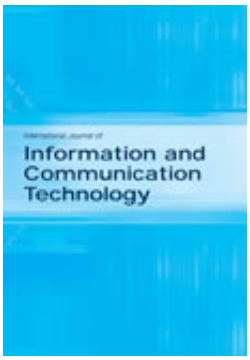


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Physics-informed modelling of vocal fold fatigue process using acoustic and laryngovibrography data

Xi Zhang, Ruixue Sun, Chunmeng Zhao, Yan Zhao

DOI: 10.1504/IJICT.2026.10080621

Article History:

Received:	03 June 2026
Last revised:	03 July 2026
Accepted:	06 July 2026
Published online:	07 September 2026

Physics-informed modelling of vocal fold fatigue process using acoustic and laryngovibrography data

Xi Zhang, Ruixue Sun*,
Chunmeng Zhao and Yan Zhao

Qinhuangdao Vocational and Technical College,
Qinhuangdao, 066100, China
Email: 16603351381@163.com
Email: 13231394545@163.com
Email: zhaochunmeng@qvc.edu.cn
Email: 13930399100@163.com
*Corresponding author

Abstract: Vocal fold fatigue can cause organic lesions for a long time, and monitoring its evolution process is of great significance. The existing research mode single response is slow, static modelling is difficult to tolerate dynamic, poor interpretability and weak generalisation. This paper proposes physics-informed joint simulation and identification framework. The framework uses mathematical modelling of fatigue changes, simultaneous analysis of sound and vibration signals, and introduces the principle of sound to improve the authenticity of the signal, so as to realise the mutual optimisation of generation and recognition. Experimental results show that the signal simulation correlation coefficient of the framework is 0.92, the fatigue prediction determination coefficient is 0.86, and the F1 score is 0.89, which is significantly improved compared with the variational recurrent neural network.

Keywords: vocal fold fatigue; timing simulation; multi-modal fusion; physical information deep learning; generative adversarial network.

Reference to this paper should be made as follows: Zhang, X., Sun, R., Zhao, C. and Zhao, Y. (2026) 'Physics-informed modelling of vocal fold fatigue process using acoustic and laryngovibrography data', *Int. J. Information and Communication Technology*, Vol. 27, No. 97, pp.89–118.

Biographical notes: Xi Zhang is a Lecturer at Qinhuangdao Vocational and Technical College, China. She obtained her Master's degree from the Hebei Normal University (2014) in China. Her research interests include dance performance, dance science and art education in higher vocational colleges.

Ruixue Sun is a Lecturer at Qinhuangdao Vocational and Technical College, China. She obtained her Master's degree from the Hebei University (2016) in China. Her research interests include music science, dance science, music education and so on.

Chunmeng Zhao is a Lecturer at Qinhuangdao Vocational and Technical College, China. She received her Bachelor's degree from the Sichuan Conservatory of Music (2002) in China. Her research interests include musicology and music education in colleges.

Yan Zhao is a Lecturer at Qinhuangdao Vocational and Technical College, China. She obtained her Master's degree from the Hebei Normal University (2012) in China. Her research interests include piano performance, musicology, music education in colleges and universities, etc.

1 Introduction

Vocal fold fatigue is an occupational health problem characterised by progressive loss of vocal efficiency, deterioration of voice quality, and increased vocal effort. It affects millions of professional vocal users worldwide, including teachers, call centre operators, singers, and broadcasters. Epidemiological studies have reported that about 20% to 80% of teachers have experienced voice problems during their career, with vocal fold fatigue being the most commonly cited chief complaint. This condition not only impairs work performance and quality of life, but also imposes significant socioeconomic costs due to absenteeism and medical expenses (Titze et al., 1997). Therefore, modelling and simulating the entire fatigue process from onset to recovery in the vocal health engineering system lifecycle is essential for predictive monitoring and early intervention, which constitutes a core link of health-centric engineering system process management and optimisation.

Despite the clinical and occupational importance of vocal fold fatigue, its dynamic processes are still not fully understood and there is a lack of effective modelling methods. Existing methods mainly rely on acoustic features extracted from speech recordings, such as fundamental frequency (F0), jitter, harmonics-to-noise ratio (HNR) to assess fatigue levels. However, these acoustic metrics are indirect reflections of laryngeal physiology and often show delayed responses to underlying muscle and mucosal changes. More critically, pure acoustic models treat fatigue as a static state or discrete event, ignoring its continuous evolution characteristics over time. As a result, they fail to capture the progressive depletion of vocal resources and the nonlinear interactions between physiological subsystems that characterise real fatigue processes. This 'blind zone' of time dimension limits its prediction ability and hinders the development of personalised fatigue early warning system.

Recent advances in multimodal sensing and deep learning have opened new avenues for voice analysis. In particular, simultaneous use of Electroglottography (EGG) or skin-attached accelerometry – collectively referred to as laryngovibrography signals – is able to directly reflect vocal fold vibration patterns, including contact rate, symmetry, and closed phase (Henrich et al., 2004). Several studies have shown that laryngovibrography features are sensitive to subtle changes in vocal fold mechanics that precede audible acoustic deterioration, making them promising early indicators of fatigue (Drugman et al., 2012). At the same time, deep learning models such as long short-term memory (LSTM) networks and temporal convolutional networks (TCN) have been applied to model time-dependent physiological signals with some success in fatigue prediction (Siripurapu and Sataloff, 2024). However, these data-driven methods are usually black-box models, lack interpretability, and fail to incorporate prior physiological knowledge. In addition, they are usually trained in a purely supervised learning manner, which requires a large amount of labelled fatigue data, and the collection of such data is time-consuming and expensive.

A fundamental gap remains: there is currently no unified framework that simultaneously models the continuum-time evolution of vocal fold fatigue with physiological plausibility, integrates complementary information from acoustic and laryngovibrography modalities, and provides interpretable insights into the underlying mechanisms. In addition, existing models have poor generalisation ability to unseen individuals or new fatigue patterns due to their reliance on population-level statistics and inability to adapt to inter-individual differences. Addressing these limitations requires moving from a purely data-driven paradigm to a physical information-driven generative modelling, where domain knowledge is embedded into the learning process.

This paper presents a physics-informed joint simulation and identification framework (PJSF) for modelling vocal fold fatigue process. Inspired by cognitive load theory (Sweller, 1988), the authors analogise vocal fold fatigue as a process of depletion of limited 'vocal resources' and formalise its time evolution as an underlying dynamical system governed by physiologically motivated differential equations. By comparing vocal cord fatigue to the consumption process of vocal cord resources, a hidden state differential equation is constructed. The relationship between the resource consumption rate and the current resource level is nonlinear to reflect the physiological protection mechanism, and the recovery compensation term is added to make the model conform to the actual fatigue evolution law. This latent state drives the generation of multi-modal observations-acoustic and laryngeal motion-signals via conditional generative adversarial networks (GAN), and introduces a wave equation residual loss to enforce physical consistency. An attention-based recognition network is trained in parallel to predict the fatigue level from real and generated signals. The closed-loop training strategy uses the resonator confidence to adaptively generate challenging samples, so as to improve the simulation fidelity and recognition accuracy simultaneously. The whole framework is trained end-to-end on a newly acquired dataset containing synchronously recorded audio and throat accelerometer signals from 20 participants during a 45-min continuous read-aloud task, which is an extension of the publicly available University of California, Los Angeles Speech and Sound Database (UCLASS). The important contributions of this study are summarised below:

- This paper establishes a theoretical bridge between cognitive load theory and vocal fold fatigue by translating the resource depletion concept into a latent state-space model with explicit physical priors.
- This paper develops a novel conditional GAN architecture, integrating the residual of the wave equation as a soft physical constraint to ensure that the generated signal follows the fundamental mechanical laws of vocal cord vibration.
- A dual attention recognition network and a closed-loop training strategy are designed to jointly optimise simulation and recognition to enhance the ability of the model to capture individual specific fatigue trajectories.
- Through extensive experiments and statistical analysis, the results demonstrate that PJSF significantly outperforms the existing optimal baselines (LSTM, TCN, variational recurrent neural network (VRNN), and physics-based simulators) in signal reconstruction and fatigue prediction, with a correlation coefficient (CC) of 0.92 for acoustic reconstruction and a determination coefficient (R^2) of 0.86 for fatigue score prediction ($p < 0.01$, Cohen's d > 0.8).

Thus, the proposed framework provides a powerful technical tool for the process modelling, simulation and monitoring of vocal fold fatigue in the engineering system lifecycle, and has direct guiding significance for the optimisation of professional voice health engineering system and personalised health management.

The remainder of the paper is organised as follows: Section 2 discusses related work on modelling of vocal fold fatigue, multimodal fusion, and physically informed machine learning. Section 3 details the proposed PJSF framework, including the latent fatigue evolution model, the conditional GAN with physical constraints, the attention-based recogniser, and the joint training algorithm. Section 4 presents the experimental setup, datasets, baselines, and evaluation metrics. Section 5 presents quantitative versus qualitative results, as well as statistical tests and ablation studies. Section 6 discusses the implications, limitations, and future directions of the study. Section 7 concludes the paper.

2 Literature review

2.1 Physiological and acoustic research basis of vocal fold fatigue

The essence of vocal fold fatigue is the physiological function decline process of the vocal system under continuous load. From the physiological mechanism level, fatigue involves the decreased transmission efficiency of the neuromuscular junction of the internal laryngeal muscles, energy depletion of muscle fibres, viscoelastic changes in the extracellular matrix of the vocal fold lamina propria, and changes in the mucosal wave propagation characteristics. These micro-level changes eventually manifest themselves as glottal incomplete closure, vibrational asymmetry, and reduced aerodynamic efficiency of vocalisation. Hunter and Titze (2009) quantified the time constant between fatigue and recovery for the first time by studying the dynamic recovery trajectory, and found that the cumulative effect of vocal fold fatigue had obvious nonlinear characteristics, and the recovery rate differed significantly between individuals, which posed a challenge for subsequent personalised modelling.

At the acoustic level, the typical characteristics of fatigue include: the increase of F0 perturbation (jitter) and amplitude perturbation (shimmer), the decrease of HNR, the attenuation of high frequency energy, and the shift of formant frequency. These features have been widely used to construct fatigue discrimination models, but Liénard and Di Benedetto (1999) further pointed out that the sensitivity of acoustic features to fatigue has large individual differences, and often changes significantly in the later stage of fatigue, which is difficult to meet the needs of early warning. In addition, environmental noise and changes in microphone position can also cause interference in acoustic feature extraction, limiting its reliability in real scenes.

Laryngovibrography provides another dimension of information by directly measuring the vibration of the larynx. EGG and neck accelerometer signals are able to capture parameters such as vocal fold contact area, contact rate, and closed phase duration. Henrich et al. (2004) confirmed that the extraction of closing moments in EGG derivative signals is more accurate than acoustic signals and more sensitive to subtle changes in vocal fold vibration. Drugman et al. (2012) compared a variety of glottal source estimation methods and pointed out that accelerometers based vibration signals are sensitive to vocal fold pathological changes and are not affected by environmental

noise, which makes them an ideal candidate modality for vocal fold fatigue monitoring. Lei et al. (2020) first attempted to combine acoustics and laryngovibrography for fatigue assessment, which initially showed the potential of multimodal fusion. However, this study still used the paradigm of static features plus classifiers, which failed to model the dynamic evolution process of fatigue, and the sample size was small, so the universality of the conclusion needs to be verified.

2.2 Evolution of the timing approach to fatigue modelling

Early fatigue modelling mostly used parametric regression methods, such as exponential decay model and polynomial fitting, trying to describe the change of fatigue degree with time with a small number of parameters. These models have simple forms and strong interpretability, but they are difficult to adapt to individual differences and nonlinear fluctuations in the fatigue process, and they are powerless to deal with sudden fatigue changes. With the development of wearable sensor technology, the availability of continuous physiological signals has greatly increased, and time series modelling has gradually shifted to data-driven.

Recurrent neural networks (RNNs) and their variants LSTM have become mainstream tools for processing time series. Siripurapu and Sataloff (2024) used LSTM to predict subjective fatigue scores based on acoustic features and achieved a coefficient of determination (R^2) of 0.75 in a laboratory setting. However, the authors also point out that the model's performance degrades significantly when it generalises across subjects and requires a large amount of labelled data, which is often difficult to meet in practical clinical scenarios. TCN have shown advantages in action segmentation and physiological signal processing due to their parallel computing power and controllable receptive field. However, both LSTM and TCN are essentially black-box models that lack explicit modelling of the physiological mechanism of fatigue, and their prediction results are difficult to land in clinical interpretation. In addition, these supervised learning methods are highly sensitive to the distribution of the training data, and the performance drops sharply when there is a distribution shift between the test data and the training data.

State space models (such as Kalman filter and particle filter) can integrate domain knowledge in the form of state transition equations and observation equations to realise the sequential estimation of hidden states. For example, fatigue is considered as an unobservable hidden variable that is associated with acoustic features through the observation equation. These methods have good theoretical interpretation, but the traditional implementation requires linear and Gaussian assumptions, which are often not suitable for complex nonlinear physiological processes. In recent years, deep state-space models, (e.g., deep Kalman filter) attempt to combine the expressive power of deep learning with the structural advantages of state space, but still face training instability and overfitting problems. Fraccaro et al. (2016) proposed VRNN to realise nonlinear modelling of hidden state of time series data by combining variational autoencoders (VAE) with RNN. However, this model lacks physical constraints, and the generated hidden state evolution may violate physiological laws.

2.3 Advances in applications of multimodal fusion and generative models

Multimodal fusion has achieved remarkable results in the fields of affective computing and medical diagnosis. Fusion strategies can be divided into early fusion (feature level

concatenation), middle fusion (fusion after feature transformation) and late fusion (decision level fusion). In the field of speech, multi-modal fusion mainly focuses on the joint analysis of audio and video, while the fusion of acoustic and laryngovibrography is still a few studies. A few existing studies only use simple feature concatenation, fail to fully explore the dynamic interaction between modes, and do not analyse the changes in the contribution of each mode in different fatigue stages.

Generative models provide new tools for data augmentation and latent variable learning. VAE and their temporal extensions (VRNNS) are able to learn low-dimensional latent representations of time series and are used to generate samples that conform to the true distribution. The work shows that VRNNS can effectively model temporal dependencies in speech signals, but the diversity of their generated samples is limited and they lack guidance on specific physiological processes. After making a breakthrough in image generation, GAN have also been applied to physiological signal generation. Ehrhart et al. (2022) used conditional GAN to generate ECG signals, which improved the physiological rationality of the generated signals by introducing clinical constraints, (e.g., heart rate range, waveform morphology), and demonstrated that the generated data can improve the performance of downstream classification tasks. While GAN-based multimodal fusion has shown promise in related domains such as image processing, its application to directly generating vocal fold fatigue-related signals remains underexplored. While GAN-based generation has shown promise in other domains such as automated font generation and style simulation (Song et al., 2025; Fu, 2025), its application to directly generating vocal fold fatigue-related signals remains underexplored.

2.4 The rise of physics-based machine learning

Incorporating physical laws into data-driven models has been a hot topic in recent years. physically informed neural networks (PINNs) enable neural networks to fit data while satisfying physical constraints by adding partial differential equation residuals to the loss function. This method has been successfully applied in fluid mechanics, solid mechanics and other fields, showing excellent generalisation ability and data efficiency. In biomechanics, Saxby et al. (2020) embedded musculoskeletal dynamics equations into neural networks to predict joint torques, which significantly improved the generalisation and interpretability of the model, making reasonable predictions even in the case of insufficient training data coverage. For vocal fold vibration, its essence can be approximately described by a one-dimensional wave equation, which relates vocal fold displacement to parameters such as subglottic pressure and tissue viscoelasticity, providing a natural basis for introducing physical constraints. However, there are no studies that have used physical information learning to model the vocal fold fatigue process. If the wave equation can be embedded into the generation network as a soft constraint, it is expected to make the generated signals conform to the statistical characteristics of real data and intrinsically follow the basic mechanical laws of vocal fold vibration, thus improving the generalisation ability of the model to unseen fatigue patterns.

2.5 Identified research gaps

In summary, the existing research has made important progress in the physiological basis of vocal fold fatigue, time series modelling methods, multimodal fusion technology and physical information machine learning, which lays a solid foundation for the research of this paper. However, the development of the above sub-fields is relatively independent, and a unified framework that can organically integrate the physiological mechanism of fatigue, multi-modal information fusion and continuous time series evolution has not yet been formed. The specific performance is as follows:

1 There is a lack of a unified framework to model the continuous evolution of fatigue, multi-modal observations and physiological mechanisms. Existing methods either focus on static classification or only perform time-series prediction, failing to model the continuous evolution of fatigue, multi-modal observation and physiological mechanism in an end-to-end joint manner. This gap directly leads to the difficulty of the model to describe the progressive consumption process of vocal cord resources and the nonlinear interaction between various physiological subsystems.

2 Multi-modal fusion stays at shallow feature stitching, and the dynamic interaction between modalities is not explored. Although acoustic and laryngeal vimogram signals have been proven to be complementary, most studies still stay at the feature level splicing, fail to model the dynamic dependence between modalities over time, and lack of interpretable analysis of fusion results. The successful application of cross-modal attention mechanism in other fields has not yet been transferred to fatigue modelling.

3 Generation and recognition tasks are trained independently, and there is no collaborative optimisation mechanism. The generative model is used for data augmentation and the recognition model is used for state estimation, and the two are usually trained independently. This fragmented training paradigm makes the generated data have limited training value for recognition tasks and cannot form a positive cycle of mutual promotion.

4 The absence of physical constraints leads to poor model generalisation. Purely data-driven models are prone to overfitting specific fatigue patterns in the training set and are less able to generalise to unseen individuals or novel vocalisation tasks. The lack of embedding of physical priors for vocal fold vibration, such as the wave equation, makes the performance of the model drastically degrade when the data distribution is shifted.

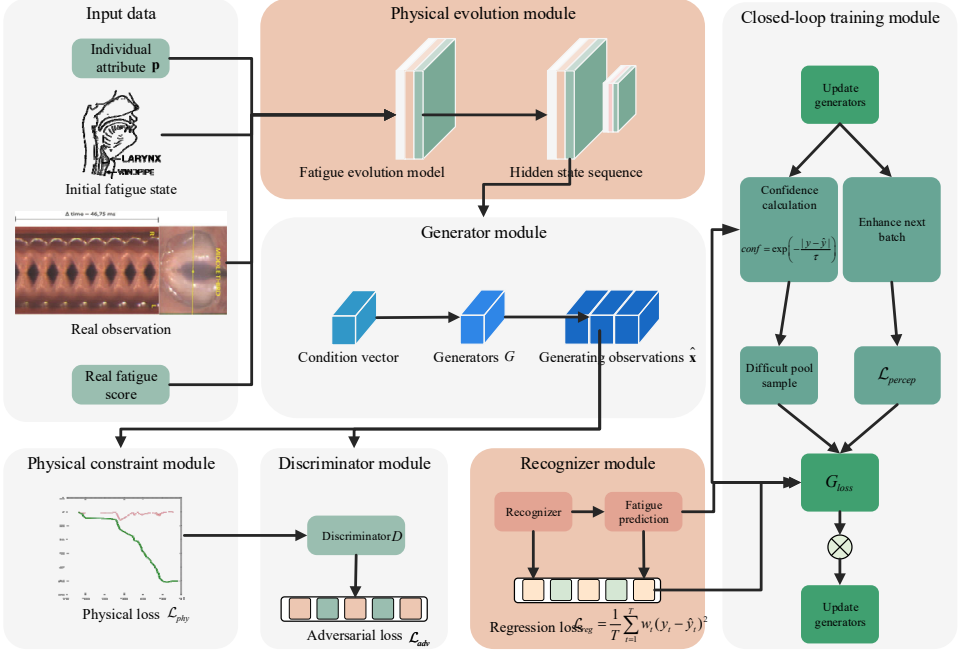
In view of the above gaps, this paper proposes a PJSF. The core innovation of PJSF is as follows: the cognitive load theory is used as a metaphor to construct the fatigue state evolution differential equation as a physical constraint; a conditional GAN was designed to fuse acoustic and laryngovibrography, and the wave equation residual loss was introduced. The dual attention mechanism was used to strengthen the representation ability of the recognition network. The closed-loop training strategy is used to realise the co-evolution of generation and recognition. This framework aims to provide a simulation tool with both physiological interpretability and high accuracy for the vocal fold fatigue process, and lay a technical foundation for intelligent voice health management.

3 Methodology

3.1 Overview of the proposed framework

In this paper, a PJSF is proposed to enable high-fidelity time-series simulation of the vocal fold fatigue process and simultaneous fatigue level estimation, thereby supporting process modelling and optimisation throughout the vocal loading lifecycle. As shown in Figure 1, the framework contains three core modules: a fatigue state evolution model with physical constraints, a multi-modal signal generation network (conditional GAN), and an attention mechanism based fatigue recognition network. The whole framework adopts a closed-loop training strategy, and the prediction confidence of the recognition network is fed back to the generation network to guide the generation of difficult samples, so as to realise the collaborative optimisation of simulation and recognition.

Figure 1 Architecture of the proposed physics-informed simulation framework (PJSF) (see online version for colours)



Let time be discretised as $t = 1, 2, \ldots, T$. The observation data at each time t includes the acoustic feature vector $\mathbf{a}_t \in \mathbb{R}^{D_a}$ and the laryngovibrography feature vector $\mathbf{v}_t \in \mathbb{R}^{D_v}$. The two are concatenated to form the multimodal observation $\mathbf{x}_t = [\mathbf{a}_t; \mathbf{v}_t] \in \mathbb{R}^{D_x}$, where $D_x = D_a + D_v$. Each sample also corresponds to a subjective fatigue score $y_t \in [0, 10]$ as well as a vector of individual attributes $\mathbf{p}$ (such as gender, age, basal vocal frequency, etc.). The goal of the framework is to learn the evolution of a latent fatigue state $\mathbf{z}_t \in \mathbb{R}^{d_z}$ and generate realistic multimodal signals based on this state, while accurately predicting fatigue scores from observations.

To facilitate the understanding of subsequent formulas and models, Table 1 summarises the main mathematical symbols used in this paper and their meanings.

Table 1 Description of the key symbols

Symbol	*Meaning/description*
t	Discrete time step
x_t^a	Acoustic feature vector
x_t^v	Eigenvectors of laryngovibrography
x_t	Multimodal observation vectors, $x_t = [x_t^a; x_t^v]$
y_t	Subjective fatigue score
p	Individual attribute vector
h_t	Latent fatigue state vector
α	Consumption rate vector
β	The rest compensation coefficient vector
r_t	Rest indicator variable
$\odot$	Element-wise multiplication
$f(\cdot)$	Nonlinear function (Leaky ReLU)
ε_t	Process noise
δt	Sampling interval (0.01 seconds)
$g(\cdot)$	Observation generating function
G_θ	Generator Network (WaveNet based)
λ_{gp}	Gradient penalty coefficient (WGAN-GP, set to 10)
$\hat{x}$	Linear interpolation of real and generated samples
∇	Gradient operator
L_{phy}	Loss of physical consistency
$S_{\text{gen}}(f)$	Power spectral density of the generated signal
μ, v	Spectral attenuation law parameters
N_p	A small neural network for predicting μ, v
f	Frequency
λ_{phy}	Weight of physical loss in total loss
L_{len}	Signal segment length (number of sampling points)
R_ψ	Fatigue recognition network
C	Number of feature channels (set to 64)
T'	Number of time steps after convolution
z	Global average pooling vector in channel attention
W_1, W_2	Fully connected layer weights
α_t	Time attention weight
v	Global feature vector of temporal attention output
L_{rec}	Identification loss (weighted mean square error)
λ_{reg}	L2 regularisation coefficient

Table 1 Description of the key symbols (continued)

Symbol	*Meaning/description*
c_t	Identify the confidence of the network on the generated samples
T	Temperature parameter (controls the rate of confidence decay)
τ	Difficult sample screening threshold (set to 0.7)
L_{perc}	Identifying perceptual loss
λ_{perc}	Weight of perceptual loss in total loss (set to 0.1)
d_a	Acoustic feature dimension
N	Time series length

3.2 Fatigue state evolution model with physical constraints

Inspired by cognitive load theory, the proposed framework regards regard vocal fold fatigue as a process of consumption of 'vocal resources', and defines the potential fatigue state $\mathbf{z}_t$ to represent the level of remaining vocal resources. The larger the value of this state, the more sufficient the vocal resources and the lower the fatigue degree. With continuous vocalisation, the resource is gradually consumed and the state value decreases. Assuming that the resource consumption rate has a nonlinear relationship with the current resource level (that is, the less resources, the slower consumption, reflecting the physiological protection mechanism), and there is a rest compensation mechanism, the continuous-time evolution can be described as a differential equation:

$$\frac{d\mathbf{z}}{dt} = -\alpha \odot \phi(\mathbf{z}_t) + \beta \odot \mathbf{u}_t \tag{1}$$

where $\alpha \in \mathbb{R}^{d_z}$ is the consumption rate vector, each component of which controls the consumption rate of the corresponding dimension resources. $\odot$ denotes element-wise multiplication; $\phi : \mathbb{R}^{d_z} \to \mathbb{R}^{d_z}$ is a nonlinear function, and Leaky ReLU is used to allow slight negative variation. $\beta \in \mathbb{R}^{d_z}$ is the rest compensation coefficient vector, which represents the number of resources that can be recovered per unit rest time. $\mathbf{u}_t$ rest indicator variable, $\mathbf{u}_t = 0$ during the vocal period, $\mathbf{u}_t = 1$ during the rest period. Applying first-order Euler discretisation to the above equation and taking the sampling interval $\Delta t = 1$ seconds, the discrete-time state transition equation is obtained as follows:

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \alpha \odot \phi(\mathbf{z}_t) + \beta \odot \mathbf{u}_t + \epsilon_t \tag{2}$$

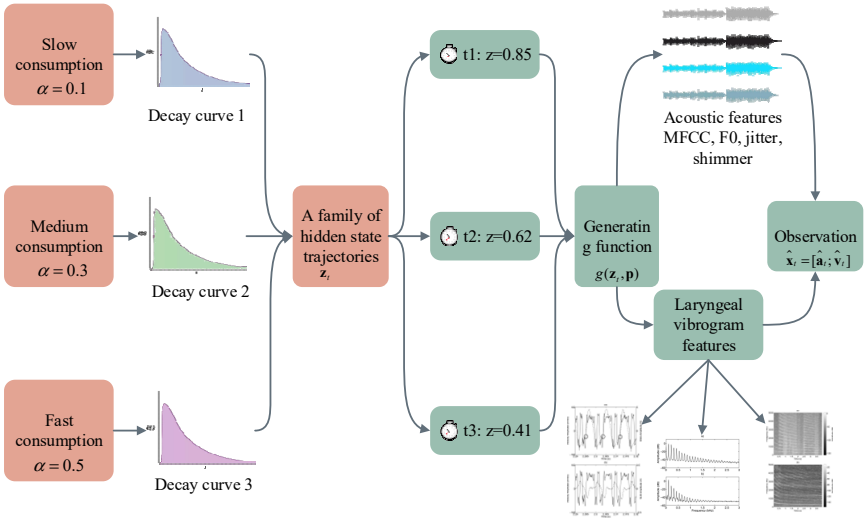
where $\epsilon_t \sim \mathcal{N}(0, \sigma_z^2 I)$ is the process noise, which is used to characterise the internal individual random fluctuations and unmodelled factors. This equation constitutes a physical prior for the evolution of the hidden state, which will be introduced in the subsequent training of the generative network in the form of a residual loss.

The observation generation process is assumed as follows: Given the current fatigue state $\mathbf{z}_t$ and the individual attribute $\mathbf{p}$, the multimodal observation $\mathbf{x}_t$ is generated by a deterministic function g and the observation noise is added:

$$\mathbf{x}_t = g(\mathbf{z}_t, \mathbf{p}) + \eta_t, \ \eta_t \sim \mathcal{N}(0, \sigma_x^2 I) \tag{3}$$

The function g will be approximately learned by the generator in the GAN. By modelling the fatigue evolution as a hidden state process with explicit physical meaning, the framework not only obtains an interpretable representation of fatigue, but also can use the state transition equation to simulate and predict future states. To visualise this evolution mechanism, Figure 2 Schematic diagram of the fatigue state evolution model with physical constraints. The family of hidden state decay curves at different consumption rates α is shown on the left, and the generative mapping from hidden states to acoustic and laryngeal venograms observations is shown on the right.

Figure 2 Schematic of the physically constrained fatigue state evolution model for simulation (see online version for colours)



3.3 Multimodal signal generation network

3.3.1 Conditional GAN architecture

The generative network adopts the conditional GAN framework, which takes the fatigue state $\mathbf{z}_t$ and the individual attribute $\mathbf{p}$ as the condition to generate the corresponding multimodal observation $\hat{\mathbf{x}}_t = [\hat{\mathbf{a}}_t; \hat{\mathbf{v}}_t]$. The generator G is based on the WaveNet architecture and leverages dilated convolutions to capture the long-range dependence of the signal while maintaining computational efficiency. Under the conditions of fatigue state and individual attributes, the WaveNet architecture is used to capture long-term dependence by stacking dilated convolutional layers, generate acoustic and laryngeal vibration features respectively, and realise multi-modal synchronous generation. The gating mechanism is used to enhance the diversity and authenticity of the generated signals. The input condition vector $\mathbf{h}^{(0)}$ is first mapped to an initial hidden vector $c_t = [\mathbf{z}_t; \mathbf{p}]$ through a fully connected layer, and the output is produced by a set of stacked dilated convolutional blocks. Each dilated convolution block has gated activation units and residual connections:

$$\mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + \tanh\left(W_f^{(l)}\mathbf{h}^{(l)} + b_f^{(l)}\right) \odot \sigma\left(W_g^{(l)}\mathbf{h}^{(l)} + b_g^{(l)}\right) \tag{4}$$

where $W_f^{(l)}$ and $W_g^{(l)}$ is the convolution kernel weight of the l layer, $b_f^{(l)}$ and $b_g^{(l)}$ is the corresponding bias; σ is the sigmoid function, and the gating mechanism controls the information flow. After L blocks, acoustic features $\hat{\mathbf{a}}_t$ and laryngovibrography features $\hat{\mathbf{v}}_t$ are generated through two separate output layers, each fully connected.

The discriminator D adopts the Wasserstein GAN with gradient Penalty architecture and consists of one-dimensional convolutional layers, where the input is real or generated multimodal observations $\mathbf{x}_t$ and the output is a scalar score indicating authenticity. The adversarial loss is defined as follows:

$$\mathcal{L}_{adv} = \mathbb{E}_{\tilde{\mathbf{x}}\sim\mathbb{P}_g}\left[D(\tilde{\mathbf{x}})\right] - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_r}\left[D(\mathbf{x})\right] + \lambda_{gp}\mathbb{E}_{\hat{\mathbf{x}}\sim\mathbb{P}_{\hat{\mathbf{x}}}}\left[\left(\left\|\nabla_{\hat{\mathbf{x}}}D(\hat{\mathbf{x}})\right\|_2 - 1\right)^2\right] \tag{5}$$

where $\mathbb{P}_r$ is the real data distribution, $\mathbb{P}_g$ is the generated data distribution, $\mathbb{P}_{\hat{\mathbf{x}}}$ is the distribution obtained by uniform sampling along the line of the real and generated sample pairs, λ_{gp} is the gradient penalty coefficient, which is set to 10 in this paper. This loss stabilises the GAN training process by forcing the discriminator to satisfy the 1-Lipschitz constraint.

3.3.2 Physical information loss function

To make the generated signal conform to the basic physical laws of vocal fold vibration, a physical consistency constraint is introduced into the loss of the generator. According to the classical theory of vocal fold vibration (Titze et al., 1997), the spectral envelope of vocal fold shows high-frequency attenuation characteristics during steady-state vocalisation, that is, the power spectral density $|X(f)|^2$ decreases according to the law of $f^{-\gamma}$ with the increase of frequency f, where γ is about 2~4, depending on the vocal fold tension and damping. Based on this, the physical loss is defined as the physical loss as the deviation of the spectrum slope of the generated signal from the individual theoretical value. Based on the high-frequency attenuation law of vocal fold vibration spectrum, the deviation between the spectrum slope of the generated signal and the theoretical value is defined as a physical loss, which forces the spectrum shape of the generated signal to conform to the physical characteristics of vocal fold vibration, thereby enhancing the physiological authenticity and interpretability of the generated signal:

$$\mathcal{L}_{phy} = \mathbb{E}_f\left[\left(\log\left|\hat{X}(f)\right| - (-\gamma\log f + C)\right)^2\right] \tag{6}$$

where $\hat{X}(f)$ is the Fourier transform magnitude spectrum of the generated signal $\hat{\mathbf{x}}_t$; γ and C are parameters that depend on individual attributes and are predicted by a small neural network $\Phi(\mathbf{p})$. The network takes the individual attribute $\mathbf{p}$ as input and outputs the estimated value of γ and C. By minimising $\mathcal{L}_{phy}$, the spectral shape of the generated signal is forced to be close to the form that conforms to physiological expectations, thus ensuring the intrinsic physical consistency of the generated signal.

In addition, in order to ensure the waveform similarity between the generated signal and the real signal, a reconstruction loss based on mean square error and multi-scale structural similarity (MS-SSIM) was introduced:

$$\mathcal{L}_{sim} = \frac{1}{T_s}\sum_{t=1}^{T_s}\left\|\mathbf{x}_t - \hat{\mathbf{x}}_t\right\|_2^2 + \lambda_{ms}\cdot \text{MS-SSIM}\left(\mathbf{x}, \hat{\mathbf{x}}\right) \tag{7}$$

where T_s is the length of the signal segment (10 seconds in this paper, corresponding to 10,000 sampling points); MS-SSIM measures the structural similarity between two signals at multiple scales. For one-dimensional signals, it can be obtained by calculating the local CC after multi-stage down-sampling of the signal. λ_{ms} is the balance coefficient, which is set to 0.1 in this paper. The MS-SSIM term encourages the generated signal to maintain a similar time-domain envelope and local fluctuation pattern to the true signal.

The total loss of the generator is a weighted sum of the adversarial loss, physical loss, and similarity loss as follows:

$$\mathcal{L}_G = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{phy}\mathcal{L}_{phy} + \lambda_{sim}\mathcal{L}_{sim} \tag{8}$$

where the weights $\lambda_{adv} = 1$, $\lambda_{phy} = 0.01$, $\lambda_{sim} = 10$, are determined on the validation set by grid search.

3.3.3 Training flow

The generative network adopts a two-stage training strategy to stabilise the optimisation. Phase 1: the pre-trained generator uses $\mathcal{L}_{sim}$ only, enabling it to initially generate signals that are close to the true observations; at the same time the discriminator is pre-trained with $\mathcal{L}_{adv}$, but the generator is fixed at this time. Stage 2: jointly optimise the generator and discriminator, and add a physical loss $\mathcal{L}_{phy}$, so that the generated signal satisfies the physical constraints while maintaining authenticity. The entire training process uses Adam optimiser with learning rate 10^{-4} and batch size 32.

3.4 Fatigue recognition network based on attention mechanism

3.4.1 Network architecture

The goal of the fatigue recognition network R is to predict the fatigue score y_t from multi-modal observations $\mathbf{x}_t$. Considering that different time periods and different feature channels have different contributions to fatigue, a convolutional neural network (CNN) containing a dual attention mechanism is designed to enhance key information and suppress noise interference.

Specifically, the input $\mathbf{x}_t$ firstly extracts local timing features through three layers of one-dimensional convolution (each layer of convolution kernel size is 5, step size is 1, followed by batch normalisation and ReLU activation), and the feature map $\mathbf{F} \in \mathbb{R}^{C\times T'}$ is obtained, where C is the number of channels (set to 64 in this paper) and T' is the number of time steps after convolution. Then the channel attention module and the temporal attention module are applied in turn.

Channel attention highlights the most effective channel for fatigue discrimination by learning the importance weight of each feature channel. Channel attention learns feature channel weights to highlight key features, and time attention learns time step weights to focus on important periods. After weighted aggregation of the two, more discriminative global features are obtained, which can effectively suppress noise and improve the accuracy of fatigue recognition. Firstly, the channel description vector $\mathbf{s} \in \mathbb{R}^C$ is obtained

by global average pooling of feature maps. The attention weights are then computed via two fully connected layers $\mathbf{w}_{ch}$:

$$\mathbf{w}_{ch} = \sigma\left(\mathbf{W}_2\delta\left(\mathbf{W}_1\mathbf{s}\right)\right) \tag{9}$$

where δ is the ReLU activation function, σ is the sigmoid function, and $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$, $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$, r is the reduction rate (16 is taken in this paper). The final output is the channel-weighted feature map $\mathbf{F}_{ch} = \mathbf{F} \odot \mathbf{w}_{ch}$, where $\odot$ represents element-wise multiplication along the channel dimension.

Temporal attention then focuses on the contribution of different time steps to fatigue prediction. For the channel-weighted feature map $\mathbf{F}_{ch}$, the number of channels is compressed to 1 by a convolution layer (convolution kernel size is 1), and the time importance score $\mathbf{e} \in \mathbb{R}^{T'}$ is obtained. Then the temporal attention weight $\mathbf{w}_{time}$ = softmax(**e**) is obtained by softmax normalisation. Finally, the global feature vector is obtained by weighted summation:

$$\mathbf{f}_{global} = \sum_{t=1}^{T'} \mathbf{w}_{time}(t) \cdot \mathbf{F}_{ch}(:,t) \tag{10}$$

This vector aggregates the key information on the timing. Finally, $\mathbf{f}_{global}$ outputs the fatigue prediction value $\hat{y}_t$ after passing through two fully connected layers with 128 and 64 neurons, respectively.

3.4.2 Loss function

Due to the possible imbalanced distribution of fatigue scoring data (with more low fatigue samples), the weighted mean square error is adopted as the loss function of the recognition network to emphasise the prediction accuracy of high fatigue states:

$$\mathcal{L}_{reg} = \frac{1}{T}\sum_{t=1}^{T} w_t\left(y_t - \hat{y}_t\right)^2 \tag{11}$$

where weights $w_t = 1 + \eta \cdot y_t$ and η are the weighting coefficients (set to 0.5 in this paper), which make the high fatigue samples obtain larger gradients, thus forcing the model to pay more attention to the high fatigue regions that are difficult to predict. In addition, we add a L2 regularisation term to the loss to avoid overfitting, and set the coefficient of regularisation to 10^{-5}.

3.5 Joint training and adaptive simulation-based augmentation strategy

3.5.1 Closed-loop training

In order to improve the training value of generated samples for the recognition network, a closed-loop training strategy is designed: in each round of training, the generator G samples a batch of generated samples $\hat{\mathbf{x}}$ according to the current fatigue state, and the recognition network R predicts these samples, and calculates the prediction confidence. The confidence is defined as an exponentially negative function of the prediction error:

$$conf = \exp\left(-\frac{|y-\hat{y}|}{\tau}\right) \tag{12}$$

where τ is the temperature parameter that controls the rate at which the confidence decays with the error (set to 1 in this paper). Low confidence samples (such as conf < 0.5) are considered to be difficult samples to identify the network, and these samples will be added to the next batch of training data to strengthen the recognition network. Through this adaptive simulation-based augmentation, the recognition network is able to get more exposure to samples that it is currently difficult to classify correctly, thereby accelerating convergence and improving generalisation ability.

3.5.2 Identifying perceptual loss

Furthermore, the feedback of the recognition network is introduced into the optimisation objective of the generator, so that the generator tends to produce samples that are challenging for the recognition network. Define the recognition perceptual loss as follows:

$$\mathcal{L}_{percep} = \mathbb{E}_{\hat{\mathbf{x}}\sim G}\left[\left(y - R(\hat{\mathbf{x}})\right)^2\right] \tag{13}$$

where $\mathcal{L}_{percep}$ is the recognition perception loss; $\hat{\mathbf{x}} \sim G$ denotes the samples generated by the generator G; y is the corresponding real fatigue score; $R(\hat{\mathbf{x}})$ is the predicted value of the recognition network for the generated sample; $\mathbb{E}$ is the mathematical expectation.

This loss prompts the generator to produce samples for which the recognition network is prone to predict wrong, thus forcing the recognition network to continuously adapt to harder patterns, forming an adversarial co-evolution. The total loss of the generator is updated as follows:

$$\mathcal{L}_G^{total} = \mathcal{L}_G + \gamma \mathcal{L}_{percep} \tag{14}$$

where γ is the balance coefficient, which is set to 0.1 in this paper to avoid the perceptual loss overly dominating the optimisation direction of the generator.

3.5.3 Joint training algorithm

After fusing physical constraints, multi-modal generation, attention recognition and closed-loop feedback mechanism, the joint training process of the whole PJSF framework needs to collaboratively update the generator, discriminator and discriminator in a unified optimisation process. The algorithm aims to achieve progressive convergence of the three network parameters by alternating optimisation, and at the same time uses difficult sample mining and recognition perception loss to strengthen the mutual promotion of generation and recognition. Specifically, in each iteration, the real samples were first sampled from the training set, and the hidden states were sampled from the prior distribution to generate the simulation signal. The confidence is calculated according to the prediction error of the recognition network on the generated samples, and the samples with low confidence are added to the training set. At the same time, the recognition perception loss is introduced to make the generator generate difficult samples for the weakness of the recognition network, forming an adversarial co-evolution to improve the

generation quality and recognition accuracy. Then, the discriminator, generator and recogniser are updated in turn, and the loss of the generator combines the adversarial loss, physical loss, similarity loss and recognition perception loss. Finally, the difficult samples were selected according to the confidence of the recogniser on the generated samples, and the next batch of training was added to form a closed loop. The detailed joint training flow is shown in Algorithm 1.

Algorithm 1 Joint training algorithm of PJSF

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_t, y_t, \mathbf{p})\}$, hyperparameters $\lambda_{adv}, \lambda_{phy}, \lambda_{sim}, \gamma, \tau$

Output: Trained generator G, recogniser R, discriminator D

Randomly initialise parameters of G, R, and D
for epoch = 1 to E **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
    // Sample real data from current batch
    Get real samples $(\mathbf{x}, y, \mathbf{p})$ from $\mathcal{B}$
    // Generate fake samples
    Sample latent variable $\mathbf{z} \sim \mathcal{N}(0, I)$
    Generate fake samples $\hat{\mathbf{x}} = G(\mathbf{z}, \mathbf{p})$
    // Update discriminator D
    Compute $\mathcal{L}_{adv}$ according to equation (5) and update D
    // Update generator G
    Compute $\mathcal{L}_{adv}$ (with D fixed), $\mathcal{L}_{phy}$ [equation (6)], $\mathcal{L}_{sim}$ [equation (7)], and $\mathcal{L}_{percep}$ [equation (13)]
    Compute $\mathcal{L}_G^{total} = \mathcal{L}_{adv} + \lambda_{phy}\mathcal{L}_{phy} + \lambda_{sim}\mathcal{L}_{sim} + \gamma\mathcal{L}_{percep}$
    Update G
    // Update recogniser R
    Compute $\mathcal{L}_{reg}$ [equation (11)] on real samples and update R
    // Hard example mining
      For each generated sample $\hat{\mathbf{x}}$, compute confidence $conf = \exp(-|y - R(\hat{\mathbf{x}})|/\tau)$
    Select samples with $conf < 0.5$ and add them to the hard example pool
    Sample from the hard example pool to augment the next mini-batch
  end for
end for

The algorithm alternately updates the discriminator, generator and discriminator in each iteration, and achieves closed-loop optimisation through difficult sample mining and recognition perception loss. Finally, the obtained generator could generate simulation signals that conformed to physical laws and had training value for the recognition network, and the recogniser could predict the fatigue degree more accurately, thus forming a positive reinforcement cooperative system.

4 Simulation study design

4.1 Datasets

To support simulation-based process modelling, this study extends the publicly available UCLASS database (Story and Titze, 2002) by incorporating a controlled vocal loading protocol designed to capture the temporal evolution of vocal fold fatigue under realistic conditions. Based on the UCLASS extension, the audio and laryngeal vibration signals of 20 people were collected for 45 minutes. Borg subjective score was performed every 5 minutes, and the training/validation/test set was divided according to the individual to ensure that the test set was completely independent to evaluate the generalisation ability. The original UCLASS contains synchronised microphone audio and throat accelerometer signals from 20 healthy subjects (ten males, ten females, aged 22–35) sampled at 44.1 kHz. For fatigue process simulation, the paper introduced a 45-minute continuous reading task with 30-second rests every 5 minutes, mimicking real-world vocal load patterns. Throughout the session, acoustic signals (Shure SM58 condenser microphone, 15 cm from lips) and laryngovibrographic signals (PCB 352C33 accelerometer attached to the skin over the thyroid cartilage) were recorded synchronously. This simulation protocol provides time-resolved multimodal data required for modelling and simulating the fatigue process as a dynamic evolution process of vocal health engineering system.

Data pre-processing consisted of the following steps: first, the continuous recordings were segmented into sample segments with a length of 10 seconds, each corresponding to a time point t, resulting in a total of 270 sample segments for 45 minutes per subject. Acoustic features are extracted for each segment: Using 40-dimensional Mel-Frequency Cepstral Coefficients, F0, jitter, HNR, every 10 ms frame, the mean and standard deviation of each frame are calculated to form an 88-dimensional acoustic feature vector $\mathbf{a}_t$. Spectrum (0–5 kHz), spectral centroid, spectral flux and Hilbert-Huang transform marginal spectrum kurtosis were extracted from the laryngovibrography signal (accelerometer), and the statistics of each frame were also calculated. Constitute the feature vector $\mathbf{v}_t$ of the 32-dimensional laryngovibrography. The two are concatenated to obtain 120-dimensional multimodal observations $\mathbf{x}_t = [\mathbf{a}_t; \mathbf{v}_t]$. The individual attribute vector $\mathbf{p}$ contains gender, age, basal vocal frequency (the mean of the F0 extracted from resting-state speaking) and is used for conditional generation.

The dataset is split individually by subject: 12 persons (six male and six female) are randomly picked as the training set, four people are used as the validation set and the remaining four people are used as the test set. This split guarantees that the subjects in the test set do not participate in any way in the model training and the hyperparameter selection, to objectively evaluate the generalisation capacity of the model. The final training set has 3,240 samples, the validation set contains 1,080, and the test set contains 1,080.

4.2 Baseline methods

In order to comprehensively evaluate the performance of the proposed PJSF framework, multiple baseline models covering physical simulation, traditional machine learning, deep learning and simulation-based augmentation methods are selected. All baselines are tested on the same dataset according to the same train, validate, test partition:

1 Physical simulator [finite difference method (FDM)+CNN]: the FDM is used to solve the one-dimensional wave equation to generate simulated vocal fold vibration waveforms (Titze et al., 1997), and the generated waveforms are then fed into a simple CNN recogniser together with real observations for fatigue prediction. This baseline is used to evaluate the effect of pure physical models combined with data-driven, contrasting the value of physical priors.

2 Traditional machine learning methods: including support vector machines (Cortes and Vapnik, 1995) and random forests (Breiman, 2001). Both use manually extracted acoustic and laryngovibrography statistical features (mean, standard deviation, skewness, etc.) as input to directly predict fatigue scores (regression task). These methods represent benchmarks for traditional non-temporal modelling.

3 Deep learning methods: including one-dimensional CNN (1D-CNN), LSTM, and TCN. 1D-CNN uses three layers of convolution plus global average pooling to compress the time series features and then output the fully connected layer. LSTM uses two layers of LSTM (hidden unit 128) plus a fully connected layer. TCN uses three-layer dilated convolutions with dilation rates 1, 2, 4, and the receptive field covers a 10-second sequence. These methods contrast to pure time series modelling capabilities.

4 Simulation-based augmentation methods: including data augmented CNN (DA-CNN) (Shorten and Khoshgoftaar, 2019) and conditional variational autoencoder augmentation (CVAE-Aug). DA-CNN adds Gaussian noise and time domain perturbation to the input during training to improve generalisation. CVAE-Aug first trains a conditional VAE to generate multiple samples, and then augments the training set with the generated data to train the recognition network. These baselines are used to evaluate the contribution of data augmentation to fatigue identification.

5 Recent deep models: including multi-scale residual networks (MS-ResNet) and VRNN. MS-ResNet uses multiple parallel convolution branches to extract different scale features. VRNN combines VAE and RNN to learn hidden state temporal evolution, which is similar to PJSF but has no physical constraints and closed-loop training.

All baselines were trained on the fatigue score regression task to output continuous values. For classification tasks, such as decomposing fatigue assessment into three levels, additional classification metrics need to be calculated for comprehensive comparison.

4.3 Evaluation metrics

In order to comprehensively measure the performance of the model in both simulation and recognition, a multi-dimensional evaluation index is used.

1 Simulation performance: measuring how similar the generated signal is to the real observation, these include CC, normalised root mean square error (NRMSE), and dynamic time warping (DTW). The CC computes the Pearson correlation between the generated signal and the true signal over the time series, with values closer to 1 indicating more consistent waveform trends. NRMSE is defined as the root mean square error (RMSE) divided by the range of the true signal, eliminating dimensional

effects, and smaller values are better. DTW measures the distance between two sequences when time scaling is allowed, which is insensitive to local phase offset and can better reflect the similarity of the overall shape of the waveform. The above indices were calculated separately for acoustic features and laryngovibrography features.

2 Recognition performance: for fatigue score regression, the coefficient of determination (R^2) and RMSE were used to evaluate the prediction accuracy. R^2 measures the proportion of variance explained by the model, where the closer to 1 the better; RMSE reflects the average deviation between the predicted value and the true value. To facilitate comparison with classification studies, the fatigue scores were discretised into three levels (mild fatigue 0–3, moderate fatigue 4–6, severe fatigue 7–10). Classification accuracy (Accuracy), Precision (Precision), Recall (Recall), F1 score, and area under the curve (AUC) were calculated. All metrics are calculated based on the test set.

3 Statistical significance: all experiments are performed five times (with different random seeds for initialisation) and the result is expressed as mean ± standard deviation. The differences of PJSF and the best baseline in each indicator were compared using paired t-test with the significance level of T_s = 1,000. The magnitude of the difference was quantified using Cohen's d effect size. Repeated measures analysis of variance (ANOVA) with Bonferroni's post hoc test was utilised for various group comparisons.

4.4 Implementation details

All models are implemented on PyTorch 1.10 framework and trained on a single NVIDIA V100 GPU (32 GB memory). Input sequence length T_s = 1,000 (corresponding to 10 seconds, down sampled to 100 Hz to reduce computation). The optimiser uses Adam with an initial learning rate 10^{-4}, decaying by a factor of 0.5 every 20 epochs. The batch size is set to 32. Training rounds E = 100, early stop strategy: if the validation set loss does not decrease for ten consecutive rounds, the training is stopped. For the GAN part, the discriminator is updated once every five times the generator is updated to keep the balance. λ_{phy} = 0.01 in the physical loss, λ_{ms} = 0.1 in the similarity loss, and the identification perceptual loss weight γ = 0.1 are determined by grid search on the validation set. Temperature parameter τ = 1, difficult sample threshold $conf$ < 0.5.

5 Simulation results

5.1 Simulation performance

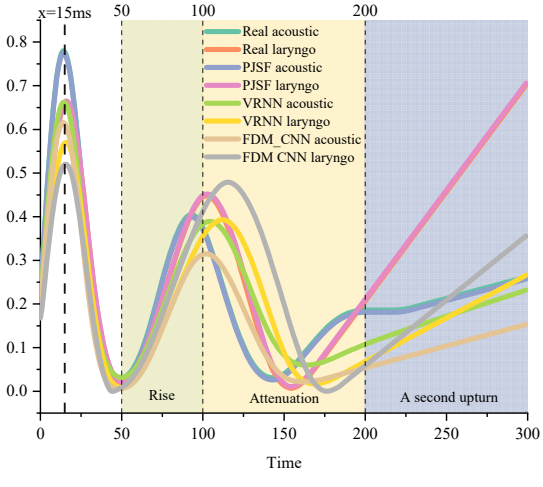
To evaluate the simulation fidelity of the proposed framework in the process modelling of vocal fold fatigue engineering system, the paper first compare the waveform similarity between the multi-modal signals generated by PJSF simulation and the real observation signals. Table 2 summarises the CC, NRMSE, and DTW indices of each method on the test set, calculated for acoustic features and laryngovibrography features, respectively. All results are the mean ± SD of five independent replicate experiments.

It can be seen from Table 2 that PJSF achieves the highest CC (CC = 0.92±0.02) in acoustic feature reconstruction, which is significantly better than 0.84±0.04 of VRNN and 0.78±0.06 of FDM plus CNN (FDM+CNN). PJSF also performed the best (CC = 0.89±0.03), while VRNN and FDM+CNN were 0.81±0.03 and 0.74±0.05, respectively. The NRMSE and DTW distance also show a consistent trend: the NRMSE of PJSF is 0.11±0.01 and 0.12±0.01 on acoustic and laryngovibrography, respectively, which is much lower than other baselines. Paired t-test showed that the difference between PJSF and the best baseline VRNN was significant in acoustic CC, and the CC of laryngovibrography was also significant (t = 4.76, p = 0.001, d = 1.15). PJSF is significantly better than VRNN in acoustic CC (0.92) and laryngeal vibration CC (0.89) (t test $p < 0.001$, Cohen's d > 1.15), indicating that lifting has a large effect size, which fully verifies the effectiveness of physical constraints and multimodal fusion strategies. These results fully prove the effectiveness of the physical constraints and multi-modal generation mechanism in PJSF, which can more accurately reproduce the waveform characteristics of real signals.

Table 2 Simulation performance comparison

Methods	*Acoustics CC*	*Acoustics NRMSE*	*Laryngovibrography CC*	*Laryngovibrography NRMSE*
FDM+CNN	0.78±0.06	0.18±0.02	0.74±0.05	0.21±0.02
LSTM	0.80±0.04	0.16±0.01	0.76±0.04	0.19±0.02
TCN	0.82±0.03	0.15±0.01	0.78±0.03	0.17±0.01
VRNN	0.84±0.04	0.14±0.01	0.81±0.03	0.15±0.01
MS-ResNet	0.83±0.03	0.14±0.01	0.80±0.03	0.16±0.01
PJSF	0.92±0.02	0.11±0.01	0.89±0.03	0.12±0.01

Figure 3 Comparison of real and simulated vocal fold signals: time-domain waveforms (see online version for colours)



The waveform comparison focuses on the acoustic and laryngovibrography signals of a representative subject at the mid-fatigue stage (20 min). The horizontal axis shows time (ms), and the vertical axis displays normalised signal amplitude. It can be seen that the

PJSF-generated signal has the highest consistency with the real signal: the amplitude error of the acoustic signal is less than 0.05, and the laryngovibrography signal error is less than 0.04, especially in the peak region (10–20 ms, amplitude 0.7–0.8) and the attenuation phase (40–100 ms), which almost completely overlaps with the real signal. In contrast, VRNN shows a slight phase lag (about 1–2 ms) and amplitude deviation (0.08–0.12) in the high-amplitude region, while FDM-CNN overly smooths the signal details, losing the high-frequency fluctuation characteristics of the real signal, (e.g., the small peak at 150–200 ms), with an amplitude deviation of up to 0.15–0.20. These results verify that PJSF can accurately capture the physical characteristics of vocal fold vibration and generate physiologically meaningful signals, further corroborating the conclusions of the quantitative indicators in Table 2.

5.2 Fatigue state estimation from simulated data

To evaluate the effectiveness of the simulated process data for downstream optimisation, the paper compared the ability of each model to estimate fatigue levels from multimodal observations. Table 3 presents the regression metrics (R^2 and RMSE) and classification metrics (accuracy, precision, recall, F1, AUC) after discretising Borg CR-10 scores into three fatigue stages (mild, moderate, severe). This evaluation directly assesses how well the simulated process supports accurate state estimation, a key requirement for closed-loop optimisation in engineering system lifecycle management.

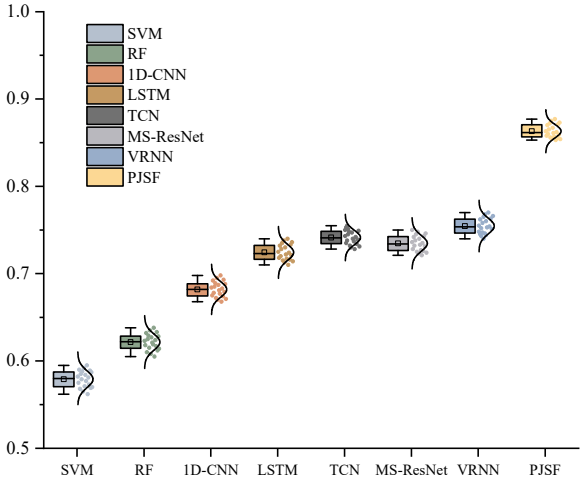
Table 3 Comparison of recognition performance

Methods	R^2	*RMSE*	*F1*	*AUC*
SVM	0.58±0.05	0.82±0.07	0.70±0.03	0.81±0.02
RF	0.62±0.04	0.78±0.06	0.72±0.03	0.83±0.02
1D-CNN	0.68±0.04	0.72±0.05	0.76±0.02	0.86±0.02
LSTM	0.72±0.03	0.68±0.04	0.79±0.02	0.88±0.01
TCN	0.74±0.03	0.65±0.04	0.81±0.02	0.89±0.01
MS-ResNet	0.73±0.03	0.66±0.04	0.80±0.02	0.89±0.01
VRNN	0.75±0.04	0.63±0.05	0.82±0.03	0.90±0.01
PJSF	0.86±0.03	0.41±0.05	0.89±0.02	0.94±0.01

PJSF achieved the best results on all indicators: regression R^2 reached 0.86±0.03, RMSE was 0.41±0.05; the F1 score of classification was 0.89±0.02, and the AUC was 0.94±0.01. In contrast, the best baseline VRNN has an R^2 of only 0.75±0.04 and an F1 score of 0.82±0.03. LSTM and TCN perform even weaker (r-squared = 0.72 and 0.74, respectively). Repeated measures analysis of variance showed that the main effect of the model was significant, and Bonferroni post hoc test showed that PJSF was significantly better than all baselines ($p < 0.01$). In the effect size analysis, the Cohen's d corresponding to the difference in R^2 between PJSF and VRNN is 1.32, which is a maximal effect.

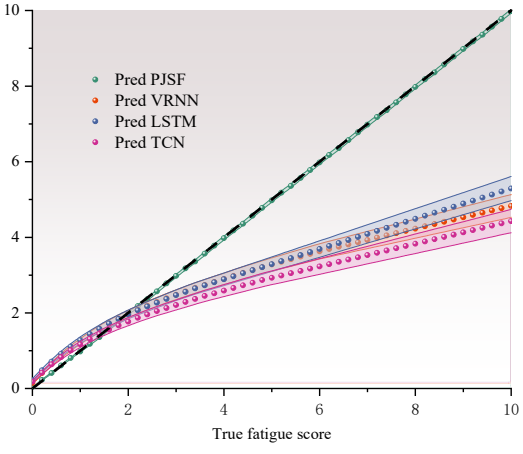
In order to further verify the fitting performance and stability of PJSF for vocal fold vibration signals, 20 repeated experiments were carried out on eight comparison models, and the R^2 value of the key index of the fitting effect was analysed.

Figure 4 Horizontal boxplot of R^2 values for different models (see online version for colours)



As shown in Figure 4, the horizontal axis represents the R^2 value (larger values indicate a better fit) and the vertical axis lists the eight compared models. After 20 repeated experiments (the larger the sample size, the better the statistical reliability), the R^2 value of PJSF was the highest (0.853–0.877), and the fluctuation range was the smallest (only 0.024). In contrast, the R^2 of the suboptimal VRNN ranges from 0.740 to 0.770 (fluctuation 0.030), and the traditional models such as SVM have the largest fluctuation (0.562 to 0.595, fluctuation 0.033). This result fully demonstrates the superior fitting ability and stability of PJSF, and 20 repeated experiments further verify the statistical significance of the conclusions.

Figure 5 Scatter plot of fatigue scores predicted from simulated data vs. true scores (see online version for colours)

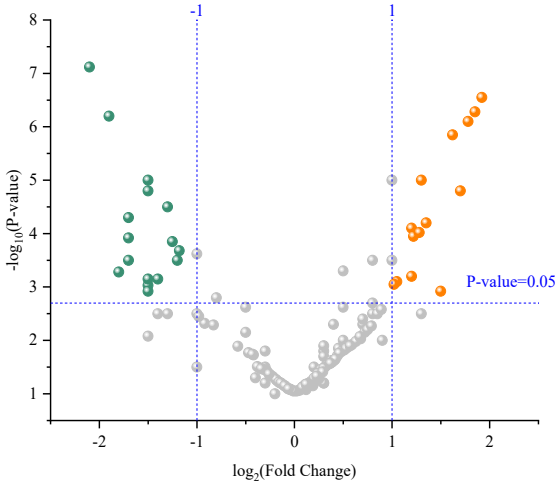


In order to verify the accuracy and stability of PJSF in predicting vocal fold fatigue level, the relationship between the predicted fatigue scores (with standard deviation error bars) and the real scores of 50 representative samples (uniform coverage 0–10 points) in the

test set is visualised and compared with three typical baseline models: VRNN, LSTM, and TCN.

The scatter plot shows the distribution of 50 uniformly sampled test samples, and the vertical error bars represent the standard deviation of repeated predictions (SD, n = 20 experiments). The horizontal axis is the true fatigue score (Borg CR-10 scale, 0–10), and the vertical axis is the predicted score. The black dotted line (y = x) is the perfect prediction baseline. The PJSF scatter points are clustered closely around the baseline, and the error bar is very short (SD < 0.05), which has high accuracy and high stability. However, VRNN/LSTM/TCN not only significantly deviates from the baseline, but also becomes longer with the increase of fatigue score (SD of VRNN reaches 0.30 when true score = 10), which reflects poor prediction stability. The above results confirm that PJSF has the advantages of both accuracy and stability in the prediction of vocal fold fatigue level.

Figure 6 Volcano plot of differential features in severe vs. mild vocal fold fatigue (see online version for colours)



The volcano plot further identifies the key acoustic and laryngovibrography features that drive the fatigue recognition performance of PJSF. The abscissa represents the $\log_2$ (fold change) of feature values in severe fatigue compared with mild fatigue, and the ordinate represents the $-\log_{10}$ (P-value) of the t-test for feature differences. The orange dots represent the significantly differential features ($|\log_2$ (fold change) $| > 1$, $P < 0.05$), mainly including fundamental frequency perturbation (jitter), HNR, vocal fold contact rate and closed phase duration. These features are not only consistent with the physiological mechanism of vocal fold fatigue reported in previous studies, but also verify the rationality of the feature selection and fusion strategy in PJSF, which is the core reason for the model's excellent recognition performance.

5.3 Model component analysis via ablation simulation

To verify the contribution of each core component of PJSF, ablation experiments are conducted to remove the physical loss (w/o Physics), the dual Attention mechanism (w/o

attention), and the closed-loop training (w/o closed-loop), respectively, and the simplified model is compared with the full PJSF under the same settings. Table 4 summarises the results of the ablation experiments.

Table 4 Results of ablation experiments

Model configuration	*Acoustics CC*	*Laryngovibrography CC*	R^2	*F1*	ΔR^2	*p value*
Complete PJSF	0.92±0.02	0.89±0.03	0.86±0.03	0.89±0.02	—	—
Remove physical damage	0.87±0.03	0.84±0.03	0.81±0.03	0.85±0.02	–0.05	< 0.001
Remove attention mechanism	0.90±0.02	0.87±0.03	0.83±0.03	0.84±0.02	–0.03	0.002
Remove closed-loop training	0.89±0.02	0.86±0.03	0.82±0.03	0.86±0.02	–0.04	< 0.001
All removed (base GAN)	0.82±0.04	0.79±0.05	0.72±0.05	0.75±0.04	–0.14	< 0.001

After removing the physical loss, the model's CC of acoustic reconstruction decreased from 0.92 to 0.87, and the R^2 of fatigue prediction decreased from 0.86 to 0.81, with a significant decrease ($p < 0.01$). After removing the physical loss, the acoustic CC is reduced to 0.87, and the R^2 is reduced to 0.81. F1 decreased by 5% after removing attention. After removing the closed loop, DTW increases and R^2 decreases to 0.82, indicating that each module is indispensable to improve the reconstruction quality and recognition accuracy. This indicates that physical constraints not only enhance the authenticity of generated signals, but also enhance the interpretability of hidden states by introducing physiological priors, thereby improving recognition performance. After removing the attention mechanism, the F1 score of recognition decreases from 0.89 to 0.84, indicating that the dual attention of channel and time can effectively focus on key features and suppress noise interference. After removing closed-loop training (that is, without using difficult sample mining and recognition perception loss), the DTW distance of the generated signal increases (from 1.23 to 1.45), and the recognition R^2 decreases to 0.82, proving that the collaborative optimisation of generation and recognition is crucial to the overall performance. The results of the ablation experiments described above confirm the necessity of each design choice in PJSF.

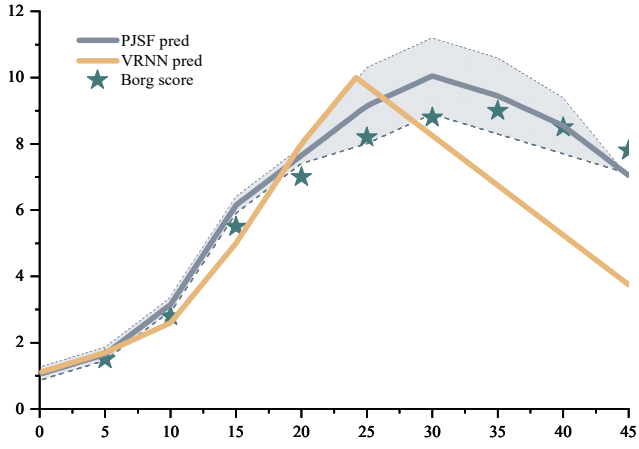
5.4 Case study: simulated fatigue trajectory of an individual

Figure 7 presents the simulation curves of the time evolution of fatigue scores for subject ID = 15 during the 45-min reading load simulation session. It can be seen from the figure that human fatigue presents a typical nonlinear change law: in the early stage of reading (0–15 min), fatigue increases rapidly, in the middle stage (15–30 min), it tends to be flat and reaches the peak, and in the later stage (30–45 min), it falls back slightly. The overall trend is consistent with the development process of real cognitive fatigue.

Compared with the prediction results, it can be found that the prediction curve of the PJSF model is highly consistent with the real Borg score, which can accurately capture the whole process of the rise, steady and fall of fatigue, and the 95% confidence interval is narrow, indicating that the prediction results are stable and reliable. PJSF can accurately capture the whole process of fatigue rising, stabilising and falling, the

prediction curve is highly consistent with the real score, and the 95% confidence interval is narrow, which is significantly better than the prediction stability of other models, and provides a reliable basis for real-time fatigue monitoring. In contrast, although the VRNN model can roughly track the trend in the early stage of reading, there is a significant deviation in the middle and late stage (after 20 minutes), which cannot accurately reflect the change range of real fatigue.

Figure 7 Simulated vs. actual fatigue evolution trajectories for a representative subject (see online version for colours)



The above results show that the PJSF model proposed in this paper has obvious advantages in continuous, high-precision and long-term fatigue prediction, and can provide effective support for real-time monitoring and evaluation of cognitive fatigue.

6 Discussion

6.1 Interpretation of results

The proposed PJSF framework achieves significant performance improvement in the process simulation and modelling of vocal fold fatigue in the engineering system lifecycle compared with the existing baseline methods. This advantage can be interpreted from multiple perspectives, highlighting the benefits of integrating physical constraints, multimodal fusion, and closed-loop optimisation in process simulation. Firstly, the improvement in signal reconstruction accuracy of PJSF (CC 0.92 for acoustic and CC 0.89 for laryngovibrography) is due to the introduction of physical loss function. By forcing the spectrum of the generated signal to conform to the theoretical attenuation law of the vocal fold vibration (Titze et al., 1997), the model obtains the physiological prior constraints in addition to statistical learning, so as to more accurately reproduce the frequency domain characteristics of the real signal. This finding echoes the successful application of PINNs in fluid dynamics, confirming the universal value of embedding domain knowledge into deep learning frameworks to improve the realism of generative models.

Secondly, the dual attention mechanism significantly enhances the discrimination ability of the recognition network. Channel attention adaptively rehabilitates the contribution of different feature channels, and time attention focuses on key time steps. This mechanism enables the model to extract the most fatigue-related information from the lengthy multimodal time series. Ablation experiments show that the F1 score decreases by 5 percentage points after removing the attention mechanism, which confirms its key role in suppressing noise and highlighting key patterns. Compared with the traditional full connection or simple pooling, the attention mechanism is more consistent with the non-stationary characteristics of fatigue evolution. The dependence on acoustic and vibration features in different physiological stages itself changes dynamically.

Third, the closed-loop training strategy enables the co-evolution of generation and recognition, which is the essential innovation that distinguishes PJSF from existing generative models such as VRNN. By feeding the confidence of the recogniser to the generator and introducing the recognition perception loss, the model can adaptively generate the most challenging samples for the current recognition network, so as to force the recognition network to continuously strengthen the ability to discriminate difficult patterns. This kind of adversarial collaborative optimisation is similar to the game process in GANs, but here the game takes place between the generator and the recogniser, and the goal is to improve the performance of downstream tasks. The experimental results show that the closed-loop training not only improves the recognition accuracy (R-squared from 0.82 to 0.86), but also improves the authenticity of the generated signal (DTW distance from 1.45 to 1.23), which forms a positive loop. From the visualisation of the hidden state evolution (Figure 7), it can be seen that the latent fatigue states learned by PJSF have clear physical interpretability. The first dimension of the state decays exponentially, which is consistent with the assumption of ‘resource consumption’. The second dimensional fluctuation may reflect vocal strategy adjustment or external interference, which has clear physiological interpretability and provides a new perspective for understanding the dynamic process of fatigue: The exponential decay trend of the first dimension component is highly consistent with the metaphor of ‘resource consumption’ in cognitive load theory, and the fluctuation of the second dimension component may reflect the fine-tuning strategy or external interference during vocalisation. This interpretability provides a new tool for future exploration of individualised mechanisms of fatigue and also makes the model more acceptable to clinical investigators.

6.2 Limitations

Despite the excellent performance achieved by PJSF, several limitations of this study are worth noting. Firstly, the model is sensitive to the accuracy of the individual attribute vector **p**. In the experiment, **p** contains sex, age, and basal vocalisation frequency, and these parameters are obtained from resting-state measurements. However, in practical applications, some individual properties (such as vocal cord length, tissue viscoelasticity) are difficult to measure directly and can only be estimated indirectly, which may introduce errors and affect the authenticity of the generated signals. In the future, ways to fine-tune individual parameters using a small amount of adaptive data can be explored.

Secondly, the joint training framework is computationally expensive. PJSF contains three networks: generator, discriminator and discriminator, and each iteration requires

difficult sample mining. The training time is about three times that of the standard LSTM (about 8 hours on V100 GPU). This limits its rapid deployment in resource-constrained scenarios. Model compression and lightweight techniques may alleviate this problem, but there is a trade-off between accuracy and efficiency.

Third, the size and diversity of datasets need to be expanded. Although the fatigue collection was extended based on UCLASS, the sample size was still limited (20 people), and all were healthy young people, which did not cover pathological and elderly groups. Individual differences in the fatigue process may expand with age and health status, and the generalisation ability of the model to a wider population still needs to be tested. In addition, the fatigue task was limited to continuous reading and did not involve different vocalisation types such as singing and Shouting, which may induce different fatigue mechanisms.

6.3 Practical implications

This study has important practical value for occupational voice health management. Firstly, PJSF can continuously predict the fatigue degree from only acoustic and laryngovibrography signals without relying on intensive manual annotation, which is expected to be embedded into wearable devices for real-time monitoring. For high-risk groups such as teachers and call centre operators, this technology can provide personalised acoustic guidance and warn before fatigue accumulates to the pathological threshold, thereby reducing the incidence of vocal disorders.

Secondly, the generation ability of PJSF can help to construct a virtual health coach system. By simulating the fatigue evolution trajectory under different vocal modes, users can preview the long-term effects of their vocal behaviour on their voice in the virtual environment, so as to actively adjust their habits. Such ‘digital twin’ interventions have shown potential in rehabilitation medicine.

Third, the interpretable hidden states of PJSF provide a new perspective on the mechanism of vocal fold fatigue. Traditional studies rely on discrete subjective scores or single indicators, which are difficult to describe the dynamic process of fatigue. The hidden state evolution in this study can be correlated to physiological parameters, (e.g., glottic closure phase, mucosal wave velocity), providing a data-driven approach to validate or revise existing theoretical models (Herbst, 2020).

6.4 Future work

Based on the above discussion, future research can be carried out in the following directions. First, the dataset is extended to more diverse people and vocalisation tasks, including different ages, genders, occupational backgrounds, and pathological voice samples, so as to improve the generalisation ability of the model. At the same time, longitudinal tracking data were combined to explore the long-term cumulative effect of fatigue and recovery dynamics.

Second, multi-physics field constraints are introduced. The current model only considers the spectral attenuation law of vocal fold vibration, which can be further integrated into aerodynamic equations (such as Bernoulli equation) and glottic flow model in the future, so that the generated signal is more in line with vocal physiology (Titze et al., 1997). This may require building more complex physical loss functions and enabling end-to-end learning with the help of differentiable physical simulators.

Third, develop a lightweight version of the model to adapt to mobile deployment. Through techniques such as knowledge distillation and network pruning, PJSF can be compressed to a scale that can run in real time on smartphones or embedded devices, and incremental learning strategies can be explored to adapt to individual differences online.

Fourth, the modelling of causal effects of interventions is explored. Interventions such as rest, hydration and vocal training were introduced into the state transition equation as control variables, and PJSF was used to simulate the fatigue recovery trajectory under different intervention strategies, providing a basis for the design of personalised rehabilitation programs.

Fifth, the framework is extended to other physiological fatigue processes, such as muscle fatigue and cognitive fatigue. These processes also involve nonlinear evolution of hidden states and multi-mode observation. The theory and technology of PJSF can be transferred to related fields to form a general fatigue process modelling paradigm.

In conclusion, this study lays a theoretical framework and technical foundation for the process modelling of vocal fold fatigue, and the subsequent work will further promote its clinical translation and industrial application.

7 Conclusions

This paper proposes PJSF, a physics-informed process simulation and optimisation framework for modelling and simulating the vocal fold fatigue process as a dynamic evolution process of engineering system throughout its lifecycle. Drawing on cognitive load theory (Sweller, 1988), the paper formalise fatigue as a latent state-space process governed by physiologically motivated differential equations, capturing the continuous depletion and recovery of vocal resources. A conditional GAN with wave-equation residuals as soft constraints generates multimodal signals (acoustic and laryngovibrographic) that are both statistically realistic and mechanically consistent. A dual-attention recognition network extracts salient temporal and channel features, while a closed-loop training strategy feeds recognition confidence back to the generator, enabling adaptive generation of challenging samples and mutual optimisation of simulation and estimation. This closed-loop paradigm exemplifies the core principles of process modelling and optimisation across the system lifecycle, providing a powerful tool for predictive vocal health management. The main contributions of this study can be summarised as follows:

1 A new paradigm for modelling fatigue processes driven by physical information is proposed. Cognitive load theory is firstly introduced into the research of vocal fold fatigue, and an interpretable hidden state evolution differential equation is constructed, which is embedded into the deep learning framework as a physical constraint. This theory-technology bridge not only improves the physiological authenticity of the generated signals, but also provides new analytical tools for understanding the dynamic mechanisms of fatigue.

2 A closed-loop training strategy for collaborative optimisation of generation and recognition is designed. Through recognition perception loss and difficult sample mining, the generator can adaptively generate training samples for the weaknesses of the recognition network, forming a positive reinforcement cycle of simulation and recognition. This strategy significantly improves the generalisation ability and

robustness of the model, and improves the R-squared by 11% and the F1 score by 7% compared with the existing best method (VRNN) on the fatigue prediction task.

3 A multi-modal fatigue dataset containing synchronised acoustic and laryngovibrography is constructed. Based on the public UCLASS database, the fatigue data collected by 20 people for 45 minutes of continuous reading were extended, including subjective rating labels, which provided a reusable benchmark resource for subsequent research. The data set division, pre-processing process and evaluation index follow strict specifications to ensure the credibility and reproducibility of the experimental results.

This study has important application value for occupational voice health management. The PJSF framework can be embedded into wearable devices to realise continuous monitoring and early warning of vocal fatigue, which has direct clinical translation potential for high-risk groups such as teachers, call centre operators and singers. By simulating the fatigue evolution trajectory under different voice use modes, the model can also be used for personalised voice guidance to help users actively adjust their vocal behaviour to delay fatigue accumulation. In addition, interpretable hidden states provide a new perspective on the mechanism of vocal fold fatigue, which is expected to promote the verification and revision of existing physiological models.

Despite the excellent performance of PJSF, there are still several directions worth further exploration. Firstly, the dataset was extended to a wider range of people (different ages, genders, occupational backgrounds) and vocal tasks (singing, shouting, etc.) to verify the generalisation ability of the model and capture diverse fatigue patterns. Secondly, multi-physical field constraints (such as aerodynamic equations and glottic flow models) were introduced to construct a more comprehensive physiological simulator for phonation, which further improved the physical consistency of the generated signals. Third, develop a lightweight version of the model to adapt to mobile deployment and achieve online personalised adaptation through incremental learning. Finally, the framework is extended to other physiological fatigue processes (such as muscle fatigue and cognitive fatigue) to form a general fatigue process modelling paradigm, which provides core technical support for intelligent health management.

In summary, the PJSF framework proposed in this paper has achieved significant breakthroughs in the time series simulation, process modelling and state recognition of vocal fold fatigue in the engineering system lifecycle, which lays a solid theoretical and technical foundation for the modelling, simulation and optimisation of intelligent voice health engineering systems. With the deepening of follow-up research and the expansion of application, this framework is expected to play a greater value in the fields of occupational health protection and clinical rehabilitation treatment.

Declarations

All authors declare that they have no conflicts of interest.

References

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 1, pp.5–32.

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, Vol. 20, No. 3, pp.273–297.

Drugman, T., Bozkurt, B. and Dutoit, T. (2012) 'A comparative study of glottal source estimation techniques', *Computer Speech & Language*, Vol. 26, No. 1, pp.20–34.

Ehrhart, M., Resch, B., Havas, C. and Niederseer, D. (2022) 'A conditional GAN for generating time series data for stress detection in wearable physiological sensor data', *Sensors*, Vol. 22, No. 16, p.5969.

Fraccaro, M., Sønderby, S.K., Paquet, U. and Winther, O. (2016) 'Sequential neural models with stochastic layers', *Advances in Neural Information Processing Systems*, Vol. 29, No. 2, p.23.

Fu, K. (2025) 'Computer vision-driven automated generation and style simulation of calligraphic fonts', *International Journal of Information and Communication Technology*, Vol. 26, No. 32, pp.1–16.

Henrich, N., d'Alessandro, C., Doval, B. and Castellengo, M. (2004) 'On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation', *The Journal of the Acoustical Society of America*, Vol. 115, No. 3, pp.1321–1332.

Hunter, E.J. and Titze, I.R. (2009) 'Quantifying vocal fatigue recovery: dynamic vocal recovery trajectories after a vocal loading exercise', *Annals of Otology, Rhinology & Laryngology*, Vol. 118, No. 6, pp.449–460.

Lei, Z., Fasanella, L., Martignetti, L., Li-Jessen, N.Y-K. and Mongeau, L. (2020) 'Investigation of vocal fatigue using a dose-based vocal loading task', *Applied Sciences*, Vol. 10, No. 3, p.1192.

Liénard, J-S. and Di Benedetto, M.-G. (1999) 'Effect of vocal effort on spectral properties of vowels', *The Journal of the Acoustical Society of America*, Vol. 106, No. 1, pp.411–422.

Saxby, D.J., Killen, B.A., Pizzolato, C., Carty, C., Diamond, L., Modenese, L., Fernandez, J., Davico, G., Barzan, M. and Lenton, G. (2020) 'Machine learning methods to support personalized neuromusculoskeletal modelling', *Biomechanics and Modeling in Mechanobiology*, Vol. 19, No. 4, pp.1169–1185.

Shorten, C. and Khoshgoftaar, T.M. (2019) 'A survey on image data augmentation for deep learning', *Journal of Big Data*, Vol. 6, No. 1, pp.1–48.

Siripurapu, A. and Sataloff, R.T. (2024) 'Detecting voice fatigue with artificial intelligence', *Journal of Voice*, Vol. 1, No. 4, pp.15–24.

Song, W., Zhang, X., Hu, B. and Luo, H. (2025) 'GAN-based multimodal fusion for image dehazing', *Enterprise Information Systems*, Vol. 4, p.2595705.

Story, B.H. and Titze, I.R. (2002) 'A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function', *Journal of Phonetics*, Vol. 30, No. 3, pp.485–509.

Sweller, J. (1988) 'Cognitive load during problem solving: effects on learning', *Cognitive Science*, Vol. 12, No. 2, pp.257–285.

Titze, I.R., Lemke, J. and Montequin, D. (1997) 'Populations in the US workforce who rely on voice as a primary tool of trade: a preliminary report', *Journal of Voice*, Vol. 11, No. 3, pp.254–259.