Glimpse Proportion Maximization for Speech Intelligibility Enhancement

-

Master's Thesis Report

Valentina Correa Bejarano

Aalborg University
Computer Engineering (AI, Vision and Sound)

Computer Engineering AI, Vision and Sound
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Glimpse Proportion Maximization for Speech Intelligibility Enhancement

Theme:
-

Project Period:
Spring Semester 2025

Project Group:
XXX

Participant(s):
Valentina Correa Bejarano

Supervisor(s):
Filippo Villani (AAU)
Jesper Jensen (AAU and Oticon)
Jan Østergaard (AAU)

Copies: -

Page Numbers: 34

Date of Completion:
August 25, 2025

Abstract:

This thesis investigates the use of Glimpse Proportion (GP) maximisation as an optimisation objective for Near-End Listening Enhancement (NELE). Unlike conventional methods based on the Speech Intelligibility Index (SII) or related metrics, the proposed approach employs a differentiable formulation of GP, enabling gradient-based optimisation under energy-preservation constraints. The method, introduced as GlimpseP, applies frequency-dependent, time-invariant spectral weighting and is evaluated across multiple datasets (DANTALE II, AEMST, TIMIT) and noise conditions (stationary and competing speaker). Results show consistent improvements in objective intelligibility metrics, with particular advantages in fluctuating noise where glimpsing cues are most perceptually relevant. Compared to established baselines such as FractileASII, the proposed method demonstrates comparable or superior performance while maintaining robustness across conditions. These findings confirm the potential of GP as a perceptually grounded optimisation target for NELE.

Contents

Preface

I would like to thank my supervisors, Filippo Villani, Jesper Jensen and Jan Østergaard for their continuous support throughout this project.
The repository containing the implementation can be found using the following link: `https://github.com/filippovillani/NELE-playground/tree/feature/glimpseOpt_added`.

Aalborg University, August 25, 2025

Valentina Correa Bejarano
<vbejar23@student.aau.dk>

Chapter 1

Introduction

1.1 Background

Being able to understand speech in noisy environments is a fundamental part of human communication. Yet, everyday acoustic scenes often contain multiple competing sound sources: background conversations in a café, traffic noise on a busy street, or environmental sounds in urban spaces. For listeners with hearing impairments, or even for normal-hearing listeners in particularly challenging conditions, speech intelligibility can degrade significantly. This challenge has led to a long tradition of research into models that predict intelligibility and into systems that can enhance it.

A particular focus has been on *Near-End Listening Enhancement* (NELE). Unlike conventional speech enhancement, which aims to recover a clean signal from a noisy mixture, NELE modifies the playback signal at the listener's side to make speech more intelligible under local noise conditions. This makes the problem both practically relevant and technically challenging, since the system cannot rely on knowledge of the exact acoustic environment. Over the past two decades, NELE has evolved from simple analytic solutions to more perceptually motivated and even machine-learning-based strategies. However, most approaches have continued to optimise for predictors such as the Speech Intelligibility Index (SII) or its derivatives, which are reliable in stationary noise but less accurate in more complex listening environments.

1.2 Problem Statement

While traditional metrics like SII and STOI have provided valuable tools for optimisation, they do not fully capture the way humans perceive speech in realistic noise. In particular, in situations with *fluctuating noise* such as competing speakers, listeners rely on brief moments where parts of the target speech become audible. This

process, described by *glimpsing theory*, motivates the use of the *Glimpse Proportion* (GP) as an intelligibility predictor. GP has been shown to correlate strongly with listening test outcomes in dynamic noise, but despite this evidence it has rarely been used as a direct optimisation target in NELE systems. Existing approaches therefore miss the opportunity to align system design more closely with perceptual principles.

1.3 Aims and Scope

This thesis addresses the above gap by exploring Glimpse Proportion as an optimisation objective for Near-End Listening Enhancement. A differentiable formulation of GP is employed, making it suitable for gradient-based optimisation while energy-preservation constraints. The approach is evaluated across multiple speech corpora and noise conditions, and compared against established baselines that optimise for SII-based metrics.

The central aim is to investigate whether optimising directly for GP can deliver consistent intelligibility improvements, particularly in non-stationary noise where traditional approaches are less reliable. More broadly, the project seeks to connect perceptual theory with practical signal processing, showing that models inspired by how humans listen in noise can be translated into effective algorithmic solutions.

Chapter 2

Literature Review

The sections that follow review State-of-the-Art knowledge on Speech intelligibility predictors, Glimpse Proportion theory, Near-End Listening Enhancement systems, and optimisation methods, and they conclude with a gap analysis that motivates the present study.

2.1 Speech Intelligibility Predictors

The accurate prediction of speech intelligibility in noisy environments has long been of important focus of speech and hearing science. Early standardised models, the Speech Intelligibility Index SII [1] [20] and the Speech Transmission Index STI [26], offer signal-based metrics that remain part of international standards. However, because they rely on long-term statistics, their accuracy declines when noise fluctuates and temporal information becomes crucial.

To overcome these drawbacks, newer measures such as the Short-Time Objective Intelligibility STOI [28] and its extension ESTOI[14], employ short-time envelope correlation analyses between reference and processed speech, delivering more robust predictions in non-stationary noise. Metrics designed for hearing-aid purposes,HASPI[15] and HASQI [16], further refine predictions by modelling auditory-periphery loss.

While these predictors represent significant progress, their performance remains inconsistent in contexts dealing with Near-End Listening Enhancement (NELE) [2] or reverberation. This has motivated research into models based on perceptual principles. For instance, models employing glimpsing theory and spectro-temporal approaches [6, 7].

2.2 Glimpse Proportion Theory

The Glimpse Proportion (GP) concept, originally introduced by [3], showed that intelligibility can be predicted from the fraction of time–frequency elements in which the local signal-to-noise ratio (SNR) exceeds a fixed threshold. Listeners effectively glimpse speech fragments in these regions and integrate them into a coherent representation.

Following studies have demonstrated a strong correlation between GP and actual intelligibility measured in listening tests, particularly in non-static noisy and reverberant environments. [3]Extensions of GP have been proposed:

- **High-Energy Glimpse Proportion (HEGP):** assigns more weight to high-energy glimpses, under the assumption that these contribute disproportionately to intelligibility [29]
- **Spectro-Temporal Glimpsing Index (STGI)**: integrates GP with modulation transfer characteristics, thereby capturing both temporal modulation and masking effects.

These developments establish glimpse-based metrics as perceptually grounded predictors that more closely mirror human performance than earlier models like SII or STI.

2.3 NELE Systems

The Near-End Listening Enhancement (NELE) is a technique that improves speech intelligibility in noisy environments by adaptively preprocessing speech signals based on noise estimates. Unlike conventional speech enhancement systems, which aim to recover a clean speech signal from noisy input. NELE systems aim to modify the playback of speech signals at the listener's side. It operates without prior knowledge of the specific acoustic environment, making the problem inherently more challenging.

Research into NELE was initiated by Sauert and Vary [23], who proposed spectral power allocation optimised under perceptual constraints, yielding intelligibility gains without raising overall signal power. Their later work incorporated explicit Speech Intelligibility Index (SII)-based optimisation while respecting audio power limitations [22], followed by recursive closed-form solutions enabling efficient real-time implementation [25]. Extensions addressed band-limited noise scenarios [24]. In parallel, Niermann introduced alternative formulations: a time-domain linear prediction approach offering extremely low-latency enhancement [19], noise-inverse speech shaping [18], and joint enhancement with far-end noise reduction [17]. These developments compared time-domain and frequency-domain solutions in terms of performance and computational cost.

More recent research has broadened the NELE landscape. Taal et al. in [27] approached NELE as an energy-constrained maximisation of an SII proxy, producing a closed-form Linear Time-Invariant (LTI) filter that shifts energy away from sub–15 dB SNR bands toward more informative regions. Villani et al. [32] extended this idea with a gammatone-based, noise-robust LTI solution whose closed-form, band-wise gains preserve overall energy while adapting to the long-term speech- and noise-spectral statistics. Fuglsig et al. [9] pursued a "minimum processing" method, minimising signal modification while maintaining intelligibility gains. Chermaz and King [2] adopted a sound engineering perspective, framing NELE design within practical perceptual and audio engineering constraints. Complementary perceptual optimisation strategies have also emerged: Crespo and Hendriks [4, 5] introduced reinforcement methods grounded in perceptual distortion measures, later refined by Hendriks et al. [12] through short-time SII-based optimisation accounting for additive noise and reverberation. These studies collectively demonstrate NELE's evolution from optimization driven by SII to diverse algorithmic, perceptual, and engineering-based strategies, while highlighting the persistent challenge of balancing enhancement in intelligibility with maintaining naturalness, robustness, and computational feasibility.

2.4 Optimisation Techniques in Audio Processing

Speech intelligibility enhancement is often treated as an optimisation problem based on perceptual metrics. Early approaches focused on simple analytic formulations: for example, Taal et al. in [27] and Hendriks et al in [11] maximised approximations of the Speech Intelligibility Index (SII) under power constraints, leading to fast, closed-form frequency shaping methods, at the cost of using coarse perceptual models. Later research introduced more computationally demanding iterative methods, such as sequential quadratic programming for spectro-temporal weights [8] or adaptive redistribution of spectral gains from SNR estimates [34], which delivered measurable improvements in intelligibility and STOI.

The latest developments move towards machine learning integration, where differentiable versions of perceptual measures, such as a smoothed Glimpse Proportion [31], are used as optimisation objectives for gradient-based methods. This allows speech processing algorithms to be trained directly within modern Machine Learning frameworks. Overall, the field balances a trade-off: lightweight closed-form or simple iterative methods are practical for real-time NELE applications, while gradient-based approaches offer potentially stronger gains but at a higher computational cost.

2.5 Gap Analysis

Research on Near-End Listening Enhancement (NELE) has shown many ways to make speech more intelligible in noisy environments, but some gaps remain. Most methods rely on traditional metrics like SII or STOI, which are useful but do not fully reflect how humans actually perceive speech in noise. Measures based on Glimpse Proportion (GP), which capture the listener's ability to pick out speech fragments, correlate better with intelligibility but have rarely been used as direct optimisation targets in NELE.

Existing NELE systems also face trade-offs. Some approaches are computationally heavy, making real-time use difficult, while others use restrictive models that may not work well in all conditions. Evaluations often focus on limited noise types, so it is unclear how these systems perform in everyday noisy situations with multiple noise sources. Additionally, few studies balance improvement with maintaining natural-sounding speech.

This project addresses these gaps by exploring the use of Glimpse Proportion as an optimisation objective for NELE. By benefitting from a differentiable GP version, to directly optimise for. This approach aims to combine the perceptual grounding of glimpse-based models with the flexibility of gradient-based optimisation, offering a potential alternative to state-of-the-art NELE systems that is both theoretically motivated and practically applicable.

Chapter 3

Methodology

As stated the goal of NELE systems is to maintain speech intelligibility and acceptable quality even in the presence of unpredictable environmental noise at the listener's location [12]. Along the same path, Glimpse Proportion, a well-established and reliable metric of intelligibility [29, 30], has been used as an objective function to optimise for intelligibility [30]. This work aims to make use of Glimpse Proportion maximisation to enhance speech intelligibility and explore its potential for achieving results comparable to state-of-the-art NELE systems.

3.1 Signal Model

In the Near-End Listening Enhancement (NELE) framework, the observed acoustic scene is modelled as a linear superposition of a clean speech signal and an interfering noise signal. Let $x(t)$ denote the clean speech waveform, and $v(t)$ the additive noise waveform. The observed signal $y(t)$ at the listener's ear is then given by

$$y(t) = x(t) + v(t). \tag{3.1}$$

This additive model assumes linear propagation of sound sources without significant non-linear distortions or reverberation. This is a commonly adopted assumption in speech-enhancement and intelligibility research due to its tractability and adequacy in many real-world scenarios

Time–Frequency Representation

To analyse the signals in the time–frequency domain, the Short-Time Discrete Fourier Transform (STDFT) is applied. For an input signal $x(t)$, its STDFT representation is:

$$S(k,l) = \sum_{n=0}^{N-1} s(n+lH)\, w(n)\, \exp\left(-j2\pi \frac{kn}{N}\right), \tag{3.2}$$

where k indexes the frequency bins, l the time frames, N is the FFT length, H the hop size, and $w(n)$ is the analysis window.

Perceptual Filterbank Representation

Although the STDFT provides uniform spectral partitioning, human auditory perception is non-uniform across frequency. The cochlea (part of the inner ear that converts sound vibrations into electrical signals for the brain to interpret) exhibits filters of roughly constant Equivalent Rectangular Bandwidth (ERB). To emulate this, the STDFT magnitude spectra are passed through a Gammatone filterbank, a standard auditory model offering more perceptually relevant frequency resolution.

Originally proposed by Schofield and Patterson et al., the gammatone filter approximates human auditory filters well in both amplitude and phase characteristics and has since become a foundational tool in auditory modelling Its implementation has been widely adopted in auditory modelling toolboxes, emphasising both biological fidelity and computational efficiency

3.2 Spectral Weighting Strategy

To improve speech intelligibility in noisy environments, a spectral weighting strategy is implemented. This strategy applies time-invariant, band-specific gains to the time-frequency (T-F) representation of the received signal to optimise the components that are crucial for perception.

Definition of Spectral Weights

Let $g(k)$ represent the gain applied to the k-th frequency channel. Given the Time–Frequency representation $S(k,l)$ and $V(k,l)$, obtained from the Gammatone filter bank decomposition of the clean speech signal $x(t)$ and noise signal $v(t)$, the enhanced representation is described as:

$$\hat{S}(k,l) = g(k) \cdot S(k,l). \tag{3.3}$$

The gain $g(k)$ is constant across time frames l for each frequency channel k, meaning that gains are frequency-dependent but time-invariant. Such an approach simplifies optimisation and avoids excessive time-based fluctuations that might introduce temporal artefacts.

Perceptual Motivation

The gain vector $g(k)$ is designed to match human hearing by functioning within an auditory-inspired filterbank model, specifically using Gammatone channels. This approach ensures that vital frequency areas for speech intelligibility receive differentiated treatment. This non-uniform spectral weighting follows auditory masking, where the weights are adjusted according to the human sensitivity across frequency bands.

Optimisation Relation with Glimpse Proportion

In this project, the spectral weighting method is directly linked to optimising for the Glimpse Proportion (GP) metric. By adjusting the spectral gain for each frequency band, the optimisation algorithm selectively boosts or reduces specific frequency components to maximise the GP. The GP measures the fraction of time-frequency units where the local signal-to-noise ratio surpasses a defined threshold, representing clear glimpses of the speech signal. This perceptually driven spectral weighting enhances speech components that contribute most to intelligibility in noisy conditions while minimising distortion and preserving perceptual relevance.

3.3 Objective Function

The core of this work is the notion of Glimpse Proportion, a perceptual metric that quantifies how many time–frequency regions glimpses of the speech signal exceed the noise level by a certain threshold, and which correlates strongly with intelligibility [3]Formally, the Glimpse proportion is defined as:

$$\mathrm{GP} = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \{\mathrm{SNR}(k,l) > \theta\}, \tag{3.4}$$

where θ is the SNR threshold for a glimpse.

The definition 3.4 shows the binary nature of the metric, hence making it non-differentiable. In order to apply optimisation, there is a need for a differentiable version of glimpse proportion. Following prior work in the context of intelligibility-oriented enhancement [31]a sigmoid function is applied to the SNR, to maintain a smooth function:

$$\sigma\left(\frac{\mathrm{SNR}(k,l)-\theta}{\beta}\right) = \frac{1}{1+\exp\left(-\frac{\mathrm{SNR}(k,l)-\theta}{\beta}\right)}, \tag{3.5}$$

where β is a scale parameter that controls the sharpness of the transition, how hard the threshold is.

Thus, the differentiable GP objective becomes:

$$\mathrm{GP} = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \sigma \left(\frac{\mathrm{SNR}(k,l) - \theta}{\beta} \right). \tag{3.6}$$

This formulation approximates the binary GP, enabling the computation of gradients with respect to optimisation variables such as the frequency band gains $g(k)$.

The optimisation goal is:

$$\max_{g(k)} \ \mathrm{GP}(g(k); S(k,l), V(k,l))$$

3.4 Constraints

In addition to maximising the differentiable Glimpse Proportion (GP), the optimisation problem requires constraints to ensure that the resulting enhanced signal remains perceptually natural, physically consistent, and comparable in energy to the original clean speech.

Energy Preservation

A key constraint is that the total energy of the enhanced signal $\hat{y}(t)$ should match that of the clean speech $x(t)$. Without such a constraint, the optimiser could trivially increase GP by uniformly amplifying all frequency bands, leading to loudness mismatches and unrealistic listening conditions. Energy preservation is formulated as

$$\sum_{k=1}^{K} g(k)^2 \sum_{l} |S(k,l)|^2 = \sum_{k=1}^{K} \sum_{l} |S(k,l)|^2 \tag{3.7}$$

where $S(k,l)$ and $\hat{S}(k,l)$ denote the clean and enhanced speech representations in the time–frequency domain.

3.5 Optimisation Strategy

The objective function defined in Section 3.3 and the constraint in Section 3.4 together form a non-linear constrained optimisation problem. The optimisation variable is the gain vector $g = [g(1), g(2), \ldots, g(K)]$ applied to the frequency channels.

Problem Formulation

The problem is summarised as:

$$\max_{g(k)} \ \mathrm{GP}(g(k); S(k,l), V(k,l))$$

$$\text{s.t.} \sum_{k=1}^{K} g(k)^2 \sum_{l} |S(k,l)|^2 = \sum_{k=1}^{K} \sum_{l} |S(k,l)|^2$$

This formulation is non-linear in both the objective and the constraint, requiring the use of numerical optimisation techniques.

Optimization Method

The **Sequential Least Squares Programming (SLSQP)** algorithm, as provided by the `scipy.optimize.minimize` package, is used to solve the constrained optimisation problem in this work. SLSQP is a gradient-based optimisation method that belongs to the family of sequential quadratic programming (SQP) techniques. In each iteration, SLSQP solves a quadratic programming sub-problem that approximates the original non-linear objective and linearises the constraints, making it highly effective for problems where both the objective and constraints are differentiable but potentially non-linear. SLSQP can directly handle both equality and inequality constraints as well as variable bounds. The algorithm iteratively refines its solution by computing search directions and step sizes, ultimately converging to a local optimum that satisfies the Karush–Kuhn–Tucker (KKT) conditions, assuming sufficient regularity.

Initialization and Convergence

The optimisation is initialised with a flat gain vector, $g(k) = 1 \; \forall k$, corresponding to no initial enhancement. This setup guarantees that the optimiser starts from the unprocessed state, avoiding bias towards particular spectral regions. The iterative SLSQP algorithm refines the gains using gradient information derived from the differentiable GP formulation. Convergence is evaluated by examining changes in the objective function and constraint fulfilment, with early stopping applied if improvements fall below a set threshold.

Application of Optimized Gains

Once the optimal gain vector $\hat{g}(k)$ is obtained, it is applied to the time–frequency representation:

$$\hat{S}(k,l) = \hat{g}(k) \cdot S(k,l). \tag{3.8}$$

This results in an enhanced time–frequency representation with spectrally weighted glimpses.

Resynthesis

The enhanced time–frequency representation is transformed back into the time domain using a filterbank-based synthesis approach consistent with the Gammatone

analysis This ensures proper reconstruction within the constraints of the filterbank implementation and preserves the perceptual alignment of the auditory-inspired analysis.

3.6 Baselines

To meaningfully evaluate the proposed method, its performance must be compared against appropriate baseline NELE systems. Speech signals are then observed under the following conditions: unprocessed, using OptimalASII, using FractileASII, and the proposed method.

Unprocessed Speech

The most basic reference point is the *unmodified speech* signal, $y(t) = x(t) + v(t)$, which remains unchanged. This establishes a performance floor and guarantees that any gains in intelligibility result from the enhancement technique itself, rather than from advantageous testing conditions.

OptimalASII

OptimalASII [27], demonstrates substantial intelligibility improvements in stationary noise conditions, particularly in speech-shaped noise scenarios, with improvements most pronounced at lower SNR conditions, for example, improving intelligibility from 17.3% to 50.6% words correct in controlled listening tests. However, the method exhibits limitations when applied to fluctuating noise sources, such as competing speaker scenarios, where, despite improved objective SII predictions, actual listening test results showed decreased intelligibility performance due to the SII's reduced reliability as a predictor for non-stationary noise. This baseline provides an important comparison point for the proposed Glimpse Proportion optimisation approach, as both methods share fundamental characteristics, including time-invariant spectral processing, energy conservation constraint, and perceptually motivated optimisation. The proposed method's use of Glimpse Proportion could potentially address OptimalASII's limitations in competing speaker scenarios.

FractalASII

In FractalASII [32], there are extensive evaluations in both stationary and fluctuating noise conditions that demonstrate the method yields significant intelligibility gains compared to unprocessed speech and other LTI-based methods like OptimalASII. In stationary white and speech-shaped noise, the method achieves up

to a 35% increase in word recognition scores at low SNRs, while in more complex, non-stationary noise scenarios, it consistently outperforms traditional spectral subtraction and Wiener filtering approaches. However, because the filter is time-invariant, it cannot track rapid noise fluctuations as effectively as adaptive or time-varying methods, leading to reduced performance in highly dynamic acoustic environments. Nonetheless, its low computational complexity and robustness to noise estimation errors make it a practical baseline for comparison against more sophisticated, time-variant enhancement algorithms.

3.7 Signal Resources

This section presents the speech and noise datasets used in the study, along with the objective metrics and conditions applied to evaluate intelligibility.

3.7.1 DANTAlE II

The Danish Matrix Test is a standardised corpus used to assess speech intelligibility by determining speech reception thresholds (SRTs) in noisy conditions. This test is part of the Matrix Sentence Tests family, which has been adopted in over 20 languages to facilitate cross-linguistic studies on intelligibility [33]. Each sentence follows the structure:

Name Verb Numeral Adjective Object

Index	Name	Verb	Numeral	Adjective	Object
0	Anders	**ejer**	ti	gamle	jakker
1	Birgit	havde	*fem*	røde	**kasser**
2	Ingrid	ser	syv	pæne	ringe
3	Ulla	købte	tre	*nye*	blomster
4	Niels	vandt	seks	fine	skabe
5	Kirsten	får	tolv	flotte	*masker*
6	Henning	solgte	**otte**	smukke	*biler*
7	Per	låner	fjorten	store	*huse*
8	**Linda**	valgte	ni	**hvide**	gaver
9	*Michael*	*finder*	tyve	sjove	planter

Table 3.1: Sentence material (DANTALE II)

Each category has ten different options, see table 3.1. The recorded sentences were formed by randomly choosing one of the ten alternatives for each word. This approach guarantees that the sentences are grammatically correct, yet their content remains unpredictable. Additionally, to form the final dataset, the recorded

sentences were segmented into individual words so the sentences would maintain realistic speech characteristics, especially co-articulation; therefore, listening experiments would be valid and reliable.

3.7.2 AEMST

The American English Matrix Sentence Test is the English-language equivalent of the Matrix test family. Like DANTALE, it employs a 50-word lexicon organised into five categories: Name, Verb, Numeral, Adjective, Object. Sentences follow the same fixed grammatical structure, ensuring comparability across languages.

Index	Name	Verb	Number	Adjective	Noun
0	Peter	Got	Three	Large	Desks
1	Kathy	Sees	Nine	Small	Chairs
2	Lucy	Brought	Seven	Old	Tables
3	Alan	Gives	Eight	Dark	Toys
4	Rachel	Sold	Four	Heavy	Spoons
5	William	Prefers	Nineteen	Green	Windows
6	Steven	Has	Two	Cheap	Sofas
7	Thomas	Kept	Fifteen	Pretty	Rings
8	Doris	Ordered	Twelve	Red	Flowers
9	Nina	Wants	Sixty	White	Houses

Table 3.2: Sentence material from matrix test (AEMST)

3.7.3 TIMIT

The TIMIT Acoustic-Phonetic Continuous Speech Corpus [10] was created by DARPA and distributed by the Linguistic Data Consortium (LDC). TIMIT differs from Matrix tests, which feature sentences with no semantic predictability. Instead, TIMIT comprises phonetically diverse read sentences intended to contain the entire range of phonemes found in American English.

TIMIT contains recordings from 630 speakers of eight major American English dialect regions. Each speaker reads ten sentences, producing a total of 6300 sentences. Even though TIMIT sentences make sense together as a normal sentence, they are still read aloud by the speakers, not as natural or spontaneous conversation.

It includes different types of sentences for speakers to read, see Table 3.4. Dialect sentences, made of two special sentences created to highlight different dialects; phonetically-compact sentences, with 450 sentences designed to cover many important pairs of speech sounds; and 1890 phonetically-diverse sentences picked from existing text sources, such as books and plays.

Dialect Region	#Male	#Female	Total
New England	31 (63%)	18 (27%)	49 (8%)
Northern	71 (70%)	31 (30%)	102 (16%)
North Midland	79 (67%)	23 (23%)	102 (16%)
South Midland	69 (69%)	31 (31%)	100 (16%)
Southern	62 (63%)	36 (37%)	98 (16%)
New York City	30 (65%)	16 (35%)	46 (7%)
Western	74 (74%)	26 (26%)	100 (16%)
Army Brat (moved)	22 (67%)	11 (33%)	33 (5%)
Total	438 (70%)	192 (30%)	630 (100%)

Table 3.3: Dialect distribution of TIMIT speakers

Sentence Type	#Sentences	#Speakers	Total	#Sentences/Speaker
Dialect	2	630	1260	2
Compact	450	7	3150	5
Diverse	1890	1	1890	3
Total	2342		6300	10

Table 3.4: TIMIT speech material

3.7.4 ISTS

Unlike typical speech recordings that have words and meaning, the International Speech Test Signal [13] is made to sound like real speech but cannot be understood as sentences or words. ISTS was created by taking short pieces of real speech recorded from six different female speakers speaking various languages: **English, Arabic, Chinese, French, German, Spanish**. These pieces were cut and smoothly stitched together to make one continuous sound. Even though it is not understandable speech, ISTS preserves long-term average speech spectrum attributes, keeping the natural rhythm, tone, and pattern of real conversations. In the current project, this dataset is used as a competitor speaker noise due to its characteristics that can create a realistic scene, similar to being in a noisy place with multiple people talking.

3.7.5 UrbanSound8k

UrbanSound8K is a collection of sounds recorded from real urban environments [21]. It has 8,732 short audio clips. The sounds go from steady noise to very variable noise and are grouped into 10 categories like **air conditioner, car horn, children playing, dog barking, drilling, engine idling, gunshot, jackhammer, siren, and street music**. These were chosen because they represent common city

Language	Age	From	Fundamental frequency [Hz]
Arabic	37	Oran, Algeria	204
English	29	USA and Germany	194
French	25	Nantes, West France	201
German	33	Oldenburg, Lower Saxony	205
Mandarin	26	Henan, Middle-East China	208
Spanish	26	Zamora, Castile and Leon	207

Table 3.5: Properties of the six selected female speakers: age, provenance, and median fundamental frequency.

noises that can interfere with hearing speech.

3.8 Evaluation Metrics and Conditions

The performance of the current enhancement method is assessed using established objective intelligibility measures, multiple noise conditions, and diverse speech datasets. This ensures that results are both reproducible and comparable with prior work in the field.

Objective Intelligibility Metrics

Three intelligibility-oriented metrics are employed:

- **Glimpse Proportion (GP):** Measures the proportion of time–frequency tiles where the local SNR exceeds a threshold. GP has been shown to correlate strongly with human intelligibility in noise.
- **High-Energy Glimpse Proportion (HEGP):** A variant of GP that emphasises high-energy regions of the speech signal, under the assumption that energetic glimpses contribute more to intelligibility
- **Speech Transmission Index–Glimpse Index (STGI):** A hybrid metric combining the Speech Transmission Index with glimpse analysis. STGI captures both modulation transfer characteristics and masking effects, offering a more comprehensive measure of intelligibility in complex noise

Together, these metrics provide a robust evaluation of both raw glimpse availability and perceptual relevance, aligning the evaluation with the theoretical motivation of this project.

Noise Characteristics

Two types of noise signals are considered in this work:

- **Speech-Shaped Noise (SSN):** A stationary noise whose spectrum matches that of speech. It provides controlled conditions, particularly when spectral overlap is high.
- **Competing-Speaker Noise (CP):** Non-stationary background noise with varying intensity and spectral content—representing realistic listening environments.

The combination of SSN and CP enables evaluation across controlled conditions, reflecting the range of listening challenges users may encounter. In experimental studies, SSN is often used to simulate steady-state masking, while real-life noises such as competing speech offer more complex intelligibility challenges.

Three different SSN samples are used. Each matches the long-term average spectrum of the corresponding speech material (DANTALE, AEMST, TIMIT). As for competing speaker noise, two samples are taken from ISTS dataset[13]. Finally, for the environmental, UrbanSound8k dataset [21]. Experiments are conducted at multiple signal-to-noise ratios (SNRs) to evaluate robustness across varying degrees of noise interference.

Dataset conditions

To ensure generality across languages, speakers, and recording conditions, three datasets are used. DANTALE, AEMST and TIMIT see Section 3.7.

Evaluation Protocol

For each condition (noise type, SNR, and dataset), intelligibility metrics (GP, HEGP, STGI) are computed for:

1. The *unprocessed speech signal* $x(t)$.
2. Enhanced signals obtained with alternative baselines (Section 3.6).
3. The proposed GP-optimized enhancement method.

This protocol is thought to guarantee equal treatment of all methods, allowing for a quantitative evaluation of how effectively the proposed system enhances intelligibility.

Chapter 4

Results

4.0.1 Model performance comparison under gp metric

Since the optimization is performed on the differentiable version of the GP, the GlimpseP model is expected to perform particularly well when evaluated using the GP metric. Table 4.1 presents the results for three models: two baselines, Unprocessed and FractileASSI, and the proposed GlimpseP method. The evaluation was conducted on the Dantale II, AEMST, and TIMIT datasets, using SSN and ISTS as competing speaker (CS) noise types.

Speech	Method	SSN				CS			
		-15 dB	-10 dB	-5 dB	0 dB	-15 dB	-10 dB	-5 dB	0 dB
Dantale	Unprocessed	0.18069	1.296	5.027	11.860	22.939	30.028	38.953	48.692
	FractileASII	8.587	15.265	22.659	31.527	35.328	42.524	51.127	60.050
	GlimpseP	**6.787**	**17.254**	**26.401**	**35.923**	**39.646**	**47.054**	**55.146**	**63.192**
AEMST	Unprocessed	1.457	4.12	9.35	17.79	27.215	35.010	43.278	52.907
	FractileASII	8.566	14.314	20.828	28.751	33.539	41.114	49.037	57.85
	GlimpseP	**8.8480**	**15.666**	**23.351**	**32.208**	**37.606**	**45.585**	**53.225**	**61.533**
TIMIT	Unprocessed	1.135	3.271	7.107	12.987	21.988	29.889	37.440	45.616
	FractileASII	7.417	13.133	20.069	28.115	32.101	40.105	47.973	56.486
	GlimpseP	**6.9668**	**13.694**	**21.542**	**30.217**	**35.529**	**43.856**	**51.657**	**59.383**

Table 4.1: Performance evaluated with glimpse propotion metric for Unprocessed, FractileASII, and GlimpseP method, for different datasets and noise conditions.

The proposed method consistently demonstrates strong performance across all datasets, noise types, and SNR levels. Even at extreme negative SNRs; for instance,

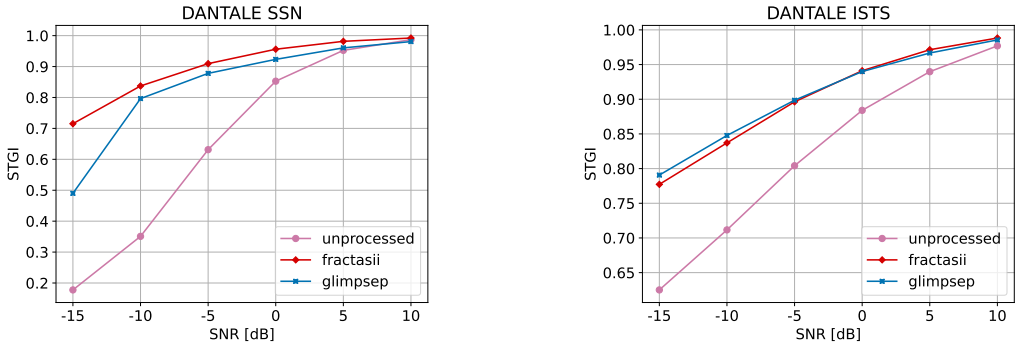
on **-15 dB SSN for DANTALE** GlimpseP achieves substantial GP values **6.787**, significantly higher than unprocessed speech and comparable to or slightly better than FractileASII. At higher SNRs, the advantages of GlimpseP become more pronounced; for example, at **-5 dB SSN for DANTALE**, GP reaches **26.401**, exceeding the FractileASII method by over 4 points. GlimpseP shows a strong advantage over both baselines, especially in highly masked conditions, indicating its effectiveness in extreme noise.While FractileASII closes the gap slightly, GlimpseP still maintains a clear improvement, demonstrating robustness across varying noise levels. Improvements are consistent for both SSN and CS noise types, suggesting the method generalises across different masking characteristics. Even in the case of TIMIT, where GlimpseP generally shows slightly lower GP than AEMST and DANTALE, likely due to variations in speech content and recording conditions, but GlimpseP still improves intelligibility consistently.

4.0.2 Model performance comparison under STGI and HEGP

The same evaluation as in Section 4.0.1 was performed, this time the analysis being under STGI and HEGP for each Dataset.

Dantale II

Figures 4.1 and 4.2 display the performance of Dantale II when evaluating the methods under STGI and HEGP



(a) Speech-Signal-Noise

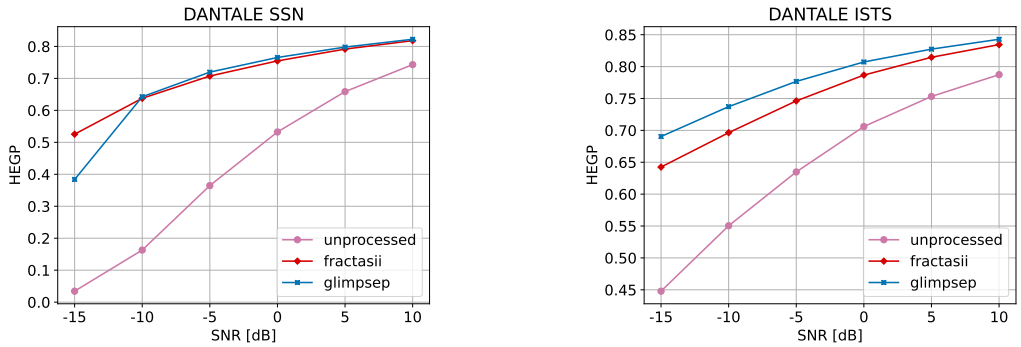
(b) Competing speaker

Figure 4.1: Performance of baselines and proposed model using Dantale II speech signal under **STGI** metric

Performance of GlimpseP evaluated on STGI in SSN (see Figure4.1a) is not as good compared to Section 4.0.1. A reason for this result could be the nature of SSN, spectrally and temporally dense. Such masking signal, challenges the process of optimising glimpses. In contrast, when the noise is a Competing Speaker,

Figure 4.1b, GlimpseP performance improves and even surpasses FractileASII. The attribution to this outcome might be to CP fluctuation characteristics, which again is a good scene for a model like GlimpseP.

Figure 4.2a, displays the a sort of expected performance of GlimpseP under HEGP. Due to the bases of the metric. It weights glimpses by their energy contribution in the speech signal. Compared to STGI, HEGP is not as sensitive to modulation chnages. Hence, the better results for GlimpseP in Figure 4.2b.

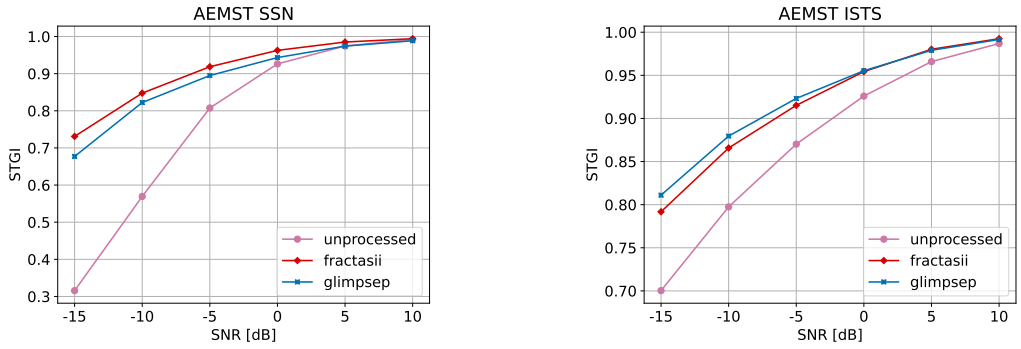


(a) Speech-Signal-Noise **(b)** Competing speaker

Figure 4.2: Performance of baselines and proposed model using Dantale II speech signal under **HEGP** metric

AEMST

Figures 4.3 and 4.4 display the performance of AEMST when evaluating the methods under STGI and HEGP, respectively.



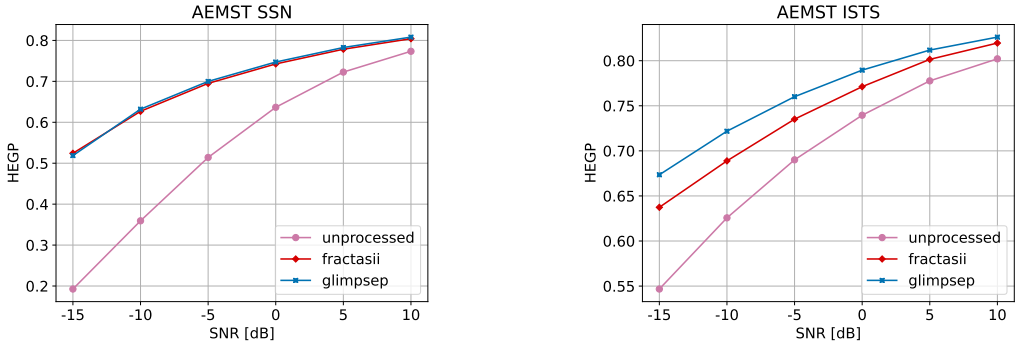
(a) Speech-Signal-Noise **(b)** Competing speaker

Figure 4.3: Performance of baselines and proposed model using AEMST speech signal under **STGI** metric

When evaluating AEMST under the STGI metric, the pattern is similar to that

observed with DANTALE II. In stationary SSN, GlimpseP performs comparably to FractileASII across SNRs, with a slight disadvantage at mid-range SNRs. However, in competing-speaker noise the advantage of GlimpseP becomes clearer, as it consistently outperforms both the baseline and the unprocessed condition. This suggests that GlimpseP is more effective at exploiting the temporal fluctuations of competing speech, which aligns with its perceptual motivation.

The HEGP evaluation for AEMST reinforces this observation. Because HEGP emphasises high-energy glimpses, the improvements from GlimpseP are more pronounced, especially under competing-speaker conditions. While FractileASII and GlimpseP remain close in stationary SSN, GlimpseP consistently shows higher scores at more adverse SNRs. This indicates that the method is particularly successful at preserving or enhancing the more energetic parts of speech, which are crucial for intelligibility.

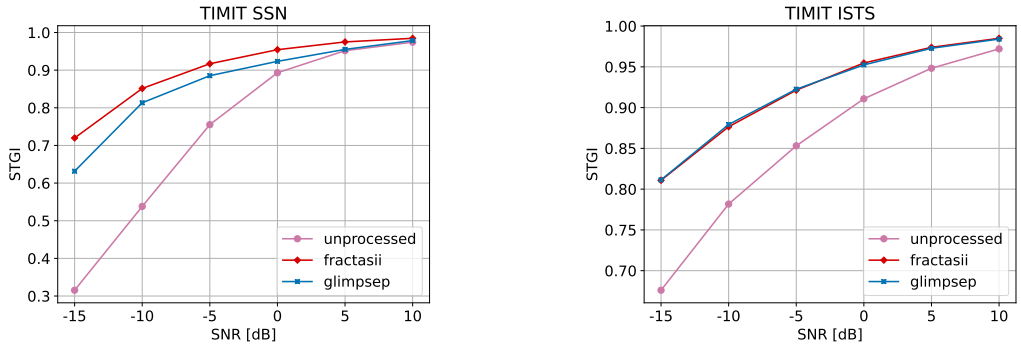


(a) Speech-Signal-Noise **(b)** Competing speaker

Figure 4.4: Performance of baselines and proposed model using AEMST speech signal under **HEGP** metric

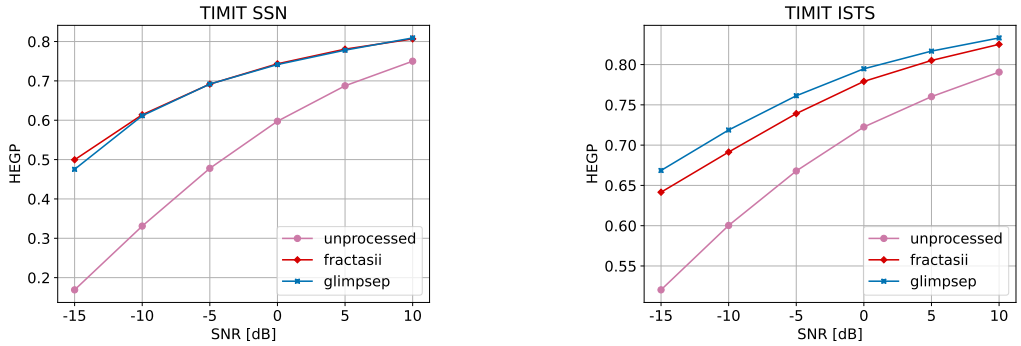
TIMIT

Figures 4.5 and 4.6 display the performance of TIMIT when evaluating the methods under STGI and HEGP, respectively.



(a) Speech-Signal-Noise **(b)** Competing speaker

Figure 4.5: Performance of baselines and proposed model using TIMIT speech signal under **STGI** metric



(a) Speech-Signal-Noise **(b)** Competing speaker

Figure 4.6: Performance of baselines and proposed model using TIMIT speech signal under **HEGP** metric

For the TIMIT dataset, the evaluation under STGI again shows mixed results. In stationary SSN, GlimpseP performs similarly to FractileASII, with only minor differences across SNRs. In contrast, under competing-speaker conditions GlimpseP shows stronger relative gains, surpassing the baseline in most cases. This pattern is consistent with the hypothesis that glimpsing-based optimisation is particularly suited to fluctuating maskers such as overlapping speech, while being less dominant in dense, stationary noise. The HEGP results for TIMIT show that GlimpseP consistently provides improvements over the unprocessed condition and performs on par with FractileASII in SSN. In competing-speaker noise, however, GlimpseP maintains a clearer advantage, especially at lower SNRs. Taken together, these results suggest that while performance on TIMIT is slightly less pronounced compared to the matrix test datasets, the general trends hold: glimpse-based optimisation yields the largest benefits in non-stationary interference.

Chapter 5

Discussion

5.1 Interpretation of Results

The experimental results presented in Chapter 4 show that the Glimpse Proportion maximisation method (GlimpseP) consistently improves objective intelligibility metrics compared to both unprocessed speech and established NELE baselines. The advantages are most apparent under severe noise conditions, where listeners are expected to rely heavily on glimpses for speech perception. For example, in **-15 dB** SSN scenarios, GlimpseP improves gp scores over both unprocessed and FractileASII signals. This supports the hypothesis that directly optimising for glimpsing is beneficial precisely in situations where only fragments of speech are available to the listener.

A closer look at the datasets reveals additional insights. For DANTALE II and AEMST, the method shows robust improvements across both stationary and fluctuating noise conditions. For TIMIT, the glimpse proportions are somewhat smaller, which may be attributed to the fact that the SSN for the evaluation is a signal that matches the long-term spectrum of speech used in Dantale II. Unlike matrix tests, TIMIT features phonetically diverse and semantically coherent sentences. Still, even in this more difficult setting, GlimpseP maintains improvements over the baselines, which suggests good generalisation across speech types.

The comparison across noise types also highlights the strengths of the approach. In stationary SSN, improvements are present but sometimes less pronounced under alternative metrics such as STGI. In contrast, in Competing Speaker noise, the advantages become clearer. This matches expectations from glimpsing theory, which was originally motivated by perception in fluctuating maskers. These results therefore reinforce the perceptual credibility of the chosen optimisation target.

5.2 Comparison with Related Work

Situating these findings in the broader literature, GlimpseP can be viewed as a natural continuation of earlier NELE research. Early systems by Sauert and Vary [23, 22] demonstrated that perceptually motivated optimisation was possible, but they remained tied to SII, which is less reliable in dynamic noise. Taal et al. [28] extended this line of work by formulating closed-form solutions that are computationally attractive, but again were limited by the SII predictor. More recent approaches, such as Villani et al.

As for FractileASII method, the name reference for [32] work in the current document, it is important to acknowlegde the role of their approach in NELE systems. They demonstrated that sophisticated intelligibility-driven optimisation can achieve state-of-the-art performance without requiring expensive computations and data requirements of e.g. deep learning approaches. Relevant to emphasize then that GlimpseP, the current work, achieved comparable results, in regards to intelligibility.

The present study differs by directly addressing a gap identified in the literature review: despite strong evidence for the predictive power of GP [3, 29], glimpse-based metrics had rarely been used as direct objectives for optimisation. By introducing an optimisation for a differentiable version of GP and energy-constrained framework, this thesis contributes a concrete demonstration of how perceptual models can be operationalised in NELE. The fact that GlimpseP performs particularly well in Competing Speaker conditions highlights the validity of this perceptual grounding.

At the same time, the limitations under STGI connect to findings by Hendriks et al. [12], who showed that incorporating modulation transfer functions and reverberation into intelligibility measures is crucial for realistic scenarios. This indicates that glimpse-based optimisation, while valuable, may be insufficient in isolation. A future direction could then be the integration of glimpsing with modulation-sensitive metrics, or the use of hybrid objectives that combine the strengths of different predictors.

Finally, it is worth noting the relationship to machine-learning-based approaches. Recent work has shown that differentiable perceptual measures can serve as training objectives for neural models [31]. While the present study does not employ Learning architectures, it demonstrates the feasibility of gradient-based optimisation with glimpse-inspired objectives, thus bridging the gap between classical NELE frameworks and modern data-driven approaches.

5.3 Limitations

Despite its promising results, the study has several limitations. First, the optimisation is designed for a single metric. Although GP has strong perceptual support, speech understanding is influenced by other cues such as temporal modulation and listener adaptation, which are not explicitly represented. The mixed results under STGI display this drawback.

Second, the evaluation relied entirely on objective predictors. While these measures are widely used and correlate with listening tests, they cannot fully replace subjective evaluations. Without human experiments, it remains unclear whether the optimal gains translate directly into perceptual benefits, or whether they introduce undesirable artefacts such as unnatural timbre or increased listening effort.

Third, the optimisation procedure itself imposes constraints. The SLSQP method guarantees convergence under smooth differentiable objectives, but it remains a local optimiser and may not find globally optimal gain patterns.

Another limitation is the use of time-invariant spectral weights. This design choice simplifies the problem and avoids temporal distortions, but it limits adaptability to rapid changes in noise. In everyday listening environments, noise often fluctuates on short timescales, and fixed gains could not be enough to capture these dynamics.

Finally, the experiments did not address computational cost or latency, both of which are critical in real-world NELE applications such as hearing aids or mobile devices. While the optimisation was tractable in an offline research setting, its applicability for real-time usage is still under consideration.

5.4 Alternative approaches

Several alternative approaches could have been explored to address the limitations above. A multi-objective optimisation framework, combining GP with STGI or with other metrics like STOI or ESTOI, could have reduced dependence on a single predictor and produced more balanced improvements. Such a framework would likely be more computationally demanding, but could yield results that generalise better across metrics and conditions.

The optimisation process itself could also be refined. Using closed-form SII solutions as an initialisation, followed by GP-based fine-tuning, might have reduced convergence time and improved stability.

Finally, including a small-scale, informal, listening test would have added important evidence about perceptual outcomes. Even if limited in scope, such a test could have clarified whether the objective gains correspond to actual improvements in intelligibility and whether the enhanced speech remained natural.

5.5 Summary

In summary, the results confirm that glimpse-based optimisation is a viable and effective strategy for NELE systems, particularly in fluctuating noise conditions where traditional SII-driven approaches struggle. The method leverages a perceptually motivated predictor, achieves consistent objective improvements, and offers a bridge between classical auditory modelling and gradient-based optimisation. At the same time, the work highlights the need to combine glimpsing with other perceptual models, to validate results through listening tests, and to address computational and real-time constraints. Taken together, these points suggest that GlimpseP represents a useful step forward.

Chapter 6

Conclusion and Future Work

6.1 Summary of Findings

The aim of this thesis was to address the gap identified in Chapter 2: although Glimpse Proportion (GP) has been shown to correlate strongly with human intelligibility in noisy conditions, it had rarely been used as a direct optimisation target in Near-End Listening Enhancement (NELE) systems. Existing approaches were largely built on the Speech Intelligibility Index (SII) or its short-time extensions, which are effective in stationary noise but less reliable in fluctuating maskers. The central research question was therefore whether intelligibility improvements could be obtained by maximising a differentiable version of GP within a constrained optimisation framework.

The results confirm that this approach is feasible and effective. The proposed method (GlimpseP) consistently outperformed unprocessed speech and matched or slightly exceeded strong baselines such as FractileASII across multiple datasets and noise conditions. The advantages were particularly pronounced in Competing Speaker scenarios, where glimpsing can be perceptually critical. This demonstrates that the method not only follows theoretical motivations but also delivers practical outcomes. In more stationary noise, improvements were still observed, although performance under STGI showed that temporal modulation aspects were not fully captured by the current formulation.

In addition to validating the potential of GP as an optimisation target, the thesis also contributed a differentiable GP formulation and an application of gradient-based optimisation in an auditory-inspired framework. These elements extend the methodological toolbox for NELE and open possibilities for further research.

6.2 Future Work

While the findings are encouraging, several limitations point to directions for future work. First, the method remains tied to a single metric. A promising extension would be multi-objective optimisation, balancing GP with STGI, STOI, or ESTOI, to better capture temporal and reverberation effects. Such an approach could reduce the dependence on any single predictor and produce more robust results.

Second, the gains in this thesis were time-invariant across frequency bands. A logical next step would be to explore time-varying or adaptive filters that respond to short-term fluctuations in noise. This could further enhance performance in dynamic environments, although it raises additional challenges in terms of complexity.

Third, all evaluations were objective. Conducting subjective listening tests is crucial for validating whether improvements measured by predictors translate into real perceptual benefits. Such tests would also allow for assessment of speech naturalness and listening effort, which are important considerations for practical use.

Finally, integration into machine-learning frameworks presents another avenue. The differentiable GP objective developed here could be embedded in training pipelines for neural enhancement models, combining perceptual grounding with the flexibility of data-driven approaches. Similarly, issues of computational efficiency and latency must be addressed to make the method suitable for real-time applications such as hearing aids or communication devices.

6.3 Closing Remarks

In conclusion, this thesis has shown that glimpse-based optimisation is a viable and promising direction for NELE. By building directly on perceptual theory, the proposed method delivers measurable improvements where needed: in fluctuating noise environments. At the same time, the work highlights that intelligibility optimisation is a multifaceted problem, requiring the combination of several predictors, validation with human listeners, and careful attention to implementation constraints. The contributions made here therefore represent both a step forward in theory-driven enhancement and a foundation for continued research in this area.

Bibliography

[1] *American National Standard Methods for Calculation of the Speech Intelligibility Index*. American National Standards Institute, 1997.

[2] C. Chermaz and S. King. "A Sound Engineering Approach to Near End Listening Enhancement". In: *Proc. Interspeech 2020*. 2020.

[3] Martin Cooke. "A Glimpsing Model of Speech Perception in Noise". In: *The Journal of the Acoustical Society of America* (2006). Department of Computer Science, University of Sheffield, UK; m.cooke@dcs.shef.ac.uk.

[4] J. B. Crespo and R. C. Hendriks. "Speech reinforcement in noisy reverberant environments using a perceptual distortion measure". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2014. DOI: 10.1109/ICASSP.2014.6855131.

[5] J. B. Crespo and R. C. Hendriks. "Speech reinforcement with a globally optimized perceptual distortion measure for noisy reverberant channels". In: *International Workshop on Acoustic Signal Enhancement*. 2014.

[6] Amin Edraki et al. "A Spectro-Temporal Glimpsing Index (STGI) for Speech Intelligibility Prediction". In: *Proceedings of Interspeech 2021*. 2021, pp. 1388–1392.

[7] Amin Edraki et al. "Spectro-temporal modulation glimpsing for speech intelligibility prediction". In: *Hearing Research* 426 (2022), p. 108620.

[8] Ali Fallah and Steven van de Par. "A Speech Preprocessing Method Based on Perceptually Optimized Envelope Processing to Increase Intelligibility in Reverberant Environments". In: *Applied Sciences* (2021). DOI: 10.3390/app112210788.

[9] A. J. Fuglsig et al. "Minimum Processing Near-End Listening Enhancement". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023). DOI: 10.1109/TASLP.2023.3282094.

[10] J. Garofolo et al. *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*. Tech. rep. 1993.

[11] R. C. Hendriks, T. Gerkmann, and J. Jensen. "DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art". In: *Synthesis Lectures on Speech and Audio Processing*. Morgan & Claypool Publishers, 2013.

[12] Richard C. Hendriks et al. "Optimal Near-End Speech Intelligibility Improvement Incorporating Additive Noise and Late Reverberation Under an Approximation of the Short-Time SII". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2020).

[13] I. Holube et al. "Development and analysis of an International Speech Test Signal (ISTS)". In: *International Journal of Audiology* (2010). DOI: 10.3109/14992027.2010.506889.

[14] Jesper Jensen and Cees H. Taal. "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2016).

[15] James M. Kates and Kathryn H. Arehart. "The Hearing-Aid Speech Perception Index (HASPI) Version 2". In: *Speech Communication* (2021).

[16] James M. Kates and Kathryn H. Arehart. "The Hearing-Aid Speech Quality Index (HASQI) Version 2". In: *Journal of the Audio Engineering Society* (2014).

[17] M. Niermann, P. Jax, and P. Vary. "Joint Near-End Listening Enhancement and far-end noise reduction". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2017. DOI: 10.1109/ICASSP.2017.7953102.

[18] M. Niermann, P. Jax, and P. Vary. "Near-End Listening Enhancement by Noise-Inverse Speech Shaping". In: *Proceedings of European Signal Processing Conference (EUSIPCO)*. 2016. DOI: 10.1109/EUSIPCO.2016.7760677.

[19] M. Niermann et al. "Time Domain Approach for Listening Enhancement in Noisy Environments". In: *ITG-Fachtagung Sprachkommunikation*. VDE Verlag GmbH, 2016.

[20] C. V. Pavlovic. "Derivation of primary parameters and procedures for use in speech intelligibility predictions". In: *Journal of the Acoustical Society of America* (1987).

[21] J. Salamon et al. "A dataset and taxonomy for urban sound research". In: *Proceedings of the 22nd ACM International Conference on Multimedia*. 2014.

[22] B. Sauert and P. Vary. "Near End Listening Enhancement Optimized with Respect to Speech Intelligibility Index and Audio Power Limitations". In: *European Signal Processing Conference*. 2010.

[23] B. Sauert and P. Vary. "Near End Listening Enhancement: Speech Intelligibility Improvement in Noisy Environments". In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. 2006. DOI: 10.1109/ICASSP.2006.1661048.

[24] B. Sauert and P. Vary. "Near-End Listening Enhancement in the Presence of Bandpass Noises". In: *ITG Conference on Speech Communication*. 2012.

[25] B. Sauert and P. Vary. "Recursive Closed-Form Optimization of Spectral Audio Power Allocation for Near End Listening Enhancement". In: *Conference on Natural Language Processing*. 2010.

[26] *Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index*. International Electrotechnical Commission, 2020.

[27] Cees H. Taal, Jesper Jensen, and Arne Leijon. "On Optimal Linear Filtering of Speech for Near-End Listening Enhancement". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2013). DOI: 10.1109/TASL.2012.2228314.

[28] Cees H. Taal et al. "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* (2011).

[29] Yan Tang and Martin Cooke. "Glimpse-based metrics for predicting speech intelligibility in additive noise conditions". In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. y.tang@salford.ac.uk, m.cooke@ikerbasque.org. Acoustics Research Centre, University of Salford; Ikerbasque, Bilbao, Spain; Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain. Salford, UK, 2017.

[30] Yan Tang and Martin Cooke. "Optimised spectral weightings for noise-dependent speech intelligibility enhancement". In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. y.tang@laslab.org, m.cooke@ikerbasque.org. Language, Speech Laboratory, Universidad del País Vasco, and Ikerbasque, Bilbao, Spain. Vitoria, Spain, 2015.

[31] Cassia Valentini-Botinhao et al. "Cepstral Analysis Based on the Glimpse Proportion Measure for Improving the Intelligibility of HMM-Based Synthetic Speech in Noise". In: *Proceedings of Interspeech*. San Francisco, USA, 2016. DOI: 10.21437/Interspeech.2016-292.

[32] Filippo Villani et al. "Near-End Listening Enhancement Using a Noise-Robust Linear Time-Invariant Filter". In: *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Aalborg, Denmark: IEEE, 2024. ISBN: 979-8-3503-6185-8. DOI: 10.1109/IWAENC61483.2024.10694258.

[33] K. Wagener et al. "Design, optimization and evaluation of a Danish sentence test in noise". In: *International Journal of Audiology* (2003).

[34] T.-C. Zorila and Y. Stylianou. “On the Quality and Intelligibility of Noisy Speech Processed for Near-End Listening Enhancement”. In: *Proc. Interspeech 2017*. 2017.