Security relevant sounds perception in cochlear implant users

Alexandre Philippon

Signal processing, computer science, acoustic
Master Atiam

2nd april 2024 - 13th september 2024

University referent teacher:
Emmanuel
Saint-James
saint-james@lip4.fr

Industrial mentor:
Adrien Daniel
addaniel@cochlear.com

Industrial mentor:
Yue Zhang
yuezhang@cochlear.com

Abstract

This work focuses on the analysis of how cochlear implants process relevant security sounds and aims to explain the poor performance of CI users in terms of recognition and detection. This analysis involves visualizing the electrodograms and spectrograms of the processed sounds, as well as comparing the acoustic features of the input and output sounds. The output sounds are generated using relevant CI vocoders. Additionally, the work proposes a design to improve CI transduction and provides experimental protocol guidelines to test its efficiency.

Acknowledgment

I'd like to express my gratitude and affection to everyone who has supported and helped me during this internship. My first thanks go naturally to my dedicated supervisors, Dr. Adrien Daniel and Dr. Yue Zhang. It is hard for me to express how thrilling and fulfilling our meetings and discussions were. Thanks to your kindness, open-mindedness, and availability, I had the opportunity to work in an extremely pleasant environment and to learn a great deal.

I would also like to extend my gratitude to the entire RD team at Cochlear Vallauris, especially to Dr. Rafael Attili Chiea and Dr. Behnam Molaee-Ardekani, who assisted me with their work and provided valuable and insightful advice.

Naturally, I also want to thank my loving family and friends, who, as always, support me in every project I undertake, even the most ambitious. Life would likely be less meaningful without the love we share, and I am truly grateful to have you all in my life.

Nomenclature

Devices and electronic

Bone anchored hearing system Bone anchored hearing system is a surgically implantable device. It sends sound waves through the bones of the skull directly into the cochlea and is target for people having conductive hearing loss

CI CI is the abbreviation of Cochlear Implant, an electronic prosthesis used to compensate the hearing loss of the wearer. It emulates the natural processing chains of hearing from the wave-detection to the electrical stimulation of neurons and is target for people suffering profound sensory-neural hearing loss.

Psycho-acoustic

Alarm detection We call in this study alarm detection the capacity of a subject to detect the presence of an alarm sound even without identifying this sound as an alarm

Alarm recognition We call in this study alarm recognition the capacity of a subject to identify a sound stimuli as an alarm and to differentiate those (police siren, fire fighter, evacuation city signal...)

Alarm audibility We call in this study alarm audibility the capacity of a subject to answer correctly at the question "Was the alarm audible" in case a mixture containing alarm sound were presented.

The current spread issue This phenomena occurs because of the tendency of electric current to spread when put in a aquatic environment, because of this, when an electrodes fires a electrical current, adjacent neurons associated with other electrodes are being stimulated causing a multiple channel interaction in term of "sound sensation".

SIR SIR stands for Speech Intelligibility Ratio, a score measuring the capacity of a subject to understand speech

SMRT SMRT stands for Spectral-temporally modulated ripple test, this test was created by David Landsberg and Justin Aronof and is design to either measure the ripple per octave threshold of the ripple rate threshold of a subject. You can find a more precise description of this test at link to site [2]

SRS SRS stands for Safety Related Sounds, a category of sounds providing safety related information (like the presence of a danger or a dangerous situation).

Signal processing

Beamforming Beamforming or spatial filtering is a signal processing technique used to select spatially a signal having multiple channel as input. This is achieved by combining elements in an antenna array in such a way that signals at particular angles experience constructive interference while others experience destructive interference. Beamforming can be used at both the transmitting and receiving ends in order to achieve spatial selectivity

CIS CIS stands for Continuous Interleaved Sampling. It's a technique for neural stimulation that consist of delaying the impulses of the different channels/electrodes in order to stimulates only one electrode at a time. It has been observed in various study according to [18] that the interaction between channels caused by the electrical spread decreases the performance of subject in speech understanding. This technique decreases the problem compared to non-CIS process, thought it implies a kind of temporal distortion which do not significantly impair CI user's speech recognition.

EDG EDG stands for electrodogram. An electrodogram is a 2D representation (1 dimension being the electrodes number, and the second the time) of the electrical stimulation provided at the electrodes array by the implant in response of a sound stimulis (depending on the coding strategy, the amplitude of stimulation can be encoded as an electrical tension (the higher the level of the sound, the higher the electrical tension of is) or as a duration of stimulation (the higher the level of the sounds, the longer the electrical stimulation is). Those representation are close to spectrograms as each electrodes is associated with a certain frequency band.

STFT STFT stands for Short time Fourier transform, it's a signal processing technique which consists of applying the Fourier transform on short temporal frame of a signal (eventually with an overlapping on those transforms) to obtain spectrum's of those short temporal frames. Joined in respect of the temporal axis, we obtain a Spectro/temporal representation of the signal also called spectrogram.

Acoustic

Reverberation Reverberation is a natural acoustic phenomenon occurring when a direct sound is undergoing multiple reflection to the point where the sound confused. In our study, we characterize reverberation using only the RT60 coefficient

Audiofeature An audio feature is an acoustic and or perceptual characteristic who can be estimate numerically through a calculation on the raw waveform of the acoustic signal it's associate with (it associates one coefficient with an

audio waveform). Depending on the level of representation, those characteristic maybe close to the physic of sounds (zero-crossing rate for example), close to a human perception (roughness coefficient) or close to both (pitch/ F0)

RIR The RIR is the abbreviation for room impulse response, the Green function of a room when exited at a certain point by a acoustic impulsion.

$RT60$ The RT60 (reverberation time to -60dB) is the time needed to a room impulse response to loose 60 dB of level comparing to the impulse intensity.

Contents

1 Introduction

"Give every man thine ear, but few thy voice" [27]. The ability to hear is undoubtedly crucial for social interaction, and research in hearing aids has primarily focused on restoring the capacity to hear and understand speech. Although this is an important issue, losing the ability to hear also means losing significant environmental information carried by sounds, such as security-relevant information or general environmental context. This work aims to study, characterize, and address the issue of low recognition of environmental sounds by cochlear implant users, particularly concerning security-relevant sounds, and to propose designs for improving this perception.

We advise our readers to first consult [18] for an introduction to the subject of cochlear implants and the history of this technology, and then read [7] for more details on current coding strategies and electronic designs.

1.1 Companies presentation

Neurelec was originally founded as a laboratory called MXM in 1977, focusing primarily on developing implanted medical devices. In 2006, to enhance the quality of the company's activities, the MXM laboratory was divided into three subsidiaries, each dedicated to a specific focus area. One of these was Neurelec, specializing in the development of cochlear implants.

In April 2013, Neurelec became part of the William Demant Holding Group, a global leader in hearing healthcare established in 1904. This group includes Oticon and Oticon Medical. With 16,500 employees in over 130 countries, the group has unparalleled access to the latest technological advances and insights into hearing care.

Oticon Medical is a renowned global company specializing in implantable hearing solutions. Its activities revolve around two types of implants: the Bone Anchored Hearing System (BAHS) and the Cochlear Implant (CI). The development and commercialization of the BAHS implant are carried out in Sweden, while the cochlear implant is managed by the Vallauris site (formerly Neurelec) in France.

The cochlear implant and the bone anchored hearing system have different functionalities and indications. Bone anchored hearing systems are primarily indicated for patients with middle ear dysfunction and preserved inner ear structures, with sound transmitted as vibrations through the skull bone. In contrast, cochlear implants are designed for patients with sensorineural hearing loss (SNHL), often related to cochlear damage.

In 2021, Oticon Medical CI had over 280 employees and generated revenue exceeding 38 million euros. The company's main clients include hospitals, private clinics, various distributors, and patients benefiting from the produced hearing aids.

All activities related to the development and market introduction of the cochlear implant system are based at the Vallauris site. More specifically, the following

departments can be found there: Research and Development (R&D), Clinical Team, Logistics, Marketing, Production, Quality, and Regulatory Affairs (RA).

1.2 Neurelec today

In May 21, 2024, Neurelec has officially become a part of Cochlear limited.
Based in Sydney, Cochlear was established in 1981 as a subsidiary of Nucleus, with funding from the Australian government. By 2022, the company controlled 50% of the cochlear implant market, with over 250,000 people having received a Cochlear implant by 2015 . Their Nucleus 22 implant gained FDA approval in 1985, becoming the first multi-channel cochlear implant to receive such approval . Despite my internship coinciding with a major transitional period (involving changes to the offices, training courses, etc.), my work was not disturb thanks to the thoughtfulness of my tutors and to our weekly updates to keep track and discuss the advancement and any issue I could have.

1.3 The R&T team

The goal of the **Research and Technology Team** is to discover and test new solutions that can be introduced into the production chain, where they are adapted and implemented by the R&D and embedded software teams. The R&T team is divided into two groups: the *AIR* specialized in Artificial Intelligence, Image processing and Robotics. and the signal processing and acoustic team, which I was a part of .

1.4 Context and interest of the project

The project finds its interest in the augmentation of CI users well being. It was financed by Oticon medical then Cochlear company and aimed to develop a better understanding of why CI users have difficulties with the recognition and perception of security relevant sounds and then to leads to the implementation of signal processing and or informatics design in order to improve this perception.

2 Literature Review and state of the art

2.1 Normal functioning of hearing

The hearing process (also called auditory transduction) can be described as follows:

1. As a sound wave occurs, it propagates through the ear canal.
2. The sound wave undergoes mechanical impedance adaptation through the middle ear, causing the eardrum and the ossicles (the small bones) to vibrate.

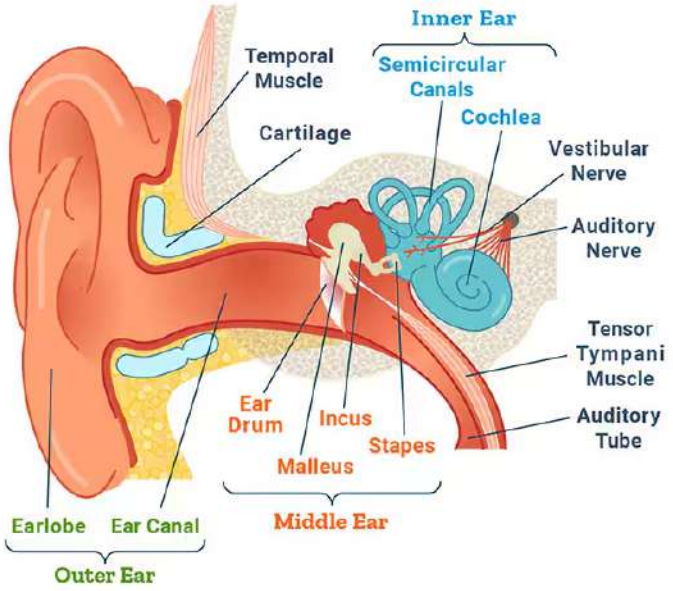


Figure 2: Anatomy of the ear (scheme from life-extension.com)

3. These vibrations cause the fluid in the scala vestibuli and scala tympani to move, making the basilar membrane within the cochlea vibrate and activating the organ of Corti.

4. Depending on the frequency of excitation, a localized part of the basilar membrane with maximal displacement will vibrate, causing the hair cells in the organ of Corti to move. These hair cells are innervated, and upon deformation, they generate neural impulses that are interpreted as sound sensations by the brain.

Within this process, localized parts of the basilar membrane are associated with specific frequencies of excitation. This property, known as tonotopy, largely explains the frequency discrimination capabilities of the ear. Neural processing in the brain leads to psychoacoustic phenomena, such as temporal and spectral masking [16].

There are three main types of hearing loss:

- Conductive Hearing Loss: This type occurs when there is a problem with the transmission of sound waves from the outer ear to the inner ear (e.g., when the eardrum is damaged). In such cases, surgery or the use of a BAHS (Link to site) can be effective solutions.

- Sensorineural Hearing Loss: This type originates in the inner ear, sensory organs, or the vestibulocochlear nerve and is prevalent according to the Pasteur

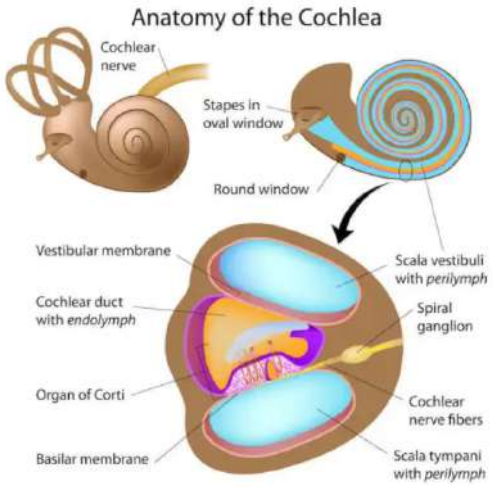


Figure 3: Focus on the inner ear (scheme from Audiocardio.com)

Institute (Link to site). The causes of sensorineural hearing loss are varied (genetic, trauma, infections, diseases, congenital malformations, etc.) and are not always treatable. Cochlear implants (CI) can address severe cases by bypassing the cochlea and directly stimulating the auditory nerves.

- Mixed Hearing Loss: This type combines aspects of both conductive and sensorineural hearing loss.

Next, we will describe the functioning of cochlear implants.

2.2 Functioning of CIs

CIs are the first examples of neuro-prosthesis that can substitute a sensory organ [20]. To elaborate, CIs allow users to hear without the sound passing through the natural auditory pathway described in the previous section, as the implant replicates the hearing process through its own chain. However, for practical reasons and efficiency in transduction, state-of-the-art CIs tend to compress auditory information and do not fully match normal auditory transduction performance in terms of dynamic range, frequency resolution, and speech recognition [7]. The scheme in Figure 4 illustrates the general signal processing chain of CIs.

This scheme is quite generic and does not capture the full range of techniques and subtleties used in audio transduction. However, some general aspects of the process, also known as coding strategy, can be described:

- When used in acoustic mode (as opposed to streaming mode, where sound input comes from streamed content), the raw output from the microphones is preprocessed to obtain a single-channel digital signal. This preprocessing

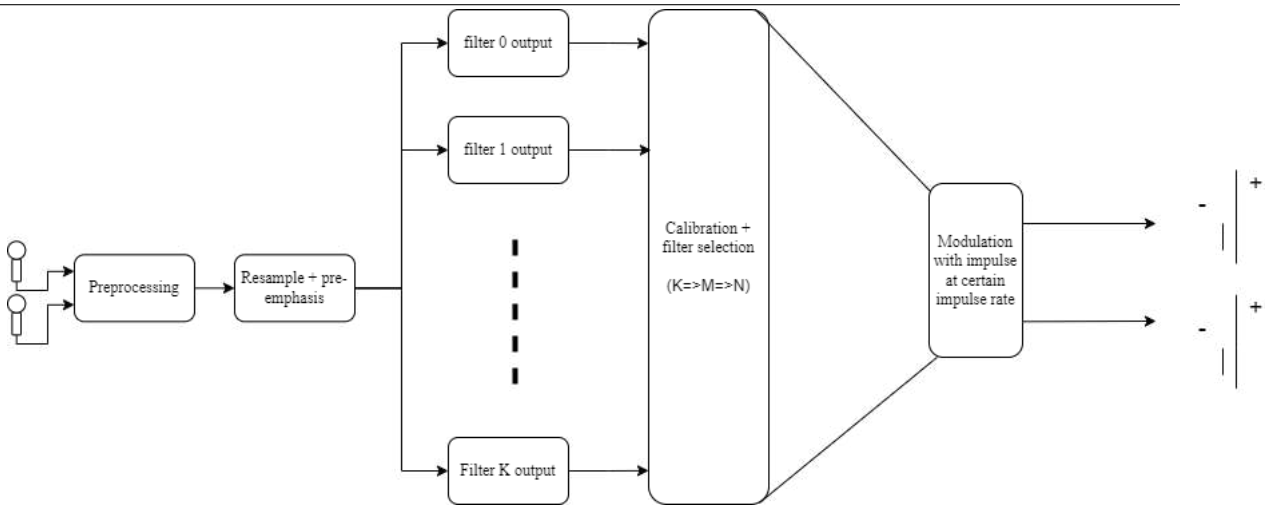


Figure 4: General processing block of CIs

may involve sampling, addition, normalization, beamforming, and some noise reduction.

- The signal is then resampled and passed through a pre-emphasis filter designed to mimic the natural mechanical filtering of sounds as they pass through the ear.

- The next stage involves band-filtering the signal. Depending on the approach, this filtering may use Fourier transforms to achieve a Short-Time Fourier Transform (STFT), band-pass filters on short audio frames to mimic the critical bands of human hearing, or other types of band-pass filters.

- A selection algorithm is then applied to choose the "bands" of frequency to encode for each frame, which are used to stimulate the cochlea via a dedicated electrode placed at the tonotopically relevant location. Note that this description does not fully capture some coding strategies like MPEAK or F0/F1/F2 [18]. We focus on the most common schemes used in recent implementations. The number of simultaneously stimulable electrodes, referred to as *Nofm*, is a variable parameter.

- The final step involves extracting the signal envelopes and modulating these envelopes with impulses generated at a constant rate. The amplitude in a given frequency band is linked with the amplitude of stimulation in the electrodes. Other encoding schemes exist; for example, Oticon Medical CIs encode amplitude using variable-duration pulses with constant amplitude. To minimize confusion due to electrical spread, pulses from different channels are subtly delayed to ensure that neurons receive the appropriate amount of stimulation. This strategy is known as CIS [18], which differs from CIs. (The frequency of

stimulation between consecutive pulses is called the pulse rate; a higher pulse rate improves amplitude envelope transmission. In our experiment, this value is fixed at 500 Hz.)

The first four parts of this process are handled by the "speech processor" of the CI. The resulting signal is then transmitted to the internal transmitter, which excites an array of electrodes placed in the cochlea at tonotopically coherent locations (or at least close to them). Note that the design of the electrode array is crucial in CI design, as it directly affects the current spread issue and the overall "electrical to neuronal" transduction.

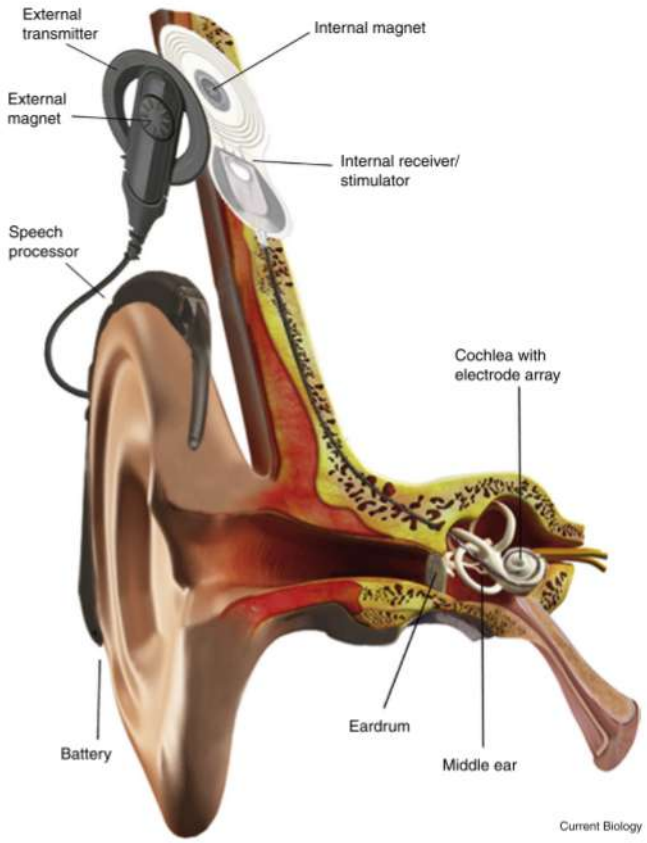


Figure 5: Scheme of an impaired patient's ear (figure from [20])

2.3 Safety related sounds

SRS (safety-related sounds) are a category of sounds carrying relevant security information for humans. Examples include alarms (especially fire alarms), human and animal vocalizations, and mechanical sounds from building sites or vehicles. We

propose a categorization based on four axes: (1) the safety rating on a scale from 1 to 5 (where 5 represents maximal danger and 1 represents no danger), (2) the need to know the source position of these sounds to ensure safety, (3) the difficulty of generating samples from a signal processing perspective, and (4) the difficulty of obtaining a database of real samples. The summarized information is shown in the following table.

	Difficulty to Obtain a Database	Difficulty to Create Samples	Need for Spatialization	Safety Rating According to [19]
Alarms	Easy	Easy	No need	4.43
Human Vocalizations	Probably Exists	Hard (Requires DL)	Need + Tracking	≈ 2.88
Animal Vocalizations	Probably Exists for Common Animals	Hard (Requires DL)	Need + Tracking	≈ 3.34
Mechanical Sounds	Uncertain	Medium to Hard	Need	≈ 4.0
Gun Sounds	Probably Exists	Easy	Need	4.62
Vehicle Sounds	Probably Exists	Medium to Hard	Need + Tracking	≈ 4.5

($\approx$) is used when the category is not explicitly addressed in [19]. The score is then estimated based on the score corresponding to sounds within the category.

These sounds have specific acoustic features that make them distinguishable using signal processing methods:

- Alarms, according to ISO 7731, should have a loudness at least +15 dBA above the ambient sound environment, frequency components between 500 Hz and 2500 Hz with a main frequency between 500 and 1500 Hz, and be modulated in amplitude with a pulse rate of 0.5 to 4 Hz. Generally, they are composed of harmonic-structured sounds. Additionally, [11] notes that some acoustic features of alarms with background noise, compared to background noise alone, are significantly correlated with a "clearly audible" human assessment (see Figure 6).

- Human vocalizations (screams, cries, rattles, etc.) are quite challenging to characterize. According to [1], the distinction between different vocalizations appears to be independent of language, although their study only includes three languages. It seems reasonable to assume that human vocalizations (such as screams and laughs) are understood and recognized across cultures. According to Table 1, key features for differentiating vocalizations include: pitch, standard deviation of inter-burst intervals (time between amplitude peaks in

the vocalization), and the first quartile of spectral energy distribution (i.e., the frequency below which a quarter of the spectral energy is gathered).

- More generally, concerning environmental sound differentiation (and not only SRS), [6] Table 3 shows that among various types of audio features, spectral and cepstral features such as MFCC and brightness are most useful for classifying environmental sounds (based on the BDlib-2 library). This is corroborated by [5], which also includes temporal cues such as zero-crossing rate. Figure 7 shows the ranking of the 10 best features for separation in [5] and [6]. However, it should be emphasized that the models used in these studies are quite far in performance from state-of-the-art models. Thus, the most relevant features for these simpler models may not be the ones allowing the best differentiation in the most advanced models.

In conclusion, the recognition of SRS and environmental sounds strongly depends on spectral and cepstral (MFCC) analysis, and to a lesser extent on temporal analysis.

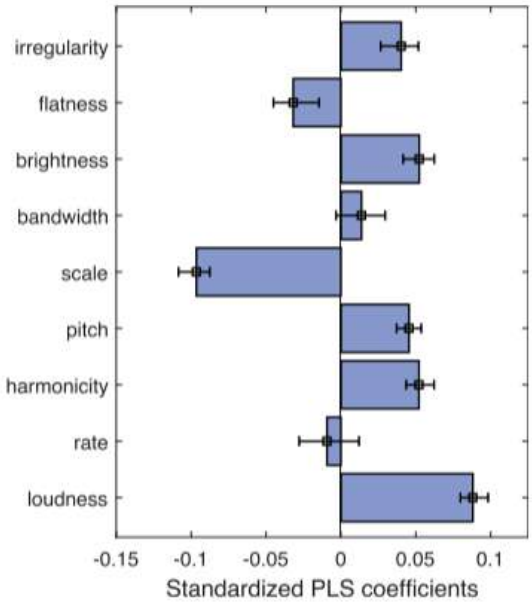


Figure 6: Coefficient of the first PLSR component. Acoustic features whose coefficients have bootstrap distributions above or below zero are considered statistically significant predictors. Figure from [11].

2.4 The Problematic Issue of Relevant Security Sounds for Cochlear Implant Users

As discussed in Section 1.2, cochlear implants (CIs) encode sound into electrical stimulation using one of the coding strategies and material configurations detailed in [7]. Initially optimized for processing speech [8], these implants have evolved to more generic designs [18], though they are still primarily evaluated based on speech perception.

Additionally, as mentioned in Section 1.3, acoustic features that are useful for differentiating environmental sounds depend on spectral (e.g., pitch, brightness, and frequential centroid) or cepstral analysis (e.g., MFCC, LPCC). Due to the constraints of CI design—such as having a limited number of electrodes or using broad band filters to decompose input sounds—CIs often evaluate spectral content imprecisely and, consequently, the cepstral content as well. This limitation likely contributes to the difficulties CI users face in recognizing speech-relevant sounds (SRS) and other environmental sounds.

Despite these challenges, some CI users, according to [21], continue to use their implants frequently (more than 5 hours a day) even without significant improvements in speech processing (see Figure 8). This suggests that obtaining environmental sound information may be perceived as a substantial improvement in their overall perception of the world. The following figure, extracted from [21], highlights this trend.

Each line in the figure represents a subject's daily CI usage, showing their average usage at the most recent follow-up and three months prior. It is clear that a significant majority of users maintain long-term use of their implants despite poor speech intelligibility ratios (SIR), likely because the ability to perceive environmental information is considered a substantial benefit, even though it does not match normal hearing perception.

In [23], the authors demonstrate that CI users have a reduced capacity to discriminate natural soundscapes across three modalities: habitat, time of day, and seasons, compared to normal hearing individuals. This limitation could potentially impact their safety.

Furthermore, [19] shows that CI users struggle to recognize certain security-relevant sounds (e.g., only 38% recognition accuracy for explosions in a quiet context, despite high confidence in the relevance and their knowledge of the sound after brief training).

Therefore, it can be hypothesized that achieving satisfactory perception of security-relevant sounds remains a challenge for CI users, and improving this aspect could provide significant benefits.

2.5 Coping Strategies for Integrating Security Sounds

To our current knowledge, there has been limited work on the integration of relevant security sounds for cochlear implant (CI) users. Currently, these sounds are often treated as noise in relation to speech, as evidenced by systems such as Oticon's SuddenSound stabilizer [**Santurette2023**] and Apple's Conversation Awareness with Loud Sound Reduction [**Apple2023**]. Additionally, even without considering the issue of removing speech-relevant sounds (SRS) from audio transduction, the signal processing aspects that contribute to the salience and audibility of SRS for CI users remain unclear. One of the goals of this work is to identify these components to develop an efficient response mechanism to alert users.

Table 3. Top 20 features for the SFb method.

Rank Order	Ranking Algorithms		
	IGR	GLM	SVM
1	MFCC1	MFCC3	MFCC3
2	ENTR	MFCC1	MFCC2
3	BRI	MFCC2	ENTR
4	ROL	SFM15	MFCC1
5	MFCC2	MFCC4	BRI
6	CEN	ZCR	CEN
7	SKEW	VAR	VAR
8	SFM12	SFM18	MFCC4
9	KURT	SFM14	SMO
10	SFM10	ROU	ROL

(a) Best features for environmental sound classification depending on the model of classification from [6].

Table 2. Ranking of the salient features

Rank Order	Ranking Algorithms	
	InfoGainAttributeEval	OneRAttributeEval
1	Brightness	ZCR
2	ZCR	SFM_12
3	SFM_12	MFCC_2
4	MFCC_1	Brightness
5	MFCC_2	LPCC_12
6	SFM_13	MFCC_1
7	Spectral Smoothness	SFM_10
8	LPCC_12	LPCC_11
9	MFCC_3	LPCC_10
10	SFM_11	Spectral Smoothness

(b) Best features for environmental sound classification depending on the method of relevance evaluation from [5].

Figure 7: The best features for environmental sound classification on BDlib.

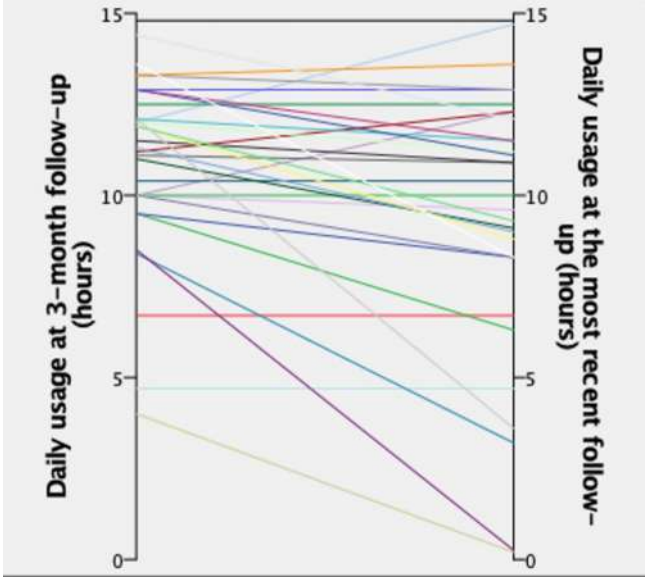


Figure 8: Daily average time of CI usage among subjects from [21]

3 Analysis of SRS in CI Processing

To address the issue of SRS recognition, we analyze the effects of CI audio transduction by comparing electrodograms (EDG) and short-time Fourier transform (STFT), and by examining important audio features. Initially, we focus on isolated SRS (due to time constraints, we worked with alarm signals) and then on SRS combined with environmental noises (in this case, we used alarms signals with specific morphed white noise, which will be described in the relevant section). Our analysis supports two claims about CI transduction that may explain the poor alarm detection and recognition experienced by CI users:

- Some spectral modulations of alarms are above the detection thresholds of CI users, making it difficult for them to recognize these alarms.
- Features important for the salience and audibility of alarms, according to [11], are degraded throughout the processing chain.

3.1 Data Collection and Process of Analysis

To begin the study, the first task was to collect relevant data. In our case, these data are sound files of SRS and environmental sounds (such as speech and vocalizations from humans and animals). We chose to use sounds from peer-reviewed databases such as:

- The Acted Emotional Speech Dynamic Database [24] [29]. This database contains speech (Greek in this case) expressed with 5 different emotions (anger, disgust, fear, happiness, and sadness). We consider speech expressing anger, disgust, and fear as SRS, and the other emotions as environmental sounds.
- The Montreal Affective Voices [4]. This database contains human vocalizations pronounced with the vowel "A," where subjects try to reproduce vocalizations associated with various emotions (such as anger, happiness, and fear). We use the stimuli associated with negative emotions as SRS and the others as environmental sounds.
- The audio used in the Luzum sound recognition study [19]. These are "everyday sounds" (vehicles, animals, objects), some of which are classified as SRS and others as environmental sounds.
- The BDlib-2 [5] [6]. This database includes 20 categories of environmental sounds such as rain, thunder, airplane, and applause. Sounds indicating danger (such as thunder, alarms, and screams) are categorized as SRS, while the rest are classified as environmental sounds.

- The VoxCeleb dataset [9], which contains speech pronounced by celebrities. This entire dataset is used as environmental sounds to construct babble noise, for example.
- The ESC-50 [26]. This database consists of a large variety (50 categories) of environmental sounds, such as keyboard typing, fire crackling, and brushing teeth.

All these sounds were sorted according to their relevance to safety. We classified sounds carrying security information (such as alarms, human vocalizations with negative emotions, and vehicle sounds) as signals, while all other sounds were classified as noise.

As mentioned earlier, we only worked with alarm sounds as SRS in this study, leaving the investigation of other SRS for future work.

To create the test sounds, denoted as stimuli used in our investigation, we followed the processing scheme shown in Figure 9.

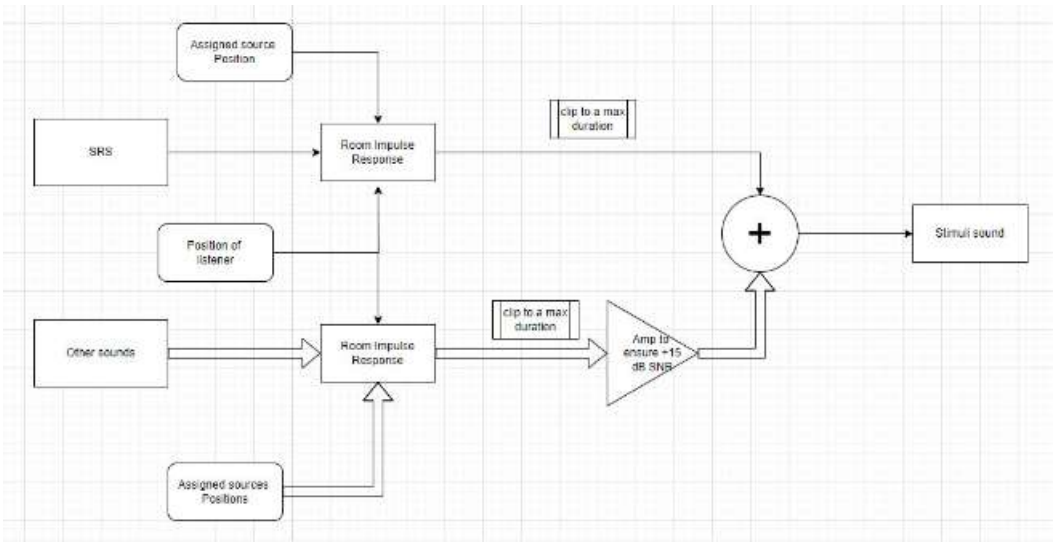


Figure 9: Process to generate the stimuli sounds

Thick arrows indicate potential multiple sources/sounds (i.e., in this process, a stimulus sound contains at most one SRS and multiple noise/environmental sounds). The general process involves applying a room effect to the SRS and the noise based on the algorithm described in [22] (an implementation of the image/source model of room impulse response). The amplitude of the noise is then controlled to obtain a stimulus sound with a specified SNR (calculated over the total duration of the signal).

Since the [22] algorithm uses the localization of the source/microphone, room dimensions, and acoustic reflections (same for all surfaces) to calculate the RIR, and uses the T60 as an input, we fixed the room dimensions and used the Eyring

formula to calculate the reflection coefficient corresponding to the desired T60. The same positions and room dimensions were used for all signals.

$$1 - C_r = \exp\left(\frac{0.163 \cdot V}{-T_{60} \sum_i S_i}\right)$$

The general nomenclature of these sound files was:

$$Signal - XX.wav - with - Ys - T60 - SNR - Z$$

where XX is the name of the original wave file, Y is the T60 value applied to the file, and Z is the SNR. In this first stage of analysis, only alarms without noise were studied (i.e., with infinite SNR).

3.2 Analysis of the SRS Processing Alone

Having correctly created those sounds, we proceed to our first analysis. We use Oticon Medical's CI process chain to obtain the EDGs of these signals and compare them with the spectrogram of the input. The goal is to gain an initial understanding of the influence of various parameters and the CI on the acoustic information.

3.2.1 Analysis of Electrodograms

The EDGs were created by varying three parameters:

- The T60 of the signal (i.e., the T60 used to create the RIR filter that processes the input signal). The values tested were 0s (simulating a free field situation), 2s (simulating a typical room environment), and 15s (simulating an over-reverberated environment such as a church).
- The Input Level, a parameter from the CI process chain that normalizes the sound to simulate different dBSPL levels. The tested values were 50 dBSPL (quiet environment level), 70 dBSPL (conversation level), and 85 dBSPL.
- The NofM, another parameter of the chain, which is the maximum number of channels (i.e., electrodes) that can be stimulated simultaneously (due to CIS, "simultaneously" refers to a certain frame of stimulation rather than absolute simultaneity).

An example of these EDGs is shown in Figure 10. In these plots, the spectrogram is shown in the background (with darker colors representing higher amplitudes). The EDG is superimposed in orange, with each line representing a channel (i.e., an electrode) and aligned with the center frequency of its channel. The amplitude is

represented by the thickness of the line, and you may notice that these plots are often superimposed.

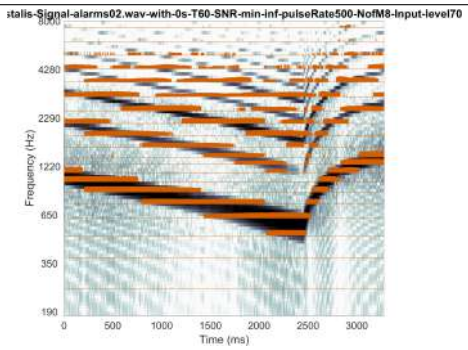
The first observation is that the cochlear implant drastically reduces spectral resolution compared to normal hearing listeners, who have up to 144 channels between 100 and 8000 Hz (considering 6 octaves and 24 steps per octave). In practice, most CI users have fewer than 20 independent channels due to their neurological condition and current spread issues.

A second observation is that the EDGs have a minimum threshold to trigger an amplitude (notably visible in Figure 10(a)) and a maximum amplitude, above which higher levels are clipped (as seen in Figure 10(c) compared to Figure 10(b)). These maximum and minimum levels are known as C and T levels (C for comfort, the maximum stimulation level, and T for threshold, the minimum stimulation level). These levels depend on the patient and the specific channel and can change over time. In summary, the dynamic range of CI users is significantly reduced by the process chain to match the user's capabilities.

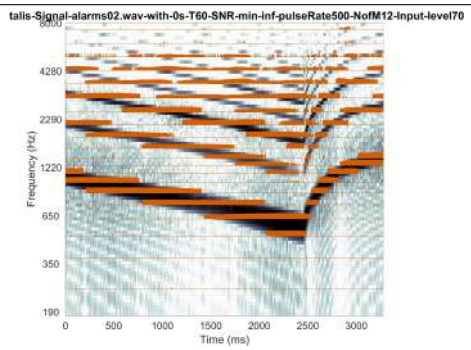
Lastly, the EDGs exhibit good temporal resolution, with a temporal frame typically consisting of 128 samples at a 16 kHz sample rate (i.e., a 2 ms frame).

Regarding parameter influence, reverberation has the most noticeable impact, blurring frequency patterns over time, particularly in high frequencies (above 2290 Hz), which can decrease alarm recognition due to the loss of harmonicity. Other parameters have more limited effects. Input levels increase EDG stimulation amplitude until clipping occurs and can trigger additional channels if the stimulation level exceeds the T levels. The NofM parameter modifies the number of available channels in a given frame but does not necessarily cause different EDGs when it variates.

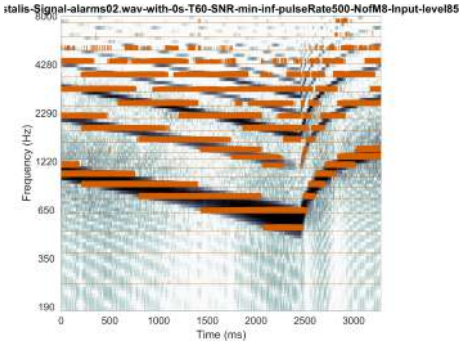
At this point, we aimed to study the most correlated feature with audibility as discussed in [11]. Specifically, we wanted to compare the features of input sounds with their "encoded version" in the EDG. However, we encountered a methodological challenge. While the EDG provides an objective representation of sound and probably close to the perception for CI users, it is not a unique representation of sound, in particular concerning spectral feature evaluation. Therefore, we decided to use vocoders to reconstruct sounds from the EDG and compare the features of these vocoder outputs with those of the input sounds.



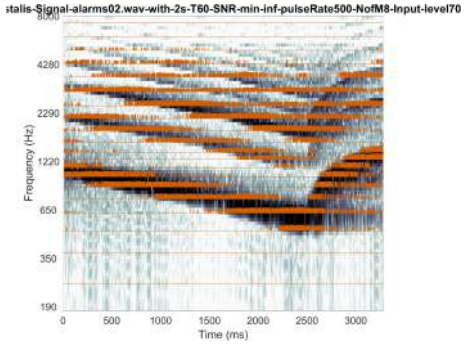
(a) Signal with 2s T60, 70 dB input level, and NofM=8



(b) Signal with 0s T60, 70 dB input level, and NofM=12



(c) Signal with 0s T60, 85 dB input level, and NofM=8



(d) Signal with 0s T60, 70 dB input level, and NofM=8

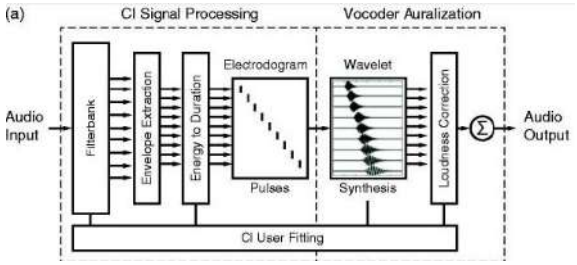
Figure 10: Examples of EDGs superimposed with the spectrogram

3.2.2 Description of the Vocoders

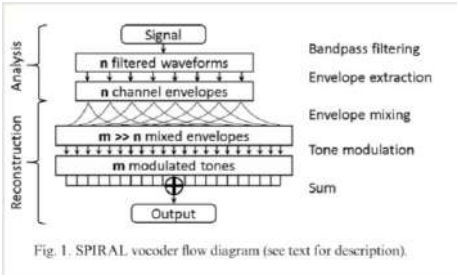
We used two different vocoders during our analysis: the sine vocoder and the spiral vocoder.

These vocoders take as input the EDG, which can be seen as an N-channel envelopes array in Figure 11 (b), and compute the audio auralization from it. The process consists of extracting the amplitude envelopes for each channel and multiplying them with a sine wave at the frequency equal to the center frequency of the corresponding channel. Note that this is done under the assumption that when a channel is triggered, it is only by a sine wave with the center frequency of that channel, which is not always the case in our situation. The sine waves are then summed across the different channels.

The spiral vocoder differs in that it adds an envelope mixing stage where neighboring channel envelopes are combined with a decrementing coefficient as they become further neighbors. This additional stage aims to mimic the current spread phenomenon, providing a closer representation of what CI users may hear. Consequently, we expect the correlation score of spiral/input features to be lower than the correlation score of sine/input features. We may also observe some effects of



(a) Scheme of the sine vocoder from [3]



(b) Scheme of the spiral vocoder from [13]

Figure 11: Concept schemes of the vocoders

spreading on the degradation of features.

3.2.3 Analysis of the Spectral Modulation

Among the features we studied (to be described in the next subsection), the scale feature was particularly important. The perception of spectral modulation has already been studied concerning CI users, notably through the use of the SMRT. We aimed to investigate whether part of the alarm's frequency modulation might be imperceptible to CI users.

To verify this, we calculated the frequency modulation spectrogram for all our signals (using the NSL Matlab toolbox). Knowing the detection threshold from previous studies, such as [15] (around 9 ripples per octave (RPO) for adults with normal hearing and around 4 RPO for CI users), we calculated the proportion of modulation components above the threshold divided by the total number of components for all time frames and signals. Figure 12 shows an example for one signal. We then obtained the rate of imperceptible modulation for CI users compared to normal hearing subjects.

The calculation follows the formula:

$$\text{Not perceived energy ratio}(t) = \frac{\sum_{n=\text{threshold}}^{\max \text{ RPO}} \text{spec modulation}(t,n)^2}{\sum_{n=\min \text{ RPO}}^{\max \text{ RPO}} \text{spec modulation}(t,n)^2}$$

We obtained a mean value of 7.97% for normal hearing subjects and 23% for CI users, with a standard deviation of 3%. This score even rises to 45.29% when calculated from internal clinical values of Oticon Medical (mean RPO at 2.145), validating our initial claim. See link to claim.

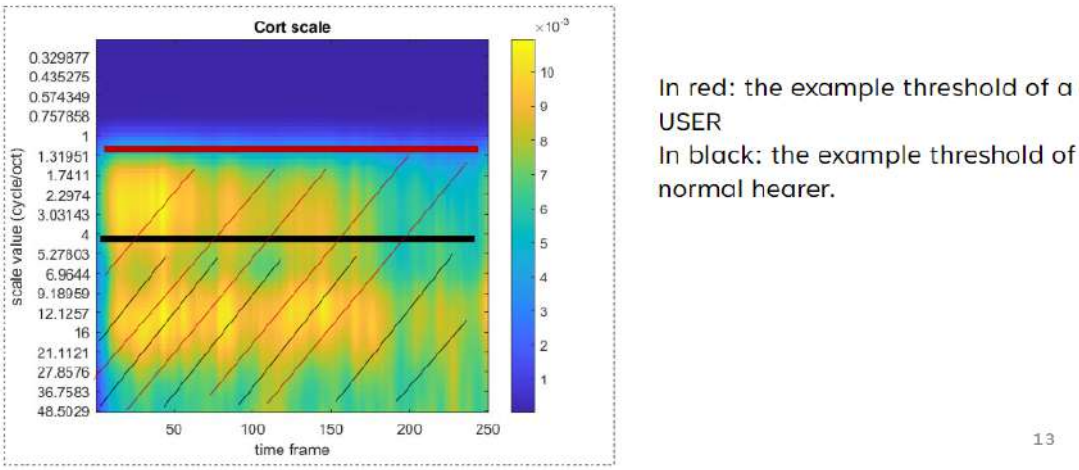


Figure 12: Example of frequency modulation spectrogram. The red hashed area represents the part of the modulation spectrum not perceived by CI users, while the black area represents that perceived by normal hearing individuals. Threshold values are arbitrary, just to illustrate the discussion.

3.2.4 Analysis of CI Vocoded Sounds Features vs. Actual Sounds Features

Let's first describe more precisely the set of features we will study:

- Loudness: Corresponding roughly to an input level defined on a time frame, extracted using the Zwicker model. Calculated from the MATLAB function 'acoustic-loudness'.
- Harmonicity: A value describing the proximity of a spectrum to a template spectrum representing a totally harmonic sound (see [14] for more details). Due to inconsistent results with our implementation following Pr. Ehlilali's original code, we instead used the harmonic-ratio measure from MATLAB, estimating this ratio using the autocorrelation method.
- Pitch: In our case, this refers to the F0 value. Due to inconsistent results with our implementation following Pr. Ehlilali's original code, we used the autocorrelation estimation method from MATLAB.
- Scale: The centroid along the frequency modulation axis of a cortical representation of frequency modulations, calculated using the NSL MATLAB toolbox.
- Brightness: The frequency centroid of a spectral profile, calculated using the formula:

$$BR(t) = \frac{\sum_f f \cdot spec(t,f)^2}{\sum_f spec(t,f)^2}$$

- Flatness: Calculated as the geometric mean of the spectrum divided by the arithmetic mean:

$$FL(t) = \frac{\sqrt[N]{\prod_{f=0}^{N-1} spec(t,f)}}{\sum_{f=0}^{N-1} spec(t,f)/N}$$

- Irregularity: Extracted as a measure of the difference in strength between adjacent frequency channels, defined as:

$$IR(t) = \frac{\sum_f (spec(t,f+1) - spec(t,f))^2}{\sum_f spec(t,f)^2}$$

These features were chosen because they are, according to [11] Figure 6, significant predictors of audibility attributes for normal hearing individuals. Thus, checking their degradation can help explain the lack of detection and recognition of alarms by CI users. They were calculated on 256-sample frames with a 128-sample overlap, aiming to use analysis parameters close to the sound processor's capacity to facilitate potential future feature estimation on this processor and to provide an initial idea of the precision to be expected.

We chose the Pearson correlation score as the metric to evaluate degradation, defined as:

$$Corr\ score(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

We hypothesize that there is a monotonic relationship between the correlation score of our input and vocoded features and a perceptual alarms audibility score. Hence, the correlation score of features provides an indication of the perceptual loss experienced by CI users (an experiment to verify this hypothesis is proposed in the Clinical Protocol section).

Additionally, we included two new parameters in our analysis:

- Number of Deactivated Electrodes (values: 1, 4, 8, 12, 14, 17). Until our hypothesis about the existence of a mapping is validated (or not), we wanted a measure of feature perception degradation by the CI. Since the effect of electrode deactivation has been more studied, we expected to establish anchors for each feature and make comparisons such as: "The mean degradation (correlation score) of the distribution of feature N is comparable to the mean

degradation of this distribution for M deactivated electrodes." Unfortunately, as we'll see, the CI behaves differently. Electrodes are deactivated "sparsely," meaning they are spaced out consistently (e.g., for 1, the 10th electrode is deactivated; for 2, the 7th and 14th electrodes, etc.).

- Compression (values: "auto", "loud", "medium", "quiet"). This parameter sets the dB HL of compression knee points for each selected channel, as shown in Figure 13. When set to "auto", the knee points are calculated automatically based on RMS. We expected that having the largest dynamic range possible would improve feature conservation (i.e., that the "auto" setting would give the best results concerning feature correlation scores).

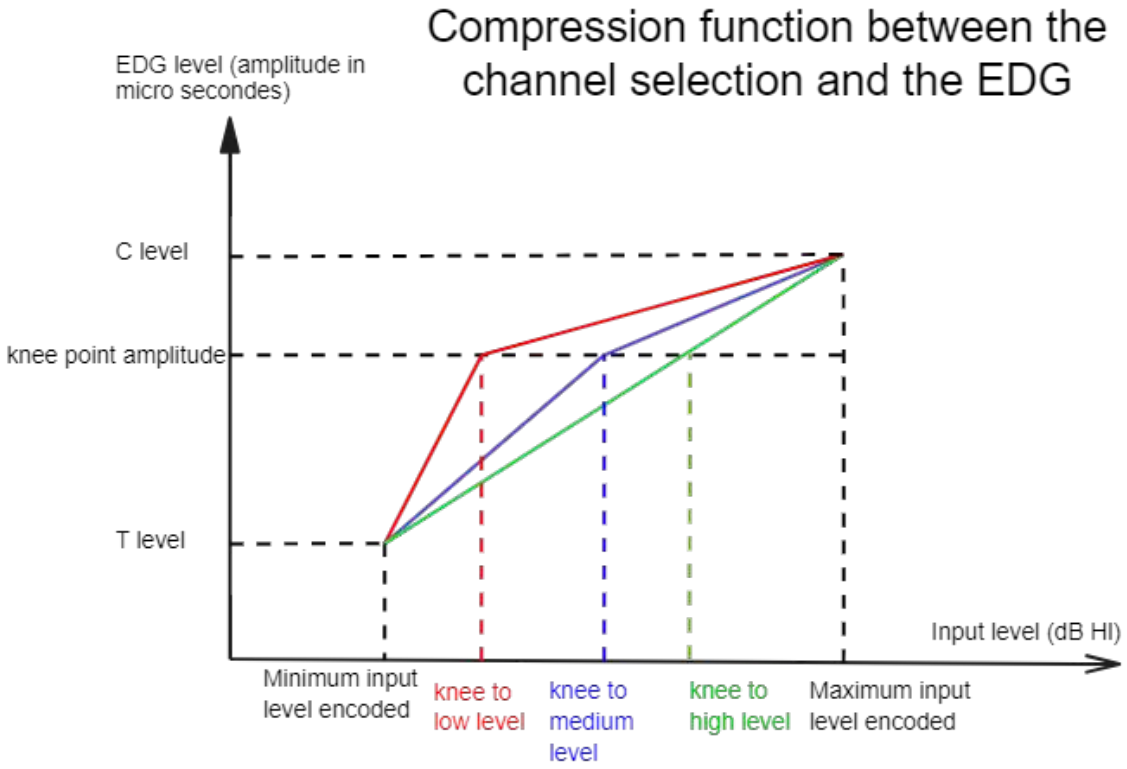


Figure 13: Example of compression function for one channel

We analyzed a total of 20 different alarm sounds, resulting in 432 configurations ($20 \times 6 \times 4 \times 2 \times 3 \times 3 = 8640$ different sound files processed). The following table summarizes the different configurations tested:

Sound File Name	Input Level (dB)	T60 (s)	Nofm	Number of Electrodes	Compression
XX.wav	50/70/85	0/2/15	8/12	1/4/8/12/14/17	auto/high/medium/low

Figure 14 describes the process of calculating the correlation score starting from an audio file. Different colors emphasize the type of data: red for audio signal, green for EDG, blue for audio features, and purple for correlation scores.

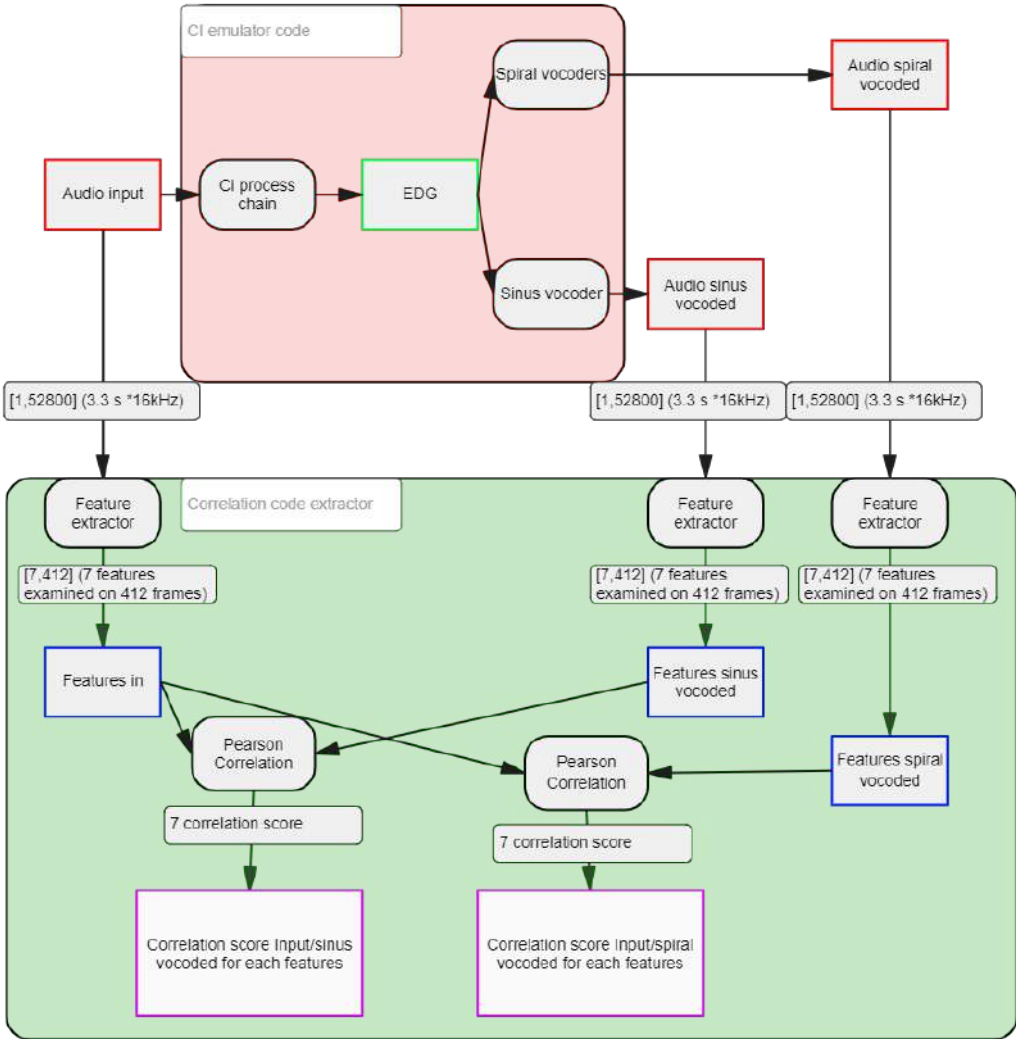


Figure 14: Process of correlation score calculation

Note that, in addition to the 7 features mentioned, we also used the STOI [28] measured between input and vocoded sounds (interpreted as a global degradation measure of the sound and of acoustic features relevant to speech perception by the vocoder).

3.2.5 Results of the Feature Correlation Analysis

We aimed to observe the global effect of the CI on acoustic features. For the 8640 files, we calculated the correlation scores of the inputs and outputs, resulting in the data shown in Figure 15.

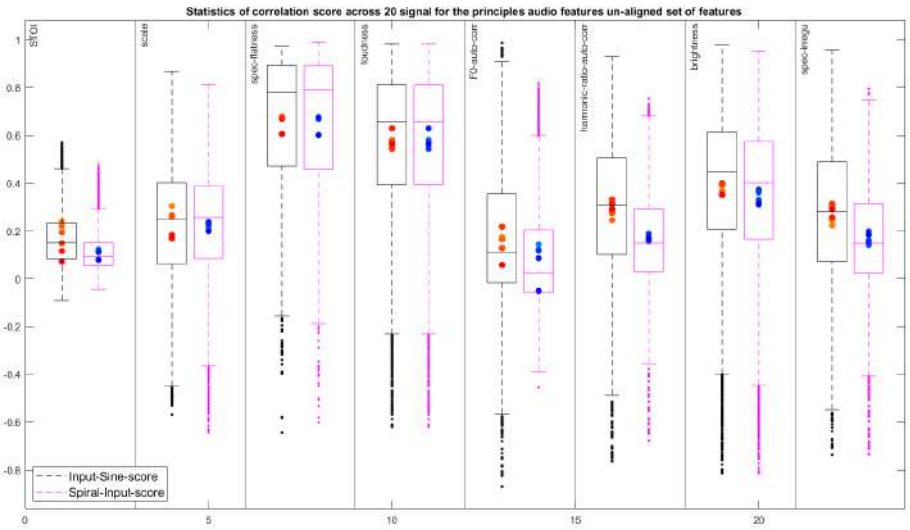


Figure 15: Statistics of correlation scores of features according to the vocoder. The Y-axis represents the correlation score, and the X-axis distinguishes the distributions based on vocoder type and feature. The features are referenced in the top left part of different columns.

As expected, the spiral-input feature correlation scores are generally lower than the input-sine correlation scores. Additionally, the data for the loudness feature are identical because the loudness of the output cannot be computed from the vocoder's output, primarily due to rescaling during the inversion of the EDG to N-channel components. The "CI loudness" is estimated by summing the pulse amplitude of the EDG over the time frame processed and the channels. The CI loudness is calculated from the EDG using the following formula:

$$loudness(t) = \sum_{channels} \sum_{n=t_0(t)}^{t_{end}(t)} scaled\_edg\_charge(chan, n)$$

where $t_0(t)$ and $t_{end}(t)$ are the first and last sample numbers of the time frame t. The scaled EDG charge is obtained by subtracting the maximum accepted value (C level) for each channel. We use a standard C and T levels set of parameters for all configurations. This calculation is consistent across both vocoders.

Furthermore, we observed some configurations with negative correlation scores. Initially, we thought this issue could be resolved by artificially synchronizing the

input and vocoded features, but this did not yield a significant difference. We assume this is part of the CI's effect on acoustic features.

Apart from spectral flatness and loudness, the correlation scores of other features are generally close to 0. We suspect that the lack of frequency channels is the reason, as these features depend on spectral content. The higher scores for flatness and brightness can be attributed to the fact that they are mean values, making them less sensitive to blurring compared to F0, which theoretically depends on a unique frequency component.

At this point, we have supported our second claim (assuming a link between the loss of correlation score between features and poor audibility of CI users for alarms) link to claim. The most intriguing aspect is the behavior of the dots.

The red and blue dots represent the means of the correlation scores of features for a fixed number of deactivated electrodes. From light to dark blue/red, there are 6 points representing means with 1, 4, 8, 12, 14, and 17 electrodes deactivated. The idea was to use these points as anchors to interpret correlation scores in terms of perception. Indeed, the effect of deactivating electrodes on perception has been studied for CI users [12] [10] (although most of these studies focus on speech perception). We expected to see higher scores with fewer deactivated electrodes (as adding electrodes increases information). This behavior is observed for the STOI and scale features. For spiral-input correlation scores, the order of these anchors is more uncertain, as eliminating channels results in a loss of information but also affects the interaction between channels.

Since the input-sine anchors for other features did not behave this way, we suspected interactions between parameters and the number of electrodes. For example, interactions between NofM and the number of deactivated electrodes are possible; above 12 deactivated electrodes, NofM makes no difference because there are fewer than 8 available channels.

We then conducted a repeated 5-way ANOVA for each feature, with the dependent variable being the correlation score and the parameters including NofM, Input Level, T-60, Compression, and the number of deactivated electrodes.

The ANOVAs (see Figure 16 and Figure 17) revealed that some parameters indeed interacted, and the nature of these interactions differed depending on the vocoding type (green indicates $p < 0.05$, i.e., the null hypothesis is rejected, while red indicates $p \geq 0.05$, i.e., the null hypothesis failed to be rejected).

The standard error bars of these plots were calculated according to [25]. Due to the number and complexity of the plots, general conclusions about interactions across features are challenging. To verify and observe the specific interactions, we plotted the standard errors with mean plots for the relevant subgroups. For clarity and conciseness, we present and describe four examples below (all these distributions were found by our ANOVA to have at least one significantly different mean distribution, with $p < 0.05$).

We clearly see that input/sine STOI scores are more organized and follow a

	STOI	scale	loudness	F0_auto_corr	harmonic_ratio_auto_corr	brightness	spec_flatness	spec_irregu
Compression	7,14913E-22	1,14708E-08	8,85251E-14	0,017982513	1,58456E-05	9,14888E-05	2,00949E-14	0,067284879
Input_level	0,31903286	0,849731677	1,53289E-08	0,803126773	0,457967172	0,001388924	5,52209E-11	0,236781338
T_60	1,10131E-06	0,380681065	1,10388E-09	0,021367065	0,007112848	0,066013060	0,00059306	0,062724581
Nofm	4,53831E-05	0,042811497	0,004005123	0,817438448	0,466867625	0,430443125	0,44971827	0,858363527
deactivated_electrodes	1,60786E-45	6,08741E-11	0,020228181	0,061308247	0,003843119	0,073549607	1,77147E-12	0,065970927
Compression:Input_level	0,034028401	0,29038879	3,08864E-13	0,112823828	0,033107907	0,030645148	1,65182E-14	0,230416312
Compression:T_60	0,080179288	0,012534381	0,003057874	0,066909444	5,34837E-14	1,1498E-10	0,134381763	1,10248E-22
Input_level:T_60	0,233138393	0,000416097	0,018748012	0,061193001	1,44761E-11	0,30062859	3,54787E-21	0,80645741
Compression:Nofm	0,419003231	0,906694128	0,000000469	0,298226733	0,820851322	0,118639178	0,71460109	0,630056433
Input_level:Nofm	8,84579E-07	0,954812551	0,000743347	0,087025538	0,274141181	0,483878501	0,621388891	0,882713964
T_60:Nofm	0,319033382	0,131680309	7,78738E-05	0,820527964	0,071481452	0,210799923	0,577081418	0,879096587
Compression:deactivated_electrodes	8,32825E-67	1,37965E-14	0,4571892	0,052776795	1,8720E-07	0,080982221	0,001960826	0,015681024
Input_level:deactivated_electrodes	0,093102808	1,84151E-13	4,48174E-19	0,84926089	0,012312504	1,51444E-09	0,01523299	1,80116E-08
T_60:deactivated_electrodes	8,14958E-24	7,70734E-06	0,004038457	6,81132E-05	0,000144185	4,42465E-09	2,38102E-21	1,70988E-08
Nofm:deactivated_electrodes	1,51926E-15	0,022225906	7,64873E-07	0,483911901	0,64498183	0,486237397	0,466009707	0,655285606
Compression:Input_level:T_60	0,31997711	0,117735835	0,003426583	0,094171451	0,005125215	0,098048899	1,94778E-19	0,055072715
Compression:Input_level:Nofm	0,005648327	0,848858274	0,011413323	0,863951747	0,976426446	0,728688573	0,85618865	0,824233282
Compression:T_60:Nofm	0,448823461	0,324142353	0,344213985	0,013012885	0,310776126	0,483391041	0,583061730	0,228061679
Input_level:T_60:Nofm	0,582418754	0,041233897	5,81292E-05	0,380246745	0,420870296	0,382947832	0,773254448	0,958824861
Compression:Input_level:deactivated_electrodes	2,42156E-19	1,12954E-08	8,77098E-13	0,031588254	0,534380967	0,080151654	0,049732813	1,46031E-08
Compression:T_60:deactivated_electrodes	8,47281E-06	0,002489677	1,28537E-05	0,012634753	0,022394301	1,42595E-06	1,02980E-07	0,080684096
Input_level:T_60:deactivated_electrodes	0,684232261	0,017032384	4,23779E-09	0,155647679	4,13839E-05	3,16263E-17	0,00253145	1,85998E-22
Compression:Nofm:deactivated_electrodes	0,088621470	0,888248988	6,15654E-11	0,186204828	0,81509855	0,119588847	0,375319836	0,611608776
Input_level:Nofm:deactivated_electrodes	8,16489E-24	0,231773593	1,32238E-08	0,370703307	0,274655750	0,685284294	0,917624817	0,928796881
T_60:Nofm:deactivated_electrodes	0,089027408	0,048512956	2,55949E-09	0,228794837	0,002688488	0,1727443	0,560929888	0,454568119
Compression:Input_level:T_60:Nofm	0,475884123	0,339697708	0,374922744	0,055703158	0,880408581	0,7805369	0,995467143	0,197375873
Compression:Input_level:T_60:deactivated_electrodes	0,154429584	0,001242969	1,70818E-09	0,01078373	0,050349183	3,78447E-06	5,40218E-07	7,79929E-07
Compression:Input_level:Nofm:deactivated_electrodes	1,47523E-12	0,921855883	6,52252E-05	0,290453903	0,896309076	0,982268648	0,790356331	0,988706899
Compression:T_60:Nofm:deactivated_electrodes	0,972339825	0,003389444	0,999816017	0,008121844	0,353411398	0,522676042	0,070105614	0,063182455
Input_level:T_60:Nofm:deactivated_electrodes	0,728162508	0,220324072	3,48318E-09	0,443671828	0,781489239	0,372713282	0,090321528	0,273651261
Compression:Input_level:T_60:Nofm:deactivated_electrodes	0,966665482	0,75	0,337752114	0,318675832	0,988056471	0,993118671	0,99882706	0,000381558

Figure 16: P-value table of the 5-way repeated measures ANOVA of the correlation input/sine scores

	STOI	scale	loudness	F0_auto_corr	harmonic_ratio_auto_corr	brightness	spec_irregu	spec_flatness
Compression	2,34E-17	0,000047	9,85E-14	0,00332756	0,012063008	1,18E-05	0,000445561	3,00424E-14
Input_level	8,71E-07	0,677432	1,53E-06	0,719724604	0,008330653	0,011737	0,612875498	2,902E-09
T_60	0,001839	0,820064	1,19E-09	8,2429E-08	0,000908477	0,001375	0,015994202	2,95302E-05
Nofm	0,403377	0,351813	0,004800	0,604249905	0,980429471	0,390015	0,055249605	0,207198471
deactivated_electrodes	5,75E-20	0,113782	0,020228	6,07852E-12	0,122524325	0,020574	0,018530057	3,00705E-13
Compression:Input_level	0,036307	0,002419	2,1E-13	0,187792256	0,025215406	0,009359	0,899042484	9,49777E-18
Compression:T_60	0,361554	0,000298	0,003058	0,009756887	6,5314E-08	1,71E-12	5,56501E-14	0,272361386
Input_level:T_60	0,372465	0,141668	0,016748	0,1151004	1,99443E-11	3,23E-05	0,003238494	2,31123E-20
Compression:Nofm	0,311725	0,767559	0,0003	0,671107723	0,576861933	0,129727	0,211107294	0,386010906
Input_level:Nofm	0,676282	0,318166	0,000743	0,706869127	0,965331592	0,515361	0,181689117	0,104483561
T_60:Nofm	0,390736	0,202819	7,77E-05	0,229976893	0,978066058	0,153592	0,251706294	0,138571639
Compression:deactivated_electrodes	1,34E-09	0,048336	0,457189	0,000198072	0,000296085	3,63E-08	5,81002E-05	9,20571E-06
Input_level:deactivated_electrodes	0,274255	0,080406	4,48E-19	0,043019114	0,896656742	5,48E-10	1,92699E-16	0,000144478
T_60:deactivated_electrodes	0,008516	0,695872	0,004036	0,027695102	8,13826E-06	2,11E-10	0,000147591	1,37098E-22
Nofm:deactivated_electrodes	0,483512	0,917573	7,65E-07	0,617987888	0,469407139	0,437411	0,011433736	0,141672905
Compression:Input_level:T_60	0,0537	0,103419	0,002427	0,00259375	0,000357743	0,016784	0,037950403	2,23503E-28
Compression:Input_level:Nofm	0,717197	0,931798	0,011413	0,507982573	0,535372948	0,259165	0,140840597	0,105487487
Compression:T_60:Nofm	0,685297	0,700145	0,944214	0,026914422	0,814140477	0,101335	0,143032204	0,21962054
Input_level:T_60:Nofm	0,245228	0,755761	5,81E-05	0,05178296	0,908089732	0,264563	0,091726623	0,3088723
Compression:Input_level:deactivated_electrodes	0,001256	0,027573	8,77E-13	5,31162E-09	0,527361893	2,33E-06	2,46296E-16	0,309357022
Compression:T_60:deactivated_electrodes	0,433345	0,784472	1,29E-05	0,046860498	0,390660278	0,000273	0,000169743	5,05146E-08
Input_level:T_60:deactivated_electrodes	0,562588	0,232956	4,24E-09	0,579669277	0,142829819	5,81E-31	1,00261E-08	1,57193E-05
Compression:Nofm:deactivated_electrodes	0,17761	0,721027	6,16E-11	0,244149708	0,055029381	0,132402	0,547460592	0,484541482
Input_level:Nofm:deactivated_electrodes	0,528946	0,865947	1,32E-08	0,837949504	0,986412772	0,430572	0,092154239	0,097438972
T_60:Nofm:deactivated_electrodes	0,111009	0,652014	2,56E-08	0,50928222	0,831102205	0,090792	0,118759142	0,074899972
Compression:Input_level:T_60:Nofm	0,650422	0,373157	0,374923	0,083842129	0,329938734	0,259957	0,994111927	0,156660104
Compression:Input_level:T_60:deactivated_electrodes	0,921189	0,394875	1,71E-09	0,009631876	0,063265579	1,9E-13	3,78529E-05	2,50959E-09
Compression:Input_level:Nofm:deactivated_electrodes	0,00209	0,989656	6,52E-05	0,3378517	0,438952415	0,349567	0,344138221	0,000535098
Compression:T_60:Nofm:deactivated_electrodes	0,883204	0,982273	0,999816	0,404197251	0,555514773	0,040462	0,197201413	0,277272116
Input_level:T_60:Nofm:deactivated_electrodes	0,466537	0,80942	3,48E-09	0,165042763	0,883344742	0,117265	0,426429317	0,487485584
Compression:Input_level:T_60:Nofm:deactivated_electrodes	0,329773	0,088698	0,337752	0,059753245	0,836949553	0,320655	0,443935315	0,000617657

Figure 17: P-value table of the 5-way repeated measures ANOVA of the correlation input/spiral scores

simpler pattern compared to input/spiral STOI scores. For example, Figure 18 (b) aligns with our expectations: higher NofM, less reverberation, and fewer deactivated electrodes result in better scores. We also note that after 12 electrodes are deactivated, NofM makes no further difference. Additionally, the amplitude of change for NofM is less than that for T60, and T60's effect is less significant than the deactivation of electrodes.

In comparison, the input/spiral Figure 18 (a) appears flatter regarding the number of electrodes, which could be due to interactions between channels. For input/spiral scores, no significant interactions were found within electrodes/NofM/T60, so we decided not to make comparisons.

Regarding F0 input-sine scores, Figure 19 (b), we do not observe a clear degradation of the correlation score until 17 electrodes are deactivated. It even seems that deactivating 12 electrodes might increase the score. We expected to see a clear interaction between input level and compression, with low input levels and quiet compression providing higher scores compared to higher input levels and louder

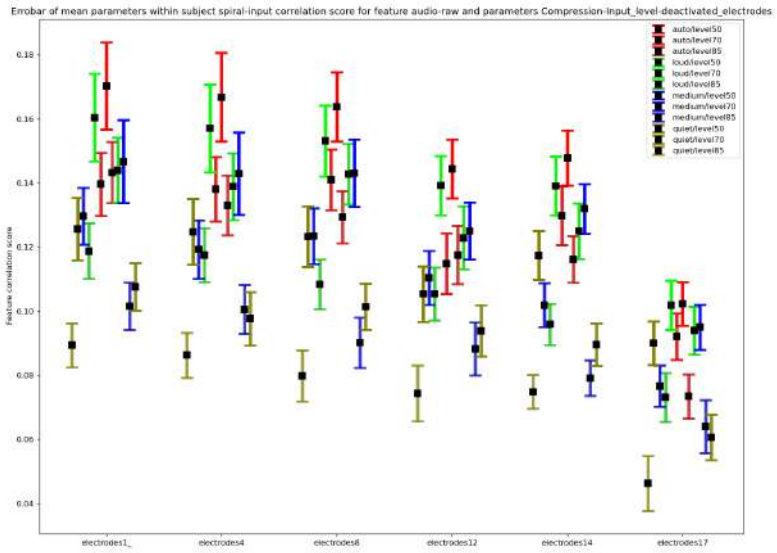
compression, and that the "auto" setting would yield the best results for feature correlation scores. These results might be explained by the fact that F0 is generally independent of input level or compression. Additionally, the alarm signals used had variable amplitude and thus variable input levels, which could introduce variability into the results.

Finally, for Figure 19 (a), we observe a significant decrease in score with 17 deactivated electrodes and a flattening of scores for other numbers of electrodes. The best correlation values changed from "auto" to "medium" and "loud" values depending on the number of electrodes, which is surprising but might result from data variability (note that we conducted a repeated measures ANOVA on 20 sounds across 432 configurations, so more varied sounds might yield stronger results).

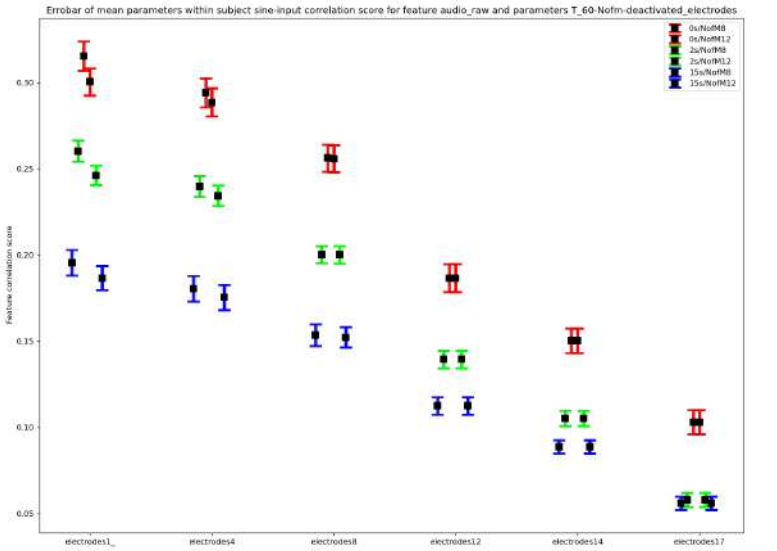
3.3 Analysis of SRS with Highly Masking Noise

The purpose of this part of the work was to analyze the behavior of CI audio transduction when noise is added. Due to time constraints, we focused our study on alarm sounds with added masking noise shaped according to the alarm spectrogram. This scenario represents a "worst-case" situation concerning environmental noise, as the noise is designed to effectively mask the alarm signal.

The shaping of the noise involves estimating the average spectrum of the input sounds over 30 ms frames, resulting in a list of spectra, each corresponding to a specific time frame. The masked noise is then constructed by filtering white noise with the estimated spectrum.

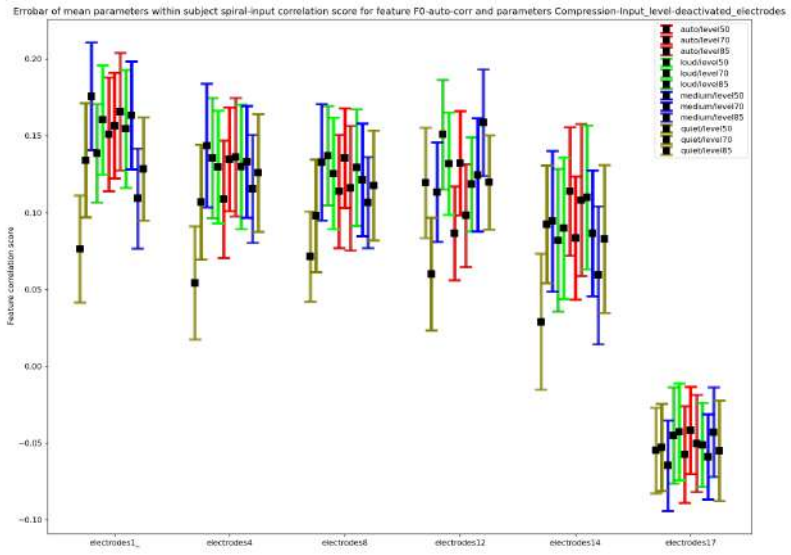


(a) Error bars of the mean for parameters compression:input-level-electrodes. The dependent variable is the spiral/input STOI score.

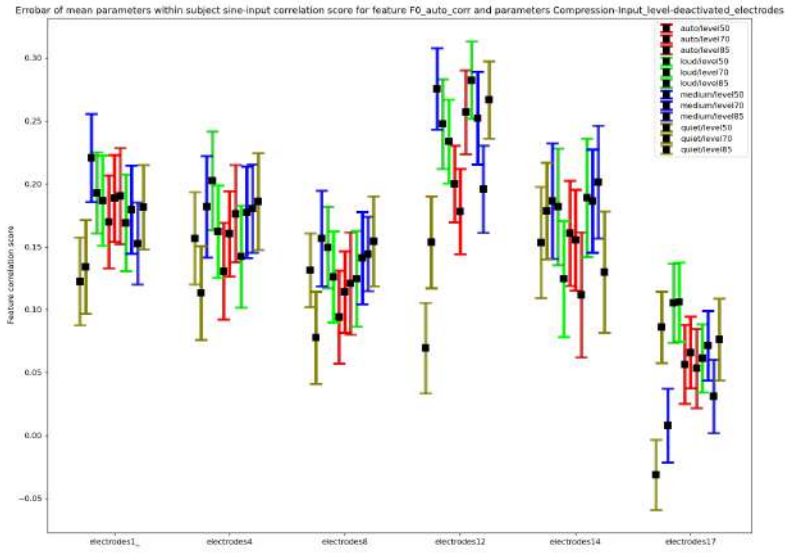


(b) Error bars of the mean for parameters T60:NofM:electrodes. The dependent variable is the sine/input STOI score.

Figure 18: Error bars of STOI scores for input-sine and input-spiral vocoders

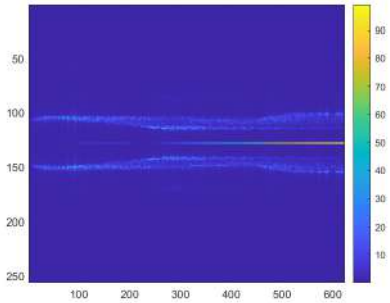


(a) Error bars of the mean for parameters compression:input-level-electrodes. The dependent variable is the spiral/input F0 correlation score.

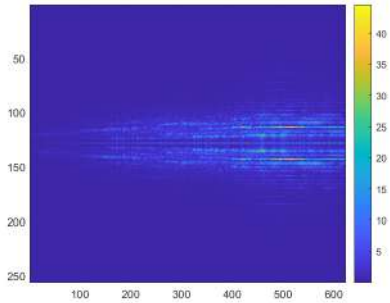


(b) Error bars of the mean for parameters compression:input-level-electrodes. The dependent variable is the sine/input F0 correlation score.

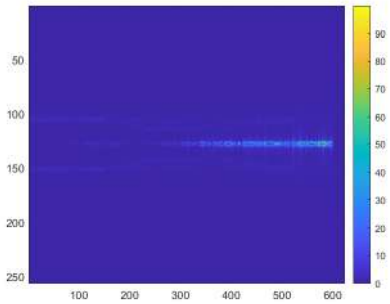
Figure 19: Error bars of F0 correlation scores for the input-sine and input-spiral vocoders



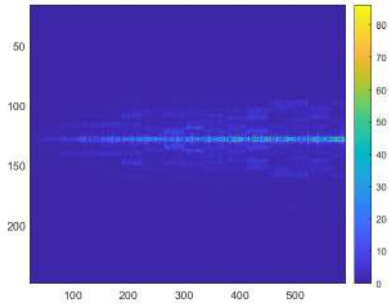
(a) First signal spectrogram example



(b) Second signal spectrogram example



(c) Masking noise spectrogram of signal 1



(d) Masking noise spectrogram of signal 2

Figure 20: Examples of signals and correspondings masking noise spectrogram, Yaxis=frequency bins,Xaxis=time frame

4 Proposed solution for improvements of the SRS perception

This part of the report cannot be described because the solution I proposed at the moment of writing is discussed to be part of a patent. (Note here that if the patent lawyer decides to drop the patent, then the description of this process will be add to this report as an appendix).

$$-\backslash\_(")/\_/-$$

5 Clinical Protocol for Testing the Proposed Solutions

The solution mentioned earlier is intended to be implemented within a coding strategy. We will refer to the classical Oticon Medical coding strategy as the "Crystalis" strategy and to this strategy with the proposed solution implemented as the "Crystalis Enhanced" strategy.

The experiments have three main objectives:

- Estimating the SNR alarm detection threshold to evaluate the effect of the Crystalis Enhanced solution.
- Estimating the recognition of alarms to evaluate the effect of the Crystalis Enhanced solution.
- Verifying our model, i.e., whether there is a monotonic mapping between input sound features and the corresponding CI vocoded features correlation scores, and a audibility perceptual score for alarm sounds for CI users, or at least for normal hearing individuals.

To achieve these objectives, we propose three experiments.

5.1 Exclusion and Inclusion Criteria

Two groups of participants should be formed:

- A control group consisting of individuals with normal hearing (NH), with hearing levels up to 25 dB HL for frequencies between 100 and 8000 Hz.
- A second group consisting of cochlear implant users (with unilateral or bilateral impairment) who have the classical Crystalis coding strategy implemented (implying that they use an Oticon Medical device).

Exclusion criteria for the normal hearing group: psycho/sensory impairments (e.g., neuro-sensory loss, cognitive impairment).

Exclusion criteria for the CI user group: any psycho/sensory impairments (e.g., neuro-sensory loss, cognitive impairment) other than hearing loss, or if the CI is not from Oticon Medical.

Additionally, subjects should have at least 1 year of implantation with a minimum of 6 months of cumulative use of their device (to ensure they are already adapted to their devices).

The following parameters concerning subjects must be recorded and will be used for analysis: duration of deafness, pre- or post-lingual hearing loss, age of the subject, unilateral or bilateral hearing loss/implantation, duration of implantation, duration of use, and type of hearing loss (etiology).

5.2 Protocol Proposed

5.2.1 Experiment 1

The purpose of this experiment is to measure the Signal-to-Noise Ratio (SNR) threshold for detecting alarm sounds. This experiment will be conducted in several blocks within two different groups (refer to Figure 22):

- A control group consisting of normal hearing subjects.
- A CI user group with Oticon Medical implants.

This test is a 3-Alternative Forced Choice (3AFC) test with a one-up/one-down adaptive procedure, where the variable parameter is the SNR, coupled with a 2×2 crossover design. The procedures for the NH group and the CI group during a session will be slightly different.

CI Test Session:

- The target signal is an alarm sound with an environmental composition of sounds (e.g., babble noise, street noises) at a controlled SNR (signal being the alarm sound).
- The reference signal is a target alarm sound with a modification (e.g., music instrument morphing, spectral symmetric inverting according to a constant frequency axis) and the same environmental composition of sounds at the same controlled SNR.

Other target and reference signals (referred to as dummy target and reference) may also be used in the testing session to avoid the subjects focusing specifically on the alarm. However, subjects' responses to these will not be considered during the adaptive one-up/one-down procedure (see Figure 21).

At each presentation, two reference signals and one target signal are presented in random order to the subject, who must differentiate the target from the references. A session will contain a maximum of 10 trials (see Figure 1). A trial ends after either 30 presentations or 5 consecutive reversals between two values. These trials should include at least 10 presentations of dummy target/reference signals presented in random order.

The SNR threshold for the trial is the mean of the last 5 reversal values.

NH Test Session:

The target and reference signals are defined similarly to the CI test session. However, after construction, the reference and target pairs will be vocoded using either the Crystalis strategy or the Crystalis Enhanced strategy (depending on the session and subgroups) and processed with the sinusoidal and spiral vocoders (for all testing sessions and subgroups) before being presented to the subjects (see Figure 22).

The same procedure will be applied to dummy target/reference pairs.

During a session, each trial will focus on one type of sound production (sine vocoded, spiral vocoded, or unprocessed) presented in random order. The SNR threshold estimation will then correspond to the SNR threshold for normal hearing subjects using vocoder X or no vocoder with coding strategy Y. Thus, each session will estimate 3 SNR thresholds: the SNR for the no vocoder coding strategy, sine vocoder, and spiral vocoder.

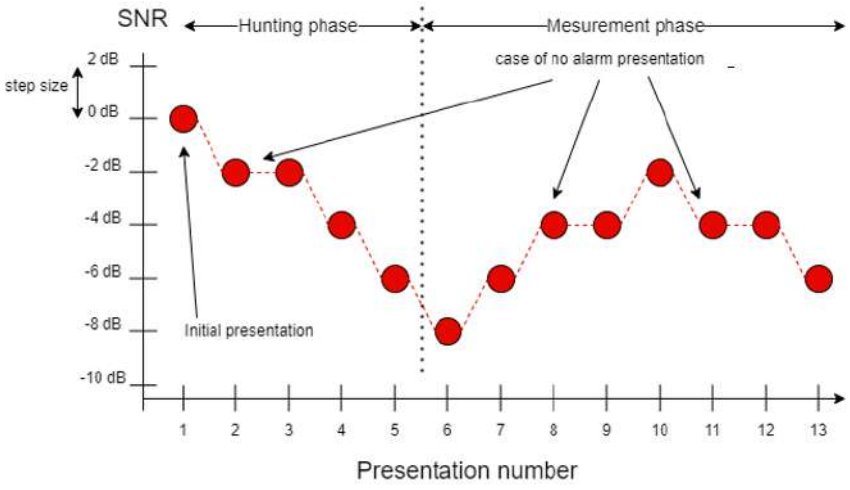


Figure 21: Example of a one-up/one-down trial

Note that the training phases for CI users will involve using their CI with the specified coding strategy for the required time before participating in the sessions.

- **Red:** Sessions testing unprocessed, Crystalis sinus vocoded, and Crystalis spiral vocoded signals.

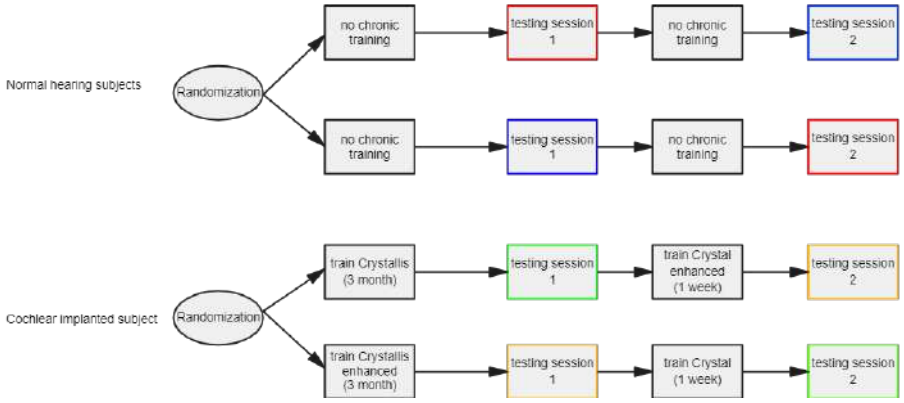


Figure 22: Session planning of Experiment 1

- **Blue:** Sessions testing unprocessed, Crystalis Enhanced sinus vocoded, and Crystalis Enhanced spiral vocoded signals.
- **Green:** Sessions for CI users under the Crystalis strategy.
- **Yellow/Orange:** Sessions for CI users under the Crystalis Enhanced strategy.

Analysis of Results and Hypotheses:

The SNR threshold for a trial will be estimated using the mean of the last 5 reversal SNR values. The mean of the estimated thresholds across trials and subjects will represent the SNR threshold for the session and subject. The result is validated only if the standard deviation (STD) estimated from at least 4 trials is less than or equal to the step size (2 dB). Some biases can be mitigated by the design, as discussed in [17].

After averaging the scores across trials and subjects, we obtain 7 SNR threshold estimates:

- NH threshold
- NH Crystalis sine vocoded threshold (NHC-sin-threshold)
- NH Crystalis Enhanced sine vocoded threshold (NHCe-sin-threshold)
- NH Crystalis spiral vocoded threshold (NHC-spi-threshold)
- NH Crystalis Enhanced spiral vocoded threshold (NHCe-spi-threshold)
- CI Crystalis threshold (CIC-threshold)
- CI Crystalis Enhanced threshold (CICe-threshold)

We expect the SNR thresholds to be in the following order:
NH threshold < *NHC-sin-threshold* < *NHCe-sin-threshold* < *NHC-spi-threshold* < *NHCe-spi-threshold* < *CICe-threshold* < *CIC-threshold*

We naturally expect that NH subjects will perform better with unprocessed sound compared to degraded sounds and will outperform CI subjects. Additionally, we anticipate that the Crystalis Enhanced design might disrupt normal hearing listeners (although the specific reasons for this will be revealed through the solution), but not to the extent that it reverses the performance benefit of the sinus vocoder compared to the spiral vocoder.

We will also compare the SNR thresholds from the green sessions and the yellow sessions to assess the impact of short adaptation times for both coding strategies. We expect the results from CI testing sessions under Crystalis Enhanced to be lower than those from sessions under the Crystalis strategy.

Finally, we anticipate that the enhanced version of the coding strategy will improve detection capabilities for CI users compared to those with the classical coding strategy.

5.2.2 Experiment 2

The second experiment, involving the same groups, is a recognition test aimed at estimating alarm recognition. Specifically, it is a 42-item computerized forced-choice, closed-set test. This test is designed to assess recognition scores based on the hearing condition of the participants.

The test should be conducted at least 2 months after the first experiment to avoid carryover effects. Additionally, the CI group should be randomly split into two subgroups: one with the Crystalis strategy and another with the Crystalis Enhanced strategy.

Participants will be asked to rate their familiarity with a set of sounds and assess the security relevance of these sounds, using a scale from 1 to 5 (where 5 represents the highest confidence or greatest security relevance, and 1 represents the lowest confidence or security relevance). They will also receive training on recognizing the sounds used in the Luzum 2023 protocol (refer to the Materials and Methods section of [19]).

We will specifically examine the recognition rate of alarm signals and compare the results for Crystalis Enhanced, Crystalis, and NH (Normal Hearing) scores. We expect the recognition score for Crystalis Enhanced to be higher than that for Crystalis but lower than the Normal Hearing score.

5.2.3 Experiment 3

The purpose of this experiment is to establish a mapping between the perceptual alarm audibility score of CI users and the correlation score of certain acoustic features of the original sounds with a vocoded version of the same sound.

The features considered are brightness, pitch, harmonicity, scale, loudness, flatness, and irregularity (defined according to Elhilali's procedure [14]).

The processing of alarm sounds will involve modifying one of these features while attempting to keep the others as constant as possible. This will result in alarm signals with controlled correlation scores for the feature in question.

Ideally, we aim to create a table for each sound and each feature as follows:

Correlation Score of Feature F with Its Modified Version	Alarm Sounds Created
1	Original Alarm
0.8	Alarm with Feature F Score 0.8
0.6	Alarm with Feature F Score 0.6
0.4	Alarm with Feature F Score 0.4
0.2	Alarm with Feature F Score 0.2

Table 1: Example of table necessary for the experiment

Note that the method for obtaining sounds with modified features will not be explicitly detailed here. Remarks about the creation of these sounds will be provided in the limitations section.

Method

The same groups are retained (i.e., a normal hearing control group and a CI group with the Crystalis strategy). The CI group is divided and tested using the same 2×2 crossover design as shown in Figure 22.

The test is a 2-Alternative Forced Choice (2AFC) test, where the target signal includes an environmental background sound and an alarm sound, while the reference signal contains only the environmental background sound.

Before the test, subjects will be trained to recognize the unprocessed alarms that will be used. During the testing session, target and reference signals will be presented in a random order, and subjects will be asked to indicate whether the alarm was clearly audible. The target signals will include both unprocessed and processed alarms, presented in random order.

Analysis of Results

We will compute the mean of the rate of correct positive responses across subjects and alarms for each modality (i.e., feature correlation score). We expect to obtain scatter plots similar to Figure 23.

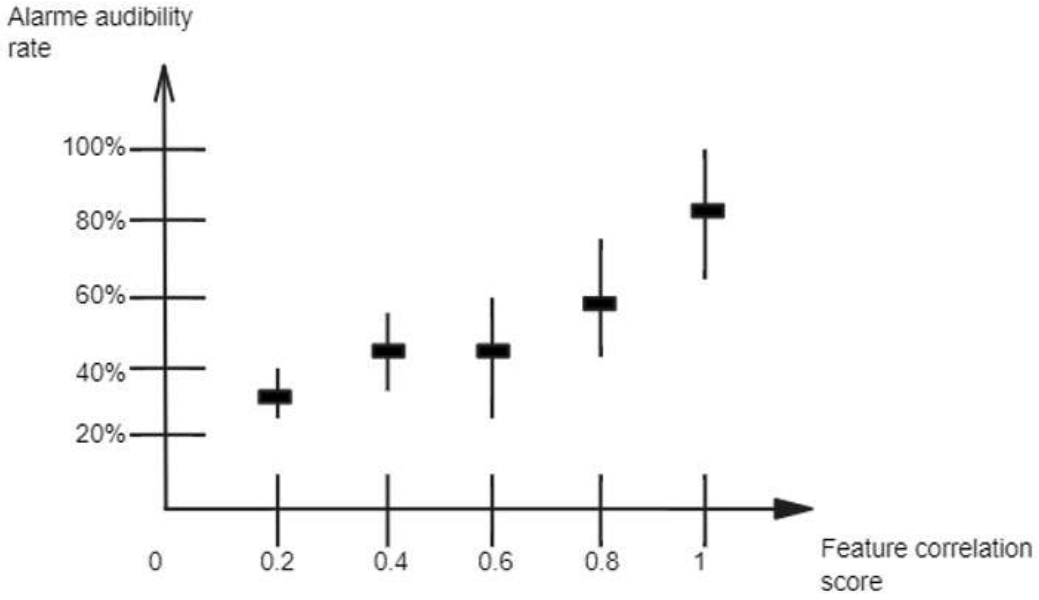


Figure 23: Example of expected mapping between the feature correlation score and the alarm audibility rates

We will use the matching procedure outlined in [28]. Other mapping functions, such as polynomial, exponential, or sigmoid functions, are also considered. Possible expressions for modeling the perceptual score based on the feature correlation include:

$$\text{perceptual score}(\text{corr score}) = a \cdot (\text{corr score})^3 + b \cdot (\text{corr score})^2 + c \cdot (\text{corr score}) + d$$

$$\text{perceptual score}(\text{corr score}) = \frac{1}{1+\exp[-(a \cdot \text{corr score} + b)]}$$

$$\text{perceptual score}(\text{corr score}) = a \cdot \text{corr score} + b$$

where a, b, c, and d are learnable parameters, optimized using a method such as gradient descent.

5.3 Limitations of Experiments

We can classify the limitations of these experiments into the following categories: participant profile, experimental setup, material (hardware and software), and balance and representativity.

5.3.1 Participant Profile

Participant variability may influence their adaptation capabilities, notably during the adaptation period of the first experiment (in terms of time required and performance achieved). To preserve participant focus, the experiment should be limited in duration. CI users unfortunately have heterogeneous characteristics, such as varying durations of deafness and hearing acuity, which we cannot control and which will likely affect the variability of our results.

5.3.2 Experimental Setup

- Reproducing the same hearing conditions across experiments can be very challenging. It is particularly critical that the sound stimuli produced in the experiment have consistent characteristics (e.g., loudness, pitch), which cannot be guaranteed without a controlled experimental environment. Using a crossover design for Experiment 1 may introduce undesirable effects, such as carryover or sequence effects [17].

- In Experiment 3, the CI might affect the acoustic features of their processed sounds. Therefore, we cannot precisely assess the effect of modifying alarm sound features on CI users' perception (the best approximation would be to use the vocoders again).

- For Experiment 3, there is not a single method to modify a signal x into another signal y such that corr-score$(x, y) = a$ for a determined value. It is reasonable to assume that among all y candidates where corr-score$(x, y) =$ constant, some will be perceived differently by CI users. Thus, the specific way CIs modify features in relation to correlation requires further study.

5.3.3 Material (Hardware and Software)

- The equipment must be as controlled as possible to ensure the repeatability of the experiments and the randomization of stimulus presentation to avoid sequential bias.

- Modifying only one acoustic feature for alarms may not be feasible for spectral-related features without affecting potentially important spectral cues. For example, modifying the pitch will also impact the brightness (spectral centroid) because changing the pitch shifts the entire spectrum.

- The modifications applied to the features will depend on the input alarm sounds, especially if the flattening of feature variations is insufficient to achieve the targeted correlation score.

5.3.4 Balance and Representativity

A small number of subjects poses a problem for representativity, making it challenging to obtain a balanced sample of every characteristic of interest (such as age, duration of CI use, and pre- or post-lingual hearing loss). A preliminary sample analysis based on the expected effect size must be conducted.

6 Conclusion and Discussion

6.1 Conclusion

In this work, we created a database of SRS and environmental sounds and developed a framework for processing these sounds, including adding reverberation and controlling the SNR. These tools have not been used to their full potential due to time and computational constraints but should be capable of providing controlled signals for future experiments. We support two claims about the reasons for the poor alarm detection capabilities of CI users. We proposed a design to improve both the detection and recognition capacities of CI users, although we cannot discuss it in detail. Finally, we provided clinical guidelines to test the efficacy of the solution in both recognition and detection tasks and to validate our hypothesis regarding the link between feature correlation scores and alarm audibility rates.

6.2 Future Work

Our approach has several areas for improvement in future experiments:

- We focused only on alarm signals, which, although important, do not necessarily represent all SRS signals. For example, alarms are harmonic signals with periodic patterns, unlike broadband noises such as vehicle noise. Additionally, for other SRS signals, the pitch feature may be difficult or even impossible to define, making it challenging to generalize our analysis to all SRS sounds.
- We used STOI as a measure of "global degradation," relying on the assumption that STOI correlates with alarm recognition rates. While STOI is highly correlated with speech intelligibility, its effectiveness as a predictor for SRS recognition or detection rates is not well established.
- We used a measure of "CI loudness" that is contentious due to questionable assumptions. For instance, integrating contributions from different channels into the loudness calculation is problematic, and adding channel contributions alone is insufficient. This measure also does not account for channel interactions caused by the current spread issue.

- Our analysis focused solely on correlation scores (i.e., the variation of feature values compared to the output). We observed that features depending on signals can vary significantly over short durations. Therefore, averaging values or calculating standard deviations could be misleading, as significant variations might be important for understanding CI effects but would be obscured by such statistical tools. However, mean value differences can be important for assessing audibility, as shown in [11], which compares the mean value of background + alarm with background only during the alarm period. We believe that correlation may still be a good predictor of audibility, as humans can recognize sequences of sounds based on temporal-frequency patterns, such as recognizing a melody even if transposed or played by a different instrument at a different tempo.

- Another concern is that our analysis was conducted on a limited number of different signals. This limitation is primarily due to software constraints, as calculating scale features and processing chains were demanding. We prioritized the number of parameters over the number of files, as adding more alarms would have introduced less variability than changes in parameters.

- Additionally, using the vocoder involves making assumptions about the EDG, specifically regarding frequency perception. Determining how well these models represent CI users' sound perception is a complex question with no universally correct answer.

7 Annex

7.1 Some other figures

We studied other features since the beginning of this work, notably because in other paper less related to our topics [5], [6] cite those features are best features to make environmental sound classification. [11] also mention roughness as a feature valuable of investigation.

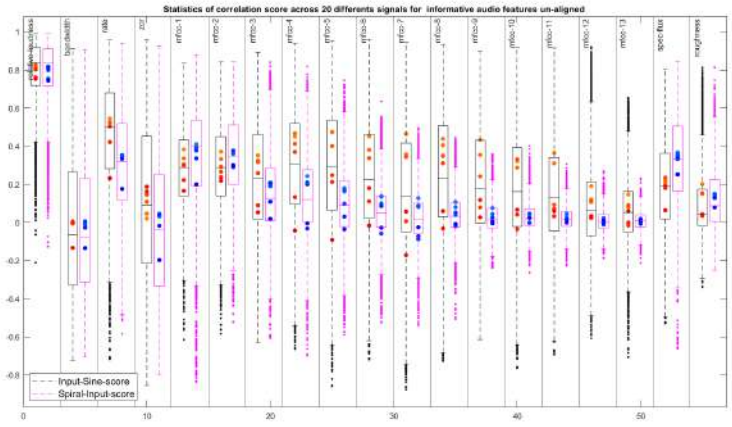


Figure 24: Statistic of correlation score for other acoustic features. On Y-axis you have the correlation score, on X-axis the different categories tested (feature/vocoder type)

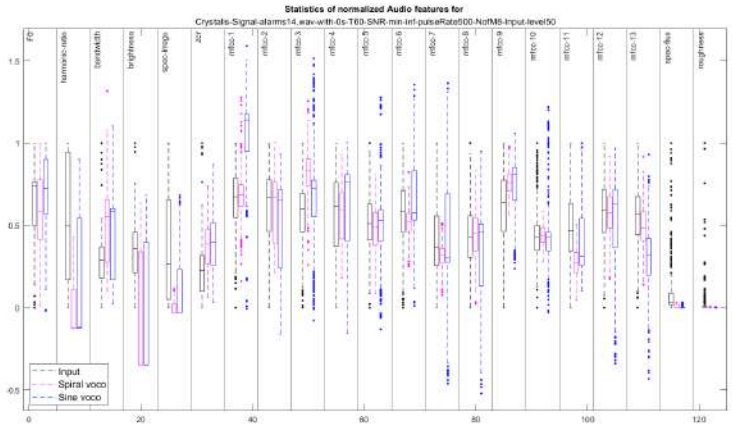


Figure 25: Example of plotting the statistic across the time of feature values for the input and the vocoders on one modality, the values are normalized to ensure that the f0 is between 0 and 1. On Yaxis, you see the means features value of a signal under a specific modality normalize for all feature to ensure the f0 values are between 0 and 1 (those values are no more interpretable by themselves). The X axis separate the different outputs (unprocessed, sine or spiral vocoded). A way to interpret this graph it is that if the distribution of the vocoded values are very different from the output, then the feature is degraded.

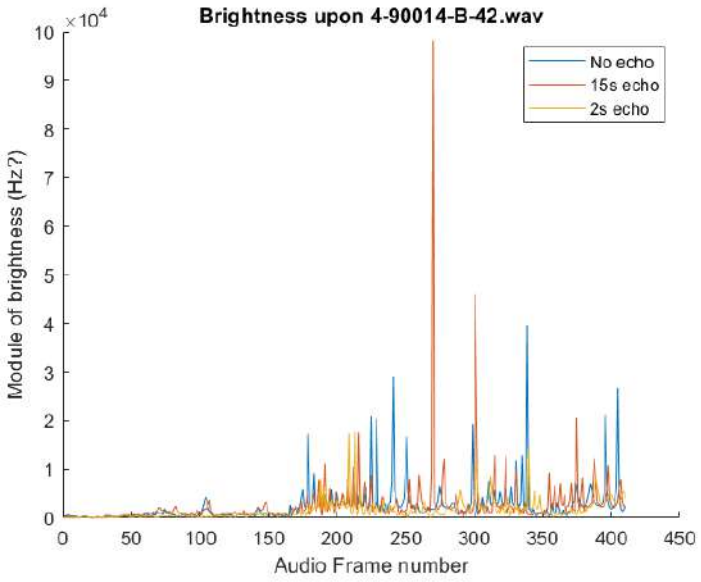


Figure 26: Brightness estimation of signal 490014-B-42.wav, we studied those kind of figure to ensure that the parameters had effects on feature and were correctly set up, note here that brightness is supposed to be a frequential centroid value for a signal at 16kHz sample rate, so the value superior to 8kHz are a bit odd

Figure 27: Example of former EDG used to our analysis, (the electrodes being numerated from apex to base of the cochlea, first number correspond to low frequency and last to high ones, so it's upside done compared to spectrogram classic plots, thanks to Dr.Rafael Attili Chiea code, we could superposed EDG and spectrogram.

References

[1] Andrey Anikin, Rasmus Bååth, and Tomas Persson. "Human Non-linguistic Vocal Repertoire: Call Types and Their Meaning." In: *Journal of nonverbal behavior* 42 (1 2018), pp. 53–80. ISSN: 0191-5886. DOI: 10.1007/s10919-017-0267-y.

[2] Justin M. Aronoff and David M. Landsberger. "The development of a modified spectral ripple test". In: *The Journal of the Acoustical Society of America* 134 (2 Aug. 2013), EL217–EL222. ISSN: 0001-4966. DOI: 10.1121/1.4813802.

[3] Sebastian A. Ausili et al. "Sound Localization in Real-Time Vocoded Cochlear-Implant Simulations With Normal-Hearing Listeners". In: *Trends in Hearing* 23 (2019). ISSN: 23312165. DOI: 10.1177/2331216519847332.

[4] Pascal Belin, Sarah Fillion-Bilodeau, and Frédéric Gosselin. "The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing". In: *Behavior Research Methods* 40 (2008), pp. 531–539. URL: https://api.semanticscholar.org/CorpusID:1463309.

[5] Vasileios Bountourakis, Lazaros Vrysis, and George Papanikolaou. "Machine Learning Algorithms for Environmental Sound Recognition: Towards Soundscape Semantics". In: *Proceedings of the Audio Mostly 2015 on Interaction With Sound*. AM '15. Thessaloniki, Greece: Association for Computing Machinery, 2015. ISBN: 9781450338967. DOI: 10.1145/2814895.2814905. URL: https://doi.org/10.1145/2814895.2814905.

[6] Vasileios Bountourakis et al. "An Enhanced Temporal Feature Integration Method for Environmental Sound Recognition". In: *Acoustics* 1.2 (2019), pp. 410–422. ISSN: 2624-599X. DOI: 10.3390/acoustics1020023. URL: https://www.mdpi.com/2624-599X/1/2/23.

[7] Robert P. Carlyon and Tobias Goehring. *Cochlear Implant Research and Development in the Twenty-first Century: A Critical Update*. Oct. 2021. DOI: 10.1007/s10162-021-00811-5.

[8] Lorenzi Christian et al. "Human Auditory Ecology: Extending Hearing Research to the Perception of Natural Soundscapes by Humans in Rapidly Changing Environments". In: *Trends in Hearing* 27 (Jan. 2023). ISSN: 23312165. DOI: 10.1177/23312165231212032.

[9] J. S. Chung, A. Nagrani, and A. Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *INTERSPEECH*. 2018.

[10] Mishaela DiNino et al. "Vowel and consonant confusions from spectrally manipulated stimuli designed to simulate poor cochlear implant electrode-neuron interfaces". In: *The Journal of the Acoustical Society of America* 140 (6 Dec. 2016), pp. 4404–4418. ISSN: 0001-4966. DOI: 10.1121/1.4971420.

[11] F. Effa et al. "Evaluating and predicting the audibility of acoustic alarms in the workplace using experimental methods and deep learning". In: *Applied Acoustics* 219 (Mar. 2024). ISSN: 1872910X. DOI: 10.1016/j.apacoust.2024.109955.

[12] Tobias Goehring et al. "The effect of increased channel interaction on speech perception with cochlear implants". In: *Scientific Reports* 11 (1 Dec. 2021). ISSN: 20452322. DOI: 10.1038/s41598-021-89932-8.

[13] Jacques A. Grange et al. "Cochlear implant simulator with independent representation of the full spiral ganglion". In: *The Journal of the Acoustical Society of America* 142 (5 Nov. 2017), EL484–EL489. ISSN: 0001-4966. DOI: 10.1121/1.5009602.

[14] Nicholas Huang and Mounya Elhilali. "Auditory salience using natural soundscapes". In: *The Journal of the Acoustical Society of America* 141.3 (Mar. 2017), pp. 2163–2176. ISSN: 0001-4966. DOI: 10.1121/1.4979055. eprint: https://pubs.aip.org/asa/jasa/article-pdf/141/3/2163/15325788/2163_1_online.pdf. URL: https://doi.org/10.1121/1.4979055.

[15] David M. Landsberger et al. "Spectral-Temporal Modulated Ripple Discrimination by Children With Cochlear Implants". In: *Ear and hearing* 39 (1 Jan. 2018), pp. 60–68. ISSN: 15384667. DOI: 10.1097/AUD.0000000000000463.

[16] Birgitta Larsby and Stig Arlinger. "Auditory Temporal and Spectral Resolution in Normal and Impaired Hearing". In: *Journal of the American Academy of Audiology* 10 (04 Apr. 1999), pp. 198–210. ISSN: 1050-0545. DOI: 10.1055/s-0042-1748481.

[17] In Junyong Lim Chi-Yeon. "Considerations for crossover design in clinical study". In: *Korean J Anesthesiol* 74.4 (2021), pp. 293–299. DOI: 10.4097/kja.21165. eprint: http://ekja.org/journal/view.php?number=8753. URL: http://ekja.org/journal/view.php?number=8753.

[18] Philipos C. Loizou. "Mimicking the human ear". In: *IEEE Signal Processing Magazine* (Sept. 1998), pp. 101–130. DOI: 10.1109/79.708543.

[19] Nathan R. Luzum et al. "Identification Accuracy of Safety-Relevant Environmental Sounds in Adult Cochlear Implant Users". In: *Laryngoscope* 133 (9 Sept. 2023), pp. 2388–2393. ISSN: 15314995. DOI: 10.1002/lary.30475.

[20] Olivier Macherey and Robert P. Carlyon. *Cochlear implants*. Sept. 2014. DOI: 10.1016/j.cub.2014.06.053.

[21] Loureiro Manuel et al. "Datalogging Findings in Adult Cochlear Implant Recipients Who Never Developed Intelligible Speech". In: *J Int Adv Otol* 20 (2024), pp. 113–118. DOI: 10.5152/iao.2024.231193. URL: www.advancedotology.org.

[22] Stephen G. McGovern. "Fast image method for impulse response calculations of box-shaped rooms". In: *Applied Acoustics* 70 (1 Jan. 2009), pp. 182–189. ISSN: 0003682X. DOI: 10.1016/j.apacoust.2008.02.003.

[23] Nicole Miller-Viacava et al. "Sensorineural hearing loss alters auditory discrimination of natural soundscapes". In: *International Journal of Audiology* (2023). ISSN: 17088186. DOI: 10.1080/14992027.2023.2272559.

[24] vryzas nikolaos et al. "speech emotion recognition for performance interaction". In: *journal of the audio engineering society* 66 (6 June 2018), pp. 457–467. DOI: 10.17743/jaes.2018.0036.

[25] Fearghal O'Brien and Denis Cousineau. "Representing Error bars in within-subject designs in typical software packages". In: *The Quantitative Methods for Psychology* 10.1 (2014), pp. 56–67. DOI: 10.20982/tqmp.10.1.p056. URL: http://www.tqmp.org/RegularArticles/vol10-1/p056/p056.pdf.

[26] Karol J. Piczak. "ESC: Dataset for Environmental Sound Classification". In: *Proceedings of the 23rd Annual ACM Conference on Multimedia*. Brisbane, Australia: ACM Press, Oct. 13, 2015, pp. 1015–1018. ISBN: 978-1-4503-3459-4. DOI: 10.1145/2733373.2806390. URL: http://dl.acm.org/citation.cfm?doid=2733373.2806390.

[27] William Shakespear. *Hamlet*. 1623.

[28] Cees H. Taal et al. "An algorithm for intelligibility prediction of time-frequency weighted noisy speech". In: *IEEE Transactions on Audio, Speech and Language Processing* 19 (7 2011), pp. 2125–2136. ISSN: 15587916. DOI: 10.1109/TASL.2011.2114881.

[29] Nikolaos Vryzas et al. "Subjective Evaluation of a Speech Emotion Recognition Interaction Framework". In: Sept. 2018, pp. 1–7. DOI: 10.1145/3243274.3243294.