Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Method for AI-Assisted Human Interpretation of Biological Signals: Analysis of Neonatal EEG

Sergi Gomez-Quintana (sgomez@umail.ucc.ie)
University College Cork
Alison O'Shea
Munster Technological University
Andreea Factor
University College Cork
Emanuel Popovici
University College Cork
Andriy Temko
University College Cork

Research Article

Keywords: Computational methods, signal processing, medical data, analysis, visualization, sonification, electroencephalogram, neonates, seizures, deep learning, explainable artificial intelligence

Posted Date: February 4th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1232994/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

A method for AI-assisted human interpretation of biological signals: analysis of neonatal EEG

Sergi Gomez-Quintana[1*], Alison O'Shea[2], Andreea Factor[3], Emanuel Popovici[1], Andriy Temko[1]
[1] Electrical and Electronic Engineering, University College Cork, Ireland
[2] Department of Computer Science, Munster Technological University, Ireland
[3] Department of Anatomy and Neuroscience, University College Cork, Ireland
sgomez@umail.ucc.ie

Abstract—

The study proposes a novel method to enable healthcare professionals interact and leverage AI decision support in an intuitive manner using auditory senses. The method suitability is assessed through acoustic detection of the presence of neonatal seizures in electroencephalography (EEG). Neurophysiologists use EEG recordings to visually identify seizures. However,neurophysiological expertise is expensive and not available 24/7, even in tertiary hospitals. Nurses, neonatologists, paediatricians can make frequent misdiagnosis when interpreting complex EEG signals. While artificialintelligence (AI) has been widely used to provide objective decision support for EEG analysis, AI decisionsare not always explainable. A solution is developed in this study to combine AI algorithms with a human-centric intuitive EEGinterpretation method. Specifically, EEG is converted to sound using an AI-driven attention mechanism. Using this method perceptual characteristics of seizure events can be heard and an hour of EEG can beanalysed in five seconds. A survey that has been conducted among targeted end-users on a publicly available dataset has demonstratedthat not only does it drastically reduce the burden of reviewing the EEG data, but also the obtained accuracy is on par with experiencedneurophysiologists trained to interpret neonatal EEG. It is also shown that the proposed communion of a medical professional and AIoutperforms AI alone through empowering the human with little or no experience to leverage AI attention mechanisms to enhance theperceptual characteristics of seizure events.

Keywords— Computational methods, signal processing, medical data, analysis, visualization, sonification, electroencephalogram, neonates, seizures, deep learning, explainable artificial intelligence

I. INTRODUCTION

In 2019, 2.4 million neonatal deaths occurred globally [1]. Most neonatal deaths occur during the first week of life, and about 1 million newborns die within the first 24 hours. The vast majority occurred in developing countries, with hypoxic-ischemic encephalopathy (HIE) due to birth asphyxia being one of the major causes of high child mortality [2]. While the incidence of encephalopathy ranges from 1 to 8 per 1000 live births in developed countries, it is as high as 26 per 1000 live births in underdeveloped countries [3]. Neonatal seizures are associated with various acute illnesses such as strokes, HIE or infections [4]. The reported incidence of such events varies from 1.5 to 3.5 out of 1000 newborns [5-7]. The mortality rate on those affected by seizures is aproximately 10% (range: 7–16%), and permanent neurological disability is estimated around 50% of those that survive [8-11]. Most of those neonatal deaths could have been prevented with optimal care [12].

Detection of seizures is a challenging clinical task. While seizures are often associated with clinical signs in children and adults, including involuntary jerking movements of the arms and legs, eye blinkings, or difficulty breathing, less than 10% are accompanied by documentable physical manifestations in the neonatal population. Monitoring the electrical brain activity through electroencephalography (EEG) is the only way to detect seizures accurately [13, 14]. While EEG has become a gold standard tool in neonatal neurophysiology and most neonatal units in the developed world have access to EEG acquisition, only some have immediate availability of neurophysiological expertise to interpret the signal [15]. In addition, EEG is a highly complex signal, even more so for neonates where the brain is still developing [16]. While multichannel EEG monitoring can last several hours to several days, seizures are infrequent events. Therefore, a simpler representation of EEG is often used to facilitate the interpretation of long and complex EEG signals, such as amplitude-integrated EEG (aEEG). This method represents the signal as temporally smoothed and energy-compressed waveforms [17]. While aEEG allows observing several hours of EEG in a single screen page, seizures become difficult to detect visually due to limited spatial coverage, attenuation of short duration seizure due to smoothing, and false alarms are frequently caused by prolonging energy artefacts [18, 19].

Raw multichannel EEG monitoring is required to detect all seizures. The raw EEG signal interpretation requires extensive clinical expertise and years of training. Multichannel EEG, especially continuous multichannel EEG obtained from long real-life clinical monitoring, is often corrupted with all sorts of artefacts, some of which can resemble waveform patterns similar to seizures. When EEG recordings last for several days, detecting of relatively rare seizure events becomes a challenging and highly time-consuming process. Even among experienced EEG experts, there is still a substantial disagreement when detecting seizure events [20]. Moreover, the level of expertise of a healthcare professional required to identify seizures scales with the complexity of the data to interpret – clinical manifestations can be interpreted by nurses, aEEG interpretation requires neonatologists, whereas full EEG requires neurophysiologists. The higher the level of experience of a healthcare professional is, the smaller the chances of its availability onsite 24/7.

Artificial intelligence (AI) has become a popular tool to assist medical professionals at interpreting EEG signals. AI aims to close the gap between the availability of interpretation expertise and detection timeliness and accuracy, particularly at detection of seizure events. For years, automated seizure detection was tackled as a machine-learning (ML) problem in a well-established two-step approach consisting of (a) the summarization EEG data into a set of hand-crafted informative characteristics called features, and (b) the usage of classifiers that learn the mapping between features and the labels. The works undertaking this approach differ mainly on the selected features and the choice of the classifier [21-26]. Contrary to these approaches, deep learning (DL) is able to directly learn the representation of the relevant information from the raw EEG data avoiding time-consuming feature engineering efforts [27-31].

AI models are not error-free. They are often subject to adversarial attacks [32, 33]. In a clinical environment, errors can result in wrong decisions concerning a diagnosis or a treatment choice, threatening and putting the overall patient's health and well-being at risk. Therefore, it is of utmost importance to understand the nature of the errors and the availability of an explainable framework to trace back those errors to underlying clinical causes is of utmost importance [32]. Explainable AI and saliency detection are the research areas that aim to mitigate the risks. Medical professionals require the option to understand how and why a machine decision has been made. However, there is a clear trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions. For example, the feature-based neonatal seizure detection approaches could provide insight into what feature contributed most to the wrong decision [34]. However, the performance of such systems is significantly lower than that of deep learning methods [35]. The black-box models like neural networks tend to extract most of the information from data to beat the state-of-the-art performance without explaining and interpreting their decisions.

Alternative methods such as EEG sonification have been developed to detect neonatal seizures [36] to assure interpretability of the clinical decisions while still simplifying the process of analysis of complex EEG signals. Sonification is the process by which data are represented through sound in order to be displayed acoustically. Sonification has been proven useful in various contexts involving pattern recognition because such patterns are implicitly mapped into sound as distinguishable rhythms or tones [37]. For example, in medicine and healthcare, all sorts of biological signals are used to monitor the patient's health, and either the presence or absence of such patterns in the data are highly informative. An example of that is the auscultation of the heart sounds through a stethoscope: a simple and yet highly informative clinical examination routine that has been used since 1838 when the stethoscope was first invented [38].

Neonatal seizures in EEG present a specific characteristic evolving rhythmic pattern with temporal and often spatial evolution [39, 40]. Morphological changes in temporal signals are often more naturally perceived by the ear than by the naked eye. Historically, in early paper-based EEG monitoring machines, seizures could be heard as rhythmic and evolving pen movements. Most of the frequential content of neonatal EEG falls in the frequency range from 0.5 to 13Hz approximately [41], and the audible spectrum in humans ranges from 20Hz to 20kHz [42], meaning that EEG can not be directly heard in its original time-series representation. The phase vocoder (PV) was firstly used for EEG seizure detection by raising the frequency content of the EEG into the audible range without affecting the time scale [43]. It has been subsequently shown in [36] that the accuracy at detecting seizure events improves as sonified EEG is progressively sped up. Sonification can be used as a tool to get additional confirmation on the presence or absence of a particular pattern in the EEG signal.

While interpretability of the decision with sonified EEG is secured and the causes of each decision can be traced back directly to the signal characteristics, analysis of EEG through sonification can still be a time-consuming process. Compressing the signal too much results in an abundance of missed events, sharing the same disadvantage with aEEG.

In this work, a new method of neonatal EEG interpretation for seizure detection, which combines the accuracy of AI methods with interpretability and intuitiveness of sonification and which allows quick, accurate and interpretable decision making is proposed. The developed technique utilizes AI as an attention mechanism. In particular, the AI probabilities are used to modulate the speed-up factor of the EEG sonification, maintaining the focus on the EEG regions of interest (possible seizures) while allowing for a quick pass over long segments of background EEG. The main contributions of this study are:

- A novel method of AI-driven spatial neonatal EEG sonification is presented;
- An extensive survey is designed and conducted online among the targeted end-users (healthcare professionals) and lay-readers (non-healthcare professionals) to assess the level of accuracy of the developed method on a publicly available dataset of neonatal seizures;
- The survey results are compared to the performance obtained by the three neonatal neurophysiologists who annotated the dataset on raw multichannel EEG.

II. AI-DRIVEN SONIFICATION ALGORITHM

To make neonatal EEG analysis more accurate, pervasive and ubiquitous, reducing the gap between the complexity of the signal and the level of expertise needed for its analysis must be reduced while ensuring transparency in the interpretation of each decision is paramount. The method outlined in this section aims to facilitate quick and a

ccurate diagnosis by healthcare professionals without dedicated EEG interpretation expertise while keeping the doctor in the centre of the decision making process.

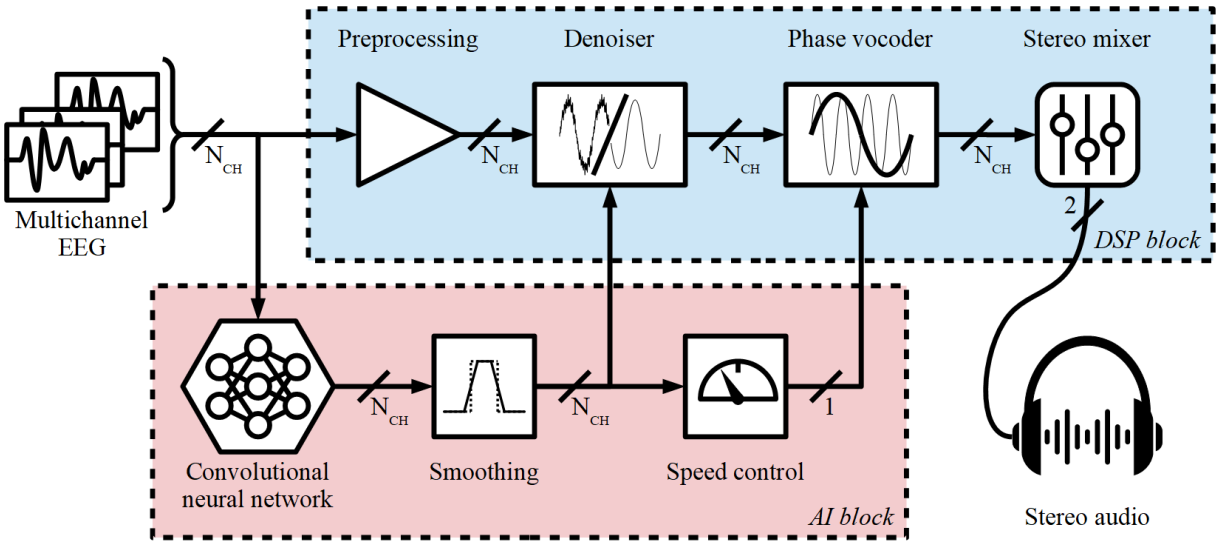

Figure 1. AI-driven sonification algorithm block diagram.

The outline of the proposed algorithm is shown in Figure 1. The algorithm takes multichannel neonatal EEG as input and delivers stereo audio that captures the content of a few hours of EEG data into a few seconds of sound. The algorithm is constituted by two fundamental blocks: the digital signal processing (DSP) block and the AI block. The DSP block converts the EEG data into audio. The EEG signals are pre-processed and denoised before the phase vocoder lifts the frequency content into an audible frequency range. While all those steps are applied in parallel to each EEG channel, the stereo mixer combines all the incoming audio tracks from each EEG channel into left-right stereo audio channels by mapping each track into a virtual 2D space to capture the relative positions of the electrodes. The AI block is used to modulate several aspects of the DSP block: it analyzes the segment of EEG data and every second provides the probability of a seizure event happening for each time segment of EEG. These probabilities control the variable time-compression rate as a function of the seizure likelihood, serving as an AI-driven attention mechanism that modulates the perceived audio accordingly.

The following subsections will detail the functionality of each subroutine.

1) DSP block: Preprocessing

The signal pre-processing stage applies a set of deterministic operations to adequate the EEG signals before it enters the sonification algorithm's subsequent stages, reducing the noise level outside the band of interest and EEG artifacts [44]. First, a combination of highpass and lowpass filters remove undesired components outside the frequency range between 0.5Hz and 16Hz. A notch filter removes the interfering 50Hz-AC component caused by the power supply of the acquisition system. Although this component is already considerably attenuated by the lowpass filter, in some cases, the 50Hz component may remain non-negligible and could potentially affect in terms of aliasing when downsampling the signal in further stages of the algorithm.

The AC component is not the only potential source of interference when dealing with EEG signals. The electrical activity of the heart often impacts EEG recordings too. It is captured similarly to EEG (by placing electrodes on the skin), and the recorded electrical signal is known as an electrocardiogram (ECG). While EEG signals are usually in the range of tens to hundreds of microvolts (μV), the ECG signals have wider amplitudes on the range of millivolts (mV), meaning that despite having a good isolation between ECG and EEG electrodes, interference from ECG into EEG can still occur. Furthermore, the ECG frequencies may fall in the same range as neonatal EEG and thus produce seizure-like patterns in EEG signals.

When the ECG signal is provided additionally to the EEG data, it can be used to estimate the amount of ECG interference in any given EEG signal. The following parametric model has been considered in this study: the ECG signal is superposed in the EEG signal as an attenuated and delayed version of the ECG reference.

$$s[t] = \hat{s}[t] + \alpha \cdot ECG[t - \delta] \tag{1}$$

where $s[t]$ represents the EEG signal corrupted by ECG, $\hat{s}[t]$ the estimation of true EEG signal, and α, δ are the model parameters for the attenuation delay respectively. Those are estimated through the location of the maximum correlation between EEG and ECG reference. Once the parameters (α, δ) are obtained, the estimation of the cleaned EEG, $\hat{s}[t]$, can be made by Eq. 1. The ECG interference amount (α) have temporal variations over long EEG signals. Therefore, the parameters of this model are re-estimated every 8s for 16s segments of EEG. The example of ECG-corrupted and cleaned EEG is shown in Figure 2.

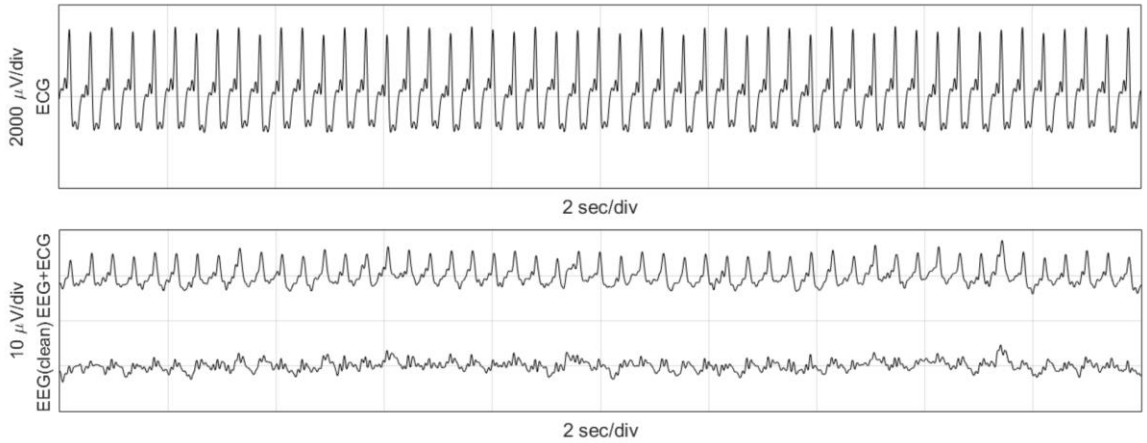

Figure 2. From top to bottom: Recorded ECG signal, recorded EEG (single channel) and cleaned EEG with ECG artifact removed.

Finally, EEG signals may have big bursts due to electrodes disconnections or other types of artefacts, but seizures typically fluctuate in the range of ±100μV. Thus, all signals are normalized to this value for the subsequent sonification states. After the signal normalization, a dynamic-range compressor is applied to reduce all values above ±1 to ensure the signal is contained in this normalized amplitude range.

2) DSP block: Denoiser

A variant of spectral subtraction which is utilized in speech processing [45] is used to reduce the amplitude of those EEG sections that do not contain seizure events. The short-time Fourier transform (STFT) is computed for the whole EEG signal $X[n,k] = STFT\{x[t]\}$. In equation (2), a spectral profile of the background sections $T[k]$ is estimated as a weighted power spectral density $P_{XX}[n,k] = |X[n,k]|^2$ with respect to the complementary value of the seizure probability given by the AI algorithm $AI[n] \in [0,1]$ (i.e. background EEG contributes more to the spectral profile estimation than the seizure events).

$$T[k] = \sum_n P_{XX}[n,k] \frac{1 - AI[n]}{\sum_n (1 - AI[n])} \tag{2}$$

Then, the following gain function is applied to the STFT:

$$G[n,k] = \begin{cases} 1 & if\ P_{XX}[n,k] \geq T[k] \\ \dfrac{P_{XX}[n,k]}{T[k]} & otherwise \end{cases} \tag{3}$$

Note that the gain function $G[n,k]$ has continuity when $P_{XX} = T$, and also $G[n,k] \leq 1$, meaning that $G[n,k]$ actually serves as an attenuation factor. Intuitively, each STFT bin would be attenuated proportionally to its own power when this is smaller than the average power of the background spectral profile (T). The denoised EEG STFT (Y) is obtained as:

$$Y[n,k] = 10^{AI[n]-1} \cdot G[n,k] \cdot X[n,k] \tag{4}$$

The first term adds an additional attenuation factor of up to 10% when seizure probability approaches zero (i.e. no seizure). Finally, the Inverse-STFT is computed to obtain the time-domain denoised EEG signal.

By using this process, any activity that deviates from the average activity in the recording is emphasized by taking into account the level of seizure-ness of such activity.

3) DSP block: Phase vocoder

Phase vocoder basics

The phase vocoder (PV) is a well-known technique to manipulate signals in the time domain by preserving their spectral properties. The PV was firstly developed to stretch speech signals in time for encoding voice data (hence the term, vocoder) [46]. Although this technique was first applied to speech, it also has been used to manipulate the various type of audio [47] and non-audio signals [48, 49]. The PV interpolates new frames between the existing frames of the STFT of the input signal (Figure 3).

The interpolation of the STFT frames is performed on both the magnitude and phase, as $STFT\{\cdot\}: \mathbb{C} \rightarrow \mathbb{C}^2$. While the magnitude can be trivially interpolated by any method (linear, quadratic, spline, etc.), the main challenge on the PV is the phase interpolation (as its name implies). The original implementation [46] ensured that the phase difference (instantaneous frequency) between each frame of the output STFT is the same as the difference between the analogous frames in the input STFT.

Let $X[n,k] = STFT\{x[t]\}$ be the 2-dimensional input *STFT*, where $0 \leq t \leq T-1$ indicates the time index of the time-series signal, and $0 \leq n \leq N-1$ and $0 \leq k \leq K-1$ stand for the 2D indexes of the STFT matrix, being N the total temporal frames of the STFT (columns) and K the total number of frequency-bins (rows).

Let $Y[m,k]$ be the STFT resulting from the PV algorithm, with $0 \leq m \leq M-1$, being M the number of frames of the resulting STFT, which is related to the original number of frames N through the chosen stretching ratio $R = M/N$. The frame index relation of m with respect to n can be expressed as $m = \lfloor R \cdot n \rceil$, and vice versa $n = \left\lfloor \frac{m}{R} \right\rceil$, where $\lfloor \cdot \rceil$ denotes the closest smallest integer. The stretching ratio is defined as positive ($0 < R < \infty$), but when $R < 1$ the signal would be contracted instead of stretched.

The magnitude of the resulting STFT $\|Y[m,k]\|$ can be obtained through linear interpolation of the input $\|X[n,k]\|$. This can be written using a recurrent expression, where the magnitude of the next frame $\|Y[m+1,k]\|$ is obtained from the previous $\|Y[m,k]\|$ and $\|X[n,k]\|$:

$$\|Y[m+1,k]\| = \|Y[m,k]\| + \frac{1}{R}\frac{d}{dn}\|X[n,k]\| \tag{5}$$

where $\|Y[0,k]\| = \|X[0,k]\|$ and $\frac{d}{dn}\|X[n,k]\| = \|X[n+1,k]\| - \|X[n,k]\|$ it is defined as the magnitude difference between two consecutive frames (i.e. discrete derivative).

The phase can be obtained using a similar recurrent equation:

$$\measuredangle Y[m+1,k] = \measuredangle Y[m,k] + \frac{d}{dn}\measuredangle X[n,k] \tag{6}$$

where $\measuredangle$ denotes the phase (angle), $\measuredangle Y[0,k] = \measuredangle X[0,k]$, and $\frac{d}{dn}\measuredangle X[n,k] = \measuredangle X[n+1,k] - \measuredangle X[n,k] = \omega_X[n,k]$ is the phase difference between two adjacent input frames, and also the instantaneous frequency. Note that in this case, the last term is not scaled by the stretching ratio R. This is to keep the phase consistency between adjacent frames.

Rearranging the terms from (6), and considering that $\frac{d}{dn}\measuredangle X[n,k] = \omega_X[n,k]$ and $\frac{d}{dm}\measuredangle Y[m,k] = \omega_Y[m,k]$ (i.e. the time-derivative of the phase is the instantaneous frequency), one arrives at the following conclusion:

$$\omega_Y[m,k] = \omega_X[n,k] \tag{7}$$

In other words, the instantaneous frequency is preserved, which reflects better the PV intentions of stretching the signal in time only without affecting its frequency content.

Finally, the time-domain stretched signal is obtained by the inverse STFT $y[t] = STFT^{-1}\{Y[m,k]\}$

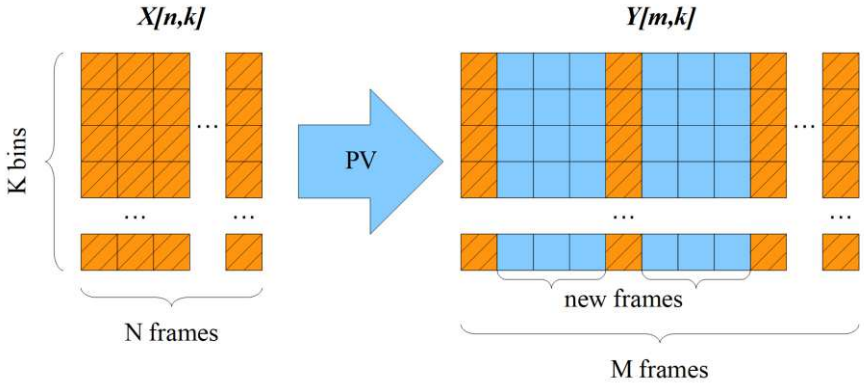

Figure 3. General scheme for input STFT X[n,k] and phase vocoder output STFT Y[m,k].

After this process, if the same sampling frequency it is kept on the obtained signal, $y[t]$, as in the original, $x[t]$, it would also keep the same spectral properties (i.e. same frequency content) proportionally stretched in time by a certain factor R as intended. However, the PV can also be exploited to shift the spectra of a signal while keeping its duration when choosing a sampling frequency R times larger than the original. In this case, the stretching ratio can also be expressed as a function of the relationship between input/output sampling frequencies as follows:

$$R = Fs_{out}/Fs_{in} \tag{8}$$

In this work, the following properties of the phase vocoder are exploited – first, to conventionally raise the spectral characteristics of EEG signals into an audible range; second, to compress the signals in time to make them shorter and therefore quicker to listen.

Wrapped versus unwrapped phases

As described in [50], the phase of a complex number $\measuredangle$ has a cyclic characteristic around 2π, and it is usually reduced to the range $\Theta = (-\pi, \pi]$ radians, known as the wrapped phase. One can achieve this by adding or subtracting multiples of 2π until the value fits inside Θ range. Mathematically speaking, the unwrapped phase can be any real number $\measuredangle \in \mathbb{R}$, and the wrapped phase is restricted to $(-\pi, \pi]$ interval $\measuredangle \in \Theta$. The wrapped phase can be computed from the unwrapped phase as:

$$(\measuredangle \mathrm{a})_\Theta = mod\{(\measuredangle \mathrm{a})_\mathbb{R} + \pi, 2\pi\} - \pi = (\measuredangle \mathrm{a})_\mathbb{R} - 2\pi \left\lfloor \frac{(\measuredangle \mathrm{a})_\mathbb{R} + \pi}{2\pi} \right\rfloor \tag{9}$$

When dealing with a single complex value series, the phase may convey the information about the total number of revolutions (unwrapped) or just the fractional part over 2π (wrapped). However, when working with complex sequences, the wrapped phase can lead to inconsistent phase differences and consequently inconsistent instantaneous frequency estimations for the case of the PV. Thus, the unwrapped version of the phase should be used for the estimation of the instantaneous frequency. The unwrapped phase is obtained by differentiating the wrapped phase, wrapping the difference and integrating back:

$$(\measuredangle \mathrm{S}_n)_\mathbb{R} = \sum_{k=0}^{n} \left[\frac{d}{dk} (\measuredangle \mathrm{S}_k)_\Theta \right]_\Theta \tag{10}$$

Linear frequency estimation

The linear interpolation of the phase leads to a propagated version of the instantaneous frequency $\omega_X[n,k]$ for $\lfloor R \cdot n \rfloor \leq m < \lfloor R \cdot (n+1) \rfloor$, which causes a potential discontinuity at $m = \lfloor R \cdot (n+1) \rfloor$ as the instantaneous frequency changes suddenly from $\omega_X[n,k]$ to $\omega_X[n+1,k]$, and $\omega_X[n,k] \neq \omega_X[n+1,k]$ in general. This issue is not relevant for most of the applications of the PV where the stretching factor may be larger than one but not much, like $R < 2$ or $R < 4$ at most. However, when $R \gg 1$ (i.e. $R = 100$), the effect of the discontinuity of the instantaneous frequency becomes more evident to the ears, which will be perceived as sudden tonal jumps.

A better estimation of the instantaneous frequency at each output frame is proposed here through the means of linear interpolation of $\omega_X[n,k]$ to ensure a continuous (smooth) transition between instantaneous frequencies of consecutive frames $(n, n+1)$:

$$\omega_Y[m,k] = \left(1 - \frac{m}{R} + n\right) \omega_X[n,k] + \left(\frac{m}{R} - n\right) \omega_X[n+1,k] \tag{11}$$

By rearranging terms, it is obtained:

$$\omega_Y[m,k] = \omega_X[n,k] + \left(\frac{m}{R} - n\right) \frac{d}{dn} \omega_X[n,k] \tag{12}$$

which leads to an updated version of the original expression in (6), with an additional term:

$$\measuredangle Y[m+1,k] = \measuredangle Y[m,k] + \frac{d}{dn} \measuredangle X[n,k] + \left(\frac{m}{R} - n\right) \frac{d^2}{dn} \measuredangle X[n,k] \tag{13}$$

As a result of enforcing first-order continuity to the instantaneous frequency with equation (11), the phase $\measuredangle Y[m,k] = \sum_{i=0}^{m} \omega_Y[i,k]$ has subsequently second-order continuity.

Variable stretching ratio

So far, it has been assumed that the stretching ratio R to be constant. But what if we want the R to be dynamic depending on the signal properties? For example, suppose that the input STFT is to be stretched by three different stretching factors R_1, R_2 and R_3. One could simply divide the input STFT into those three chosen parts and apply the PV consecutively using R_1, R_2 and R_3, respectively. An example is shown in Figure 4.a, where the input STFT consists of 3 consecutive tones at 8, 16 and 32Hz of 10 seconds long each. A time-dependent stretching factor R affects the output STFT differently; the duration of the first tone becomes twice as larger as in the input (R=2). The second tone remains the same length (R=1). The length of the third tone is halved (R=0.5).

By taking this approach to the limit, one could consider that each input frame n has a well-defined value of $R[n]$ and make use of the PV equations (5) and (13) to extrapolate the magnitude and phase of the output STFT $Y[m,k]$ for a given $X[n,k]$. Figure 4.b, shows a chirp signal being stretched continuously under this principle.

By doing that, we can achieve a variable-speed PV (VSPV) since the stretching ratio is inversely proportional to the "speed" ($VS[n]$), while also raising the frequency content:

$$R[n] = \frac{Fs_{out}/Fs_{in}}{VS[n]} \tag{14}$$

Where in particular, the EEG signals in this work need to be raised from a sampling frequency of $Fs_{in} = 32Hz$ to the audio sampling frequency $Fs_{out} = 16000Hz$.

Equation (14) allows for decoupling the total stretching ratio into (a) the amount of frequency raising required (numerator) and (b) the actual temporal compression (denominator). The variable speed $VS[n]$ in the VSPV is used in EEG sonification to reduce the reviewing effort time, regardless of the input/output sampling frequencies. AI is used to control the speed factor $VS[n]$, which follows the exponential rule on the given equation:

$$VS[n] = V_{max}\left(\frac{V_{min}}{V_{max}}\right)^{AI[n]} \tag{15}$$

where $V_{max} = 3600$, (1 hour compressed of EEG compressed as 1 second of audio for background sections) and $V_{min} = 60$ (1 minute of EEG compressed as 1 second of audio for seizure sections) are the maximum and minimum velocities (speeds), $AI[n] \in [0,1]$.

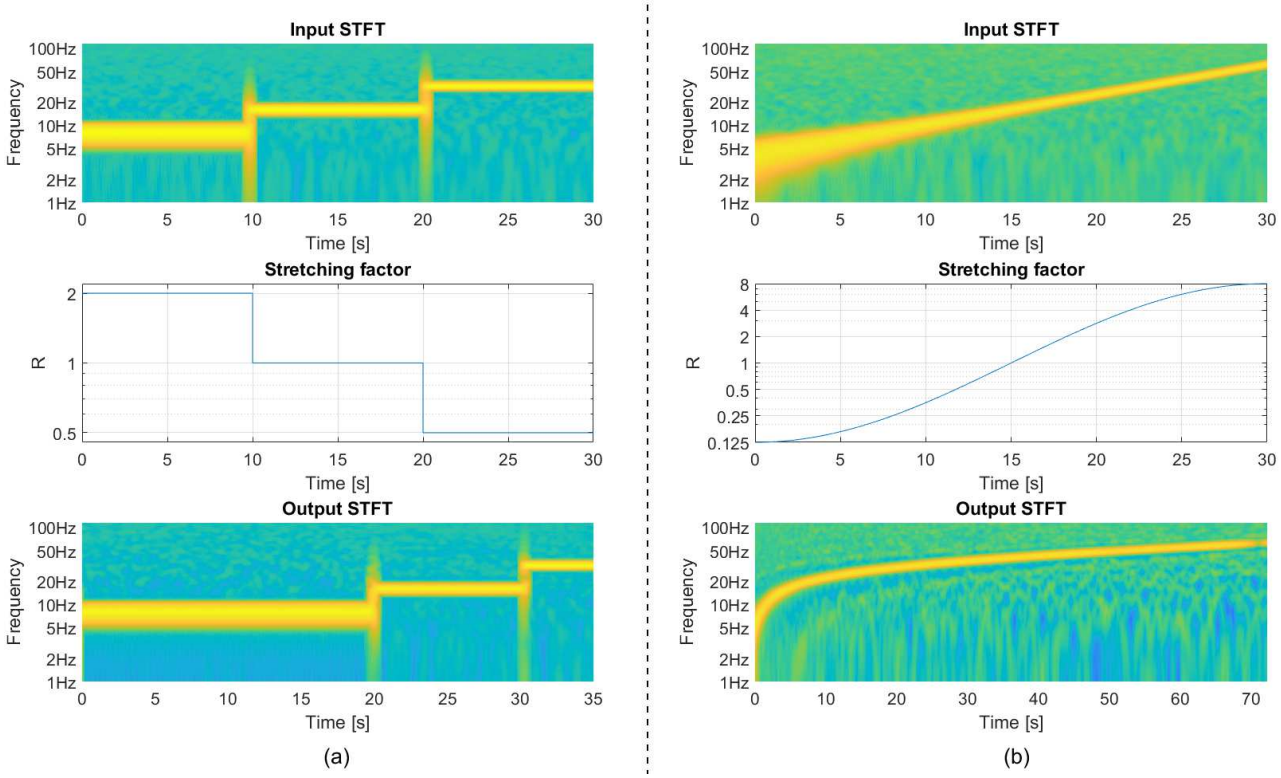

Figure 4. VSPV examples. (a) Tones of 8, 16 and 32Hz stretched/contracted individually. (b) Chirp signal being stretched continuously.

Given the variable speed, the total duration, T_o, of the output signal can be precomputed beforehand as a function of the total duration of the input signal, T_i, and $VS[n]$:

$$T_o = T_i \frac{1}{N} \sum_{n=0}^{N-1} VS[n] \tag{16}$$

Moreover, the expansion and compression rates (ER, CR) are defined as the ratio of those durations:

$$ER = \frac{1}{CR} = \frac{T_o}{T_i} = \frac{1}{N} \sum_{n=0}^{N-1} VS[n] \tag{17}$$

where T_i, T_o can be measured either in seconds or the number of STFT frames. In a typical use-case of this work, the duration of the obtained signal would often be shorter than the original ($T_o < T_i$) the CR is frequently utilized to measure the capacity of condensing EEG data into sound.

4) AI block: Seizure detection model

The seizure probability obtained from an AI algorithm is used in two of the stages of the proposed sonification algorithm: (a) Denoiser and (b) VSPV. Any seizure detection algorithm that provides a probabilistic output per segment and channel of EEG could potentially be used. In this work, the deep learning neonatal seizure detection algorithm is utilized [51]. The model consists of a fully convolutional neural network (FCNN) architecture, which runs on top of raw EEG data and does not require hand-crafted features. The algorithm has been extensively evaluated to show the stage-of the-art generalization error and was also validated on the publicly available Helsinki dataset (same dataset used in this work), showing an area under the curve performance of 0.956 at detecting seizure events.

The FCNN algorithm takes 8s-long fragments of the pre-processed multichannel EEG signal, which is filtered and downsampled to 32Hz, similar to the pre-processing stage in this study. For each segment, the model then generates a probabilistic output per channel. For this study, an overlap of 7s is used to achieve a seizure-probability throughput of 1Hz. By doing so, one probabilistic output per second is given in the form of a multichannel time series. After that, the probabilities are smoothed in each channel and merged into an overall probabilistic output by taking the maximum probability across channels:

$$AI[n] = \max_{CH} \{AI[ch, n]\} \tag{18}$$

Originally [51], a moving-average filter of 60s was used to smooth the probabilistic output in order to balance the rate of false alarms with the rate of detected events to maximize the accuracy of seizure detection. In this work, in order to allow more events (false and correct) to pass through the smoothing filter, a 15s median filter was utilized to allow more short seizure events in the context of seizure sonification. As shown in Figure 1, the smoothed probabilities are used in the denoising stage and in the VSPV where the overall probability obtained from Eq. (18) is combined with the geometric transformation from Eq. (15) to be used to control the speed factor.

5) DSP block: Stereo mixer

Up to this point, the signals for each EEG channel have been converted into sound independently, obtaining, as a result, an audio signal for each EEG channel. Those audio signals need to be mixed into a standard stereophonic sound system so most standard devices can be play it. In other words, a number of independent sources of information (audio from EEG channels) need to be mapped into 2 representing the left and right audio channels. There are many ways to achieve this purpose, depending on the degree of complexity of the model utilized to undertake this goal. Perhaps the most straightforward approach is to map the left/right hemisphere channels into left/right stereo channels, respectively [43]. However, if for any reason the listener loses one of the stereo channels, half of the EEG information is lost too. From this perspective, an approach that utilizes the stereo attribute as a complementary but non-essential part of the auditory display is preferred. More complex models more adequately approximate the so-called head-related transfer function (HRTF) [52]. These models aim to mimic how humans perceive sound sources as a function of the relative distance, azimuth, elevation and frequency [53, 54]. With HRTF models, the lack of one stereo channel does not suppress the sound source completely. However, the degree of complexity in such cases is unnecessary or even unwanted for the scope of this work. First, due to the attenuation of propagation proportional to the squared of the source distance, some channels would be perceived louder than others, and in this work, all EEG channels are considered equally important. Second, the EEG channels can be displayed in a 2D plane, disregarding the elevation dimension. Lastly, the transfer function to mimic front/back works well for well-known, and natural sounds such as speech, but the listener is a priori not familiarized with sonified EEG. Thus, the main takeaway from these works is the azimuthal dimension. This allows for the localization of the sound source direction [55], thus enabling the listener to differentiate from left, central or right hemisphere sonified electrode pairs.

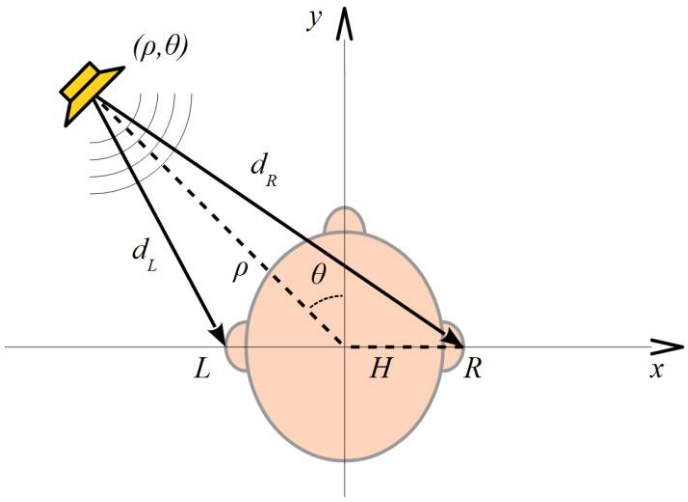

Figure 5. Precedence effect due to interaural delay.

Table 1. Polar coordinates of electrodes. For simplicity and due to the symetry of the montages, the table just shows the left-hemisphere pairs. The right hemisphere pairs have just the same values with negative sign on the azimuthal angle.

8-Channel:										
Chan.	F3–C3			C3–O1			T3–C3		C3–Cz	
ρ/H	5.2			6.3			7.4		2.6	
θ	60º			140º			90º		90º	
18-Channel:										
Chan.	Fp1–F3	F3–C3	C3–P3	P3–O1	Fp1–F7	F7–T3	T3–T5	T5–O1	Fz–Cz	Cz–Pz
ρ/H	8.0	5.1	5.1	8.0	9.5	9.6	9.6	9.5	2.4	2.4
θ	30º	60º	120º	150º	35º	70º	110º	145º	0º	180º

A simple model based exclusively on the precedence effect due to the interaural time delay (ITD) is used to accomplish this goal by virtually situating each audio source in a 2D plane in relation to an imaginary listener [56]. The model assumes a listener with right and left ears (*L* and *R*) separated by a certain opposed distance H from the origin of coordinates (head centre), and that the sound source is at a distance ρ within a certain azimuthal angle θ defined positive towards left (Figure 5). The model assumes the simplest approach where there are no obstacles between the sound source and the ears (i.e. no attenuation) and that the sound reaches the ears by direct propagation only, with no reflections.

Under this assumption, one can obtain through trigonometry the equations for the left and proper propagation delays $d_{L/R}$ as:

$$d_{L/R} = \frac{1}{c}\sqrt{\rho^2 + H^2 \pm \rho \cdot H \cdot sin\theta} \tag{19}$$

where the positive sign is considered for left (L). The constant, c, indicates the propagation speed of sound in air, which is about 340m/s, and H indicates the radius of the head, which was rounded to 10cm.

Under this principle, each audio signal corresponding to each EEG channel is then virtually placed on a specific set of coordinates, (ρ, θ), to obtain a pair of delays, $d_{L/R}$, associated with each channel (Table 1). These positions were chosen to be proportional to the actual physical positions (same θ with augmented ρ) of the electrode pairs in the EEG montages utilized in this study: 8-channel and 18-channel montages shown in Figure 5.

Once these delays are obtained, the stereo audio is obtained by applying the corresponding L/R delays to each channel and summing them up onto each stereo channel.

$$s_{L/R}[t] = \sum_{ch=1}^{N_{CH}} a\left[t - d[ch]_{L/R}, ch\right] \tag{20}$$

where $a[t, ch]$ is the time-series matrix of the sonified data, and $s_{L/R}[t]$ the left/right stereo audio.

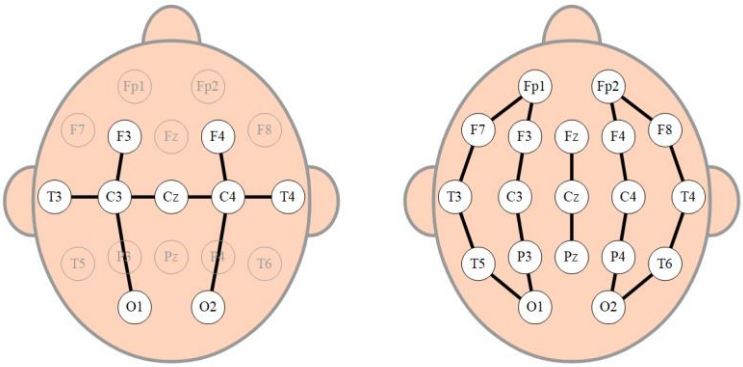

Figure 6. Aproximate locations on the scalp of the 8 and 18 channel montages for EEG acquisition.

III. EXPERIMENTAL DESIGN AND METRICS

1) Dataset

The publicly available neonatal EEG dataset provided by the Helsinki University Hospital is used in this study [57]. This dataset was obtained from 79 neonates admitted to the NICU, and recordings were performed due to suspicion of seizures. It contains excerpts of 1-2h multichannel EEG recordings at a sampling frequency of 256 Hz stored in a referential montage with 19 electrodes positioned as per the international 10–20 standard, including a recording reference at the midline. The presence of seizures in these patients was annotated independently by three experts with over ten years of experience in visual interpretation of neonatal EEG. The annotations are represented by the timestamps for the onsets and the duration of the seizures, with no information about the specific spatial location (EEG channel) of the seizure event. All annotated seizure events were at least 10 seconds long. According to the annotations, by consensus regarding the presence or absence of seizures in each recording, 39 neonates had seizures (at least one seizure event detected by all experts), 22 were seizure-free by consensus, the diagnosis of the remaining 18 patients varied among experts.

2) Design of the survey

A total of 79 multichannel 1-2h EEG recordings were sonified and converted into 79 audio samples, respectively. The level of accuracy of the developed AI-driven sonification algorithm was assessed in a survey. The survey participants were asked to distinguish the recordings with any sign of seizures from those with no sign of seizures by just listening to the audio samples. In other words, in each recording, the survey only asked about the presence/absence of seizures regardless of the actual number of seizure events or their duration.

An online platform was developed to conduct the survey (http://sergigomezquintana.github.io/EEGsoundSurvey). As shown in Figure 7, a simple graphical user interface was designed to display the sonified examples and allow a participant to provide an answer and move to the next sample.

A subset of five audio samples were chosen as training examples for the participants: one containing an obvious seizure, two with less obvious seizures (short and low amplitude), one with normal background EEG and one containing ECG artefact. The examples were carefully selected to show a good representation of the wide variety of sounds associated with seizure and no-seizure activity with the least amount of samples possible.

In an initial exploratory phase of the study, a group of 5 internal non-clinical participants evaluated the developed AI-driven sonification algorithm in several variations to tune various algorithmic hyperparameters. The following variations were considered in this study with the settings chosen for the external survey outlined in bold:

a) Sonification **with** and without an ECG removal algorithm to reduce the interference from ECG.

b) Changing the shape of the probability density function from uniform to U-shape vs **no reshaping**.

c) **Minimum speed-factor** (for seizure-epochs) ranging between **60** and 20.

d) Effect of reducing/increasing the **number of EEG channels** from **8** to 2/18, respectively.

After this initial evaluation, the best performing variation of the algorithm was selected to be evaluated among the clinical cohort of targeted end-users using the developed online platform.

3) Evaluation metrics

The kappa statistic measures the interrater reliability and agreement level between different annotators [58]. While Cohen's kappa originally measured the interrater reliability between two raters, the Fleiss kappa is an adaptation of Cohen's kappa for three or more raters. In both cases, the kappa statistic is defined as:

$$\kappa = \frac{p - p_e}{1 - p_e} \tag{21}$$

where p is the relative observed agreement among raters, measured as the number of instances where raters agree divided by the total number of instances, and p_e is the probability of agreement by chance.

The confidence interval of 95% of this measure is given by:

$$CI(\alpha) = \kappa \pm Z_s(n, \alpha) \cdot \frac{SD_\kappa}{\sqrt{n}} \tag{22}$$

where $Z_s(n, \alpha)$ indicates the z-score for $n = 79$ samples (patients) and an estimation error $\alpha = 0.05$. SD_κ indicates the standard deviation for kappa, given by:

$$SD_\kappa = \sqrt{\frac{p(1-p)}{(1-p_e)^2}} \tag{23}$$

The kappa score can range from -1 to +1. The level of agreement can be interpreted depending on the kappa value range, from slight (0.01–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80) and almost perfect agreement (0.81–1.00).

While comparing the level of agreement between sets of different raters or annotators, it is important to provide statistical significance in order to show that a certain distribution (in this particular case, given by kappa mean and variance) are sufficiently similar. For this purpose, the student's t-test is used to obtain the t-statistic of the two sample distributions to compare. For the particular case where both distributions have the same number of samples, the t-statistic reduces to the following expression:

$$t = \frac{\kappa_1 - \kappa_2}{\sqrt{\frac{SD_{\kappa_1}^2 + SD_{\kappa_2}^2}{n}}} \tag{24}$$

where κ_i and SD_{κ_i} are the mean and standard deviations of the kappa statistic of each distribution. The p-value can then be derived from the t-statistic using the inverse of the cumulative density function (CDF) of a t-students distribution of n degrees of freedom.

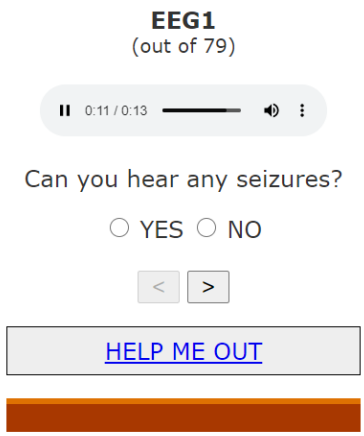

Figure 7. Snapshot of web survey to assess the AI-driven sonification algorithm.

The performance of the sonification algorithm is evaluated at the patient level, meaning that only the presence or absence of seizures is being evaluated. Thus, in order to establish a proper comparison with the AI algorithm alone, a single probabilistic output per patient was obtained by taking the maxima across time, i.e. taking the maximum across EEG channels and time over the smoothed probabilistic output (using the original 60s moving average filter):

$$AI = \max_t\{AI[t]\} = \max_t\left\{\max_{ch}\{AI[ch,t]\}\right\} \tag{25}$$

The area under the receiver-operating curve (AUC ROC) [59, 60] is used to measure the performance of a binary-classification AI algorithm. In general, this metric is computed from the predictions given by a model with respect to the ground truth (true value). In this work, the AI-probabilistic output obtained from (25) is used as a prediction, while the majority vote among the three annotators was used as a ground truth. In the best-case scenario, the ROC curve would converge into a squared-shape of area equal to one, and worst-case scenario (predictions obtained randomly), the curve would approximate a right triangle of area equal to 0.5 (equivalent to flipping a coin). Thus, the expected AUC score for the AI algorithm would range between 0.5 to 1.

IV. RESULTS

The survey was conducted through the web page with the link disseminated by e-mail invitations. Twenty-three experienced clinicians have taken part in the survey and answered all audio questions (without any access to the EEG signal). The survey participants' answers were first contrasted with the experienced annotators who used a full continuous video EEG setup to annotate the signal.

Table 2 shows the kappa statistics for inter-rater agreement among four annotators. The first three are original EEG annotators from the dataset, whereas the majority vote represents the fourth annotator among all survey participants. The mean and the 95% confidence intervals are computed for Cohen's kappa scores. The level of agreement between the three original annotators and the sonification is no different from the agreement level among the three original annotators.

Figure 9 shows the mean and the 95% confidence interval of the Fleiss' kappa for inter-rater agreement using any three annotators at each time. The results similarly indicate no statistically significant difference between the accuracy obtained by an experienced EEGer and the healthcare professionals with access to sonified EEG only, with all p-values higher than 0.457.

Figure 8 shows the performance obtained from the AI sonification survey compared to the accuracy of the AI algorithm alone in terms of the AUC. The red/black mark indicates the performance obtained by the clinical (23) and non-clinical cohort (12), respectively. It can be seen that the performance of AI-driven sonification is superior to any point in the curve of the AI algorithm alone, implying that the additional benefits of sonification and a human interpreter in the loop are advantageous in terms of the accuracy of detecting patients suffering from seizures. The clinical cohort achieves higher sensitivity (0.89 vs 0.83) but lower specificity (0.78 vs 0.91) than that of the non-clinical cohort.

With an intent to have a deeper dive into the analysis of errors obtained with AI sonification, Table 3 shows the confusion matrix of the AI-driven sonification (clinical cohort only), with a total of 7 false alarms (false positives) and five undetected seizure patients (false negatives).

Focusing on false negatives from Table 3, Table 4 shows the statistics of different characteristics of the seizure-present recording in terms of seizure duration, EEG amplitude of the seizures, the number of annotators agreeing and the AI probabilistic output, to compare detected (42 patients) and undetected patients (5 patients). The AUC scores which are reported for each characteristic can be interpreted as the quantitative measure of importance towards the separation between detected and undetected seizure patients. The missed patients can be mainly attributed to low AI probabilistic output (an AUC of 0.981) and the low amplitude of the EEG (an AUC of 0.861).

Focusing on false positives from Table 3, Table 5 shows the statistics of different characteristics of the non-seizure patients in terms of the number of annotators disagreeing on the health status and the AI probabilistic output, to compare the detected (25 non-seizure patients) and wrongly detected as seizure (5 patients).

Table 6 presents the performance of AI sonification as a function of various parameters of the system during the development stage when evaluated using the internal survey of 5 participants before the parameters were set and released to the external survey. The baseline setting consists of no ECG removal, no probability function transformation and min speed set to 60 (seizure events are compressed by a speed factor of 60).

Table 7 presents the performance of AI sonification as a function of the number of EEG channels utilized as input. Two-channel settings utilize only F4-C4 and F3-C3 EEG channels for sonification. The pairs of electrodes present in 8 and 18-channel montages can be found in Figure 6. The best performance is obtained with an 8-channel montage.

Table 2. Cohen's kappa scores for annotator pairs and AI sonification algorithm (Kappa ± CI95).

	Annotator 1	Annotator 2	Annotator 3	AI sonification
Annotator 1	1	0.819±0.130	0.651±0.176	0.710±0.161
Annotator 2	0.819±0.130	1	0.628±0.180	0.686±0.166
Annotator 3	0.651±0.176	0.628±0.180	1	0.503±0.205
AI sonification	0.710±0.161	0.686±0.166	0.503±0.205	1

Table 3. Confusion matrix for the AI sonificaiton majority vote.

Predicted \ Actual	Positive	Negative
Positive	42	7
Negative	5	25

Table 4. Average ± CI95 for duration, amplitude rms and average AI probability for correctly detected and missed seizures. The last row shows AUC score of a given feature (duration, rms, probability) as an indication of how discriminative the feature is on its own for separation between detected and undetected patients.

Seizure patients (N=47)	Duration (s)	Amplitude RMS (μV)	# Annotators agreeing	Seizure probability (AI)
Detected (N=42)	496±157	67.8±28.2	2.90±0.0925	0.790±0.0746
Missed (N=5)	279±207	18.6±8.18	2.400±0.629	0.228±0.1058
AUC	0.595	0.861	0.752	0.981

Table 5. Average ± CI95 for AI probability for correctly detected seizures and missed seizures. Last row shows AUC score as a reference on how related are the features with the errors.

Non-seizure patients (N=32)	# Annotators disagreeing	Seizure probability (AI)
Detected (N=25)	0.280±0.188	0.287±0.075
False alarms (N=7)	0.428±0.477	0.526±0.125
AUC	0.574	0.851

Table 6. The influence of the parameters on the performance of AI sonification, in terms of sensitivity, specificity and compression rate.

Setting	Sensitivity	Specificity	Compression Rate
Baseline	0.787	0.949	727.4
ECG removal	**0.830**	**0.951**	727.4
Uniform distribution probabilities	0.809	0.864	359.3
Min Speed = 20	0.809	0.927	330.1

Table 7. Performance for different number of EEG channels.

# EEG channels	Sensitivity	Specificity
2	0.617	0.935
8	**0.830**	**0.951**
18	0.851	0.851

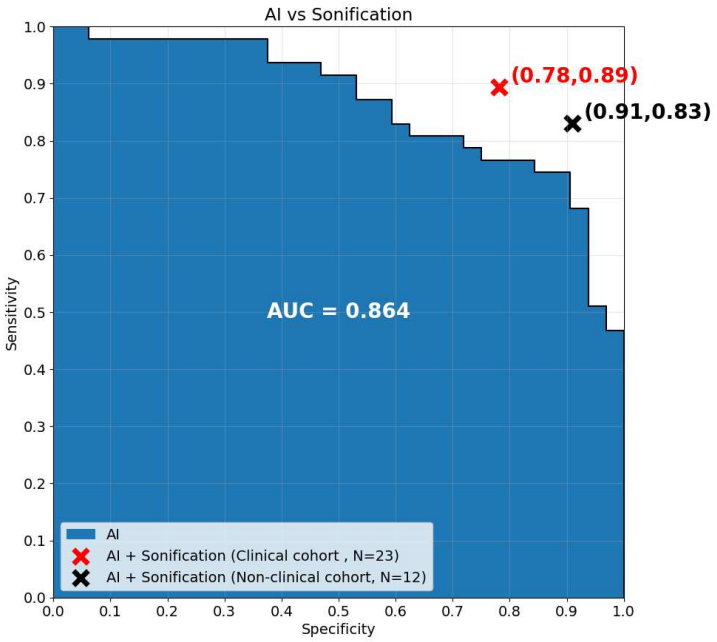

Figure 8. Area under the curve for the AI probabilistic output and sensitivity/specificity of the majority vote of the survey participants.

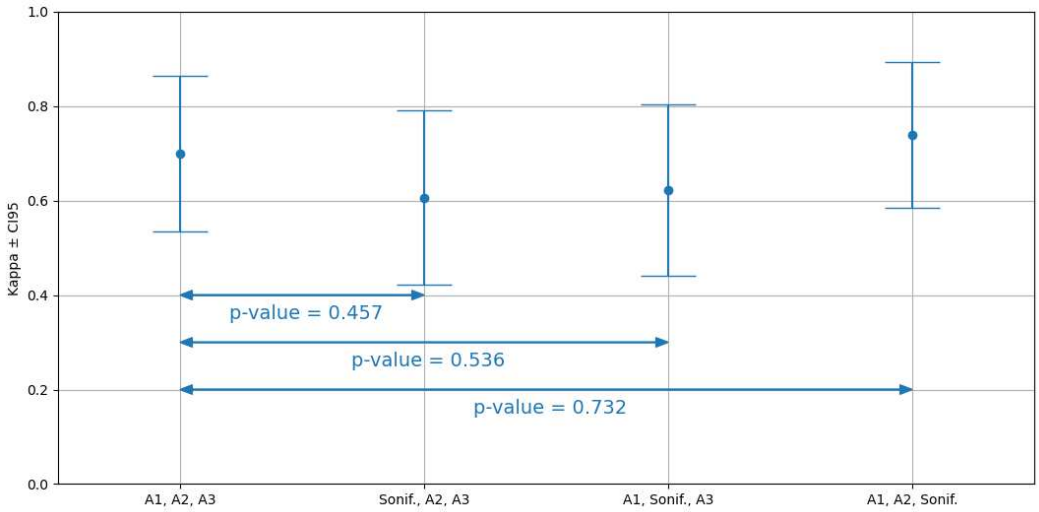

Figure 9. Fleiss kappa while exchanging one annotator by the AI sonification survey results and its p-values.

V. Discussion

EEG monitoring is considered the gold standard for detecting seizures in newborns [61]. However, continuous EEG monitoring in clinical settings is limited due to the lack of specialists onsite [15]. In addition, interpretation of EEG is a challenging task, and neurophysiologists are not available 24/7 to the clinical team. In the developing world, the expertise availability is even scarcer [62], especially in the first 24h after birth.

Without timely neurophysiological support, healthcare professionals must rely on a simplified aEEG version of EEG or clinical manifestations to detect possible brain injuries. Unfortunately, both approaches are inaccurate in detecting seizures [13, 18, 19, 63]. The misdiagnosis does not improve even when continuous electroencephalography is available, but the timely interpretation is not. For example, in the secondary data analysis of 2 European multicenter cohort studies from 2 clinical trials [64], with continuous EEG monitoring available, 20% of newborns with seizures were misdiagnosed, with some not given antiseizure medication at all and others given before seizures started. Similarly, 258 newborns received antiseizure medication in a clinical setting, whereas only 154 newborns were confirmed with seizures retrospectively using EEG. From those wrongly diagnosed, 47 infants received multiple antiseizure medication doses throughout the study.

Timely intervention in reducing the severity of brain damage due to encephalopathy is very important. In the same study [64], of those who received medication correctly only, 17% received it within 1 hour of seizure onset. It was shown that seizures treated within 1 hour of seizure onset subsequently had the lowest seizure burden when compared with infants who received antiseizure medication after 1 hour of seizure onset.

Scaling up the skills of healthcare professionals in managing asphyxia in newborns has been ranked as the second-highest of research priorities for improving newborn health and birth outcomes by 2025 [65].

With the increasing use of prolonged neonatal cEEG monitoring and insufficient neonatal neurophysiology expertise, additional support from automated seizure detection algorithms could be the solution. While a number of high-accuracy, automated, AI-driven solutions for seizure detection were reported [24, 34, 51] to maximize the use and exploit the benefits of such an automated system, the AI models used have to be explainable and keep the medical professional in the loop. These are the key requirements as physicians are unlikely to put their judgement aside in favour of an automated process or machine [66]. Explainable AI is a key enabler that augments the physician's capabilities without taking away the decision making. By developing human-in-the-loop AI, we may enable an ideal symbiosis of human experts and AI models, exploiting the strengths of both while at the same time overcoming their respective limitations [67, 68].

Hearing and vision represent two of the most used special senses by medical professionals. Humans improved their hearing abilities from the most basic functions of acoustically detecting natural hazards (such as storms or predator noises) to the development and subsequent perception of speech and music. These are some of the most complex forms of communication and expression known amongst animals. Interestingly, acoustic perception can sometimes be even more intuitive than sight when receiving sensory information. An example of that is language itself, which is naturally self-learned by hearing during the first year of life, whereas reading (which uses the sense of sight) needs to be explicitly taught, and partly because of that, it is considered to be a less natural way to acquire information [69]. The same idea can be explored when interpreting long time-series data such as EEG. However, the task is even more complex due to multiple concurrent channels of EEG, which need to be read/analyzed simultaneously.

Earlier works in EEG sonification simply increased the sampling frequency of the EEG signal to raise its frequency content to the audible range, assuming the collateral effect of time compression [70, 71]. The increased factor of the sampling frequency can be inconvenient as too large factors can make seizure events too short and therefore missed. More sophisticated methods have been developed since these earlier studies. Sonification based on tone synthesis are described in [72, 73]. A musical approach where EEG amplitude is mapped into musical notes was developed [74]. Another method used EEG as a modulator on a voicelike synthesizer [75]. However, those approaches do not convey the totality of the EEG information but are rather partially guided by EEG. The phase vocoder (PV) was firstly used for EEG seizure detection by raising the frequency content of the EEG into the audible range without affecting the time scale [43]. In the same context of seizure detection, a subsequent study compared the PV with a novel FM/AM sonification, showing interesting results on why the perception of seizures increases with quicker playbacks [36].

However, all the previous works in EEG sonification considered a constant time-compression factor, meaning that EEG data are compressed in time always at the same rate. Because seizures can be rare events, the constant time-compression factor results in a trade-off between the higher level of perception of short seizure events at the expense of reviewing long recordings versus having less listening overhead at the cost of missing occasional short seizure events. In this study, a solution for a quick preview of large multichannel EEG recordings is presented based on the PV as in [36, 43], with an addition of an AI guided time-compression factor which serves as an attention mechanism [51]. In this manner, the accuracy of the AI method is combined with interpretability and intuitiveness of sonification to allow for quick and accurate decision making. In further subsections, we discuss the advantages and drawbacks of the AI sonification method, comparing its speed and accuracy with that of full EEG annotators and with AI alone, alongside a deep analysis of the errors made.

A. AI sonification vs human annotators

The Fleiss and Cohen kappa scores presented in Table 2 and Figure 9 indicate that the accuracy obtained by a human analyzing the sonified 1-2h segment of EEG is not different to that of the experienced EEG interpreters analyzing full multichannel EEG. In [20], three experienced EEG experts reviewed over 4000h of multichannel EEG in 120h, which corresponds to the rate of ~2 mins per hour of EEG. The number of identified neonates with seizures was different for each reviewer, while the agreement on seizure events was only 78%. The AI-driven sonification compressed 1 hour of EEG to just 5s of audio (approximate average across the whole dataset) by focusing on the important segment of EEG, while remarkably achieving the same level of performance with no training required.

One of the main bottlenecks of the more pervasive usage of EEG monitoring systems is the lack of availability of interpretation expertise 24/7. Partly because of that, methods like aEEG aimed to simplify visual interpretation. However, some focal, low-amplitude, and brief seizures may be missed [63]. The general cost of the equipment can be an additional limitation, especially in the developing world. These limitations have triggered the research towards the brain stethoscope project; a low-cost and intuitive solution for quick EEG review [76-78].

A technological framework to assist the early detection of seizure events is proposed here; acquired EEG data would be analyzed through an always-on AI system that would raise the alarm to notify the clinical staff in the NICU immediately. The presented EEG acoustic interpretation would then be available on-demand for the clinical personnel to review the seizure event (reported by AI) intuitively, with an accuracy proven to be on par with trained EEG experts at detecting seizure patients. That would enable higher confidence for the clinical personnel to call for a full review of EEG while reducing the burden of the EEG review process amongst paediatric neurologists.

B. AI-driven sonification vs AI alone

As shown in the previous section, the ground truth for this dataset is subject to interobserver disagreement. However, to further analyze the algorithm performance, the single representation of the ground truth needs to be chosen, and this dataset was represented with consensus annotations across the three neurophysiologists who annotated the EEG.

When considering a human interpreter in an AI-driven decision support system, it is important to quantify the gain obtained by the designed system with respect to the task performed by AI alone. From Figure 8, it can be seen that there is a clear benefit from having the human in the loop of decision making. The performance is improved with respect to the AI system alone, but more importantly, the designed method can interpret the signal in an intuitively.

DL methods do not always learn intuitively understandable discrimination mechanisms, so they are often referred to as black-box AI systems [79]. Although the general trend for ML and AI is to develop autonomous systems capable of detaching from having a human supervisor in the loop [80], medicine is exceptional in this regard: understanding how and why a machine decision has been made is essential among medical professionals [81]. In addition to that comes the new European General Data Protection Regulation (GDPR), in which black-box methods will not be allowed to be used in business if they cannot explain why a decision has been made [82]. Here, the paradigm is different; AI is not meant to decide on the data directly but to be used as an attention mechanism to assist doctors' data interpretation. From this perspective, the proposed system is GDPR-compliant, but it also contributes to making AI more approachable and trusted among medical professionals, improving human-computer interactions.

While both clinical and non-clinical cohorts outperform the AI system, it is interesting that the clinical cohort achieves significantly higher sensitivity (0.89 vs 0.83) whereas the non-clinical cohort has significantly higher specificity (0.78 vs 0.91). The higher sensitivity obtained with the clinical cohort implies inherited bias of trained healthcare professionals to catch the disease. This bias affects the hearing so that even small pitch changes result in an alarm. This allows detecting seizure patients who were not identified by the non-clinical cohort. In contrast, the non-clinical cohort is much more conservative and only raises the alarm when confidently hearing something abnormal, with fewer false alarms than the clinical cohort.

C. Analysis of errors

The accuracy with respect to the consensus ground truth is shown in Table 3 in the form of a confusion matrix, from where a sensitivity (0.89) and specificity (0.78) can be inferred. The reported sensitivity values for detection of seizure presence in a patient for an aEEG, which also aims to simplify standard EEG visual analysis for non-EEG experts in the NICU, ranges from 0.22 to 0.57, with no false positives among the control group [83]. However, the ictal cohort of patients was significantly larger than the control group. Other studies with a more balanced but smaller cohort have shown that aEEG might result in overdiagnosis with respect to standard EEG, with high sensitivities (0.80) but specificities as low as 0.50 [84]. Our sound-based solution has been evaluated in a relatively balanced dataset (47 records with seizures vs 32 without any), showing strength in both sensitivity and specificity.

Only five training examples were preselected as training for the users, trying to be sufficiently representative of the diversity of each class (seizure versus background). Still, these may not be sufficient to depict all the possible variations on complex EEG signals. EEG requires a lot of understanding of its underlying characteristics to be accurately interpreted. Previous studies showed that the sensitivity at visually detecting seizure events could be between 2 to 6 times higher on those with previous experience at EEG analysis

[85]. In this regard, it is still remarkable that the achieved accuracy with such few training examples could improve further with a wider training set.

1) Misses

The detection of the presence of seizure events can be seen as a function of a) the duration of the audio that covers seizure events b) the discriminatory perceptual patterns in each seizure event, such as pitch changes and the volume.

The duration of audio, in turn, is a function of the own duration of the seizure events and the AI compression factor, which relies on the ability of AI to detect those events. From Table 4, it can be seen that the duration of the correctly detected seizures is on average, twice longer than the duration of the missed seizures. The detection of short seizures has been historically challenging for both visual interpretation and automatic detection; the inter-rater agreement is typically lower on short seizures [20], and the performance of the seizure detection algorithms is often impacted by the length of the seizure, showing detection rates of 70% for seizures shorter than 1min vs more than 95% for seizures longer than 4 min [24].

The role of the amplitude in the misdetection of seizure patients is estimated to be larger than the role of duration in this study, with an AUC of 0.861 compared to 0.595 for the duration. The amplitude of the EEG signals is significantly larger for the detected patients. In this sense, the aEEG has proven to be less effective at detecting low-amplitude seizure events [63].

The AI probability plays by far the biggest role in this algorithm as it affects both the duration and the volume, with an AUC of 0.981. It can be seen that it is much lower on the missed patients than that on the detected. While the algorithm relies on AI to start with, it is shown to perform better than AI alone. The human ear's perception can combine the benefits of being exposed to perceptual seizure characteristics in audio such as pitch changes and duration and compensate each other when one fails to convey discrimination.

Finally, more annotators agree on correctly detected seizure patients than on missed seizure patients, with an AUC of 0.752. This indicates that some of the errors of the algorithm are of subjective nature, and some annotators might find them to be correct instead of errors.

2) False alarms

It can be seen from Table 5 that in the wrongly detected seizure patients, more annotators disagree than on the correctly identified non-seizure patients. However, an AUC of that characteristic is much lower than that of the missed seizure patients from Table 4 (0.574 vs 0.752), indicating that this cannot be considered as the main factor in false alarms. The AI probability is lower on the correctly identified non-seizure patients (an AUC of 0.851). Again, the AUC is smaller than that of missed seizures reported in Table 4. False alarms are a product of detecting something abnormal in the audio stream and making a judgment call whether this abnormality is sufficiently different from the background, often forgetting that the pitch evolution must be present for it to be labelled as a seizure. False alarms can be further decreased with more training examples, which will increase the coverage of heard situations and thus increase the listener's confidence and competence. In this study, the results are obtained with virtually no training, only five indicative examples used for training.

D. Parameter tuning during the algorithm development phase

As with any new method, AI sonification requires several parameters to be tuned. The tuning had to be done empirically since it is difficult to analytically assess the influence of internal settings on the resultant audio's perceptual and discriminatory aspects.

It was found that while pitch evolution was indicative of seizures, the presence of constant pitch itself was perplexing to the listeners. The detection of the presence of evolving patterns was improving with more training examples. However, giving more examples to train comes at the cost of removing examples from the test. It also increases the overall survey duration. The main contributing factor to the constant pitch was found to be the presence of seizure-like artifacts, like sweating or ECG artifacts [44]. Incorporating the ECG artifact removal algorithm was proposed to attenuate the ECG interference. The resultant sound contained fewer seizure-like sounding segments. Interestingly, the removal of ECG interference was expected to improve the specificity with fewer false positives. While the specificity was marginally improved, sensitivity's largest gain went from 0.787 to 0.830. This confirms why it is difficult to analytically predict how perceptually the ear will respond to the changes in the algorithm.

Adjusting other parameters that control the duration of the resultant audio, such as the probability distribution or the Min Speed constant, did not result in any increase in Sensitivity or Specificity, while resulting in a significant increase in the audio duration, with the compression factor decreasing from 727.4 to 359.3 for probability, and to 330.1 for Min Speed, showing that the method is relatively robust to parametric change. Conversely, in [18] aEEG was evaluated on non-expert, showing that there is a sensitivity/specificity trade-off as a function of the aEEG resolution (speed, mm/min), with 38/92% sens/spec at low-resolution (1 mm/min) and 55/61% at higher-resolution (5 mm/min), demonstrating a stronger dependence on the parametric display setting of aEEG.

The public EEG dataset used in this study comes in the 18-channel montage. From the results from Table 7, it is clear that the usage of 8-channel montage gives clear performance advantages over the 2-channel montage for both Sensitivity and Specificity. The usage of 18-channel improves sensitivity from 0.830 to 0.851. In fact, an example was found (EEG sample #77) where the seizures are only audible in P3-O1 and C3-P3 channels, which are not included in the standard 8-channel montage. However, being

exposed to 18 sources of information reduces specificity from 0.951 to 0.851. Thus the performance gains were not conclusive to switch to the full 18-channel montage, and we kept the simpler 8-channel montage in the clinical survey.

E. Limitations of the study

A public dataset of neonatal EEG was utilized in this study. The dataset is composed of 79 multichannel 1-2h EEG recordings annotated by three experts. The usage of the dataset allows for complete replication of the results obtained in this study, the reproduction of the resultant sounds by other researchers and assessing the level of agreement with other annotators. Certain limitations are inherited with this dataset. The data are preselected. Further study needs to be conducted to assess the performance in a real clinical environment by retrospectively analyzing long unedited continuous EEG and mimicking the scenario outlined in this study.

The clinical cohort of people comprises experienced healthcare professionals, primarly neonatologists with 10+ years of experience. Participation in the survey was conducted through personal invitation. Because a particular audience was targeted, the resultant clinical cohort is of moderate size. However, the survey requires a certain time commitment, and healthcare professionals are busier than ever, facing the difficulties brought by the pandemic of COVID-19. During the survey design, many initiatives were researched to limit the listening burden by randomly subsampling the sounds for the survey from the dataset or increasing the speed of sonification to reduce the number of audio questions. These initiatives were found to insignificantly shorten the survey duration while posing extra challenges in the analysis of results and test power. In the end, it was agreed to survey all 79 examples. The 23 clinicians who participated in the survey answered all 79 questions. To compensate for the moderate number of participants, the majority vote was utilized to represent the accuracy of the AI-driven sonification. Through the analysis of the results, it was revealed that in 76 out of 79 examples for the decision to be changed between seizure to non-seizure categories, at least two more participants would be required with all suggesting the decision which is opposite to the current majority vote decision. Only in 3 examples, 1 or 2 extra votes could swing the decision to the opposite. This indicates that the majority vote label is stable in the vast majority of examples. Also, 1/3 (33%) of the unstable examples produced a wrong classification after the majority vote, but just 15% of the stable examples turned into a wrong classification.

The non-clinical cohort of 12 participants was mainly designed to provide an indication of the performance obtained with lay users. The non-clinical cohort may not fully understand the underlying physiological processes for diagnosing seizures or the medical consequences of such decisions. The invitation of the non-clinical cohort to participate in the survey was mostly driven by curiosity regarding the level of the performance and the location of the operating point. Twelve participants were sufficient to show a similar level of performance and the existence of a different operating point on the curve of AI-driven sonification.

In the comparison of the AI-driven sonification performance with the EEG annotators, it can be argued that the latter had access to the video and EEG, which allowed for the influence of many observable physiological artifacts whereas those had to be taken care of in an algorithmic manner with the AI-driven sonification.

The study only evaluates the ability of the AI-driven sonification algorithm with the human interpreter to detect the presence or absence of seizures in a given EEG segment of typically 1-2h duration. The algorithm is not designed to detect individual seizure events or report their duration. The latter is much more critical for prognostication purposes which can be done offline and for which pure AI-based seizure detection algorithms can be used.

F. Availability of the sounds, code and AI models

The resultant sounds of the AI-driven sonification algorithm generated from the open-source dataset can be downloaded via github https://github.com/SergiGomezQuintana/EEGsoundSurvey/tree/main/html/audio

VI. Conclusions

A new method that allows for acoustic interpretation of complex neonatal EEG brain signals is proposed. A survey has been conducted to assess the level of accuracy of this method among clinical personnel with no EEG interpretation experience. The overall accuracy of detecting the seizure presence in an EEG recording was measured as Sensitivity of 0.89 and Specificity of 0.78, evaluated on a clinical cohort of twenty-three healthcare professionals. This compares favourably with the reported aEEG accuracy for the same task. In fact, the obtained accuracy is comparable with trained EEG interpreters with years of specialized experience. The method facilitates the review of 1h of EEG in just 5s, while the time typically spent on EEG visual assessment is 2min per hour of EEG. The AI-driven solution outperforms the AI algorithm alone by allowing the human interpreter to leverage both the seizure characteristics and AI-driven attention mechanisms. The intuitiveness behind acoustic interpretation allows this method to be used almost off-the-shelf – just five examples were presented to the participants prior to the survey, implying less than 1min of training for obtaining these results.

The AI-driven sonification shows a great potential to simplify the task of detecting patients suffering from seizures. While the method is mainly oriented towards non-EEG expert medical professionals in the frontline in the NICU, neurologists and neurophysiologists can similarly benefit from acoustic interpretation to speed up visual EEG assessment.

The proposed method serves as a human-computer interface by providing an extra layer of explainability to potentially any AI algorithm. Future research is needed to quantify the added value of acoustical interpretation of EEG as a function of AI algorithm performance.

ACKNOWLEDGEMENTS

We are grateful to all survey participants for their time and suggestions. The authors of this paper acknowledge the funding sources that supported this study: Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No (18/CRT/6223), and Science Foundation Ireland INSIGHT Centre for Data Analytics, Grant No (2/RC/2289-P2).

REFERENCES

[1] WHO. Newborns: improving survival and well-being. *World Health Organization* https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality (2020)

[2] Lawn, J. E., Manandhar, A., Haws, R. A. & Darmstadt, G. L. Reducing one million child deaths from birth asphyxia - A survey of health systems gaps and priorities. *Health Research Policy and Systems* **5**, 1–10 (2007)

[3] Douglas-Escobar, M. & Weiss, M. D. Hypoxic-ischemic encephalopathy: a review for the clinician. *JAMA pediatrics* **169**, 397–403 (2015)

[4] Delanty, N., Vaughan, C. J. & French, J. A. Medical causes of seizures. *Lancet* **352**, 383–390 (1998)

[5] Eriksson, M. & Zetterström, R. Neonatal convulsions incidence and causes in the Stockholm area. *Acta Paediatrica* **68**, 807–811 (1979)

[6] Lanska, M. J., Lanska, D. J., Baumann, R. J. & Kryscio, R. J. A population-based study of neonatal seizures in fayette county, kentucky. *Neurology* **45**, 724–732 (1995)

[7] Ronen, G. M., Penney, S. & Andrews, W. The epidemiology of clinical neonatal seizures in Newfoundland: A population-based study. *Journal of Pediatrics* **134**, 71–75 (1999)

[8] Scher, M. S., Painter, M. J., Bergman, I., Barmada, M. A. & Brunberg, J. EEG diagnoses of neonatal seizures: Clinical correlations and outcome. *Pediatric Neurology* **5**, 17–24 (1989)

[9] McBride, M. C., Laroia, N. & Guillet, R. Electrographic seizures in neonates correlate with poor neurodevelopmental outcome. *Neurology* **55**, 506–513 (2000)

[10] Nagarajan, L., Palumbo, L. & Ghosh, S. Neurodevelopmental outcomes in neonates with seizures: A numerical score of background encephalography to help prognosticate. *Journal of Child Neurology* **25**, 961–968 (2010)

[11] Uria-Avellanal, C., Marlow, N. & Rennie, J. M. Outcome following neonatal seizures. *Seminars in Fetal and Neonatal Medicine* **18**, 224–232 (2013)

[12] Oza, S., Lawn, J. E., Hogan, D. R., Mathers, C. & Cousens, S. N. Neonatal cause-of-death estimates for the early and late neonatal periods for 194 countries: 2000–2013. *Bulletin of the World Health Organization* **93**, 19–28 (2014)

[13] Murray, D. M. *et al.* Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures. *Archives of Disease in Childhood-Fetal and Neonatal Edition* **93**, F187–F191 (2008)

[14] Boylan, G. B., Stevenson, N. J. & Vanhatalo, S. Monitoring neonatal seizures. *Seminars in Fetal and Neonatal Medicine* **18**, 202–208 (2013)

[15] Boylan, G. B., Burgoyne, L., Moore, C., O'Flaherty, B. & Rennie, J. M. An international survey of EEG use in the neonatal intensive care unit. *Acta Paediatrica, International Journal of Paediatrics* **99**, 1150–1155 (2010)

[16] Husain, A. M. Review of neonatal EEG. *Neurodiagnostic Journal* vol. 45 12–35 (2005)

[17] Rakshasbhuvankar, A., Paul, S., Nagarajan, L., Ghosh, S. & Rao, S. Amplitude-integrated EEG for detection of neonatal seizures: a systematic review. *Seizure* **33**, 90–98 (2015)

[18] Rennie, J. M. *et al.* Non-expert use of the cerebral function monitor for neonatal seizure detection. *Archives of Disease in Childhood - Fetal and Neonatal Edition* **89**, F37–F40 (2004)

[19] Zhang, L., Zhou, Y.-X., Chang, L.-W. & Luo, X.-P. Diagnostic value of amplitude-integrated electroencephalogram in neonatal seizures. *Neuroscience Bulletin* **27**, 251–257 (2011)

[20] Stevenson, N. J. *et al.* Interobserver agreement for neonatal seizure detection using multichannel EEG. *Annals of Clinical and Translational Neurology* **2**, 1002–1011 (2015)

[21] Aarabi, A., Wallois, F. & Grebe, R. Automated neonatal seizure detection: A multistage classification system through feature selection based on relevance and redundancy analysis. *Clinical Neurophysiology* **117**, 328–340 (2006)

[22] Temko, A., Thomas, E., Boylan, G., Marnane, W. & Lightbody, G. An SVM-based system and its performance for detection of seizures in neonates. in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009* 2643–2646 (2009)

[23] Thomas, E. M., Temko, A., Lightbody, G., Marnane, W. P. & Boylan, G. B. Gaussian mixture models for classification of neonatal seizures using EEG. *Physiological Measurement* **31**, 1047–1064 (2010)

[24] Temko, A., Thomas, E., Marnane, W., Lightbody, G. & Boylan, G. EEG-based neonatal seizure detection with Support Vector Machines. *Clinical Neurophysiology* **122**, 464–473 (2011)

[25] Boashash, B., Boubchir, L. & Azemi, G. A methodology for time-frequency image processing applied to the classification of nonstationary multichannel signals using instantaneous frequency descriptors with application to newborn EEG signals. *Eurasip Journal on Advances in Signal Processing* **2012**, 1–21 (2012)

[26] Stevenson, N. J. *et al.* An Automated System for Grading EEG Abnormality in Term Neonates with Hypoxic-Ischaemic Encephalopathy. *Annals of Biomedical Engineering* **41**, 775–785 (2012)

[27] Mirowski, P., Madhavan, D., LeCun, Y. & Kuzniecky, R. Classification of patterns of EEG synchronization for seizure prediction. *Clinical Neurophysiology* **120**, 1927–1940 (2009)

[28] Truong, N. D. *et al.* Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural Networks* **105**, 104–111 (2018)

[29] Hussein, R., Palangi, H., Ward, R. K. & Wang, Z. J. Optimized deep neural network architecture for robust detection of epileptic seizures using EEG signals. *Clinical Neurophysiology* **130**, 25–37 (2019)

[30] Roy, Y. *et al.* Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering* **16**, 37 (2019)

[31] O'Shea, A., Lightbody, G., Boylan, G. & Temko, A. Neonatal seizure detection using convolutional neural networks. in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP* vols 2017-Septe 1–6 (2017).

[32] Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy 2021, Vol. 23, Page 18* **23**, 18 (2020)

[33] Su, J., Vargas, D. V. & Sakurai, K. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation* **23**, 828–841 (2019)

[34] Ahmed, R., Temko, A., Marnane, W., Lightbody, G. & Boylan, G. Grading hypoxic–ischemic encephalopathy severity in neonatal EEG using GMM supervectors and the support vector machine. *Clinical Neurophysiology* **127**, 297–309 (2016)

[35] O'Shea, A., Lightbody, G., Boylan, G. & Temko, A. Investigating the Impact of CNN Depth on Neonatal Seizure Detection Performance. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* **2018**, 5862–5865 (2018)

[36] Gomez, S. *et al.* On sound-based interpretation of neonatal EEG. *29th Irish Signals and Systems Conference, ISSC 2018* (2018)

[37] Barrass, S. & Kramer, G. Using sonification. *Multimedia Systems* **7**, 23–31 (1999)

[38] Laennec, R. & Forbes, J. *A Treatise on the Diseases of the Chest, and on Mediate Auscultation*. (1838).

[39] Rose, A. L. & Lombroso, C. T. Neonatal Seizure states. *Pediatrics* **45**, (1970)

[40] Clancy, R. R. & Legido, A. The Exact Ictal and Interictal Duration of Electroencephalographic Neonatal Seizures. *Epilepsia* **28**, 537–541 (1987)

[41] Kitayama, M. *et al.* Wavelet analysis for neonatal electroencephalographic seizures. *Pediatric Neurology* **29**, 326–333 (2003)

[42] Purves, D. *et al.* Neuroscience. in (Sinauer Associates, 2001).

[43] Temko, A., Marnane, W., Boylan, G., O'Toole, J. M. & Lightbody, G. Neonatal EEG audification for seizure detection. in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014* 4451–4454 (2014).

[44] White, D. M. & Van Cott, A. C. EEG artifacts in the intensive care unit setting. *Neurodiagnostic Journal* vol. 50 8–25 (2010)

[45] Boll, S. F. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **27**, 113–120 (1979)

[46] Flanagan, J. L. & Golden, R. M. Phase Vocoder. *Bell System Technical Journal* **45**, 1493–1509 (1966)

[47] Laroche, J. & Dolson, M. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing* **7**, 323–332 (1999)

[48] Temko, A. Estimation of heart rate from photoplethysmography during physical exercise using Wiener filtering and the phase vocoder. in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* vols 2015-Novem 1500–1503 (2015).

[49] McGee, R. & Rogers, D. Musification of Seismic Data. in *International Conference on Auditory Display* 201–204 (2017).

[50] Dolson, M. Phase Vocoder: A tutorial. *Computer Music Journal* **10**, 14–27 (1986)

[51] O'Shea, A., Lightbody, G., Boylan, G. & Temko, A. Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture. *Neural Networks* **123**, 12–25 (2020)

[52] Brown, C. P. & Duda, R. O. Efficient HRTF model for 3-D sound. in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics* (1997).

[53] Algazi, V. R., Duda, R. O., Duraiswami, R., Gumerov, N. A. & Tang, Z. Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America* **112**, 2053–2064 (2002)

[54] Kapralos, B., Jenkin, M. R. & Milios, E. Virtual audio systems. *Presence: Teleoperators and Virtual Environments* **17**, 527–549 (2008)

[55] Middlebrooks, J. C. Sound localization. in *Handbook of Clinical Neurology* vol. 129 99–116 (2015).

[56] Brown, A. D., Stecker, G. C. & Tollin, D. J. The Precedence Effect in Sound Localization. *Journal of the Association for Research in Otolaryngology 2014 16:1* **16**, 1–28 (2014)

[57] Stevenson, N. J., Tapani, K., Lauronen, L. & Vanhatalo, S. A dataset of neonatal eeg recordings with seizure annotations. *Scientific Data* **6**, 1–8 (2019)

[58] McHugh, M. L. *Interrater reliability: the kappa statistic*. *Biochemia Medica* vol. 22 (2012)

[59] Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* **30**, 1145–1159 (1997)

[60] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874 (2006)

[61] Pisani, F. & Pavlidis, E. The role of electroencephalogram in neonatal seizure detection. *Expert review of Neurotherapeutics* **18**, 95–100 (2018)

[62] Haider, B. A. & Bhutta, Z. A. Birth asphyxia in developing countries: current status and public health implications. *Current problems in pediatric and adolescent health care* **36**, 178–188 (2006)

[63] Hellström-Westas, L., Rosén, I., de Vries, L. S. & Greisen, G. Amplitude-integrated EEG Classification and Interpretation in Preterm and Term Infants. *NeoReviews* **7**, e76–e87 (2006)

[64] Pavel, A. M. *et al.* Neonatal Seizure Management: Is the Timing of Treatment Critical? *The Journal of Pediatrics* (2021)

[65] Yoshida, S. *et al.* Setting research priorities to improve global newborn health and prevent stillbirths by 2025. *Journal of Global Health* **6**, (2016)

[66] Kundu, S. AI in medicine must be explainable. *Nature Medicine* **27**, 1328–1328 (2021)

[67] Patel, B. N. *et al.* Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digital Medicine 2019 2:1* **2**, 1–10 (2019)

[68] Schuller, B., Virtanen, T., Riveiro, M., Rizos, G. & Jing, H. Towards Sonification in Multimodal and User-friendlyExplainable Artificial Intelligence; Towards Sonification in Multimodal and User-friendlyExplainable Artificial Intelligence. in *Proceedings of the 2021 International Conference on Multimodal Interaction* (2021).

[69] Liberman, I. Y. & Shankweiler, D. Speech, the Alphabet, and Teaching to Read. in *NIS Conference-on the Theory and Practice of Beginning Reading Instruction, Learning Research and Development Center* (1976).

[70] Olivan, J., Kemp, B. & Roessen, M. Easy listening to sleep recordings: Tools and examples. *Sleep Medicine* **5**, 601–603 (2004)

[71] Khamis, H., Mohamed, A., Simpson, S. & McEwan, A. Detection of temporal lobe seizures and identification of lateralisation from audified EEG. *Clinical Neurophysiology* **123**, 1714–1720 (2012)

[72] Hermann, T. *et al.* Sonifications for EEG data analysis. in *Proceedings of the 2002 International Conference on Auditory Display* (2002).

[73] Baier, G., Hermann, T. & Stephani, U. Event-based sonification of EEG rhythms in real time. *Clinical Neurophysiology* **118**, 1377–1386 (2007)

[74] Loui, P., Koplin-Green, M., Frick, M. & Massone, M. Rapidly Learned Identification of Epileptic Seizures from Sonified EEG. *Frontiers in Human Neuroscience* **8**, 820 (2014)

[75] Parvizi, J., Gururangan, K., Razavi, B. & Chafe, C. Detecting silent seizures by their sound. *Epilepsia* **59**, 877–884 (2018)

[76] Poveda, J., O'Sullivan, M., Popovici, E. & Temko, A. Portable neonatal EEG monitoring and sonification on an Android device. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* 2018–2021 (2017)

[77] O'Sullivan, M. *et al.* Neonatal EEG Interpretation and Decision Support Framework for Mobile Platforms. in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* vols 2018-July 4881–4884 (2018).

[78] Gomez, S. *et al.* An EEG analysis framework through AI and sonification on low power IoT edge devices. in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* 1–4 (2021).

[79] Alain, G. & Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644* (2016)

[80] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE* **104**, 148–175 (2016)

[81] Holzinger, A., Biemann, C., Pattichis, C. S. & Kell, D. B. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923* (2017)

[82] Sartor, G. & Lagioia, F. The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. *European Parliamentary Research Service* 1–100 (2020)

[83] Shellhaas, R. A., Soaita, A. I. & Clancy, R. R. Sensitivity of Amplitude-Integrated Electroencephalography for Neonatal Seizure Detection. *Pediatrics* **120**, 770–777 (2007)

[84] Evans, E., Koh, S., Lerner, J. T., Sankar, R. & Garg, M. Accuracy of amplitude integrated EEG in a neonatal cohort. *Archives of Disease in Childhood - Fetal and Neonatal Edition* **95**, F169–F173 (2010)

[85] Abend, N. S., Dlugos, D. & Herman, S. Neonatal seizure detection using multichannel display of envelope trend. *Epilepsia* **49**, 349–352 (2008)

AUTHOR CONTRIBUTIONS STATEMENT

All authors reviewed and contributed writing the manuscript. S.G., E.P. and A.T. designed the algorithms and performed the engineering experiments. A.OS, A.F., S.G, E.P. and A.T. performed the early reviews of the algorithms which led to the optimisations presented in this paper. A.T. and A.OS provided the AI model used in this paper. A.F. contributed also with medical expertise. S.G. gathered, reviewed and analized the data under the guidance of A.T. and E.P., and prepared the tables and figures.

ADDITIONAL INFORMATION

The authors have no competing interests as defined by Nature Research, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

FIGURE LEGENDS

Figure 1. AI-driven sonification algorithm block diagram.

Figure 2. From top to bottom: Recorded ECG signal, recorded EEG (single channel) and cleaned EEG with ECG artifact removed.

Figure 3. General scheme for input STFT X[n,k] and phase vocoder output STFT Y[m,k].

Figure 4. VSPV examples. (a) Tones of 8, 16 and 32Hz stretched/contracted individually. (b) Chirp signal being stretched continuously.

Figure 5. Precedence effect due to interaural delay.

Figure 6. Aproximate locations on the scalp of the 8 and 18 channel montages for EEG acquisition.

Figure 7. Snapshot of web survey to assess the AI-driven sonification algorithm.

Figure 8. Area under the curve for the AI probabilistic output and sensitivity/specificity of the majority vote of the survey participants.

Figure 9. Fleiss kappa while exchanging one annotator by the AI sonification survey results and its p-values.

TABLE LEGENDS

Table 1. Polar coordinates of electrodes. For simplicity and due to the symetry of the montages, the table just shows the left-hemisphere pairs. The right hemisphere pairs have just the same values with negative sign on the azimuthal angle.

Table 2. Cohen's kappa scores for annotator pairs and AI sonification algorithm (Kappa ± CI95).

Table 3. Confusion matrix for the AI sonificaiton majority vote.

Table 4. Average ± CI95 for duration, amplitude rms and average AI probability for correctly detected and missed seizures. The last row shows AUC score of a given feature (duration, rms, probability) as an indication of how discriminative the feature is on its own for separation between detected and undetected patients.

Table 5. Average ± CI95 for AI probability for correctly detected seizures and missed seizures. Last row shows AUC score as a reference on how related are the features with the errors.

Table 6. The influence of the parameters on the performance of AI sonification, in terms of sensitivity, specificity and compression rate.

Table 7. Performance for different number of EEG channels.