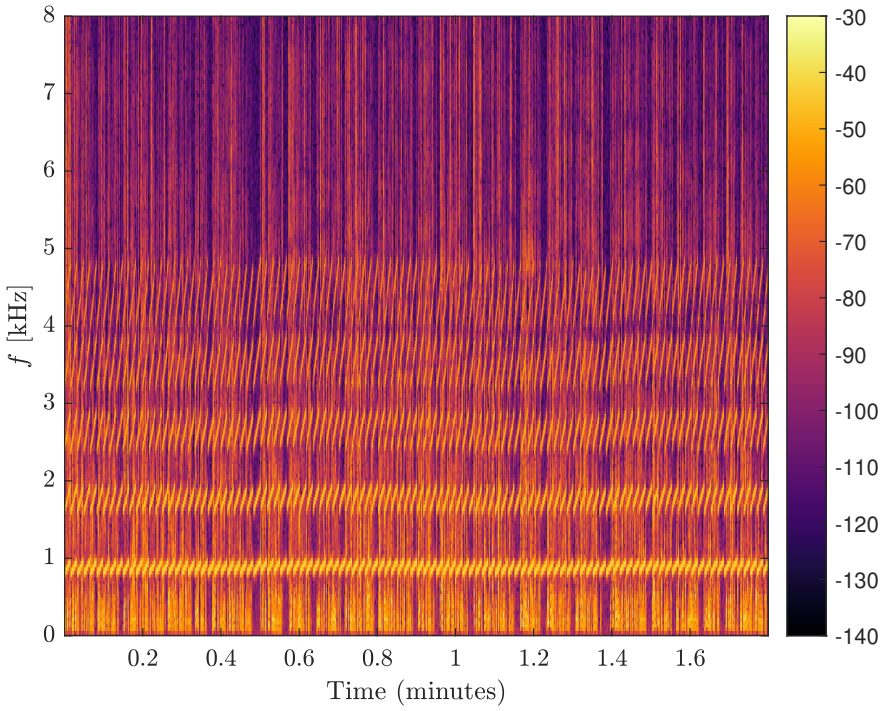


A Model-Based Approach to Tonal Interference Suppression:

An LPC and Cepstral Filtering Framework for Alarm Noise Reduction in Single-Channel Speech Signals in the Context of Emergency Calls

Master's thesis in Sound and Vibration

Richard Heikkilä
Gustav Nilsson

DEPARTMENT OF ARCHITECTURE AND CIVIL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Model-Based Tonal Interference Suppression

An LPC and Cepstral Filtering Framework for Alarm Noise Reduction in Single-Channel Speech Signals from Emergency Calls

Richard Heikkilä
Gustav Nilsson

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
Gothenburg, Sweden 2025

A Model-Based Approach to Tonal Interference Suppression:
An LPC and Cepstral Liftering Framework for Alarm Noise Reduction
in Single-Channel Speech Signals in the Context of Emergency Calls

Richard Heikkilä and Gustav Nilsson

Supervisor: Sigmund Olafsen, Brekke & Strand Akustik
Supervisor: Jens Ahrens, Division of Applied Acoustics
Examiner: Jens Ahrens, Division of Applied Acoustics

Master's Thesis 2025
Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Spectogram of a speech signal contamined with harmonic alarm noise.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

A Model-Based Approach to Tonal Interference Suppression:
An LPC and Cepstral Liftering Framework for Alarm Noise Reduction
in Single-Channel Speech Signals in the Context of Emergency Calls
Richard Heikkilä and Gustav Nilsson
Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology

Abstract

When alarm call center operators receive emergency calls, the caller is often located close to a nearby activated fire alarm or smoke detector. In many cases, this means that alarm noise has a strong presence in the call received by the operator. This is an issue for several reasons. First, it exposes the operator to the risk of hearing damage, second, it degrades the quality of the speech, thereby increasing the listening effort and the cognitive demand.

This thesis acknowledges this problem and applies it to the context of a real emergency call center. Therefore, alarm noise levels in the headsets were measured at the Oslo Fire and Rescue Service in Oslo, Norway. The noise levels in the headsets were found to be above the Norwegian action limit values. In an attempt to propose a solution to this problem, two model-based methods for suppressing tonal noise in emergency calls were proposed. The techniques used in this work are grounded in the source-filter separation approach: Cepstral liftering, and linear predictive coding (LPC). Based on these techniques, two filter frameworks were programmed in MATLAB. In the first proposed method, Proposal *A*, tonal interference suppression was applied solely by cepstral liftering. In Proposal *B*, this technique was iterated and combined with LPC-based analysis.

The performance of the proposed filters was discussed, based on analytical results computed in MATLAB, including waveforms, pitch detection, and spectograms. The filters were further evaluated by objective quality measures (WSS, LLAR and $\text{fwSNR}_{\text{seg}}$. To extend the analysis, a subjective comparative listening test was conducted. The results of the listening test indicated that even though the proposed models were partially successful in suppressing tonal noise components, this was accomplished at the cost of a decreased speech quality. This proved to have a negative impact on the overall perception and the original degraded audio signal was preferred in most cases over the filtered one. The filters were also compared to the built-in filter in the videoconferencing software Webex, which uses AI technology, based on data. Compared to the proposed filters, Webex was found to be superior in suppressing the tonal noise and, at the same time, it managed to preserve the speech quality to a great extent.

More work is needed, to achieve effective suppression of tonal interference while preserving speech quality through a model-based approach. Large data sets are required for the fine-tuning of parameters to ensure efficiency and adaptability to different scenarios. It is suggested that further work investigates the implementation of data-driven technologies and the use of trained neural networks.

Keywords: Tonal Noise Suppression, Noise Filtering, Linear Predicitve Coding, LPC, Cepstrum, Cepstral Liftering, Tonal Interference, Objective Quality Measures, Speech Quality, Listening Test.

Acknowledgements

We would like to express our gratitude to our supervisor Sigmund Olafsen at Brekke & Strand Akustik for his encouragement, thoughtful insights and valuable guidance throughout the work on our master's thesis. Special thanks also go to Roy Kristoffersen at Oslo Fire and Rescue Services, who has provided great help regarding on-site measurements and insights related to the working environment for alarm operators. Thank you for welcoming us to the 110-Central. Everyone at Brekke & Strand in Oslo and Gothenburg, thank you for allowing us in your facilities, lending equipment and for your general assistance. Furthermore, we are grateful for the practical assistance provided by Stefan Zillekens and everyone at HEAD Acoustics GmbH. Finally, we would like to thank our supervisor and examiner at Chalmers University of Technology, Jens Ahrens, for his invaluable support, for giving constructive feedback, and for helping us acquire the tools we needed to complete this work.

Richard Heikkilä & Gustav Nilsson, Gothenburg, June 2025

AI Disclosure

We acknowledge the use of ChatGPT 4o, (https://chat.openai.com/) to generate function modules, automatic plotting and exportation of figures from the hard-coded scripts for tonal noise suppression in speech signals. Typical prompts had the following structure:

"Based on my hard-coded script attached below this prompt, generate a function module and call logic for my explicit pitch tracker-section.", or

"Based on my hard-coded snippets for figure plotting attached below this prompt, generate a function module for figure plotting, LaTeX-formatted annoations/labels, and figure export".

The outputs from these typical prompts were used to enhance the workflow during the iterative development of the noise suppression techniques. ChatGPT 4o was further used for figure and table formatting during report writing in LaTeX, and it also functioned as one of our tools for search of literature during research.

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AI	Artificial Intelligence
BSS	Blind Source Separation
DFT	Discrete Fourier Transform
DNN	Deep Neural Networks
DSP	Digital Signal Processing
$\text{fwSNR}_{\text{seg}}$	Frequency-Weighted Segmental Signal-to-Noise Ratio
FFT	Fast Fourier Transform
ICA	Independent Component Analysis
LLR	Log-Likelihood Ratio
LPC	Linear Predictive Coding
NRF	Noise Reduction Factor
NRT	Noise Reduction Technique
SFM	Spectral Flatness Measure
SPL	Sound Pressure Level
STFT	Short-Time Fourier Transform
VoIP	Voice over Internet Protocol
WSS	Weighted Spectral Slope
WOLA	Window Overlap Add

Nomenclature

γ	Power exponent
a_k	Prediction coefficient
$\vec{a}_c$	LPC-vector of clean reference signal
$\vec{a}_p$	LPC-vector of processed signal
$a[n]$	Additive tonal alarm signal
$A(\omega)$	Fourier transform of the additive tonal alarm signal
$A(z)$	Analysis filter
c_i	Sensitivity coefficient
$dB_{\max}$	Maximum output of the whole signal
$dB(j,m)$	Output of critical band j in frame m
$dB_{\text{loc.max}}(j,m)$	Output of spectral peak closest to critical band j in frame m
$dB_{c,\text{SPL}}$	Overall sound pressure level of clean reference signal
$dB_{p,\text{SPL}}$	Overall sound pressure level of processed signal
$e[n]$	Excitation signal
$E(\omega)$	Fourier transform of the excitation signal
$E(z)$	Fourier transform of the residual
f	Frequency
f_0	Fundamental frequency of alarm
f_s	Sampling frequency
$h[n]$	Vocal tract impulse response
$H(\omega)$	Fourier transform of the vocal tract impulse response
k	Coverage factor of 1.65 for the 95 % CI
$K_{\max}$	Constant with the value of 20 to calculate the weighting function for the WSS
$K_{\text{loc.max}}$	Constant with the value of 1 to calculate the weighting function for the WSS
K_{SPL}	Constant related to the overall sound pressure level for the WSS

L_{EX}	Occupational noise level
$L_{f,\mathrm{eq}}$	Equivalent sound pressure level over one third-octave bands
L_p	Sound pressure level
$L_{p,\mathrm{A,eq}}$	A-weighted equivalent sound pressure level
$L_{p,\mathrm{A},f,\mathrm{eq}}$	A-weighted equivalent sound pressure level over one third-octave bands
$L_{p,\mathrm{C,peak}}$	Peak C-weighted sound pressure level
p	Sound pressure
p_{A}	A-weighted sound pressure
$p_{\mathrm{C,peak}}$	C-weighted peak sound pressure
p_{ref}	Reference sound pressure of $2 \cdot 10^{-5}$ Pa
$\tilde{p}$	Root mean square of the sound pressure
$\boldsymbol{R}_c$	Autocorrelation matrix for the clean reference signal
$s[n]$	Speech signal
$\hat{s}[n]$	Speech sample approximated by a linear combination of past samples
$S(\omega)$	Fourier transform of the speech signal
$S_c(j,m)$	Spectral slope for clean reference signal at critical band j in frame m
$S_p(j,m)$	Spectral slope for processed signal at critical band j in frame m
$S(z)$	Transfer function of filter
T_0	Reference time of 8 hours
T_{e}	Measurement time
u	Combined standard uncertainty
u_i	Standard uncertainty
U	Expanded uncertainty
$W(j,m)$	Weighting function for the WSS and $\mathrm{fwSNR}_{\mathrm{seg}}$ for critical band j in frame m
$W_{\mathrm{max}}(j,m)$	Factor in the equation for the weighting function of the WSS
$W_{\mathrm{loc.max}}(j,m)$	Factor in the equation for the weighting function of the WSS
$x[n]$	Discrete time-signal
$X[k]$	DFT of the discrete time-signal
$\lvert X(j,m)\rvert$	Magnitude spectrum for the clean signal
$\lvert \hat{X}(j,m)\rvert$	Magnitude spectrum for the processed signal

Contents

List of Figures

List of Tables

1

Introduction

1.1 Background

The master thesis is carried out in collaboration with the Norwegian acoustics consultancy firm Brekke & Strand. The assignment is to investigate the acoustical environment at an alarm center for Oslo's Fire and Rescue Service. The alarm center, located in Oslo, Norway, is a critical work place in the sense that it is of great importance that operators can perform their tasks with as little interruption as possible. It is also crucial that the employees have a low general stress level to allow them to fully concentrate on their tasks. Annoyance and stress are both highly linked to noise [1] and prolonged exposure to noise is associated with various health issues such as hypertension and other types of cardiovascular diseases [2].

Because it is an alarm center, many of the calls they receive are corrupted by tonal noise from fire alarms, smoke detectors or sirens. The mixture of clean speech, background noise and tonal interference from alarms (sketched in Figure 1.1) causes problems related to intelligibility and health. If the speech quality is degraded, the operator needs to increase the volume of the signal. This operation increases the level of the tonal noise, which can cause unhealthy work conditions in the headsets for the operators. Brekke & Strand have previously made measurements at Oslo's Fire and Rescue, and according to their measurements, a high level of noise between 1.25 - 3 kHz could be observed. This is a problem for several reasons. The high-frequency noise may cause annoyance and stress for the operators, which can impair their ability to focus on their tasks, as well as have a long-term negative effect on their overall well-being. Due to these reasons, it is of high importance that the problems are addressed.

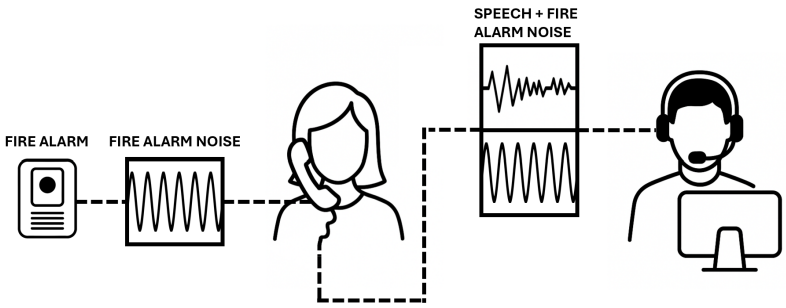


Figure 1.1: The fire alarm or smoke detector noise gets transmitted through the phone of the caller to the receiving call operator who receives a mixed signal that consists of both speech and alarm noise.

1.2 Aim

The aim of the thesis is to investigate how the working environment at an emergency call center can be improved from an acoustical point of view, such that the requirements are fulfilled under Norwegian standards [3]. For this thesis, the focus is placed on the issue of high-frequency noise in the headsets. This work will propose a model-based approach to suppression of tonal interference, investigating two techniques: Cepstral liftering and Linear Predictive Coding (LPC-analysis). The proposed filters will be tested and evaluated, including comparisons with a state-of-the-art data-driven technique used in the video conference software Webex.

1.3 Objectives

The following objectives serve as the framework for the thesis:

- Research current knowledge and previous work on tonal noise suppression in degraded speech signals.
- Perform full-day measurements of sound levels in the headsets over 1 week to capture occupational noise levels corresponding to a full work day in the work environment.
- Perform a task-based measurement during an alarm call at the emergency call center, to capture occupational noise levels corresponding to one hour in the work environment.
- Prototype design of digital filters for adaptive suppression of tonal noise, using cepstral liftering and LPC-analysis.
- Compute analytical filtering results of the proposed filter, the state-of-the-art filter and the raw sound file.
- Compute analytical evaluation metrics, using existing standards: WSS, LLR, $\text{fwSNR}_{\text{seg}}$.
- Conduct listening tests for subjective evaluation of the performance of the filter prototypes.

1.4 Limitations

To keep the work within the scope of the thesis, and to account for a limited time frame, the project is restricted to several limitations:

- The measurements were taken during one simulated alarm call, due to the uncertainty of incoming real-world calls fulfilling required conditions.
- Two alarm types were considered during the work: frequency-sweeping stationary alarms, and constant tone-alarms operating in T3-temporal pattern.
- Filter prototypes were programmed using MATLAB and are not feasible for real-time applications.
- The filter design relied solely on a model-based approach, excluding data-driven or AI techniques, which may limit generalisability.
- The filters were tested on constructed test files, due to complications during the measurement stated in 5.2.2.

1.5 Societal, Ethical and Ecological Aspects

It is of great interest from a societal viewpoint that a workplace such as an alarm center can function as efficiently as possible. This is due to how vital it is for the infrastructure that the emergency services function without any obstacles. This can be applied to other types of workplaces with similar societal functions.

From an ethical standpoint, a key concern in an alarm center is the handling of sensitive and confidential information. Alarm calls often involve personal details, medical conditions, criminal incidents, and other highly sensitive data that require strict confidentiality. Moreover, the secrecy of internal operations within the alarm center itself is crucial. Information about response protocols, system vulnerabilities, and operational strategies must be protected to prevent misuse by unauthorized parties. To account for these concerns, sound recordings were only captured during the simulated alarm call measurement.

As mentioned, noise has also proven to be a potential health risk and reduction of noise exposure in the work place could prevent such potential health risks due prolonged noise exposure. Improved acoustical conditions such as those proposed in this thesis could therefore be argued to fall into the category of sustainable development [4].

2
Literature Review

The initial stage of this work is a literature review. It serves as the foundation of the work, gathering previous research and current knowledge within the scope of the thesis. The other purpose of a literature review is to identify and illuminate knowledge gaps in the topic. Relevant literature was identified using search engines such as Chalmers Library and Scispace. The selected sources focus on tonal noise suppression and the modelling of speech and noise in digital signals, encompassing both foundational contributions and contemporary developments in the field. To ensure scientific rigour, literature published in peer-reviewed journals or presented at reputable conferences was preferred.

To delimit the research for this work into specific subjects, the following research questions are used as a guideline:

- How are humans affected by noise in the workplace in general?
- What are the current regulations to noise exposure in office environments, and particularly in headsets?
- What work has previously been done to research the issue on noise in call centers?
- Which techniques for noise reduction for suppression of tonal noise exist, and how can they be implemented?
- Which noise reduction techniques are most suitable for suppression of tonal noise in the frame of an alarm center?
- How can speech intelligibility and speech quality be analytically evaluated?

2.1 General Topic & Overview

There are already regulations established, both worldwide and on national levels, to govern general health aspects at workplaces, as well as guidelines and engineering standards to ensure safe sound conditions within them. These regulations are based on psychological research, where the human perception of sound, usually a subjective matter, plays the deciding role. It has been proven that noise in the work environment can have harmful physical and psychological effects on humans [5]. This matter is interdisciplinary, involving psychology and acoustics among other fields. To serve the scope of the project, the literature review focuses on two main fields. First comes a study of the existing theory and applications of digital signal processing techniques for tonal noise reduction in speech signals corrupted by tonal

noise. The second part reviews relevant methods to evaluate the performance of processing techniques used to enhance corrupted speech signals.

2.2 Acoustic Environments and Health

The link between acoustic and health is the underlying theme behind the demand for the type of research the thesis conducts on behalf of Brekke & Strand. Previous research and work within the field of environmental psychology is therefore of interest, to understand the perception of sound environments from a humanistic perspective. To obtain insights, beneficial for the contexts of general background and human psychology, master theses from Ellefsen [6] and Kovanen Trangmyr [7] were studied with interest. However, it is important to underscore that this thesis is written by students in acoustic engineering and there is no ambition of attempting to conclude from the perspective of psychology or psychoacoustics.

2.3 Noise Levels in Work Environments

In 1999, the World Health Organization (WHO) established standards for safe noise levels in community environments [5]. The guidelines address various specific contexts, including noise exposure through headphones. This aspect is particularly relevant to this work, as the case focuses on the acoustic conditions of the work environment in which alarm operators receive and manage calls in headsets. According to WHO, both adults and children should not be exposed to sound levels from headphones that exceed an equivalent daily A-weighted noise level of 70 dBA, which corresponds to a one-hour exposure at 85 dBA. The purpose of this regulation is to avoid hearing impairment. Additionally, to avoid acute hearing impairment, exposure to sound levels that exceed 110 dBA should always be avoided. Since the case studied in this thesis involves work activities demanding high cognitive performance and communication, where noise is part of the work environment, regulations to prevent hearing impairment are inadequate, hence the need for further regulations for specific work environments. The thesis' scope is based on a work environment in Norway, thus there are specific regulations for sound pressure levels and room acoustic parameters that all workplaces must meet. Those were established by the Norwegian Ministry of Labour and Social Inclusion in 2011 [8] and the values are presented in Chapter 4.

2.4 Previous Work on Noise in Call Centers

Several previous studies have been conducted on the topic of noise in call centers. One study from 2002 concluded that the general noise levels could be considered safe in the call centers that were studied based on current regulations and that the risk of hearing damage there was "extremely low" [9]. However, they still acknowledged

that the use of headsets could convey a risk of dangerous noise levels due to potential acoustic shocks if not having some kind of protection against it [9]. Another study from 2018 found that even though the operators that were surveyed expressed fatigue, this could not be linked to the noise exposure from using headsets and the measured noise levels were within the current regulations [10]. However, a study from 2010 did find that the call centers that were surveyed had problems with noise levels that exceeded permissible limits [11]. It was also found that one of the main noise sources was the headsets [11]. A study conducted in 2003 by Fulcrum Voice Technologies and the University of Hertfordshire emphasizes the common issue with alarm signal noise that is transmitted from the caller to the call operator at alarm centers [12]. They bring attention to how this noise may both cause annoyance for the operator and as well as reduce the intelligibility of the call [12]. The study included an implementation of a noise suppressing filter targetting the alarm noise which proved to be successful in improving the speech intelligibility of the signal [12]. More details on how the noise suppression works in this case is described in Section 2.5.4.

Based on these studies, as they have varying conclusions, it appears that there is no definite consensus among researchers on the magnitude of the problem of noise in call centers. However, these studies seem to base their conclusions on regulative values from WHO as described previously in Section 2.3. As will be shown in Chapter 4 the regulations for this work are stricter. Additionally, previous work done within the field of environmental psychology by Kovanen Tangmyr together with Brekke & Strand and the Oslo Fire and Rescue Service, show that alarm call operators suffer from high noise levels in headsets. This problem is caused by degraded quality of the incoming signal, forcing the operator to work under loud noise levels. This causes impaired ability to perform tasks, make decisions, and hearing damage [7]. Moreover, the study by Fulcrum Voice Technologies and the University of Hertfordshire draws attention to the special needs and noise problems that exist at alarm centers in particular, showing that it is a topic worth studying further. Additionally, they present that perceivable and quantifiable improvements are obtainable through digital signal processing, which is in line with the objective of this thesis. With the results from the previous work carried out for the Oslo Fire and Rescue Services by Kovanen Tangmyr, this work has a clear problem in focus.

2.5 Noise Reduction Techniques for Tonal Noise Suppression

Speech signals transmitted over telephone lines, recorded in open environments, or captured under emergency conditions are frequently corrupted by background noise [13]. This degradation can severely impact speech intelligibility and overall communication effectiveness, especially in critical contexts such as emergency call centers. To this day, a wide array of noise reduction techniques (NRTs) have been developed, grounded in different signal processing paradigms, assumptions, and representations.

Enhancement of speech quality is closely tied to noise reduction, but the relationship is not always straightforward. In many cases, suppressing background noise improves perceptual quality while simultaneously introducing artifacts that degrade intelligibility [14]. The challenge becomes particularly complex in high-stakes settings such as emergency communication, where listener fatigue, perceptual effort, and distorted speech cues can impair decision-making [7], [15]. These scenarios often involve uncorrelated, additive noise, sometimes tonal or modulated, that overlaps with speech in time and frequency. The design of effective enhancement systems must therefore address both perceptual and objective metrics, balance distortion and suppression, and operate under real-time constraints.

This section examines key NRTs relevant to tonal noise suppression and speech enhancement. These methods range from early spectral approaches that operate in the frequency domain to more recent adaptive, statistical, and data-driven techniques. Each method is assessed in terms of its underlying assumptions, theoretical formulation, effectiveness in practical scenarios, and limitations in handling complex noise conditions. Since this thesis is limited to a model-based approach for tonal noise suppression, the goal is to find the most suitable filtering techniques for two major alarm types listed below:

- **Sweeping alarms**, which are frequency-modulated and operate at a constant amplitude.
- **Frequency-stable alarms**, which are amplitude-modulated and operate at a constant frequency.

2.5.1 Spectral Subtraction and Wiener Filtering

Spectral subtraction and Wiener filtering are among the most established techniques in single-channel speech enhancement. Both operate in the frequency domain and share the goal of attenuating additive noise while preserving speech intelligibility. They rely on the assumption that speech and noise are statistically independent, and that the noise spectrum can be reliably estimated—typically during segments where speech is absent.

2.5.1.1 Spectral Subtraction.

Initially developed in analog form by Schroeder [16], [17] and implemented digitally by Boll [13], spectral subtraction subtracts an estimate of the noise magnitude spectrum from the observed noisy signal. The method is computationally simple and effective under stationary noise conditions, such as background hums or HVAC noise. However, it assumes that the noise remains constant or slowly varying, and that clean speech does not overlap spectrally with the noise.
A notable drawback is the introduction of perceptual artefacts known as *musical noise* [14]—random, tonal residuals arising from spectral over- or under-subtraction. These artefacts often degrade the perceived quality of speech even when intelligibility is improved.

2.5.1.2 Wiener Filtering.

Wiener filtering, formulated in the context of minimum mean square error (MMSE) estimation [18], offers a more statistically grounded alternative. Rather than subtracting a fixed noise estimate, it applies frequency-dependent gains based on local signal-to-noise ratios. This allows smoother suppression and avoids some of the harsher artefacts seen in spectral subtraction.

Despite this theoretical advantage, Wiener filtering shares core limitations. It assumes uncorrelated noise and speech, stationary noise statistics, and accurate knowledge of the noise power spectral density (PSD). In practical terms, this translates to similar weaknesses when faced with non-stationary or speech-like noise. Simulation studies [19] have demonstrated a clear trade-off between noise reduction and speech distortion: as noise attenuation increases, so does the risk of degrading important speech cues—especially when speech and noise overlap spectrally.

Benesty *et al.* [18] introduced metrics such as the Speech Distortion Index (SDI) and Noise Reduction Factor (NRF) to quantify this trade-off. Their results confirmed that even statistically optimal filters must balance intelligibility against perceptual quality.

Both methods depend on prior knowledge or estimation of noise characteristics. Their performance drops in dynamic or unpredictable environments—such as those involving alarm tones, overlapping speakers, or background activity typical in telephony systems. These constraints motivate the development of more adaptive or blind suppression strategies, which are less reliant on stationary assumptions and explicit noise models, as discussed in the following sections.

2.5.2 Blind Source Separation and Independent Component Analysis

Blind Source Separation (BSS) refers to a family of methods aimed at isolating individual source signals from observed mixtures without relying on explicit prior knowledge of the source or noise characteristics. This paradigm offers a contrast to traditional noise reduction techniques that depend on spectral models or stationarity assumptions. A widely used approach within BSS is Independent Component Analysis (ICA), which separates signals by maximizing their statistical independence, typically under the assumption that the sources are non-Gaussian and mutually independent.

ICA has been applied to speech enhancement in complex acoustic environments such as multi-speaker conversations, teleconferencing, and mobile communications [20]. Unlike techniques like Wiener filtering, which require prior estimation of the noise power spectral density, ICA estimates both the mixing and source signals by exploiting higher-order statistical structure in the observed data.

In the implementation described by Obillaneni *et al.* [20], neural networks are used to refine the separation matrix iteratively, guided by the minimization of mean square error (MSE) between the observed signal and a reconstructed noise estimate. This learning-based strategy improves robustness in noisy or reverberant environments and enhances speech clarity.

While ICA has demonstrated significant gains in signal-to-noise ratio and intelligibility, especially in multi-channel setups, its effectiveness depends on several assumptions and conditions. The core assumption of source independence may not strictly hold in conversational speech over a single-channel telephone line. It typically performs best in scenarios with as many sensors (microphones) as sources. Additionally, neural ICA methods require either extensive training data or computational resources for iterative adaptation.

2.5.3 Wavelet-Based Denoising

Wavelet-based denoising, introduced by Donoho in 1995 [21], offers a time–frequency approach to noise suppression. By transforming a signal into wavelet coefficients—representing localised frequency content—noise can be attenuated using soft-thresholding, where low-magnitude coefficients are suppressed and high-magnitude ones are slightly shrunk. This preserves important speech features while reducing transient, non-stationary noise [22].

The technique is known for its balance between noise reduction and signal integrity, but its performance depends heavily on appropriate threshold selection. While wavelet denoising has shown promise, particularly in recent hybrid models that combine wavelet-domain processing with deep neural networks [23]–[25], the data-driven application drifts out of the scope of this thesis. Instead, the focus remains on model-based filtering techniques more directly aligned with tonal suppression.

2.5.4 A Model-Based Approach: Linear Predictive Coding (LPC)

Linear Predictive Coding (LPC) models each speech sample as a linear combination of past samples, producing a filter that captures the spectral envelope (vocal tract) and a residual representing excitation [26]. This structure enables effective decomposition of voiced speech into formants and pitch, and supports low-complexity synthesis and modification.

Gül *et al.* [12] proposed a dynamically adaptive LPC-based method to suppress tonal alarm interference in telephony speech. Their approach targets stationary sinusoids by averaging LPC coefficient vectors over several frames and subtracting the resulting bias from each frame's LP coefficients. The de-biased coefficients are then used to re-synthesise the speech, attenuating the tonal energy absorbed into the original filter poles, without modifying the residual.

The method was validated on real emergency call data from the Kent Fire Brigade (UK), achieving up to 34 dB tonal attenuation with only 6% processor usage on a TMS320C6711 DSP. However, its effectiveness depends on the assumption that tonal interference is stationary and captured in the filter poles. This may not hold for modulated or transient tonal noise. Further limitations include residual contamination, instability in LPC estimation, and sensitivity to model order selection [27].

Despite these constraints, LPC-based suppression provides a computationally efficient and interpretable framework for tonal noise reduction, particularly suitable for real-time telephony scenarios. Its structure is of interest to the theoretical foundation for the suppression models developed in this thesis.

2.5.5 From Spectral Filtering to Source–Filter Separation

While spectral subtraction, Wiener filtering, and LPC are effective techniques against stationary and some non-stationary noise, they are not tailored to highly structured tonal interference. Alarms with harmonic spacing, whether stationary (e.g., tones in T3-temporal pattern) or frequency-modulated (e.g., sweeping alarms), pose a unique challenge due to spectral overlap with voiced speech and the inability of traditional methods to separate source excitation from filter characteristics.

Cepstral alanysis, rooted in the source–filter model of speech production, offers a way to address this limitation by separating excitation (pitch, harmonics) from the spectral envelope of the vocal tract [28], [29]. This approach is on a theoretical level particularly suited for frequency-periodic interference with dynamic pitch, such as sweeping alarms, where interference appears as regular quefrency components in the cepstrum.

2.5.5.1 Cepstral Liftering for Tonal Interference Suppression

The cepstrum is computed by taking the inverse Fourier transform of the log magnitude spectrum:

$$\text{cepstrum}(n) = \mathcal{F}^{-1}\{\log|\mathcal{F}\{x(t)\}|\}, \tag{2.1}$$

This transformation linearises convolution and separates periodic components (e.g., excitation harmonics) from the smoother vocal tract envelope. Liftering (cepstral-domain filtering) enables suppression of specific quefrency ranges:

- **Low-pass liftering** preserves formants and removes pitch detail;
- **High-pass or band-stop liftering** targets rapidly periodic components, such as the pitch or harmonics of sweeping alarms [28].

This makes cepstral liftering attractive for alarm suppression: tonal interference with regular spectral spacing maps to distinct quefrency peaks, which can be selectively attenuated—especially when guided by pitch confidence, spectral flatness, or harmonic salience [30].

Cepstral liftering is most effective under the assumption of local quasi-stationarity and clear harmonic periodicity [29]. Challenges include:

- **Resolution trade-offs:** Accurate quefrency resolution requires long analysis windows, which reduces responsiveness to pitch changes.
- **Overlap with speech:** Harmonic interference and voiced speech can share quefrency bins, risking speech distortion during suppression.
- **Dependence on guidance:** Unguided or overly aggressive liftering can degrade speech naturalness; dynamic liftering may become unstable without reliable pitch tracking.

In summary, cepstral liftering is a powerful but specialised tool. It is not a general-purpose denoising method, but under appropriate conditions, particularly for sweeping tonal alarms, it enables targeted suppression while preserving speech intelligibility. For this reason, it can serve as a core component in the suppression strategies developed in this thesis.

2.5.6 Selection of a Suitable Method for Tonal Noise Suppression

The reviewed literature spans a broad range of noise reduction techniques (NRTs), from classic frequency-domain filtering (e.g., spectral subtraction and Wiener filtering) to more adaptive and structure-aware methods such as blind source separation, wavelet denoising, and parametric modeling. While many of these approaches are effective against broadband or transient noise, they are not designed for structured tonal interference—especially alarm signals with harmonic or modulated characteristics.

In emergency communication scenarios, the dominant noise is not rarely tonal, either stationary (e.g., T3 alarms) or sweeping (e.g., frequency-modulated fire alarms). These signals overlap spectrally with voiced speech, and their regular structure makes them resistant to general-purpose suppression. Within this context, two model-based methods stand out:

- **Linear Predictive Coding (LPC)** models the spectral envelope of speech and enables suppression of stable tonal components by modifying the prediction filter. Prior work by Gül *et al.* demonstrated that tonal interference absorbed into the LPC model can be attenuated without modifying the excitation signal.
- **Cepstral liftering**, in contrast, suppresses periodic interference in the que-

frency domain by targeting harmonic spacing, making it especially suitable for modulated tones such as sweeping alarms.

While each method has limitations—LPC in handling modulated tones, and cepstral liftering in isolating overlapping speech harmonics—their conceptual complementarity makes them promising candidates for hybrid use. LPC offers time-domain parametric modeling; cepstral liftering exploits harmonic regularity in the spectral domain.

This thesis investigates both methods independently and in combination, with the aim of suppressing tonal alarm interference under real-time and telephony constraints. The overarching goal is to assess whether a hybrid strategy can offer greater robustness and speech quality than either method alone. Based on the initial research carried out in the literature review, the following objectives will guide the scope of this work:

- **1.** To what extent can cepstral liftering suppress tonal alarm signals without introducing speech artifacts or reducing intelligibility?
- **2.** How does the performance of LPC-based alarm suppression compare to cepstral liftering when applied to stationary tonal interference?
- **3.** Can a hybrid or sequential approach combining LPC and cepstral liftering offer improved noise suppression while preserving speech integrity?

These questions are addressed through the design, implementation, and evaluation of experimental speech enhancement systems. Each system is assessed in terms of its ability to attenuate tonal noise, maintain intelligibility, and preserve speech quality under realistic communication constraints.

2.6 Rating of Speech Quality

There are several parameters that can affect the quality of speech. In this work the topic of speech quality is primarily of interest because of the corruption of the signal of the incoming call, which is due to the presence of alarm noise close to the caller. However, even though the speech intelligibility of the room is also of importance when assessing the general acoustical environment of the alarm center, in this thesis it is first and foremost of interest to be able to measure the intelligibility before and after the alarm noise has been filtered out of the signal, in order to estimate the improvement. As shall be discussed, what is referred to as listening effort is also of relevance.

When rating speech, both objective and subjective methods can be useful. As mentioned in Section 2.4, the study by Fulcrum Voice Technologies and the University of Hertfordshire showed that the speech intelligibility of the speech signal could be improved by implementing a noise suppressing filter. They based this conclusion on results that were obtained by using two different objective evaluation methods of the speech intelligibility; the *Itakura-Saito* measure and the *weighted*

spectral slope measure, abbreviated as IS and WSS [12]. In a study conducted in 2008 where nine different objective measurement methods of speech intelligibility were compared, including the IS and WSS, it was concluded that the WSS was the most accurate alternative [31]. The WSS measure was first developed and presented in 1982 by Dennis H. Klatt in an attempt to create a robust method that could compute the speech intelligibility without having to account for smaller peaks between larger peaks in the spectrum [32]. On how to calculate the WSS, see Section 3.3.1. Another speech quality measure which was developed and presented by Yi Hu and Philipos C. Loizou in 2008 is the frequency-weighted segmental signal-to-noise ratio, abbreviated as $\text{fwSNR}_{\text{seg}}$ [33]. In the same paper, the $\text{fwSNR}_{\text{seg}}$ was presented, it was found that, together with the Log-Likelihood Ratio (LLR), it was the best alternatives after the PESQ (Perceptual Evaluation of Speech Quality) when assessing the overall speech quality [33]. Even though PESQ is the recommended method by ITU-T (The International Telecommunication Union Telecommunication Standardization Sector) to measure speech quality in telephony speech, they found that the $\text{fwSNR}_{\text{seg}}$ and LLR were almost as accurate, but with a much lower computational effort [33]. Another paper presented in 2013 that compared PESQ, SNR_{seg} (segmented signal-to-noise ratio), $\text{fwSNR}_{\text{seg}}$ and the C_{ovl} measure (composite measure combining the MOS (Mean Opinion Score), LLR and WSS as $C_{\text{ovl}} = 1.594 + 0.805 \cdot \text{MOS} + 0.512 \cdot \text{LLR} + 0.007 \cdot \text{WSS}$) concluded that the $\text{fwSNR}_{\text{seg}}$ correlated the most with subjective measures when assessing the speech intelligibility [34]. Based on the findings presented in these papers, the WSS, $\text{fwSNR}_{\text{seg}}$ and LLR are deemed suitable for to scope of this work and they will be used in this thesis to evaluate the effeciency of the filters. On how to calculate the $\text{fwSNR}_{\text{seg}}$ and LLR see Section 3.3.2 and Section 3.3.3 respectively.

In addition to the intelligibility or quality of a speech signal, one can also speak of the *listening effort* required to comprehend a speech signal. In contrast to speech intelligibility criteria such as STI (Speech Transmission Index) which is a standardized objective measurement basis for determining the speech intelligibility in rooms, or the WSS, $\text{fwSNR}_{\text{seg}}$ or LLR, the listening effort is an opinion based variable used to determine the effort required for the comprehension of speech [35]. One paper from 2018 concluded that attempting to decipher speech with reduced comprehensibility notably affects several "cognitive operations required for both linguistic and non-linguistic tasks" [36]. It is therefore of interest to be able to measure it. Ensuring a low listening effort is of high relevance in the context of this thesis since the call operators need to be able to fully concentrate without having to put unnecessary effort into trying to understand the one calling. In 2021 The European Telecommunications Standards Institute produced a technical specification for methods to objectively assess the listening effort of speech called ETSI TS 103 558 [35]. However, the model requires a reference source, which means that it needs to be able to compare the corrupted signal with the clean signal. This might not necessarily pose a problem in general but it requires a more complex measurement process. For this work however the clean signal cannot be known. In November 2024 the German company HEAD acoustics GmbH launched the software tool LEAP [37]. It is engineered to predict the perceived listening effort on a scale of 1 to 14, with 1 indicating

no effort and 14 indicating complete noise [38]. In contrast to ETSI TS 103 558, LEAP is based on a single-ended algorithm [38]. This means that it does not need to know the reference signal [38]. LEAP was initially supposed to be implemented in this thesis to evaluate the filters, however, technically it showed to be complex to apply it for this type of context. Instead, a listening test was conducted where the participants were asked to state their experienced listening effort among other subjective parameters (see Section 5.7 and Section 6.3).

3
Theory

This chapter begins by defining some basic acoustical concepts that are used in this thesis. The chapter continues by describing the theory behind the signal processing techniques and methodologies which are applied in this thesis. The chapter then concludes by describing the theory behind the objective quality measures that were used.

3.1 Foundational Acoustics

The following section briefly describes some of the most fundamental parameters in acoustics that are of relevance in this work.

3.1.1 Sound Pressure Level

The sound pressure level L_p is calculated as

$$L_p = 10 \log \left(\frac{\tilde{p}}{p_{\text{ref}}} \right)^2 \tag{3.1}$$

where $p_{\text{ref}} = 2 \cdot 10^{-5}$ Pa is the reference sound pressure and threshold of hearing at 1000 Hz and $\tilde{p}$ is the root-mean-squared sound pressure [39].

3.1.2 A- and C-Weighting

Humans perceive the loudness of sounds differently at different frequencies. To address this, weighting filters are commonly applied [39]. Two of the most common ones are the A- and C-weighting filters, which are the ones that are used in this thesis. The A- and C-weighted sound pressure levels are often denoted as $L_{p,\text{A}}$ and $L_{p,\text{C}}$ with the units dBA and dBC [39]. A-weighting is most commonly applied while C-weighting is applied to sounds with higher sound pressure levels [39]. Other weighting filters exist as well, such as the B- and D-weighting filters, although these are no longer commonly applied [40]. The A-, B-, C- and D-weighting curves are shown in Figure 3.1.

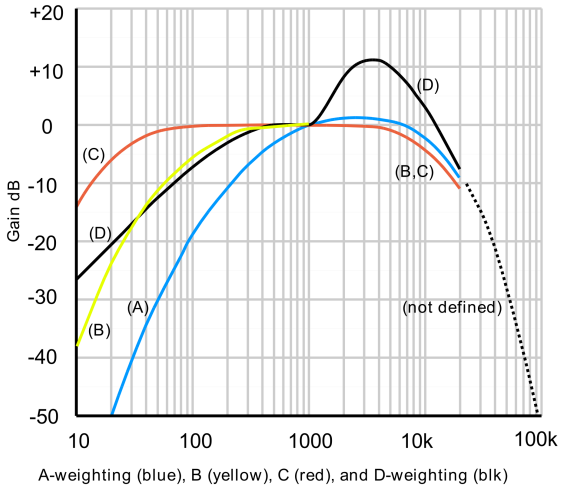


Figure 3.1: A-, B-, C- and D-weighting curves [41].

3.1.3 Occupational Noise

The occupational noise level $L_{\mathrm{EX,8h}}$ is calculated as

$$L_{\mathrm{EX,8h}} = L_{p,\mathrm{A,eq}T_\mathrm{e}} + 10\log\left(\frac{T_\mathrm{e}}{T_0}\right) \tag{3.2}$$

where T_e is the duration of the measurement in hours, T_0 is the reference time of 8 hours and $L_{p,\mathrm{A,eq}T_\mathrm{e}}$ is the A-weighted equivalent continuous sound pressure level for the time period T_e [42]. It is defined as

$$L_{p,\mathrm{A,eq}T_\mathrm{e}} = 10\log\left(\frac{\frac{1}{T_\mathrm{e}}\int_{t_1}^{t_2} p_A^2(t)\,dt}{p_0^2}\right) \tag{3.3}$$

with p_A being the A-weighted measured sound pressure level and t_1 and t_2 the start and end of the measurement period T_e [42].

For a measurement duration of 8 hours the occupational noise $L_{\mathrm{EX,8h}}$ will be equal to $L_{p,\mathrm{A,eq}T_\mathrm{e}}$ as T_e equals T_0. The occupational noise level for any other time duration may be calculated as

$$L_{p,\mathrm{A,eq}T_\mathrm{e}} = L_{\mathrm{EX},T_\mathrm{e}} = L_{\mathrm{EX,8h}} - 10\log\left(\frac{T_\mathrm{e}}{T_0}\right) \tag{3.4}$$

3.1.4 Peak Sound Pressure Level

For this thesis it will be necessary to calculate the peak sound pressure level. This is defined as

$$L_{p,\mathrm{C,peak}} = 10\log\left(\frac{p_{\mathrm{C,peak}}}{p_\mathrm{ref}}\right)^2 \tag{3.5}$$

where $p_{\mathrm{C,peak}}$ is the C-weighted maximum sound pressure that only occurs instantaneously [42].

3.2 Signal Processing for Tonal Noise Suppression

3.2.1 Discrete-Time Signal Processing in Speech Analysis

Speech is inherently analog but is processed digitally in most modern systems. A discrete-time signal $x[n]$ is obtained by sampling at a constant interval $T_s = 1/f_s$, typically at 8–16 kHz for speech. Discrete signals can be analyzed in the time domain, for waveform and predictive models, or in the frequency domain, where spectral content is exposed for noise suppression and enhancement.

Since speech is non-stationary, short-time analysis is used to assume local stationarity within overlapping frames. The Discrete Fourier Transform (DFT) provides a frequency-domain representation of each frame:

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j2\pi kn/N} \tag{3.6}$$

This process, known as the Short-Time Fourier Transform (STFT), reveals how energy evolves over time and frequency. Harmonic peaks from voiced speech and broad spectra from unvoiced or noisy segments can be analyzed to identify speech structures or interference patterns.

3.2.2 Frame-Based Processing and Windowing

The STFT relies on windowing functions to reduce spectral leakage. A well-chosen window balances frequency resolution and side-lobe suppression—crucial for tonal noise suppression, where interference must be distinguished from adjacent spectral features. This thesis employs the Hann window:

$$w[n] = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) \tag{3.7}$$

The Hann window offers a practical trade-off: moderate frequency resolution and 31 dB side-lobe attenuation. It is used consistently in both analysis and overlap-add synthesis stages throughout the work.

3.2.3 Relevance to Tonal Noise Suppression

Frame-based, time–frequency representations underpin all suppression techniques used in this thesis. While LPC operates in the time domain and cepstral analysis in the spectral domain, both rely on accurate short-time representations of speech signals. Frame-based STFT analysis provides the temporal localization and spectral resolution necessary to support these parametric models. These techniques assume local stationarity and depend on accurate spectral estimates within each windowed

frame. The following sections introduce the source–filter model and transform-domain representations that form the theoretical core of the proposed tonal suppression methods.

3.2.4 Speech Production and the Source–Filter Model

Human speech is the result of a dynamic acoustic process that can be modeled as a cascade of two main components: an excitation source and a time-varying filter [43]. This conceptual framework is known as the *source–filter model*. In this model, voiced speech is produced by quasi-periodic vibrations of the vocal folds, which generate a pulse-like excitation signal. This excitation travels through the vocal tract (a resonant cavity formed by the throat, mouth, and nasal passages), which shapes the spectral content of the sound.

Mathematically, the speech signal $s[n]$ can be represented as the convolution of an excitation signal $e[n]$ with the vocal tract impulse response $h[n]$:

$$s[n] = e[n] * h[n] \tag{3.8}$$

This convolutional relationship indicates that the spectral content of speech is shaped by both the source (pitch, periodicity) and the filter (formants, resonances). In the frequency domain, this becomes a product:

$$S(\omega) = E(\omega) \cdot H(\omega) \tag{3.9}$$

where $S(\omega)$, $E(\omega)$, and $H(\omega)$ are the Fourier transforms of $s[n]$, $e[n]$, and $h[n]$, respectively. This multiplicative structure is key to both cepstral and LPC-based techniques, as each seeks to isolate either the excitation or the filter [44]. This structure allows us to model speech as the combination of fine-grained periodicity (from the source) and a smoother spectral envelope (from the filter).

3.2.5 Linear Predictive Coding (LPC)

Linear Predictive Coding (LPC) models speech as the output of an all-pole filter excited by a residual source signal. For each sample $s[n]$, LPC assumes it can be approximated by a linear combination of its past p values:

$$s[n] = \sum_{k=1}^{p} a_k \, s[n-k] + e[n] \tag{3.10}$$

Here, a_k are the linear prediction coefficients and $e[n]$ is the residual (also called the prediction error or excitation). In the z-domain, this relationship defines an all-pole filter:

$$S(z) = \frac{E(z)}{A(z)} = \frac{E(z)}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.11}$$

LPC coefficients are typically computed on a frame-by-frame basis using the autocorrelation method and Levinson–Durbin recursion, minimising the total squared

prediction error within each frame. The residual $e[n]$ captures the excitation characteristics: periodic for voiced speech and noise-like for unvoiced segments.
In the presence of additive periodic noise (e.g., alarm tones), the LPC vector of the contaminated signal $\tilde{\mathbf{a}}_i$ contains both speech and noise components:

$$\tilde{\mathbf{a}}_i = \mathbf{a}_i + \mathbf{a}_{\text{noise}} \tag{3.12}$$

To suppress the interference, the bias $\mathbf{a}_{\text{noise}}$ is estimated across multiple frames as the mean vector:

$$\hat{\mathbf{a}}_{\text{noise}} = \frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{a}}_i \tag{3.13}$$

This estimate is then subtracted from each frame's coefficients:

$$\hat{\mathbf{a}}_i = \tilde{\mathbf{a}}_i - \hat{\mathbf{a}}_{\text{noise}} \tag{3.14}$$

Instead of reconstructing the clean speech from a separately cleaned residual—which itself may still be contaminated—the improved synthesis applies the refined LPC coefficients $\hat{\mathbf{a}}_i$ directly to the original signal:

$$\hat{s}[n] = \sum_{k=1}^{p} \hat{a}_k \, s[n-k] \tag{3.15}$$

This implements a noise-suppressing FIR filter designed per frame to cancel the biasing effect of tonal interference, while preserving the underlying speech structure [12].

3.2.6 Cepstral Analysis and the Cepstrum

Cepstral analysis is a spectral-domain technique grounded in the source–filter model of speech, but its mathematical structure also has the potential to lend itself to modeling structured tonal interference such as alarm signals. The key idea is to transform convolution in the time domain into addition in a transformed domain, enabling component-wise separation.The goal is to isolate and suppress periodic components—such as alarms—by manipulating the cepstral coefficients associated with the excitation, while preserving those associated with the vocal tract envelope,

Let a clean speech signal be modeled as the convolution of an excitation $e[n]$ and a vocal tract response $h[n]$:

$$x[n] = e[n] * h[n] \tag{3.16}$$

This becomes multiplicative in the frequency domain:

$$X(\omega) = E(\omega) \cdot H(\omega) \tag{3.17}$$

Applying the logarithm of the magnitude spectrum converts the product into a sum:

$$\log |X(\omega)| = \log |E(\omega)| + \log |H(\omega)| \tag{3.18}$$

Taking the inverse Fourier transform yields the real cepstrum:

$$c[n] = \mathcal{F}^{-1}\left\{\log\left|\mathcal{F}\{x[n]\}\right|\right\} \tag{3.19}$$

This homomorphic transformation [45] results in an additive representation of source and filter components in the *quefrency domain*. In this domain:

- **Low quefrencies** correspond to the slowly varying spectral envelope (e.g., formants).
- **High quefrencies** represent rapidly periodic components (e.g., pitch or tonal interference).

This separation allows periodic interference, such as pitch and harmonics of a voice or an alarm, to be modeled and manipulated independently of the speech envelope. In tonal alarm suppression, this property forms the basis for targeted cepstral-domain filtering.

3.2.7 Cepstral Liftering

Cepstral liftering modifies a signal in the cepstral domain by applying a window function, here called a *lifter*, to its cepstral coefficients. It is analogous to filtering in the frequency domain, allowing enhancement or suppression of specific quefrency regions.

In tonal noise suppression, the goal is to attenuate the high-quefrency components corresponding to structured interference, while preserving the speech envelope represented at low quefrencies. Common strategies include:

- **Low-pass liftering** – preserves formant structure; suppresses pitch or tonal detail.
- **High-pass liftering** – isolates pitch-related or tonal features.
- **Band-stop (notch) liftering** – suppresses interference localized at specific quefrency bands.

Adaptive liftering, guided by pitch tracking or harmonic confidence, can be utilized to balance suppression precision. This is important to consider, to avoid degradation of voiced speech, which also exhibits high-quefrency harmonic structure.

3.2.8 Modeling Alarm Components in Mixed Signals

In practical applications, speech signals are often corrupted by tonal interference, such as alarm tones. These signals are characterized by strong harmonic structure and fixed or slowly modulated pitch. Unlike speech excitation, which varies over time, alarms tend to produce relatively stable harmonic spacing, making them highly detectable in the cepstral domain.

A mixed signal can be modeled as:

$$x[n] = s[n] + a[n] \tag{3.20}$$

where $s[n]$ is the clean speech and $a[n]$ is the additive tonal alarm signal. In the frequency domain:

$$X(\omega) = S(\omega) + A(\omega) \tag{3.21}$$

While the logarithm in the cepstrum is not strictly linear, the periodicity introduced by $A(\omega)$ is often dominant enough to produce distinct features in the log-magnitude spectrum. Thus, the real cepstrum becomes:

$$c[n] = \mathcal{F}^{-1}\left\{\log|X(\omega)|\right\} \tag{3.22}$$

In this case, tonal interference manifests as sharp peaks at specific quefrencies $n_0 \approx f_s/f_0$, where f_0 is the fundamental frequency of the alarm and f_s is the sampling frequency [45]. Harmonics appear at integer multiples $2n_0, 3n_0, \ldots$. These high-quefrency features are separable from the low-quefrency speech envelope, making them accessible for suppression through band-stop or adaptive liftering. This structured, additive representation is what enables tonal alarms to be modeled as distinct components in cepstral-domain speech enhancement pipelines [45].

3.3 Objective Quality Measures

The following equations for the computation of the WSS, $\text{fwSNR}_{\text{seg}}$ and LLR are adapted from [33] since the assumptions that are made (e.g. filter application, frame size et cetera) are the same as the ones adapted into the Matlab scripts from [46] that were used in this work [33].

3.3.1 Weighted Spectral Slope Distance (WSS)

As mentioned, the WSS was originally defined in [32]. For this work, the WSS is expressed in the following form as desrcibed in [33]:

$$\text{WSS} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K}\left(W(j,m)\left(S_c(j,m)-S_p(j,m)\right)^2\right)}{\sum_{j=1}^{K}W(j,m)} \tag{3.23}$$

where M is the number of frames of the signal and K is the number of bands. The weighting function $W(j,m)$ for band j and frame m is obtained as originally defined in [32] as

$$W(j,m) = \frac{W_{c,\max}(j,m)W_{c,\text{loc.max}}(j,m) + W_{p,\max}(j,m)W_{p,\text{loc.max}}(j,m)}{2}. \tag{3.24}$$

$W_{\max}(j,m)$ is defined for the clean signal c, and processed signal p, respectively as

$$W_{\max}(j,m) = \frac{K_{\max}}{K_{\max} + dB_{\max} - dB(j,m)} \tag{3.25}$$

where $K_{\max}$ is a constant, $dB_{\max}$ is the maximum output of the whole signal and $dB(j,m)$ is the output of band j and frame m. Analogously, for both the clean reference signal c, and processed signal p, $W_{\text{loc.max}}(j,m)$ is defined as

$$W_{\text{loc.max}}(j,m) = \frac{K_{\text{loc.max}}}{K_{\text{loc.max}} + dB_{\text{loc.max}}(j,m) - dB(j,m)} \tag{3.26}$$

where $K_{\text{loc.max}}$ is a constant, $dB_{\text{loc.max}}(j,m)$ is the output of the spectral peak that is closest to band j in frame m and $dB(j,m)$ remains the same as in Eq. (3.25). The constants $K_{\max}$ and $K_{\text{loc.max}}$ were experimentally found in [32] to give the best correlation with reality with $K_{\max} = 20$ and $K_{\text{loc.max}} = 1$.

In Eq. (3.23), $S_c(j,m)$ and $S_p(j,m)$ are the spectral slopes for band j and frame m and they are calculated as [32]

$$S(j,m) = dB(j+1,m) - dB(j,m) \tag{3.27}$$

meaning the spectral slope for frame m in band j is the difference between the output $dB(j+1,m)$ in band $j+1$ and output $dB(j,m)$ in band j.

In [32] it was also suggested that the term $K_{\text{SPL}}(dB_{c,\text{SPL}} - dB_{p,\text{SPL}})$ could be added to Eq. (3.23) to take into account the difference in overall sound pressure level between the two signals. However, it was found that $K_{\text{SPL}} = 0$ gives the best correlation with reality [32] which means that the whole term $K_{\text{SPL}}(dB_{c,\text{SPL}} - dB_{p,\text{SPL}})$ equals 0.

3.3.2 Frequency-Weighted Segmental Signal to Noise Ratio ($\text{fwSNR}_{\text{seg}}$)

The frequency-weighted segmental signal-to-noise ratio, $\text{fwSNR}_{\text{seg}}$, is just like the WSS, and as the name suggests, a frequency-weighted objective quality measure. It is calculated as [33]:

$$\text{fwSNR}_{\text{seg}} = \frac{10}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} \left(W(j,m) \log \left(\frac{|X(j,m)|^2}{\left(|X(j,m)| - |\hat{X}(j,m)| \right)^2} \right) \right)}{\sum_{j=1}^{K} W(j,m)}. \tag{3.28}$$

As in Eq. (3.23) for WSS, Eq. (3.28) includes a weighting function $W(j,m)$. However for $\text{fwSNR}_{\text{seg}}$ this weighting function is defined as

$$W(j,m) = |X(j,m)|^{\gamma} \tag{3.29}$$

where $|X(j,m)|$ is the clean reference signal's magnitude spectrum for band j in frame m and γ is an exponent that can be adjusted for optimal correlation [33]. In [33] it was experimentally found that $\gamma = 0.2$ gave the best correlation. $|\hat{X}(j,m)|$ is the processed signal's magnitude spectrum.

3.3.3 Log-Likelihood Ratio (LLR)

Compared to the WSS and the $\text{fwSNR}_{\text{seg}}$ which are both frequency-weighted measures, LLR is an LPC-based measure [33]. The LLR is calculated as

$$\text{LLR} = \log\left(\frac{\vec{a}_p \boldsymbol{R}_c \vec{a}_p^T}{\vec{a}_c \boldsymbol{R}_c \vec{a}_c^T}\right) \tag{3.30}$$

where $\vec{a}_p$ is the LPC-coefficient vector for the processed signal and $\vec{a}_c$ is the LPC-coefficient for the clean reference signal while $\vec{a}_p^T$, $\vec{a}_c^T$ are the transposed versions of the same coefficients and $\boldsymbol{R}_c$ is the autocorrelation matrix for the clean reference signal [33].

4

Regulations and Measurement Standards

This chapter outlines the relevant regulatory and measurement standards that govern the practical aspects of this study, ensuring compliance with established guidelines and methodologies.

4.1 Regulation- and Action Limits

4.1.1 Sound Pressure Levels

According to the Norwegian Ministry of Labour and Social Inclusion, workplaces are required to ensure that employee exposure to sound remains within prescribed limits, which vary depending on the specific purpose and nature of the work environment. The action values for noise exposure are established in Table 4.1.

Table 4.1: Action- and limit values established by the Norwegian Ministry of Labour and Social Inclusion [8].

Group	Action Value for Noise Exposure	Purpose of Regulation
1	Lower: $L_{\mathrm{EX,1h}} = 55$ dBA	The conversation should be able to be conducted effortlessly.
2	Lower: $L_{\mathrm{EX,1h}} = 70$ dBA	Requirements for concentration and attention are imposed.
3	Lower: $L_{\mathrm{EX,8h}} = 80$ dBA	Mitigate the risk of hearing damage by limiting the noise level in the area.
-	Upper: $L_{\mathrm{EX,8h}} = 85$ dBA and $L_{p,\mathrm{C,peak}} = 130$ dBC	Mitigate the risk of hearing damage through the use of hearing protection.

For groups 1 and 2, the equivalent noise level is measured over one hour. For group 3 and the limit value, the equivalent noise level is measured over a full work shift (8

hours). For shifts lasting 12 hours (which is the case at the alarm central), the limit value is reduced to 83 dB to represent the same exposure level. This adjustment accounts for the logarithmic difference between 12 and 8 hours times 10, which is approximately 2 dB. Additionally, the peak level as described in Section 3.1.4 must not exceed 130 dBC.

4.1.1.1 Recommended Levels

According to the Norwegian Labour Inspection Authority's Regulations on the Performance of Work, §14-6 Special Measures Against Noise When Action Values Are Exceeded, noise exposure should be reduced by at least 10 dB below the lower action value. This requirement is typically applied only to groups 1 and 2. This, together with the 12-hour measurement duration gives the following target values as displayed below in Table 4.2.

Table 4.2: Recommended action- and limit values based on the 12-hour workday.

Group	Action Value for Noise Exposure	Purpose of Regulation
1	Lower: $L_{\mathrm{EX,1h}} = 45$ dBA	Requirements for concentration and attention are imposed.
3	Lower: $L_{\mathrm{EX,12h}} = 78$ dBA	Mitigate the risk of hearing damage by limiting the noise level in the area.

4.1.1.2 Measurement Uncertainties

In accordance with ISO 9612:2009, statistical uncertainties of the measurements need to be taken into account [42]. The expanded uncertainty U is used to express the one-sided 95 % confidence interval as $L_{\mathrm{EX,12h}} + U$ [42]. This means that $L_{\mathrm{EX,12h}} + U$ will be above 95 % of the values. U is calculated as

$$U = ku \tag{4.1}$$

where the coverage factor k is 1.65 for the 95 % confidence interval and u is the combined standard uncertainty [42]. The squared combined standard uncertainty is defined as

$$u^2 = \sum c_i^2 u_i^2 \tag{4.2}$$

where u_i is the standard uncertainty and c_i is the sensitivity coefficient [42].

4.2 Measurement Standards

The work aims to measure the acoustic conditions in the work environment in the operator control room and to assess them against the action and limit values established in the Norwegian standard NS 8175:2019. To measure SPL this work will follow the international measurement standards ISO 9612:2019.

4.2.1 ISO 9612:2009

The noise exposure measurement in the operator control room is intended to meet the requirements established in the international measurement standard ISO 9612:2009, an engineering method for measuring occupational noise exposure in work environments. Underscored requirements are pre-analysis of the work environment (including identification of jobs and their corresponding tasks), detailed description of the measurement procedure, and quality checking of the measurements performed. The standard also provides methods for estimating the uncertainties in the measured noise levels.

5
Methods

The methods used in this work include measurement methods in the field (the open-office environment and the headsets), as well as the signal processing methods for the enhancement of sound conditions the operator headsets. The chapter follows a practically chronological structure, beginning with a work analysis before describing the measurement procedures, the filtering methods and the evaluation methodology.

5.1 Work Analysis: Operator Control Room and Headset Measurements

Before conducting the measurements in the alarm center, there is a requirement to make a work analysis [42]. This preparation procedure involves steps to ensure an appropriate approach according to the international measurement standards mentioned in Chapter 4. A thorough work analysis can additionally help possible troubleshooting and act as a guide for further measurements if any aberration from the standards occurs in the first measurements.

5.1.1 Orientation

The initial step of the work analysis is the orientation. It can be interpreted as a pre-analysis of the field of measurements, or as the structure of the work analysis. In this work, the orientation is used as the latter option. The steps of the work analysis are listed below:

- Identify jobs and specific work activities at the workplace.
- Determine homogenous noise exposure groups.
- Determine the nominal workday for the homogenous noise exposure groups.
- Identify significant noise sources.
- Choose a measurement strategy.
- Plan the measurements.

5.1.2 Work Activities and Jobs

The initial phase of the work analysis involves systematically mapping various job roles and their associated work activities. Within the alarm center, three primary roles can be identified: workplace manager, shift leader, and operator. This study

focuses on the acoustic conditions within the operator control room; therefore, the following work activities are specifically related to the role of an operator. The following activities are related to the job role of an operator:

- Receiving and managing incoming calls
- Monitoring and responding to alarms in the control room
- Communicating via headset with callers and other alarm centers
- Coordinating with colleagues, shift leaders, and management
- Logging and reporting significant events or operational issues.

5.1.3 Homogenous Noise Exposure Groups

The operators are gathered in the same group, because they can be considered to perform equal tasks and to be involved in similar work activities. The groups are previously presented in Table 4.1, and concerning the type of work environment, the suitable choice is Group 1. This choice has to be verified and confirmed by the employees and the employers. The required noise exposure level of 55 dBA is determined to facilitate conditions for effortless concentration and attention during work. It is furthermore considered that the workers in Group 1 have no control over existing sound and noise sources.

5.1.4 Choice of Measurement Strategy

In ISO 9612:2009 three strategies exist for conducting measurements: task-based-, job-based- and full-day measurements [42]. The task-based measurement strategy is applicable for contexts where noise-producing tasks can be clearly defined, while the job-based measurement strategy is suitable for when the tasks are difficult to describe or for when no work analysis has been made [42]. The full-day measurement strategy is the recommended method for complex or unpredictable noise environments [42]. Since the problem with alarm noise occurs irregularly and unpredictably, full-day measurements will be complemented with a task-based measurement in this work.

5.2 Measurements

This section describes the procedure followed for measurements at the Fire and Rescue Service 110-Central. The measurements were divided into two categories:

- Full-day measurements of equivalent noise levels in the sound environment inside on-ear headsets.
- Task-based measurement of noise levels inside the headsets during a stress-test involving an emergency call.

Due to the risk of handling sensitive information, no recordings were taken during the full-day measurements, instead, only sound level data was stored. A simu-

lated emergency call was conducted to ensure the possibility of recording sound during the task-based measurement. Thus, no sensitive personal information could be stored. The measurements were according to the guidelines and appeals stated in ISO 9612:2009, but adapted to the specific situation of the headset environment. The standards are tuned for room measurements, demanding interpretation corresponding to the headset environment in this case. The headset used was of on-ear type, meaning that the space inside the headset is not fully separated from the room. This unsealed environment is extended even more due to the resulting alternative microphone setup stated in Section 5.2.2. The measurement procedures are described in detail in Section 5.2.4 and Section 5.2.5, including considerable deviations from the standard guidelines.

5.2.1 Equipment List

Table 5.1 lists the equipment for full-day measurements in the emergency call center office:

Table 5.1: List of equipment used for the 1-week log of full-day measurements.

Equipment	Type	Serial Number
1 Field data recorder	Nor140 (BS21)	1407977
1 Microphone	B&K Type 1209A	23815
1 Sound calibrator	Nor1251	35012
1 Artificial head	HMS II.3 LN-HEC	17032133
1 Stand	HTB VI	15740595
1 Wireless on-ear Headset	EPOS SDW 60	Unknown
1 Cable	BNC	-

The equipment listed in Table 5.2 was used for the task-based alarm call measurement in the emergency call center office:

Table 5.2: List of equipment used for the task-based measurements during a simulated emergency call.

Equipment	Type	Serial Number
1 Field data recorder	Nor140 (BS21)	1407977
1 Field data recorder	Nor140 (BS15)	1406406
1 Microphone	B&K Type 1209A	23815
1 Microphone	B&K Type 1209A	12043
1 Sound calibrator	Nor1251	35012
1 Artificial head	HMS II.3 LN-HEC	17032133
1 Stand	HTB VI	15740595
1 Wireless on-ear Headset	EPOS SDW 60	Unknown
2 Cables	BNC	-

5.2.2 Measurement Complications Disclosure

The field measurements were planned to be conducted with the built-in microphones of the HMS II.3 artificial head to capture valid sound levels experienced in human ears. However, during setup and calibration it was observed that the microphones were defective, capturing background noise levels above 80 dB. Due to this complication, an alternative setup was required. The unfortunate situation offered one solution: mounting the B&K microphones on the ear canals of the HMS II.3, directed as good as possible towards the headset membranes. This solution obviously introduced increased measurement errors: first, the enclosure initially formed by the headset and the ear was now open. Secondly, the receiver was now located closer to the sound source, which could affect the captured noise levels. Despite these uncertainties, measurement attempts were made, with the risk of eventually deeming the measurements invalid. This risk was considered particularly high for the measurement of signal levels during emergency calls.

5.2.3 Calibration

Before each measurement, the used microphones were calibrated with the Norsonic Sound Calibrator at reference SPL 114 dB. To deem measurements as valid, the calibration of a microphone must be within ± 0.5 dB of the reference.

5.2.4 1-Week Log: Full-Day Measurements

The full-day measurements took place between 2025/04/01 - 2025/04/08 in the emergency hotline 110-Central of the Oslo Fire and Rescue Services. Noise levels in a wireless on-ear headset were logged during 1 week, with a few occasional interruptions due to battery discharge.

The measurement was conducted in a closed office space with 2 workstations. However, to mimic real conditions at work, the door was left open to the hallway outside. As the headset is not fully covering the ear, the soundfield surrounding the ear is considered to include the room.

The Nor140 (BS21) unit, equipped with B&K microphone, mounted at the location of the left ear of the HMS II.3 logged noise levels constantly during the week. The data recorder was configured to sync the measured sound levels every second over a period of 9 minutes and 58 seconds. With 1 second for buffer and 1 additional second to trigger the next period, a new measurement period was started every 10 minutes. The data was logged at a 12 kHz sample rate and each 10 minute-period was saved as an NBF-file, which was converted to Microsoft Excel.

5.2.5 Task-Based Measurement: Headset Signal Levels during Emergency Alarm Call

The task-based measurement was taken on 2025/04/09. The measurement plan was set together with Sigmund Olafsen (Brekke & Strand), Roy Kristoffersen (Oslo BRE), and Miriam Askeland Thuen (Oslo BRE). Due to the uncertainty of when an incoming call would occur in the real world, it was decided to simulate such a situation. This solution additionally respects the undesired handling of sensitive personal information. The measurement was carried out in the same room, which was used during the 1-week measurement. The caller triggered both a fire alarm and a smoke detector alarm during the call, and the signal was measured by the microphones and stored by the Nor140 BS15 (right ear), and Nor140 BS21 (left ear) at 44.1 kHz sample rate, 16-bit resolution and syncronised each second. The alarm call lasted 4 minutes and 41 seconds and the data was stored in one NBF-file.

5.2.6 Measurement Data Acquisition

All measurement data was saved to the field data recorders on SD-cards as NBF-files. They were transferred to the computers via NorXfer, a software able to handle NBF-files and to convert the data to Microsoft Excel. Due to the high dynamic range of NBF-files, a software like Norreview is is usually required for audio playback. It was therefore decided to record in 16-bit resolution to make them playable on any sound interface. The list below presents the quantities acquired from the measurements.

- $L_{\mathrm{A,eq}}$
- $L_{f,\mathrm{eq}}$
- $L_{\mathrm{A}f,\mathrm{max}}$
- $L_{\mathrm{C,peak}}$

Here, f is the 1/3rd octave bands between 6.3 Hz - 20 kHz.

5.3 Tonal Noise Suppression in Emergency Communication: Filtering Method

This section of the methodology chapter describes the procedures for the design of the proposed filter prototypes. It starts with an overview and explanation of the model-based approach, followed by a chronological step-by-step structure in the same chronological order as the stages of the processing chain. As stated in Chapter 2, the used methods are based on theoretical framework for cepstral alanysis and LPC-analysis. Two proposals have been designed, and their deviations from each other are clearly distinguished under each subsection.

5.3.1 Filtering Objectives and Methodological Approach

This section outlines the methodology developed to suppress tonal alarm interference in speech signals. Such alarm sounds are typically periodic and often overlap temporally and spectrally with voiced speech, posing a challenge to both intelligibility and listening effort.

Two distinct alarm categories are addressed:

- **Frequency-sweeping (non-stationary) alarms**, where the fundamental frequency and its harmonics change over time;
- **Constant-frequency (stationary) alarms**, where tonal interference remains fixed or slowly varying within frames.

The filtering strategies are based on the theoretical framework of source–filter modeling presented in Chapter 3, with particular emphasis on cepstral alanysis for harmonic detection and Linear Predictive Coding (LPC) for filter modeling and resynthesis. The work presents two approaches:

- **Proposaed filter A** is solely based on cepstral liftering, while
- **Proposed filter B** utilises a combination of ceptral alanysis and LPC-based suppression.

The remainder of this section is structured by chronological steps in the DSP-chain. All filtering algorithms were implemented in MATLAB using the *Signal Processing Toolbox* to support windowing, spectral analysis (STFT and cepstrum), and predictive filtering (LPC). Additionally, HEAD acoustics ArtemiS Suite 16.7 was used during the measurement phase to compute SPL trends and generate spectrogram representations for offline inspection.

5.3.2 Pre-Processing

Input Signal
The present implementation of the system loads a user-specified sound file and converts it to a single-channel signal for processing. This temporary design facilitates iterative testing of preprocessing and filter behavior across multiple recordings. During development, one clean speech signal was used against two different tonal alarm signals separately. See Table 5.3 below for detailed information.

Table 5.3: Charachteristics of tonal alarm noise used during filter development.

Alarm Cateogry	Fundamental	Harmonics	Temporal Pattern
Frequency-sweeping	0.8 – 1 kHz	Yes (Sweeping)	Constant
Time-patterned	3.36 kHz	Yes	T3

Framing
All implementations process the signal using short-time analysis. The input is

framed into overlapping segments using a 25 ms Hann window with an 8 ms hop size (68% overlap). This setup was chosen iteratively to balance temporal resolution with frequency selectivity and for smooth overlap-add reconstruction post filtering. For consistency, Hann windowing is used in the resynthesis-step. Figure 6.18 shows a visual presentation of the frame handling. A visualised windowing scheme is shown in Figure 6.18.

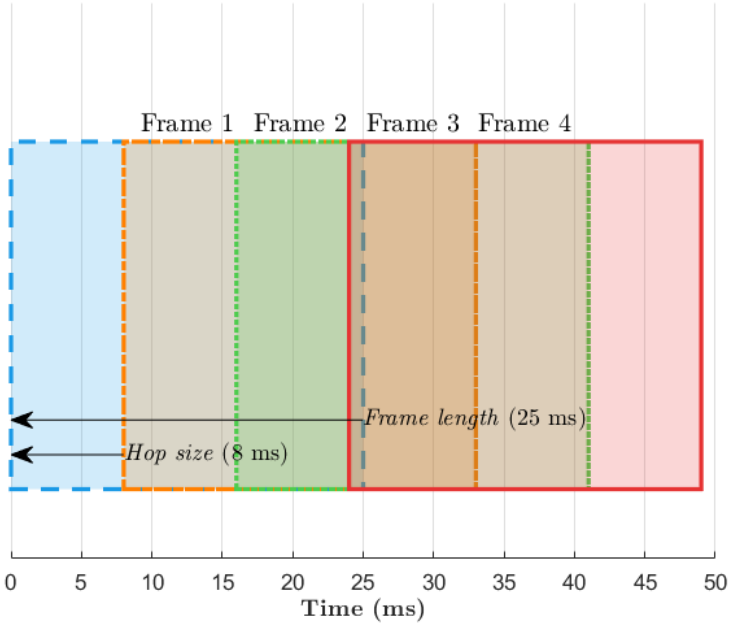


Figure 5.1: Signal framing with 25 ms frame length and 8 ms hop size (68% overlap). The figure displays how the first four frames relate to each other.

5.3.3 Overview and Processing Strategy

The suppression system developed for sweeping alarm signals operates on short-time frames and uses pitch information (estimated via cepstral analysis) to guide the adaptive filter. These alarms are characterized by frequency-varying, harmonically spaced tones, and suppressing them requires a filter that is capable of tracking pitch trajectories. The proposed framework integrates several techniques into a hybrid architecture:

- **Pitch tracking**: Estimated from cepstral analysis and used to guide suppression of the fundamental and the harmonics.
- **Tonal frame detection**: Based on pitch detection and spectral flatness.
- **Suppression domains**: Cepstral liftering, LPC residual filtering, or a combination of both.
- **Reconstruction and gain restoration**: Suppressed frames are rescaled and overlap-added to the output signal.

The system was developed iteratively, beginning with *Version 1* (cepstral liftering

only), then incorporating LPC-based filtering in *Version 2*, and eventually converging on a hybrid filtering strategy in *Version 3*. The following subsections describe the design and logic of each processing stage in detail.

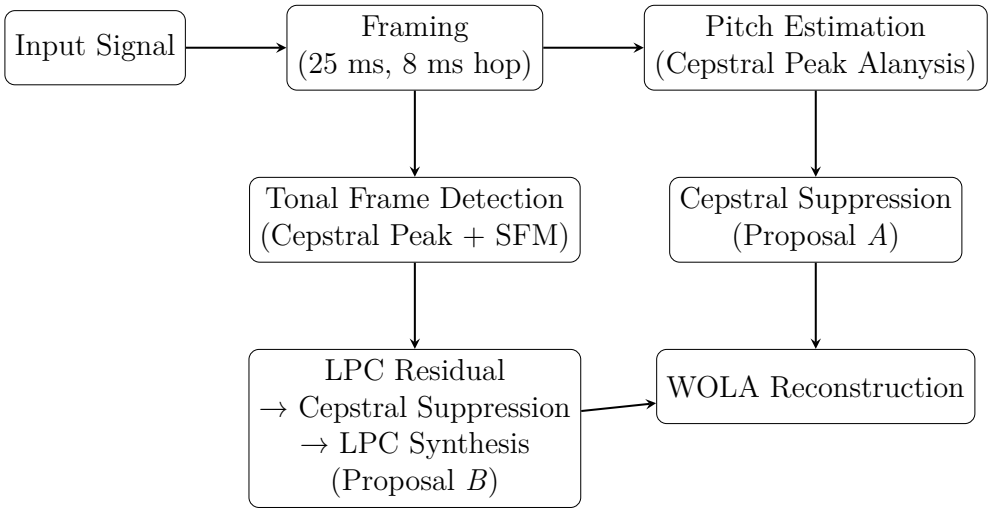


Figure 5.2: Block diagram of the proposed alarm suppression systems (A and B).

5.3.4 Pitch Tracking and Tonal Frame Classification

The pitch tracking module is designed to detect time-varying fundamental frequencies typical of sweeping alarms. The initial logic is reflected in Proposal *A*, and its evolution is present in Propsal *B*.

Proposal A: Basic Cepstral Peak Detection

Each frame undergoes cepstral analysis, where the pitch is estimated from the peak quefrency within a defined range (corresponding to 500–1500 Hz). To stabilize tracking, temporal smoothing is applied for each frame. If the pitch is undefined or zero, the most recent valid estimate is retained. The first-pass stage of the processing chain can thus be summarised accordingly:

- Compute cepstrum from log-magnitude FFT.
- Estimate pitch from the highest peak in the quefrency band corresponding to the investigated frequency region of 500-4000 Hz.
- Apply recursive temporal smoothing.

Proposal B: Hybrid Decision Using SFM

In the second iteration, a two-factor decision scheme is introduced. A frame is considered tonal if:

- A pitch in the search range mentioned above is detected, and
- SFM is below the mean value.

This is a simplified concept that allows the system to distinguish between harmonically rich and modulated alarm tones, while ignoring broadband or speech-like aperiodic noise.

5.3.5 Adaptive Suppression Techniques

This section describes the evolution and rationale behind the suppression applied to tonal noise in the signal.

Proposal A: Cepstral Liftering with Harmonic Scaling

In the first iteration, suppression is applied directly in the quefrency domain. The pitch estimate determines the quefrency of the fundamental frequency, and harmonics are modeled by its integer multiples. These harmonics are attenuated by damping cepstral coefficients around each peak. Equal attenuation factor is used on all harmonics, to test if cepstral liftering can be used to effectively manipulate periodic alarm components without distorting speech components. However, the suppression width (expressed as quefrency values) scales with harmonic index, due to increased Δf with increasing harmonic index.

Proposal B: LPC Residual Suppression

The second approach investigates the combination of cepstral liftering and LPC. In Proposal B, cepstral suppression is applied in the residual domain, estimated by a programmed LPC-filter. For each tonal frame:

- A 14^{th}-order LPC-vector is estimated.
- The frame is inverse-filtered to extract the excitation (residual).
- Cepstral suppression (as described in Proposal A) is applied to the residual.
- The modified residual is resynthesized using the pole-pruned stabilised LPC filter.

This approach attempts to isolate tonal energy embedded in the excitation source while preserving the vocal tract filter (where formant structure resides). This method can introduce potential instability if the LPC model becomes ill-conditioned. To address this, reflection margin stabilization and pole pruning are used to ensure filter robustness during synthesis.

LPC Filter Stabilization

In frames where LPC-based suppression is applied, reconstruction requires re-synthesis through the frame's LPC filter. However, early testing revealed that unstable LPC filters could cause either large output spikes or complete dropouts. To address this, the following stabilization measures were implemented:

- Reflection coefficient margins are enforced to ensure all poles remain within a safe radius.

- Pole pruning is applied to remove spurious roots near the unit circle.

These safeguards are considered to ensure that the re-synthesis remains perceptually smooth and numerically stable.

5.3.6 Frame Reconstruction

The final output signal is constructed via window overlap-add (WOLA) synthesis using the same Hann window and hop size as in the analysis stage. This enables smooth transitions between frames and minimizes discontinuities at frame boundaries. The WOLA-procedure is applied accordingly:

- The compensated frame is multiplied by the original analysis window.
- The frame is overlap-added into its corresponding position in the output signal buffer.
- After all frames are processed, the output signal is normalized to prevent clipping.

5.4 Webex: Adaptive Noise Suppression using AI

This work intends to compare the proposed filtering methods presented in Section 5.3 to state-of-the-art technology within the scope of the thesis. Today, adaptive noise suppression is typically used in digital communication systems, such as video conferencing softwares. For this work, Webex by Cisco is used as the reference level of filter performance. This section describes the steps taken to implement Webex adaptive noise suppression on the test files (see Section 5.3.1 and Section 5.3.2).

5.4.1 Virtual Audio Interface: LoopBeAudio

To play sound files internally on the computer, one can use a virtual audio interface. It replaces physical audio interfaces and can be used as both an audio input- and output device. In this work, the virtual audio interface *LoopBeAudio* by Daniel Schmitt was used. It is a virtual audio device to transfer audio between computer programs, digitally, without any quality loss [47].

5.4.2 Noise Suppression in Webex

The noise suppression test was then conducted within a Webex video call. The following audio settings were used:

- Microphone input: LoopBeAudio (internal driver)
- Audio output: LoopBeAudio (internal driver)
- Noise Suppression: On (default mode)

Next, a recording of the video call was started. The test file was played on the default media player used on a personal computer (DELL), and the adaptive noise

suppression was automatically applied. The recording was stopped and saved as a .mp4-file which was manually converted into a 16-bit, 44.1 kHz .wav-file.

5.5 Normalisation of Sound Levels before Quality Evaluation

To achieve feasible comparability of signal quality between the original and processed sound files, audio normalisation was carried out individually. First, each file was normalised to its maximum signal level to avoid audio clipping. Then, as a further step, all sound files were staged to equal rms-levels and true peak values.

5.6 Computation of Objective Quality Measures in MATLAB

The calculations of the objective quality measures WSS, $\text{fwSNR}_{\text{seg}}$ and LLR were made in Matlab using various scripts from [46]. Before the computation could be made, the audio signals were converted from stereo to mono format. The audio signals that were analysed were 1 minutes and 48 seconds long. Each of the of the scripts used the original uncorrupted speech signal as the clean reference signal.

For the computation of the WSS the following is applied in the Matlab script [46]:

- 25 critical bands in the 4 kHz bandwidth ranging from a center frequency of 50 Hz with bandwidth of 70 Hz for the first one to a center frequency of 3597.63 Hz and a band width of 346.136 Hz for the 25th.
- Gaussian filters are used as the critical band filters.
- Hanning windows are applied to the frames.
- As suggested in [32] it uses the values of $K_{\text{max}} = 20$, $K_{\text{loc.max}} = 1$ and $K_{\text{SPL}} = 0$.

For the computation of the $\text{fwSNR}_{\text{seg}}$ the following is implemented in the script [46]:

- The same critical bands were used as for the WSS.
- Gaussian filters are used as the critical band filters.
- Hanning windows are applied to the frames.
- Each spectrum is normalized to have the area of 1.
- The power exponent γ is set to 0.2 when computing the weighting function $W(j, m)$ as suggested in [33] and described in Section 3.3.2.

For the computation of the LLR the following is used in the script [46]:

- An LPC-order of 10 for signals with sampling frequencies below 10 kHz, for a signal with a sampling frequency above that, an LPC-order of 16 is used.
- Hanning windows are applied to the frames.
- The Levinson-Durbin algorithm is used to compute the autocorrelation matrix.

5.7 Listening Test

In order to more accurately evaluate the filters' effect on the test signals and especially the subjective aspects of the effects, a listening test was conducted. During the test, which was conducted over three days, 16 test persons - 9 males and 7 females between the ages 21 and 30 and with normal hearing - got to listen to different versions of the same audio signal. They did this as they were seated alone in a quiet room without any distractions. The duration of the test was around 15 to 20 minutes per test person.

The audio signal they got to listen was a five second segment with a male reading the following section from *The Rainbow Passage* from [48]:

> *Aristotle thought the rainbow was caused by reflection of the sun's rays by the sun*[1]

The test persons were exposed to this five second snippet both with the addition of the sweeping fire alarm noise and the addition of the frequency stable smoke detector noise respectively. These two files were then presented as they were, but also through the various filters that have been described. They were presented in the form of a paired comparison A/B half matrix test in the HEAD acoustics' jury testing software SQala. Before the test started, they were instructed as follows:

> You will be asked to rate sound A and sound B based on various parameters. You can listen to each sound once. You can only rate the sound after you have listened to each sound. Press on the play buttons for A and B to listen to the sounds. You will then get to rate them based on the questions you will be asked.

The test then consisted of comparisons of each of the files based on these six questions in the following order:

1. Which signal requires MORE effort from you for you to understand the speech?
2. Which signal makes you MORE frustrated trying to understand the speech?
3. Which signal do you perceive to have the MOST unnatural speech?
4. Which signal do you perceive as the loudest?
5. Which signal do you perceive as the MOST annoying?
6. Overall, which signal do you consider to be the worst?

[1] The sentence is originally: *Aristotle thought* ***that*** *the rainbow was caused by reflection of the sun's rays by the* ***rain*** [48].

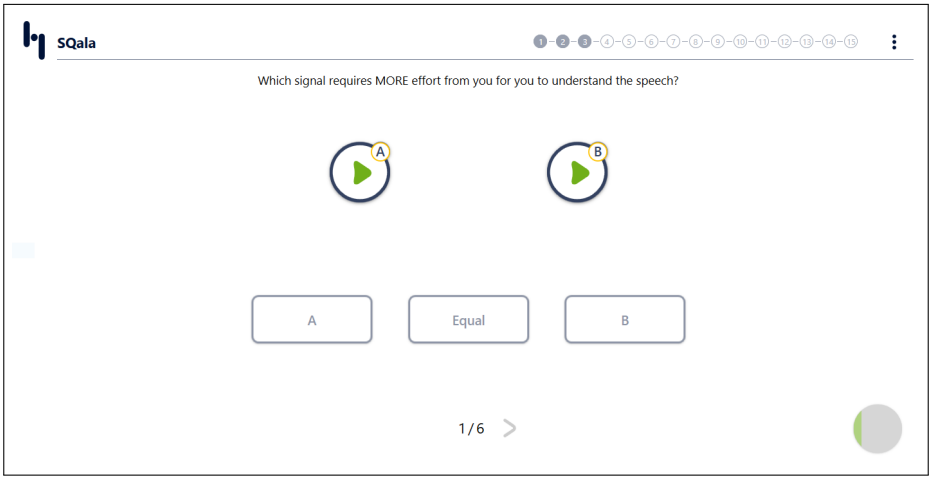


Figure 5.3: The participants of the listening test got to answer the question as A, B or Equal (A=B).

For the comparison between the four fire alarm degraded signals, each question resulted in six paired comparisons as shown in Table 5.4 below.

Table 5.4: Example of one test person's answers to the question: "Which signal requires MORE effort from you for you to understand the speech?" for the fire alarm degraded signals in the form of an A/B half matrix.

	B: Original	**B: Webex**	**B: Filter A**	**B: Filter B**
A: Original	-	Equal	B	B
A: Webex	-	-	B	Equal
A: Filter A	-	-	-	A
A: Filter B	-	-	-	-

For the comparison between the three smoke detector noise degraded signals, each question resulted in three paired comparisons as shown in Table 5.5 below.

Table 5.5: Example of one test person's answers to the question: "Which signal requires MORE effort from you for you to understand the speech?" for the smoke detector degraded signals in the form of an A/B half matrix.

	B: Original	**B: Webex**	**B: Proposed filter**
A: Original	-	A	B
A: Webex	-	-	B
A: Proposed filter	-	-	-

6
Results

This chapter holds the results obtained from on-site measurements, analysis and filtering in MATLAB, and the evaluation of filter performance.

6.1 Alarm Filtering

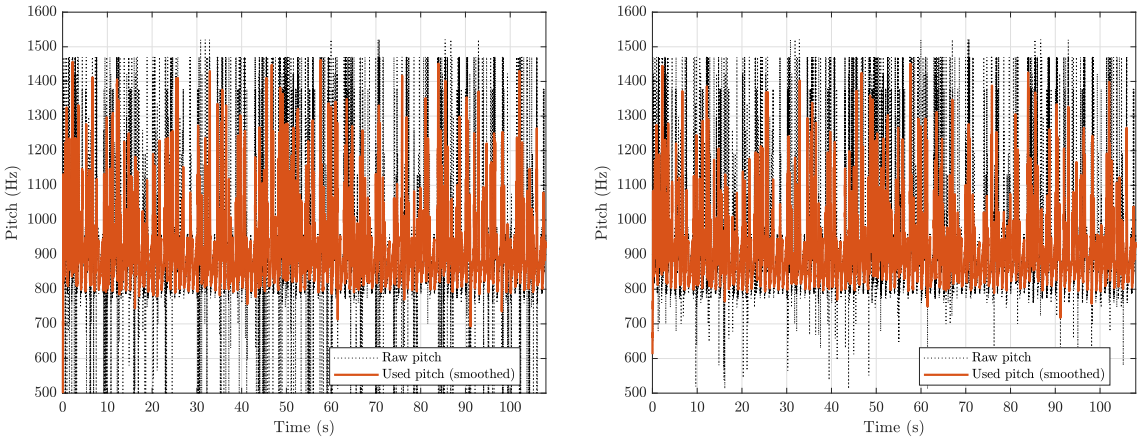
Filtering was carried out on two different speech- and alarm combinations. The processing results for the case of the frequency-sweeping alarm are presented in 6.1.1, while the corresponding results of the frequency-stable, T3-temporal alarm are shown in 6.1.2.

6.1.1 Sweeping Alarm: 0.8 - 1 kHz

The proposed filters were analysed directly in MATLAB by plotting pitch tracker-detections, spectograms of the signal, and its resulting waveform. Figures 6.2 and 6.3 display a comparison between the unprocessed signal, the filtered signal (Proposal *A* and Proposal *B*), and the state of the art filter provided by Webex.

6.1.1.1 Pitch Tracking

The pitch trackers were given a search range between 500 - 1500 Hz. A distinction can be made between filter A (Figure 6.1a) and B (Figure 6.1b, where the raw pitch (unsmoothed) has fewer bursts below 800 Hz. However, the smoothed versions do not show a great difference. The scores above 1000 Hz indicate the instability of the pitch tracking module, potentially leading to over-processing of non-alarm content in the signal, and under-processing of the alarm in affected frames.



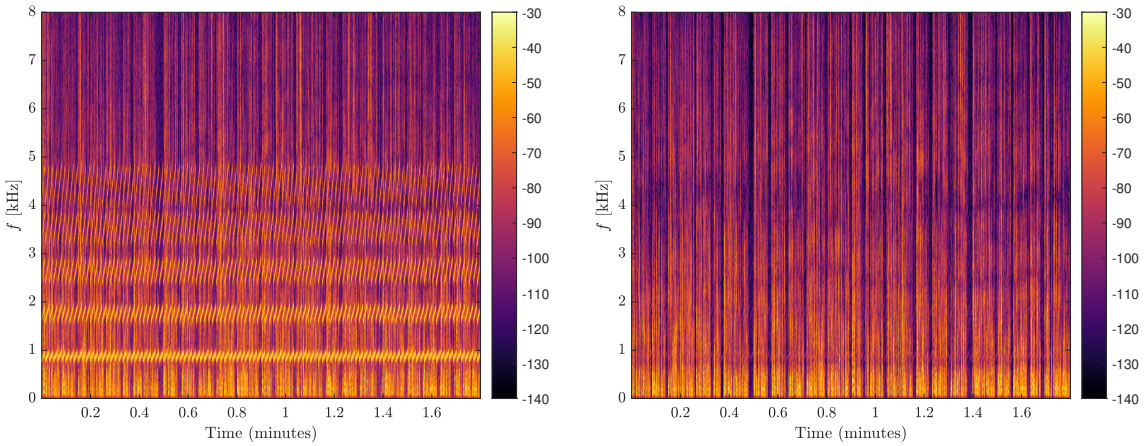
(a) Proposed filter A.

(b) Proposed filter B.

Figure 6.1: The integrated pitch tracker, detecting pitch over time. The figure displays both raw and smoothed values.

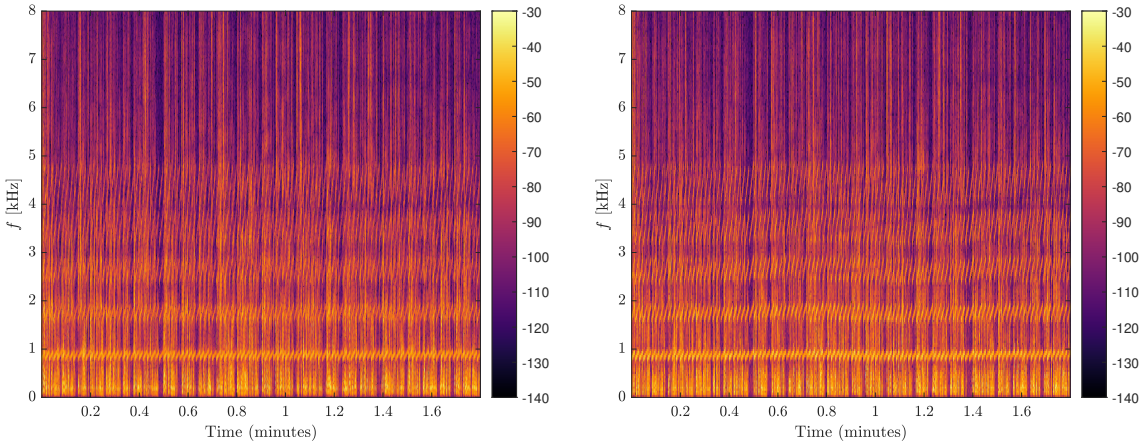
6.1.1.2 Spectograms

The spectograms in Figure 6.2 reveal no considerable difference between filter A and B (Figures 6.2c and 6.2d). Compared to the unprocessed signal visible in Figure 6.2a, gentle suppression is applied on the alarm fundamental and its harmonics. Speech below 1 kHz remains relatively unaffected. However, the Webex suppressor (Figure 6.2b) stands superior, where speech remains intact and alarm harmonics are almost completely removed.



(a) Original.

(b) Webex suppression.



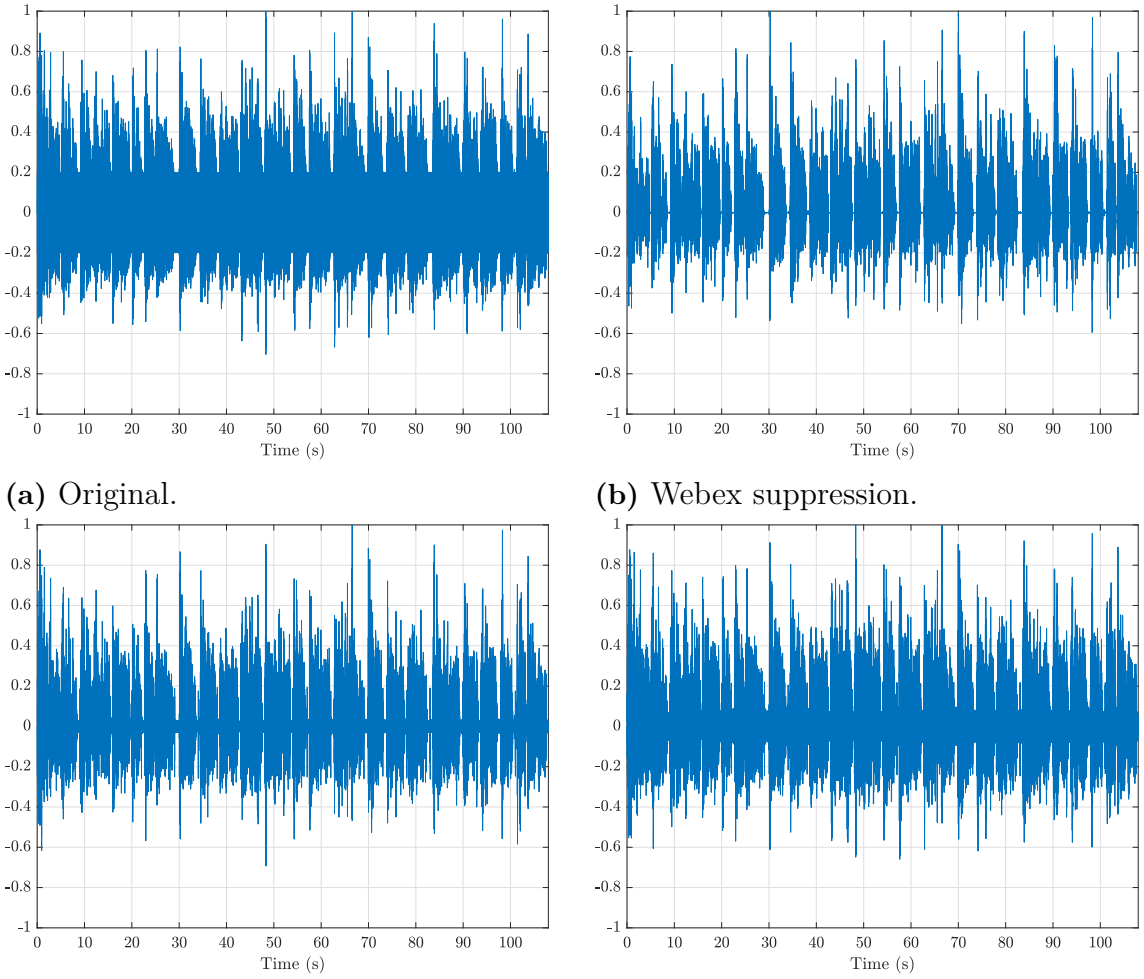
(c) Proposed filter A

(d) Proposed filter B.

Figure 6.2: Spectogram comparison between original and processed signals.

6.1.1.3 Waveforms

The waveforms have been compared and they are displayed in Figure 6.3 below. Transients are generally preserved in all filters, but there is not much to say on how well the speech quality is preserved from these plots. However, the constant sweeping alarm can be distinguished in regions where speech is absent. The alarm tone is visible as a spine between major transients. This can reveal the relative amount of suppression of alarm components in the signal. Once again, the Webex filter (Figure 6.3b) proves to be the most efficient filter, keeping speech-absent frames almost silent. By comparing filter A and B, the thinner waveforms between transients in Figure 6.3c reveal more effective suppression of alarm noise in filter A.



(a) Original.

(b) Webex suppression.

(c) Proposed filter A.

(d) Proposed filter B.

Figure 6.3: Waveform comparison between original and processed signals.

6.1.2 Frequency-Stable Alarm: 3.36 kHz

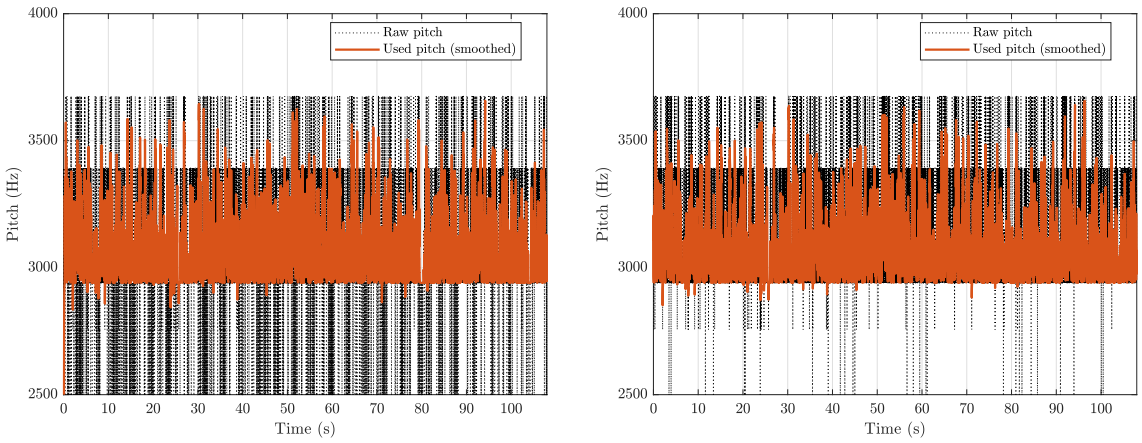
In this case, the alarm is at a constant frequency of 3.36 kHz and operates according to the T3-temporal pattern:

- **ON** for 0.5 seconds,
- **OFF** for 0.5 seconds,
- **ON** for 0.5 seconds,
- **OFF** for 0.5 seconds,
- **ON** for 0.5 seconds,
- **OFF** for 1.5 seconds.

After the final pause, the pattern repeats itself. At a rate of approximately every 20 seconds (depending on the time stamp for each repetition), the level of the alarm increases with 6 dB, resulting in the horn-like shape of the waveform in Figure 6.6.

6.1.2.1 Pitch Tracking

Regarding the performance of the pitch tracker module when searching for a frequency-stable alarm (Figure 6.4, a similar ascertainment can be done as for the case of a frequency-sweeping alarm (Figure 6.1). The raw pitch tracker is more accurate in filter *B* than *A*. However, the smoothing process of the pitch detection leads to a negligible difference between the pitch trackers.



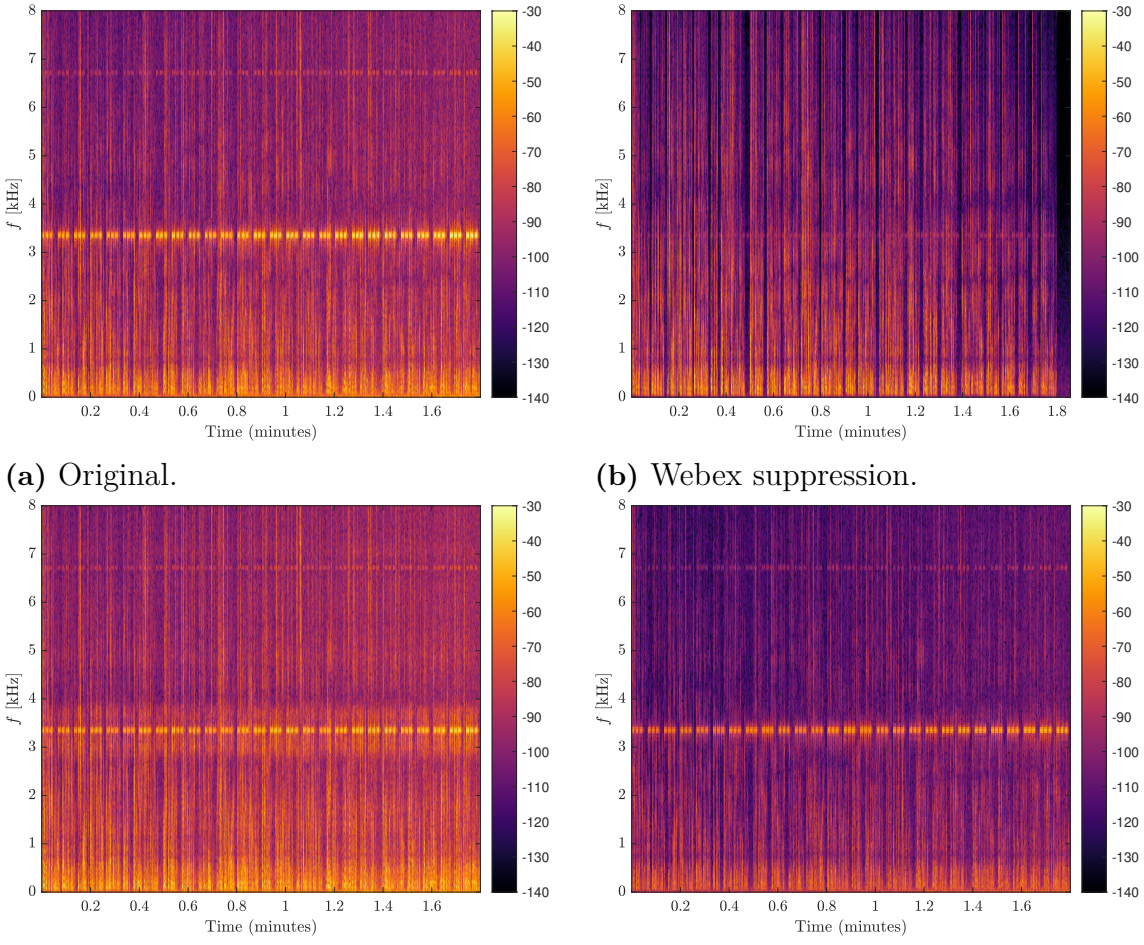
(a) Proposed filter *A*.

(b) Proposed filter *B*.

Figure 6.4: The integrated pitch tracker, detecting pitch over time. The figure displays both raw and smoothed values.

6.1.2.2 Spectograms

Compared to the similar performance of the pitch tracking modules, a greater difference between filter performances can be observed when studying the spectograms in Figure 6.5 below. Filter B (Figure 6.5d) acts more aggressively on the signal compared to filter A. Both the alarm and general signal has been suppressed more heavily, hinting that the LPC-filter has greater influence on the signal in this case, compared to the case of a sweeping alarm. Filter 6.5c applies gentle suppression of the alarm tone, as the level increase (visible in Figure 6.5a) is dampened. Another effect of the filter is smearing of the fundamental alarm frequency, visible as bright shades above and below 3.36 kHz. By comparing the filters with the Webex filter, it becomes clear that the latter is more effective. In Figure 6.5b, the alarm components are almost completely removed, similar to Figure 6.2b.



(a) Original.

(b) Webex suppression.

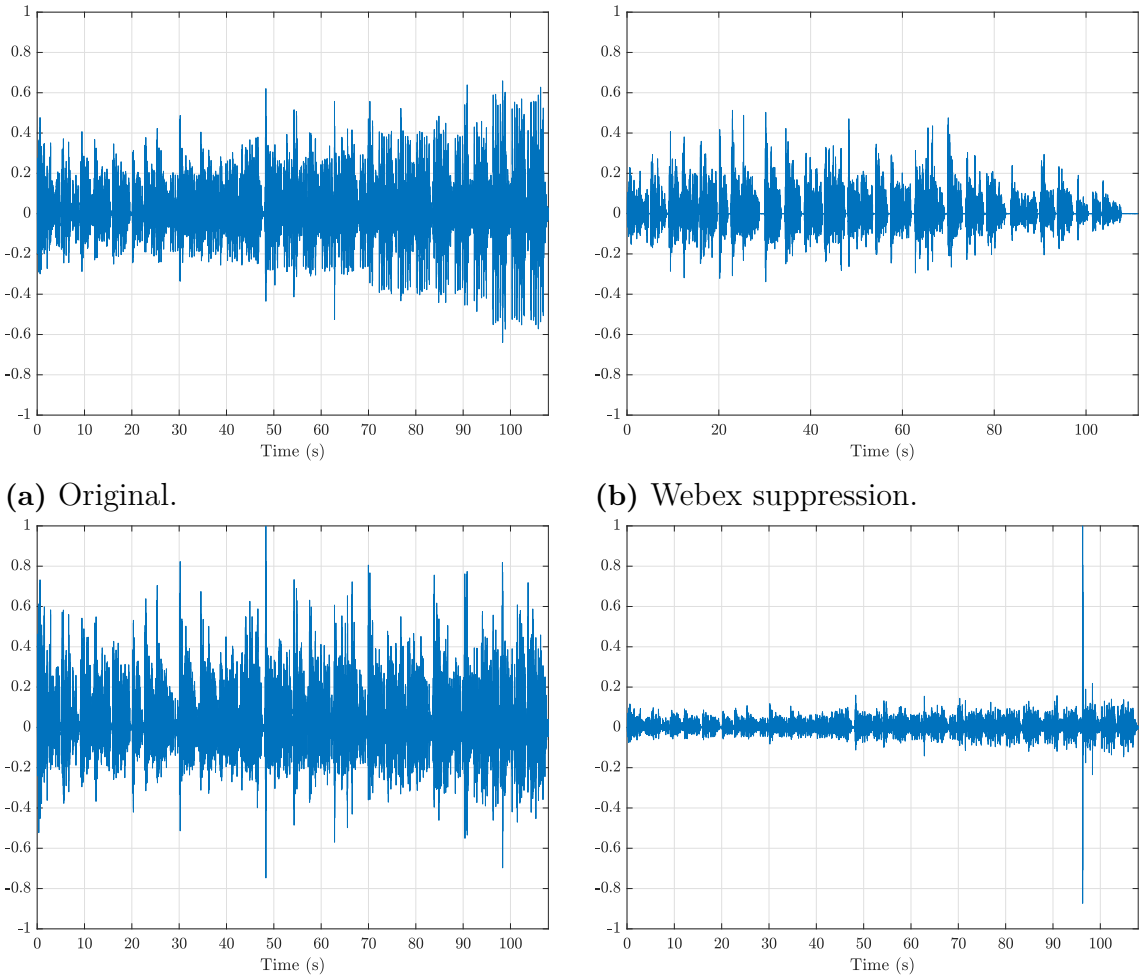
(c) Proposed filter A

(d) Proposed filter B.

Figure 6.5: Spectogram comparison between original and processed signals.

6.1.2.3 Waveforms

The difference in the filtering results of filter A and B becomes even clearer in Figure 6.6 below. Filter B causes over-suppression on the entire signal (Figure 6.6d, where even the transients are heavily suppressed. In this case, filter A is favourable when it comes to preservation of the original signal. However, the amount of suppression on the alarm noise can not be evaluated in Figure 6.6c, as alarm noise can be represented both transients, in addition to other regions.



(a) Original.

(b) Webex suppression.

(c) Proposed filter A.

(d) Proposed filter B.

Figure 6.6: Waveform comparison between original and processed signals.

6.2 Objective Quality Evaluation

The objective quality measures for three out of the four audio signals that were degraded due to the addition of the sweeping fire alarm noise and two out of the three audio signals that were degraded due to the addition of the frequency stable smoke detector noise were computed. This could not be done for the signals that were filtered through Webex. This is primarily due to the problem that occurs when playing and recording the signal through Webex in that it slightly warps the signal in

the time domain. This causes the filtered signal to be out of sync with the reference signal which is required when computing the following numerical measures. The results for the original fire alarm noise degraded signal, as well the same signal filter through proposed filter A and proposed filter B and the original smoke detector noise degraded signal and the same signal filtered through the proposed filter. For both the WSS and LLR a higher numerical value means a higher level of degradation and for the $\text{fwSNR}_{\text{seg}}$ and lower value means a higher level of degradation.

Table 6.1: The computed numerical values of WSS, $\text{fwSNR}_{\text{seg}}$ and LLR for the original audio signals as well as the ones applied with the proposed filters. Note that it wasn't possible to compute the numerical values for the signals that were filtered through Webex.

Audio signal	**WSS**	**$\text{fwSNR}_{\text{seg}}$**	**LLR**
Original (fire alarm noise)	290.6488	2.6294	1.0169
Webex (fire alarm noise)	-	-	-
Proposed filter A (fire alarm noise)	66.3059	12.7058	0.3383
Proposed filter B (fire alarm noise)	151.3129	8.1685	0.6800
Original (smoke detector noise)	69.9050	11.0998	0.6437
Webex (smoke detector noise)	-	-	-
Proposed filter (smoke detector noise)	64.0734	9.7468	0.9345

Ranking according to WSS (best to worst)

1. Proposed filter (smoke detector noise) - 64.0734
2. Proposed filter A (fire alarm noise) - 66.3059
3. Original (smoke detector noise) - 69.9050
4. Proposed filter B (fire alarm noise) - 151.3129
5. Original (fire alarm noise) - 290.6488

Ranking according to $\text{fwSNR}_{\text{seg}}$ (best to worst)

1. Proposed filter A (fire alarm noise) - 12.7058
2. Original (smoke detector noise) - 11.0998
3. Proposed filter (smoke detector noise) - 9.7468
4. Proposed filter B (fire alarm noise) - 8.1685
5. Original (fire alarm noise) - 2.6294

Ranking according to LLR (best to worst)

1. Proposed filter A (fire alarm noise) - 0.3383
2. Original (smoke detector noise) - 0.6437
3. Proposed filter B (fire alarm noise) - 0.6800
4. Proposed filter (smoke detector noise) - 0.9345
5. Original (fire alarm noise) - 1.0169

According to these objective quality measures, the filter that gets the best results is proposed filter A, as it ranks the best according to both the LLR and the $\text{fwSNR}_{\text{seg}}$ at the same time as it ranks the second best according to the WSS measure. The signal with the second overall best ranking is the orignal smoke detector noise degraded signal, as it ranks the second best according to both the $\text{fwSNR}_{\text{seg}}$ and the LLR. For the WSS it gets ranked on the third place. The third overall best signal could be argued to be the filtered smoke detector noise degraded signal as it ranks on place 3 according to the $\text{fwSNR}_{\text{seg}}$ and on place 4 according to the LLR but as the least degraded signal according to the WSS. Overall, the second most degraded signal according to these measures could be argued to be the fire alarm noise degraded signal filtered through proposed filter B as it ranks as the second most degraded according to both the WSS and the $\text{fwSNR}_{\text{seg}}$ and on place 3 for the LLR. The most degraded signal is the original fire alarm noise degraded signal as it ranks the worst according all the objective quality measures.

6.3 Listening Test

In this section, an overview of the results of each of the questions that were answered in the listening test are presented. Here the "fire alarm signal" refers to the frequency sweeping alarm while "smoke detector signal" refers to the frequency stable signal.

6.3.1 Effort

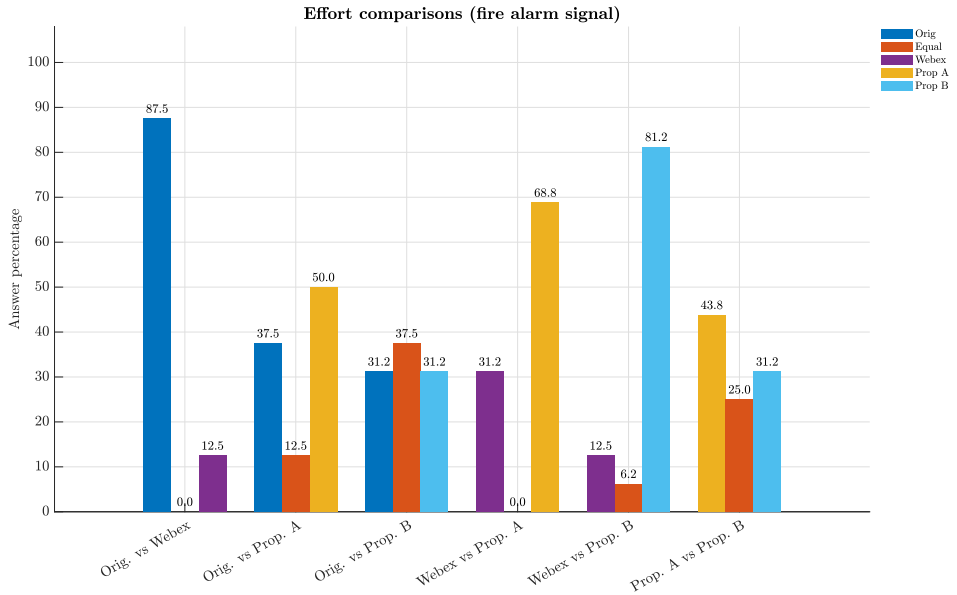


Figure 6.7: Answers to the question "Which signal requires MORE effort from you for you to understand the speech?" for the fire alarm signals.

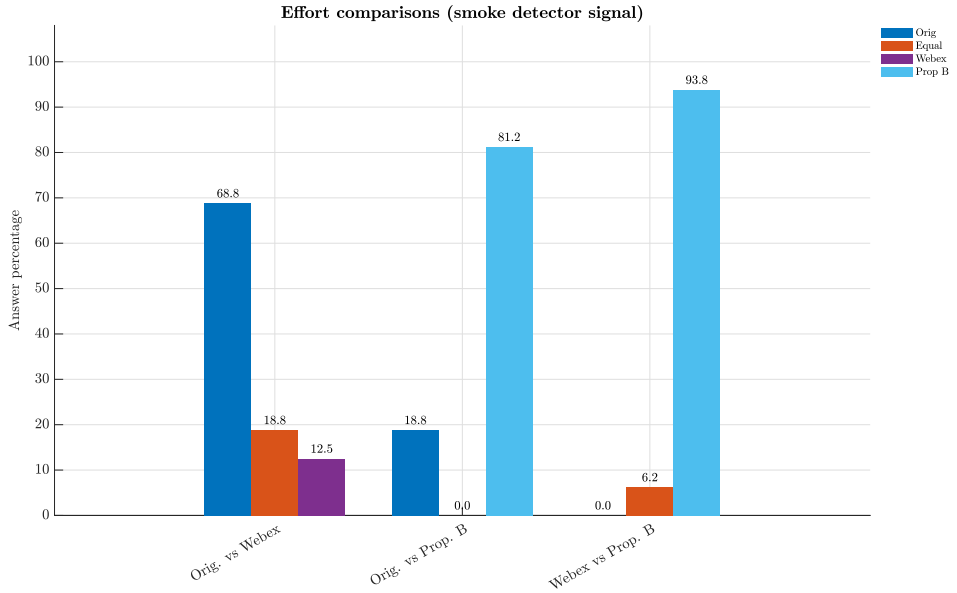


Figure 6.8: Answers to the question "Which signal requires MORE effort from you for you to understand the speech?" for the smoke detector signals.

Based on the question "Which signal requires MORE effort from you for you to understand the speech?" in the listening test. Which regards how the the audio signals were perceived relative to each other in regards to the perceived listening effort. For this category, for the fire alarm degraded audio signals, proposed filter *A* did the worst in the listening test as it was perceived to require more effort in order to understand the speech than all the other signals. At least this is the case when observing all the paired comparisons in isolation. Proposed filter *B* was the second worst as it performed worse than all the other audio signals except the original signal where it was perceived to require an equal amount of effort. However when comparing how proposed filter A and proposed filter B measured against the signal that was filtered through Webex, 68.8 % perceived proposed filter *A* to require more effort and 31.2 % less effort, whereas 81.2 % perceived proposed B to require more effort, 12.5 % less effort and 6.2 % an equal amount of effort. However, when paired against each other 43.8 % perceived proposed filter *A* to require more effort than proposed filter *B* and 25.0 % perceived an equal amount of effort, which appears contradictory. It is thus inconclusive which of these two filters that is worse in regards to how they affect the perceived listening effort. At the least they appear to have a similar effect. They also do not differ largely from the original fire alarm degraded signal. 50.0 % perceived proposed filter *A* to be worse than the original, 37.5 % considered the original signal to be worse than proposed filter *A* and 12.5 % considered them to be equal. 31.2 % considered proposed filter B to require more effort than the original while the same percentage considered the original to require more effort than proposed filter *B*. 37.5 % thought they required an equal amount of effort. The only really clear result in this category for the fire alarm degraded signals is that Webex was perceived as better than the three other signals.

For the smoke detector degraded audio signals it is evident that Webex shows the

most promising results. It is also clear that the audio signal that is filtered through the proposed filter is worse in terms of required listening effort than both the original audio signal and the audio signal that is filtered through Webex.

6.3.2 Frustration

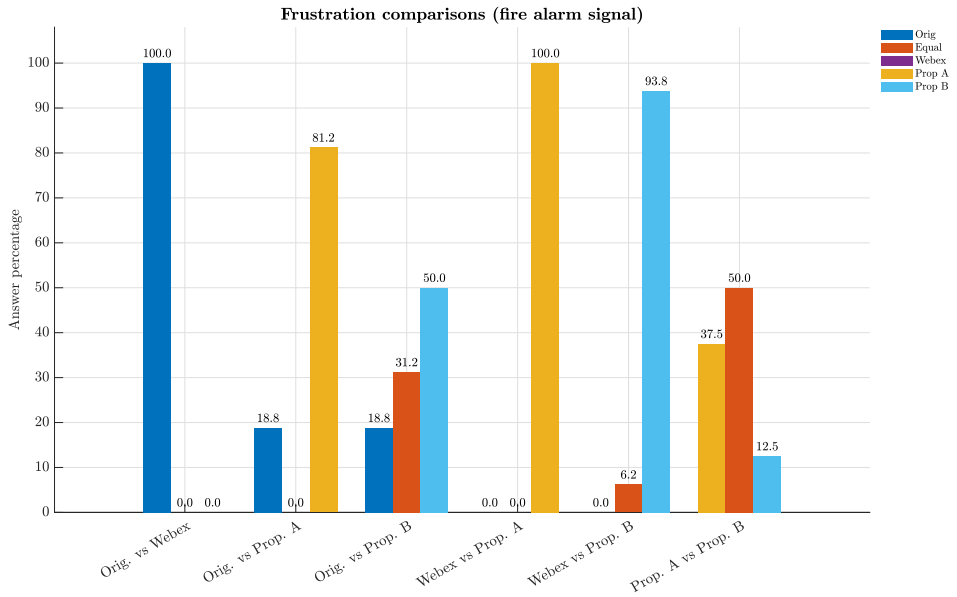


Figure 6.9: Answers to the question "Which signal makes you MORE frustrated trying to understand the speech?" for the fire alarm signals.

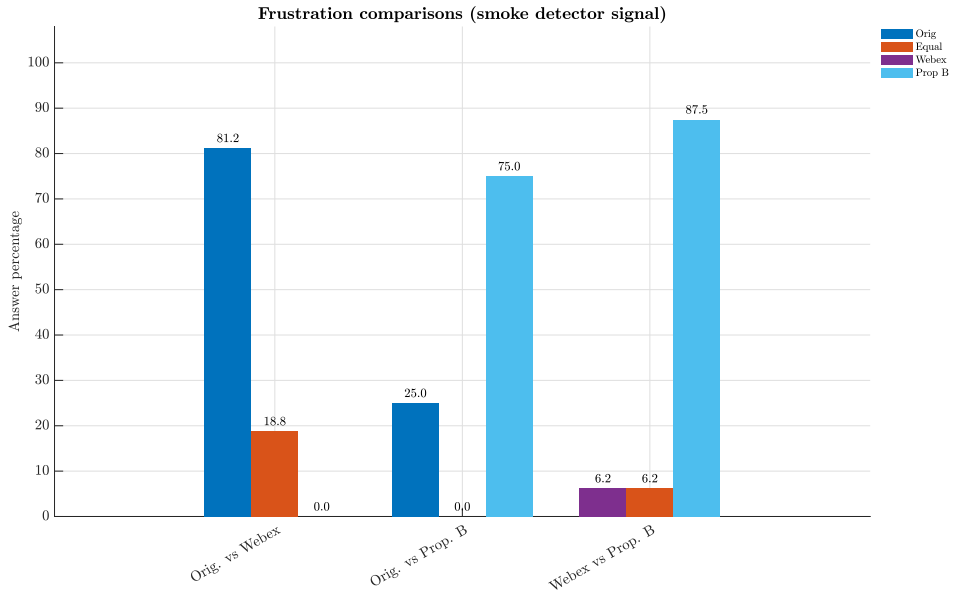


Figure 6.10: Answers to the question "Which signal makes you MORE frustrated trying to understand the speech?" for the smoke detector signals.

The evaluation of how the different audio signals compared relating to how much frustration they cause was based on the question "Which signal makes you MORE frustrated trying to understand the speech?" in the listening test. The results for the caused frustration level are more clear than for the listening effort. Here it is evident that Proposed filter A performs the worst. 81.2 % thought that it was worse than the original signal while the remaining percentage thought that the original was the worst. Only 50.0 % thought that proposed filter B was worse than the original, while 31.2 % thought it caused in an equal level of frustration. All participants in the listening test considered proposed filter *A* to be more frustrating than Webex and 93.8 % thought that proposed filter B was worse than Webex with 6.2 % rating them as equally frustrating. When comparing proposed filter *A* and proposed filter *B*, 37.5 % rated proposed filter A as more frustrating with 50.0 % rating them as equal. In contrast to the results from the effort category, it is here more clear that proposed filter *A* is the least preferred alternative, as the results consistently show that.

As with the listening effort, Webex was clearly the best alternative, both for the fire alarm degraded noise signal and the smoke detector noise degraded signal. 100 % thought that the original fire alarm signal made them more frustrated than the signal that was filtered through Webex. For the smoke detector degraded signal 87.5 % considered it to cause them more frustration than the signal filtered through Webex. 75.0 % became more frustrated by the signal that was filtered through the proposed filter than the original signal. The remaining 25.0 % considered the original signal to be more frustrating.

6.3.3 Unnaturalness

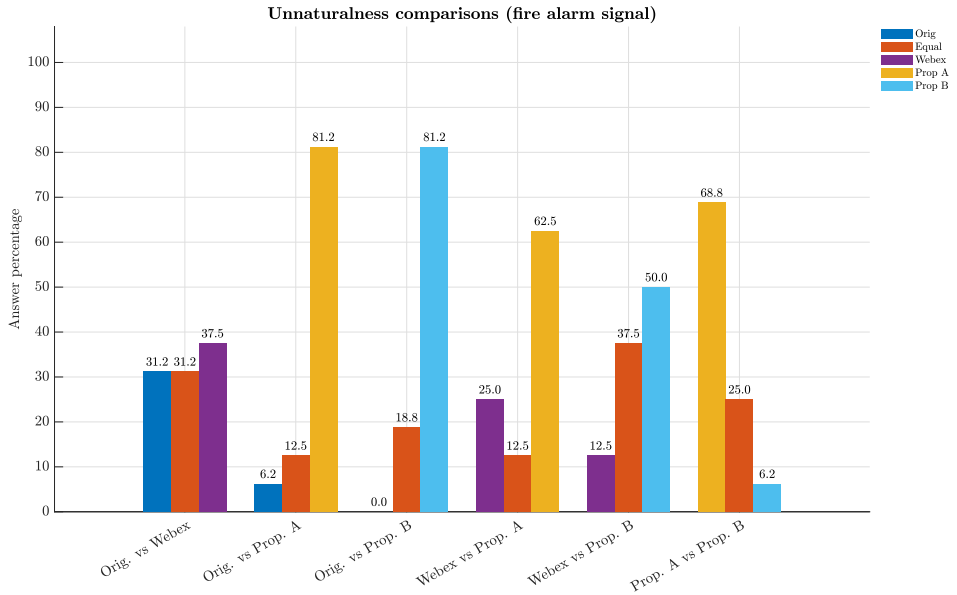


Figure 6.11: Answers to the question "Which signal do you perceive to have the MOST unnatural speech?" for the fire alarm signals.

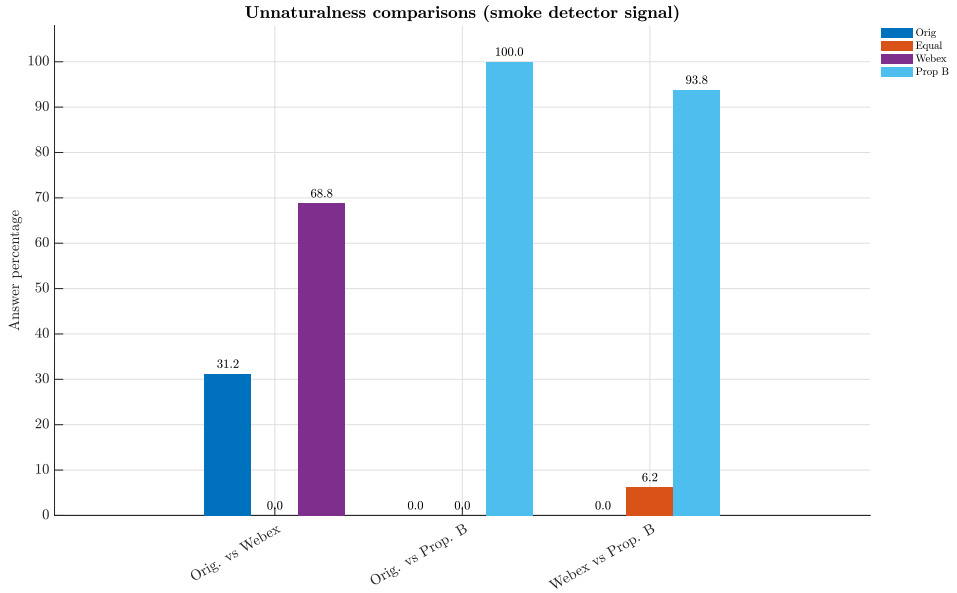


Figure 6.12: Answers to the question "Which signal do you perceive to have the MOST unnatural speech?" for the smoke detector signals.

The evaluation of the category of the perceived unnaturalness of the speech is based on the responses to the question "Which signal do you perceive to have the MOST unnatural speech?" in the listening test. As with the frustration category it is clear that proposed filter A performs the worst in this category as well. 81.2 % considered it to have more unnatural speech than the original signal with 12.6 % rating them as having equally unnatural speech. 62.5 % considered proposed filter A to have more unnatural speech than Webex while 12.5 % rated them equally. 68.8 % considered proposed filter A to have m ore unnatural speech than proposed filter B with 25.0 % rating them as having equally unnatural speech. While it is clear that Webex was perceived as considerably better than both proposed filter A and proposed filter B, it did not fully manage to maintain the natural speech of the original signal as 37.5 % considered it to have more unnatural speech than the original signal and 31.2 % considered the original signal to have more unnatural speech. The remaining 31.2 % considered them to have equally unnatural speech. However, this weighting towards Webex having more unnatural speech than the original is very slight and it cannot be concluded based on this alone that Webex in fact does have noticeably more unnatural speech. The results rather suggest that they have similar speech quality.

For the smoke detector signals it is however more clear that Webex is considered to produce more unnatural speech as 68.8 % considered it to be worse than the original speech signal while the remaining 31.2 % considered the original signal to have more unnatural speech. When comparing the Proposed filter for the smoke detector signal it is evident that Webex is the better filter as 93.8 % considered the Proposed filter to be worse with the remaining 6.2 % rating them as equal.

6.3.4 Loudness

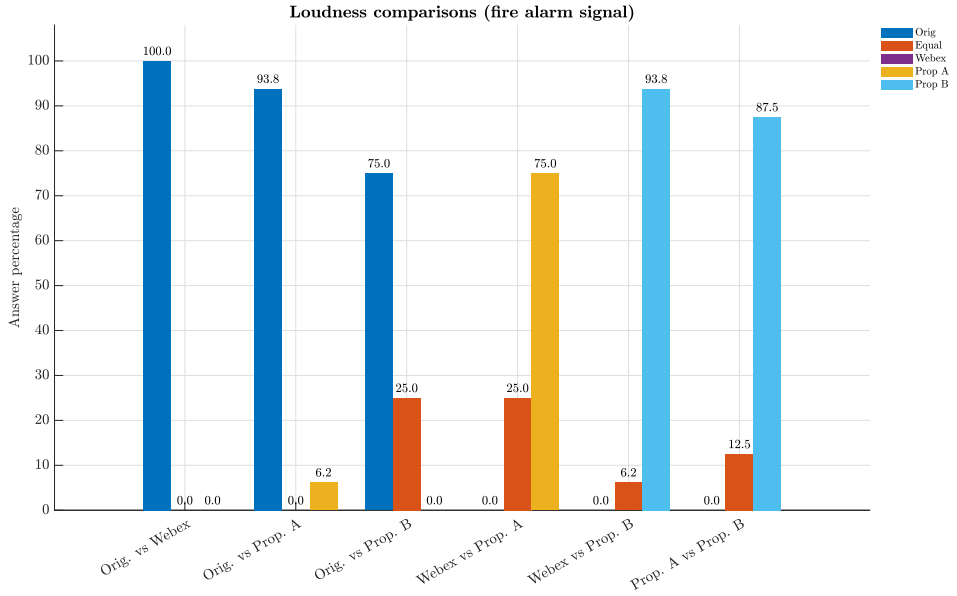


Figure 6.13: Answers to the question "Which signal do you perceive as the loudest?" for the fire alarm signals.

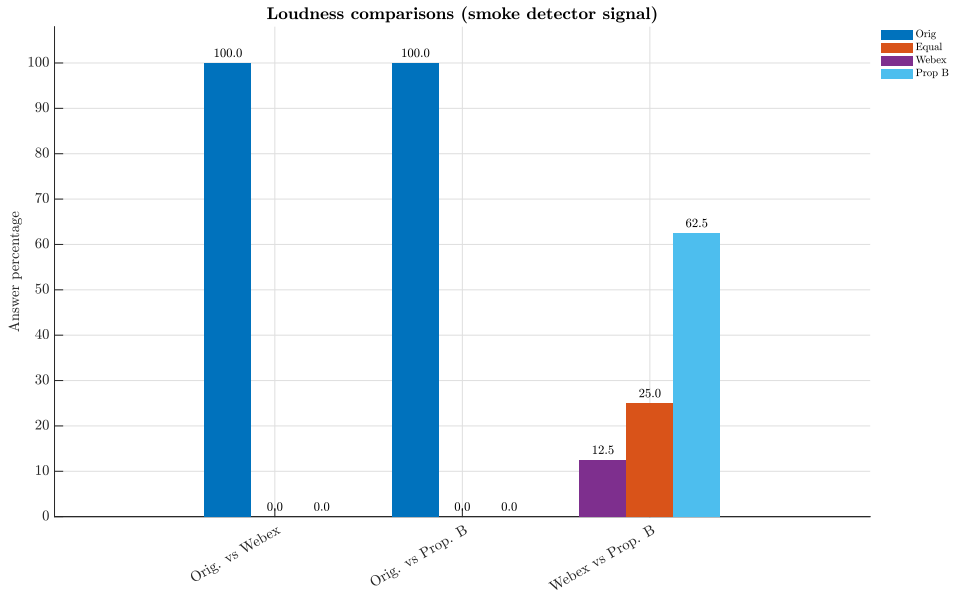


Figure 6.14: Answers to the question "Which signal do you perceive as the loudest?" for the smoke detector signals.

The perceived loudness was evaluated based on the question "Which signal do you perceive as the loudest?" in the listening test. For the fire alarm signal it is clear that the original signal is perceived as the loudest version. All the filters manage

to produce a perceivable suppression of the noise that the fire alarm produces. It is clear however, that Webex does it the best as 100.0 % perceived the original signal to be louder than the one filtered through Webex. 75.0 % considered Webex to obtain a lower loudness level than proposed filter *A* (25.0 % considered them to have an equal loudness level) and 93.8 % considered Webex to be quieter than proposed filter *B*. The second best filter for the fire alarm signal was proposed filter *A* as 93.8 % perceived it as quieter than the original signal (6.2 % thought proposed filter *A* was louder) while 75 % considered proposed filter *B* to produce a lower loudness level than the the original signal (25.0 % thought they were equally loud). Additionally, when comparing proposed filter *A* and proposed filter *B*, 87.5 % considered proposed filter *B* to have a higher loudness level, while 12.5 % considered them to be equally loud.

For the smoke detector signal, Webex also performs the best as all the listening test participants considered it to be quieter than the original signal. Even though this was also the case for the Proposed filter as all the participants also considered it to be quieter than the original signal, 62.5 % considered the signal filtered through Webex to be quieter than the signal that was filtered through the proposed filter (25.0 % thought they maintained an equal loudness level).

6.3.5 Annoyance

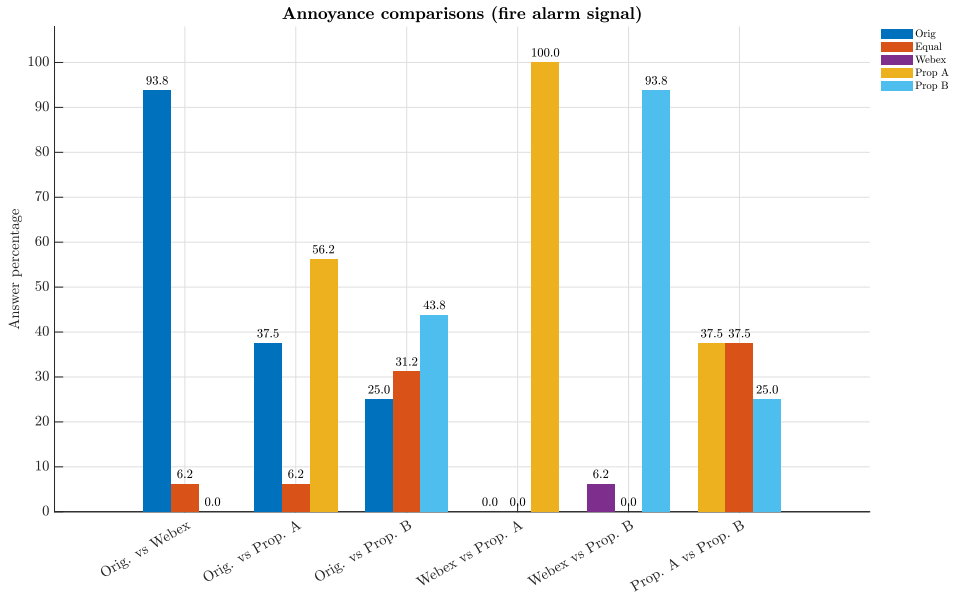


Figure 6.15: Answers to the question "Which signal do you perceive as the MOST annoying?" for the fire alarm signals.

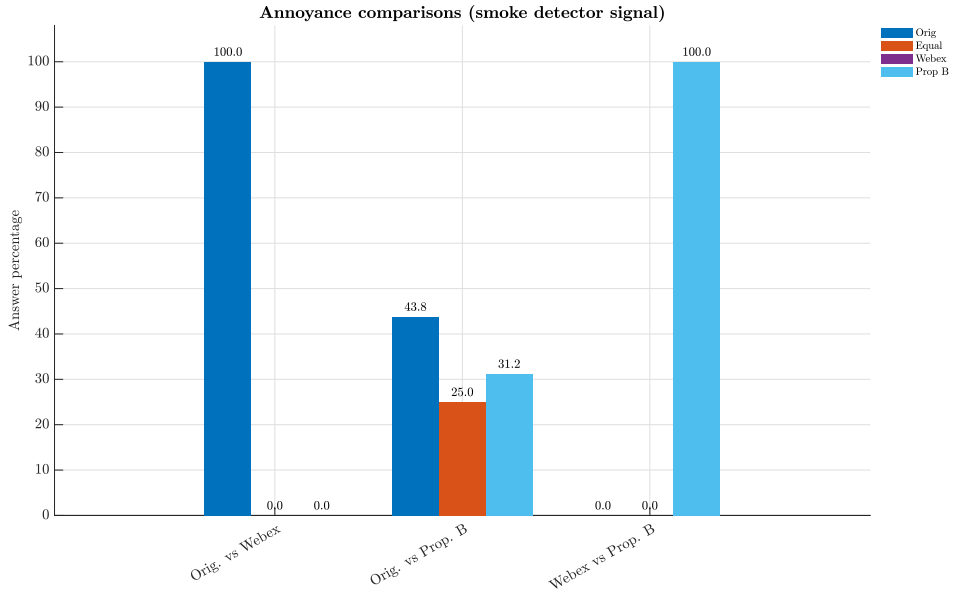


Figure 6.16: Answers to the question "Which signal do you perceive as the MOST annoying?" for the smoke detector signals.

The perceived annoyance was evaluated based on the question "Which signal do you perceive as the MOST annoying?" in the listening test. As with the other categories, Webex here shows the most promising results. For the fire alarm signal, when it was filtered through Webex it was perceived as less annoying by almost everyone than the original signal and when it was filtered through the other filters. The original signal was perceived as the second least annoying as both proposed filter A and proposed filter B were perceived as slightly more annoying. When comparing proposed filter A and proposed filter B, the results show that there is a slight weighting towards proposed filter A being perceived as more annoying (37.5 % considered proposed filter A to be more annoying while 25.0 % considered proposed filter B to be more annoying and 37.5 % considered them to be equally annoying). However, since this bias towards proposed filter A being more annoying is not of substantial size, it cannot be concluded for sure that this is the case.

For the smoke detector signal it is also clear that Webex has the best filtering method. All of the test participants considered the signal that was filtered through Webex to be less annoying than the unfiltered one. When comparing the proposed filter with the original signal, the proposed filter is slightly preferred as 43.8 % considered the original signal to be more annoying while 31.2 % considered the signal that was filtered through the proposed filter to be more annoying and the remaining 25.0 % thought they were equally annoying. This is not a considerable difference and a firm conclusion cannot be drawn regarding this comparison. When comparing the proposed filter with the Webex filtering method it is clear that the Webex filter is better as all participants found the signal to be more annoying when it was filtered through the proposed filter.

6.3.6 Overall

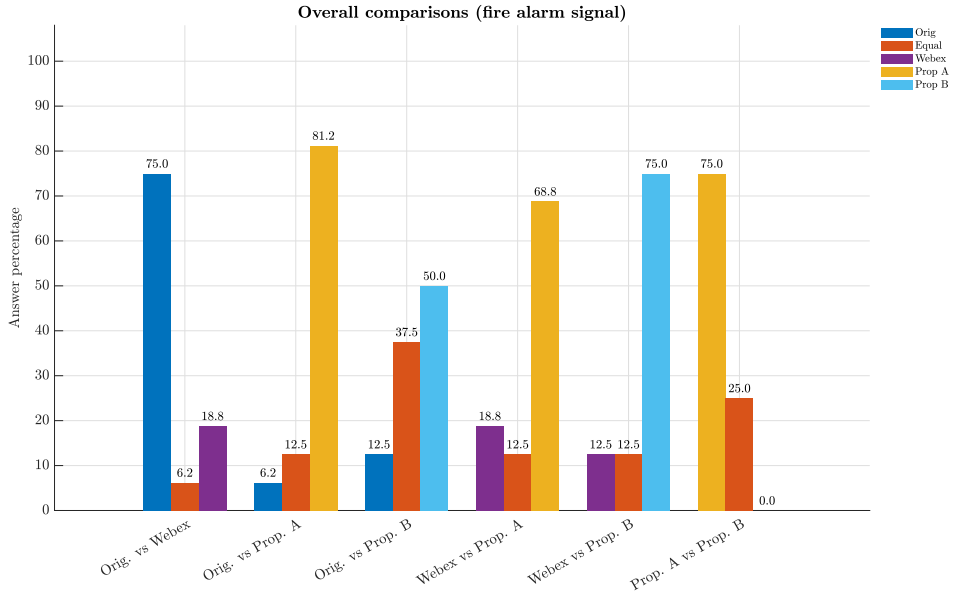


Figure 6.17: Answers to the question "Overall, which signal do you consider to be the worst?" for the fire alarm signals.

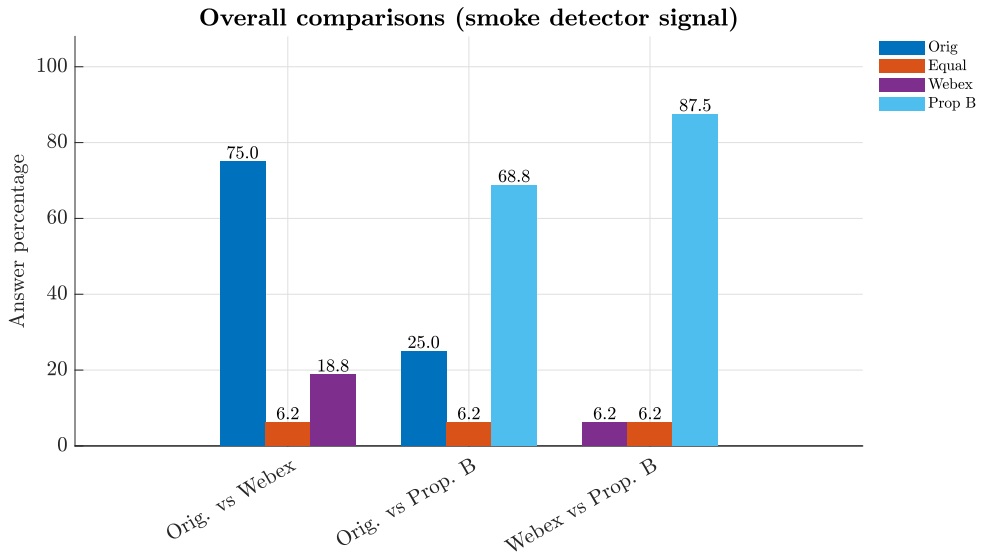


Figure 6.18: Answers to the question "Overall, which signal do you consider to be the worst?" for the smoke detector signals.

The overall quality of the signal is based on the question "Overall, which signal do you consider to be the worst?" in the listening test. For the fire alarm signal it is evident that Webex has the overall best filter. 75.0 % found the original signal to be worse overall than the filter that was filtered through Webex. 68.8 % found the signal that was filtered through proposed filter A to be worse than the signal that

was filtered through Webex. 12.5 % found them to be of equal overall quality. 75.0 % found the signal that was filtered through proposed filter B to be worse than the when it was filtered through Webex. 12.5 % found them to be of equal overall quality. Both proposed filter A and proposed filter B were considered worse than the original signal - 81.2 % found proposed filter A to be worse (12.5 % thought they were of equal quality) and 50.0 % found proposed filter B to be worse (37.5 % thought they were of equal quality). When comparing proposed filter A and proposed filter B, 75.0 % found proposed filter A to be worse overall, with 25.0 % ranking them as having equal quality.

For the smoke detector noise degraded signal it is evident that Webex performs the best as 75.0 % found the original signal to be worse overall and 6.2 % considering them to be equal. 87.5 % found the proposed filter to be worse than Webex with 6.2 % rating them as having equal overall quality. The proposed filter was deemed the worst overall filter as 68.8 % also found it to be worse than the original signal with

6.3.7 Summary

The results are evident in that it shows that the built-in filter in Webex obtains the signal that is most preferred by the majority of the listening test participants. This is the case for both the fire alarm degraded signal and the smoke detector degraded signal. For the fire alarm signal, the second most preferred signal is the original unfiltered signal. Even though both proposed filters manage to reduce the overall loudness level of the signal, they fail to maintain the quality of the speech and this seems to suggest that most listening test participants found them to sound worse than not only the built-in filter in Webex but also the original unfiltered signal. This suggests that the low speech quality that results from the proposed filters is so considerable that it is perceived as worse than the presence of the fire alarm noise or smoke detector noise. The built-in filter in Webex however manages to reduce the loudness significantly better than the proposed filters, both for the fire alarm degraded signal and the smoke detector degraded signal at the same time as it manages to maintain a nearly as good speech quality as the original signal. Conversely, when comparing proposed filter A and proposed filter B, the results from the listening test shows that proposed filter A is perceived as worse than proposed B in all categories except loudness. The results suggest that this is also due to the fact that it causes the speech quality to worsen with an increased noise suppression as proposed filter A has a better capacity than proposed filter B in that aspect. The results are clear as they show that the built-in filtered method in Webex is far superior to any of the proposed filters and that it does manage to improve the degraded signals in all of the aspects that were tested in the listening test.

7
Discussion

The proposed suppression strategies for sweeping tonal alarms aim to remove periodic interference without degrading speech quality. Two variants were tested: Approach *A*, based on frame-wise cepstral liftering, and Approach *B*, a hybrid method combining liftering with LPC analysis. This section functions as a discussion on the work, and is structured according to the list below.

- Analytical results from MATLAB (7.1)
- Objective quality evaluation and listening test results (7.2)
- Methodological reflections (7.3)
- Further work (7.4)

The discussion chapter is structured chronologically according to the listed topics, starting with the analytical results in Section 7.1.

7.1 Analytical Results of Proposed Filters

The analytical results, presented in Chapter 6, are discussed in this section, based on Figure 6.1 to 6.6.

7.1.1 Sweeping Alarm

Filter *A* uses harmonic attenuation guided by pitch estimates based on cepstral peaks only. Harmonics are suppressed in the cepstrum, with increasing width applied to higher-order harmonics. Figures 6.3 and 6.2 show that gentle suppression of alarm harmonics is applied, while maintaining transient speech below 1 kHz. However, its performance is closely bound to the accuracy of the pitch tracker. When the estimated pitch deviates or fluctuates, the algorithm either fails to suppress the alarm or introduces artifacts due to over-suppression in undesired regions of the cepstrum. While the smoothing factor on the raw pitch detection helps the precision of the suppressor, the accuracy is still too low for effective suppression of the tonal noise and it leaks over undesired regions, including speech formants of higher order.

Proposed filter *B* adds a decision layer that classifies frames as tonal or non-tonal based on cepstral peak and spectral flatness measure (SFM). Tonal frames undergo LPC analysis, and suppression by cepstral liftering is applied on the estimated LP-coefficients. By comparing the spectograms and waveforms of Propsal *A* and Proposal *B* in Figure 6.2 and 6.3, the first achieves clearer suppression of the tonal alarm

harmonics. One reason could be that the LPC model misinterprets speech-alarm mixtures. Furthermore, if the coefficients in the cleaned LPC-vectors still contain tonal alarm noise after suppression, the noise will be re-introduced during the resynthesis. In addition to the new possible pitfalls, the suppression stage of proposed filter *B* still depends on the accuracy of the pitch tracker, which has proven unstable according to Figure 6.1b.

7.1.2 T3-Temporal Stationary Alarm

Compared to the first case, where the amount and width of suppression were relatively similar despite the different processing chains, the second case shows distinct differences between the filter outputs. Like in case 1, Filter *A* applies gentle suppression on the alarm fundamental (Figure 6.5c). However, the alarm component is still clearly visible in the spectogram after suppression, as well as its vague harmonic. The maximum levels have been attenuated, dampening the increasing amplitude over time. The difference is smaller in the beginning of the file, and this can be explained by the fixed threshold for the cepstral peak detection of the system, causing the pitch detector to pass frames without triggering the suppression. This pitfall adds to the error caused by the pitch tracker.

A clear distinction regarding filter performance can be done for Filter *B*. The reason to the heavily suppressed signal visible in Figure 6.5d could be the initial LPC-estimation. Two possible explanation stands out: first, the LPC-estimation isolates speech- and alarm components more accurately from low-level background noise, compared to case 1. This would lead to an estimated signal consisting of speech and alarm only. Even though filter *B* applies cepstral suppression to the cleaned, resynthesised signal on each frame, an unfavourable ratio between speech and tonal noise is obtained.

7.1.3 Speech Quality and Suppression Trade-offs

A consistent challenge across all methods is balancing suppression depth against speech integrity. While the goal is to remove alarm noise without impairing intelligibility, tonal alarm noise often overlaps with speech formants or pitch-related components. Filter *A* is designed to be selective, targeting harmonics tied to the tracked pitch. This means that voiced phonemes that overlap with the estimated alarm frequency can be unintentionally attenuated if they fit the alarm model. Thus, a combination of undesired suppression of speech formants, and the lack of suppression of alarm noise on affected frames. A further source of error is the mathematical versus the real-world relation between alarm fundamental and harmonics in the cepstral domain. A clear cepstral peak is typically found on the quefrency corresponding to the spectral peak of the fundamental in the frequency domain. However, harmonics which are present in the spectral domain, may not show strong periodicity in the cepstral domain, making them difficult to detect and suppress, which increases the risk of unintentional suppression of other components.

To preserve subjectively perceived speech quality, the decision to limit the suppression amount was taken. Together, the results demonstrate that deterministic, model-based methods can provide interpretable suppression. Yet their ability to adapt to overlapping conditions, where voiced speech or harmonics interfere with the fundamental of the alarm, remains limited. This positions data-driven approaches as a necessary complement to model-based approaches.

7.1.4 Limitations and Deployment Considerations

Several limitations emerged during evaluation that constrain generalisability and deployment readiness. First, the suppression logic was designed to operate on isolated frames. With no long-term memory or temporal modelling, rapid modulations or short tonal bursts are handled inconsistently. Pitch tracking errors propagate into the suppression logic and lead to either insufficient attenuation or speech distortion.

The binary tonal/non-tonal frame classifier used in Approach B is too coarse for complex mixtures. In real speech, tonal and non-tonal features can co-occur, requiring a more nuanced or probabilistic decision framework.

From a deployment standpoint, neither method currently accounts for directional cues, reverberation, background noise or spectral colouring captured by real telephone microphones. These factors significantly alter alarm characteristics and may affect suppression precision in a real-world scenario.

7.1.5 Model-Based vs. Data-Driven Approaches

The filters developed in this thesis follow a model-driven approach, relying on the source–filter model of speech and frequency-domain signal representations. Their structure allows direct control over filter shape, suppression width, and detection thresholds. This ensures transparency, low complexity, and ease of tuning. However, model-based filters require explicit design logic for each possible signal composition, which limits their use if models are too simplified. Furthermore, they do not learn from previous data, and their detection performance degrades in the presence of competing harmonics, fluctuating SNR, or mixed signal types. In contrast, data-driven approaches can learn robust representations from noisy examples and generalise across varied alarm types and environments. In fact, these solutions consider one thing that a model never does: the real world. A functional model-based approach would need to be fine-tuned for a vast amount of conditions: different alarm types, signal levels, SNR-conditions, different voice characteristics, and so on. Eventually, this kind of modelled system would become data-driven, and that is one reason to why AI-based methods should be investigated for further work or development within this scope.

7.2 Objective Quality Evaluation and Listening Test Results

A comparison between the results from the objective quality evaluation and the listening test clearly shows that there is a discrepancy between the two evaluation methodologies. The objective quality measures and the experienced quality seem to have a low correlation. The subjective data from the listening test suggests that the objective quality measures actually fail to take the speech quality into consideration when computing the overall quality of the signal and thus fail at what was the initial purpose of introducing them. This can be observed when comparing the listening test results for categories such as unnaturalness (Section 6.3.3) with the results from the objective quality measures. In the listening test, for the category of unnaturalness where the test participants were asked to answer the question "Which signal do you perceive to have the MOST unnatural speech?" the filtered signals did the worst, except for the fire alarm noise degraded signal that was filtered through Webex as it performed only insignificantly worse compared to the unfiltered signal. For the smoke detector noise degraded signal that was filtered through Webex, the negative effect on the speech quality was more perceivable, however not as much as for the proposed filter. The listening test results show that, for the proposed filters, there seems to be a strong correlation between the unnaturalness of the speech and the effort, frustration and annoyance - despite the reduced loudness of the fire alarm or smoke detector noise. This proves that despite the relatively reduced loudness of the fire alarm or smoke detector noise bu these filters, the reduced speech quality and speech intelligibility is so considerable that the original louder signals are still perceived as better. This is however not the case for the signals that were filtered through Webex as the perceived naturalness of the speech is much more preserved.

Compared to the listening test results, the results from the objective quality measures seem to relate substantially more to the overall loudness of the degraded signal, which here mostly means the loudness of the fire alarm or smoke detector. When comparing the quality of the fire alarm noise degraded signal when it that was filtered through proposed filter A and when it was proposed through proposed filter B, proposed filter A had the best quality according to all three objective measures. However this in contradictory to the findings of the listening test as it shows conflicting results. It therefore seems that the quality that the objective quality measures compute mostly relates to the the relative loudness. It could therefore be discussed whether these kinds of measures are appropriate when the degradation in the form of background noise is so considerable as in this case - at least this is the case for the fire alarm noise degraded signal. When reviewing the results for the objective quality evaluation of the smoke detector noise degraded signal, the original degraded signal is gets a better score than the one filtered through the proposed filter, at least for the $\text{fwSNR}_{\text{seg}}$ and LLR which does correlate with the results of the listening test. However, to conclude, the accuracy of these objective evaluation methods for these kinds of signal degradations is up for discussion as the results are varying and sometimes contrary to the subjective listening test results, which in this case are deemed much more reliable and statistically representable of how humans actually

perceive such degradations.

7.3 Methodological Reflections

This section of the discussion reflects upon methodological decisions, including the use of test files during filter design and evaluation, and the particular choice of using cepstral alanysis and LPC-analysis for tonal noise suppression.

7.3.1 Using Test Files instead of Real-World Recordings

A critical constraint in this study was the inability to perform reliable in-ear acoustic measurements using the HMS II.3. As a result, suppression development was based on artificially mixed test files. This removed the proposed systems from real acoustic conditions. Critical factors such as reverberation, telephony DSP artefacts, background noise, and microphone colouration were absent. A solution to mimic real-world conditions could have been pre-processing, such as equalisation with a low-pass filter on the entire signal, before running it through the filter, but this methodology would still be speculutive.

Consequently, suppression thresholds and pitch tracking models may have been tuned under idealised or worse conditions, compared to reality. In real deployments, these parameters may behave differently, resulting in lower suppression performance or unintended artefacts. However, the artificial setup enabled reproducible evaluation for a listening test, allowing evaluation of filter performance and comparing them to unprocessed conditions and one state of the art-technique.

7.3.2 The Choice of Theoretical Modelling using the Cepstrum and LPC-Analysis

The idea to use the cepstrum as the processing domain is motivated by its theoretical feasibility to separate sounds that operate in the same frequency range. In an ideal case, the designed pitch tracker would be able to isolate tonal alarm components from voiced speech and speech harmonics in the same region. This approach was combined with LPC-analysis due to the successful work carried out by Gül, *et al.* [12], showing that pure linear predictive coding was a feasible method for source-filter separation. By using cepstral liftering as the targeted suppression method, the idea was to refine the LPC-based cleaning of a corrupted speech signal. However, as previously discussed in 7.1 several reasons may add to the need of further work on this approach.

7.4 Further Work

Future work should prioritise evaluation under real-world acoustic conditions. This includes successful headset-based recordings in operational environments, contamination with real alarm sounds and room effects at the source, and transmission

through VoIP or mobile communication systems. Such a scenario was prepared, as mentioned in 5.2.2, but the attempt failed.

To generalise across alarm systems, suppression models must handle greater variety: modulated multi-tone alarms, overlapping sources, or short-duration tonal bursts. While model-based filters can be extended with more suppression branches or classifiers, the possible use of large datasets would be time-consuming and would lead to inefficient work. A promising direction is to introduce AI-based techniques, such as deep neural networks (DNNs) into the detection pipeline, using them to identify alarm types, isolate speech, and predict suppression masks. Such systems could be trained using test data combined with real-world recordings. In the context of this work, the investigated techniques of cepstral liftering and LPC-analysis prove to produce interpretable results aligned with theoretical assumptions and could thus act as a foundation for further development, incorporating DNNs.

8

Insights & Conclusion

This thesis investigated the application of model-based methods for attenuating tonal alarm interference from speech recordings in emergency call center scenarios. Specifically, cepstral liftering, LPC-analysis, and their combination were used. The focus on these methods was motivated by their interpretability, computational efficiency, and theoretical grounding in the source–filter model of speech. In theory, cepstral liftering offers harmonic selectivity with limited interference to broadband speech, while LPC-based residual suppression targets periodic excitation components by separating them from the spectral envelope. A hybrid system was developed to explore their complementarity.

Theoretical expectations were partially met. The filters achieved targeted attenuation of tonal components under controlled conditions using a test file consisting of male speech and different types of tonal alarms. The filters showed robustness in preserving speech transients and overall dynamics of the original sound files. However, their performance were limited due to unreliable pitch estimation. Conditions in which alarms and speech overlapped spectrally proved difficult to process without distorting the speech formants, leading to the trade-off between gentle suppression and perceived speech quality. Residual-domain suppression introduced additional artifacts when LPC modeling was inaccurate, particularly in speech-dominated low-frequency regions.

In comparison to a commercial data-driven solution provided by Webex, the model-based systems fell short. Webex demonstrated stronger attenuation in silent frames, better adaptation to varying alarm levels, and fewer perceptual artefacts during speech. These outcomes suggest that data-driven approaches, likely leveraging deep neural networks, are more effective in handling the complexity and variability of real-world audio environments. The results from the listening test demonstrated that Webex was perceived as superior to the proposed filters on all the subjective parameters that were tested. Furthermore, the perceived speech quality for the fire alarm noise degraded signal was nearly unaffected by being filtered through Webex. For the smoke detector noise degraded signal the speech degradation was more perceivable when filtered through Webex, but still better than for the proposed filter.

Based on these findings, the following conclusions are drawn:

1. While model-based suppression methods can deliver interpretable, perceived tonal noise reduction, their practical utility is limited by instability in pitch

detection, coarse classification logic, and lack of adaptability.
2. The combination of cepstral and LPC-based techniques prove to help the flexibility but also increase the risk of introducing artefacts, particularly when applied to mixed or ambiguous signal content.
3. Though the proposed cepstral peak tracking module was able to catch the pattern characterised by frequency-sweeping alarms, it was not robust enough to efficiently sort out false positives.
4. The data-driven solution provided by Webex outperforms the proposed methods in both suppression strength and speech quality, probably due to its ability to adapt suppression strategies based on learned patterns and contextual cues.
5. For organisations aiming to reduce tonal alarm interference in speech communication, especially in high-stakes environments, purely model-based methods are unlikely to offer sufficient robustness and quality.
6. A serious consideration should be given to integrating a data-driven solution—either via adoption of a commercial engine or development of a custom, perceptually tuned neural suppressor—potentially in combination with interpretable signal-model components to retain controllability.

In summary, this thesis contributes a functional and explainable tonal suppression framework, but also demonstrates the limitations of model-based filtering in realistic and varied conditions. The results point clearly toward hybrid or data-driven approaches as the most promising path forward for high-performance speech enhancement in operational settings.

Bibliography

[1] World Health Organization, *Burden of disease from environmental noise - Quantification of healthy life years lost in Europe*. 2011.

[2] T. Münzel, F. P. Schmidt, S. Steven, J. Herzo, A. Daiber, and M. Sørensen, "Environmental noise and the cardiovascular system," *Journal of the American College of Cardiology*, 2018. DOI: 10.1016/j.jacc.2017.12.015.

[3] Standard Norge, *NS 8175:2019: Lydforhold i bygninger - Lydklasser for ulike bygningstyper (Acoustic conditions in buildings - Sound classification of various types of buildings)*. 2019.

[4] S. Bhamra. "How does noise pollution impact sustainability?" Accessed: 2025-05-31. (2024), [Online]. Available: https://environbuzz.com/how-does-noise-pollution-impact-sustainability/#:~:text=Finally%2C%20noise%20pollution%20can%20also,by%20significantly%20impacting%20environmental%20quality..

[5] World Health Organization, "Guidelines for community noise," World Health Organization, Geneva, Switzerland, Tech. Rep. a68672, 1999, Outcome of the WHO-expert task force meeting held in London, UK, April 1999. Based on the document "Community Noise" (1995)., p. 141. [Online]. Available: https://www.who.int/publications/i/item/a68672.

[6] E. T. Ellefsen, "Noise and personalization: The effect of a personalized workplace on the perception of noise," Master's thesis, Inland School of Business and Social Sciences, 2024.

[7] M. K. Tangmyr, "The impact of noise on a psychosocial relationship: Qualitative analysis," Master's thesis, Inland School of Business and Social Sciences, 2023.

[8] Ministry of Labour and Social Inclusion, *Regulations concerning action and limit values for physical and chemical agents in the working environment and classified biological agents*, In force from 2013-01-01. Last amended: FOR-2024-04-05-581., Dec. 6, 2011. [Online]. Available: https://www.arbeidstilsynet.no/en/laws-and-regulations/regulations/regulations-concerning-action-and-limit-values//.

[9] J. A. Patel and K. Broughton, "Assessment of the noise exposure of call centre operators," *The Annals of Occupational Hygiene*, 2002. DOI: 10.1093/annhyg/mef091.

[10] T. Venet, A. Bey, P. Campo, *et al.*, "Auditory fatigue among call dispatchers working with headsets," *International Journal of Occupational Medicine and Environmental Health*, 2018. DOI: 10.13075/ijomeh.1896.01131.

[11] B. Swagowska, "Noise at workplaces in the call center," *Archives of Acoustics*, 2010. DOI: 10.2478/v10168-010-0024-2.

[12] Y. Gül, A. M. Ariyaeeinia, and O. Dewhirst, "A new approach to reducing alarm noise in speech," *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003*, 2003. DOI: 10.21437/Eurospeech.2003-393.

[13] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979. DOI: 10.1109/TASSP.1979.1163209.

[14] J. Benesty, S. Makino, and J. Chen, "Introduction," in *Speech Enhancement*, J. C. Jacob Benesty and Y. (Huang, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 1–8, ISBN: 978-3-540-27489-6. DOI: 10.1007/3-540-27489-8_1. [Online]. Available: https://doi.org/10.1007/3-540-27489-8_1.

[15] F. H. Bess and B. W. Y. Hornsby, "Listening can be exhausting—fatigue in children and adults with hearing loss," *Ear and Hearing*, vol. 35, no. 6, pp. 592–599, 2014. DOI: 10.1097/AUD.0000000000000099. [Online]. Available: https://doi.org/10.1097/AUD.0000000000000099.

[16] M. R. Schroeder, "U.s. patent no. 3,180,936," Filed Dec. 1, 1960, Apr. 1965.

[17] M. R. Schroeder, "U.s. patent no. 3,403,224," Filed May 28, 1965, Sep. 1968.

[18] J. Benesty, J. Chen, Y. (Huang, and S. Doclo, "Study of the wiener filter for noise reduction," in *Speech Enhancement*, J. C. Jacob Benesty and Y. (Huang, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 9–41, ISBN: 978-3-540-27489-6. DOI: 10.1007/3-540-27489-8_2. [Online]. Available: https://doi.org/10.1007/3-540-27489-8_2.

[19] R. Martin, "Statistical methods for enhancement of noisy speech," in *Speech Enhancement*, J. C. Jacob Benesty and Y. (Huang, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 43–65, ISBN: 978-3-540-27489-6. DOI: 10.1007/3-540-27489-8_1. [Online]. Available: https://doi.org/10.1007/3-540-27489-8_1.

[20] S. Obillaneni, P. S. P. Wei, S. Kar, A. Khan, and J. Karhade, "Ica based noise reduction in mobile phone speech communications," in *Proceedings of the IEEE International Conference on Computer Communication and Artificial Intelligence (CCAI)*, 2024, pp. 326–331. DOI: 10.1109/CCAI59530.2024.10545965. [Online]. Available: https://doi.org/10.1109/CCAI59530.2024.10545965.

[21] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995. DOI: 10.1109/18.382009.

[22] I. Daubechies, "The wavelet transform," in *Ten Lectures on Wavelets*, Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM), 1992, pp. 1–16. DOI: 10.1137/1.9781611970104.ch1.

[23] W. Zhao, Z.-H. Tan, and D. Wang, "Wavelet-based dnn for speech enhancement under non-stationary noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1107–1119, 2020.

[24] C. Li, Y. Zhang, and S. Liu, "A wavelet-domain u-net for speech enhancement," *IEEE Signal Processing Letters*, vol. 29, pp. 359–363, 2022.

[25] P. Bao, M. Yang, and Q. Liu, "Attention-guided wavelet-based speech enhancement using deep learning," *Digital Signal Processing*, vol. 111, p. 102 978, 2021.

[26] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. Springer, 1976, ISBN: 978-3540075633.

[27] D. Ellis, *Linear prediction (lpc) - lecture notes*, `http://www.ee.columbia.edu/~dpwe/e4896/`, E4896 Music Signal Processing, Columbia University, 2013.

[28] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proceedings of the IEEE*, vol. 65, no. 10, pp. 1428–1443, 1977. DOI: `10.1109/PROC.1977.10760`.

[29] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 95–106, 2004. DOI: `10.1109/MSP.2004.1338295`.

[30] S. He, W. Rao, J. Liu, *et al.*, "Speech enhancement with intelligent neural homomorphic synthesis," *arXiv preprint arXiv:2210.15853*, 2022. [Online]. Available: `https://arxiv.org/abs/2210.15853`.

[31] W. M. Liu, K. A. Jellyman, N. W. D. Evans, and J. S. D. Mason, "Assessment of objective quality measures for speech intelligibility," *Interspeech*, 2008. DOI: `10.21437/Interspeech.2008-220`.

[32] D. H. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing, Paris, France*, 1982. DOI: `10.1109/ICASSP.1982.1171512`.

[33] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 16, 2008.

[34] K. Kondo, "Estimation of speech intelligibility using objective measuresestimation of speech intelligibility using objective measures," *Applied Acoustics*, vol. 74, 2013.

[35] ETSI, "Speech and multimedia transmission quality (stq); methods for objective assessment of listening effort," ETSI, Tech. Rep. ETSI TS 103 558 V1.3.1, 2021.

[36] J. E. Peelle, "Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior," *Ear and hearing*, 2018. DOI: `10.1097/AUD.0000000000000494`.

[37] HEAD acoustics. "Listening effort prediction made easy: Head acoustics integrates leap into acqua 6.1.100." Accessed: 24-Feb-2025. (2024), [Online]. Available: `https://www.head-acoustics.com/news-events/press-releases/press-details/show/head-acoustics-integrates-leap-into-acqua`.

[38] HEAD acoustics, "Acopt 38 - option leap – listening effort prediction from acoustic parameters," HEAD acoustics, Tech. Rep., 2024, Accessed: 24-Feb-2025. [Online]. Available: `https://www.head-acoustics.com/news-events/`

press-releases/press-details/show/head-acoustics-integrates-leap-into-acqua.
[39] M. Kleiner, *Acoustics and Audio Technology*. J Ross Publishing, 2011.
[40] Gracey Associates. "Acoustic glossary." Accessed: 18-Feb-2025. (2025), [Online]. Available: https://www.acoustic-glossary.co.uk/frequency-weighting.htm?.
[41] Lindosland at English Wikipedia, *A graph of the A-, B-, C- and D-weightings across the frequency range 10 Hz – 20 kHz*, [Online; accessed 10-Feb-2025], 2011. [Online]. Available: https://en.wikipedia.org/wiki/A-weighting.
[42] Swedish Institute for Standards, "Acoustics – determination of occupational noise exposure – engineering method (iso 9612:2009)," Swedish Institute for Standards, Tech. Rep. SS-EN ISO 9612:2009, 2009.
[43] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.
[44] B. Gick, I. Wilson, and D. Derrick, "Source–filter theory of speech," in *Oxford Research Encyclopedia of Linguistics*, Oxford University Press, 2013. [Online]. Available: https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-894.
[45] B. Chen, *Cepstral processing and source–filter separation*, 2022. [Online]. Available: https://speech.ee.ntu.edu.tw/~hwchiu/NTNU_Speech_Signal_Representations.pdf.
[46] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, Taylor & Francis Group, 2013, ISBN: 9781138075573.
[47] D. Schmitt, *Nerds.de audio & midi particles*, https://nerds.de/en/loopbeaudio.html, Accessed: 2025-02-21, 2025.
[48] G. Fairbanks, *Voice and articulation drillbook*. Harper & Row, 1960, ISBN: 0060419903.

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY