the harmonics and sub-harmonics of this fundamental frequency.

While speech isn't perfectly periodic, fundamental frequency variations and harmonics create a spectro-temporal structure visible in the PD features. The primary rationale for using PD features alongside spatial features is to leverage the periodicity features as a robust footprint of speech signals in a noisy mixture [45, 46]. This approach allows the neural network to pinpoint voiced speech segments while simultaneously mapping their associated CPS features to the talker's DOA.

Magnitude spectrograms provide rich spectro-temporal information about formant frequencies and harmonic content, making them common in DOA estimation systems [12]. To explore the benefit of spectral features in combination with spatial features for multi-talker DOA estimation, as an alternative to periodicity features, we use the magnitude spectrogram of the same microphone used for PD feature extraction. We take the magnitude of the microphone's STFT coefficients up to the Nyquist frequency, i.e., $k = 0, 1, \cdots, K/2$, for L consecutive time frames. This results is an input magnitude spectrogram of shape $L \times (K/2 + 1)$.

4.4 CNN-based DOA estimation systems

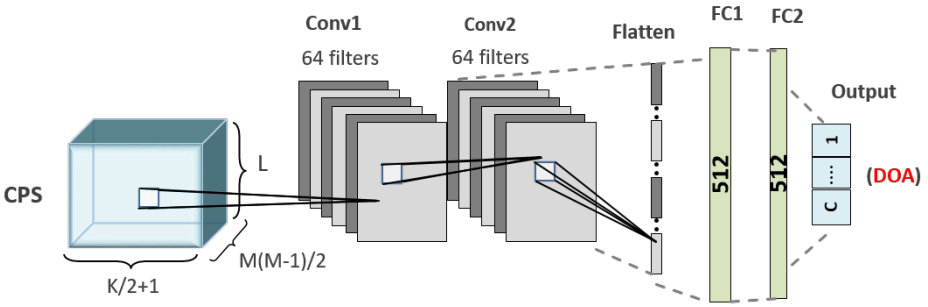
This section outlines the CNN-based DOA estimation systems. The baseline systems are discussed in Section 4.4.1, which employ either only the CPS phase or a combination of the CPS phase and magnitude spectrogram as input. Section 4.4.2 presents our proposed system which utilizes the CPS phase combined with PD features as input. We also explore alternative design choices for this two-stage architecture. Finally, Section 4.4.3 analyzes the computational complexity of all considered systems.

All systems share the same training and DOA estimation procedures. The key difference between our proposed system and the baselines lies in the two-stage CNN architecture and the combination of CPS phase and PD features. For training, each training example consists of a block of L consecutive time frames, i.e., we employ block-level labeling and each CNN generates its output for the whole block. A key assumption is that the DOA remains constant within this block of L frames when assigning a ground truth label.

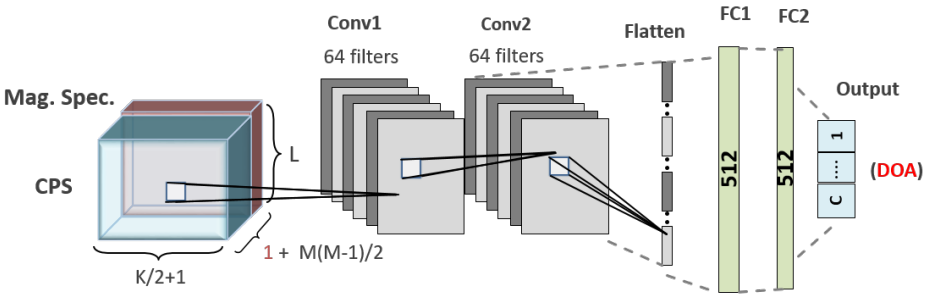
For DOA estimation, with $\mathcal{I}$ active sources, we first find the $\mathcal{I}$ DOA classes $\phi_j, j = 1, \cdots, \mathcal{I}$ with the highest probability values in the posterior probability map $\boldsymbol{P}$. To refine these discrete DOA classes into continuous estimates, we estimate the sound source DOA by employing parabolic interpolation [151] on three DOA classes centered around ϕ_j, i.e., ϕ_{j-1}, ϕ_j and ϕ_{j+1}. As a result, this approach allows for a more accurate DOA estimation with a higher spatial resolution. Please note that during testing, consecutive input features for all systems overlap by $L - 1$ frames, yielding a new posterior probability map for every new frame.

4.4.1 Baseline systems

Neural networks based on CNNs have proven highly effective for DOA estimation and sound source localization [12]. Fig. 4·2 depicts the baseline systems consisting of a CNN using only the CPS phase or a combination of the CPS phase and magnitude spectrogram as input. The CNN architecture in both baseline systems begins with a cascade of two convolutional blocks (*Conv1* and *Conv2*). Each block consists of a 2D convolutional layer, followed by batch normalization



(a) With only CPS phase



(b) With CPS phase and magnitude spectrogram

Figure 4·2: Baseline DOA estimation systems using (a) only spatial feature (CPS phase), and (b) spatial and spectral features (CPS phase and magnitude spectrogram).

and a ReLU activation layer. Only *Conv2* incorporates a max pooling layer after the ReLU. Next, the concatenated outputs of *Conv2* serve as an intermediate feature vector and are fed into two fully-connected blocks (*FC1* and *FC2*). These blocks each comprise a fully connected dense layer with batch normalization, ReLU activation, and dropout layers. Finally, the output layer employs C sigmoid activation functions to generate the posterior probability map for the C independent DOA classes.

To improve CNN performance, we employ layer normalization [149] (without an affine transformation) directly on the input features before the first convolutional block. This normalization targets the CPS phase and magnitude spectrogram separately. It's important to note that this has been implemented in addition to the batch normalization used within the convolutional and fully-connected blocks of the CNNs.

4.4.2 Proposed system

Fig. 4·3 illustrates our proposed multi-talker DOA estimation system, which combines PD features (cf. Section 4.3.2) with the CPS phase (cf. Section 4.3.1) as input to a two-stage CNN. Inspired by [168], our system features a PD feature reduction stage before joint processing with CPS phase. Within this reduction stage, we use 1×1 convolutions [164] to transform the N-channel PD features into a single-channel PD saliency feature for each time-frequency bin. This aims to decrease input depth (number of channels) while preserving the time-frequency resolu-

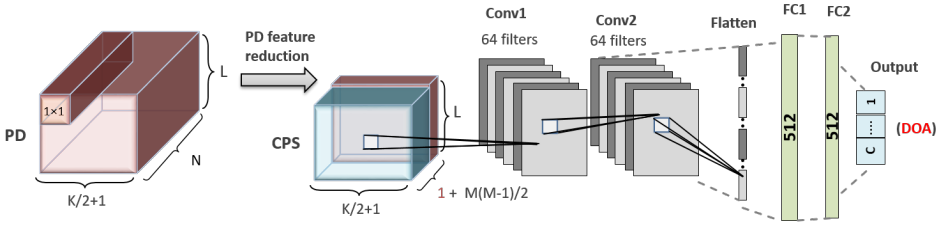


Figure 4·3: Proposed system using the CPS phase and PD features as input. The PD features undergo dimensionality reduction via 1×1 convolutions to create compact PD saliency features. These are then combined with CPS phase features as input to the convolutional blocks for joint processing and extraction of spectro-temporal patterns related to source DOA.

tion of PD features. In the following stage, these PD saliency features are jointly processed with CPS features using convolutional filters.

The second stage of our proposed system shares the same architecture as our baseline systems (Section 4.4.1). To process the combined CPS phase and PD saliency features, we use convolutional blocks (*Conv1* and *Conv2*), each composed of a 2D convolutional layer, batch normalization, and a ReLU activation layer. *Conv2* also includes max pooling after the ReLU. The outputs of the *Conv2* block are then concatenated and fed as an intermediate feature vector into two fully-connected blocks (*FC1* and *FC2*). Each block features a fully-connected dense layer, batch normalization, a ReLU activation layer, and a dropout layer. Finally, similar to the baseline systems, the output layer uses C sigmoid activation functions to generate the posterior probability map for the C independent DOA classes. We employ the same input layer normalization scheme as our baseline systems, normalizing the PD and CPS phase separately. It should be noted that both stages of the proposed system are trained jointly.

Employing convolutional kernels across L consecutive time frames and $K/2+1$ frequency bins allows for different design approaches to capturing spectro-temporal dependencies. In particular, we have made two main design choices for the two-stage CNN architecture, employed for the proposed system.

As our first design choice, we consider a combination of kernel, dilation, and max pooling sizes that reduces the temporal dimension of the input features to a single value at the output of the last convolutional block. This essentially captures temporal correlation solely through the convolutional path. Consequently, the intermediate feature vector at the input of the fully-connected path primarily contains elements representing different frequencies. Our CNN design is based on the assumption that convolutional blocks effectively capture temporal dependencies in the input features, while fully-connected blocks best capture global patterns across frequencies.

In multi-talker scenarios, neighboring frequency bins may contain dominant activity from different speakers. Hence, previous works [98, 102] have used convolutional kernels that separately process each frequency bin to take benefit from the widely adopted assumption of W-disjoint orthogonality [163]. As the second design choice, we preserve frequency resolution at the output of the convolutional path (and hence, capturing global patterns across frequencies merely via the fully-connected path). We expect that this approach may lead to a better DOA estimation

performance compared to joint learning of time-frequency dependencies using 2D convolutional kernels.

With an interest in designing a computationally efficient system in this study, the proposed system utilizes a temporal kernel size of 7, a dilation rate of 2, and a temporal max Pooling size of 2. This CNN architecture was chosen to help reduce computational costs while maintaining performance. The large kernel size of 7 captures long-range temporal dependencies in the input features. Using dilation in the convolutions expands the receptive field without increasing the number of parameters. The max pooling then further downsamples the temporal dimension to reduce computations in subsequent layers. This specific combination implements our first design choice while capturing large temporal contexts within just two convolutional blocks (*Conv1* and *Conv2* in Fig. 4·3). To implement our second design choice, we ensure convolutional kernels and max pooling operate exclusively across the time dimension. In the following, we investigate our main design choices by considering alternative approaches.

First, aiming at investigating different approaches to capture temporal dependencies, we employ convolutional blocks with different combinations of dilation and max pooling. This results in reducing the temporal dimension into different numbers of features for each filter at the output of the convolutional path (compared to the single feature in the proposed system). Consequently, both convolutional and fully-connected paths contribute to capturing temporal dependencies. We compare our proposed system with two additional two-stage CNN configurations: one with a dilation size of 2 and no max pooling (temporal dimension of 2), and the other with neither dilation nor max pooling (temporal dimension of 8). Similar to the proposed system, theses two two-stage configurations use convolutional kernels exclusively across time (see the second and third systems in Table 4.1).

Second, we explore the usage of kernels that span both time and frequency dimensions using different kernel sizes across frequencies. To do so, the proposed system with the kernel size of 7x1 (only temporal processing) is paired with two alternative two-stage systems that employ 2D kernels across both time and frequency with the sizes of 7x2 and 7x3 (see the fourth and fifth systems in Table 4.1). It should be noted that the temporal dimension in all three systems is reduced to a single value. To prevent information loss and focus on adjacent frequencies, we avoid dilated kernels across the frequency dimension.

To prevent temporal information loss, in all systems, the first convolutional block (*Conv1*) uses neither dilation nor max pooling. The subsequent convolutional block (*Conv2*) in the alternative systems may incorporate max pooling and/or a dilation rate of 2 across time. All systems include convolutional blocks with 64 filters and fully-connected blocks with 512 neurons.

4.4.3 Computational complexity

Table 4.1 shows the number of trainable parameters and multiply-accumulate operations (MACs), both in millions for the proposed system, alternative two-stage systems using different temporal reduction and spectro-temporal strategies, as well as the two baseline systems. The number of parameters, i.e. the model size, influences the memory required to store the model, while MACs provide an estimate of the arithmetic computations, which inherently affects energy consump-

Table 4.1: Number of trainable parameters and multiply-accumulate operations (MACs) of different systems.

CNN Architecture	Kernel/ Dilation	Max Pool	Temp. Dim.	MACs (M)	Param. (M)
Prop. two-stage	7x1/2x1	2x1	1	11.5	3
Two-stage w/ only Dil.	7x1/2x1	No	2	14.2	5.6
Two-stage w/o MaxP. & Dil.	7x1/No	No	8	44	21.6
Two-stage w/ Spectro-temp. proc. 1	7x2/2x1	2x1	1	19.3	3
Two-stage w/ Spectro-temp. proc. 2	7x3/2x1	2x1	1	26.8	2.9
Baseline w/ only CPS phase	7x1/2x1	2x1	1	10.7	3
Baseline w/ CPS phase & Mag. Spec.	7x1/2x1	2x1	1	11.2	3

tion.

To investigate the effect of different dilation and max pooling strategies (across time) on the complexity of the two-stage systems (cf. Fig. 4·3), we consider the first three systems in Table 4.1. Using different dilation rates and max pooling results in varying degrees of temporal reduction within the convolutional path (*Conv1* and *Conv2*), which consequently leads to intermediate feature vectors with different sizes. This manipulation yields configurations with differing computational complexities, where the complexity scales in proportion to the degree of reduction in the intermediate feature vector. For example, the proposed system which utilizes the maximum temporal dimension reduction (temporal dimension size of 1) has the minimum computational complexity of 11.5 million MACs, compared to 14.2 and 44 million MACs for the two other two-stage CNN configurations. Similarly, it has the smallest memory footprint (3 million parameters) in contrast to 5.6 and 21.6 million parameters for the two other systems.

To compare spectro-temporal processing strategies, we pair the proposed system (using 1D temporal kernels of 7x1) with alternative two-stage CNNs employing 2D kernels spanning both time and frequency (sizes 7x2 and 7x3). As Table 4.1 reveals, while the number of parameters remains relatively comparable across different frequency kernel sizes, larger kernels lead to increased MACs. All three systems maintain the same temporal kernel size (7), resulting in a temporal dimension reduction to a single value. Their difference lies in the frequency kernel size (1, 2, and 3). Due to frequency dimension reduction within the convolutional path, the modified systems have fewer intermediate features. This translates to slightly fewer trainable parameters in the system using 7x3 kernels.

Apart from the PD feature reduction stage, the architecture of the proposed system (Fig. 4·3) closely mirrors the baseline systems (Fig. 4·2). Since the proposed and baseline systems use the same kernel, dilation and max pooling sizes, as well as the same number of filters (64), they have the same number of intermediate features. Additionally, all systems use 512 neurons in the fully-connected blocks. This results in a comparable number of trainable parameters, predominantly determined by the fully-connected path. Please note that the PD reduction

stage has a negligible impact on the total number of trainable parameters (3 million). However, the PD feature reduction stage and distinct inputs to *Conv1* slightly increase the MACs of the proposed system (11.5 million) compared to the baseline systems employing the CPS phase alone (10.7 million) or in combination with magnitude spectrogram (11.2 million)

In summary, among the two-stage CNN configurations explored, the proposed system's focus on temporal feature reduction (to a single value) together with independent frequency processing yields the lowest computational cost and smallest model size. Furthermore, with a slight increase in computational need, the model size of the proposed system remains comparable to the baselines. To the best of our knowledge, direct implementation in current hearing devices is not feasible. Achieving this may require further model optimization, quantization, and pruning, which are aspects beyond the scope of this study.

4.5 Experimental setup

This section presents experiments evaluating the performance of the multi-talker DOA estimation systems described in Sections 4.4.2 and 4.4.1. We detail the employed datasets in Section 4.5.1. Sections 4.5.2 and 4.5.3 describe procedures for generating training and evaluation data. In Section 4.5.4, implementation details of input features are provided. Section 4.5.5 presents CNN training procedures and hyperparameters. Finally, Section 4.5.6 outlines the evaluation metrics used to assess the performance.

4.5.1 Datasets

We used Speech signals of 462 and 168 speakers from the TIMIT dataset [152] (including both male and female speakers) for training and validation, respectively. For evaluation, speech signals from the validation TIMIT dataset were used as source signals. A database of multichannel BRIRs [1] was used to generate data for training and evaluation. We considered a binaural hearing aid setup consisting of $M = 4$ microphones, where the front and rear microphones (approximate microphone distance of 15 mm) in both left and right hearing aids were used. The database in [1] provides BRIRs for anechoic conditions for different source-to-head distances and $C = 72$ directions in the azimuthal plane (5° resolution). It additionally includes BRIRs for two reverberant environments: a cafeteria ($T_{60} \approx 1.3$ s) and a courtyard ($T_{60} \approx 0.9$ s). Noisy binaural microphone signals were generated by convolving source signals with BRIRs and mixing the resulting clean binaural microphone signals with background noise. Training was conducted in noisy anechoic conditions, while evaluation was conducted in noisy reverberant environments.

4.5.2 Training data

For training, clean binaural microphone signals were generated by convolving speech signals with anechoic BRIRs for each of the 72 directions at a fixed 3 m source-to-head distance. The noisy binaural microphone signals were generated by mixing the clean binaural microphone signals with simulated binaural diffuse noise at SNRs ranging from −5 dB to +20 dB in 5 dB steps. This noise was generated by convolving uncorrelated speech-shaped noise (ICRA database [154])

Table 4.2: Summary of the training data

Source signals	Speech (TIMIT)
Environment	Anechoic [1]
Background noise	Simulated diffuse noise
SNR	-5 dB to $+20$ dB in 5 dB steps
Source-to-head distance	3 m
Source positions	72 positions in the horizontal plane

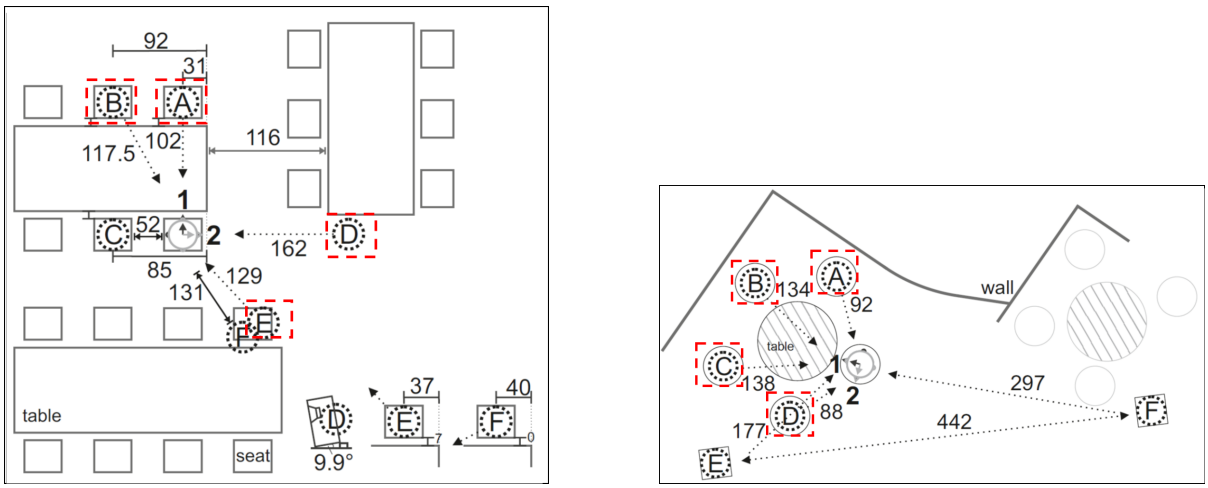
Table 4.3: Summary of the evaluation data

Source signals	Speech (TIMIT)
Environment	Cafeteria ($T_{60} \approx 1.3$ s) and courtyard ($T_{60} \approx 0.9$ s) [1]
Background noise	Recorded noise
SNR	-5 dB to $+10$ dB in 5 dB steps
Source-to-head distance	$1 - 1.6$ m
Source positions	4 source positions with 2 head orientations in each environment

with anechoic BRIRs, and summing all resulting binaural signals from 72 directions. Training examples included all 72 directions at six different SNRs. In a data pre-processing step, a simple oracle broadband energy-based VAD was used to select segments with sufficient speech content, ensuring meaningful data contributed to the loss function. Training examples consisted of blocks of $L = 20$ consecutive time frames (corresponding to 105 ms). We generated a *training set* of 1.9 million examples (approximately 55.4 hours) and a *validation set* of 200000 examples (approximately 5.8 hours). A summary of the training data is presented in Table 4.2.

4.5.3 Evaluation data

We evaluated the performance of all systems for static source scenarios in reverberant environments. Source signals, background noise signals, acoustic conditions, and source positions were distinct from those used during training. Table 4.3 summarizes the evaluation setup and data generation. Two real environments were used, a cafeteria environment with a reverberation time of approximately 1300 ms, and a courtyard environment with a reverberation time of approximately 900 ms. The clean binaural microphone signals were generated by convolving the speech source signals with reverberant BRIRs[1]. Figs. 4·4a and 4·4b illustrate the room configurations. In each environment, we considered four source positions (depicted with dashed boxes) with two head orientations measured for each position. Two-source and three-source scenarios were created for each environment by combining all possible pairs and triplets of source positions across the two head orientations, resulting in 12 two-source and 8 three-source scenarios. Performance was evaluated at SNRs from -5 dB to $+10$ dB in unmatched background conditions, where recorded cafeteria babble noise and courtyard ambient noise [1] were used. A total number of 150 speech utterances (each with a length of 2 s) randomly chosen from 30 unique male and female speakers were selected from the validation TIMIT dataset.



(a) Cafeteria (static scenario)

(b) Courtyard (static scenario)

Figure 4·4: Evaluation setups for static scenarios, adapted from [1]. In the cafeteria, source positions A, B, D, E were considered, while in the courtyard, source positions A, B, C, and D were considered. Dashed arrows extending from each source position towards the head indicate the head location. Head orientations are indicated by the numerals 1 and 2, which are placed close to the head icon.

4.5.4 Implementation details

All signals were sampled at 16 kHz. CPS phase and magnitude spectrogram features were calculated using an STFT with a Hann window of length $K = 160$ (corresponding to 10 ms) and a hop size of $D = 80$ corresponding to 5 ms), yielding 81 STFT frequency bins. For each of the 6 microphone pairs, CPS features were computed over a block of $L = 20$ consecutive time frames, resulting in a CPS input feature of size $20 \times 81 \times 6$.

In this paper, we consider the front microphone of the left hearing aid as the reference microphone for the PD and magnitude spectrogram feature extraction. The magnitude spectrogram was also calculated over 20 STFT time frames and 81 frequency bins, resulting in an input feature size of 20×81, aligning with the CPS input dimensions for joint spectro-temporal processing.

PD features were computed using a 4-th order GTFB implementation [48] with 61 bands, a group delay of 256, and minimum and maximum center frequencies of 60 Hz and 7200 Hz, respectively. By choosing the maximum and minimum fundamental frequencies as 320 Hz and 70 Hz, respectively, the range of fundamental period candidates for PD feature extraction lies between 3.1 ms and 14.3 ms for $N = 180$ period candidates. The comb filter gain was chosen to be $\alpha = 0.7$. After adjusting the frequency resolution of the PD features to match the STFT (cf. Section 4.3.2), the input PD features had a size of $20 \times 81 \times 180$, aligning them with the spectro-temporal dimensions of the CPS input features.

4.5.5 Training and network hyperparameters

All systems were implemented using PyTorch [156]. Convolutional blocks in all CNNs employed 64 filters with a stride size of 1×1. When used, max pooling had a size of 2×1 with strides of the same size. The training was conducted using the Adam optimizer [157], a binary cross-entropy loss function, an initial learning rate of 10^{-4}, a mini-batch size of 128, and a dropout rate of 0.5. We employed early stopping regularization that terminated training if the validation loss did not improve for 10 epochs. A variable learning rate scheduler was also used, halving the learning rate if the validation loss did not improve for 2 epochs. The maximum training epoch number was set to 100. Each epoch randomly selected 1.9 million non-repeating examples from the training set. Mini-batches were assembled randomly, drawing examples from various SNR conditions and DOA classes. Validation loss was computed using a fixed set of 200000 examples from the validation set. The validation data were not seen by the network during the training.

4.5.6 Evaluation metrics

We evaluated the DOA estimation performance of all systems in terms of accuracy (Acc.) [14, 98]. For a signal block l containing a mixture of $\mathcal{I}$ sources, an estimated DOA for the i-th source ($\hat{\theta}_i^l$) is considered accurate if the absolute error between that and the oracle DOA(θ_i^l) is less than 5°, i.e., the minimum angular resolution of the database in [1]. The accuracy is defined as

$$\text{Acc.} = \frac{1}{\mathcal{L}\mathcal{I}} \sum_{l=1}^{\mathcal{L}} \sum_{i=1}^{\mathcal{I}} H(5 - \left|\hat{\theta}_i^l - \theta_i^l\right|) \times 100, \tag{4.7}$$

where $\mathcal{L}$ denotes the total number of signal blocks, and H denotes the Heaviside step function. Please note that the Heaviside step function is defined here such that it returns 1 if the absolute error is less than 5° (an accurate estimate), and 0 otherwise (an inaccurate estimate).

4.6 Results and discussion

In this section, we will present and analyze the performance evaluation results of the proposed system and the alternative two-stage configurations employing PD and CPS phase features, along with baseline systems using either the CPS phase or the combination of the CPS phase and magnitude spectrogram. We assessed the performance of all systems in different reverberant environments with different background noises for both static two-talker and three-talker scenarios in terms of accuracy. Section 4.6.1 compares the proposed system to the two alternative two-stage systems using different temporal dilation and max pooling strategies. In Section 4.6.2, we compare the performance of the proposed system and the two alternative two-stage systems using different spectro-temporal processing strategies. Finally, Section 4.6.3 discusses the performance evaluation of the proposed and baseline systems.

4.6.1 Different temporal reduction strategies

For two-source and three-source scenarios in two reverberant environments (cafeteria and courtyard), Fig. 4·5 shows the accuracy at different SNRs for the proposed system, and the two

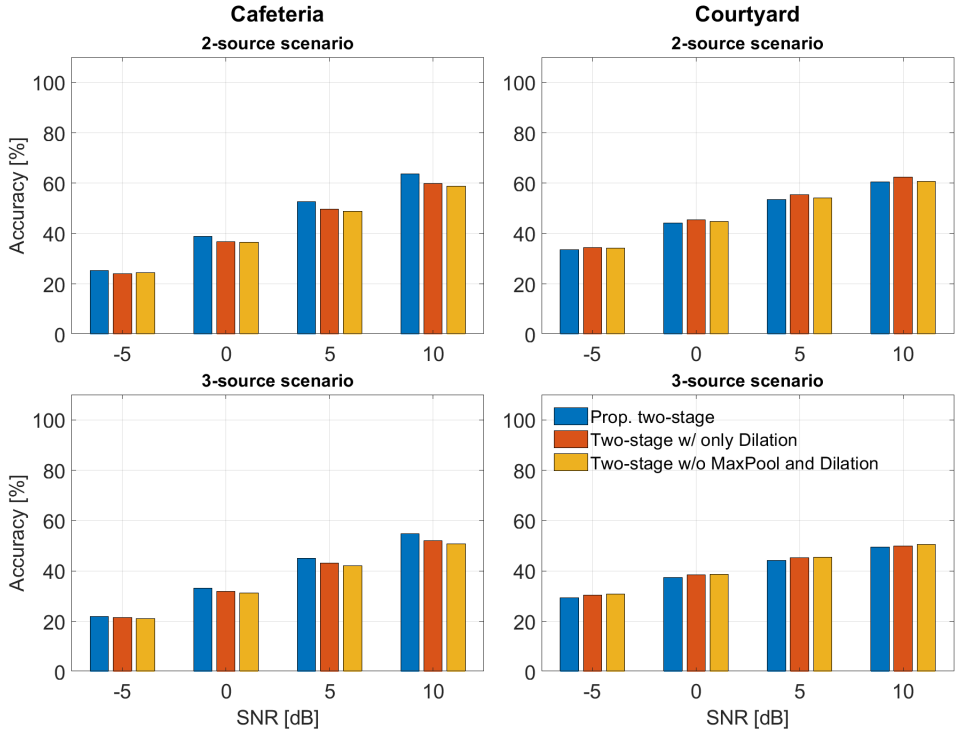


Figure 4·5: Accuracy of the proposed system and the two two-stage CNN configurations using different temporal feature reductions for static source scenarios in two reverberant environments (cafeteria and courtyard) for different SNRs and two-talker and three-talker scenarios. The proposed system (indicated by the blue bar) employs a dilation and max pooling size combination that leads to a temporal dimension of size 1 for each filter output, while the red and orange bars correspond to counterpart configurations, resulting in temporal dimensions of sizes 2 and 8, respectively. All systems use the combination of PD and CPS phase as input, without any max pooling or convolutional kernels across frequencies.

alternative two-stage systems (Section 4.4.2) using different temporal feature reduction strategies. The blue bars indicate the evaluation results of the proposed system employing dilated kernels and max pooling across time that leads to a temporal dimension of size 1 for each filter of the *Conv*2 block (cf. Fig. 4·3). The red and orange bars correspond to alternative configurations, one with dilated kernels but no max pooling, leading to a temporal dimension of 2, and another with neither max pooling nor dilation across time, yielding a temporal dimension of 8 (as detailed in Table 4.1).

In particular, we intend to test the hypothesis that the temporal dependencies in the input features can be effectively captured merely by convolutional blocks, while the frequency resolution of features is preserved. The latter is ensured by using no Max Pooling and no convolutional kernels across frequencies. This would essentially mean that global patterns across frequencies are exclusively captured by the fully-connected blocks.

It can be observed that in the cafeteria environment, the proposed system clearly outperforms the other two-stage configurations. In the courtyard environment, all three systems perform

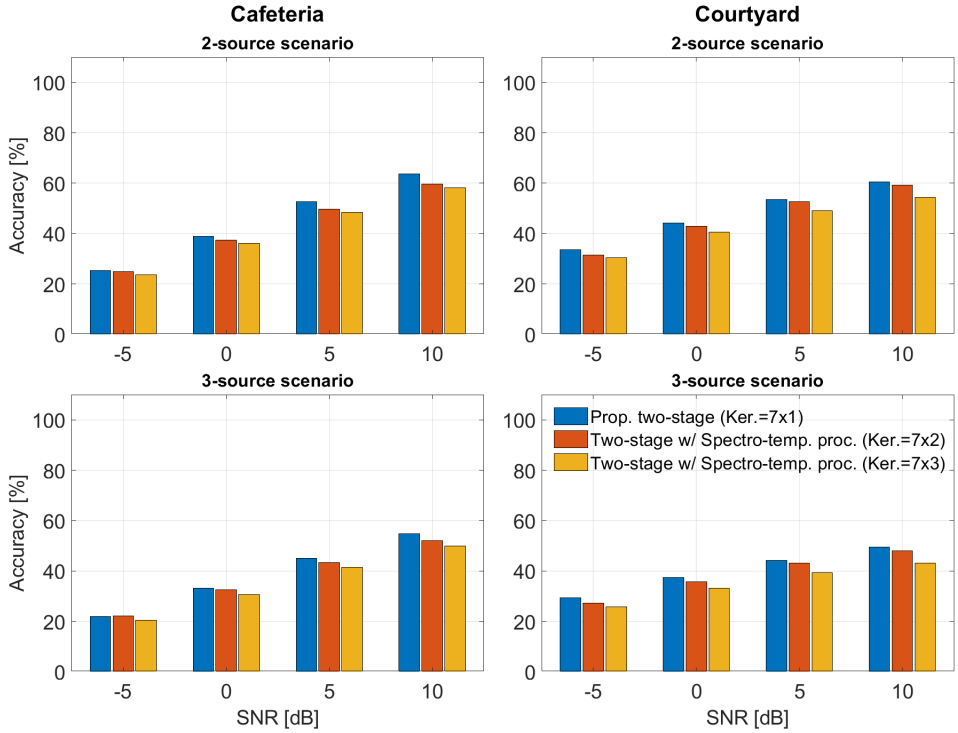


Figure 4·6: Accuracy of the proposed system and the two two-stage CNN configurations using different spectro-temporal processing for static source scenarios in two reverberant environments (cafeteria and courtyard) for different SNRs and two-talker and three-talker scenarios. The proposed system (indicated by the blue bar) employs a kernel size of 1 across frequencies (no frequency correlation), while the red and orange bars correspond to counterpart configurations, using kernel sizes of 2 and 3 across frequencies. All systems use the combination of PD and CPS phase as input, and employ kernels of size 7 with a dilation rate and max pooling size of 2 across time, and without any max pooling across frequencies.

comparably, with slightly higher accuracy for the two-stage configuration using no temporal max pooling or dilated kernels (yellow bar), nonetheless at the expense of significantly more computational cost and model size (cf. Table 4.1).

It is important to note that using a model with the highest computational complexity does not necessarily lead to a better performance in terms of accuracy. In fact, the proposed system, which captures temporal dependencies in the input features through convolutional blocks, proves to be a favorable approach due to its more efficient configuration with less computational complexity and model size.

4.6.2 Different Spectro-temporal filtering strategies

In this section, in comparison to the proposed system that only employs dilated convolutional kernels of size 7 and dilation rate of 2 across time, we investigate the potential benefit of using kernels across both time and frequency in alternative two-stage systems using kernels of sizes 2 and 3 across frequencies, while utilizing the same temporal kernel sizes as the proposed system.

In all systems, the temporal dimension across convolutional blocks is reduced to one.

For two-source and three-source scenarios in two reverberant environments (cafeteria and courtyard), Fig. 4·6 shows the accuracy at different SNRs for the three two-stage systems using the PD and CPS phase as inputs through different spectro-temporal feature processing. The blue bars indicate the evaluation results of the proposed system using the kernel size of 7x1, while the red and orange bars represent the alternative two-stage system configurations using kernel sizes of 7x2 and 7x3, respectively.

It can be clearly observed that in all environments and SNR conditions for the two-talker and three-talker scenarios, the proposed system performs comparably to or better than the alternative two-stage systems. This demonstrates that the two-stage CNN using PD and CPS features does not benefit from the joint spectro-temporal filtering using 2D kernels. This also suggests that while capturing the temporal dependencies solely through the convolutional blocks (i.e. temporal dimension reduction to a single value), it is more effective to process each frequency independently through the convolutional path. In this way, the fully-connected layers alone can effectively learn global patterns across frequencies, rather than having both the convolutional and fully-connected blocks contribute to learning these patterns. It is particularly notable when considering the additional computational load from the joint spectro-temporal processing in the convolutional path (cf. Table 4.1), which further demonstrates the benefit of independently processing each frequency.

4.6.3 Comparison against baseline systems

In this section, we evaluate the advantage of incorporating PD features in combination with the CPS phase as a spatial feature in our proposed system (cf. Fig. 4·3), compared to baseline systems that use either the CPS phase or a combination of CPS phase and magnitude spectrogram as a spectral feature (cf. Fig. 4·2). All systems are implemented using the same convolutional kernel, dilation, and max pooling strategies (cf. Table 4.1). The key distinction is that our proposed system includes a feature reduction stage before merging the PD saliency features and CPS phase features. In contrast, the baseline system using the spectral and spatial features directly combines the magnitude spectrum with CPS phase features.

For two-source and three-source scenarios in two reverberant environments (cafeteria and courtyard), Fig. 4·7 depicts the accuracy at different SNRs for the proposed system and the two baseline systems. The blue bars indicate the evaluation results of the proposed system using the PD and CPS phase as input, while the red and orange bars represent the baseline systems using the combination of the magnitude spectrogram and CPS phase, and the one using the CPS phase as input, respectively.

For all conditions and environments, the proposed system clearly benefits from using the PD features in combination with CPS features, when compared to the two baseline systems. For example, for an SNR of 0 dB in the courtyard environment, for two-source scenarios, the benefit of using PD features is approximately 4% points compared to the baseline system using only the CPS phase, and 5% points compared to the baseline system using the magnitude spectrogram and CPS phase. For three-source scenarios, the benefit of using PD features is approximately

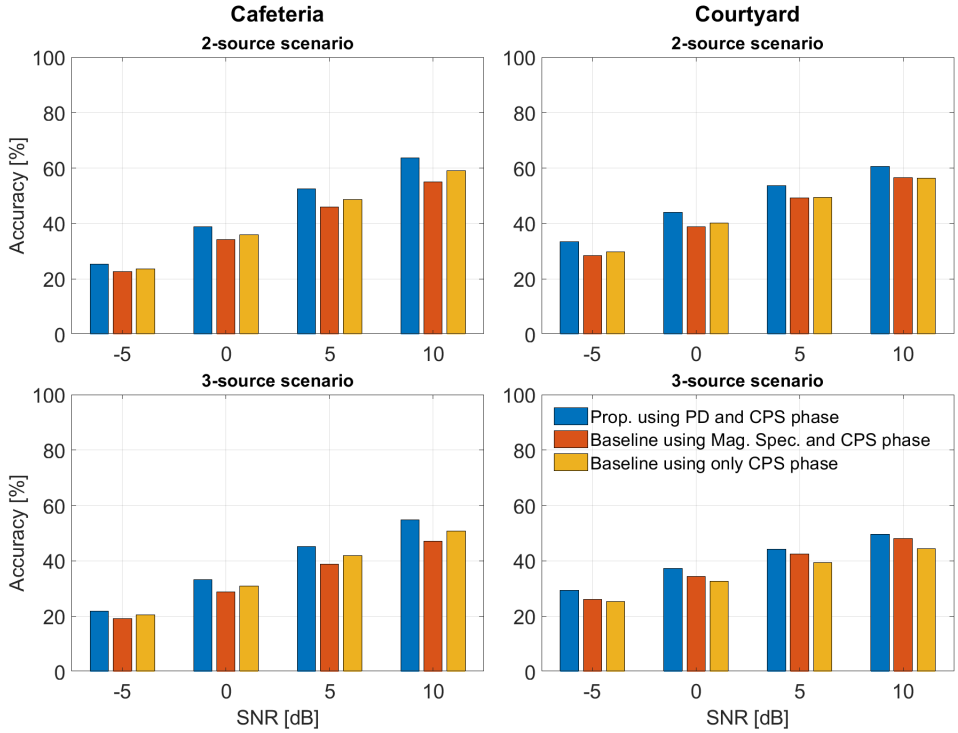


Figure 4·7: Accuracy of the proposed system, and the two baseline systems for static source scenarios in two reverberant environments (cafeteria and courtyard) for different SNRs and two-talker and three-talker scenarios. The proposed system (indicated by the blue bar) employs the PD and CPS phase as input, while baseline systems specified by the red and orange bars employ the combination of magnitude spectrogram and CPS phase, and the CPS phase, respectively. All systems employ a convolutional kernel size of 7 with dilation rate and max pooling size of 2 across time, without using any max pooling or convolutional kernels across frequencies.

5% points compared to the baseline system using only CPS phase, and 3% points compared to the baseline system using the magnitude spectrogram and CPS phase.

We can also observe from Fig. 4·7 that the benefit of using PD features increases with SNR. For example, comparing the performance of the proposed system and the baseline system using only the CPS phase in the cafeteria environment exhibits that, for the two-source scenarios, this benefit increases from about 1% points at -5 dB to about 5% points at 10 dB SNR condition. At higher SNRs, the impact of background noise is reduced, thus emphasizing the periodicity characteristics of speech and making the periodicity features more distinguishable. This improved discriminability enhances the ability of the proposed system to effectively use PD features in conjunction with spatial CPS features, allowing the PD features to contribute more meaningfully to the accuracy of the DOA estimation.

Including the magnitude spectrogram in combination with the CPS phase seems to be advantageous merely in the courtyard environment, in particular, for the thee-source scenario. This observation suggests that, unlike PD features, the usage of the magnitude spectrogram as a spectral feature in combination with the CPS phase does not offer significant benefits for

DOA estimation for the considered settings and environments when compared to using only CPS phase features. On the one hand, While PD features provide a clear indication of the source's harmonic structure, the magnitude spectrogram provides a broad spectral representation that may not be as effective in isolating the specific characteristics of speech needed for accurate DOA estimation, or may require a much more sophisticated network architecture to capture these characteristics. On the other hand, PD features are less susceptible to noise that does not share the harmonic structure of the sound sources of interest, while magnitude spectrogram features are more general and can capture both the speech signal and noise without distinguishing between them, making it harder to identify speech sources in noisy environments.

4.7 Conclusion

This paper investigated the effectiveness of combining periodicity and spatial features for multi-talker DOA estimation in binaural hearing aids using a two-stage convolutional neural network (CNN) architecture. The proposed system utilized periodicity degree (PD) features as spectral features in combination with cross-power spectrum (CPS) phase as spatial features.

Several design choices for the two-stage CNN architecture were explored, including different strategies for temporal feature reduction through dilation and max pooling, as well as spectro-temporal filtering approaches using convolutional kernels of varying sizes. Experimental results demonstrated that the proposed system which effectively captures the temporal dependencies within the convolutional blocks alone, while independently processing each frequency, leads to the best performance. Furthermore, the proposed system offers advantages in terms of computational complexity compared to alternative configurations.

The results demonstrated that the proposed system outperformed baseline systems which utilized either CPS features or a combination of CPS and magnitude spectrogram features in terms of accuracy for both two-talker and three-talker scenarios in two reverberant environments and across various SNR conditions. The proposed system achieved this improvement without requiring significantly higher computational complexity or model size compared to the baseline systems.

This study paves the way for advancements in sound source localization and speech enhancement for binaural hearing aids. By combining periodicity and spatial features, the research demonstrates the potential for more accurate DOA estimation and broader improvements in various speech-related tasks. Moreover, this study underscores the importance of feature selection in designing systems for complex auditory scene analysis, particularly in noisy and reverberant environments and multi-source scenarios.

Future work could explore the adaptation and integration of the proposed system for real-time processing pipelines. Additionally, further research could investigate using PD features to enhance the spatial features by taking alternative approaches other than the direct combination of features. For instance, by exploiting periodicity features for learning-based mask estimation techniques, which might potentially achieve even better DOA estimation performance.

Chapter 5

Conclusions and Future Research

This chapter summarizes the key contributions outlined in the thesis and explores potential directions for future research. We highlight the main findings and suggest possible extensions that could serve as a continuation of the work conducted in this thesis.

5.1 Conclusions

By accurately determining the DOA of desired speech sources, hearing aids can selectively amplify sounds from that direction while suppressing noise from other directions. This targeted approach improves speech intelligibility for hearing aid users, especially in noisy environments. The main objective of the thesis was to improve the accuracy of DOA estimation in binaural hearing aids, particularly in challenging acoustic conditions such as noisy and reverberant environments, by leveraging auditory-inspired periodicity features as a distinctive characteristic of speech signals in combination with spatial features within deep learning frameworks, specifically CNNs.

Chapters 2 through 4 focused on advancing binaural DOA estimation systems, with each chapter building on the previous one to address increasingly complex auditory environments and refine deep learning approaches. Chapter 2 introduced a speech-aware model for single-talker DOA estimation in complex auditory scenes, proposing a dual-path CNN. Building on these findings, Chapter 3 proposed a more efficient network architecture, a two-stage CNN with feature reduction, to streamline spectro-temporal processing while maintaining the benefit of speech-aware DOA estimation in single-talker scenarios. Finally, Chapter 4 extended these methods to tackle multi-talker DOA estimation, by adapting the model architecture to handle multiple talkers effectively and enabling efficient spectro-temporal processing. The model architectures in each chapter were based on CNNs, incorporating optimized architectural modifications to enhance computational efficiency and accuracy. The evolution from a dual-path CNN to a two-stage CNN with feature reduction highlights this progression. Various combinations of spatial and periodicity features were considered as input features to different CNN architectures. Spatial features include both broadband (GCC-PHAT) and narrowband (magnitude, phase, real, and imaginary components of CPS) representations. Periodicity features were represented in different forms, including broadband, subband-averaged, and narrowband PD features.

All chapters in this thesis share several core features that shape a cohesive approach to advancing binaural DOA estimation in this thesis. A key theme across all three chapters is integrating periodicity features with spatial features. This strategy proves considerably more effective than using spatial features alone or with a VAD. This combination enhanced DOA estimation accuracy for different spatial features across varying auditory tasks and model architectures. For all systems, the DOA estimation problem was formulated as a classification task in the horizontal plane, with the full 360° azimuthal range divided into 72 discrete directions, each with a 5° angular resolution. All chapters rely on a database of multichannel BRIRs for data generation. This database provides a wide range of BRIRs for various environments, source positions, and head orientations. A binaural hearing aid setup with four microphones was considered, including two microphones (front and rear) on each hearing aid. All systems employed supervised learning approaches and were trained using labeled audio data with diverse sound source directions and SNR conditions in an anechoic environment using simulated diffuse noise as background noise. All systems were evaluated using separate sets of BRIRs recorded in real reverberant environments like a cafeteria, courtyard, or office with diverse source positions, with background noise recordings with different SNRs. The types of background noise varied, in-

cluding simulated diffuse noise, recorded cafeteria babble, and courtyard ambient noise. Varied training and evaluation conditions were used to assess the model's ability to generalize and adapt to different acoustic environments. By systematically building upon each chapter's findings, the thesis demonstrates a clear progression toward robust and efficient binaural DOA estimation systems for both single-talker and multi-talker scenarios. The integration of periodicity features and the refinement of CNN architectures represent significant contributions to the field of DOA estimation for hearing aids.

In **Chapter 2**, we proposed a novel approach for single-talker binaural DOA estimation called *speech-aware* DOA estimation. This method aimed to accurately estimate the DOA of a single talker only when speech is present, without requiring a separate VAD. The key innovation was the integration of auditory-inspired periodicity features with spatial features for different feature combinations. This combination, as hypothesized in Chapter 1, addressed the limitations of systems that solely rely on spatial features and require separate VADs to handle speech inactivity. The chapter explored various feature combinations, including broadband GCC-PHAT with broadband PD and narrowband CPS features (real and imaginary parts or magnitude and phase components) with subband-averaged PD. These features were fed into dual-path CNN architectures, where separate branches processed the spatial and periodicity features before combining their outputs for DOA estimation. These combinations resulted in different proposed systems for speech-aware DOA estimation. The use of dual-path CNN architecture allowed for the separate processing of spatial and periodicity features, potentially allowing each branch to specialize in extracting relevant information from its respective input. The proposed systems were trained using speech (TIMIT dataset) and non-speech signals (ESC50 dataset). This allowed the network to differentiate between speech and non-speech, a crucial aspect of the speech-aware design. Evaluations were conducted in cafeteria and courtyard environments using recorded background noises at different SNRs and speech source signals. The key performance metrics used were accuracy, mean absolute error, precision, and recall. These metrics provide a comprehensive view of the system's capabilities in estimating the DOA and detecting the presence of speech. Evaluations across static and dynamic single-speaker scenarios demonstrated that incorporating PD features consistently improved DOA estimation accuracy and reduced angular error, outperforming baseline systems using the same spatial features with a cascaded pitch-based VAD. These findings strongly supported the hypothesis that combining periodicity and spatial features enhances the robustness of DOA estimation. All proposed speech-aware systems achieved near-perfect precision. This indicates a low likelihood of false positives, suggesting that the systems effectively distinguish speech from non-speech. While demonstrating excellent precision, the proposed systems exhibited lower recall than the integrated VAD of the baseline systems, particularly at low SNRs. However, it is essential to remember that the proposed systems are not designed solely for speech detection but for speech-aware DOA estimation. The lower recall indicates that the systems are more conservative in classifying segments as speech. A closer analysis revealed that the speech-aware system primarily estimates DOA in segments with a high degree of periodicity, potentially to avoid estimating DOA from noisy or unreliable segments. Chapter 2 effectively established the benefit of incorporating periodicity features for speech-aware DOA estimation.

In **Chapter 3**, building on the benefits of combining spatial and periodicity features as discussed in Chapter 2, we proposed integrating narrowband spatial and narrowband periodicity features directly into the input of the CNN, rather than feeding them into separate CNN branches. We also refined the speech-aware single-talker DOA estimation system by introducing a more computationally efficient two-stage CNN architecture. This chapter utilized the real and imaginary components of CPS and narrowband PD features as input. The chapter's key contribution was the implementation of a feature reduction stage using 1×1 convolutions to address the high dimensionality and sparsity of the narrowband PD features in the proposed two-stage CNN architecture. This reduction created compact PD saliency features, facilitating spectro-temporal processing within the CNN. This chapter employed a similar experimental methodology to Chapter 2, by training the proposed systems (with and without feature reduction) using both speech signals and non-speech signals. The training was conducted in simulated anechoic conditions, and evaluation was conducted in reverberant environments (cafeteria and courtyard) using diverse background noise recordings at various SNRs. The performance of the proposed systems in terms of accuracy and MAE was compared against a baseline system using a CNN with only CPS features and a pitch-based VAD. Both proposed systems consistently outperformed the baseline system in terms of DOA estimation accuracy and mean absolute error. Importantly, the two-stage CNN with feature reduction achieved the better performance despite having significantly fewer trainable parameters compared to the proposed system without feature reduction, and comparable to the baseline system. This highlighted the effectiveness of the feature reduction stage in capturing relevant information from the sparse PD features and improving computational efficiency. These findings are further underscored in Appendix A, where the performance of the proposed and baseline systems are evaluated in single-talker scenarios with non-speech interference. This investigation demonstrates the robustness and efficacy of the two-stage CNN in real-world scenarios. Chapter 3 demonstrated the benefits of incorporating periodicity features for speech-aware DOA estimation in CNN architectures, extending beyond those proposed in Chapter 2 through a more efficient approach. This finding is particularly relevant for implementing DOA estimation in hearing aids, where processing power and battery life are critical constraints. The feature reduction stage in the two-stage CNN can be viewed as a form of attention mechanism, guiding the network to focus on the most salient glimpses of the PD features, potentially mimicking the capabilities of human auditory scene analysis. This mechanism is hypothesized to improve robustness in noisy environments by reducing the impact of irrelevant information within the PD features.

Building upon the advancements made particularly in Chapter 3, in **Chapter 4**, we tackled the task of multi-talker DOA estimation by adapting the two-stage CNN architecture to handle multiple simultaneous talkers. We proposed incorporating narrowband PD features and CPS phase features as input. Consistent with earlier chapters, training was conducted in simulated anechoic conditions across different SNR conditions with a simulated diffuse noise. Evaluations employed realistic reverberant environments with recorded background noise at different SNRs in static multi-talker scenarios with two and three simultaneous speakers. The chapter investigated different strategies for capturing spectro-temporal dependencies within the two-stage CNN, including using dilated kernels, max pooling, and various kernel sizes across the frequency

dimension. These architectural choices aimed to find an efficient and effective CNN configuration that could accurately estimate multiple DOAs without significantly increasing model size or computational complexity. The findings revealed that the two-stage CNN processes each frequency independently and performs comparably to or better than two-stage CNNs with joint spectro-temporal processing. This suggests that, in multi-talker scenarios, considering correlations across frequencies within the convolutional blocks might not necessarily lead to improved performance. Independently processing each frequency simplifies the model and reduces computational complexity. The best performance was achieved by a the proposed two-stage CNN that independently processes each frequency bin while using dilated kernels with max pooling across time. This configuration employs convolutional kernels exclusively to capture temporal dependencies. The performance of the proposed two-stage CNN with this configuration was compared against two baseline systems, one utilizing only the CPS phase and another combining the CPS phase with magnitude spectrograms. Evaluations demonstrated that the proposed two-stage CNN consistently outperformed both baseline systems across all evaluated SNRs and environments. These results highlight the effectiveness of combining PD and spatial features for multi-talker DOA estimation within a carefully designed two-stage CNN architecture. The system achieved these improvements without requiring a substantially larger model size or increased computational complexity compared to the baseline systems. Chapter 4 successfully extends the benefits of integrating periodicity and spatial features, previously demonstrated for single-talker scenarios, to the more challenging task of multi-talker DOA estimation. It underscores the importance of carefully considering temporal and spectro-temporal processing strategies within the CNN to achieve optimal performance. Developing a computationally efficient architecture in this chapter opens up possibilities for implementing multi-talker DOA estimation in resource-constrained devices like hearing aids.

The difference in task definitions for single- and multi-talker DOA estimation led to different approaches in the output coding of the CNNs and the data labeling process. Chapters 2 and 3 focused on speech-aware single-talker DOA estimation, training the networks to identify speech segments and estimate DOAs only during those segments. By incorporating a detection class alongside the DOA classes in the output layer, this approach inherently integrated speech detection into the DOA estimation framework, thereby eliminating the need for a separate VAD. By training the CNN with a mixture of speech and non-speech signals, the network learns to associate the presence of a strong harmonic structure in the PD features with speech segments. This is a crucial aspect of the speech-aware models, allowing the network to perform speech detection implicitly. The speech-aware approach in these chapters modeled the output as a joint probability distribution over the detection and DOA classes using a softmax activation function, assuming that these classes are mutually exclusive. Unlike this approach, the multi-talker systems did not aim to condition the DOA estimation on speech detection. Instead, the proposed multi-talker system modeled the output using only the DOA classes, assuming that sound source locations are mutually independent, thereby allowing each output to independently predict the probability of a source at its corresponding DOA class. Unlike the softmax function, which assumes mutual exclusivity among classes and forces the probabilities to sum to one, multiple sigmoid activation functions in the output layer of multi-talker systems enable the model to

represent multiple active DOAs simultaneously. The network is trained exclusively on speech signals without non-speech examples during training. Each training example is labeled with the DOAs of the active speech sources. This simplifies the data generation and labeling process, as there is no need to generate or label non-speech data. However, the network must learn to associate the intricate harmonic pattern of speech signals with the relevant spatial information without the contrasting information provided by non-speech signals.

5.2 Suggestions for further research

The primary focus of this thesis was to investigate the effectiveness of combining periodicity features with standard spatial features within deep learning frameworks. We explored various CNN architectures and design choices to optimize both performance and computational efficiency for the proposed feature combinations. The findings demonstrated the benefit of including periodicity features for binaural DOA estimation while identifying several promising directions for future research.

Alternative approaches for reference microphone selection in PD feature extraction: In this thesis, the PD features were extracted using a single reference microphone signal, specifically the front microphone of the left hearing aid. The front microphone is oriented toward the frontal hemisphere, where desired sound sources are most likely located. This orientation allows the front microphone to capture a stronger direct sound component and possibly a higher SNR than the rear microphone. However, this choice does not account for the potential variations in signal quality, SNR, or various acoustic conditions that may affect different microphones. The optimal reference microphone selection was not within the scope of this thesis. Future research could explore strategies for dynamically selecting the reference microphone based on real-time assessments of signal quality or periodicity measures. By continuously evaluating the SNR or periodicity features at each microphone, the system could select the microphone that currently provides the most reliable PD features. This adaptive approach would enhance the robustness of the PD feature extraction process in varying acoustic conditions. Alternatively, PD features could be extracted from multiple microphones and combined instead of relying on a single reference microphone to form a more robust representation. Techniques such as averaging the PD features or weighted combining based on signal quality metrics could be investigated. This approach can potentially improve the system's resilience to noise and interference affecting individual microphones. Incorporating PD features from multiple microphones or implementing dynamic selection strategies may increase computational demands. Future work should focus on optimizing algorithms to balance the benefits of enhanced PD feature extraction with the constraints of processing power and energy consumption, which is especially important for real-time applications in hearing aids.

Extesnsions to the speech-aware approach for simultaneous speech/non-speech detection and DOA estimation: The speech-aware DOA estimation approach presented in Chapters 2 and 3 of the thesis is designed to estimate the DOA of a single talker while inher-

ently handling speech inactivity by including a detection class in addition to the DOA classes. The labeling strategy of this approach during the training phase associates each training example with only one active class (either a specific DOA class for speech or the detection class for non-speech), effectively treating the problem as a single-label classification task. By labeling all DOA classes as zero for non-speech examples, regardless of the actual direction from which the non-speech sound originates, the network is not trained to utilize the DOA information for non-speech sounds, essentially discarding the spatial information of non-speech sounds during training. Future research could extend the current approach to detect and localize both speech and non-speech sources simultaneously, by using auditory-inspired periodicity features in combination with spatial features alongside modifications to the network architecture and training methodology. First, the output layer of CNN needs to be expanded to support the detection and localization of non-speech events in addition to speech. This expansion could be tailored to specific task assumptions, such as scenarios involving a single talker and a single non-speech event, or multiple talkers and multiple non-speech events. The latter, which is a more general task, could include a set of non-speech DOA classes and removing the detection class altogether. The neural network would then have two sets of DOA classes, one set for speech and another set for non-speech sounds. Both sets would use sigmoid activation functions in the output layer, allowing each output neuron to independently represent the probability of its corresponding class. This modification transforms the problem into a multi-label classification task, enabling the network to predict multiple classes simultaneously. Including non-speech DOA classes preserves the spatial information that was previously lost for non-speech sounds. For speech-only examples, the speech DOA class corresponding to the speech source's direction is labeled as one, and all other speech DOA classes are labeled as zero. All non-speech DOA classes are also labeled as zero. For non-speech-only examples, the non-speech DOA class corresponding to the non-speech sound's direction is labeled as one, all other non-speech DOA classes are labeled as zero, and all speech DOA classes are labeled as zero. In cases where both speech and non-speech sounds are present simultaneously, the corresponding DOA classes in both speech and non-speech sets are labeled as one. Implementing this modification involves careful consideration of the network architecture, training data preparation, and feature extraction. The training dataset should include a large set of examples containing speech, non-speech, and possibly mixed sounds from various DOAs and acoustic conditions.

Alternative approaches to utilize periodicity features in DOA estimation: In this thesis, we mainly investigated two approaches for the fusion of periodicity and spatial features. In Chapter 2, two parallel branches of convolutional blocks process the periodicity and spatial features, each leading to some feature embeddings. These embeddings were concatenated and then used as input to fully-connected blocks. The primary role of the convolutional blocks in both branches is to extract local patterns from the input data in terms of feature embeddings. Fully-connected blocks take all the learned embeddings (periodicity and spatial) and combine them in a way that optimally *weighs* their importance, learning global patterns in the features extracted for the DOA estimation task. We may consider this weighting as a sort of *implicit attention* mechanism on the spatial embeddings. However, this dual-path architecture may limit

the network's ability to learn joint representations due to separate local feature learning in each branch.

To allow for joint learning of feature interactions, in Chapter 3, we considered a different fusion approach and proposed combining both spatial and periodicity features directly in the input. In particular, the two-stage CNN first learns a salient and compact representation from high-dimensional PD features in each time-frequency bin. These PD saliency features are then combined with the spatial features and processed via the convolutional blocks. The convolutional blocks that follow the fusion of PD saliency and spatial features could naturally learn to emphasize certain spectro-temporal regions of the spatial features based on the PD values. Regions with higher PD saliency values may play a more significant role in how the network processes spatial cues, effectively leading to an *implicit attention* mechanism where certain parts of the spatial features may be weighted more heavily than others, depending on their alignment with PD saliency features. While this architecture has shown improvements in DOA estimation performance, future research could investigate whether combining PD saliency and spatial features as input to a CNN is optimal for binaural DOA estimation. The binaural DOA estimation may benefit from introducing an *explicit attention* mechanism to ensure that periodicity features influence spatial features in a more structured way. A promising approach is a masking-based feature fusion [116] that could be tailored for future studies, where PD features could be used to estimate time-frequency masks that emphasize speech-dominated regions. These masks could be applied to spatial features, effectively filtering the spatial cues based on speech presence before convolutional processing. The mask values, derived from PD features, may indicate the likelihood of speech in each time-frequency bin. By multiplying the spatial features with this mask, regions with low speech presence or periodicity are suppressed, providing a refined input to the network.

In addition to these suggestions, which could further improve performance or extend the applicability of periodicity features to other auditory tasks, there are several other potential avenues for exploration.

In this thesis, all models primarily use a dataset [1] of measured BRIRs from a single artificial head. This approach may not capture the full diversity of real-world and individual differences in HRTFs for different hearing aid users, potentially limiting the models' generalization capabilities for deployment. BRIRs from various artificial heads with different shapes and sizes can help account for individual anatomical variations. The training data generation was also performed in an anechoic environment using BRIRs of sources located in different directions with a fixed source-to-head distance. Expanding the BRIR database to encompass source positions with different distances to the hearing aid setup and different reverberation times can enhance the models' robustness and generalizability to diverse acoustic environments.

The computational complexity of the proposed systems, while being efficient compared to alternative configurations, is still demanding for direct implementation in current hearing aid devices. Further model optimization techniques, such as model quantization and pruning, could be explored to reduce the computational requirements and facilitate real-time processing in hearing aid devices.

The thesis primarily employs CNN architectures for DOA estimation. While CNNs effec-

tively capture spatial and temporal patterns in audio features, alternative architectures such as RNNs, long short-term memory networks (LSTMs), or attention-based models might better capture temporal dependencies and handle dynamic scenarios with moving sources. Future research could investigate these architectures or hybrid models that combine CNNs with RNNs or attention mechanisms to potentially improve performance.

For hearing aid users to fully benefit, the proposed methods must be integrated with existing hearing aid signal processing pipelines. This may include considering how the DOA estimation outputs can be effectively utilized by beamforming algorithms, noise reduction techniques, and other components of the hearing aid processing chain. Future research could focus on integration and real-time implementation, ensuring that the proposed methods contribute to overall improvements in hearing aid performance.

Appendix A

Appendix to Chapter 3

A.1 Single-talker DOA estimation in the presence of non-speech interference

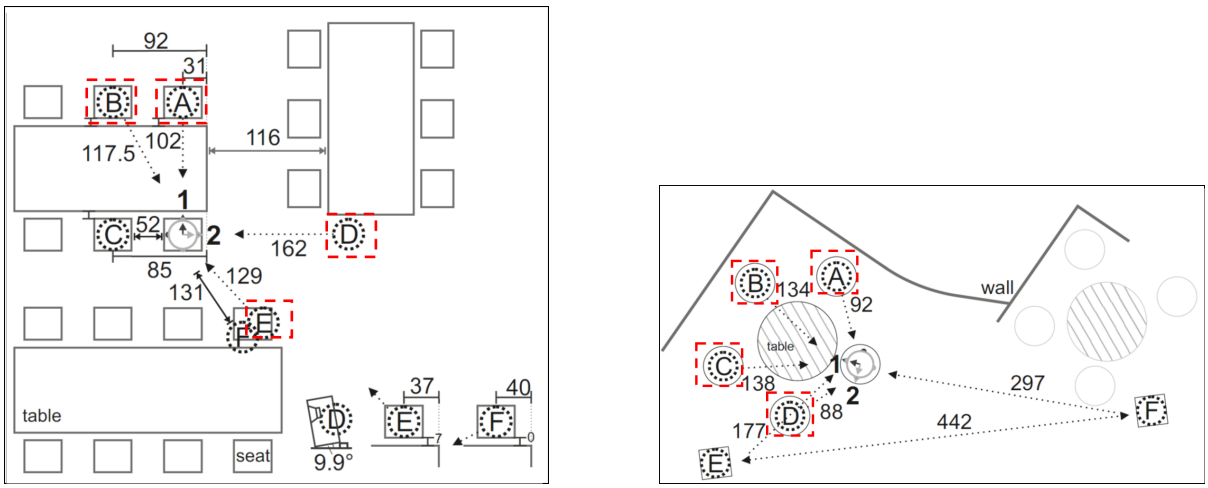
In this appendix, we present an additional experiment for Chapter 3. In Chapter 3, we proposed two speech-aware DOA estimation systems using the combination of narrowband PD and CPS features as input features (cf. Fig. 3·3), and compared their DOA estimation performance against a baseline system consisting of a CNN using only spatial CPS features as input cascaded with a pitch-based VAD (cf. Fig. 3·2). The performance evaluation was conducted in single-talker scenarios in the presence of recorded background noise. In this appendix, we consider the same three systems and evaluate them for single-talker scenarios in the presence of non-speech directional interference. All implementation details, including system configurations and training data, remain identical to those described in Chapter 3. No additional training has been conducted, and only the evaluation data has been modified. Specifically, the recorded background noise in Chapter 3 has been replaced with non-speech directional interference to generate the noisy microphone signals for this evaluation.

A.1.1 Evaluation data

The evaluation data were generated for static-source scenarios in two real environments, cafeteria and courtyard, with a reverberation time of approximately 1300 ms and 900 ms, respectively. All systems were evaluated at SNRs ranging from -5 dB to $+10$ dB in 5 dB steps. A total of 150 speech segments randomly chosen from 30 unique male and female speakers (each with a length of 1 s) from the validation TIMIT [152] were used as speech source signals. To generate noisy binaural microphone signals, we randomly selected non-speech sounds (e.g. chainsaws, washing machines, and chirping birds) from the validation ESC50 dataset [153] (each with a length of 1 s). In each environment, we considered BRIRs of two head orientations for four source positions as depicted in Fig. A·1. All four source positions for each head orientation were considered as the location of a single talker. For each talker position and speech segment, a position from the remaining three was randomly chosen as the location of the non-speech interference. The noisy binaural microphone signals were generated by convolving the speech and non-speech source signals with their respective BRIRs and mixing the resulting clean binaural microphone signals with the binaural interference signals at different SNRs.

A.1.2 Performance measures

The DOA estimation performance of all systems was evaluated in terms of accuracy and MAE, as defined in Chapter 3 (cf. (3.5) and (3.6)) and in Chapter 2 (cf. (2.15) and (2.16)). Additionally, the speech detection performance of the rVAD [138] in the baseline system and the speech DOA detection performance of the proposed systems were evaluated using the precision (P) and recall (R) metrics, as defined in Chapter 2 (cf. (2.17) and (2.18)). To compute precision and recall, a simple oracle broadband energy-based VAD was employed to identify speech segments, similar to the method in Chapter 2. Together, these metrics provide a comprehensive evaluation of the DOA estimation performance across all systems. Precision reflects the proportion of detected speech segments that are indeed correct. Obtaining high precision is important as the non-



(a) Cafeteria **(b)** Courtyard

Figure A·1: Evaluation setups, adapted from [1]. Static source positions *A*, *B*, *D*, and *E* were evaluated in the cafeteria, and positions *A*, *B*, *C*, and *D* in the courtyard. In both environments, dashed arrows originating at the source positions indicate the head position. The numbers 1 and 2, located near the head, denote the two head orientations. All distances are in centimeters.

speech sound could be mistakenly identified as speech (i.e. false positives), leading to erroneous estimates. Recall, on the other hand, represents the system's ability to capture all true speech segments, showing how well it detects speech.

A.1.3 Results and discussion

Fig. A·2 depicts the DOA estimation performance in terms of accuracy and mean absolute error at different SNRs for the two proposed speech-aware systems (cf. Section 3.4.2) and the baseline system (cf. Section 3.4.1). The blue bars and red bars show the metrics of the proposed systems with and without feature reduction, respectively, while the yellow bars represent the baseline system's metrics.

Comparing the performance of the two proposed systems against the baseline, it is evident that the integration of narrowband PD with the CPS feature significantly enhances single-talker DOA estimation accuracy in the presence of non-speech interference, across all SNR conditions in both environments. For example, in the cafeteria environment at 0 dB SNR, the proposed systems achieve accuracies of 68% (with feature reduction) and 64% (without feature reduction), compared to only 40% accuracy for the baseline. This improvement translates to a 24–28 percentage point benefit in accuracy over the baseline system. Similarly, in the courtyard environment at 0 dB SNR, the proposed systems demonstrate a 16–21 percentage point advantage in accuracy over the baseline. It is also evident that the proposed system with feature reduction (two-stage CNN) outperforms the one without feature reduction. These findings align with those presented in Chapter 3 in the presence of background noise, further demonstrating the effectiveness of the proposed speech-aware systems, especially the two-stage CNN, across various acoustic scenarios. Moreover, while the benefit is significantly large at low

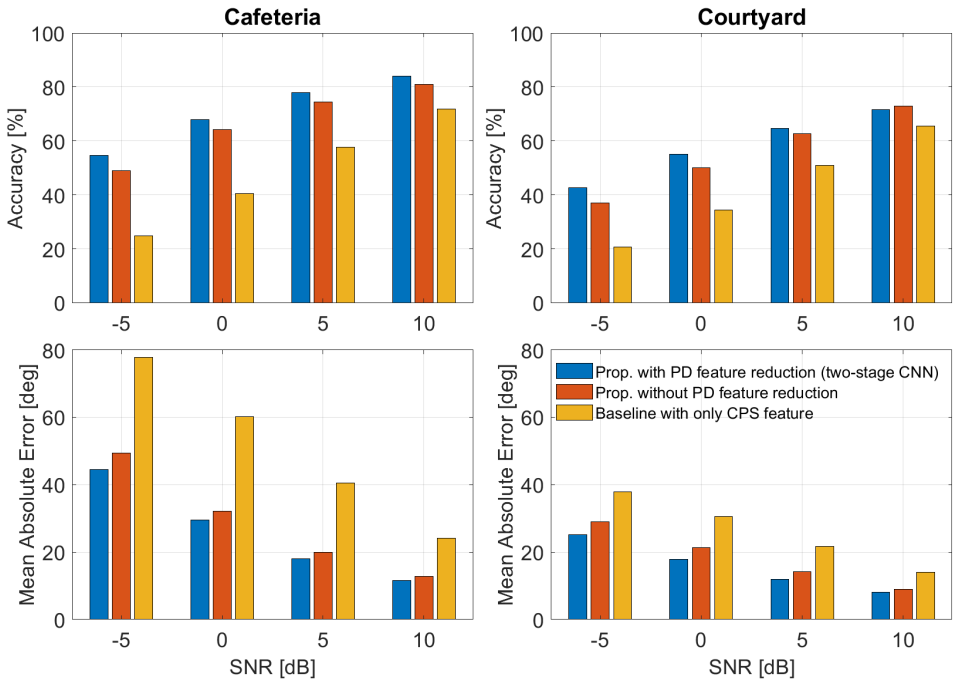


Figure A·2: The DOA estimation performance of the two proposed systems and the baseline system which were presented in Chapter 3. All systems are evaluated in terms of accuracy and mean absolute error for single-talker scenarios with *directional non-speech interference* across various SNR conditions in two reverberant environments (cafeteria and courtyard).

SNRs (−5 dB and 0 dB), it also persists toward high SNRs (5 dB and 10 dB).

The only difference between the experiment described in this appendix and the one in Chapter 3 is the use of non-speech interference instead of recorded background noise signals. This modification allows for a direct comparison of DOA estimation performance in the presence of non-speech interference (cf. Fig. A·2) with results obtained in the presence of background noise from Chapter 3 (cf. Fig. 3·4). The comparison indicates that non-speech interference presents a more challenging scenario for all systems compared to the recorded background noises used in Chapter 3. For example, in the courtyard environment, the DOA estimation accuracy of the two proposed systems with and without feature reduction decreases from approximately 80% (cf. Fig. 3·4) to 55% and 50% (cf. Fig. A·2), respectively, while the baseline system's accuracy declines from 73% to 34%. This demonstrates that although the performance of all systems declines with non-speech interference, the baseline system, which relies on a pitch-based VAD, is more sensitive to acoustic conditions, particularly the type of noise or interference. This effect is even more pronounced in the cafeteria environment where, at 0 dB SNR, the accuracy of the proposed systems only declines by approximately 1% point, whereas the baseline system's accuracy decreases by approximately 20% points. These comparisons suggest that the proposed systems provide more robust DOA estimation performance across various acoustic conditions in single-talker scenarios.

Fig. A·3 depicts the speech detection performance measures, the precision and recall, for all three systems in the two reverberant environments. Both metrics have a possible range from 0

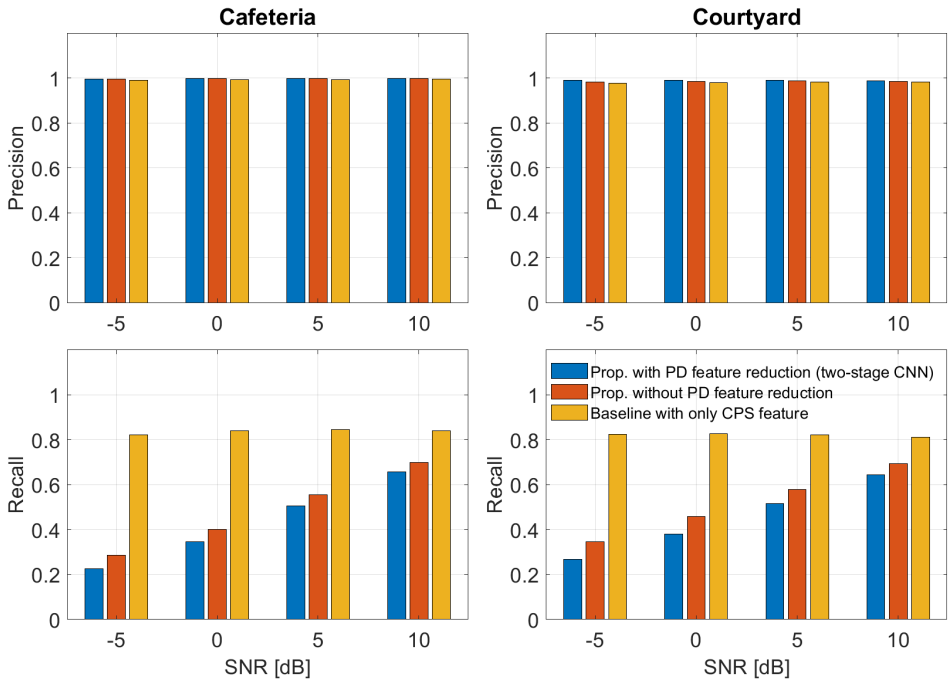


Figure A·3: The speech detection performance of the two proposed systems and the pitch-based VAD integrated into the baseline system from Chapter 3, evaluated in terms of precision (P) and recall (R) for single-talker scenarios with directional non-speech interference across various SNR conditions in two reverberant environments (cafeteria and courtyard).

to 1. It can be observed that all systems exhibit high precision close to 1. This means that there is a low likelihood of the systems falsely detecting portions of the signal for DOA estimation. When compared to the rVAD system, which is specifically designed for speech detection, the proposed systems demonstrate either better or comparable precision across all tested conditions. However, it is also observed that the proposed systems have a lower recall compared to rVAD. This means that while they are excellent at avoiding false positives, they may miss some true speech segments, leading to higher false negatives. This outcome can be attributed to the design focus of the proposed systems. Unlike the rVAD that is designed for the speech detection task, the proposed systems are designed primarily for speech DOA estimation. It should be noted that the proposed systems, including the detection class in the output, are not exclusively trained for speech detection, which explains the lower recall. In the context of DOA estimation, missing some speech segments is an acceptable trade-off if those segments do not contribute to accurate DOA estimation. These findings perfectly align with those in Section 2.6.2 in Chapter 2, further underscoring the robustness of the speech-aware DOA estimation approach across different network architectures and acoustic scenarios. Overall, the findings highlight the potential of the proposed systems for real-world applications where noise is not limited to background noise but includes various directional non-speech sounds.

References

[1] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, p. 298605, Jul. 2009.

[2] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[3] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[4] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[5] V. Hamacher, U. Kornagel, T. Lotter, and H. Puder, *Binaural Signal Processing in Hearing Aids: Technologies and Algorithms*. John Wiley and Sons, Ltd, 2008, ch. 14, pp. 401–429.

[6] T. Wittkop, S. Albani, V. Hohmann, J. Peissig, W. S. Woods, and B. Kollmeier, "Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 684–699, Jul. 1997.

[7] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 515–528, Dec. 2017.

[8] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz NY, USA, Oct. 2015, pp. 1–5.

[9] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, Jan. 2010.

[10] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[11] M. Farmani, M. S. Pedersen, Z. Tan, and J. Jensen, "Bias-compensated informed sound source localization using relative transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1275–1289, Jul. 2018.

[12] P. A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, Jul. 2022.

[13] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, Tokyo, Japan, Nov. 2013, pp. 2927–2932.

[14] N. Ma, T. May, and G. J. Brown, "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*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[15] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1335–1345, Aug. 2019.

[16] J. Wang, J. Wang, K. Qian, X. Xie, and J. Kuang, "Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–16, Feb. 2020.

[17] W. He, P. Motlicek, and J. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, May 2018, pp. 74–79.

[18] C. Pang, H. Liu, and X. Li, "Multitask learning of time-frequency CNN for sound source localization," *IEEE Access*, vol. 7, pp. 40 725–40 737, Mar. 2019.

[19] Z. Q. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 178–188, Jan. 2019.

[20] B. Yang, H. Liu, and X. Li, "Learning deep direct-path relative transfer function for binaural sound source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3491–3503, Oct. 2021.

[21] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, Apr. 2020.

[22] A. Küçük, A. Ganguly, Y. Hao, and I. M. S. Panahi, "Real-time convolutional neural network-based speech source localization on smartphone," *IEEE Access*, vol. 7, pp. 169 969–169 978, Nov. 2019.

[23] H. Hammer, S. E. Chazan, J. Goldberger, and S. Gannot, "Dynamically localizing multiple speakers based on the time-frequency domain," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–10, Apr. 2021.

[24] D. Salvati, C. Drioli, and G. L. Foresti, "Localization and tracking of an acoustic source using a diagonal unloading beamforming and a kalman filter," in *LOCATA Challenge Workshop - Satellite Event IWAENC*, Tokyo, Japan, Sep. 2018.

[25] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound.* MIT press, 1994.

[26] B. C. Moore, *An introduction to the psychology of hearing.* Brill, 2012.

[27] L. Chittka and A. Brockmann, "Perception space—the final frontier," *PLOS Biology*, vol. 3, no. 4, p. e137, Apr. 2005.

[28] J. Blauert and J. Braasch, *The Technology of Binaural Understanding.* Heidelberg, Germany: Springer-Verlag, 2020.

[29] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, Oct. 2003.

[30] L. Rayleigh, "On our perception of sound direction," *Philosophical Magazine*, vol. 13, no. 74, pp. 214–232, 1907.

[31] J. W. S. B. Rayleigh, *The theory of sound.* Macmillan, 1896.

[32] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, Mar. 1992.

[33] J. Blauert, *Spatial hearing: the psychophysics of human sound localization.* MIT press, 1997.

[34] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2219–2236, May 2002.

[35] C. Searle, L. Braida, D. Cuddy, and M. Davis, "Binaural pinna disparity: another auditory localization cue," *The Journal of the Acoustical Society of America*, vol. 57, no. 2, pp. 448–455, Feb. 1975.

[36] G. J. Brown and D. Wang, *Separation of Speech by Computational Auditory Scene Analysis.* Springer Berlin Heidelberg, 2005, pp. 371–402.

[37] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[38] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, P. Rice *et al.*, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[39] G. Hu and D. Wang, "An auditory scene analysis approach to monaural speech segregation," *Topics in acoustic echo and noise control*, pp. 485–515, 2006.

[40] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. CAZALS, K. HORNER, and L. DEMANY, Eds. Pergamon, Jun. 1992, pp. 429–446.

[41] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, 1993.

[42] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

[43] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, Nov. 2000.

[44] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Phoenix, AZ, USA, Mar. 1999, pp. 929–932 vol.2.

[45] J. Luberadzka, H. Kayser, and V. Hohmann, "Making sense of periodicity glimpses in a prediction-update-loop— a computational model of attentive voice tracking," *The Journal of the Acoustical Society of America*, vol. 151, no. 2, pp. 712–737, Feb. 2022.

[46] A. Josupeit, N. Kopčo, and V. Hohmann, "Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2911–2923, May. 2016.

[47] A. Josupeit and V. Hohmann, "Modeling speech localization, talker identification, and word recognition in a multi-talker setting," *The Journal of the Acoustical Society of America*, vol. 142, no. 1, pp. 35–54, Jul. 2017.

[48] Z. Chen and V. Hohmann, "Online monaural speech enhancement based on periodicity analysis and a priori SNR estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1904–1916, Nov. 2015.

[49] J. C. R. Licklider, "A duplex theory of pitch perception," *The Journal of the Acoustical Society of America*, vol. 23, pp. 147–147, Jan. 1951.

[50] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *The Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2866–2882, Jun. 1991.

[51] M. Slaney and R. F. Lyon, "A perceptual pitch detector," in *Proc. IEEE International conference on acoustics, speech, and signal processing (ICASSP)*, Albuquerque, NM, USA, Apr. 1990, pp. 357–360.

[52] R. Lyon, "Computational models of neural auditory processing," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 9, San Diego, CA, USA, Mar. 1984, pp. 41–44.

[53] J. Rouat, Y. C. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Communication*, vol. 21, no. 3, pp. 191–207, Apr. 1997.

[54] R. P. Carlyon and T. M. Shackleton, "Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms?" *The Journal of the Acoustical Society of America*, vol. 95, no. 6, pp. 3541–3554, Jun. 1994.

[55] M. Wu, D. Wang, and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Orlando, FL, USA, May 2002, pp. I–369–I–372.

[56] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[57] L. N. Tan and A. Alwan, "Multi-band summary correlogram-based pitch detection for noisy speech," *Speech communication*, vol. 55, no. 7-8, pp. 841–856, Sep. 2013.

[58] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, Mar. 2006.

[59] C. J. Darwin, "Listening to speech in the presence of other sounds," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1493, pp. 1011–1021, Mar. 2008.

[60] T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Communication*, vol. 27, no. 3, pp. 209–222, Apr. 1999.

[61] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using multipitch analysis and iterative parameter estimation," in *Proc. IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2001, pp. 83–86.

[62] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 255–266, Feb. 2008.

[63] C. J. Darwin and R. W. Hukin, "Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2316–2324, Oct. 1997.

[64] J. F. Culling and Q. Summerfield, "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," *The Journal of the Acoustical Society of America*, vol. 98, no. 2, pp. 785–797, Aug. 1995.

[65] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2914–2919, May. 1999.

[66] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "A speech fragment approach to localising multiple speakers in reverberant environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, May 2009, pp. 4593–4596.

[67] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503–1512, Jul. 2012.

[68] J. Blauert, *The technology of binaural listening.* Heidelberg, Germany: Springer-Verlag, 2013.

[69] I. Tashev, *Sound Capture and Processing: Practical Approaches.* Wiley, Jul. 2009.

[70] G. R. Popelka, B. C. Moore, R. R. Fay, and A. N. Popper, *Hearing aids.* Springer, 2016, vol. 56.

[71] G. J. B. R. Stern and D. Wang, *Binaural Sound Localization*. Wiley-IEEE Press, 2006, p. 147–185.

[72] D. Hammershøi and H. Møller, *Binaural technique—Basic methods for recording, synthesis, and reproduction*. Berlin, Heidelberg: Springer, 2005, pp. 223–254.

[73] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz NY, USA, Oct. 2001, pp. 99–102.

[74] H. Wierstorf, M. Geier, and S. Spors, "A free database of head related impulse response measurements in the horizontal plane with multiple distances," in *Proc. Audio Engineering Society Convention 130*. Audio Engineering Society, May 2011.

[75] W. G. Gardner and K. D. Martin, "HRTF measurements of a kemar," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, Jun. 1995.

[76] R. Bomhardt, M. de la Fuente Klein, and J. Fels, "A high-resolution head-related transfer function and three-dimensional ear model database," in *Proc. of Meetings on Acoustics*, vol. 29, no. 1. AIP Publishing, Jun. 2016.

[77] N. Madhu and R. Martin, *Acoustic Source Localization with Microphone Arrays*. John Wiley and Sons, Ltd, 2008, ch. 6, pp. 135–170.

[78] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science and Business Media, 2013.

[79] T. May, S. van de Par, and A. Kohlrausch, "Binaural localization and detection of speakers in complex acoustic scenes," in *The Technology of Binaural Listening*, J. Blauert, Ed. Heidelberg, Germany: Springer-Verlag, 2013, pp. 397–425.

[80] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[81] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Mar. 2008, pp. 2565–2568.

[82] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, vol. 53, no. 5, pp. 592–605, May–Jun. 2011.

[83] D. Fejgin and S. Doclo, "Assisted RTF-vector-based binaural direction of arrival estimation exploiting A calibrated external microphone array," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[84] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint itd/ild model and head movement tracking," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 430–434.

[85] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, Providence RI, USA, Aug. 2000.

[86] S. A. Vorobyov, A. B. Gershman, and K. M. Wong, "Maximum likelihood direction-of-arrival estimation in unknown noise fields using sparse sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 1, pp. 34–43, Jan. 2005.

[87] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[88] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband music: Opportunities and challenges for multiple source localization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2007, pp. 18–21.

[89] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[90] R. Takeda and K. Komatani, "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA, USA: IEEE, Apr. 2017, pp. 2217–2221.

[91] R. Takeda, Y. Kudo, K. Takashima, Y. Kitamura, and K. Komatani, "Unsupervised adaptation of neural networks for discriminative sound source localization with eliminative constraint," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB, Canada: IEEE, Apr. 2018, pp. 3514–3518.

[92] W. He, P. Motlicek, and J.-M. Odobez, "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, UK: IEEE, May 2019, pp. 770–774.

[93] R. Opochinsky, B. Laufer-Goldshtein, S. Gannot, and G. Chechik, "Deep ranking-based sound source localization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, Oct. 2019, pp. 283–287.

[94] M. J. Bianco, S. Gannot, E. Fernandez-Grande, and P. Gerstoft, "Semi-supervised source localization in reverberant environments with deep generative modeling," *IEEE Access*, vol. 9, pp. 84 956–84 970, Jun. 2021.

[95] W. He, P. Motlicek, and J.-M. Odobez, "Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1303–1317, Feb. 2021.

[96] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. European Signal Processing Conference (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 1462–1466.

[97] W. He, P. Motlicek, and J.-M. Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," in *Proc. Interspeech 2018*, Hyderabad, India, Sep. 2018, pp. 312–316.

[98] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, Mar. 2019.

[99] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 451–455.

[100] H. Sundar, W. Wang, M. Sun, and C. Wang, "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 4642–4646.

[101] J. Ding, Y. Ke, L. Cheng, C. Zheng, and X. Li, "Joint estimation of binaural distance and azimuth by exploiting deep neural networks," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2625–2635, Apr. 2020.

[102] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in CNN based multisource DOA estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1594–1608, Mar. 2021.

[103] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *Computer Speech and Language*, vol. 75, p. 101360, Sep. 2022.

[104] P. Goli and S. van de Par, "Deep learning-based speech specific source localization by using binaural and monaural microphone arrays in hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1652–1666, Apr. 2023.

[105] L. Wang, Z. Jiao, Q. Zhao, J. Zhu, and Y. Fu, "Framewise multiple sound source localization and counting using binaural spatial audio signals," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, May 2023, pp. 1–5.

[106] Q. Yang and Y. Zheng, "DeepEar: Sound localization with binaural microphones," *IEEE Transactions on Mobile Computing*, vol. 23, no. 1, pp. 359–375, Jan. 2024.

[107] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention*. Audio Engineering Society, May 2015.

[108] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, "On sound source localization of speech signals using deep neural networks," in *Proc. Deutsche Jahrestagung Akustik (DAGA)*, Mar. 2015, pp. 1510–1513.

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

[110] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 405–409.

[111] S. Chakrabarty and E. A. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, Oct. 2017, pp. 136–140.

[112] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[113] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based multiple DoA estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, Feb. 2019.

[114] W. Mack, U. Bharadwaj, S. Chakrabarty, and E. A. Habets, "Signal-aware broadband DOA estimation using attention mechanisms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 4930–4934.

[115] T. N. T. Nguyen, W.-S. Gan, R. Ranjan, and D. L. Jones, "Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2626–2637, Aug. 2020.

[116] W. Mack, J. Wechsler, and E. A. Habets, "Signal-aware direction-of-arrival estimation using attention mechanisms," *Computer Speech and Language*, vol. 75, p. 101363, Sep. 2022.

[117] Q. Hu, N. Ma, and G. J. Brown, "Robust binaural sound localisation with temporal attention," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5.

[118] E. Vargas, J. R. Hopgood, K. Brown, and K. Subr, "On improved training of CNN for acoustic source localisation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 720–732, Jan. 2021.

[119] L. Perotin, A. Défossez, E. Vincent, R. Serizel, and A. Guérin, "Regression versus classification for neural network based audio source localization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, Oct. 2019, pp. 343–347.

[120] Z. Tang, J. D. Kanu, K. Hogan, and D. Manocha, "Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks," in *Proc. Interspeech 2019*, Sep. 2019, pp. 654–658.

[121] P. Cooreman, A. Bohlender, and N. Madhu, "CRNN-based multi-DOA estimator: Comparing classification and regression," in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, Sep. 2023, pp. 156–160.

[122] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 300–311, Nov. 2020.

[123] F. B. Gelderblom, Y. Liu, J. Kvam, and T. A. Myrvoll, "Synthetic data for DNN-based DOA estimation of indoor speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 4390–4394.

[124] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[125] F. Chollet, *Deep learning with Python*. Simon and Schuster, 2021.

[126] S. J. Prince, *Understanding deep learning*. MIT press, 2023.

[127] E. Thuillier, H. Gamper, and I. J. Tashev, "Spatial audio feature discovery with convolutional neural networks," in *Proc. IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Calgary, AB, Canada: IEEE, Apr. 2018, pp. 6797–6801.

[128] W. Zhang, Y. Zhou, and Y. Qian, "Robust DOA estimation based on convolutional neural network and time-frequency masking," in *Proc. Interspeech 2019*, Sep. 2019, pp. 2703–2707.

[129] N. Liu, H. Chen, K. Songgong, and Y. Li, "Deep learning assisted sound source localization using two orthogonal first-order differential microphone arrays," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 1069–1084, Feb. 2021.

[130] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Direction of arrival estimation of sound sources using icosahedral CNNs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 313–321, Nov. 2022.

[131] Q. Li, X. Zhang, and H. Li, "Online direction of arrival estimation based on deep learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB, Canada: IEEE, Apr. 2018, pp. 2616–2620.

[132] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, Dec. 2020.

[133] L. Comanducci, M. Cobos, F. Antonacci, and A. Sarti, "Time difference of arrival estimation from frequency-sliding generalized cross-correlations using convolutional neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 4945–4949.

[134] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., Dec. 2017.

[135] C. Schymura, T. Ochiai, M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, and D. Kolossa, "Exploiting attention-based sequence-to-sequence architectures for sound event localization," in *Proc. European Signal Processing Conference (EUSIPCO)*. IEEE, Jan. 2021, pp. 231–235.

[136] C. Schymura, B. Bönninghoff, T. Ochiai, M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, and D. Kolossa, "PILOT: Introducing transformers for probabilistic sound event localization," in *Interspeech 2021*, Aug. 2021, pp. 2117–2121.

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

[138] Z. H. Tan, A. K. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Computer Speech and Language*, vol. 59, pp. 1–21, Jan. 2020.

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

[140] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in *Proc. Interspeech 2005*, Sep. 2005, pp. 369–372.

[141] J. M. Kates, "Classification of background noises for hearing-aid applications," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 461–470, Jan. 1995.

[142] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Communication*, vol. 52, no. 1, pp. 41–60, Jan. 2010.

[143] R. Varzandeh, K. Adiloğlu, S. Doclo, and V. Hohmann, "Exploiting periodicity features for joint detection and DOA estimation of speech sources using convolutional neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 566–570.

[144] D. Fejgin and S. Doclo, "Comparison of binaural RTF-vector-based direction of arrival estimation methods exploiting an external microphone," in *Proc. European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, Aug. 2021, pp. 241–245.

[145] G. Hu and D. Wang, "Segregation of unvoiced speech from nonspeech interference," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 1306–1319, Aug. 2008.

[146] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France, Sep. 2014, pp. 99–103.

[147] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 2386–2390.

[148] S. Popham, D. Boebinger, D. P. Ellis, H. Kawahara, and J. H. McDermott, "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," *Nature communications*, vol. 9, no. 1, pp. 1–13, Dec. 2018.

[149] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[150] Y. Wu and K. He, "Group normalization," in *Proc. of the European Conference on Computer Vision (ECCV)*, Munich, Germany, Sep. 2018.

[151] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proc. International Computer Music Conference (ICMC)*, Champaign/Urbana, IL, USA, Aug. 1987, pp. 290–297.

[152] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[153] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. ACM Conference on Multimedia*, Brisbane, Australia, Oct. 2015, pp. 1015–1018.

[154] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology*, vol. 40, no. 3, pp. 148–157, 2001.

[155] M. Park, "Models of binaural hearing for sound lateralisation and localisation," Ph.D. dissertation, University of Southampton, Oct. 2007.

[156] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, Vancouver, BC, Canada, Dec. 2019, pp. 8026–8037.

[157] D. K. P and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, May 2015.

[158] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.

[159] N. K. Desiraju, S. Doclo, and T. Wolff, "Efficient multichannel acoustic echo cancellation using constrained tap selection schemes in the subband domain," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, pp. 1–16, Sep. 2017.

[160] G. Grimm, H. Kayser, M. Hendrikse, and V. Hohmann, "A gaze-based attention model for spatially-aware hearing aids," in *Proc. ITG Conference on Speech Communication*, Oldenburg, Germany, Oct. 2018, pp. 1–5.

[161] A. Aroudi and S. Doclo, "Cognitive-driven binaural beamforming using EEG-based auditory attention decoding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 862–875, Jan. 2020.

[162] D. Fejgin and S. Doclo, "Coherence-based frequency subset selection for binaural RTF-vector-based direction of arrival estimation for multiple speakers," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sep. 2022.

[163] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[164] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[165] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[166] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Feb. 2015.

[167] D. A. Krause, G. García-Barrios, A. Politis, and A. Mesaros, "Binaural sound source distance estimation and localization for a moving listener," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 996–1011, Dec. 2024.

[168] R. Varzandeh, , S. Doclo, and V. Hohmann, "A two-stage CNN with feature reduction for speech-aware binaural DOA estimation," in *Proc. European Signal Processing Conference (EUSIPCO)*, Helsinki, Finland, Sep. 2023, pp. 241–245.

[169] ——, "Speech-aware binaural DOA estimation utilizing periodicity and spatial features in convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1198–1213, Jan. 2024.

List of Publications

Peer-reviewed Journal Papers

[J2] R. Varzandeh, S. Doclo, V. Hohmann, "Improving multi-talker binaural DOA estimation by combining periodicity and spatial features in convolutional neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, 2025, 5 (2025).

[J1] R. Varzandeh, S. Doclo, V. Hohmann, "Speech-aware binaural DOA estimation utilizing periodicity and spatial features in convolutional Neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1198-1213, 2024.

Peer-reviewed Conference Papers

[C2] R. Varzandeh, S. Doclo, V. Hohmann, "A two-stage CNN with feature reduction for speech-aware binaural DOA estimation," in Proc. *European Signal Processing Conference (EUSIPCO)*, Helsinki, Finland, 2023, pp. 241-245.

[C1] R. Varzandeh, K. Adiloğlu, S. Doclo, V. Hohmann, "Exploiting periodicity features for joint detection and DOA estimation of speech sources using convolutional neural networks," in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 566-570.