CHALMERS
UNIVERSITY OF TECHNOLOGY

Data-Driven Speech Recovery in a Fiber-Optic Polarization-Based Sensing System

Master's thesis in Electrical Engineering

Jun Cao

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's Thesis 2026

Data-Driven Speech Recovery in a Fiber-Optic Polarization-Based Sensing System

Jun Cao

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Communication Systems Group
Chalmers University of Technology
Gothenburg, Sweden 2026

Data-Driven Speech Recovery in a Fiber-Optic Polarization-Based Sensing System
Jun Cao

Supervisor: Zicong Jiang, Department of Electrical Engineering
Examiner: Christian Häger, Department of Electrical Engineering

Master's Thesis 2026
Department of Electrical Engineering
Communication Systems Group
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Data-Driven Speech Recovery in a Fiber-Optic Polarization-Based Sensing System
Jun Cao
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Optical fibers are inherently sensitive to external acoustic vibrations, which can modulate the local birefringence via the elasto-optic effect, imposing perturbations onto the state of polarization (SOP) of the transmitted light. This creates an unintended sensing channel: for example, speech spoken near the fiber may leak into the SOP trajectory and can potentially be recovered by an eavesdropper. This thesis develops, analyzes, and validates a speech recovery framework that operates directly on SOP obtained from the output of a fiber link.

A waveplate fiber channel model is adapted that incorporates the effect of speech on the fiber. Building on this model, a three-stage reproducible speech recovery pipeline is proposed, consisting of preprocessing, demodulation, and enhancement. While confirming their effectiveness, the simulation results show that different demodulation methods give comparable performance, indicating that the primary bottleneck does not lie in the choice of these methods.

Building on this insight, hardware experiments are conducted in an optical fiber laboratory using a kilometer-scale single-mode fiber spool as the acoustic sensor. The same pipeline framework used in the simulation study is applied. To further improve the performance, a data-driven speech enhancement method based on a convolutional neural network (CNN) is explored using experimental data, achieving a substantial improvement in perceptual speech quality while preserving intelligibility. Both simulation and experimental results provide consistent support for the fiber channel model, while the experiments further reveal practical performance limitations.

Keywords: fiber sensing, state of polarization, speech enhancement, convolutional neural network.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Zicong Jiang for his patient guidance, constructive feedback, and encouragement throughout this project. His insights have been invaluable in shaping both the research direction and the quality of this work.
I am also grateful to my examiner Christian Häger for his thoughtful comments and academic support. Special thanks go to Sam William O'Brien, Erik Börjeson, and Magnus Karlsson for their generous assistance with the optical fiber laboratory experiments and for many stimulating discussions. I would also like to thank Rick Maarten Butler for his professional advice on thesis writing and theory formalization.
I extend my appreciation to all members of the Communication Systems Group at Chalmers University of Technology for creating an open and inspiring research environment.
Finally, I am deeply thankful to my family and friends—especially Yu Xia and Zhou Shan—for their unwavering encouragement and warm support throughout my studies.

Jun Cao, Gothenburg, June 2026

List of Acronyms

Below is the list of acronyms used throughout this thesis, listed in alphabetical order.

ASE	Amplified Spontaneous Emission
BN	Batch Normalisation
BPF	Bandpass Filter
CNN	Convolutional Neural Network
DAS	Distributed Acoustic Sensing
DNN	Deep Neural Network
DSP	Digital Signal Processing
FPGA	Field-Programmable Gate Array
ICA	Independent Component Analysis
ISTFT	Inverse Short-Time Fourier Transform
PESQ	Perceptual Evaluation of Speech Quality
PMD	Polarization Mode Dispersion
PSD	Power Spectral Density
ReLU	Rectified Linear Unit
SI-SDR	Scale-Invariant Signal-to-Distortion Ratio
SMF	Single-Mode Fiber
SNR	Signal-to-Noise Ratio
SOP	State of Polarization
Spec-Sub	Spectral Subtraction
STFT	Short-Time Fourier Transform
STOI	Short-Time Objective Intelligibility
SVD	Singular Value Decomposition
T-F	Time-Frequency
WER	Word Error Rate

Mathematical Notation and Definitions

Bold lowercase letters denote column vectors; bold uppercase letters denote matrices. $(\cdot)^*$ is complex conjugate, $(\cdot)^\top$ transpose, and $(\cdot)^\dagger$ conjugate transpose. For discrete-time quantities, scalars are written with a square-bracket argument $x[k]$, while vectors and matrices carry a subscript $\mathbf{S}_k$, $\mathbf{H}_k$. A time series of vectors is collected into a matrix by stacking them as columns; e.g., $\mathbf{S} = [\mathbf{S}_0, \ldots, \mathbf{S}_{K-1}] \in \mathbb{R}^{3\times K}$.

Symbol	**Description**	**Unit / Type**
Mathematical notation		
i	Imaginary unit, $i^2 = -1$	—
$\mathrm{Re}[\cdot]$	Real part of a complex number	—
$\mathrm{Im}[\cdot]$	Imaginary part of a complex number	—
$\{\cdot\}$	Sequence	—
Scalars		
t	Time	s
k	Discrete time sample index, $k = 0, \ldots, K-1$	—
K	Total number of time samples	—
z	Axial position along the fiber	m
L	Total fiber length	m
N	Number of fiber sections	—
Δz	Length of one fiber section, $\Delta z = L/N$	m
ℓ	Fiber section index, $\ell = 1, \ldots, N$	—
$\mathcal{L}$	Index set of speech-perturbed sections	—
$x(t)$	Continuous-time speech signal	—
f_s	Speech sampling rate	Hz
$x[k]$	Discrete speech sample, $x[k] \triangleq x(k/f_s)$	—

(continued)

Symbol	Description	Unit / Type
$\mathbf{x}$	Speech waveform vector, $\mathbf{x} = [x[0], \ldots, x[K-1]]^\top$	$\mathbb{R}^K$
$\hat{\mathbf{x}}$	Recovered speech waveform vector	$\mathbb{R}^K$
ξ	Speech-to-phase sensitivity coefficient	rad
θ_ℓ	Rotation angle of section ℓ	rad
$\Delta\phi_\ell$	Differential phase of section ℓ	rad
b	Birefringence parameter	m^{-1}
L_{beat}	Beat length	m
α_ℓ	Fast-axis orientation of section ℓ	rad
L_{corr}	Correlation length	m
σ_{d}	Std. of axis-angle increment	rad
σ_n	Std. of the ASE noise per Jones component	—
S_0, S_1, S_2, S_3	Stokes parameters	—
Vectors		
$\mathbf{E}, \mathbf{u}, \mathbf{r}$	Jones vectors	$\mathbb{C}^2$
$\mathbf{n}$	ASE noise	$\mathbb{C}^2$
$\mathbf{S}_k = [S_1, S_2, S_3]^\top$	Normalized Stokes vector at index k	$\mathbb{R}^3$
$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$	Vector of standard Pauli matrices	$(\mathbb{C}^{2\times2})^3$
Matrices		
$\mathbf{S}$	Stokes trajectory matrix, $\mathbf{S} = [\mathbf{S}_0, \ldots, \mathbf{S}_{K-1}]$	$\mathbb{R}^{3\times K}$
$\tilde{\mathbf{S}}$	Preprocessed Stokes trajectory matrix	$\mathbb{R}^{3\times K}$
$\mathbf{J}_\ell$	Jones matrix of fiber section ℓ	SU(2)
$\mathbf{H}$	Total fiber channel matrix	SU(2)
$\mathbf{R}(\theta)$	Rotation matrix, $\left[\begin{smallmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{smallmatrix}\right]$	SO(2)
$\sigma_1, \sigma_2, \sigma_3$	Pauli matrices: $\sigma_1 = \left[\begin{smallmatrix}0 & 1 \\ 1 & 0\end{smallmatrix}\right]$, $\sigma_2 = \left[\begin{smallmatrix}0 & -i \\ i & 0\end{smallmatrix}\right]$, $\sigma_3 = \mathrm{diag}(1, -1)$	$\mathbb{C}^{2\times2}$
Σ_ℓ	Birefringence orientation matrix at section ℓ	$\mathbb{C}^{2\times2}$
Sets and spaces		
$\mathcal{S}^2$	Poincaré sphere (unit sphere in $\mathbb{R}^3$)	manifold

(continued)

Symbol	Description	Unit / Type
SU(2)	Special unitary group (2×2 unitary, $\det = 1$)	group
SO(3)	Special orthogonal group (rotations of $\mathbb{R}^3$)	group

Contents

List of Acronyms **ix**

Mathematical Notation and Definitions **xi**

List of Figures **xvii**

List of Tables **xix**

1 Introduction **1**
1.1 Background . . . 1
1.2 Problem Statement . . . 2
1.3 Related Work . . . 2
1.4 Contributions . . . 4
1.5 Thesis Outline . . . 4

2 System Model **5**
2.1 Jones Vectors and Jones Matrices . . . 5
2.2 Stokes Parameters and the Poincaré Sphere . . . 6
2.3 System Overview . . . 7
2.4 Perturbation Model . . . 8

3 Methods **15**
3.1 Speech Recovery Pipeline . . . 15
3.2 Performance Metrics . . . 18
3.3 Data-driven Speech Enhancement . . . 19

4 Results **23**
4.1 Simulation . . . 23
4.1.1 Simulated SOP . . . 23
4.1.2 Monte Carlo Results . . . 24
4.2 Experiment . . . 24
4.2.1 Experimental Setup . . . 25
4.2.2 Experimental SOP . . . 26
4.2.3 Without Data-Driven Enhancement . . . 26
4.2.4 With Data-driven Enhancement . . . 28

5 Discussion and Conclusion **31**

5.1 Simulation vs. Experiment 31
5.1.1 Fiber Sensitivity 31
5.1.2 Laboratory Noise 31
5.2 Limitations 31
5.3 Conclusion 33

Bibliography **35**

List of Figures

1.1 Schematic of the unintended information leakage in a fiber-optic communication system. 1

2.1 Poincaré sphere. $\mathbf{S}_k$ is represented as a point on the unit sphere $\mathcal{S}^2$. . 7
2.2 System framework of this thesis. 7
2.3 Signal chain from speech sample x to the observable instantaneous SOP $\mathbf{S}$. 8
2.4 Waveplate fiber model Equation (2.12): N equal-length sections and local fast-axis orientations α_ℓ. Shading marks the speech-perturbed interaction window $\mathcal{L}$. 9
2.5 Example realization of the cumulative angle α_ℓ from Equation (2.11) (discrete Wiener along sections). Here the angle is plotted without folding onto $[0, \pi)$ for visualization purposes. 10
2.6 Example realization of the time-varying interaction window offset w_k defined in Equation (2.18). 12
2.7 Speech-induced SOP perturbation on the Poincaré sphere. The reference vector is the unperturbed SOP and the perturbed vector is the SOP after a single section in the interaction window applies its speech-modulated rotation. The arc indicates the half-angle $\Delta\phi_\ell/2$ that enters the Jones-space matrix in Equation (2.19); on the Stokes sphere this corresponds to a rotation by the full angle $\Delta\phi_\ell$ about the section axis $\hat{n}_\ell$. 12

3.1 Speech recovery pipeline. $\mathbf{S}$ is the Stokes trajectory and $\tilde{\mathbf{S}}$ is the unit-normalized Stokes trajectory. $\mathbf{y}$ is the demodulated recovered speech signal (usually noisy) and $\hat{\mathbf{x}}$ is the enhanced speech signal (final output). 15
3.2 DNN-based speech enhancement. STFT and ISTFT are defined in Equation (3.15) and Equation (3.17). The spectrogram is defined in Equation (3.16). 19
3.3 The CNN architecture. The dashed box (Layer 3) indicates a dilated convolution. Key parameters are shown in Table 3.1. 20

4.1 Simulated SOP example for a single fiber realization. 23
4.2 The experimental setup for the fiber sensing system. 25
4.3 Experimental SOP example corresponding to Figure 4.2. 26

4.4 Power spectral density (PSD) of the Stokes parameter S_1 when the speech is absent: (a) simulated Gaussian Jones noise propagated to S_1 and (b) measured laboratory background. 27
4.5 Waveform comparison between the original speech and the baseline recovered speech. 28
4.6 Experimental recovery chain Section 4.2.4 using the signal notation of Figure 3.1. Demodulation is fixed to baseline Section 3.1; enhancement replaces the Butterworth bandpass Equation (3.3) by CNN Section 3.3. 28
4.7 Training and validation loss curves of the CNN-based speech enhancement model. 29
4.8 Spectrogram comparison for a representative test recording. 30

5.1 Target full-system architecture: SOP is derived from the equalizer weight matrix $\mathbf{W}_k$ of a coherent receiver, requiring no additional sensing hardware. 32

List of Tables

3.1 CNN layer parameters (frequency × time). 21

4.1 Monte Carlo metrics (mean ± std, 50 trials). Mean scores average over 10 VCTK speeches (five male, five female) per trial. 24
4.2 Experimental performance averaged over 25 testing cases (mean ± std). 28
4.3 Speech enhancement performance. 30

1
Introduction

1.1 Background

Optical fiber technology was developed for high-capacity communication systems, offering low loss, wide bandwidth, and immunity to electromagnetic interference [1]. Beyond data transmission, optical fibers are inherently sensitive to external perturbations such as temperature variations and acoustic vibrations, enabling a wide range of sensing applications [2, 3, 4, 5]. Every deployed fiber link is therefore both a communication channel and a distributed sensor [6]. This dual-use nature raises a problem as shown in Figure 1.1: acoustic signals near a fiber can leak into the transmitted signals, creating an unintended channel for eavesdropping [7].

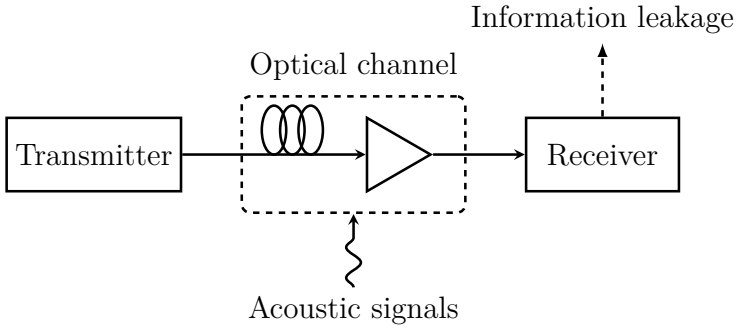


Figure 1.1: Schematic of the unintended information leakage in a fiber-optic communication system.

Human speech is a representative target, since it often carries sensitive private information. Acoustic perturbations from human speech induce mechanical strain in the optical fiber, which modifies the fibers refractive index, leading to corresponding changes in fiber birefringence via the photoelastic effect, thereby modulating the phase of the transmitted optical carrier [8]. The feasibility of speech recovery from the resulting phase shifts has been demonstrated [9, 10, 11]. Conventional distributed acoustic sensing (DAS) approaches [12] rely on dedicated hardware, which is costly to deploy within existing telecommunication infrastructure. Polarization-based sensing exploits variations in the state of polarization (SOP) of transmitted light as a signature for external perturbations. As SOP is already estimated in coherent transceivers, this approach requires no additional hardware. Kouiani et al. [13] demonstrated speech recovery from SOP measurements under controlled laboratory conditions, which serves as the most closely related prior work to this study.

The recovered acoustic signal is inherently degraded by noise such as amplified spontaneous emission (ASE) from amplifiers, laser phase noise and environmental acoustic noise [14]. The noise can be strongly coloured and spectrally overlaps the speech band, and therefore cannot be removed by simple linear filtering. Data-driven speech enhancement approaches (particularly neural network-based) have shown considerable promise in mitigating such degradation in fiber acoustic sensing systems. [15, 16, 17, 18], yet their application to polarization-based speech recovery remains unexplored. This gap leaves the true severity of the eavesdropping threat unestablished. The ultimate goal is to prevent such unintended information leakage; however, designing reliable countermeasures [19] first requires a thorough understanding of the attack itself—its physical mechanisms, recovery pipeline design, and achievable performance limits. This thesis focuses on exploring the feasibility of polarization-based speech recovery, rather than on countermeasure design.

1.2 Problem Statement

Speech signals can modulate the SOP at the output of an optical fiber link. Accurate recovery is challenging, as the SOP may encode speech-induced perturbations only implicitly through its time-varying trajectory, the signal-to-noise ratio is low, and no established processing pipeline or evaluation framework exists for this task.

> *To what extent can the speech signal be reconstructed from the SOP trajectory with acceptable quality and intelligibility?*

Concretely, three challenges must be resolved:

1. How can SOP dynamics be linked to speech-induced perturbations?
2. How can a practical speech recovery pipeline be developed and validated?
3. How can speech recovery performance be objectively evaluated and improved?

1.3 Related Work

DAS has emerged as one of the most widely deployed techniques for characterizing and exploiting acoustic and vibrational perturbations along fiber-optic infrastructure. Owing to its distributed nature, high sensitivity and spatial resolution, DAS has been investigated for a wide range of applications in seismic monitoring [3, 20], traffic flow detection [21], human activity identification [22], and pipeline leakage detection [23]. DAS was also shown to be effective for human voice recognition in [9] using spiral-shaped optical fiber coils, which served as an optical microphone. A comprehensive state-of-the-art review of DAS applications is provided in [24]. Beyond dedicated deployments, there is growing interest in integrating distributed fiber sensing with existing telecommunication systems [25, 26, 27]. However, conventional backscatter-based DAS requires dedicated hardware and faces fundamental limitations in sensing range, as the inherently weak Rayleigh backscatter is further degraded by fiber attenuation and ASE noise introduced by amplifiers in deployed links [24]. These constraints motivate the search for alternative sensing approaches.

To overcome the limitations of backscatter-based DAS, forward-transmission sensing approaches have been explored [4, 28, 29], where sensing is performed directly on the communication signal propagating along the fiber, offering a cost-effective and deployment-friendly solution. Such work tracks different physical observables—notably optical phase fluctuations and changes in the SOP—each offering sensitivity to external perturbations depending on the dominant coupling mechanism of the disturbance.

Among these, polarization-based sensing is particularly suitable to coherent communication systems: the SOP can be directly extracted from the Jones matrix [30] already estimated during polarization demultiplexing, using the equalizer coefficients [31]requiring no dedicated sensing hardware. From a practical implementation perspective, Zhan et al. [4] showed that the SOP can serve as the sensing signal for seismic wave and ocean wave detection on a 10,000-km submarine cable. To ground this sensing mechanism in theory, Mecozzi et al. [32] modeled how physical perturbations alter the local birefringence along the fiber and showed that the SOP deviation spectrum could reproduce the spectrum of the acoustic source, provided the mechanical coupling between the source and the fiber is linear. On the terrestrial side, polarization-based sensing was validated on a deployed fiber for event detection [33, 29], revealing the potential for network-wide health monitoring [34]. Real-time SOP monitoring for cable-break detection in live terrestrial optical networks was achieved using an FPGA-based coherent transceiver [35].

Of particular relevance to this thesis is the recovery of human speech signals directly from the SOP. Kouiani et al. [13] demonstrated that speech can be detected and recovered from SOP measurements using a polarimeter in a controlled laboratory environment. However, the recovery accuracy appears to be limited, possibly due to pipeline design and experimental setup limitations—a shortcoming reflected in the reliance on correlation metrics rather than comprehensive speech fidelity measures.

Data-driven methods have recently been applied to improve speech quality in fiber acoustic sensing systems. Shang et al. [18] proposed a complex convolution recurrent network operating on complex spectral mappings as a post-processing stage for speech recovery in a DAS system. Chai et al. [15] applied a complex-valued convolutional neural network & long short-term memory model to speech enhancement in an extrinsic Fabry–Perot interferometric fiber acoustic sensor, demonstrating that the deep learning post-processing paradigm generalizes across different fiber sensing techniques. Self-supervised and noise-to-noise learning strategies have been explored for general DAS signal denoising in geophysical applications, removing the need for many paired clean-noisy training data [17, 16]. Taken together, these results show that data-driven speech enhancement consistently improves speech quality across different fiber sensing systems—motivating its application to our polarization-based case.

While polarization-based sensing has been explored as a hardware-free complement to DAS for certain fiber-optic acoustic sensing tasks, its application to speech recovery remains largely unexplored: existing work [13] demonstrates experimental feasibility but provides neither a theoretical model nor a software simulation to guide the pipeline framework, and evaluates performance primarily through correlation metrics rather than perceptual quality measures. Furthermore, the data-driven

speech enhancement techniques proven effective in DAS and other fiber sensing systems have not yet been applied to polarization-based speech recovery.

1.4 Contributions

The main contributions of this thesis are as follows:

- We propose a reproducible speech recovery pipeline with SOP as the input and speech as the output.
- We develop a simulation as a design sandbox to explore and motivate pipeline architecture choices prior to hardware experiments.
- We conduct hardware experiments to validate the feasibility of the proposed pipeline under a controlled lab environment.
- We evaluate speech recovery performance using comprehensive perceptual quality metrics.
- We propose a data-driven speech enhancement method to improve the recovery performance, and validate its effectiveness using experimental data.

1.5 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 presents the theoretical foundation: Jones and Stokes formalisms, the system model, and the perturbation model linking speech to SOP rotations on the Poincaré sphere. Chapter 3 describes the three speech recovery pipelines, the performance metrics, and the CNN-based speech enhancement architecture. Chapter 4 reports the Monte Carlo simulation study and the hardware experiment results, including the effect of data-driven enhancement. Chapter 5 discusses the findings, states the limitations, and draws conclusions.

2

System Model

This chapter develops the theoretical foundation for the thesis. The central object is the SOP of light propagating through a single-mode fiber (SMF): a point on the Poincaré sphere whose trajectory encodes the speech perturbation applied to the fiber.

2.1 Jones Vectors and Jones Matrices

Light is a transverse electromagnetic wave. When the light is constrained to propagate along the z-axis, x and y electric field components represent the polarization state of this z-propagating light [36]. The Jones vector [37] is defined for fully polarized light as a two-component complex vector

$$\mathbf{E} = \begin{bmatrix} E_x \\ E_y \end{bmatrix} \in \mathbb{C}^2, \tag{2.1}$$

where E_x and E_y are the complex amplitudes of the x- and y-polarized electric field components, respectively. Throughout this thesis, Jones vectors are denoted by bold lowercase letters; in particular, $\mathbf{u}$ denotes the transmitted (input) Jones vector and $\mathbf{r}$ the received (output) Jones vector.
The Jones matrix $\mathbf{J} \in \mathbb{C}^{2\times 2}$ describes the behavior of a linear optical element, relating an input Jones vector $\mathbf{u} = [u_x, u_y]^\top$ to an output Jones vector $\mathbf{r} = [r_x, r_y]^\top$ by

$$\mathbf{r} = \mathbf{J}\,\mathbf{u} = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix} \begin{bmatrix} u_x \\ u_y \end{bmatrix}. \tag{2.2}$$

For a lossless optical element, by convention, $\mathbf{J}$ belongs to the special unitary group $\mathrm{SU}(2) = \{\mathbf{J} \in \mathbb{C}^{2\times 2} : \mathbf{J}^\dagger\mathbf{J} = \mathbf{I}, \det \mathbf{J} = 1\}$. For a cascade of N such elements, the total Jones matrix is the ordered product[1]

$$\mathbf{J} = \mathbf{J}_N\mathbf{J}_{N-1}\cdots\mathbf{J}_1, \tag{2.3}$$

which remains in SU(2) by closure of the group under matrix multiplication.

[1]The matrices are ordered right-to-left: $\mathbf{J}_1$ is the first element encountered by the propagating light and therefore acts on $\mathbf{u}$ first.

2.2 Stokes Parameters and the Poincaré Sphere

The SOP of a light wave is described by four real-valued measurable quantities S_0, S_1, S_2, S_3 known as the *Stokes parameters* [30]. In terms of the electric field components E_x and E_y from Section 2.1, they are defined as

$$\begin{aligned} S_0 &= |E_x|^2 + |E_y|^2, && \text{(total optical intensity)} \\ S_1 &= |E_x|^2 - |E_y|^2, && \text{(horizontal vs. vertical linear polarization)} \\ S_2 &= 2\,\mathrm{Re}[E_x E_y^*], && (+45^\circ \text{ vs. } -45^\circ \text{ linear polarization}) \\ S_3 &= 2\,\mathrm{Im}[E_x E_y^*]. && \text{(right- vs. left-circular polarization)} \end{aligned}$$

For *fully polarized* light—the case throughout this thesis—the four parameters satisfy

$$S_1^2 + S_2^2 + S_3^2 = S_0^2. \tag{2.4}$$

Dividing S_1, S_2, S_3 by S_0 normalizes the representation so that $S_0 \equiv 1$. The normalized Stokes parameters are then computed from the observed Jones vector $\mathbf{r} = [r_x,\, r_y]^\top$ as [30]

$$S_1 = \frac{|r_x|^2 - |r_y|^2}{|r_x|^2 + |r_y|^2}, \tag{2.5}$$

$$S_2 = \frac{2\,\mathrm{Re}[r_x r_y^*]}{|r_x|^2 + |r_y|^2}, \tag{2.6}$$

$$S_3 = \frac{2\,\mathrm{Im}[r_x r_y^*]}{|r_x|^2 + |r_y|^2}. \tag{2.7}$$

Since $S_0 = 1$ after normalization, the SOP is fully characterized by the remaining three parameters. Throughout this thesis, all Stokes parameters are understood in this normalized sense.

In practice the SOP is sampled at K discrete time steps indexed by $k = 0, 1, \ldots, K-1$. At each index k, the three normalized Stokes parameters $S_1[k], S_2[k], S_3[k]$ are computed via Equation (2.5)–Equation (2.7) from a Jones vector $\mathbf{r}_k = [r_x[k],\, r_y[k]]^\top \in \mathbb{C}^2$. Here the square-bracket argument $[\,k\,]$ denotes the value of a scalar sequence at time index k. Collecting the three values into a column vector defines the *normalized Stokes vector* at index k:

$$\mathbf{S}_k = [S_1[k],\, S_2[k],\, S_3[k]]^\top \in \mathbb{R}^3. \tag{2.8}$$

By Equation (2.4), normalization forces $\|\mathbf{S}_k\| = 1$, so $\mathbf{S}_k$ lies on the unit sphere $\mathcal{S}^2 \subset \mathbb{R}^3$, known as the *Poincaré sphere* (Figure 2.1). Each point on $\mathcal{S}^2$ corresponds to a unique SOP: the north and south poles represent right- and left-circular polarization, respectively; the equator represents all linear polarizations; and intermediate latitudes correspond to elliptical polarizations [36].

Stacking the Stokes vectors column-wise defines the *Stokes trajectory matrix* in this thesis:

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}_1^\top \\ \mathbf{s}_2^\top \\ \mathbf{s}_3^\top \end{bmatrix} = \begin{bmatrix} S_1[0] & S_1[1] & \cdots & S_1[K-1] \\ S_2[0] & S_2[1] & \cdots & S_2[K-1] \\ S_3[0] & S_3[1] & \cdots & S_3[K-1] \end{bmatrix} \in \mathbb{R}^{3\times K}, \tag{2.9}$$

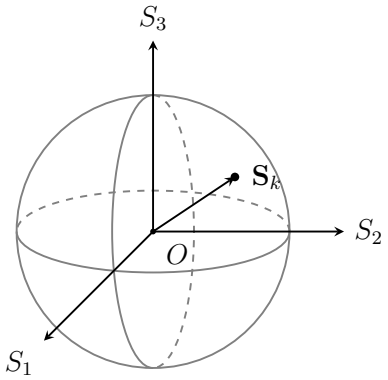


Figure 2.1: Poincaré sphere. $\mathbf{S}_k$ is represented as a point on the unit sphere $\mathcal{S}^2$.

where row i is the time series of the i-th Stokes component $\mathbf{s}_i = [S_i[0], \ldots, S_i[K-1]]^\top \in \mathbb{R}^K$, and column k is the instantaneous Stokes vector $\mathbf{S}_k = [S_1[k], S_2[k], S_3[k]]^\top \in \mathbb{R}^3$.

2.3 System Overview

The system studied in this thesis is summarized in Figure 2.2. A continuous-wave (CW) optical probe $\mathbf{u} = [1,\, 0]^\top$ (fixed x-polarized input) is launched into the fiber. The fiber channel $\mathbf{H}_k$ is affected by the speech signal (see Section 2.4), producing a time-varying received Jones vector

$$\mathbf{r}_k = \mathbf{H}_k\, \mathbf{u} + \mathbf{n}_k, \qquad \mathbf{n}_k \sim \mathcal{CN}(\mathbf{0},\, \sigma_n^2 \mathbf{I}_2), \tag{2.10}$$

where $\mathbf{n}_k \in \mathbb{C}^2$ is circularly symmetric complex Gaussian noise modeling the ASE, with variance σ_n^2 per Jones component and $\mathbf{I}_2$ the 2×2 identity matrix. Equation (2.10) describes the instantaneous snapshot at time index k. Over a recording window of K steps, the SOP computation block evaluates Equation (2.5)–Equation (2.7) at each k and stacks the results column-wise to form the Stokes trajectory matrix $\mathbf{S} \in \mathbb{R}^{3\times K}$ defined in Equation (2.9). The speech recovery pipeline then maps the full matrix $\mathbf{S}$ to an estimated waveform $\hat{\mathbf{x}} \in \mathbb{R}^K$, described in detail in Chapter 3.

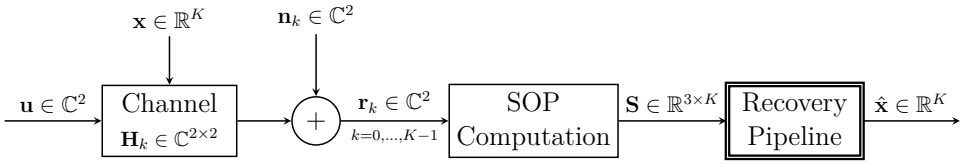


Figure 2.2: System framework of this thesis.

Remark. The broader motivation for this line of research is that a deployed coherent transceiver already estimates the SOP as a byproduct of polarization demultiplexing, making this sensing as a free ability. Integrating the recovery pipeline with a real coherent receiver—where SOP is derived from the equalizer weight matrix

rather than a dedicated polarimeter [33]—is left as future work and discussed in Chapter 5. The polarimeter-based setup adopted here isolates and focuses on the speech recovery pipeline itself, enabling a clean evaluation of the signal-processing design independently of transceiver hardware constraints.

2.4 Perturbation Model

This section develops the perturbation model that maps a single speech sample $x \in \mathbb{R}$ to an instantaneous SOP vector $\mathbf{S} \in \mathbb{R}^3$ at the fiber output[2]. The fiber is treated as a lossless single-mode fiber (SMF) supporting two orthogonal polarization modes; in a real fiber, these two modes propagate with slightly different effective refractive indices because of residual core asymmetry, bending, and mechanical stress—a property known as *birefringence* [30]. The Jones-vector and Stokes formalisms of Sections 2.1 and 2.2 apply in this regime.
The chain is summarized in Figure 2.3: (i) speech induces a small differential phase shift $\Delta\phi_\ell$ in selected fiber sections; (ii) this enters each section's Jones matrix $\mathbf{J}_\ell$; (iii) the per-section matrices cascade into the total channel matrix that maps the input probe $\mathbf{u}$ to a noisy received Jones vector $\mathbf{r}$; and (iv) $\mathbf{r}$ is converted into the observable SOP $\mathbf{S}$ via Equation (2.5)–Equation (2.7).

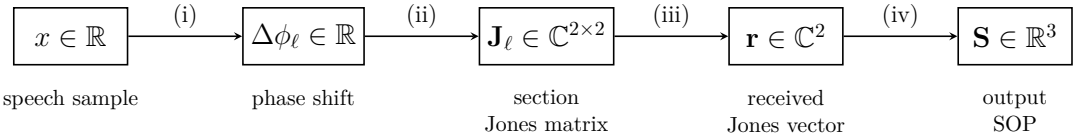


Figure 2.3: Signal chain from speech sample x to the observable instantaneous SOP $\mathbf{S}$.

The remainder of this section is organised in the natural construction order of the model: first the split-step decomposition of the fiber (Section 2.4.1); then the per-section Jones matrix in the absence of speech (Section 2.4.2); the speech-induced birefringence perturbation (Section 2.4.3); how the perturbation enters the per-section matrix (Section 2.4.4); the cascaded channel and received signal (Section 2.4.5); and finally a frozen-axis property of the resulting SOP trajectory that is exploited later in this thesis (Section 2.4.6).

2.4.1 Split-Step Fiber Model

The fiber of total length L is divided into N contiguous sections of equal length $\Delta z = L/N$,[3] indexed by $\ell = 1, \ldots, N$ (Figure 2.4). Section ℓ is treated as a uniform linear waveplate whose fast birefringence axis—the transverse eigendirection

[2]Only within this section, $\mathbf{S}$ denotes the instantaneous Stokes vector at a single time step; elsewhere it denotes the trajectory matrix $\mathbf{S} \in \mathbb{R}^{3\times K}$ of Equation (2.9). The discrete time index k is likewise suppressed throughout this section, except where time-varying quantities (such as $\mathcal{L}_k$) are explicitly introduced.

[3]Notation note: in the original split-step scheme of [38], two distinct step sizes appear—a macroscopic outer step Δz in his eq. (21), used to alternate a chromatic-dispersion factor with a polarization transfer matrix $\mathbf{M}(\omega)$, and a microscopic inner step δz in his eqs. (24)–(25), used

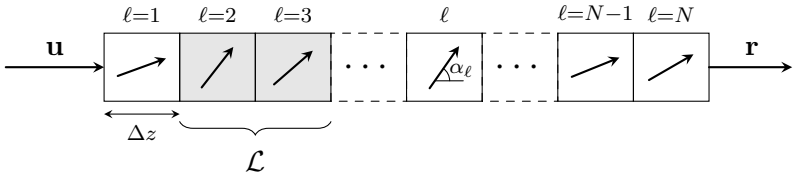


Figure 2.4: Waveplate fiber model Equation (2.12): N equal-length sections and local fast-axis orientations α_ℓ. Shading marks the speech-perturbed interaction window $\mathcal{L}$.

along which one linear polarization eigenmode propagates with higher phase velocity (lower effective refractive index) than its orthogonal "slow" counterpart—makes an angle $\alpha_\ell \in [0, \pi)$ with a fixed reference direction in the fiber section. Although the physical fast-axis angle $\alpha(z)$ varies continuously along the fiber, it does so on a characteristic distance set by the polarization correlation length L_{corr}, defined as the length over which $\alpha(z)$ becomes statistically uncorrelated with its initial value [38]. Choosing the section length $\Delta z \ll L_{\text{corr}}$ therefore guarantees that $\alpha(z)$ changes only slightly across any one section: as quantified by Equation (2.11) below. Fabrication imperfections and environmental stress rotate the birefringence axis essentially randomly along the fiber; accordingly, $\{\alpha_\ell\}$ is modelled as a discrete Wiener (random-walk) process [38]:

$$\alpha_\ell = \sum_{j=1}^{\ell} \Delta\alpha_j, \qquad \Delta\alpha_j \sim \mathcal{N}\big(0,\, \sigma_{\text{d}}^2\big), \qquad \sigma_{\text{d}} = \sqrt{\frac{\Delta z}{2\, L_{\text{corr}}}}. \tag{2.11}$$

A representative realization is shown in Figure 2.5. The orientation angles drift slowly with environmental conditions, on timescales of minutes to hours, and are therefore treated as a frozen random sequence within any single speech-recovery processing window of ∼20 ms.
Only a bounded region along the modelled fiber picks up speech-induced birefringence at levels formalized in Equation (2.16). The *interaction window* $\mathcal{L}$ in Figure 2.4 is the contiguous block of section indices experiencing speech-controlled coupling,

$$\mathcal{L} = \{\ell_{\min},\, \ell_{\min}+1,\, \ldots,\, \ell_{\max}\} \subseteq \{1, \ldots, N\}, \qquad 1 \le \ell_{\min} \le \ell_{\max} \le N, \tag{2.12}$$

while sections with $\ell \notin \mathcal{L}$ inherit no extra phase shift from Equation (2.16). Since section ℓ tiles the interval $[(\ell-1)\Delta z,\, \ell\Delta z]$ of $[0, L]$, the window occupies a contiguous physical span of width $D = |\mathcal{L}|\,\Delta z$, where $|\mathcal{L}| = \ell_{\max} - \ell_{\min} + 1$, centred at a coordinate z_c.

2.4.2 Per-Section Jones Matrix

Each section accumulates a relative phase $\Delta\beta\,\Delta z$ between its two birefringence eigenmodes, where $\Delta\beta = \beta_x - \beta_y$ is the differential propagation constant. Introduc-

to express $\mathbf{M}(\omega)$ as a product of constant-α_j waveplates. Since the present model propagates no chromatic dispersion, the outer split-step is absent: only the polarization product remains, and a single step size suffices. We use the symbol Δz for this single per-section length, which is the counterpart of Marcuse's microscopic δz, *not* of his macroscopic Δz.

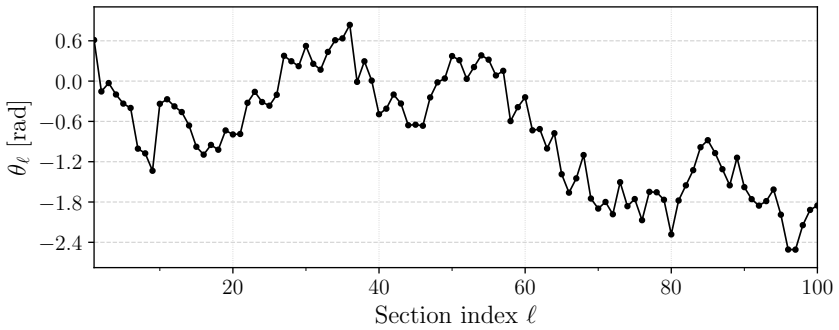


Figure 2.5: Example realization of the cumulative angle α_ℓ from Equation (2.11) (discrete Wiener along sections). Here the angle is plotted without folding onto $[0, \pi)$ for visualization purposes.

ing the birefringence parameter $b = \Delta\beta/2$ and the per-section rotation angle

$$\theta_\ell^{(0)} := b\,\Delta z, \tag{2.13}$$

the per-section propagator follows from the coupled fiber-propagation equation of [38] under three reductions that are physically appropriate for the sensing setup considered in this thesis. First, chromatic dispersion is dropped: the probe is a CW carrier whose only spectral content originates from the audio-bandwidth modulation of the birefringence, and the group-velocity dispersion accumulated over a kilometre-scale fiber at such narrow bandwidths is many orders of magnitude below any effect that influences the SOP. Second, Kerr nonlinearity is dropped because we consider a small probe power (usually less than 5dBm). Third, the frequency-dependency is removed; this is appropriate because in a deployed coherent receiver the SOP is estimated from the equalizer weight matrix at the single carrier frequency. Under these three reductions, Marcuse's [38] per-section transfer matrix reduces to

$$\mathbf{J}_\ell = \cos\theta_\ell^{(0)}\,\mathbf{I}_2 + i\sin\theta_\ell^{(0)}\,\Sigma_\ell = \begin{pmatrix} \cos\theta_\ell^{(0)} + i\sin\theta_\ell^{(0)}\cos 2\alpha_\ell & i\sin\theta_\ell^{(0)}\sin 2\alpha_\ell \\ i\sin\theta_\ell^{(0)}\sin 2\alpha_\ell & \cos\theta_\ell^{(0)} - i\sin\theta_\ell^{(0)}\cos 2\alpha_\ell \end{pmatrix}, \tag{2.14}$$

where the birefringence orientation matrix at section ℓ

$$\Sigma_\ell = \sigma_3\,\cos 2\alpha_\ell + \sigma_1\,\sin 2\alpha_\ell \tag{2.15}$$

encodes the fast-axis direction in terms of the Pauli matrices σ_1, σ_3. Geometrically, $\mathbf{J}_\ell$ rotates the Stokes vector by an angle $2\theta_\ell^{(0)} = \Delta\beta\,\Delta z$ about the axis $\hat{n}_\ell = (\sin 2\alpha_\ell,\, 0,\, \cos 2\alpha_\ell)^\top$.

Remark. The split-step decomposition (Section 2.4.1) and per-section polarization propagator Equation (2.14) summarize standard fiber polarization propagation within [38], specialized here with the simplifying assumptions mentioned in Section 2.4.2. Beginning with Section 2.4.3, acoustic stress is discussed to emulate speech-related leakage onto the observable SOP; the interaction window Equation (2.12), the resulting phase rule Equation (2.16) with sliding extension Equations (2.17) and (2.18), and the subsequent inclusion of $\Delta\phi_\ell$ into Equation (2.14)

are modelling choices contributed by this thesis. The idea to relate the acoustic pressure to the rotation angle of SOP on the Poincaré sphere is motivated by [33].

2.4.3 Speech-Induced Birefringence Perturbation

Acoustic waves near the fiber strain the cladding and modulate the optical refractive index through the elasto-optic effect [8]. At the level of polarization eigenmodes, this coupling perturbs the *difference* in effective refractive index between the two orthogonal linear polarizations—equivalently the local differential propagation constant $\Delta\beta$ that quantifies linear birefringence (Section 2.4.2). Over a section of fixed length Δz, a small change $\Delta\beta$ produces an additional phase shift $\Delta\phi_\ell \approx \Delta\beta\,\Delta z$ between the fast and slow eigenmodes relative to the unperturbed fiber; this is the same extra term that appears in step (i) of Figure 2.3, where it is abbreviated as "phase shift" for the block diagram. In the small-perturbation regime relevant here ($|x| \ll 1$, which can always be enforced by a constant rescaling of the recorded speech amplitude), $\Delta\phi_\ell$ is taken linear in the speech amplitude. With $\mathcal{L}$ supplied by Equation (2.12) and Figure 2.4, only sections close enough acoustically participate,

$$\Delta\phi_\ell = \begin{cases} \xi\, x, & \ell \in \mathcal{L}, \\ 0, & \ell \notin \mathcal{L}, \end{cases} \tag{2.16}$$

where the coupling coefficient ξ absorbs all material and geometric constants. Two assumptions are implicit in Equation (2.16): (a) within each section the elasto-optic coupling is linear in x, valid in the small-perturbation regime; (b) the speech amplitude is approximately uniform across the interaction window. The window is parameterised by a fixed width $D = |\mathcal{L}|\,\Delta z$, and by a centre position z_c.

Time-varying interaction window. One may therefore model modest motion of the acoustic source parallel to the fiber by jittering only the centre about z_c. Let $z_\ell^{\mathrm{m}} := (\ell - \frac{1}{2})\Delta z$ denote the midpoint of section ℓ, and take $z_c \in \mathbb{R}$ as the reference centre fixed when the acoustic source is undisplaced.

$$\mathcal{L}_k = \left\{ \ell \in \{1, \dots, N\} \;:\; z_c + w_k - \frac{D}{2} \;\le\; z_\ell^{\mathrm{m}} \;<\; z_c + w_k + \frac{D}{2} \right\}, \tag{2.17}$$

where $w_k \in [-w_{\max}, w_{\max}]$ translates the coupling window along $\pm z$. In words, Equation (2.17) admits every section whose midpoint lies inside $[z_c + w_k - D/2,\, z_c + w_k + D/2)$—a fixed-length interval centred on $z_c + w_k$. Offsets $\{w_k\}$ are generated by a centred random walk constrained to $\pm w_{\max}$:

$$w_k = w_{\max} \cdot \frac{\tilde{w}_k - \bar{\tilde{w}}}{\max\limits_j |\tilde{w}_j - \bar{\tilde{w}}|}, \qquad \tilde{w}_k = \sum_{j=0}^{k-1} \varepsilon_j, \qquad \varepsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{2.18}$$

In Equation (2.18), indices $k = 1, \dots, K$ label one trace; $\bar{\tilde{w}} = \frac{1}{K}\sum_{m=1}^{K} \tilde{w}_m$ is the mean from $\{\tilde{w}_m\}$, $\max_j$ runs over $j \in \{1, \dots, K\}$, and rescaling enforces $|w_k| \le w_{\max}$ (see Figure 2.6). Sliding the window Equation (2.17) makes $\mathcal{L}_k$ time-varying so the channel is no longer invariant from sample to sample: the synthetic SOP trajectory gains mild non-stationary structure and becomes more diverse than under a fixed interaction window, deliberately providing a stronger test of recovery performance.

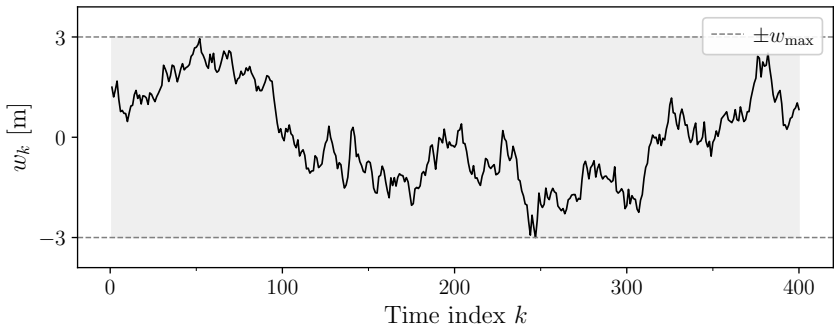


Figure 2.6: Example realization of the time-varying interaction window offset w_k defined in Equation (2.18).

2.4.4 Speech As the Rotation Angle

Speech enters the model by replacing the unperturbed rotation angle $\theta_\ell^{(0)}$ in Equation (2.14) by a speech-modulated angle. Adding the differential phase shift $\Delta\phi_\ell$ from Equation (2.16) to the unperturbed accumulated differential phase $2\theta_\ell^{(0)} = \Delta\beta\,\Delta z$ and dividing by two to convert back to the parameter that appears in $\mathbf{J}_\ell$,

$$\theta_\ell = \theta_\ell^{(0)} + \tfrac{1}{2}\,\Delta\phi_\ell = b\,\Delta z + \tfrac{1}{2}\,\xi\,x \qquad (\ell \in \mathcal{L}), \tag{2.19}$$

and substituting θ_ℓ for $\theta_\ell^{(0)}$ in Equation (2.14) yields the speech-modulated section Jones matrix. On the Poincaré sphere, the perturbation adds a small extra rotation by the angle $\Delta\phi_\ell$ in Figure 2.7.

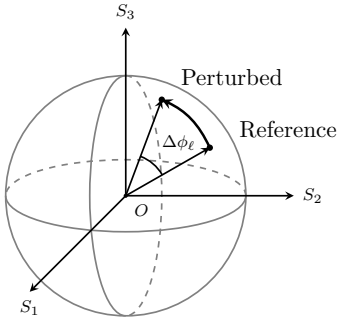


Figure 2.7: Speech-induced SOP perturbation on the Poincaré sphere. The reference vector is the unperturbed SOP and the perturbed vector is the SOP after a single section in the interaction window applies its speech-modulated rotation. The arc indicates the half-angle $\Delta\phi_\ell/2$ that enters the Jones-space matrix in Equation (2.19); on the Stokes sphere this corresponds to a rotation by the full angle $\Delta\phi_\ell$ about the section axis $\hat{n}_\ell$.

The forward model depends on ξ and x only through their product: any rescaling $x \to c\,x$, $\xi \to \xi/c$ with $c > 0$ leaves Equation (2.19), and hence the entire SOP trajectory, invariant. The product ξx is therefore the only experimentally observable combination, and ξ cannot be separately identified from the absolute amplitude of x through SOP measurements alone. Since in simulation x is read from a speech

recording whose overall gain is itself a free parameter, we fix ξ by convention rather than measurement and adopt

$$\xi = 2\,b\,\Delta z, \tag{2.20}$$

which substitutes into Equation (2.19) to give the compact final form

$$\theta_\ell = \begin{cases} b\,\Delta z\,(1+x), & \ell \in \mathcal{L}, \\ b\,\Delta z, & \ell \notin \mathcal{L}. \end{cases} \tag{2.21}$$

In this convention x acts as a dimensionless fractional modulation of the local birefringence strength: $x = 0$ recovers the unperturbed section, while $|x| \ll 1$ produces a small angular perturbation of magnitude $b\,\Delta z\,|x|$ proportional to the instantaneous speech sample. The simulations in Chapter 4 use Equation (2.21) substituted into Equation (2.14).

2.4.5 Total Channel Matrix

Cascading the N per-section matrices in propagation order gives the total channel matrix

$$\mathbf{H} = \mathbf{J}_N\,\mathbf{J}_{N-1}\cdots\mathbf{J}_1, \tag{2.22}$$

which evolves on two timescales: a slow Wiener drift of $\{\alpha_\ell\}$ on the order of minutes to hours (treated as frozen within a recovery window, see Section 2.4.1), and a fast audio-rate modulation of θ_ℓ for $\ell \in \mathcal{L}$ through the speech sample x (Equation (2.21)). The probe Jones vector $\mathbf{u} = [1, 0]^\top$ propagates through $\mathbf{H}$ and is corrupted by ASE noise to produce the noisy received Jones vector $\mathbf{r}$ specified by the channel model in Equation (2.10). The observable instantaneous SOP is then obtained from $\mathbf{r}$ via the Jones-to-Stokes conversion of Equation (2.5)–Equation (2.7), completing the forward chain of Figure 2.3.

2.4.6 Frozen-Axis Property

A useful property of the construction above is that, within a single speech-recovery processing window, the speech-induced perturbation rotates the instantaneous SOP about an *effective rotation axis* $\hat{\mathbf{n}} \in \mathcal{S}^2$ that is fixed in time. To see why, consider a perturbed section $\ell \in \mathcal{L}$ contributing the small extra Stokes-space rotation by angle $\Delta\phi_\ell$ about $\hat{n}_\ell$. The downstream sections $\mathbf{J}_N \cdots \mathbf{J}_{\ell+1}$ act on the Stokes sphere as a rotation that maps $\hat{n}_\ell$ to a fixed unit vector $\hat{\mathbf{n}}$. Since $\hat{\mathbf{n}}$ depends only on the slow-drifting $\{\alpha_j\}$ and not on x, it is effectively constant across a speech frame. The demodulation algorithms developed in Chapter 3 exploit this frozen-axis property to recover the speech as a one-dimensional projection of the SOP trajectory along $\hat{\mathbf{n}}$.

3

Methods

3.1 Speech Recovery Pipeline

The Stokes trajectory matrix $\mathbf{S} \in \mathbb{R}^{3\times K}$ (Equation (2.9)) records the full SOP evolution as the speech signal perturbs the fiber. Given $\mathbf{S}$, the task is to estimate the speech signal $\mathbf{x}$. Motivated by [13], the recovery pipeline can be decomposed into three stages:

1. **Preprocessing** ensures that the SOP trajectory lies on the Poincaré sphere.
2. **Demodulation** is a dimensionality reduction problem $\mathbb{R}^{3\times K} \longrightarrow \mathbb{R}^{K}$.
3. **Enhancement** suppresses this residual noise while preserving perceptual speech quality.

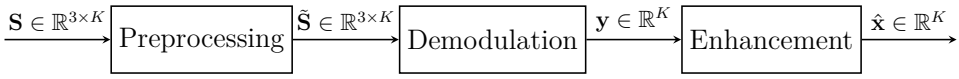


Figure 3.1: Speech recovery pipeline. $\mathbf{S}$ is the Stokes trajectory and $\tilde{\mathbf{S}}$ is the unit-normalized Stokes trajectory. $\mathbf{y}$ is the demodulated recovered speech signal (usually noisy) and $\hat{\mathbf{x}}$ is the enhanced speech signal (final output).

Three demodulation methods are discussed in this thesis: baseline method (the most direct method), Independent Component Analysis (ICA) method (the prior work [13] used), and proposed method. The preprocessing and enhancement stages are shared by all three pipelines.

Baseline method

The baseline method exploits the fact that acoustic perturbations modulate the fiber birefringence, causing a small oscillation of the SOP around its reference point. One of the three Stokes components captures the projection of this oscillation most directly; the baseline identifies and uses that component.

Normalization. Each Stokes vector is projected onto the unit Poincaré sphere:

$$\tilde{\mathbf{S}}_k = \frac{\mathbf{S}_k}{\|\mathbf{S}_k\|}, \qquad k = 0, \ldots, K-1. \tag{3.1}$$

Component selection. The component with the highest variance is selected as the demodulated speech signal:

$$j^* = \arg\max_{j\in\{1,2,3\}} \mathrm{Var}(\tilde{\mathbf{s}}_j), \qquad \mathbf{y} = \tilde{\mathbf{s}}_{j^*}, \tag{3.2}$$

because this component is the one perturbed most by the speech.

Bandpass filter. A zero-phase order-3 Butterworth bandpass filter [39] with passband 100–3200 Hz [13] is applied to suppress out-of-band noise:

$$\hat{\mathbf{x}} = \mathrm{BPF}(\mathbf{y}). \tag{3.3}$$

ICA method

Blind source separation via ICA has been employed in prior polarimetric sensing work [13] and is included here as a reference method. FastICA [40] treats the three Stokes channels as a linear instantaneous mixture of statistically independent sources and recovers them by maximizing non-Gaussianity.

Normalization. Same as Equation (3.1).

Prefiltering. Each row of $\tilde{\mathbf{S}}$ is bandpass-filtered with passband 100–3200 Hz to enhance the correlation of the separated components with the original speech [13]:

$$\tilde{\mathbf{s}}'_i = \mathrm{BPF}(\tilde{\mathbf{s}}_i), \qquad i = 1, 2, 3, \tag{3.4}$$

yielding the pre-filtered matrix $\tilde{\mathbf{S}}' = [\tilde{\mathbf{s}}'_1, \tilde{\mathbf{s}}'_2, \tilde{\mathbf{s}}'_3]^\top \in \mathbb{R}^{3\times K}$.

Decomposition. ICA itself works with any mixture dimension; here only three polarization observables $[\tilde{\mathbf{s}}'_1, \tilde{\mathbf{s}}'_2, \tilde{\mathbf{s}}'_3]$ available. Under the ICA decomposition model, this gives three recovered components:

$$\mathbf{C} = \mathbf{W}\,\tilde{\mathbf{S}}', \qquad \mathbf{C} \in \mathbb{R}^{3\times K}, \tag{3.5}$$

where $\mathbf{W} \in \mathbb{R}^{3\times 3}$ is estimated by the FastICA algorithm. The rows of $\mathbf{C}$ are the separated components $\mathbf{c}_i, i = 1, 2, 3$.

Source selection. Kurtosis measures the departure from Gaussianity of a distribution. Since ICA is applied to centred observations, each separated component $\mathbf{c}_i$ has zero empirical mean by construction. For a zero-mean sequence $\mathbf{g} = [g[0], \ldots, g[K-1]]^\top \in \mathbb{R}^K$, the excess kurtosis is estimated as

$$\kappa(\mathbf{g}) = \frac{\frac{1}{K}\sum_{k=0}^{K-1} g[k]^4}{\left(\frac{1}{K}\sum_{k=0}^{K-1} g[k]^2\right)^2} - 3. \tag{3.6}$$

By the central limit theorem, linear mixtures of independent sources tend towards Gaussianity ($\kappa \approx 0$); recovering an unmixed component therefore restores its original non-Gaussian character. Speech is super-Gaussian with $\kappa > 0$ with its sparse, impulsive amplitude distribution. We therefore select the ICA component with the largest kurtosis as the speech estimate:

$$j^* = \arg\max_{j \in \{1,2,3\}} \kappa(\mathbf{c}_j), \qquad \mathbf{y} = \mathbf{c}_{j^*}. \tag{3.7}$$

Bandpass filter. Same as Equation (3.3).

Proposed method

The proposed method exploits the geometric structure of the small-angle birefringence perturbation corresponding to our perturbation model described in Section 2.4.

Normalization. Same as Equation (3.1).

Reference point. The reference point $\mathbf{S}_0$ is estimated as the unit-normalized column mean of $\tilde{\mathbf{S}}$:

$$\mathbf{S}_0 = \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|}, \qquad \boldsymbol{\mu} = \frac{1}{K}\,\tilde{\mathbf{S}}\,\mathbf{1}_K \in \mathbb{R}^3, \tag{3.8}$$

where $\mathbf{1}_K \in \mathbb{R}^K$ is the all-ones vector.

Speech as a rotation angle. The perturbation model (Section 2.4) establishes that the speech signal linearly modulates the birefringence of the perturbed fiber sections, so that in Stokes space the output SOP undergoes a rotation by angle $\theta_k \propto x[k]$. Since the effective rotation axis $\hat{\mathbf{n}}$ is approximately fixed across all time steps, in the small-angle regime this gives

$$\tilde{\mathbf{S}}_k \approx \mathbf{S}_0 + \theta_k\,(\hat{\mathbf{n}} \times \mathbf{S}_0), \tag{3.9}$$

where $\hat{\mathbf{n}} \perp \mathbf{S}_0$. The goal is therefore to invert this relationship: recover the sequence of rotation angles θ_k, which directly yields $\mathbf{y}$.
Taking the cross-product of $\mathbf{S}_0$ with Equation (3.9), the rotation information is encoded compactly as

$$\mathbf{v}_k \;=\; \mathbf{S}_0 \times \tilde{\mathbf{S}}_k \;\approx\; \theta_k\,\hat{\mathbf{n}}. \tag{3.10}$$

Each $\mathbf{v}_k \in \mathbb{R}^3$ is a computable quantity, but it includes two unknowns: the rotation angle θ_k and the axis $\hat{\mathbf{n}}$. Stacking $\mathbf{V} = [\mathbf{v}_0 \mid \cdots \mid \mathbf{v}_{K-1}] \in \mathbb{R}^{3\times K}$ gives

$$\mathbf{V} \approx \hat{\mathbf{n}}\,\boldsymbol{\theta}^\top, \qquad \boldsymbol{\theta} = [\theta_0, \ldots, \theta_{K-1}]^\top. \tag{3.11}$$

Joint estimation of axis and speech via SVD. The SVD of $\mathbf{V}$ solves the approximation problem jointly for $\hat{\mathbf{n}}$ and $\boldsymbol{\theta}$:

$$\mathbf{V} = \mathbf{U}\,\boldsymbol{\Sigma}\,\mathbf{W}^\top, \qquad \hat{\mathbf{n}} = \mathbf{U}_{:,0}. \tag{3.12}$$

The leading left singular vector $\hat{\mathbf{n}} = \mathbf{U}_{:,0}$ estimates the rotation axis. Projecting $\mathbf{V}$ onto the estimated axis recovers the speech estimate:

$$\mathbf{y} = \mathbf{V}^\top \hat{\mathbf{n}}. \tag{3.13}$$

Bandpass filter. Same as Equation (3.3).

3.2 Performance Metrics

Speech quality is multi-dimensional: a recovered signal may closely follow the waveform of the reference yet remain perceptually degraded, or it may be intelligible while still corrupted by background noise. No single scalar metric captures all of these aspects. The prior work [13] evaluates recovery performance using the Pearson correlation coefficient [41], which measures linear waveform similarity but has no validated link to the intelligibility or perceived quality of the recovered speech. To obtain a more complete picture, this thesis employs three objective performance metrics that together cover waveform-level fidelity, perceptual quality, and intelligibility.

Scale-Invariant Signal-to-Distortion Ratio (SI-SDR). Let $\mathbf{x} \in \mathbb{R}^K$ denote the clean reference speech and $\hat{\mathbf{x}} \in \mathbb{R}^K$ the estimated speech, both zero-mean. The SI-SDR is defined as [42]

$$\text{SI-SDR}(\mathbf{x}, \hat{\mathbf{x}}) = 10 \log_{10} \frac{\|\gamma \mathbf{x}\|^2}{\|\gamma \mathbf{x} - \hat{\mathbf{x}}\|^2}, \quad \gamma = \frac{\hat{\mathbf{x}}^\top \mathbf{x}}{\|\mathbf{x}\|^2}, \tag{3.14}$$

where $\gamma\mathbf{x}$ is the best-fit projection of the estimate onto the reference, and the denominator is the residual distortion power. By projecting out the optimal scale factor γ, SI-SDR is invariant to any global amplitude rescaling of $\hat{\mathbf{x}}$—a property important here because the recovered waveform carries an arbitrary gain. SI-SDR is expressed in decibels (dB), with higher values indicating better reconstruction, and is widely adopted as the standard evaluation metric in the speech separation and enhancement literature [43, 44, 45].

Short-Time Objective Intelligibility (STOI). Waveform-level distortion does not directly predict whether a listener can understand the recovered speech, since the auditory system can tolerate additive noise as long as key speech spectral cues remain intact. STOI [46] measures intelligibility by working on short-time spectral envelopes. STOI scores lie in $[0, 1]$, with 1 indicating perfect intelligibility. The measure is specifically designed and validated for speech processed by noise-reduction algorithms, and correlates well with intelligibility scores from listening tests, making it the standard intelligibility measure in the speech enhancement community [47, 48, 45]. In this thesis, STOI is computed using `pystoi`.

Perceptual Evaluation of Speech Quality (PESQ). Even intelligible speech can be perceived as poor quality if it contains audible noise or tonal distortions. PESQ [49] measures this by aligning the clean reference and the degraded signal in time, passing both through a psychoacoustic model, computing a perceptual disturbance, and mapping the result to a predicted mean opinion score. Scores range from -0.5 to 4.5, with 4.5 corresponding to perfect quality. A systematic comparison of objective speech quality measures in [50] shows that PESQ achieves a correlation of 0.89 with subjective quality ratings of noise-reduced speech—the highest among other standard measures—making it the most reliable objective quality predictor for noise-reduction algorithms. In this thesis, PESQ is computed using `pesq`.

3.3 Data-driven Speech Enhancement

Non-learning methods such as bandpass filtering or spectral subtraction [51] can be effective when the noise is additive, Gaussian, and stationary, and when the signal-to-noise ratio (SNR) is sufficiently high. However, in a real fiber-optic sensing system, the environmental sensitivity of the fiber produces residual noise that can be non-additive, non-Gaussian, and non-stationary. Moreover, the effective SNR at the output of the demodulation stage can be low due to our small-perturbation scenario. Under these conditions, classical denoisers are unable to suppress structured interference without simultaneously degrading the speech content [50], leading to a fundamental limitation in the quality and intelligibility of the recovered speech. The deep neural network (DNN) offers a remarkable solution to overcome this limitation [47, 52, 53, 54]. As shown in Figure 3.2, the key idea is to learn a mapping from a noisy spectrogram to a clean one, exploiting the fact that speech and noise occupy different, learnable patterns in the time-frequency (T-F) domain (spectrogram representation).

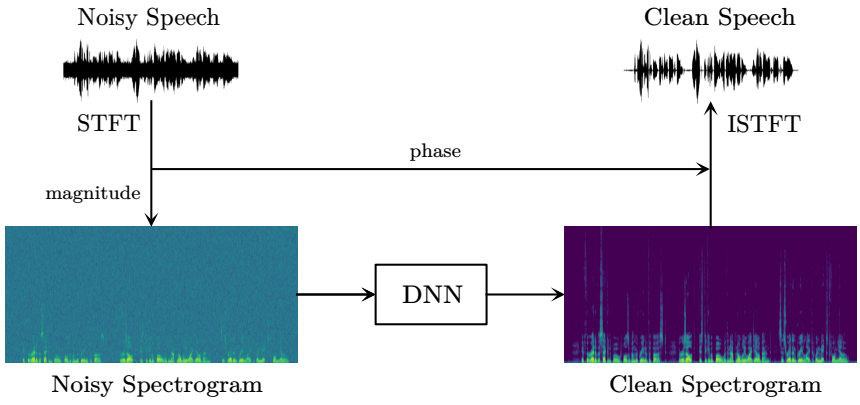


Figure 3.2: DNN-based speech enhancement. STFT and ISTFT are defined in Equation (3.15) and Equation (3.17). The spectrogram is defined in Equation (3.16).

Spectrogram Representation. The Short-Time Fourier Transform (STFT) of a discrete-time signal $y[k]$ with a window function $w[m]$ of length M and hop size

H is

$$Y[n, f] = \sum_{m=0}^{M-1} y[nH + m]\, w[m]\, e^{-i2\pi fm/M}, \quad f = 0, \ldots, \lfloor M/2 \rfloor, \tag{3.15}$$

where n is the frame index and f is the discrete frequency bin. Each column $Y[n, \cdot]$ is the spectrum of a short segment centered around time nH/f_s.
The *spectrogram* is the squared magnitude in the STFT domain

$$\mathcal{P}[n, f] = |Y[n, f]|^2. \tag{3.16}$$

The original signal is reconstructed via the Inverse Short-Time Fourier Transform (ISTFT), which inverts each frame with an Inverse Discrete Fourier Transform and recombines the frames by overlap-add:

$$\hat{y}[nH + m] = \frac{1}{M} \sum_{f=0}^{M-1} Y[n, f]\, e^{\,i2\pi fm/M}, \quad m = 0, \ldots, M-1, \tag{3.17}$$

where overlapping frames are summed and normalized to recover $\hat{y}[k]$.
Spectral magnitudes $|Y|$ can span a large dynamic range across time-frequency bins; a logarithmic amplitude map compresses that range. Following the precedent work [55], the log-magnitude is defined by

$$\tilde{Y}[n, f] = \log\Big(1 + |Y[n, f]|\Big), \tag{3.18}$$

where the additive unity keeps the logarithm bounded as $|Y[n, f]| \to 0$.
For notation brevity, we will use the matrix form $\mathbf{Y}$ to denote the noisy STFT magnitude $Y[n, f]$, $\hat{\mathbf{X}}$ to denote the enhanced STFT magnitude $X[n, f]$, and $\tilde{\mathbf{Y}}$ to denote the log-magnitude $\tilde{Y}[n, f]$. These representations are arranged as a single-channel tensor $\in \mathbb{R}^{1\times T\times F}$ because each time–frequency bin carries only that one real feature.

Convolutional Neural Network (CNN) CNN is suited to this kind of speech enhancement tasks because spectrograms can be viewed as images containing strong time and frequency correlations [56]. Motivated by [57], the CNN architecture used in this thesis is described in Figure 3.3. The input is the log-magnitude tensor $\tilde{\mathbf{Y}}$ defined in Equation (3.18).

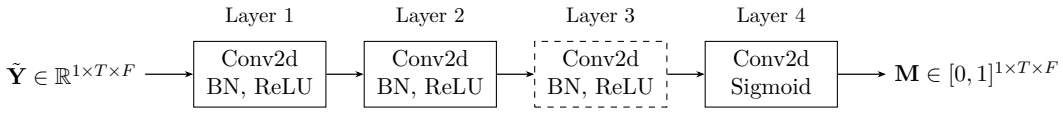


Figure 3.3: The CNN architecture. The dashed box (Layer 3) indicates a dilated convolution. Key parameters are shown in Table 3.1.

2-D convolution (Conv2d). Because speech spectrograms carry structure simultaneously along the frequency and time axes, each layer applies 2D convolutional filters to extract local time-frequency patterns. Each layer is parameterized by C_{in}

input channels and C_{out} output channels: C_{in} is the number of feature maps received from the previous layer, and C_{out} is the number of independent filter banks applied, determining the richness of the learned representation.

Batch Normalisation (BN). After each convolutional layer except the output, batch normalization normalizes feature maps over the batch dimension to stabilize gradient flow and enable faster, more robust training.

Rectified Linear Unit (ReLU). ReLU activations ($\max(0, x)$) introduced after each BN layer enable the network to represent nonlinear mappings.

Dilated convolution (Layer 3). Capturing long-range temporal context with standard convolutions requires either large kernels or many stacked layers. Dilated convolutions [58] address this by inserting uniform gaps between kernel elements, expanding the receptive field exponentially without adding parameters.

Table 3.1: CNN layer parameters (frequency × time).

Layer	C_{in}	C_{out}	Kernel	Dilation	Padding
1	1	32	3×3	1×1	1×1
2	32	64	3×3	1×1	1×1
3	64	64	1×7	1×2	0×6
4	64	1	1×1	1×1	0×0

Soft mask and time domain reconstruction. A *soft mask* $\mathbf{M}$ assigns a continuous gain $\in [0, 1]$ to each bin (n, f) [47]. $\odot$ denotes the element-wise product between two matrices. The enhanced magnitude matrix $\hat{\mathbf{X}}$ can be obtained by applying the mask to the original magnitude:

$$\hat{\mathbf{X}} = \mathbf{M} \odot \mathbf{Y}. \tag{3.19}$$

Reusing the phase provided by the noisy speech $\mathbf{y}$, the enhanced speech signal $\hat{\mathbf{x}}$ can be reconstructed using the ISTFT Equation (3.17).

Training objective. Training penalizes the masked log-magnitude features. Let $\tilde{\mathbf{Y}}_x, \tilde{\mathbf{Y}}_y \in \mathbb{R}^{T \times F}$ denote the clean and noisy log-magnitude spectrograms, where subscript x refers to the clean reference speech and y to the noisy mixture fed to the enhancement stage. During training, $\mathbf{M}$ is the output mask in Figure 3.3 when the input is $\tilde{\mathbf{Y}}_y$.

The training loss combines a mean absolute error on masked log features with a

temporal smoothness term on $\mathbf{M}$:

$$\begin{aligned}\mathcal{L}(\theta) = &\underbrace{\frac{1}{TF}\sum_{n=0}^{T-1}\sum_{f=0}^{F-1}\Big|\big(\mathbf{M}\odot\tilde{\mathbf{Y}}_y\big)[n,f]-\tilde{\mathbf{Y}}_x[n,f]\Big|}_{\mathcal{L}_{\text{rec}}}\\ &+\lambda\,\underbrace{\frac{1}{(T-1)F}\sum_{n=0}^{T-2}\sum_{f=0}^{F-1}\Big|\mathbf{M}[n{+}1,\,f]-\mathbf{M}[n,\,f]\Big|}_{\mathcal{L}_{\text{smooth}}},\quad \lambda=0.02.\end{aligned} \tag{3.20}$$

Here $\mathcal{L}_{\text{rec}}$ pulls the masked noisy log spectrum toward the clean target, while $\mathcal{L}_{\text{smooth}}$, serving as the regularization term, averages the absolute mask increments along time to prevent abrupt changes. The coefficient λ is usually small to keep $\mathcal{L}_{\text{rec}}$ dominant; $\lambda = 0.02$ here is an empirical value.

Spectral Subtraction Baseline. To evaluate the benefit of the data-driven CNN, we compare it with spectral subtraction [51], which is regarded as a classical baseline among spectral speech enhancement methods. It is an unsupervised algorithm that estimates the noise power spectrum from speech-absent frames and subtracts it from the noisy speech spectrum. Concretely, let $|\tilde{X}(\omega)|^2$ denote the short-time power spectrum of the noisy signal and $|\hat{N}(\omega)|^2$ the estimated noise power spectrum averaged over silent frames; the enhanced spectrum is obtained as

$$|\hat{X}(\omega)|^2 = \max\Big(|\tilde{X}(\omega)|^2 - |\hat{N}(\omega)|^2,\,0\Big), \tag{3.21}$$

and the enhanced waveform is reconstructed by inverse STFT reusing the phase of the noisy signal.

Remark. This data-driven block is deployed only in the experimental pipeline Chapter 4; the Monte Carlo study injects simpler Gaussian noise impairments, so classical bandpass filtering is enough there.

4

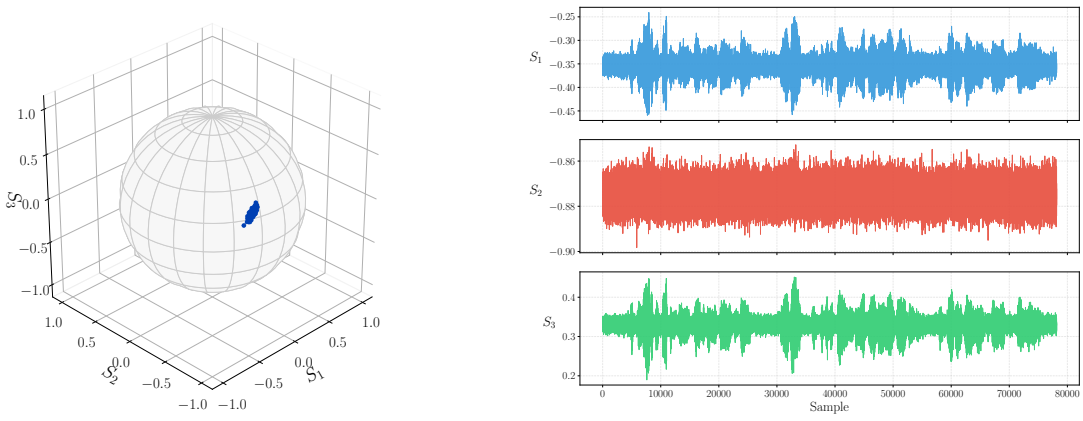
Results

4.1 Simulation

The simulation study is designed to evaluate the demodulation step of the recovery pipelines described in Section 3.1. In this controlled setting, the only noise source is the ASE noise model of Equation (2.10), i.e. complex circular Gaussian noise $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_2)$.

Choice of the perturbation speech signals. Speech signals used in simulation are drawn from the VCTK dataset [59], choosing five male and five female talkers. The aim is to make sure the results can generalize not only to different fiber realizations but also to different speech signals.

4.1.1 Simulated SOP



(a) Simulated SOP trajectory on the Poincaré sphere.

(b) Simulated noisy Stokes parameters in time domain.

Figure 4.1: Simulated SOP example for a single fiber realization.

Figure 4.1 illustrates a representative realization of the simulated SOP. The left figure shows the trajectory on the Poincaré sphere: speech drives the operating point in a small neighbourhood of a reference $\mathbf{S}_0 \in \mathbb{R}^3$. The right figure shows the corresponding Stokes time series; speech modulation is visible under ASE noise. Throughout the simulation study, the interaction window is allowed to slide along

the fiber following Equation (2.17), so the participating section set $\mathcal{L}_k$ varies from sample to sample, which broadens the SOP trajectory.
In this realization, S_1 and S_3 show larger speech-related fluctuations than S_2, but this ordering is not universal. Equation (3.9) only states that, when $\hat{\mathbf{n}}$ is approximately frozen, small-angle motion is aligned with $\hat{\mathbf{n}} \times \mathbf{S}_0$, so the relative visibility of S_1, S_2, and S_3 is whichever projection of that vector is largest. Here the effective axis is nearly aligned with $[0, 1, 0]^\top$, so the oscillation lies mainly in the S_1–S_3 plane with little spread along S_2; another fiber draw or reference could instead emphasize S_2 or a different combination.

4.1.2 Monte Carlo Results

To obtain statistically reliable performance estimates, Monte Carlo trials are conducted. Each trial draws a random seed controlling the birefringence axis α_ℓ, the perturbed window position, and the ASE noise realization; for each draw, SI-SDR, STOI, and PESQ are evaluated on every speech signals. Repeating this procedure gives the summarized results as mean ± std in Table 4.1:

Table 4.1: Monte Carlo metrics (mean ± std, 50 trials). Mean scores average over 10 VCTK speeches (five male, five female) per trial.

	SI-SDR	STOI	PESQ
Baseline	5.210 ± 3.351	0.745 ± 0.048	1.467 ± 0.083
ICA	4.781 ± 2.694	0.741 ± 0.040	1.474 ± 0.082
Proposed	$\mathbf{6.027 \pm 2.627}$	$\mathbf{0.757 \pm 0.038}$	$\mathbf{1.491 \pm 0.078}$

Table 4.1 shows that the proposed method achieves the highest mean across all three metrics once speaker diversity has been pooled. The improvement is consistent but modest: the proposed method gains approximately 0.8 dB in SI-SDR, 0.012 in STOI, and 0.024 in PESQ over the baseline. Notably, the proposed method also exhibits the smallest standard deviation in these metrics, reflecting its robustness.
The modest performance improvements reflect that *the demodulation step is not the performance bottleneck*: as shown in Figure 4.4a, the ASE noise is simple enough in its pattern that the highest-variance Stokes component already aligns with the perturbation direction $\hat{\mathbf{n}} \times \mathbf{S}_0$, so baseline component selection recovers a demodulated signal of similar quality.

4.2 Experiment

Unlike fiber sensing based on physical contacts [4, 60], where mechanical vibrations couple directly into the fiber, the acoustic perturbation considered here propagates through air before reaching the fiber. Airborne sound pressure waves can induce only a weak strain on regular telecommunication fibers, producing a small SOP oscillation that is difficult to detect. To accumulate a measurable perturbation, a

kilometer-scale fiber spool is used as the sensing element, following a similar strategy to that adopted by [13].

4.2.1 Experimental Setup

As shown in Figure 4.2, the experimental system consists of three main components: a laser source (Santec TSL-570), a 1750 m single mode fiber spool acting as the acoustic sensor, and a polarimeter (Novoptel PM1000) for continuous SOP measurement. The laser output is launched into one end of the fiber spool and the polarimeter is connected to the other end, recording the full Stokes trajectory $\mathbf{S}$ at a fixed sampling rate throughout the experiment. A loudspeaker is placed at a controlled distance (around 20 cm) from the spool and driven by the speech signal to be recovered. The acoustic pressure waves radiated by the loudspeaker perturb the fiber birefringence, producing the SOP oscillation described in Section 2.4.

Both the polarimeter data acquisition and the loudspeaker playback are controlled by MATLAB, providing two practical advantages. First, the SOP recording is automated and continuous, eliminating many unnecessary and repeated manual operations during data collection. Second, the playback and acquisition are triggered together within the same MATLAB session, achieving a rough synchronization between the loudspeaker output and the recorded Stokes trajectory; this temporal alignment is essential for computing performance metrics such as SI-SDR.

Remark. The experimental setup is different from our model described in Section 2.4, as well as the simulation study in Section 4.1. Here we intend to enhance the sensitivity of the fiber while applying the same pipelines for speech recovery. The key idea is to validate and compare the speech recovery potential of different pipelines in a real-world scenario.

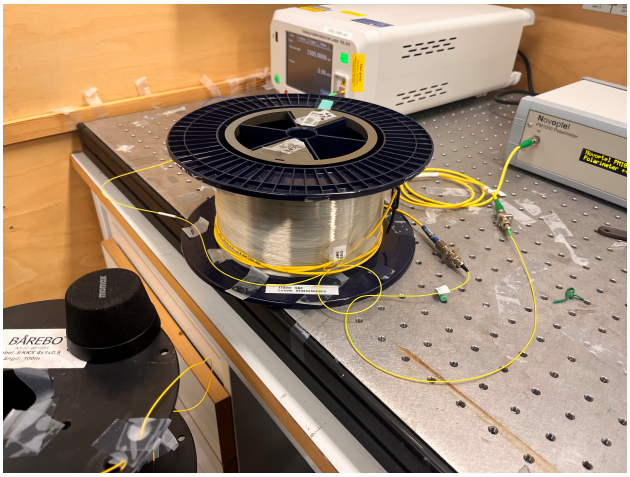
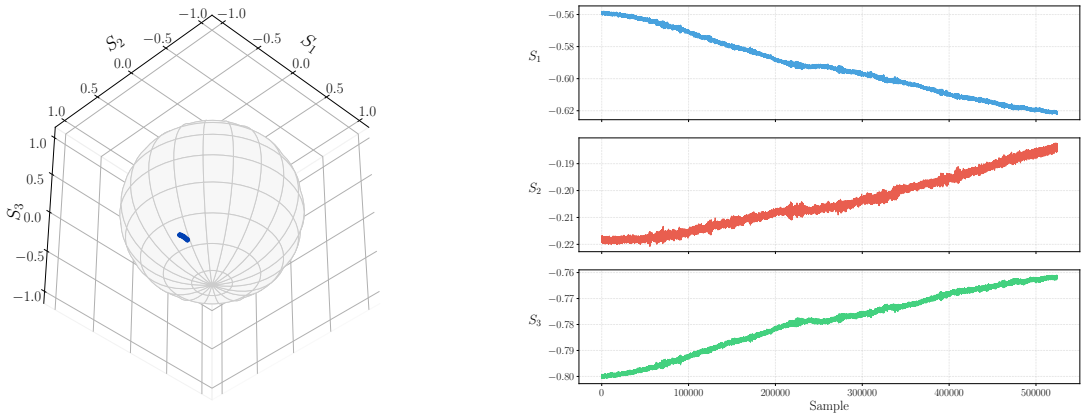


Figure 4.2: The experimental setup for the fiber sensing system.

4.2.2 Experimental SOP

The illustrative experimental SOP in Figure 4.3a and Figure 4.3b is acquired by playing back one example waveform from VCTK. The playback is sampled at $f_s = 8\,\text{kHz}$ after preprocessing.



(a) Experimental SOP trajectory on the Poincaré sphere.

(b) Experimental noisy Stokes parameters.

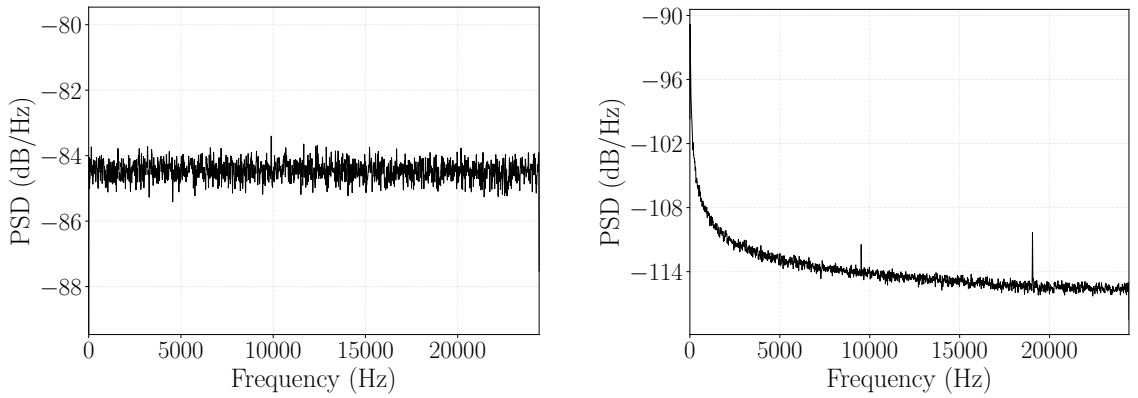
Figure 4.3: Experimental SOP example corresponding to Figure 4.2.

Several observations can be made from this example:

- A 1750 m fiber spool is enough to accumulate a visible SOP oscillation.
- Unlike the simulated example in Figure 4.1, the SOP trajectory is an arc with a strong baseline drift. The most likely reasons are (1) the mechanical relaxation of the fiber spool used in the lab; (2) the thermal effect drift accumulated in the fiber spool.
- The speech-induced SOP oscillation is not as prominent as in the simulated example, which can be attributed to the mismatch between the experimental and simulated noise models. Figure 4.4b shows the power spectral density (PSD) of the Stokes parameter S_1 measured with the loudspeaker silent, i.e., background noise only; S_2 and S_3 exhibit a similar spectral shape and are omitted for brevity. Unlike the spectrally flat ASE noise assumed in the simulation, the experimental background noise is strongly coloured: its energy is concentrated in the range of approximately 100–4000 Hz, which coincides precisely with the main energy band of speech. This spectral overlap means that the background noise cannot be suppressed by a simple bandpass filter without also attenuating the speech content, explaining the reduced signal-to-noise ratio compared to the simulation and the mismatch in demodulation performance between the two settings.

4.2.3 Without Data-Driven Enhancement

After applying the same three recovery pipelines as in the simulation study and the data alignment, Table 4.2 reports the mean and standard deviation of each metric



(a) Simulated noise on S_1.

(b) Measured background noise on S_1.

Figure 4.4: Power spectral density (PSD) of the Stokes parameter S_1 when the speech is absent: (a) simulated Gaussian Jones noise propagated to S_1 and (b) measured laboratory background.

over 25 SOP recordings using one fixed speech example and Figure 4.5 shows one example waveform (first 5 s) comparison between the original speech and the baseline recovered speech.

Compared to the simulation study in Table 4.1, the first noteworthy observation is that the baseline method outperforms the other two methods. The reasons can be attributed to three assumption mismatches between the experimental setup and the simulation study:

- The proposed method requires a good estimate of a static reference point. However, from the experimental SOP example in Figure 4.3, the baseline drift makes the reference point move along the arc on the Poincaré sphere.
- The degradation in the SI-SDR indicates that the distortion brought by the noise in the lab environment is different from the simulated Gaussian noise, which has been illustrated above in Figure 4.4b.
- The ICA method assumes statistically independent, non-Gaussian sources. In the simulation this assumption is approximately satisfied because the only structured source is the speech signal corrupted by spectrally flat Gaussian noise. In the experiment, however, the coloured background noise introduces multiple correlated, non-Gaussian components into the Stokes trajectory simultaneously.

The second observation concerns the waveform-level similarity between the recovered and the original speech signal. As shown in Figure 4.5, despite the imperfect noise suppression, the baseline-recovered waveform is temporally aligned with the ground-truth speech and preserves a similar amplitude envelope structure. This is a meaningful result: it indicates that the acoustic perturbation introduced by the speech does *not* produce strong nonlinear effects on the SOP, but rather a roughly linear, proportional modulation of the Stokes trajectory consistent with the small-angle perturbation model of Section 2.4. The fact that the speech signal is recoverable at the waveform level, even in the presence of the coloured experimental

noise, provides evidence for both the validity of the perturbation model and the effectiveness of the recovery pipeline.

Table 4.2: Experimental performance averaged over 25 testing cases (mean ± std).

	SI-SDR	STOI	PESQ
Baseline	$\mathbf{-11.16 \pm 5.51}$	$\mathbf{0.6244 \pm 0.0141}$	$\mathbf{1.3898 \pm 0.0148}$
ICA	-12.21 ± 5.24	0.5869 ± 0.0090	1.3690 ± 0.0230
Proposed	-11.23 ± 5.51	0.6224 ± 0.0136	1.3886 ± 0.0104

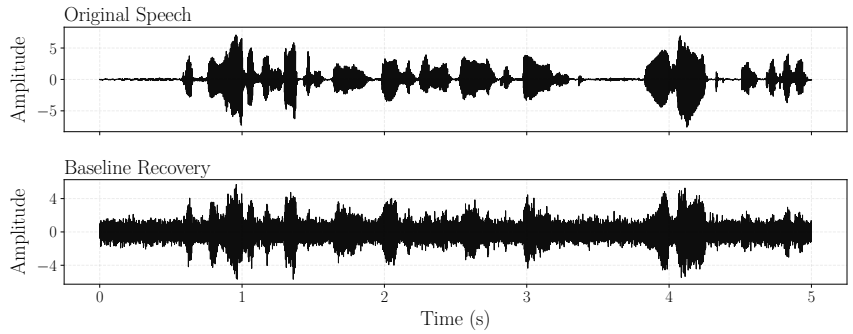


Figure 4.5: Waveform comparison between the original speech and the baseline recovered speech.

4.2.4 With Data-driven Enhancement

Since the performance bottleneck is not the demodulation step but the noise in the lab environment, we used the CNN-based speech enhancement described in Section 3.3 to improve the performance as a post-processing step in substitution of the original bandpass filter in Section 3.1. In this stage, the *baseline* demodulation is chosen as the demodulation step.

Figure 4.6: Experimental recovery chain Section 4.2.4 using the signal notation of Figure 3.1. Demodulation is fixed to baseline Section 3.1; enhancement replaces the Butterworth bandpass Equation (3.3) by CNN Section 3.3.

Dataset. This experiment was conducted on the same VCTK dataset by playing the speech signals selected from the dataset near the fiber spool, considering a balanced male and female choice of speeches. MATLAB automates this measurement process in the lab. After preprocessing and baseline demodulation, the noisy speech recordings can be obtained from the SOP measurements. The audio dataset

is prepared by pairing the noisy speech recordings with the corresponding original speech recordings, giving 270 paired clean-noisy utterances. Because the two acquisition chains run at different rates (clean reference at 48 kHz, polarimeter-derived recordings at 48.828 kHz), both signals are resampled to a common 16 kHz working rate—a standard choice for speech enhancement that retains the speech-relevant band while keeping the spectrogram input compact. A fixed-seed random split partitions the 270 pairs into 184 training, 32 validation, and 54 test utterances (roughly 68/12/20 %); the validation set is used only for early-stopping monitoring and the test set is held out for final evaluation.

Training Setup. Each utterance is cropped to a 4 s segment (64 000 samples) so that all spectrogram tensors share the same shape. The STFT uses an $N_{\mathrm{FFT}} = 512$ Hann window with hop length 128 (8 ms per frame), giving a 257-bin log-magnitude input as defined in Equation (3.18). The CNN is optimized with Adam [61] (initial learning rate 3×10^{-4}) under the loss defined in Equation (3.20). Batches of 16 utterances are drawn with reshuffling each epoch, giving roughly $\lceil 184/16 \rceil = 12$ gradient updates per epoch, and the model is trained for 30 epochs.

The Loss Curve. Figure 4.7 shows the training and validation loss curves over the training epochs. The training loss decreases steadily and the validation loss follows closely, indicating that the model converges without significant overfitting.

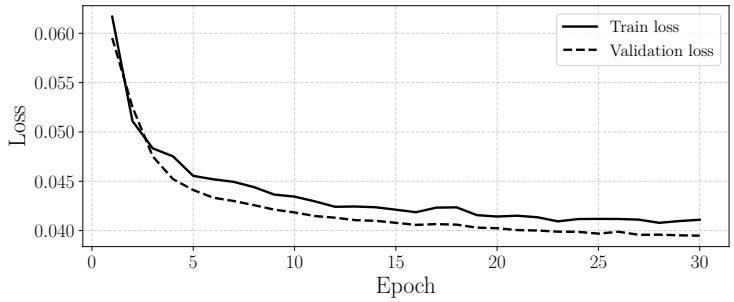


Figure 4.7: Training and validation loss curves of the CNN-based speech enhancement model.

Results. Table 4.3 reports the performance metrics. The Spec-Sub method achieves better SI-SDR but the speech quality and intelligibility are not as good as the CNN-based method. The data-driven CNN-based method achieves a significant improvement in PESQ from 1.398 to 1.927 while maintaining a good STOI value.
A representative spectrogram comparison is shown in Figure 4.8. The spectral subtraction baseline partially suppresses this floor but introduces the characteristic musical noise artefacts visible as isolated bright spots across the spectrogram. Moreover, some low-frequency speech content is also attenuated because of the spectral overlap between the speech and the noise. The CNN output recovers the spectrogram better, preserving the fundamental speech band with fewer isolated artefacts, consistent with the PESQ and STOI gains reported in Table 4.3.

Table 4.3: Speech enhancement performance.

	SI-SDR	STOI	PESQ
Noisy	−14.857	0.775	1.398
Spec-Sub	**−5.958**	0.765	1.726
CNN	−6.421	**0.787**	**1.927**

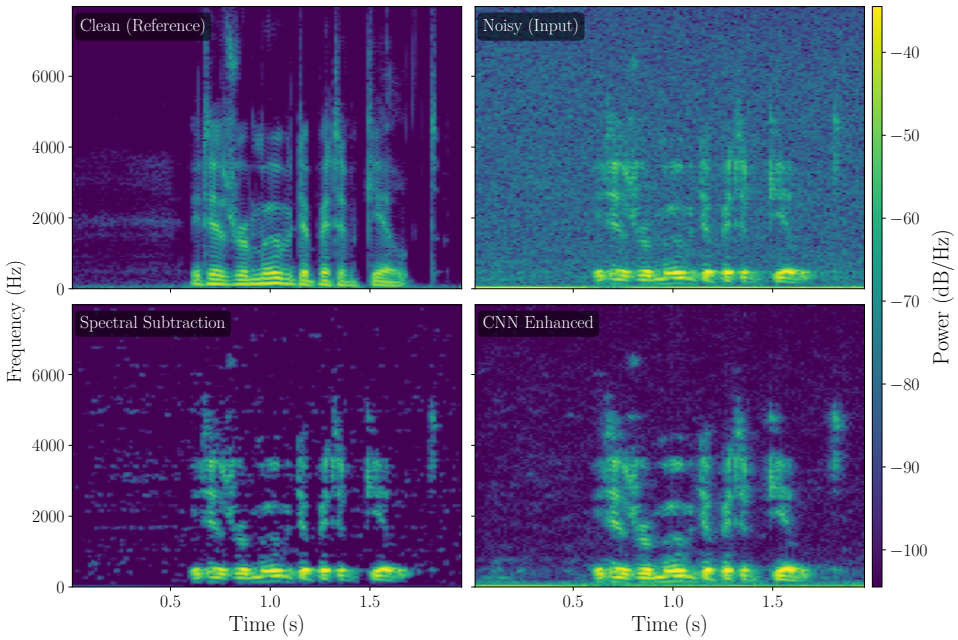


Figure 4.8: Spectrogram comparison for a representative test recording.

5

Discussion and Conclusion

5.1 Simulation vs. Experiment

5.1.1 Fiber Sensitivity

The Monte Carlo study Section 4.1 builds on the perturbation model Section 2.4. There one may rescale speech, shorten or lengthen the perturbed fiber segments, and move the trajectory on the Poincaré sphere Figure 2.1. The laboratory setup Figure 4.2 cannot mirror all of those operations: coupling strength depends together on fiber length, how the fiber spool is mounted and routed, and where the loudspeaker sits relative to the spool. If polarization-based speech recovery is ever deployed outside a tightly controlled bench—for example in an ordinary office—we may need practical sensitivity of the fiber. That motivation also explains why experiments emphasize spool-length pooling Figure 4.2: it strengthens the polarization footprint when airborne coupling alone would otherwise stay weak [11].

5.1.2 Laboratory Noise

Minor hardware choices could shift the residual noise floor even when the same spool was reused. Repeating captures on different days altered the conditions that metrics between Table 4.2 and Table 4.3 should not be read as comparable snapshots from one stationary experiment—the SOP measurements were not collected under identical hardware states.

When we pay attention to speech intelligibility, or spectrogram representations in Figure 4.8, data-driven enhancement in Section 3.3 and Figure 3.3 becomes attractive. Similar convolutional or pretrained speech-enhancement models [15, 18] could follow the same recipe. The limiting factor is usually high-quality paired data captured under realistic scenarios, if we consider superised learning. Automated acquisition workflows tested during this thesis reduce manual overhead; remaining tasks include understanding dominant interference sources and enlarging labelled datasets across lasers and recording days.

5.2 Limitations

1. **Integration with a coherent transceiver.** The ultimate deployment scenario for the attack studied in this thesis is one where the adversary extracts the SOP directly from the equalizer weights of an existing coherent receiver,

requiring no additional hardware [31, 33]. As illustrated in Figure 5.1, the full system consists of two layers: the lower layer is main communication channel, and the upper layer is the side sensing channel where the equalizer weight matrix $\mathbf{W}_k$ is passed to an SOP estimation block to produce the Stokes trajectory fed into the recovery pipeline.

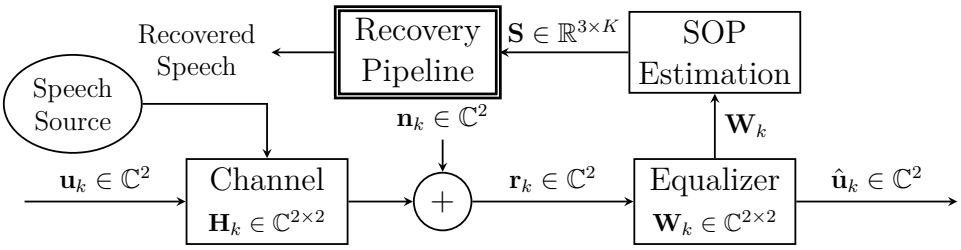


Figure 5.1: Target full-system architecture: SOP is derived from the equalizer weight matrix $\mathbf{W}_k$ of a coherent receiver, requiring no additional sensing hardware.

In this thesis, a standalone polarimeter is used instead, which decouples the speech recovery pipeline from transceiver hardware and allows a focused evaluation of the signal-processing design. Building a coherent communication testbed that estimates the SOP from equalizer coefficients is the most important next engineering step; it would also reveal the noise characteristics specific to equalizer-based SOP estimation and their impact on recovery performance.

2. **Fiber sensitivity:** Fiber sensitivity is a critical factor affecting the feasibility of effective speech recovery. For an indoor fiber environment, shorter fiber length should be used to explore the limitation of the recovery pipeline. For example, [11] applied a sensory receptor using a fiber wound on a cylinder to enhance the fiber sensitivity.
3. **Performance metrics:** Evaluating human speech quality is a challenging problem [62]. Different metrics can be weakly related to one another. For example, even if the speech is corrupted by strong noise (low SNR), the speech can still be recognized by a human listener (high STOI) as long as the main frequency components are preserved. Moreover, if an information-level metric is needed, a speech-to-text engine can be used to evaluate the Word Error Rate (WER) [63]. Moreover, merging the perceptual metrics into the design of loss functions beyond $\mathcal{L}_{\text{rec}}$ is proven to be an effective approach in speech enhancement [64, 65].
4. **End-to-end data-driven pipeline:** The current pipeline isolates the demodulation (SOP $\rightarrow$ noisy speech) and the enhancement (noisy speech $\rightarrow$ enhanced speech) steps. A better method is to design an end-to-end data-driven pipeline to extract the speech signal (or any other type of acoustic perturbation) directly from the SOP trajectory (or other available information to get at the optical receiver), which requires further theoretical and experimental investigation.

5.3 Conclusion

This thesis demonstrated the feasibility of recovering speech from the SOP trajectory of a fiber-optic link. The recovery pipeline is grounded in a signal-transform chain that translates the physical perturbation into a tractable signal-processing problem: the elasto-optic effect encodes speech as a small rotation of the Stokes vector about a fixed axis. The simulation study verified this three-stage pipeline design (preprocessing, demodulation, and enhancement) and identified the performance bottleneck, motivating the data-driven enhancement stage in the experimental data processing. The hardware experiment confirmed that the pipeline generalizes to real-world conditions: the recovered waveforms retain the amplitude envelope of the original speech, and the CNN enhancement achieves a PESQ improvement from 1.398 to 1.927 using a small-scale experimental dataset. Together, the simulation and the experiment form a mutually supporting validation: the simulation provides a basic validation of the pipeline design under a controlled noise model, while the experiment tests the full pipeline and quantifies the gain from data-driven enhancement—both evaluated with the same three objective metrics (SI-SDR, STOI, PESQ) to allow direct cross-referencing. The results provide a reproducible pipeline architecture for future work on polarization-based speech recovery.

Bibliography

[1] G. P. Agrawal, *Fiber-Optic Communication Systems*. John Wiley & Sons, Ltd, 2021.

[2] J. M. Lopez-Higuera, L. Rodriguez Cobo, A. Quintela Incera, and A. Cobo, "Fiber Optic Sensors in Structural Health Monitoring," *Journal of Lightwave Technology*, vol. 29, pp. 587–608, Feb. 2011.

[3] J. B. Ajo-Franklin, S. Dou, N. J. Lindsey, I. Monga, C. Tracy, M. Robertson, V. Rodriguez Tribaldos, C. Ulrich, B. Freifeld, T. Daley, and X. Li, "Distributed Acoustic Sensing Using Dark Fiber for Near-Surface Characterization and Broadband Seismic Event Detection," *Scientific Reports*, vol. 9, p. 1328, Feb. 2019.

[4] Z. Zhan, M. Cantono, V. Kamalov, A. Mecozzi, R. Müller, S. Yin, and J. C. Castellanos, "Optical polarization–based seismic and water wave sensing on transoceanic cables," *Science*, vol. 371, pp. 931–936, Feb. 2021.

[5] S. Donadello, C. Clivati, A. Govoni, L. Margheriti, M. Vassallo, D. Brenda, M. Hovsepyan, E. K. Bertacco, R. Concas, F. Levi, A. Mura, A. Herrero, F. Carpentieri, and D. Calonico, "Seismic monitoring using the telecom fiber network," *Communications Earth & Environment*, vol. 5, p. 178, Apr. 2024.

[6] S. Yin, P. B. Ruffin, and F. T. S. Yu, eds., *Fiber Optic Sensors*. Boca Raton: CRC Press, 2 ed., Dec. 2017.

[7] V. V. Grishachev, "Detecting Threats of Acoustic Information Leakage Through Fiber Optic Communications," *Journal of Information Security*, vol. 3, pp. 149–155, Apr. 2012.

[8] J. H. Cole, R. L. Johnson, and P. G. Bhuta, "Fiber-optic detection of sound," *The Journal of the Acoustical Society of America*, vol. 62, pp. 1136–1138, Nov. 1977.

[9] C. R. Zamarreño, C. Martelli, R. Daciuk, G. Dutra, U. J. Dreyer, J. C. Cardozo Da Silva, I. R. Matias, and F. J. Arregui, "Distributed optical fiber microphone," in *2017 IEEE SENSORS*, pp. 1–3, Oct. 2017.

[10] H. Hao, Z. Pang, G. Wang, and B. Wang, "Indoor optical fiber eavesdropping approach and its avoidance," *Optics Express*, vol. 30, pp. 36774–36782, Sept. 2022.

[11] "Hiding an Ear in Plain Sight: On the Practicality and Implications of Acoustic Eavesdropping with Telecom Fiber Optic Cables."

[12] Z. He and Q. Liu, "Optical Fiber Distributed Acoustic Sensors: A Review," *Journal of Lightwave Technology*, vol. 39, pp. 3671–3686, June 2021.

[13] H. G. Kouiani, S. Straullu, R. Ambrosone, E. Virgillito, and V. Curri, "Covert Speech Detection via Polarization Dynamics in 10 Gbps IMDD Optical Fiber

Links," in *2025 European Conference on Optical Communications (ECOC)*, pp. 1–4, Sept. 2025.

[14] P. Dejdar, O. Mokry, M. Cizek, P. Rajmic, P. Munster, J. Schimmel, L. Pravdova, T. Horvath, and O. Cip, "Characterization of sensitivity of optical fiber cables to acoustic vibrations," *Scientific Reports*, vol. 13, p. 7068, May 2023.

[15] S. Chai, C. Guo, C. Guan, and L. Fang, "Deep Learning-Based Speech Enhancement of an Extrinsic Fabry–Perot Interferometric Fiber Acoustic Sensor System," *Sensors (Basel, Switzerland)*, vol. 23, p. 3574, Mar. 2023.

[16] S. Lapins, A. Butcher, J.-M. Kendall, T. S. Hudson, A. L. Stork, M. J. Werner, J. Gunning, and A. M. Brisbourne, "DAS-N2N: Machine learning Distributed Acoustic Sensing (DAS) signal denoising without clean data," *Geophysical Journal International*, vol. 236, pp. 1026–1041, Dec. 2023.

[17] M. van den Ende, I. Lior, J.-P. Ampuero, A. Sladen, A. Ferrari, and C. Richard, "A Self-Supervised Deep Learning Approach for Blind Denoising and Waveform Coherence Enhancement in Distributed Acoustic Sensing Data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, pp. 3371–3384, July 2023.

[18] Y. Shang, J. Yang, W. Chen, J. Yi, M. Sun, Y. Du, S. Huang, W. Zhao, S. Qu, W. Wang, L. Lv, S. Liu, Y. Zhao, and J. Ni, "Speech signal enhancement based on deep learning in distributed acoustic sensing," *Optics Express*, vol. 31, p. 4067, Jan. 2023.

[19] J.-W. Chang, K. Sun, D. Xia, X. Zhang, and F. Koushanfar, "EveGuard: Defeating Vibration-based Side-Channel Eavesdropping with Audio Adversarial Perturbations," Apr. 2025.

[20] M. R. Fernández-Ruiz, M. A. Soto, E. F. Williams, S. Martin-Lopez, Z. Zhan, M. Gonzalez-Herraez, and H. F. Martins, "Distributed acoustic sensing for seismic activity monitoring," *APL Photonics*, vol. 5, p. 030901, Mar. 2020.

[21] H. Liu, J. Ma, W. Yan, W. Liu, X. Zhang, and C. Li, "Traffic Flow Detection Using Distributed Fiber Optic Acoustic Sensing," *IEEE Access*, vol. 6, pp. 68968–68980, 2018.

[22] Z. Peng, H. Wen, J. Jian, A. Gribok, M. Wang, S. Huang, H. Liu, Z.-H. Mao, and K. P. Chen, "Identifications and classifications of human locomotion using Rayleigh-enhanced distributed fiber acoustic sensors with deep neural networks," *Scientific Reports*, vol. 10, p. 21014, Dec. 2020.

[23] Y. Duan, L. Liang, X. Tong, B. Luo, and B. Cheng, "Application of pipeline leakage detection based on distributed optical fiber acoustic sensor system and convolutional neural network," *Journal of Physics D: Applied Physics*, vol. 57, p. 105102, Dec. 2023.

[24] A. M. Markom, S. Saharudin, and M. H. Hisham, "Systematic review of fiber-optic distributed acoustic sensing: Advancements, applications, and challenges," *Optical Fiber Technology*, vol. 94, p. 104293, Nov. 2025.

[25] M.-F. Huang, M. Salemi, Y. Chen, J. Zhao, T. J. Xia, G. A. Wellbrock, Y.-K. Huang, G. Milione, E. Ip, P. Ji, T. Wang, and Y. Aono, "First Field Trial of Distributed Fiber Optical Sensing and High-Speed Communication Over an Operational Telecom Network," *Journal of Lightwave Technology*, vol. 38, pp. 75–81, Jan. 2020.

[26] J. M. Marin, I. Ashry, O. Alkhazragi, A. Trichili, T. K. Ng, and B. S. Ooi, "Simultaneous distributed acoustic sensing and communication over a two-mode fiber," *Optics Letters*, vol. 47, pp. 6321–6324, Dec. 2022.

[27] H. He, L. Jiang, Y. Pan, A. Yi, X. Zou, W. Pan, A. E. Willner, X. Fan, Z. He, and L. Yan, "Integrated sensing and communication in an optical fibre," *Light: Science & Applications*, vol. 12, p. 25, Jan. 2023.

[28] E. Ip, Y.-K. Huang, G. Wellbrock, T. Xia, M.-F. Huang, T. Wang, and Y. Aono, "Vibration Detection and Localization Using Modified Digital Coherent Telecom Transponders," *Journal of Lightwave Technology*, vol. 40, pp. 1472–1482, Mar. 2022.

[29] J. Fang, M.-F. Huang, S. Kotrla, T. J. Xia, G. A. Wellbrock, J. A. Mundt, T. Wang, and Y. Aono, "Field Trial of High-Sensitivity Forward-Transmission Sensing for Real-World Event Detection Over Live Urban Fiber Networks," *Journal of Lightwave Technology*, vol. 44, pp. 1167–1177, Feb. 2026.

[30] J. P. Gordon and H. Kogelnik, "PMD fundamentals: Polarization mode dispersion in optical fibers," *Proceedings of the National Academy of Sciences*, vol. 97, pp. 4541–4550, Apr. 2000.

[31] S. J. Savory, "Digital Coherent Optical Receivers: Algorithms and Subsystems," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 16, pp. 1164–1179, Sept. 2010.

[32] A. Mecozzi, M. Cantono, J. C. Castellanos, V. Kamalov, R. Muller, and Z. Zhan, "Polarization sensing using submarine optical cables," *Optica*, vol. 8, pp. 788–795, June 2021.

[33] S. Pellegrini, L. Minelli, L. Andrenacci, G. Rizzelli, D. Pilori, G. Bosco, L. D. Chiesa, C. Crognale, S. Piciaccia, and R. Gaudino, "Overview on the state of polarization sensing: Application scenarios and anomaly detection algorithms," *Journal of Optical Communications and Networking*, vol. 17, pp. A196–A209, Feb. 2025.

[34] C. J. Carver and X. Zhou, "Polarization sensing of network health and seismic activity over a live terrestrial fiber-optic cable," *Communications Engineering*, vol. 3, p. 91, July 2024.

[35] M. Mazur, D. Wallberg, L. Dallachiesa, E. Börjeson, R. Ryf, M. Bergroth, B. Josefsson, N. K. Fontaine, H. Chen, D. T. Neilson, J. Schröder, P. Larsson-Edefors, and M. Karlsson, "Real-Time Monitoring of Cable Break in a Live Network using a Coherent Transceiver Prototype," in *2024 Optical Fiber Communications Conference and Exhibition (OFC)*, pp. 1–3, Mar. 2024.

[36] R. A. Chipman, W.-S. T. Lam, and G. Young, *Polarized Light and Optical Systems*. Optical Sciences and Applications of Light, Boca Raton London New York: CRC Press, 2019.

[37] R. C. Jones, "A New Calculus for the Treatment of Optical SystemsI Description and Discussion of the Calculus," *Journal of the Optical Society of America*, vol. 31, p. 488, July 1941.

[38] D. Marcuse, C. Manyuk, and P. Wai, "Application of the Manakov-PMD equation to studies of signal propagation in optical fibers with randomly varying birefringence," *Journal of Lightwave Technology*, vol. 15, pp. 1735–1746, Sept. 1997.

[39] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*. Pearson Prentice Hall, 4th ed., 2006.

[40] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, pp. 411–430, June 2000.

[41] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[42] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?." https://arxiv.org/abs/1811.02508v1, Nov. 2018.

[43] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," Apr. 2018.

[44] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is All You Need in Speech Separation," Mar. 2021.

[45] W. Jiang, K. Yu, and F. Wen, "Unsupervised Speech Enhancement Using Optimal Transport and Speech Presence Probability," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4445–4455, 2024.

[46] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2125–2136, Sept. 2011.

[47] Y. Wang, A. Narayanan, and D. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849–1858, Dec. 2014.

[48] T. Hussain, M. Diyan, M. Gogate, K. Dashtipour, A. Adeel, Y. Tsao, and A. Hussain, "A Novel Speech Intelligibility Enhancement Model based on CanonicalCorrelation and Deep Learning," Feb. 2022.

[49] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, pp. 749–752 vol.2, May 2001.

[50] P. C. Loizou, *Speech Enhancement: Theory and Practice*. USA: CRC Press, Inc., 2nd ed., 2013.

[51] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.

[52] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 7–19, Jan. 2015.

[53] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," June 2018.

[54] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement," June 2021.

[55] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Processing Letters*, vol. 21, pp. 65–68, Jan. 2014.

[56] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-r. Mohamed, G. Dahl, and B. Ramabhadran, "Deep Convolutional Neural Networks for Large-scale Speech Tasks," *Neural Networks*, vol. 64, pp. 39–48, Apr. 2015.

[57] S. R. Park and J. Lee, "A Fully Convolutional Neural Network for Speech Enhancement," Sept. 2016.

[58] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," Apr. 2016.

[59] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017.

[60] C. S. Monteiro, T. D. Ferreira, and N. A. Silva, "High-Precision Acoustic Event Monitoring in Single-Mode Fibers Using Fisher Information," June 2025.

[61] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." https://arxiv.org/abs/1412.6980v9, Dec. 2014.

[62] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, A. Kumar, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt, and Y. Qian, "URGENT Challenge: Universality, Robustness, and Generalizability For Speech Enhancement," in *Interspeech 2024*, pp. 4868–4872, Sept. 2024.

[63] W. Li, R. Spolaor, C. Luo, Y. Sun, H. Chen, G. Zhang, Y. Yang, X. Cheng, and P. Hu, "Acoustic Eavesdropping From Sound-Induced Vibrations With Multi-Antenna mmWave Radar," *IEEE Transactions on Mobile Computing*, vol. 24, pp. 7693–7708, Aug. 2025.

[64] J. M. Martin-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A Deep Learning Loss Function Based on the Perceptual Evaluation of the Speech Quality," *IEEE Signal Processing Letters*, vol. 25, pp. 1680–1684, Nov. 2018.

[65] S.-W. Fu, C.-F. Liao, and Y. Tsao, "Learning with Learned Loss Function: Speech Enhancement with Quality-Net to Improve Perceptual Evaluation of Speech Quality," *IEEE Signal Processing Letters*, vol. 27, pp. 26–30, 2020.

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY