

Article

A Heliumspeech Unscrambling Method Based on Deep Learning with Phonetic Multi-Objective Optimization

Shibing Zhang ^{1,2} , Kenan Zhou ¹ and Yingdong Hu ^{1,*} 

¹ School of Information Science and Technology, Nantong University, Nantong 226019, China; zhangshb@ntu.edu.cn (S.Z.); 2430310023@stmail.ntu.edu.cn (K.Z.)

² School of Artificial Intelligence Applications, Shanghai Institute of Commerce & Foreign Languages, Shanghai 201399, China

* Correspondence: huyd@ntu.edu.cn

Abstract

Saturated diving plays an important role in fields such as navigation operations, ocean development, military oceanography, and maritime rescue and is indispensable for the marine economy. Heliumspeech communication is an essential component of deep-sea saturated diving operations and serves as the sole means of communication in such environments. This paper presents a heliumspeech unscrambling method for saturated diving based on phonetic multi-objective optimization using deep learning. The method consists of a heliumspeech correction network and a heliumspeech unscrambling network. First, a phonetic multi-objective optimization algorithm is used to design the correction network, which reduces the demand for large heliumspeech training datasets. Then, a saturated diving working language heliumspeech corpus is used to train the heliumspeech unscrambling network. Finally, the unscrambling network processes heliumspeech using a cognitive transfer learning algorithm. During unscrambling, the network continuously converts unscrambled heliumspeech into sample data and adds them to a supervised database, forming a closed-loop control system that dynamically adjusts the network parameters to adapt to variations in heliumspeech signals. This approach not only reduces the deep learning neural network's reliance on large training datasets but also enhances unscrambling performance, particularly under dynamically changing diving depths. Simulation experiments demonstrate that the method effectively unscrambles heliumspeech with a low word error rate and fast convergence.

Keywords: heliumspeech communications; heliumspeech unscrambling; deep learning; neural networks; multi-objective optimization; phonetic alphabets



Academic Editors: Hubert Zarzycki and Lukasz Jelen

Received: 8 June 2026

Revised: 19 July 2026

Accepted: 22 July 2026

Published: 28 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

As global resource shortages intensify, the development and utilization of marine resources are becoming increasingly important. Thus, the 21st century is the era of the ocean economy. According to preliminary estimates, humanity will obtain more than 50% of its energy and means of production, essential for human beings, from the ocean. This concept has significantly promoted deep-sea activities and the development of related technologies worldwide, such as deep-sea diving, deep-sea mining, and deep-sea exploration. It has also facilitated the application of saturation diving technology in various fields, such as maritime transportation, marine industry and marine energy [1]. Maritime activities are becoming increasingly frequent and expanding into broader, farther, and deeper areas.

However, they cannot be separated from saturated diving. Due to the peculiarity of deep-sea operations, many tasks cannot be completed by manned submersible or underwater vehicles, instead requiring professional saturation divers to enter the water directly and be exposed to high-pressure environments for direct operations.

Saturated diving is a specialized technique [2] in which, due to physiological reasons, divers must breathe helium–oxygen mixture gas (approximately 95% helium and 5% oxygen) during deep-sea saturated diving operations instead of general air. Further, the deeper the diving depth, the lower the ratio of oxygen. The physical properties of the helium–oxygen gas mixture are very different from conventional air, greatly affecting the pronunciation of divers' vocal chambers. Under this high-pressure environment, their speech becomes severely distorted, resulting in a strange “Donald Duck” sound known as heliumpspeech that is impossible to understand without special corrective technology. This makes communication during saturated diving difficult, directly affecting these operations and even threatening the lives of saturation divers. This is a typical saturated diving heliumpspeech problem [3].

As early as 1962, the U.S. Navy carried out the first saturated diving experiment using a helium–oxygen gas mixture, discovering this voice distortion in heliumpspeech communications [4]. They noticed that the resonant frequency of heliumpspeech increases by a certain proportion. Subsequently, France and Japan implemented the Precontinent and Sealab program, respectively, to simulate sea dwelling, both carrying out exploration on the heliumpspeech communication technology in saturated diving [5]. In 1967, Morrow first reported a heliumpspeech test on the intelligibility of syllable words and the mis-unscrambling of consonants in the case of the helium–oxygen gas mixture at normal atmospheric pressure [6]. It was pointed out that the helium–oxygen gas mixture has completely different sound transmission characteristics from conventional air, which, coupled with high pressure, completely changes the resonance characteristics of saturation divers' vocal tracts, reducing speech intelligibility. In 1968, Giordano et al. studied the relation between the intelligibility of monosyllabic English words and the depth of diving operation (the pressure of the saturated diving) with a helium–oxygen gas mixture. They found that when the depth exceeded 100 m, speech became almost unintelligible [7].

Further research has shown that exploring the mechanism of voice distortion in saturation divers using only complex acoustic models generated by conventional voice is far from sufficient [8]. Lunde studied the formant frequencies, bandwidths, and amplitudes of heliumpspeech in a high-pressure environment using acoustic transmission-line vocal tract models. He modified the classic Fant–Lindquist formula for resonance peak frequency shift and gave the formulas for the bandwidth and amplitude shift of heliumpspeech [9].

Brubaker et al. studied the distortion of English vowels and consonants in a helium–oxygen environment, showing that energy is lost in both, and that the energy loss in consonants is greater [10]; this reduced the understanding of heliumpspeech. Belcher and Hatlestad used linear prediction to compare the resonance peak frequencies and bandwidths of isolated vowels of heliumpspeech at normal atmospheric pressure with those at diving depths of 54 m, 120 m, 300 m, and 500 m, finding that the frequency is shifted nonlinearly and that the bandwidth is widened significantly [11].

Although there are many varieties of heliumpspeech unscrambling technologies, their signal processing methods are mainly divided into two types: time-domain processing and frequency-domain processing [12]. Time-domain processing techniques include playback, digital encoding, and time-domain segment processing, the latter being the most commonly used [13]. It can correct the resonance peak frequency distortion of heliumpspeech and even correct the fundamental frequency distortion. The autocorrelation between heliumpspeech segments can be used to reduce the effect of noise on heliumpspeech unscrambling [14],

and linear prediction and homomorphic filtering can also be used to estimate the acoustic impulse response of heliumpspeech in order to achieve more accurate correction of its spectral envelope [15]. Frequency-domain processing usually includes the frequency subtraction method, vocoder technique, frequency-domain segment processing, and signal analysis method. In the frequency subtraction method, the heterodyne technique is used to subtract the frequency component of the distortion characteristics of heliumpspeech from the original heliumpspeech signal spectrum to reduce the resonance peak frequency [16]. Richards proposed a heliumpspeech enhancement algorithm based on the short-time Fourier transform [17] that can arbitrarily map the heliumpspeech spectrum envelope without modifying the pitch. Duncan and Jack presented another heliumpspeech unscrambling system that consists of residual excitation linear predictive encoders [18]. It uses the time-domain–frequency-domain joint-processing method to improve the clarity and naturalness of the speech decoded.

However, over the past three decades, research on heliumpspeech has encountered bottlenecks. Although heliumpspeech unscrambling technology has been continuously studied, little substantial progress has been made [19,20]. The successful application of artificial intelligence in the industry has brought new opportunities to fully solve this problem. Artificial intelligence algorithms such as deep learning focus more on incremental concept learning and dynamic knowledge processing in dynamic environments, which makes it easier to achieve multi-objective optimization [21,22].

Nayem and Williamson proposed an attention-based enhancement approach that joins learned speech embedding vectors and a speech enhancement module to enhance noisy speech. In the enhancement module, a quantized language model is used to enforce spectral constraints to achieve better perceptual evaluation of speech quality (PESQ) [23].

Krishna et al. explored self-supervised representation learning based on a hidden unit clustering (HUC) framework to improve the speaker invariance of the learned representations. The learned “time–frequency” representations are further processed with long short-term memory (LSTM) layers, and the HUC framework is used to train the model for learning semantically rich speech representations [24].

However, most deep learning algorithms are based on big data and require massive training data (corpora) as the foundation for learning, and traditional neural networks often face challenges such as insufficient feature extraction or overfitting under small-sample conditions. Yue et al. combined a convolutional neural network featuring an efficient channel attention mechanism with an extreme gradient boosting regressor, and proposed a hybrid neural network. In this hybrid network, the powerful feature extraction capability, coupled with the attention mechanism’s ability to focus on key features, enables the network to extract features efficiently. This solves the problems of insufficient feature extraction or overfitting under small-sample conditions [25]. To address the issue of sample scarcity, Chen et al. proposed a spectrogram-based heliumpspeech unscrambling method that consists of an acoustic model and a language model. The spectrogram features of heliumpspeech, such as the fundamental frequency, harmonics, and peaks, are input into the acoustic model, and the heliumpspeech signals are converted to phonetic sequences. The language model then decodes the phonetic sequence into a word sequence using deep learning [26].

Saturated diving is a specialized field where divers have to endure enormous physiological and psychological pressure when living in a saturated diving chamber or working in deep-sea. It is impossible for saturation divers to work or live in a high-pressure helium–oxygen environment for long periods of time, and thus unrealistic to obtain a massive corpus of heliumpspeech recordings. Li et al. designed a heliumpspeech correction method based on generative adversarial networks (GANs) [27] in which an adaptive speech segmen-

tation algorithm and a fusion loss function are introduced to solve the problems of pitch period distortion and formant shift in heliumpspeech. Although some progress has been made in unscrambling and recognizing isolated and continuous words in heliumpspeech using neural networks and deep learning, further research is needed on learning individual characteristic knowledge and semantic unscrambling algorithms for heliumpspeech based on small samples.

According to the characteristics of working language and limited working vocabulary in deep-sea saturated diving operations, we designed a saturated diving heliumpspeech unscrambling method based on deep learning neural networks. It uses the multi-objective optimization algorithm to accelerate the convergence of neural networks and unscramble heliumpspeech more effectively. The main contributions are summarized as follows:

- Based on phonetic alphabets, a finite impulse response (FIR) filter is combined with multi-objective optimization to correct heliumpspeech, reducing the size requirements for heliumpspeech training data.
- The heliumpspeech correction network adopts a phonetic alphabet multi-objective optimization algorithm based on an effective area in order to optimize the correction network parameters, globally optimizing heliumpspeech correction.
- Based on the iterative process of cognitive transfer learning, a closed-loop mechanism is designed to adapt to changes in diving depth, ensuring that the heliumpspeech unscrambling network's performance is superior to its predecessors.

The rest of this paper is organized as follows. Section 2 analyzes heliumpspeech characteristics and saturated diving operation characteristics. Section 3 describes the heliumpspeech unscrambling method. Some simulation results are given to evaluate the method in Section 4. Finally, some conclusions are drawn in Section 5.

2. Analysis of Saturated Diving Heliumpspeech Characteristics

2.1. Analysis of Heliumpspeech Distortion

From a phonetic perspective, heliumpspeech distortion is a result of complex interactions in the vocal tract system. In normal speech, the waveform of the speech signal is determined by three factors: the complex interaction between the shapes of the vocal tract from the vocal cords to the lips and nose, the waveform of the excitation wave injected into the vocal tract, and the speed of the exhaled air. The structure of the vocal tract, controlled by the vocal cords, tongue, jaw, and lips, as shown in Figure 1, determines the acoustic characteristics of human voice and the spectral characteristics of the sound source. All of the acoustic energy from the sound source enters the vocal tract, forms vocal tract resonance, and concentrates the acoustic energy on certain frequencies.

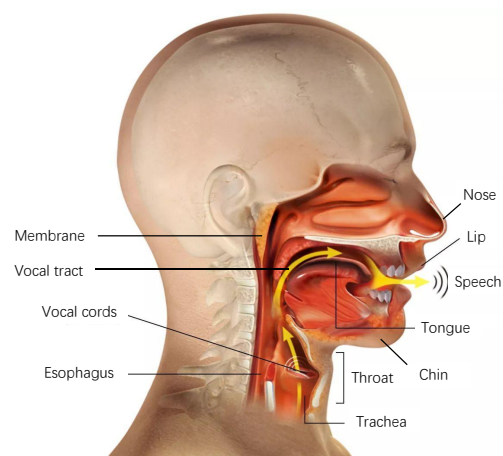


Figure 1. Structure of human vocal organs.

In a high-pressure helium–oxygen environment, the sound produced by saturation divers is affected by both helium gas and high pressure. As is known, the lowest resonance frequency of voices $f_0 = c/\lambda$, where c and λ are the speed and wavelength of the sound in the gas, respectively [28]. The speed of sound in normal air is 240 m/s, while it is 967 m/s in helium gas. As the ratio of helium in the helium–oxygen gas mixture breathed by saturation divers increases, the resonance peak frequency of the divers' voices shifts upwards. On the other hand, when a person is in a high-pressure gas environment, although their vocal speed does not change significantly, their speech clarity decreases with an increase in gas pressure, exhibiting a distinct nasal tone. In a high-pressure environment, the effect of vocal tract wall vibration can no longer be ignored due to the increase in gas density. This coupling between the vocal tract wall and the sound wave causes the resonance peak to shift toward higher frequencies. Under the combined effects of helium gas and high pressure, the resonance peak frequency of saturation divers' voices shifts significantly upward. In addition, the resonance peak bandwidth, the amplitude of high-frequency resonance peaks, and the fundamental frequency also change [19].

2.2. Characterization of Heliumspeech Distortion

Generally, heliumspeech distortion in saturated diving is mainly characterized by the frequency-domain characteristics of speech, such as resonance peak frequency, bandwidth, and amplitude. However, the frequency-domain representation of heliumspeech for each word is different, and the vocal tract of the diver varies with environmental pressure. Thus, it is very difficult to fully characterize heliumspeech using frequency-domain representations.

Phonemes are the smallest units of voice. Based on the articulatory actions of voice syllables, one articulatory action corresponds to one phoneme. Therefore, phonemes are also the smallest linear units of heliumspeech distortion from the perspective of sound quality. Research has shown that when the depth of saturated diving operations increases from 0 to 50 m, the intelligibility of three types of phonemes—stop sound, friction sound, and sliding sound—significantly decreases. The clarity of stop sound decreases the most significantly, followed by friction sound and sliding sound. These changes in heliumspeech provide us with new clues for the extraction and characterization of phoneme characteristics in saturated diving heliumspeech.

2.3. Characterization of Saturated Diving Operation

Saturated diving operations are highly specialized and have two distinct characteristics. First, the pressurization period in the preparation stage of saturated diving operations has a long duration of around 2 days, where the saturation divers are gradually acclimatized to the environment required for saturated diving operations. Second, saturated diving operations have their own industry-specific terminology, in which the vocabulary use is limited.

These two characteristics of saturated diving operations provide a new idea for unscrambling heliumspeech: we can use the phonetic alphabet and working word lexicon to correct heliumspeech and train heliumspeech unscrambling networks, respectively, and, during the pressurization stage of saturated diving operations, we have sufficient time to complete the deep learning training of the heliumspeech unscrambling system.

3. Unscrambling of Heliumspeech

Therefore, we propose a heliumspeech unscrambling system that is composed of a heliumspeech correction network and a heliumspeech unscrambling network, as shown in

Figure 2, where the former completes the pre-correction of saturated diving heliumspeech and the latter completes the unscrambling of the heliumspeech.

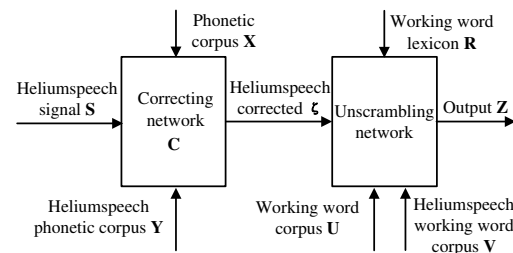


Figure 2. Diagram of heliumspeech unscrambling.

First, according to the working language and regulatory requirements of saturated diving operations, a phonetic alphabet lexicon $\mathbf{P} = \{p_n(t)\}$, which corresponds to the working language, and a commonly used working word lexicon $\mathbf{R} = \{r_m(t)\}$ are established, where $n = 1, 2, \dots, N$, $m = 1, 2, \dots, M$, N is the number of phonetic alphabets in the phonetic alphabet lexicon and M is the number of words in the working word lexicon. The phonetic alphabet lexicon $\mathbf{P}$ is read aloud by divers in both a normal atmospheric environment and the target saturated diving environment in order to generate the standard phonetic alphabet voice corpus $\mathbf{X} = \{x_n(t)\}$ and the heliumspeech phonetic alphabet voice corpus $\mathbf{Y} = \{y_n(t)\}$. $\mathbf{X}$ and $\mathbf{Y}$ are used as the fitting and fitted objectives, respectively, of the multi-objective optimization in the correction network. In the same way, the working word lexicon $\mathbf{R}$ is also read aloud by divers in both environments to generate the normal working word corpus $\mathbf{U} = \{u_m(t)\}$ and the heliumspeech working word corpus $\mathbf{V} = \{v_m(t)\}$.

Secondly, under the given fitting conditions, the correction network parameters $\mathbf{C}$ are set according to the results of the multi-objective optimization based on phonetic alphabets.

Thirdly, we take the normal working word corpus $\mathbf{U}$ as the labeled data and the heliumspeech-signal-corrected ζ as the unlabeled data to train the unscrambling network based on deep learning.

Finally, the heliumspeech correction network and the unscrambling network are joined to unscramble the heliumspeech, while the unscrambling network is further optimized according to the unscrambling results of heliumspeech.

3.1. Correcting Heliumspeech

To correct heliumspeech, we use an FIR filter with L impulse response coefficients as the correction network to compensate the frequency distortion of heliumspeech, as shown in Figure 3. The impulse response of the filter is given by

$$h(t) = \sum_{l=1}^L c_l \delta(t - lT), \quad (1)$$

where c_l , $l = 1, 2, \dots, L$ are the impulse response coefficients of the filter and T is the sampling interval of the input signal of the filter.

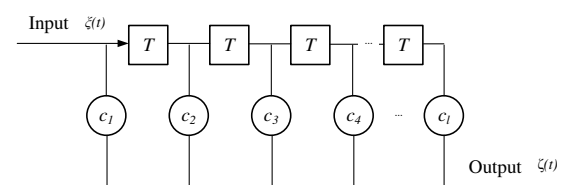


Figure 3. Diagram of FIR filter.

If the input of the filter is $\tilde{\zeta}(t)$, then the output signal of the filter (correction network) is

$$\zeta(t) = \int_0^t \tilde{\zeta}(\tau) h(t - \tau) d\tau, \quad t \geq 0 \quad (2)$$

To compensate for heliumpspeech distortion and optimally correct heliumpspeech, we use a multi-objective optimization algorithm based on the phonetic alphabet in order to design the impulse response coefficients of the filter $\mathbf{C} = \{c_l\}$, enabling the frequency-domain characteristic of the FIR filter to closely approximate the frequency-domain compensation function of heliumpspeech.

To this end, we take the standard phonetic alphabet voice corpus $\mathbf{X}$ as the fitting objective and the heliumpspeech phonetic alphabet voice corpus $\mathbf{Y}$ as the fitted objective in order to optimize the parameters $\mathbf{C}$ of the filter, minimizing the fitting error between $\mathbf{X}$ and ζ as follows:

$$\begin{aligned} &\text{Find : } \mathbf{C} \\ &\text{s.t. } \begin{cases} \text{Objective 1 : } \min \|x(t)_1 - \zeta(t)_1\|_{\mathbf{O}} \\ \vdots \\ \text{Objective } n : \min \|x(t)_n - \zeta(t)_n\|_{\mathbf{O}} \\ \vdots \\ \text{Objective } N : \min \|x(t)_N - \zeta(t)_N\|_{\mathbf{O}} \end{cases} \end{aligned} \quad (3)$$

where

$$\zeta_n(t) = \int_0^t y_n(\tau) h(t - \tau) d\tau, \quad t \geq 0 \quad (4)$$

and $\mathbf{O}$ is the goodness of fit.

In the multi-objective optimization process, we introduce a particle swarm optimization (PSO) learning mechanism in the chemotaxis operation to select the adaptive step size and direction of bacterial movement, and combine the non-dominated sorting and crowded distance sorting mechanism to achieve multi-objective optimization. By designing a global best selection method and boundary control method, a dynamic saving ratio is adopted in the replication operation to both improve the competitiveness of excellent individuals in iteration and enhance the uniformity and diversity of non-dominated solutions in the effective area. Finally, the fuzzy decision is used to select the best compromise for obtaining the global best solution.

Since the optimization in (3) is multi-objective for multiple phonetic alphabet voices, the error surface is more complex, and there may be multiple local optimal solutions. To address this, the effective area is used to achieve efficient optimal solutions in this optimization in order to ensure global optimality [25].

According to the saturation diving operation depth, we set the search area of the optimal solution as

$$\begin{cases} J_{1\min} \leq \|x(t)_1 - \zeta(t)_1\| \leq J_{1\max} \\ J_{2\min} \leq \|x(t)_2 - \zeta(t)_2\| \leq J_{2\max} \\ \vdots \\ J_{N\min} \leq \|x(t)_N - \zeta(t)_N\| \leq J_{N\max} \end{cases} \quad (5)$$

set the effective area for the optimal solution as

$$\begin{cases} J_{1\min} \leq \|x(t)_1 - \zeta(t)_1\| \leq D_1 \\ J_{2\min} \leq \|x(t)_2 - \zeta(t)_2\| \leq D_2 \\ \vdots \\ J_{N\min} \leq \|x(t)_N - \zeta(t)_N\| \leq D_N \end{cases} \quad (6)$$

and set the filtering area for the optimal solution as

$$\begin{cases} J_{1\min} \leq \|x(t)_1 - \zeta(t)_1\| \leq E_1 \\ J_{2\min} \leq \|x(t)_2 - \zeta(t)_2\| \leq E_2 \\ \vdots \\ J_{N\min} \leq \|x(t)_N - \zeta(t)_N\| \leq E_N \end{cases} \quad (7)$$

where $J_{1\min}$, $J_{2\min}$, $\dots$, $J_{N\min}$, $J_{1\max}$, $J_{2\max}$, $\dots$, and $J_{N\max}$ are the search area boundaries of the optimal solution for the optimization; $J_{1\min}$, $J_{2\min}$, $\dots$, $J_{N\min}$, D_1 , D_2 , $\dots$, and D_N are the effective area boundaries of the optimal solution for the optimization; and $J_{1\min}$, $J_{2\min}$, $\dots$, $J_{N\min}$, E_1 , E_2 , $\dots$, and E_N are the filtering area boundaries of the optimal solution for the optimization. The search area is larger than the effective area, and the effective area is larger than the filtering area. Therefore, the evaluation criteria for multi-objective optimization focus on the uniformity and diversity of the Pareto optimal solution set in the effective area; that is, the goal of multi-objective optimization is to concentrate the Pareto optimal solutions on the effective area and distribute uniformly as much as possible. Regarding the detailed Pareto optimal solutions, please refer to reference [29].

Evidently, the selection of effective and filtering areas affects the optimal solution: when the effective area and filtering area are too large, the obtained Pareto solution set may not meet the actual needs of heliumpspeech unscrambling; when they are too small, the solution set may fall into a local optimum. Considering the complexity of practical application, the range of effective and filtering areas is determined based on the parameter settings in the heliumpspeech unscrambling network and the actual range of non-dominated solutions obtained after iteration [29].

3.2. Unscrambling Heliumpspeech

The heliumpspeech unscrambling network in Figure 2 is a deep belief network (DBN), which is a deep neural network based on restricted Boltzmann machines (RBMs). This DBN may be made up of multiple layers of RBMs. During its training process, data are used to infer the hidden layers, which are then treated as the data for the next higher layer; that is, the hidden layer of the previous RBM serves as the visible layer for the next one. This requires that each RBM layer be fully trained before moving on to the next layer, until the final layer is trained. The DBN uses unsupervised learning to retain original characteristics and reduce dimensionality, and then employs supervised learning for fine-tuning to optimize classification. The training process includes layer-by-layer training of RBMs and final adjustment in the BP network.

We use two layers of RBMs and one layer of BP to construct a deep belief network in order to unscramble heliumpspeech, as shown in Figure 4. The DBN trains each RBM layer in sequence to ensure that as much characteristic information as possible is retained when the characteristic vectors are mapped to different characteristic spaces. The BP network takes the output characteristic vector of the last RBM layer as the input characteristic vector for the BP network and trains the entity relationship classifier in a supervised manner.

Meanwhile, the error information is propagated from bottom to top to each RBM layer through the backpropagation network to fine-tune the entire DBN network.

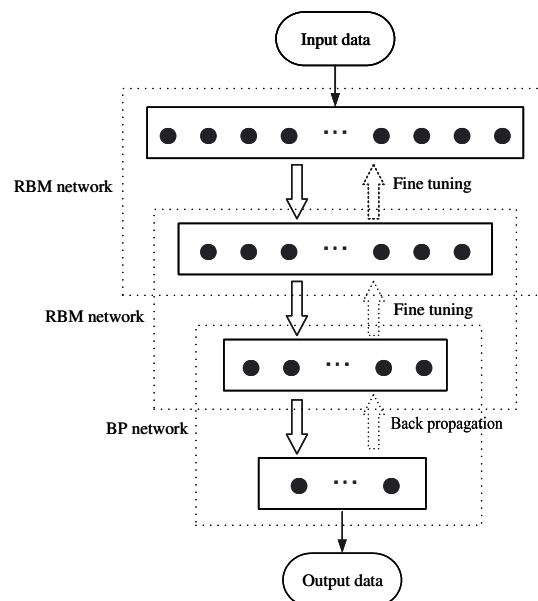


Figure 4. Diagram of deep belief network.

We design a cognitive transfer learning algorithm in the heliumpspeech unscrambling network that can solve the lack of training data (labeled data) for the deep learning of saturated diving heliumpspeech and reduce the effect of dynamic changes in saturated diving operation depth on heliumpspeech unscrambling.

Figure 5 illustrates the binary (Label A and Label B) cognitive transfer learning iterative process that we designed. The heliumpspeech unscrambling network unscrambles the heliumpspeech corrected by the correction network to form an initial decision boundary, as shown in Figure 5a. It then unscrambles and labels the diver's corrected heliumpspeech (unlabeled data) based on the initial decision boundary and updates the label database. Subsequently, the output layer of the heliumpspeech unscrambling network is retrained using the new label data to form a new decision boundary, as shown in Figure 5b. Then, the heliumpspeech unscrambling network continues to unscramble the heliumpspeech (new unlabeled data) using the trained model, as shown in Figure 5c, while continuing to label the unscrambled heliumpspeech and update the label database. The heliumpspeech unscrambling network continues to retrain the output layer based on the new labels to form a new decision boundary (or final decision boundary), as shown in Figure 5d, and repeats this process of heliumpspeech unscrambling. This is a closed-loop process that continues until the diver completes the saturation diving operation and decompresses out of the chamber.

It should be noted that Figure 5 is only a schematic diagram of the iterative process of the cognitive transfer learning algorithm used in the unscrambling network; our actual iteration includes many labels, which correspond to the working word lexicon R , instead of only two.

During heliumpspeech unscrambling, we use the working word lexicon R to filter out incorrect predictions before adding their errors to the training set. If the unscrambled output word is in R , we consider it correct, and the error is fed back to the network; otherwise, the error is not fed back to the network, in order to avoid error propagation.

This closed-loop cognitive transfer learning shown in Figure 5 can transform the heliumpspeech (unlabeled samples) into the labeled data, which then become a part of the training data, effectively addressing the issue of insufficient training data in the cognitive

learning process. Moreover, it can continuously integrate information about dynamic changes in saturation diving depth into the labeled data. Therefore, it enables an adaptive adjustment in heliumpspeech unscrambling parameters.

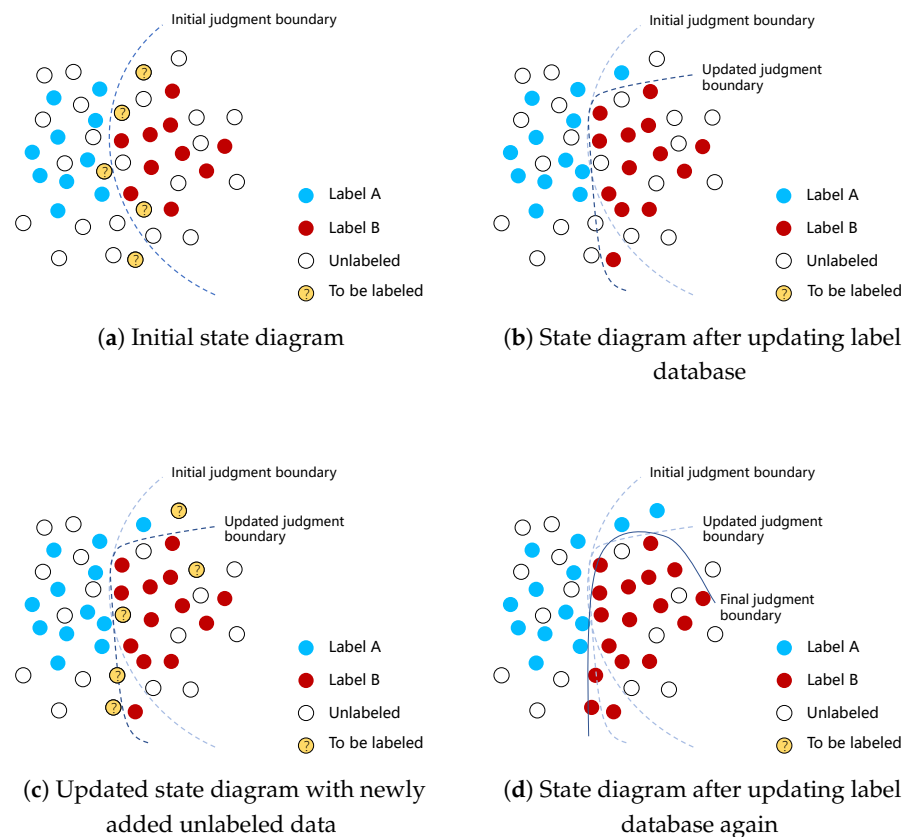


Figure 5. Iterative process of cognitive transfer learning.

3.3. Heliumpspeech Unscrambling Algorithm

The heliumpspeech unscrambling process described above can be divided into two stages: heliumpspeech correction and heliumpspeech unscrambling. In the correction stage, we use a phoneme-based multi-objective optimization algorithm to achieve basic compensation of heliumpspeech; in the unscrambling stage, we use the cognitive transfer learning algorithm to unscramble heliumpspeech. Heliumpspeech unscrambling, however, is a closed-loop system that runs throughout the entire saturated diving operation. During the unscrambling process, the heliumpspeech unscrambling network continuously converts the unscrambled heliumpspeech into sample data, adds them to the training database, and then uses the updated sample database to further train the unscrambling network. This process of unscrambling, updating, and training is repeated cyclically, forming a closed loop. Thus, the heliumpspeech unscrambling algorithm based on deep learning with multi-objective optimization (DLMOO) can be summarized as follows:

Step 1. Construct the heliumpspeech unscrambling system: Select an FIR filter as the correction network and a deep learning neural network as the heliumpspeech unscrambling network to construct a saturated diving heliumpspeech unscrambling system.

Step 2. Construct the phonetic alphabet lexicon: Based on the working language of saturated diving operations, construct a corresponding phonetic alphabet lexicon $\mathbf{P} = \{p_n\}$, where $n = 1, 2, \dots, N$, and N is the number of phonetic alphabets in the phonetic alphabet lexicon.

Step 3. Construct the working word lexicon: According to the requirements of saturated diving operation standards, construct a working word lexicon $\mathbf{R} = \{r_m\}$, where $m = 1, 2, \dots, M$, and M is the number of words in the working word lexicon.

Step 4. Generate the standard phonetic alphabet voice corpus: Under a normal atmospheric environment, the diver reads aloud each phonetic alphabet p_n in the lexicon $\mathbf{P}$ to generate the standard phonetic alphabet voice corpus $\mathbf{X} = \{\mathbf{X}_n\}$.

Step 5. Generate the standard working word corpus: Under a normal atmospheric environment, the diver reads aloud each word r_m in the working word lexicon $\mathbf{R}$ to generate the standard working word corpus $\mathbf{U} = \{\mathbf{U}_m\}$.

Step 6. Generate the heliumspeech phonetic alphabet voice corpus: Under an environment of the saturated diving operation target depth, the diver reads aloud each phonetic alphabet p_n in the lexicon $\mathbf{P}$ to generate the heliumspeech phonetic alphabet voice corpus $\mathbf{Y} = \{\mathbf{Y}_n\}$.

Step 7. Generate the heliumspeech working word voice corpus: Under an environment of the saturated diving operation target depth, the diver reads aloud each word r_m in the working word lexicon $\mathbf{R}$ to generate the heliumspeech working word corpus $\mathbf{V} = \{\mathbf{V}_m\}$.

Step 8. Set the correction network parameters $\mathbf{C}$: Using the heliumspeech phonetic alphabet voice corpus $\mathbf{Y} = \{\mathbf{Y}_n\}$ under the target depth as the input signal for the correction network and the standard phonetic alphabet voice corpus $\mathbf{X} = \{\mathbf{X}_n\}$ as the fitting learning objectives, optimize the parameters $\mathbf{C}$ of the heliumspeech correction network according to (3) under the given phonetic fitting degree $\mathbf{O}$ to minimize the error between $\{\mathbf{X}_n\}$ and $\{\zeta_n\}$.

Step 9. Train the unscrambling network: Using the standard working word corpus $\mathbf{U}$ as the labeled data and the corrected heliumspeech working word corpus $\mathbf{V}$ as the unlabeled data, train the unscrambling network.

Step 10. Unscramble heliumspeech: After the correction network has been configured and the unscrambling network has been trained, use the cognitive transfer learning algorithm to unscramble heliumspeech. The heliumspeech $\mathbf{S}$, spoken by the divers, is corrected by the heliumspeech correction network and then unscrambled by the heliumspeech unscrambling network.

Step 11. Optimize the unscrambling network: Based on the unscrambling results, further optimize the parameters of the heliumspeech unscrambling network until the divers complete the saturated diving operation and exit the cabin.

Step 12. End.

4. Simulation Results and Analysis

4.1. Simulation Parameter Settings

Since no heliumspeech database is publicly available, we used a self-constructed heliumspeech database. We selected English as the saturation diving working language and constructed a phonetic alphabet lexicon $\mathbf{P}$ with 48 international phonetic alphabets for English, including 20 vowels and 28 consonants. According to the requirements of the saturation diving operation specifications of Shanghai Salvage Company, we constructed a working word lexicon $\mathbf{R}$ for saturated diving operations that consists of 500 words and phrases such as “ship, cabin, deck, sea, water, temperature, on, I, pressure,” and so on.

The heliumspeech test data consist of 10 sentences and two texts. Within these 10 sentences are 206 words, with a duration of 2 min and 30 s. Each text is approximately 500 words long and lasts for about 6 min. The heliumspeech test data were collected at saturation diving depths of 200 m, 220 m, and 250 m, respectively, with recordings performed using high-sensitivity microphones and obtained by three professional saturation divers reading the phonetic alphabet lexicon and working word lexicon. The dataset was recorded in WAV format with a 16 kHz sampling rate. Then, these data were segmented into smaller

speech fragments for speech processing and analysis, facilitating model training and testing. The heliumpspeech database is described in more detail in [27].

In our simulation, we set the number of FIR filter impulse response coefficients to $L = 49$, which is larger than the number of phonetic alphabets. The network architecture parameters of the DBN are given in Table 1 and the training hyperparameters are given in Table 2.

Table 1. Network architecture parameters of DBN.

Layer Type	Input	Hidden 1	Hidden 2	Output
Number of Units	2352	1024	512	500
Activation Function	Linear	Sigmoid	Sigmoid	Softmax

Table 2. Training hyperparameters.

Hyperparameter	Pre-Training Learning Rate	Fine-Tuning Learning Rate	Batch Size	Pre-Training Epochs	Fine-Tuning Epochs	Momentum	Weight Initialization	Regularization (L2)
Value	0.01	0.005	48	50	30	0.02	Xavier	0.0005

4.2. Simulation Results Analysis

We simulated our heliumpspeech unscrambling algorithm based on DLMOO and compared it with the heliumpspeech unscrambling algorithms based on deep neural networks (DNNs) [21] and adaptive audio-based metric generative adversarial networks (AMGANs) [27] with the same training and test data. We used the word error rate (WER) to measure the performance of the heliumpspeech unscrambling algorithm as follows:

$$\text{WER} = \frac{\text{Number of error words}}{\text{Total number of words}} \times 100\%. \quad (8)$$

Table 3 shows the unscrambling performances of the three heliumpspeech unscrambling algorithms. We tested heliumpspeech unscrambling for three divers in a saturated diving depth environment of 250 m. It is clear that the WERs of our DLMOO algorithm is the lowest: from the average WERs of three different divers in Table 3, it can be seen that the unscrambling performance of our proposed algorithm is better than the DNN algorithm by 15.92% and AMGAN algorithm by 8.02%. Moreover, these three algorithms exhibit different robustness to divers' heliumpspeech, where that of our algorithm is the highest; that is, our algorithm is the most adaptable to different divers.

Table 4 compares the robustness of three different heliumpspeech unscrambling algorithms to different saturated diving depths, where we simulated heliumpspeech unscrambling at depths of 200, 220, and 250 m. The WERs in Table 4 are the average WERs of three divers' heliumpspeech. It is shown that our proposed algorithm outperforms the other two in different saturated diving depths, and that the unscrambling WER of the DLMOO algorithm is much lower. This indicates that our algorithm has good adaptability to different diving depths.

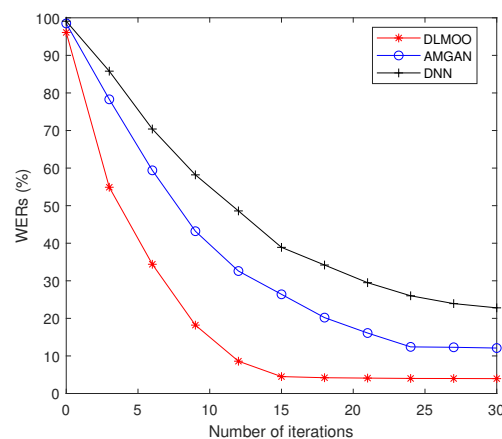
Table 3. Unscrambling WERs (%) of three divers' heliumpspeech with different algorithms.

Algorithms	WER of Diver 1	WER of Diver 2	WER of Diver 3	Average WER	Mean Squared WER
DLMOO	4.15	3.96	4.03	4.05	0.08
AMGAN	10.98	12.54	12.68	12.07	0.77
DNN	18.32	19.12	22.52	19.97	1.82

Table 4. Unscrambling WERs (%) of three different diving depths with different algorithms.

Algorithms	WER of 200 m	WER of 220 m	WER of 250 m	Average WER	Mean Squared WER
DLMOO	3.91	4.01	4.05	3.99	0.06
AMGAN	8.08	10.54	12.07	10.23	1.64
DNN	14.32	16.12	19.97	16.80	2.36

Figure 6 presents the convergence of the DLMOO unscrambling algorithm. As the number of training iterations increases, the WERs of the three algorithms decrease, with that of the DLMOO algorithm decreasing the fastest. It takes only 15 iterations for the WER of DLMOO to stabilize, while it takes 24 iterations for AMGAN and 30 iterations for DNN. The convergence time of our algorithm is less than half that of the other two; this is because we use a deep learning multi-objective optimization algorithm in our heliumpspeech unscrambling, which accelerates convergence of the algorithm and greatly reduces the training period and data requirements. Thus, it is more suitable for machine learning using small sample data, such as saturated diving heliumpspeech unscrambling.

**Figure 6.** Comparison of convergence of three algorithms.

From Tables 3 and 4, and Figure 6, it can be concluded that the heliumpspeech unscrambling algorithm based on DLMOO has the superior unscrambling performance and better robustness to different divers and diving depths.

We also carried out experiments in Chinese under the same conditions, as shown in Tables 5 and 6, which indicate that both experiments result in almost the same heliumpspeech unscrambling performances.

Table 5. Unscrambling WERs (%) of three divers' heliumpspeech with Chinese.

Algorithms	WER of Diver 1	WER of Diver 2	WER of Diver 3	Average WER	Mean Squared WER
DLMOO	4.09	3.98	4.11	4.06	0.06
AMGAN	12.08	12.44	11.59	12.04	0.35
DNN	18.91	19.73	20.48	19.71	0.64

Table 6. Unscrambling WERs (%) of three different diving depths with Chinese.

Algorithms	WER of 200 m	WER of 220 m	WER of 250 m	Average WER	Mean Squared WER
DLMOO	3.97	4.12	4.06	4.05	0.06
AMGAN	9.91	10.62	11.32	10.62	0.58
DNN	15.41	16.61	18.23	16.75	1.16

5. Conclusions

In this paper, we proposed a saturated diving heliumspeech unscrambling method based on deep learning with multi-objective optimization, which includes a heliumspeech correction network and a heliumspeech unscrambling network. We used a multi-objective optimization algorithm based on phonetic fitting to set the parameters of the heliumspeech correction network and reduce heliumspeech training data size requirements. Moreover, we used a cognitive transfer learning algorithm to enable the heliumspeech unscrambling network to better adapt to changes in diving depth. It forms a closed loop to continuously adjust the parameters of the heliumspeech unscrambling network, allowing the unscrambling algorithm to adapt to dynamic changes in heliumspeech signals. This method introduces multi-objective optimization and cognitive transfer learning for heliumspeech unscrambling, reducing learning sample data size requirements and improving the robustness of the heliumspeech unscrambling algorithm to different divers and diving depths. It is especially suitable for heliumspeech unscrambling in a complex saturated diving operating environment.

However, we only trained the network on the 500 words in the working word lexicon, which may reduce the generalization performance of our method for various accents and speaking styles. Our future work should focus on enhancing the dataset's diversity.

Author Contributions: Conceptualization, S.Z.; Methodology, S.Z. and K.Z.; Software, S.Z. and K.Z.; Validation, S.Z. and K.Z.; Formal analysis, S.Z.; Investigation, S.Z.; Writing—original draft, Y.H.; Writing—review & editing, Y.H.; Visualization, Y.H.; Supervision, Y.H.; Funding acquisition, S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China grant number 62371261.

Data Availability Statement: Data available on request due to restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Narayanan, R.; Sreelekshmi, K.; Keerthi, T. Marine resources and sustainable utilization. In *Conservation and Sustainable Utilization of Bioresources; Sustainable Development and Biodiversity*; Springer: Singapore, 2023; pp. 581–596.
2. Fang, Y.Q.; Wu, S.K.; Wang, C.P.; Liu, X.B.; Wang, S.F.; Chen, H.T.; Yu, F.T.; Xiao, W.B. Conventional helium-oxygen diving training in sea and its medical support. *J. Navy Med.* **2007**, *1*, 8–9. [[CrossRef](#)]
3. Hunter, E. Problems of diver communication. *IEEE Trans. Audio Electroacoust.* **1968**, *16*, 118–120. [[CrossRef](#)]
4. Holywell, K.; Harvey, G. Helium Speech. *J. Acoust. Soc. Am.* **1964**, *36*, 210–211. [[CrossRef](#)]
5. Maclean, D.J. Analysis of Speech in a Helium-oxygen Mixture under Pressure. *J. Acoust. Soc. Am.* **1966**, *39*, 1238–1239. [[CrossRef](#)]
6. Morrow, C.T. Speech in Deep-submergence Atmospheres. *J. Acoust. Soc. Am.* **1971**, *50*, 715–728. [[CrossRef](#)] [[PubMed](#)]
7. Giordano, T.A.; Rothman, H.B.; Hollien, H. Helium speech unscramblers—A critical review of the state of the art. *Shengxue Xuebao* **1964**, *1*, 32–40.
8. Lindblom, B.; M, S.K. On the Role of Formant Transitions in Vowel Recognition. *J. Acoust. Soc. Am.* **1967**, *42*, 830–843. [[CrossRef](#)] [[PubMed](#)]
9. Lunde, P. Acoustic transmission-line analysis of formants in hyperbaric helium speech. In *Proceedings of the ICASSP'85, IEEE International Conference on Acoustics, Speech, and Signal Processing*; IEEE: Piscataway, NJ, USA, 1985; pp. 1141–1144.
10. Brubaker, R.S.; Wurst, J.W. Spectrographic Analysis of Divers' Speech during Decompression. *J. Acoust. Soc. Am.* **1968**, *43*, 798–802. [[CrossRef](#)] [[PubMed](#)]
11. Belcher, E.O.; Hatlestad, S. Formant Frequencies, Bandwidths, and Qs in Helium Speech. *J. Acoust. Soc. Am.* **1983**, *74*, 428–432. [[CrossRef](#)] [[PubMed](#)]
12. Giordano, T.; Rothman, H.; Hollien, H. Helium speech unscramblers—A critical review of the state of the art. *IEEE Trans. Audio Electroacoust.* **1973**, *21*, 436–444. [[CrossRef](#)]
13. Stover, W.R. Technique for Correcting Helium Speech Distortion. *J. Acoust. Soc. Am.* **1967**, *41*, 70–74. [[CrossRef](#)] [[PubMed](#)]

14. Suzuki, J.; Nakatsui, M. Translation of helium speech by splicing of autocorrelation function. *J. Radio Res. Lab.* **1976**, *23*, 229–234. [\[CrossRef\]](#)
15. Duncan, G. Correction of the Helium Speech Effect by Short-Time Autoregressive Signal Processing. Ph.D. Thesis, University of Edinburgh, Edinburgh, UK, 1983.
16. Takasugi, T.; Suzuki, J. Translation of helium speech by the use of ‘analytic signal’. *J. Radio Res. Lab.* **1974**, *21*, 61–69.
17. Richards, M. Helium speech enhancement using the short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **2003**, *30*, 841–853.
18. Duncan, G.; Jack, M. Residually excited LPC processor for enhancing helium speech intelligibility. *Electron. Lett.* **1983**, *19*, 710–711. [\[CrossRef\]](#)
19. Cheng, S.; Li, S.; Deng, H. Study on the algorithm of helium speech enhancement. *Appl. Acoust.* **2004**, *23*, 15–19.
20. Zhang, S.; Guo, L.; Li, H.; Bao, Z.; Zhang, X.; Chen, Y. A survey on heliumspeech communications in saturation diving. *China Commun.* **2020**, *17*, 68–79. [\[CrossRef\]](#)
21. Li, D.; Zhang, S.; Guo, L.; Chen, Y. Helium speech correction algorithm based on deep neural networks. In *Proceedings of the 2020 International Conference on Wireless Communications and Signal Processing (WCSP)*; IEEE: Piscataway, NJ, USA, 2020; pp. 99–103.
22. Li, M.; Li, D.; Yan, P.; Zhang, S. Isolated Words Recognition Algorithm Based on Deep Learning in Heliumspeech Unscrambling. In *Proceedings of the 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*; IEEE: Piscataway, NJ, USA, 2021; pp. 1–5.
23. Nayem, K.M.; Williamson, D.S. Attention-based speech enhancement using human quality perception modeling. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 250–260. [\[CrossRef\]](#)
24. Krishna, V.; Sai, T.; Ganapathy, S. Representation learning with hidden unit clustering for low resource speech applications. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 1036–1047. [\[CrossRef\]](#)
25. Yue, J.; Wu, F.; Wang, X.; Feng, P.; Zhuo, J.; Cui, H.; Yan, J.; Duan, S. A novel method for H₂S concentration prediction under small sample based on ECA-1DCNN-XGBR. *IEEE Sens. J.* **2024**, *24*, 20167–20176. [\[CrossRef\]](#)
26. Yonghong, C.; Shibing, Z.; Dongmei, L. Helium speech recognition method based on spectrogram with deep learning. *Big Data Cogn. Comput.* **2025**, *9*, 136. [\[CrossRef\]](#)
27. Li, H.; Chen, Y.; Ji, H.; Zhang, S. A helium speech correction method based on generative adversarial networks. *Big Data Cogn. Comput.* **2024**, *8*, 158. [\[CrossRef\]](#)
28. Van den Berg, J. Myoelastic-aerodynamic theory of voice production. *J. Speech Hear. Res.* **1958**, *1*, 227–244. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Zhang, S.; Ji, X.; Guo, L.; Bao, Z. Multi-objective bacterial foraging optimization algorithm based on effective area in cognitive emergency communication networks. *China Commun.* **2021**, *18*, 252–269. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.