

RESEARCH ARTICLE

Neural Cough Counter: A Novel Deep Learning Approach for Cough Detection and Monitoring

ZONGYAO FENG¹, KONSTANTIN MARKOV¹, (Member, IEEE),
JUNPEI SAITO², AND TOMOKO MATSUI³, (Senior Member, IEEE)

¹Division of Information Systems, The University of Aizu, Aizuwakamatsu, Fukushima 965-0006, Japan

²Department of Pulmonary Medicine, Fukushima Medical University, Fukushima 960-1247, Japan

³The Institute of Statistical Mathematics, Tachikawa, Tokyo 190-0014, Japan

Corresponding author: Konstantin Markov (markov@u-aizu.ac.jp)


This work was supported in part by Japan Society for the Promotion of Science under Grant 20K12080.

ABSTRACT Cough is a common symptom associated with respiratory diseases and its analysis plays a crucial role in monitoring the health conditions of affected persons. Traditional cough detection approaches largely fail to identify single cough boundaries when continuous coughs are present, consequently limiting their suitability for effective cough monitoring. In this research, we propose a novel deep learning system for the efficient detection and monitoring of cough events in audio recordings. Our detection pipeline consists of three key steps. First, we perform voice activity detection to eliminate audio silences and focus on relevant segments. Next, we employ a cough classification technique to identify the presence of cough within those audio segments. Finally, we implement cough event detection using a high-performance classification-regression fusion method. Our approach differs from the traditional audio event detection methods in several notable ways: 1) we incorporate a teacher-student framework for the training of our detection model; 2) instead of relying on specific audio features such as MFCC or Mel Spectrogram, our end-to-end system takes the raw audio signal directly as input and outputs the cough boundary timings; 3) the proposed method is general enough to be used for various other sound event monitoring tasks. Our detection model demonstrated strong performance and robustness on both the in-house and public datasets, by achieving cough event detection error-rate scores of 0.31 and 0.32, respectively, which is several times lower than other models. The comparative cough monitoring evaluation of our approach against systems such as the Leicester Cough Monitor and XGBoost demonstrates our method's superiority by achieving the lowest average hourly symmetric mean absolute error (sMAPE) of 8.48%. The code is available at <https://github.com/FengZongyao/Neural-Cough-Counter>.

INDEX TERMS Cough classification, cough detection, cough monitoring, deep learning, signal processing.

I. INTRODUCTION

Cough is a spontaneous defensive mechanism of our body that aims to clear the airway and throat when irritants are present in the respiratory tract. A cough consists of three phases, with the first being an inspiratory phase during which air is inhaled into the lungs. The next two phases bear resemblance to a pump duster, where the air is pressed against the closed glottis through muscle

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda .

contraction in the compressive phase, and then expelled in the expiratory phase to produce the familiar cough sound [1]. Generally, cough serves as a crucial diagnostic indicator, being a common symptom related to many respiratory diseases [2], including chronic obstructive pulmonary disease, asthma, and, not to mention, COVID-19 and its variants [3], [4], [5]. In practice, cough frequency monitoring enables doctors to track the disease severity and verify the effectiveness of the treatment [6], [7]. However, traditional cough diagnosis typically relies on subjective reports of patients, such as the Leicester Cough Questionnaire [8]

or the Cough-Specific Quality of Life Questionnaire [9], which may not provide accurate information. In contrast, an automatic and precise cough monitoring system will not only better assist professionals with disease diagnosis and treatment but also benefit the development of new healthcare services.

The fundamental component of a cough monitoring system is cough detection. It falls under the category of sound event detection tasks, aiming to identify and pinpoint the target event and its corresponding temporal location within the audio signal. The typical approach for sound event detection begins with the extraction of audio features, followed by the utilization of a classifier to determine whether the target event of interest is present within the input signal [10]. When the goal is to simply assess the presence of a specific sound event within a short audio segment, the result of this step is sufficient and the task is commonly known as segment-based sound event detection. In other scenarios where identifying the boundaries of the target events is essential, an additional step is required, where the event's onset and offset timing have to be determined. This is usually achieved by performing frame-wise audio tagging [11]. However, for comprehensive cough monitoring, this approach suffers from low performance issues, primarily due to missed boundaries in multiple continuous cough events.

Cough detection inherits the challenges of sound event detection along with its own. Unlike speech recognition, sound event detection is in practice, task-specific, as there is no established ontology to universally define sound events. In addition, because of the additive nature of sounds, when multiple sound sources are active at the same time, the detection task becomes more challenging. Furthermore, the distance between the sound source and the recording device, as well as the biases introduced by different types of recording devices, are also adding difficulty to the task [11]. The challenges for detecting cough itself include the large data variation and the difficulty to identify single cough boundaries in continuous coughs, which is crucial for cough monitoring. The characteristics of cough sounds depend on many factors, such as differences in the airway, lungs, and vocal cords. Regardless of the cause of these differences, they lead to different kinds of cough sounds. For example, coughs from obstructed airways, like bronchiectasis or asthma patients, tend to have stronger energy and shorter duration, while coughs from interstitial lung disease patients have the longest duration, and coughs from healthy airways and lungs tend to have a higher frequency [12]. When the final objective is to monitor the frequency of coughs, the system has to be able to identify a single cough boundaries, especially for cases involving multiple continuous coughs. For instance, when two continuous coughs are present, the traditional frame-tagging approach may detect one cough, when in fact there are two coughs. This may occur because the interval between continuous coughs is too short for such an approach to work reliably.

To address the aforementioned challenges and to provide a comprehensive and automated solution, we propose a novel end-to-end deep neural network (DNN) based system which we call the 'Neural Cough Counter'. The key contributions of this study are as follows:

- **Combination of transfer learning with distillation training for improved frame-level cough detection:** We harness the capabilities of transfer learning by fine-tuning a pre-trained speech model for feature extraction. This process in combination with distillation training enhances the quality and robustness of the features used for cough detection, making our model more powerful and reliable.
- **Innovative Fusion Model for Cough Detection:** Our design of a regression and classification fusion model offers an innovative approach to cough boundary detection. This fusion model successfully overcomes the challenges inherent in traditional frame-tagging approaches. It not only enhances the identification of single cough events within continuous cough sequences but also provides a foundation for precise cough monitoring.
- **Application to a Range of Sound Event Detection Tasks:** While our primary focus is on cough detection and monitoring, our proposed framework can be potentially applied to a wide range of sound event monitoring tasks. By leveraging deep learning techniques, our model showcases adaptability that could extend to various sound event detection applications.

By incorporating these innovations, our study paves the way for more accurate and automated healthcare diagnostics and monitoring systems, contributing to health-related research.

II. RELATED WORKS

In the past years, several algorithms for cough detection have been developed. Matos et al. [13] proposed a keyword-spotting approach using Hidden Markov Model (HMM) with Mel frequency cepstral coefficients (MFCC) for cough detection. Some other works focused on improving the performance with more efficient features. Le and Wu [14] replaced the MFCC with the Sub-band Energy Cepstrum Coefficient (SECC), and adopted Continuous HMM (CHMM) model to handle cough data. Liu et al. [15] used Gammatone Cepstral Coefficients feature to substitute the MFCC for cough detection. Monge-Álvarez, et al. [16] utilized local Hu moments as features and a k-NN classifier. These works however deeply depend on the feature engineering step. Compared to end-to-end models, their development is more time consuming and requires task-specific knowledge for handcrafting features and parameter tuning. Since the rise of deep learning, some studies have addressed the cough detection problem with deep learning models. Liu et al. [17] adopted a two-step detection framework, an HMM based segmentation model followed by

a Multi-Layer Perceptron (MLP) based classification model carried out the task on MFCC inputs. Amoh and Odamé [18] proposed the DeepCough, in which the spectrogram of the audio signal is first calculated. Next, they treated the spectrogram as an image and applied the well-known convolutional neural network (CNN) from the computer vision domain to classify the cough segments. In addition, to model the coughs' temporal dependencies, they further deployed a recurrent neural network (RNN) in their following work [19]. Specifically, a gated recurrent unit (GRU) network encodes a sequence of spectrogram frames, then a long short-term memory (LSTM) network with a dense classifier assigns each frame a cough label. The recent CoughNet proposed by Rashid et al. [20], availed of a convolutional recurrent neural network (CRNN) to detect cough from mel-spectrograms with different window sizes. In particular, after splitting the mel-spectrogram into frames, a 1d-CNN layer is used to extract frame-wise features. The feature space is reduced by a max-pooling layer, the following LSTM layer is used to learn the temporal relations, and a dense layer is used for classification. Although these deep learning-based works achieved satisfactory performance, their cough detection is performed either as segment-based or frame-based classification, which does not meet the requirement for cough monitoring. Even though the spectrogram is a popular choice, using the raw waveform as input has shown some advantages for self-supervised models [21], [22], [23].

Lately, transfer learning and knowledge distillation have emerged as powerful techniques for improving the performance and efficiency of deep learning models. Transfer learning allows leveraging knowledge gained from model pre-training on large-scale datasets to enhance performance on downstream tasks with limited labeled data. This technique has been successfully applied in the sound event detection domain [24], [25], [26]. On the other hand, knowledge distillation enables transferring knowledge from a large, complex teacher model to a more efficient student model by learning from the combination of soft and hard labels [27], [28], [29]. Both these techniques benefit the development of better models and we utilize them in our study.

Three studies that meet the requirements for cough monitoring were selected and used for comparison with our work. The Leicester Cough Monitor (LCM) [30] is an ambulatory cough detection system. It utilizes a wearable microphone necklace that records sound continuously onto a digital recorder. The system employs a Hidden Markov Model trained with MFCC features. However, the manual system output refinement step, where a person is asked to supply information to improve the system's performance by visual and auditory analysis of 60-80 output sounds, makes the system only semi-automatic. Orlandic et al. [31] proposed a methodology utilizing an XGBoost classifier with a root mean square (RMS) energy threshold to segment cough events from the input audio. By setting an appropriate threshold, this approach aims to effectively distinguish cough

from non-cough sounds. Another work by Simou et al. [32] introduced a universal system for cough detection using a Long Short-Term Memory (LSTM) module combined with spectral energy-based onset detection enabling better cough events identification. Recently, Vankatesh et al. [33] proposed the You Only Hear Once (YOHO) model for sound event detection, which demonstrated better performance than CRNN in Music-Speech detection task. While not specifically designed for cough detection, the YOHO model's superior performance makes it a relevant method for comparison with our system.

III. SYSTEM DESCRIPTION

An overview of our cough detection and monitoring system is presented in Fig. 1. It takes waveform audio as input and outputs the time boundaries of the detected coughs according to the following steps:

- (1) Given an audio recording, voice activity detection is first performed to remove the silent parts of the audio. This way, the amount of data for further processing is greatly reduced. This step results in a set of audio segments containing speech, cough, noise, etc.

- (2) Next, we perform a classification of the segments from the previous step. Here we aim to identify the segments that contain cough events. This step improves the performance of the system, as it removes most of the non-cough segments and reduces the potential false positives in the cough detection step.

- (3) Finally, we run our cough detection model on each cough segment identified by the previous step and locate the beginning and ending time of every cough event.

A. VOICE ACTIVITY DETECTION

Voice activity detection (VAD) aims at extracting segments containing human voices from continuous audio. This allows silences, noises, and other non-voiced sounds to be removed from further processing. In our preliminary experiments, we tried out the popular enterprise-grade voice activity detector, the Silero VAD [34]. However, it removed cough segments as well, with only human speech left as the result, which certainly goes against the task. Therefore, we built a traditional energy-based VAD using the *pydub* Python package. More specifically, given an audio signal, we calculate the average amplitude decibels relative to the full scale (dBFS) of the entire signal. This is a relative measure of the target signal amplitude compared to the maximum possible amplitude in the digital audio system. Subsequently, we scan through the audio with a moving window to locate segments that have a dBFS value below a preset threshold, which are regarded as silences.

B. COUGH SEGMENT CLASSIFICATION

Cough classification refers to determining the presence of a cough within an audio segment. It can be formulated as a binary classification problem. Our objective here is to train a model that accurately determines whether a given

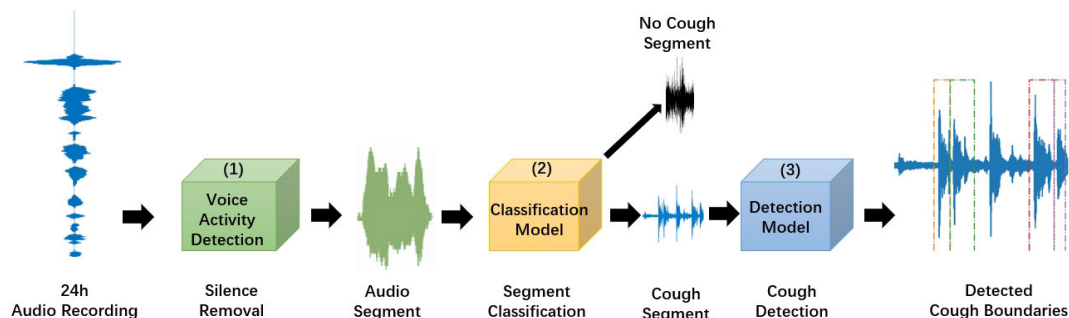


FIGURE 1. The pipeline of the proposed system in this study follows the steps below: (1) Voice Activity Detection is firstly performed on 24 hours of audio recording for silence removal, resulting in a set of audio segments. (2) Segment Classification is subsequently performed to obtain the cough segments. (3) Cough detection is finally carried out on the cough segments to obtain the cough boundaries.

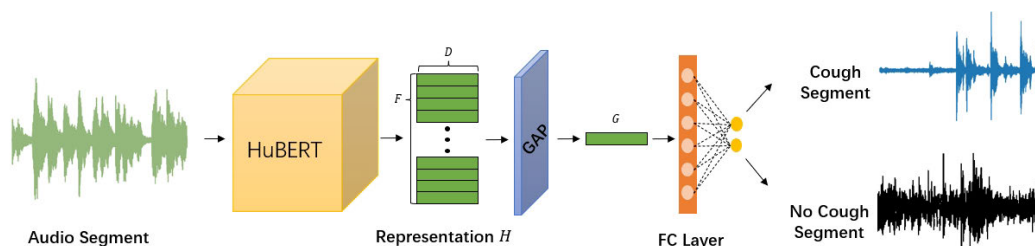


FIGURE 2. Segment classification model using HuBERT with Global Average Pooling (GAP) and a Fully Connected (FC) Layer. Initially, HuBERT extracts the representation H from the input audio segment. Subsequently, the GAP layer aggregates F frames into a global vector G . Finally, through an FC Layer binary classification is performed to distinguish between cough and non-cough segments.

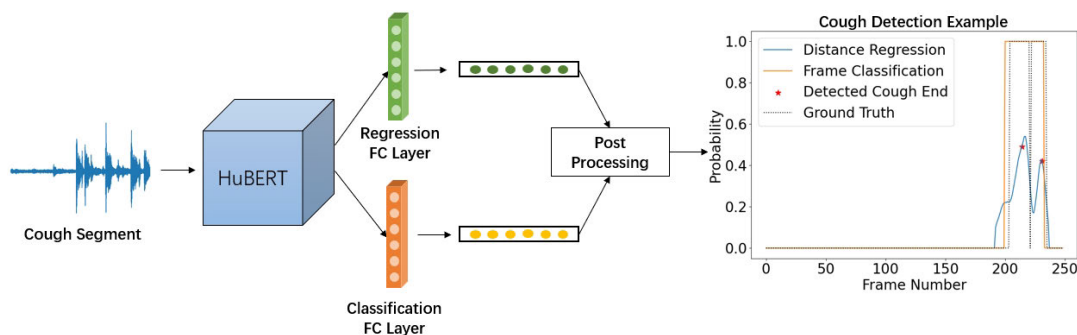


FIGURE 3. Cough Detection Model with HuBERT-based feature extraction, and two Fully Connected (FC) Layers for Regression and Classification. HuBERT extracts frame-based high-level representations from cough segments. The classification FC layer identifies cough frames for boundary determination, while the regression FC layer calculates the distances to determine cough event endings, generating a saw-tooth output for continuous coughs. Post-processing involves peak detection on smoothed regression output, refining predicted cough boundaries.

audio segment contains coughs or not. For feature extraction, we employ the HuBERT [22], which is a self-supervised pretrained speech model. It excels at extracting meaningful and contextually rich representations from raw audio data, which can then be used as input for various downstream audio tasks. As depicted in Fig. 2, the model takes an audio segment and outputs a tensor of shape $(1, F, D)$, capturing F frames with a D -dimensional representation for each frame. To make segment level prediction, we aggregate the frames through a global average pooling layer along the time dimension, resulting in a compact global representation G of shape $(1, D)$. Subsequently, a single fully connected (FC) layer is applied for classification, producing two scalar

outputs representing the predicted class label probability. The whole model is trained in a supervised manner where HuBERT is fine-tuned for the cough segment classification task.

C. COUGH DETECTION

Our cough detection model is outlined in Fig. 3. Given a cough segment, another HuBERT model, different from the one used in the cough segment classification, takes the waveform input and extracts frame-level signal representation. Next, the representation is fed to both the classification FC and regression FC layers. The former performs frame-level cough classification. The latter estimates

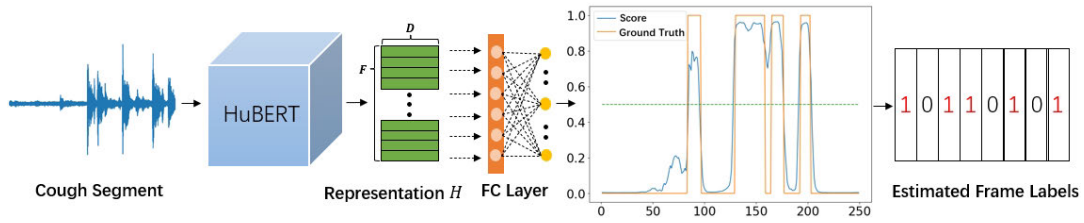


FIGURE 4. Frame-wise classification using HuBERT with a Fully Connected (FC) Layer, Model Prediction (Blue Curve), and Labels (Orange Line). HuBERT processes the input cough segment to derive the representation $H(F * D)$. This representation is passed to a FC layer, which outputs cough probability for each frame.

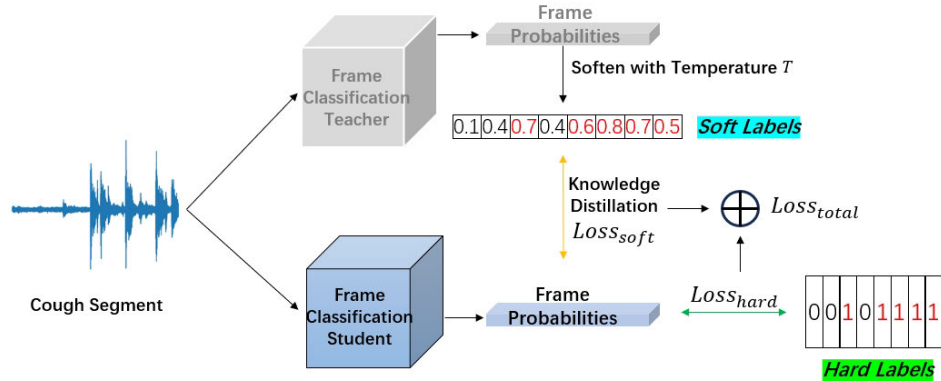


FIGURE 5. Knowledge Distillation Process for Cough Frame Classification: The Teacher model outputs frame probabilities that are softened using a temperature parameter T , resulting in soft labels. The total loss $Loss_{total}$ is computed as a weighted sum of the soft label loss $Loss_{soft}$ and the hard label loss $Loss_{hard}$, facilitating the student model’s learning from both explicit and implicit knowledge.

the distance between the current frame and the starting frame of the corresponding cough. It produces a saw-tooth shaped output when encountering the presence of continuous coughs, where the saw-tooth peak represents the ending frame of a cough event. In the post-processing stage, peak detection is carried out on the smoothed output from the regression FC layer. Finally, the model outputs the timing of cough event centers, which are calculated from obtained cough boundaries. Building this detection model takes three steps: (1) frame-wise classification training, (2) distillation training, and (3) distance estimation training. These are described in detail below:

1) FRAME-WISE CLASSIFICATION

Cough detection aims to identify the boundaries of cough events within an audio input. To accomplish this, we first employ a frame-level classification, where the model predicts the class label for each frame and then the event-level cough result is obtained by aggregating the frame results. As indicated by Fig. 4, we train a cough frame classification model that takes cough segment waveform as input. Same as in section III-B, HuBERT was adopted as a feature extraction module. Frame-level labels are used for fine-tuning HuBERT, which outputs the high-level representations (H) of F frames and D -dimension for each frame ($1, F, D$). Subsequently, the representations are passed into a fully connected layer to perform frame-wise classification. The output of this layer is a tensor of shape $(1, F, 2)$, where 2 represents the number

of classes we are considering. In this case, we utilize the SoftMax activation function to classify each frame into cough or non-cough classes.

2) DISTILLATION TRAINING

To improve the frame-level classification performance of the model we applied distillation training [35], as depicted in Fig. 5. Here, the frame classification model acts as a Teacher while the Student is initialized with a copy of the same model. During the Student training, the output probabilities z_i generated by the Teacher model are softened by the distillation temperature parameter T and serve as soft labels p_i in addition to the original hard labels:

$$p_i = \frac{\exp \frac{z_i}{T}}{\sum_j \exp \frac{z_j}{T}} \tag{1}$$

The total loss function L [35] is a weighted sum of two components: a KL divergence soft loss $Loss_{soft}$ and a standard Binary Cross Entropy (BCE) hard loss $Loss_{hard}$:

$$L = \alpha * Loss_{soft} + (1 - \alpha) * Loss_{hard} \tag{2}$$

The parameter α controls the soft and hard loss balance in the overall training objective. This allows the model to simultaneously learn from the fine-grained information provided by soft labels and the binary information from the hard labels. From our observation, during the early training stages, the KL divergence loss significantly outweighs the BCE loss, reducing substantially its effectiveness.

To maintain a balanced relationship between these two losses, we introduce an adaptive scaling method, as shown in Eq.(3). In combination with Eq.(2), it provides a piecewise dynamic loss weighting scheme, ensuring a harmonized handling of the diverse loss components.

$$L = \begin{cases} \log(Loss_{soft} + 1) + Loss_{hard}, & \text{if } \frac{Loss_{soft}}{Loss_{hard}} \geq \frac{\alpha}{1 - \alpha} \\ Eq.(2), & \text{otherwise} \end{cases} \quad (3)$$

By incorporating soft labels through knowledge distillation and considering both hard and soft losses, the model can benefit from a more comprehensive learning approach, leading to improved performance in cough detection tasks.

3) DISTANCE REGRESSION

In addition to frame classification, we propose a novel distance regression approach to address the challenge of separating continuous coughs. While the frame classification branch predicts binary values (1 or 0) for each frame, it cannot differentiate between individual coughs within a continuous multi-cough sequence event, because it would produce only a sequence of ones. To overcome this limitation, a parallel linear layer branch is introduced to perform relative distance regression. Here, each frame is assigned a label representing its relative frame distance from the starting frame of its corresponding cough event. As shown in Fig. 6, given three continuous cough events, the corresponding distance regression labels form a saw-tooth wave-like ground truth pattern, in which each peak of the saw-tooth represents the ending frame of one cough event. This regression branch enables the model to capture the temporal progression of cough events and to distinguish between different cough events within continuous cough sequences. As illustrated in Fig. 7, the frame classification model is frozen during the training phase, focusing solely on training the regression branch. Moreover, a mask covering cough event(s) is generated and applied to both the regression outputs and the true regression labels. This selective approach ensures that only cough frames actively contribute to the distance regression training process. Algorithm.1 provides details about the labels and mask generation. During training, we use the mean absolute error (MAE) loss function, because the labels are real values within the [0, 1] range. During inference, the output values from the regression branch are masked based on the positive output values index obtained from the classification branch. This selection mechanism allows the model to retrieve the regression values associated with the identified cough frames and infer the relative distances between the frames, enabling better separation and characterization of continuous cough events.

4) POST PROCESSING

Upon retrieving the raw output from the distance regression branch, our goal is to identify the peaks of the saw-tooth

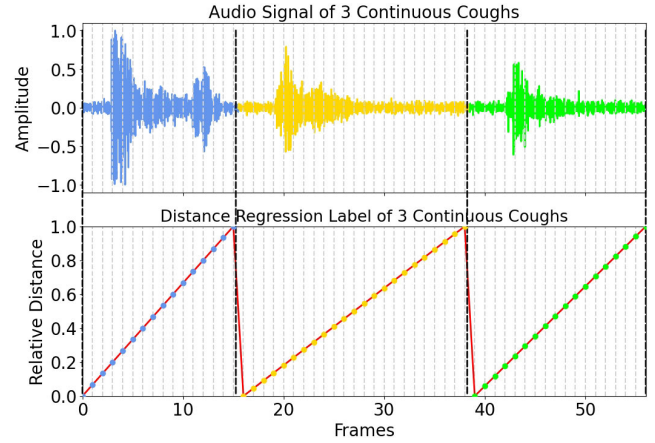


FIGURE 6. Distance Regression Labels (below) for one audio segment containing three continuous coughs (above), each of which is highlighted in a different color. Each dot in the regression label, matched in color with the respective cough in the figure, denotes the length of normalized distance from the current frame to the starting frame of the cough. Collectively, these labels form a saw-tooth pattern, highlighted by the red line.

Algorithm 1 Generation of Distance Regression Labels and Corresponding Mask for Training

Require: *Cough_Boundaries*, *frame_duration*, *total_frames*

▷ Inputs required for algorithm

```

1: regression_labels ← zeros(total_frames)    ▷ Initialize
   Distance Regression labels
2: mask ← zeros(total_frames)                ▷ Initialize mask
3: for each onset, offset in Cough_Boundaries do
4:   start_frame ← floor(onset/frame_duration)
5:   end_frame ← ceil(offset/frame_duration)
6:   cough_length ← end_frame − start_frame
7:   for frame_index ← start_frame to end_frame do
8:     current_cough_frame ← frame_index −
       start_frame
9:     regression_labels[frame_index] ←
       current_cough_frame/cough_length
10:    mask[frame_index] ← 1
11:   end for
12: end for
13: return regression_labels, Boolmask

```

waves, as each peak represents the end of a cough event. Initially, we use a Savitzky-Golay filter [36] to smooth the output and reduce the noise around the local peaks. Subsequently, a continuous wavelet transform (CWT) is applied and the local maxima from the CWT output are then identified as the peaks of interest. A cough segment is split into multiple coughs if detected peak(s) fall within its boundaries. Detected coughs that are less than the minimum cough duration obtained from the dataset analysis, are disregarded. Finally, the cough centers are calculated using the detected boundary of each cough.

D. EVALUATION METRICS

Overall, we assess the effectiveness of our models through the utilization of key evaluation metrics, such as AUROC

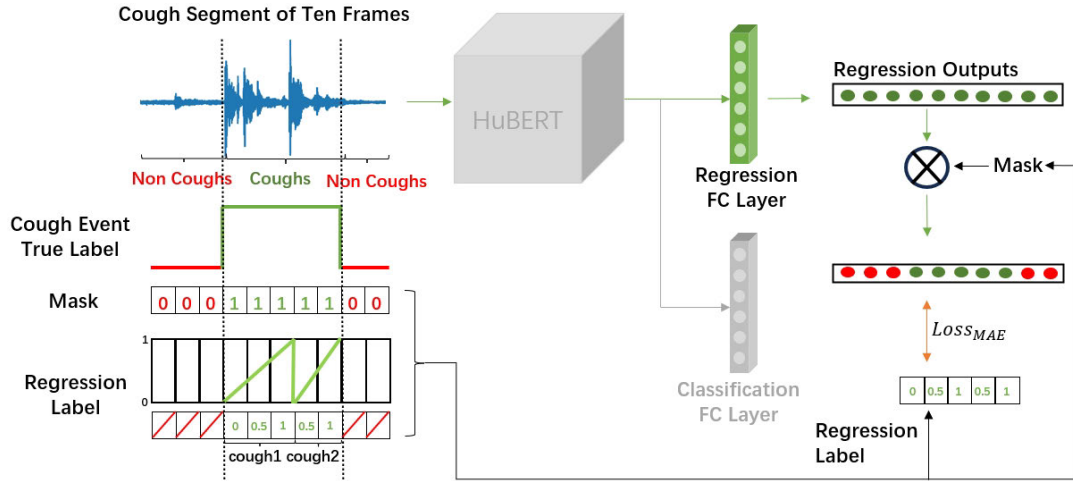


FIGURE 7. Distance regression training process for the Regression Fully Connected (FC) layer with the frame classification branch frozen (indicated in gray). Ground truth cough boundaries generate corresponding distance regression labels, accompanied by a mask that identifies cough frame indices. The Regression FC layer predicts outputs matching the number of input frames, which are then filtered by the mask to concentrate on relevant cough frames before the calculation of the MAE Loss.

(area under the receiver operating characteristic), Accuracy, Precision, Recall, and F1-Score. These metrics enable us to comprehensively evaluate the performance of our models for both the cough classification and cough detection tasks. In this study, we assume that the task of cough monitoring is to count cough events for a predefined period, e.g., one hour. Although the standard mean absolute percentage error (MAPE) can be used to evaluate the normalized count error, when the coughs are counted per hour, there could be zeros in the ground truth value y_i , which will cause division by zero in Eq.(4). To avoid this, we utilize the $sMAPE^{100}$ [37] metric defined in Eq.(5). As shown in Fig. 8, the difference between MAPE and $sMAPE$ lies mainly when the model over-predicts, with 100% error the MAPE reaches unity while the $sMAPE$ is at about 0.33.

$$MAPE = \frac{1}{n} \sum_i^n \frac{|\hat{y}_i - y_i|}{y_i} \quad (4)$$

$$sMAPE^{100} = \frac{1}{n} \sum_i^n \frac{|\hat{y}_i - y_i|}{|\hat{y}_i| + |y_i|} \quad (5)$$

Table 1 presents the metrics utilized in different tasks. When measuring the cough event level detection performance, the Sound Event Detection Evaluation (sed_eval) toolbox [38] is used to calculate the Precision, Recall, F1-Score and Error Rate. These metrics are calculated based on the temporal position overlapping between the system’s prediction and the ground truth, and a time collar or a tolerance is allowed for the onset and offset w.r.t. the ground truth event duration.

IV. EXPERIMENTS AND RESULTS

A. DATASETS

1) IN-HOUSE DATASET

Cough data were collected from 11 patients with various respiratory diseases at Fukushima Medical University over

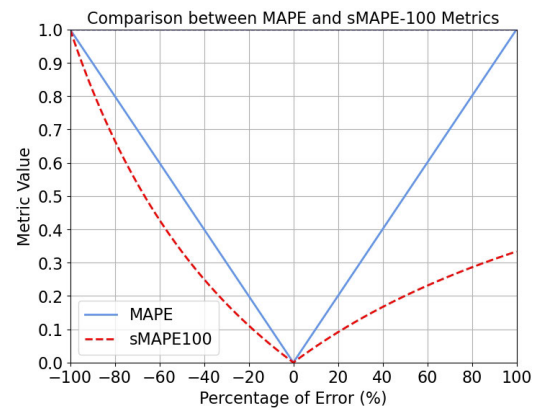


FIGURE 8. Comparison between MAPE and $sMAPE^{100}$ metrics. When the model over-predicts, with 100% error the MAPE reaches unity while the $sMAPE$ is at about 0.33.

four years and three months. Each patient was asked to wear a portable audio recording device for 24 hours (except for patient 9, who was recorded for 18 hours), capturing their daytime activities and sleep. For cough labeling, a semi-automatic procedure was employed [39]. Initially, 2747 coughs were labeled through visual and auditory analysis using Audacity [40]. A GMM-HMM based model, trained with these manually labeled coughs, was then used for cough segment classification on the remaining data. The segments predicted by the model were manually reconfirmed and labeled. Each label includes the start and end times of a single cough event, accurate to 10 milliseconds. In cases of continuous cough sounds, the end time of one cough event serves as the start time of the next. In total, there are 6453 cough labels in our dataset.

To evaluate the performance of our models with as much data as possible, a nested cross-validation approach is employed. Figure 9 illustrates the data splitting process on our In-House dataset, where three test patient data sets are

TABLE 1. Evaluation metrics utilized for assessing the performance of models in segment classification, frame classification, cough detection, and monitoring tasks. The choice of metrics is tailored to the requirements and characteristics of each task.

Task	Metrics						
	Accuracy	F1-Score	Precision	Recall	AUROC	Error-Rate	sMAPE ¹⁰⁰
Cough Segment Classification	✓	✓	-	✓	✓	-	-
Cough Frame Classification	✓	✓	✓	✓	✓	-	-
Cough Event Detection	-	✓	✓	✓	-	✓	-
Cough Monitoring	-	-	-	-	-	-	✓

created such that the total count of coughs is roughly the same forming the outer three-fold CV loop. This approach allows for a comprehensive evaluation by considering the difference between gender, disease, age, recording device, and ambient sound. To determine the best set of hyper-parameters for each model, an inner LOOCV loop is implemented utilizing the Optuna [41] tool. Once the parameter search is completed, the optimal parameters are used to train the model on all patients except those included in the test set. The trained model is then evaluated on the corresponding test set to assess its performance. This process is repeated for all three test sets, resulting in three models with their respective performances on each set. To obtain the final performance metric, the scores from the three models are averaged.

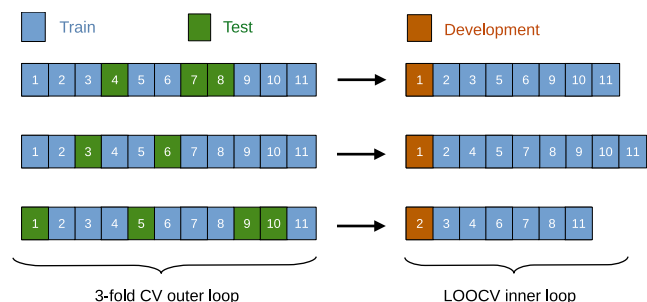


FIGURE 9. The data from 11 patients are split such that all three folds have a similar number of coughs in the test set. Within each fold, Leave-One-Out Cross-Validation (LOOCV) is used for hyperparameter optimization.

2) EDGE-AI COUGH COUNTING DATASET [42]

The Edge-AI Cough Counting (Edge-AI) dataset is a comprehensive collection of recordings spanning approximately 4 hours, with each recording ranging from several to over ten seconds. It includes a total of 4,300 manually annotated cough events sourced from 15 subjects. In addition to the cough events, the dataset also incorporates a diverse range of non-cough sounds, background noises, and motion scenarios. The data are recorded by a chest-mounted wearable device with two microphones: one facing toward the subject’s body, and the other one facing away from the body. In this study, only the recordings from the body-facing microphone are used. For the Edge-AI dataset, we applied a five-fold testing procedure, where each fold consists of the data from three subjects for test and two for validation. The subjects were assigned manually to each fold in a manner that ensured similar total cough counts across all folds. This dataset is particularly challenging because it contains a large amount of multiple consecutive cough event segments.

B. SILENCE REMOVAL

The silence threshold is set to 1.7 times the average signal dBFS value of the current signal, and the rolling window length is chosen to be 0.1 seconds. To keep the coherence of the signal silences shorter than 300ms are preserved. As a result, out of the total 199 hours of recordings from our In-House dataset, we eliminated 48.2 hours of silence, ensuring that all coughs were retained.

As for the Edge-AI dataset [42], no silence was detected with our settings, so we used the original audio segments as they are.

C. MODEL TRAINING

In our experiments, we used the HuBERT base model pre-trained on 960 hours of the LibriSpeech audio dataset. It has 95 million parameters and generates a sequence of frame representations for the input audio signal with a frame rate of 20ms. During the training of the proposed models, including the HuBERT fine-tuning, the Adam optimizer, and the ReduceLRonPlateau scheduler are applied. The scheduler is set to reduce the learning rate by 40 percent if there is no improvement in the validation loss after 3 epochs. In addition, the early stopping mechanism is applied during training with a patience of 6 epochs. The checkpoint of the model’s parameters with the lowest validation loss is saved as the final model.

D. SYSTEM BASELINE

The Convolutional Recurrent Neural Network (CRNN) is a popular model in sound event detection and is used as a baseline in this work. Our CRNN model takes a mel-spectrogram as input and outputs class probabilities. The mel-spectrogram is obtained by the Fast Fourier Transform (FFT) with a length of 400, frame hop size of 322 samples, and a filter bank of 64 mel-spaced filters. The CRNN model comprises three CNN blocks designed to capture both local and global spectral patterns. Each convolution block consists of a 2D Convolution layer with a kernel size of 3, followed by a Batch Normalization layer, a Rectified Linear Unit (ReLU) activation function, a Dropout layer with a rate of 0.3, and a 2D Average Pooling layer with a kernel size of (1,4). The CNN output is then fed into two layers of Bidirectional GRU, each with 64 hidden units and a 0.2 dropout rate. The output of the BiGRU, with a shape of (1, frames, classes), is directed to a dense layer to produce the classification result. For the audio segment classification task, the time axis of the

TABLE 2. Segment level cough classification results of our model and the CRNN baseline for the first cross-validation fold.

	F1-Score (%)	Accuracy (%)	Recall (%)	AUROC (%)	# of Parameters
CRNN baseline	93.52	96.94	93.58	98.40	0.23M
Classification Model	96.51	98.69	96.87	99.22	94.37M

TABLE 3. Frame level cough classification results of our models and the CRNN baseline. Numbers represent the mean and std of the 3-fold CV experiments.

	F1-Score (%)	Precision (%)	Recall (%)	Accuracy (%)	AUROC (%)
CRNN baseline	82.87 ± 1.62	85.32 ± 4.42	82.00 ± 4.53	91.70 ± 1.08	95.52 ± 0.79
Teacher Frame Model	87.71 ± 1.70	88.01 ± 3.69	87.90 ± 1.59	94.21 ± 1.83	97.09 ± 0.83
Student Frame Model	89.60 ± 2.00	89.72 ± 2.78	89.54 ± 1.31	95.13 ± 1.61	97.42 ± 0.72

BiGRU output is averaged before the dense layer to obtain the segment-wise score.

E. SEGMENT CLASSIFICATION

Following common practices [17], [19], [24], audio segments are split or padded to a fixed length before being passed to the models. To ensure a fair comparison between the CRNN and the other models in terms of time series modeling capabilities, it is set to 5 seconds, which equals to 250 time-steps. The training of the segment classification model was performed with a batch size of 16. For the fine-tuning, a learning rate of $1e-5$ was used for the HuBERT model, while the FC layer was optimized using a learning rate of $1e-3$. The training process was conducted for a maximum of 50 epochs with the CrossEntropy loss function. For the baseline CRNN model training, a larger batch size of 512 was selected, and the learning rate was set to $1e-3$. The remaining experimental settings for the CRNN model were consistent with those used for the segment classification model training.

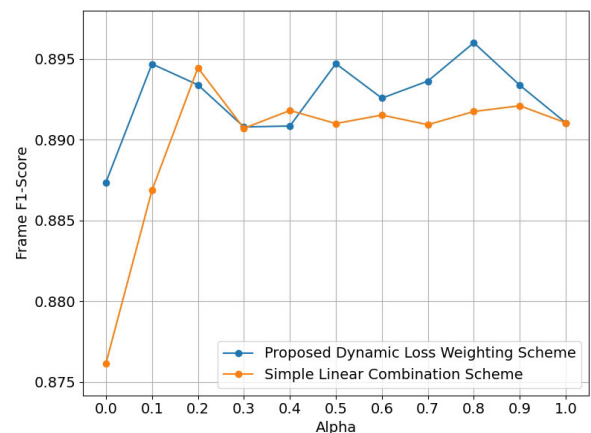
Although important, the cough segment classification is just an intermediate step in our system, so the evaluation was done on the first cross-validation fold only. As indicated by Table 2, the results reveal that our model outperforms the CRNN baseline in all evaluation metrics. It is important to note that despite the CRNN's significantly smaller parameter size, it achieved a high F1-score of 93.52%. Nevertheless, the observed 3% improvement in F1-Score by our model is significant and highlights the benefits of fine-tuning HuBERT on cough data and leveraging its larger capacity for improved acoustic representation, justifying its larger parameter size.

F. COUGH FRAME CLASSIFICATION

The CrossEntropy loss was used as the objective function in the cough frame classification experiment, with a batch size of 50. The training process was performed for a maximum of 50 epochs. For fine-tuning the HuBERT, a learning rate of $1e-5$ was utilized, while the dense classifier was trained using a learning rate of $1e-3$. These specific learning rates were optimal for the training process and achieved the best performance. Similarly, the CRNN model training process utilized a learning rate of $1e-3$. The remaining experimental settings were kept consistent with the configuration of

our frame classification model. We further improved the performance of our frame model using distillation training.

In addition, the effectiveness of our proposed dynamic loss weighting scheme Eq.(3) is validated by exploring how the balance between soft and hard losses affects the model performance. Fig. 10 compares the averaged frame classification F1-score of our proposed scheme against the simple linear combination scheme of Eq.(2) across different α values on our In-House dataset. The results demonstrate that the former consistently outperforms the latter for most α values. When α is zero, Eq.(2) relies solely on the hard loss, while Eq.(3) maintains a combination of both losses. That is why the curves have different values at $\alpha = 0$. At $\alpha = 1$, Eq.(3) collapses into Eq.(2), resulting in identical performance. The performance advantage of the proposed weighting scheme is particularly large when more emphasis is placed on the soft loss ($\alpha > 0.5$), where it maintains higher F1-scores. Finally, the best parameter α in Eq.(3) was found

**FIGURE 10.** Loss weighting schemes comparison in terms of averaged frame classification 3-fold CV F1-Score on our In-House dataset.

to be 0.8 along with distillation temperature $T=5$, indicating the relative importance of the teacher's predictions.

The results from Table 3 highlight the superior performance of the proposed frame models in cough frame classification compared to the CRNN baseline model. The improved performance of the student model over the teacher model demonstrates the efficacy of knowledge distillation and the benefit of learning from the soft labels.

TABLE 4. Comparison of cough event detection results of our proposed cough detection model, our frame-level classification model, the CRNN baseline, and the YOHO model. The estimated cough boundaries are evaluated using `sed_eval` [38] with a time collar of 0.2 seconds. Table entries are the mean and standard deviation of the 3-fold cross-validation experiment on the In-House Dataset, and a 5-fold CV on the Edge-AI Dataset.

Method	In-house Dataset				Edge-AI Dataset [42]			
	F1-Score (%)	Precision (%)	Recall (%)	Error Rate	F1-Score (%)	Precision (%)	Recall (%)	Error Rate
YOHO [33]	48.27 ± 4.13	50.63 ± 9.03	45.57 ± 3.97	0.99 ± 0.17	12.63 ± 4.58	13.04 ± 4.92	14.29 ± 7.21	1.89 ± 0.35
CRNN baseline	62.6 ± 7.18	74.67 ± 7.59	54.83 ± 10.13	0.64 ± 0.11	21.09 ± 6.61	25.53 ± 7.29	18.98 ± 6.74	1.40 ± 0.20
Frame Model	68.0 ± 6.66	78.73 ± 6.33	60.13 ± 7.73	0.56 ± 0.11	47.49 ± 11.08	57.47 ± 10.93	40.67 ± 10.90	0.88 ± 0.15
Detection Model	83.53 ± 0.94	87.27 ± 3.21	80.37 ± 3.69	0.31 ± 0.02	84.34 ± 3.98	83.22 ± 2.66	85.67 ± 6.43	0.32 ± 0.07

TABLE 5. Cough monitoring results on the In-House dataset in terms of hourly $sMAPE^{100}$. Our proposed cough detection model is referred to as the 'Detection Model', while the frame-level classification model is termed the 'Frame Model'. The 'CRNN' is our baseline system. The first three columns correspond to the test patient ID, the number of coughs, and the recording length respectively.

Patient	Cough Counts	Hours	Hourly symmetric Mean Absolute Percentage Error ($sMAPE^{100}$) on the In-House dataset					
			Detection Model	Frame Model	CRNN	LCM [30]	XGBoost [31]	Universal(tuned) [32]
1	772	24	2.02	4.94	3.26	37.92	46.05	32.58
3	1298	24	3.56	6.97	5.05	25.62	42.29	28.51
4	475	6	8.96	9.29	19.51	36.56	48.24	36.94
5	448	24	11.49	12.86	15.25	12.11	40.51	43.44
6	623	24	7.94	13.45	16.03	49.77	52.78	50.00
7	234	6	7.57	10.67	21.96	40.33	19.33	34.91
8	749	1	11.21	24.11	33.16	59.08	37.56	39.61
9	266	18	7.55	11.09	12.69	25.80	41.87	45.70
10	229	24	16.07	17.89	18.28	61.26	43.07	49.81
Average Hourly $sMAPE^{100}$			8.48	12.36	16.13	38.72	41.30	40.17

G. DISTANCE REGRESSION AND COUGH EVENT DETECTION

For the distance regression training, the parameters of the student frame classification model were frozen while the distance regression FC layer (see Fig. 7) was learned with a batch size of 100, learning rate of 7.3e-3 and MAE loss function. To estimate the distance prediction error, we evaluated the model in the 3-fold CV setting. The obtained average MAE was 0.053 which corresponds to ±1 frame error for a 400ms long cough event. This indicates the effectiveness of the distance regression branch in accurately estimating the cough frame distances.

The regression branch training completes the construction of our final cough detection model. During the evaluation using `sed_eval` toolbox [38], the time collar was set to 0.2 seconds. The estimated cough boundaries from our cough detection model and YOHO are used directly, while the cough boundaries from our frame model and the CRNN baseline are obtained from sequences of positive frame-level cough predictions. We did not evaluate the XGBoost [31] or the Universal [32] for cough detection because they are not able to output cough event boundaries. The cough detection results are summarized in Table 4 and a visualization of detected cough boundaries from our cough detection model is shown in Fig. 11. As evident from Table 4, our proposed cough detection model significantly outperforms the other methods on both datasets. It demonstrates robustness and consistency across different folds of the in-house dataset and maintains its superior performance on the more challenging and diverse Edge-AI [42] dataset. Our frame-level classification model and the CRNN baseline, which rely on sequences of positive frame-level cough predictions, achieve lower performance compared to our cough detection model. This highlights the

importance of the distance regression branch in accurately estimating cough event boundaries. The YOHO model, despite outperforming CRNN in detecting longer events like music and speech [33], shows the lowest performance among the compared methods for cough detection. This can be attributed to the relatively short duration of cough events and its failure to distinguish consecutive coughs. These results validate the effectiveness of our proposed cough detection model in accurately identifying cough events in audio recordings. The model's strong performance across both datasets demonstrates its generalizability and potential for real-world applications in cough monitoring systems.

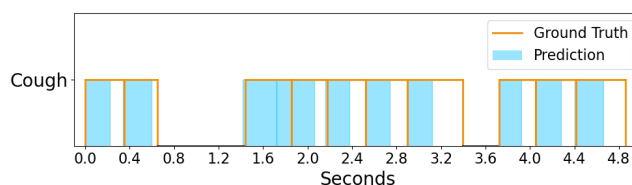


FIGURE 11. Visualization of detected cough boundaries on the Edge-AI Dataset [42] by the proposed cough detection model. The orange lines refer to the ground truth boundaries while the blue bars refer to the predicted boundaries.

H. COUGH MONITORING

As mentioned above, we assume that the cough monitoring task is to count the coughs in the patient recordings and produce a histogram of the hourly cough frequencies. In all cases, cough counts are obtained by first performing cough detection and then determining the number of coughs in every one-hour range. To provide a comprehensive performance comparison, in addition to our proposed models and the system baseline, we also evaluated two other approaches

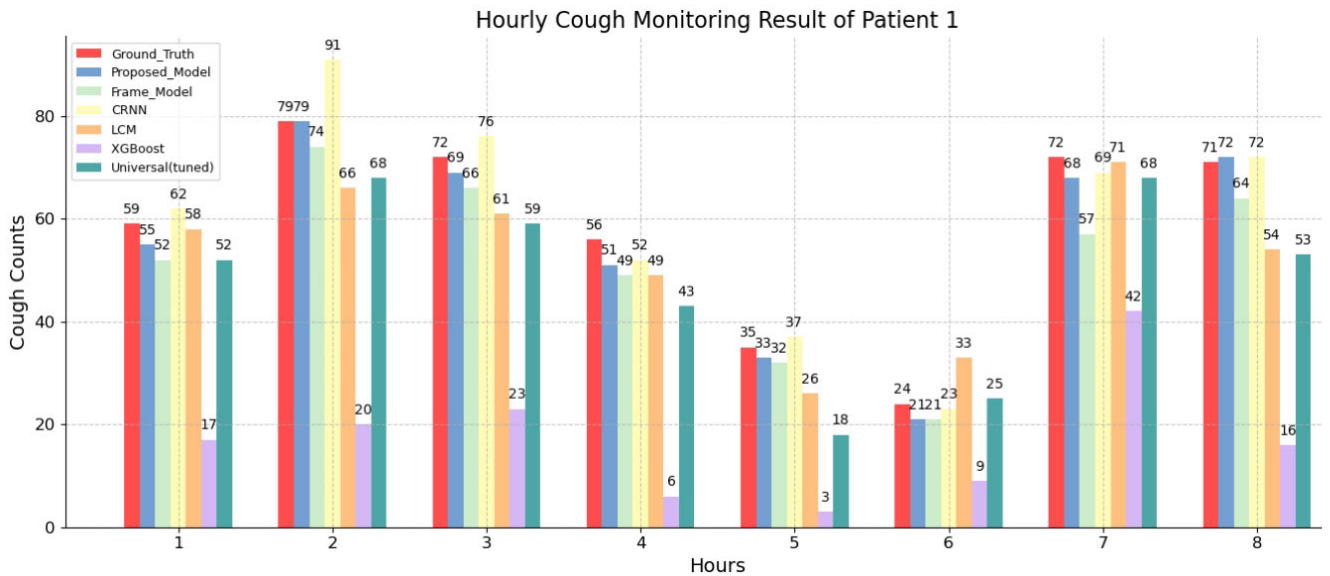


FIGURE 12. Histogram of hourly cough counts for the first 8 hours of Patient 1's recording obtained by different methods.

mentioned in Section II. For the universal system of Simou et al. [32], as suggested in the paper, we adjusted the predefined hyper-parameters to set an appropriate threshold for evaluation. This adapted version is referred to as “Universal-tuned” in the results. We also trained an XGBoost based model from our data following the methodology from [31]. It was used for segment level cough classification while the cough events were detected by applying a threshold on the RMS energy envelope as described in the paper. The last cough monitoring method we use for comparison is the Leicester Cough Monitor (LCM) [30]. The cough frequencies from the LCM for each patient were provided by Fukushima Medical University. Table 5 summarizes the cough monitoring performance of all the described methods in terms of $sMAPE^{100}$ for each test patient. The results reveal a significantly lower error rate achieved by our model than all other approaches. This indicates its effectiveness in accurately monitoring cough events and providing more precise estimations of the hourly cough counts.

An example of an hourly cough count histogram for the first 8 hours of the Patient 1 data is presented in Fig. 12. As can be seen, counts obtained from our proposed detection model (blue bars) are the closest to the ground truth (red bars).

V. CONCLUSION AND FUTURE WORK

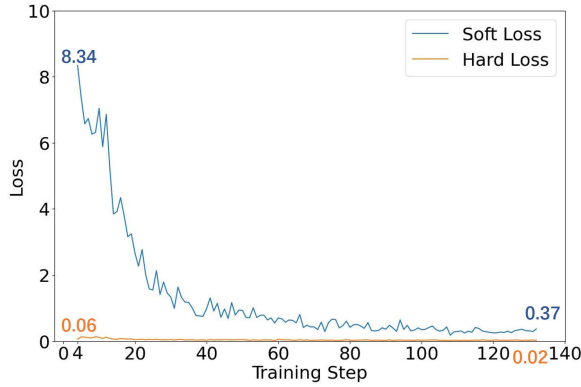
In this research, we have addressed the important task of cough detection and monitoring using deep learning-based sound event detection techniques. Through extensive experiments and analysis, we demonstrated the superiority of our classification model compared to the baseline CRNN model in terms of various evaluation metrics. Our model exhibited enhanced performance in accurately identifying and distinguishing cough segments, showcasing the effectiveness of using pre-trained models and fine-tuning them on target

cough data. Furthermore, we explored the use of knowledge distillation to train a student model with improved frame level classification performance resulting in more accurate cough event boundaries detection. In addition, to overcome the limitations of existing studies in cough detection and monitoring, we proposed an innovative regression and classification fusion approach. This approach enabled us to train a cough detection model, which exhibits remarkable capabilities in detecting the boundaries of individual cough events within continuous cough instances. We rigorously evaluated our model on both the in-house and public datasets, demonstrating its robustness and generalizability. The strong performance of our model on the public dataset highlights its potential for real-world applications and its ability to handle diverse cough data from different sources. Finally, we conducted experiments to compare the performance of different models in cough monitoring. The results highlighted the exceptional performance of our model, surpassing the system baseline and other models such as the LCM [30], XGBoost [31], and the Universal [32] approach in estimating the hourly cough counts with much lower hourly average $sMAPE$ values.

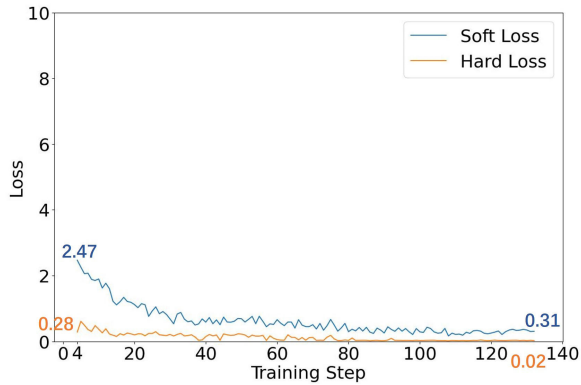
The proposed cough detection and monitoring approach has the potential to impact various domains beyond respiratory health, such as urban planning, environmental monitoring, and animal welfare. The methodology's independence on the specific data type makes it adaptable to any sound event detection task that requires counting consecutive events. Moreover, the insights gained from this research can be extended to other audio-based monitoring applications, such as detecting abnormal sounds in machinery or analyzing animal vocalizations for signs of distress, fostering advancements in the broader field of acoustic signal processing. Future work could focus on enhancing the robustness and generalizability of our approach by exploring

more diverse datasets and investigating the integration of our method with other modalities.

**APPENDIX A
DYNAMIC LOSS WEIGHTING SCHEME JUSTIFICATION
AND ANALYSIS**



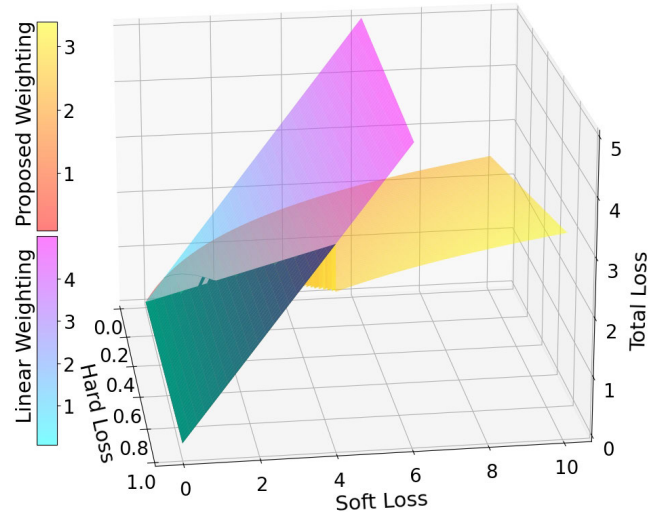
(a) Soft and Hard loss values using the simple linear combination scheme with $\alpha = 0.5$.



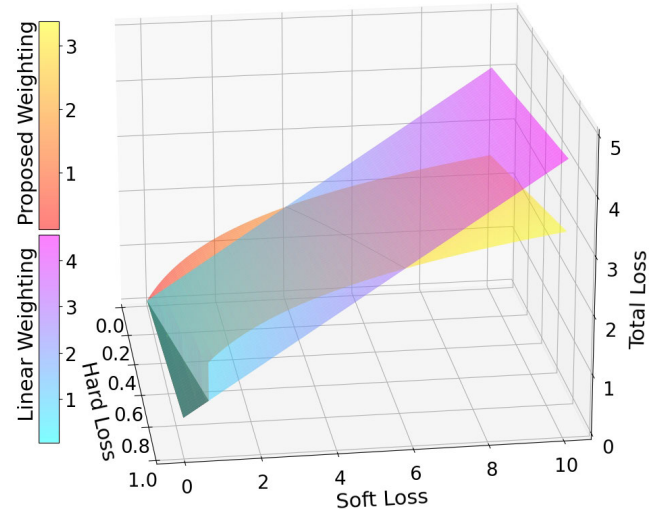
(b) Soft and Hard loss values using the proposed weighting scheme with $\alpha = 0.5$.

FIGURE 13. Soft and Hard loss values obtained by the proposed dynamic loss weighting scheme and the simple linear combination scheme during the first 140 training steps on the 1st CV fold of our In-House dataset, where $\alpha = 0.5$. (For better visual clarity, the loss curves are plotted starting from the 5th training step, although the differences are larger in the first four training steps).

The proposed dynamic loss weighting scheme was developed to address a significant challenge observed during the distillation training process. We noticed that the KL divergence soft loss was substantially larger than the BCE hard loss, especially during the first training steps. An example of the loss shape during the training on 1st CV fold of our In-House dataset is given in Fig. 13. The α is set to 0.5 to maintain the same scale on both losses. This imbalance posed a problem for the simple linear combination scheme, as the soft loss would dominate the training process, potentially hindering the model’s ability to learn effectively from both the soft and hard labels. To address this issue, we proposed a dynamic loss weighting scheme that scales the soft loss when it significantly outweighs the hard loss. This is achieved by



(a) Total loss surfaces of both weighting schemes when $\alpha = 0.8$. (For better visual clarity, the portion of the linear scheme surface above the value of 5 is not shown in the figure.)



(b) Total loss surfaces of both weighting schemes when $\alpha = 0.4$

FIGURE 14. Comparison of total loss surface between the proposed dynamic loss weighting scheme (color gradient from orange to yellow) and the simple linear combination scheme (color gradient from blue to purple) for simulated soft and hard loss values. The green part of the surface shows the region where the proposed scheme is equivalent to the simple linear combination.

taking a logarithm of the loss as:

$$\log(Loss_{soft} + 1)$$

Unity is added to ensure that the scaled loss is always positive. Such logarithmic scaling is applied only when the ratio $Loss_{soft}/Loss_{hard}$ is bigger than some predefined threshold:

$$\frac{Loss_{soft}}{Loss_{hard}} \geq th$$

This, however, is inconvenient in practice since such a threshold has no upper bound and is difficult to tune. To solve this problem, we use a threshold of the form:

$$th = \frac{\alpha}{1 - \alpha}$$

where α takes values from the $[0, 1]$ range. Thus, our proposed weighting scheme becomes:

$$L = \begin{cases} \log(Loss_{soft} + 1) + Loss_{hard}, & \text{if } \frac{Loss_{soft}}{Loss_{hard}} > \frac{\alpha}{1-\alpha} \\ \alpha * Loss_{soft} + (1 - \alpha) * Loss_{hard}, & \text{otherwise} \end{cases}$$

To analyze the behavior of this weighting scheme and to compare it with the conventional linear combination, we visualized the total loss surface for simulated $Loss_{soft}$ and $Loss_{hard}$ values ranging from 0 to 10 and from 0 to 1 respectively. Fig. 14 shows the total loss surface when $\alpha = 0.8$ and $\alpha = 0.4$ for both schemes. With the linear combination scheme, it is clear that the soft loss dominates the hard loss, especially for bigger α . In contrast, with the proposed scheme, the total loss is less influenced by the soft loss increase and is independent of α as long as the ratio $Loss_{soft}/Loss_{hard}$ is bigger than the threshold. The area where both schemes are equivalent (the green portion of the surface in the figure) occupies the space of low soft loss values only and its size can be controlled by the value of α . Both plots in Fig. 14 show that the proposed scheme effectively scales both the soft and hard losses, dynamically limiting their ratio during the training. In other words, the proposed weighting scheme ensures that neither loss component dominates the training process, allowing the model to benefit from both the fine-grained information in soft labels and the ground truth in hard labels.

REFERENCES

- [1] R. S. Irwin, "Cough. A comprehensive review," *Arch. Internal Med.*, vol. 137, no. 9, pp. 1186–1191, Sep. 1977.
- [2] A. Proaño et al., "Dynamics of cough frequency in adults undergoing treatment for pulmonary tuberculosis," *Clin. Infectious Diseases*, vol. 64, no. 9, pp. 1174–1181, May 2017.
- [3] L.-Q. Li, T. Huang, Y.-Q. Wang, Z.-P. Wang, Y. Liang, T.-B. Huang, H.-Y. Zhang, W. Sun, and Y. Wang, "Covid-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis," *J. Med. Virol.*, vol. 92, no. 6, pp. 577–583, 2020.
- [4] J. R. Maurer, "Cough and sputum production are associated with frequent exacerbations and hospitalizations in COPD subjects," *Yearbook Pulmonary Disease*, vol. 2010, pp. 101–102, Jan. 2010.
- [5] P. V. Dicpinigaitis, "Chronic cough due to asthma: ACCP evidence-based clinical practice guidelines," *Chest*, vol. 129, no. 1, pp. 75S–79S, 2006.
- [6] R. S. Irwin and J. M. Madison, "The diagnosis and treatment of cough," *New England J. Med.*, vol. 343, no. 23, pp. 1715–1721, 2000.
- [7] A. A. Raj and S. S. Birring, "Clinical assessment of chronic cough severity," *Pulmonary Pharmacol. Therapeutics*, vol. 20, no. 4, pp. 334–337, Aug. 2007.
- [8] S. S. Birring, "Development of a symptom specific health status measure for patients with chronic cough: Leicester cough questionnaire (LCQ)," *Thorax*, vol. 58, no. 4, pp. 339–343, Apr. 2003.
- [9] C. T. French, R. S. Irwin, K. E. Fletcher, and T. M. Adams, "Evaluation of a cough-specific quality-of-life questionnaire," *Chest*, vol. 121, no. 4, pp. 1123–1131, Apr. 2002.
- [10] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 379–393, Feb. 2018.
- [11] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Process. Mag.*, vol. 38, no. 5, pp. 67–83, Sep. 2021.
- [12] G. Rudraraju, S. Palreddy, B. Mamidgi, N. R. Sripada, Y. P. Sai, N. K. Vodnala, and S. P. Haranath, "Cough sound analysis and objective correlation with spirometry and clinical diagnosis," *Informat. Med. Unlocked*, vol. 19, Mar. 2020, Art. no. 100319.
- [13] S. Matos, S. S. Birring, I. D. Pavord, and D. H. Evans, "Detection of cough signals in continuous audio recordings using hidden Markov models," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 1078–1083, Jun. 2006.
- [14] S. Le and W. Hu, "Cough sound recognition based on Hilbert marginal spectrum," in *Proc. 6th Int. Congr. Image Signal Process. (CISP)*, vol. 3, Dec. 2013, pp. 1346–1350.
- [15] J.-M. Liu, M. You, G.-Z. Li, Z. Wang, X. Xu, Z. Qiu, W. Xie, C. An, and S. Chen, "Cough signal recognition with gammatone cepstral coefficients," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, Jul. 2013, pp. 160–164.
- [16] J. Monge-Álvarez, C. Hoyos-Barceló, P. Llesco, and P. Casaseca-de-la-Higuera, "Robust detection of audio-cough events using local hu moments," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 1, pp. 184–196, Jan. 2019.
- [17] J.-M. Liu, M. You, Z. Wang, G.-Z. Li, X. Xu, and Z. Qiu, "Cough detection using deep neural networks," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2014, pp. 560–563.
- [18] J. Amoh and K. Odame, "DeepCough: A deep convolutional neural network in a wearable cough detection system," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2015, pp. 1–4.
- [19] J. Amoh and K. Odame, "Deep neural networks for identifying cough sounds," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 5, pp. 1003–1011, Oct. 2016.
- [20] H.-A. Rashid, A. N. Mazumder, U. P. K. Niyogi, and T. Mohsenin, "CoughNet: A flexible low power CNN-LSTM processor for cough sound detection," in *Proc. IEEE 3rd Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Jun. 2021, pp. 1–4.
- [21] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 33, 2020, pp. 12449–12460.
- [22] W.-N. Hsu, B. Bolte, Y. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021.
- [23] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.
- [24] B. Swaminathan, M. Jagadeesh, and S. Vairavasundaram, "Multi-label classification for acoustic bird species detection using transfer learning approach," *Ecological Informat.*, vol. 80, May 2024, Art. no. 102471.
- [25] S. R. Kothinti and M. Elhilali, "Multi-rate modulation encoding via unsupervised learning for audio event detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2024, no. 1, p. 19, Apr. 2024.
- [26] S. Baur, Z. Nabulsi, W.-H. Weng, J. Garrison, L. Blankemeier, S. Fishman, C. Chen, S. Kakarmath, M. Maimbolwa, N. Sanjase, B. Shuma, Y. Matias, G. S. Corrado, S. Patel, S. Shetty, S. Prabhakara, M. Muyoyeta, and D. Ardila, "HeAR—Health acoustic representations," 2024, *arXiv:2403.02522*.
- [27] Y. Xiao and R. K. Das, "Dual knowledge distillation for efficient sound event detection," 2024, *arXiv:2402.02781*.
- [28] D. Kim, M.-S. Baek, Y. Kim, and J.-H. Chang, "Improving target sound extraction with timestamp knowledge distillation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 1396–1400.
- [29] A. M. Tripathi and K. Paul, "Data augmentation guided knowledge distillation for environmental sound classification," *Neurocomputing*, vol. 489, pp. 59–77, Jun. 2022.
- [30] S. S. Birring, T. Fleming, S. Matos, A. A. Raj, D. H. Evans, and I. D. Pavord, "The Leicester cough monitor: Preliminary validation of an automated cough detection system in chronic cough," *Eur. Respiratory J.*, vol. 31, no. 5, pp. 1013–1018, Jan. 2008.
- [31] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," 2020, *arXiv:2009.11644*.
- [32] N. Simou, N. Stefanakis, and P. Zervas, "A universal system for cough detection in domestic acoustic environments," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 111–115.

- [33] S. Venkatesh, D. Moffat, and E. R. Miranda, "You only hear once: A YOLO-like algorithm for audio segmentation and sound event detection," *Appl. Sci.*, vol. 12, no. 7, p. 3293, Mar. 2022.
- [34] S. Team. (2021). *Silero VAD: Pre-Trained Enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier*. [Online]. Available: <https://github.com/snakers4/silero-vad>
- [35] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
- [36] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.
- [37] J. S. Armstrong, *Long-Range Forecasting: From Crystal Ball to Computer*, vol. 348. New York, NY, USA: Wiley, 1985.
- [38] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, p. 162, May 2016.
- [39] Y. Kawashima, "Cough recognition system based on DNN," M.S. thesis, Division Inf. Syst., Univ. Aizu, Japan, 2020.
- [40] Audacity Team. (2021). *Software is Copyright 1999–2021 Team. The Name Audacity is a Registered Trademark*. [Online]. Available: <https://www.audacityteam.org/>
- [41] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2623–2631.
- [42] L. Orlandic, J. Thevenot, T. Teijeiro, and D. Atienza, "A multi-modal dataset for automatic edge-AI cough detection," in *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Sydney, NSW, Australia, Jul. 2023, pp. 1–7. [Online]. Available: <https://ieeexplore.ieee.org/document/10340413/>



JUNPEI SAITO received the bachelor's and Ph.D. degrees from Fukushima Medical University, in 1996 and 2004, respectively, with a focus on fractional exhaled nitric oxide as a non-invasive biomarker of asthma. He is currently a Senior Lecturer with the Department of Pulmonary Medicine, Fukushima Medical University. After graduation, he was trained in internal and respiratory medicine at Fukushima Medical University Hospital and obtained several board certifications, including internal medicine, allergology, respiratory medicine, respiratory endoscopy, and asthma in Japan. At the same time, he performed postdoctoral training at Fukushima Medical University. He has authored and co-authored over 80 scientific articles in the field of allergy and respiratory. He holds a patent regarding hydrogen sulfide in exhaled gas. His current research interests include the evaluation of new biomarkers in COPD and asthma. In particular, he is interested in cough monitoring, fractional exhaled nitric oxide (FeNO) as one of type 2 biomarkers, and hydrogen sulfide (H₂S) as a candidate for non-type 2 biomarkers. He has received several awards, including travel awards from the American Thoracic Society, European Respiratory Society, Fukushima Medical University Award, and the Clinical Asthma Conference Award. He has served on the editorial board of several journals, including *Japanese Journal of Allergology* and *Japanese Journal of Respiriology*.



ZONGYAO FENG received the B.S. degree in computer science from Beijing University of Technology, China, in 2016, and the M.S. degree in computer science and engineering from The University of Aizu, in 2023, Japan, where he is currently pursuing the Ph.D. degree. His current research interests include machine learning, biomedical signal processing, time-series analysis, computational neuroscience, graph neural networks, and self-supervised learning.



KONSTANTIN MARKOV (Member, IEEE) was born in Sofia, Bulgaria. He received the degree (Hons.) from Saint Petersburg State Polytechnic University, Russia, and the M.Sc. and Ph.D. degrees in electrical engineering from Toyohashi University of Technology, Japan, in 1996 and 1999, respectively. After graduation, he worked for several years as a Research Engineer at the Communication Industry Research Institute, Sofia. In 1999, he joined the Research Development Department, ATR, Japan. In 2000, he became an Invited Researcher at the ATR Spoken Language Translation (SLT) Research Laboratories. Later, he became a Senior Research Scientist at the Acoustics and Speech Processing Department, ATR-SLT. In 2009, he joined the Human Interface Laboratory, Information Systems Division, The University of Aizu, Japan, where he is currently leading the AutoAI Research Cluster. His research interests include audio signal and text processing, deep learning, and AI. He is a member of ISCA. In 1998, he received the Best Student Paper Award from the IEICE Society.



TOMOKO MATSUI (Senior Member, IEEE) received the Ph.D. degree in computer science from Tokyo Institute of Technology, Tokyo, Japan, in 1997. From 1988 to 2002, she was a Researcher at several NTT laboratories, focusing on speaker and speech recognition. From 1998 to 2002, she was a Senior Researcher with the Spoken Language Translation Research Laboratory, ATR, Kyoto, focusing on speech recognition. In 2001, she was an Invited Researcher with the Acoustic and Speech Research Department, Bell Laboratories, Murray Hill, NJ, USA, working on identifying effective confidence measures for verifying speech recognition results. She is currently a Professor at The Institute of Statistical Mathematics, Tokyo, working on statistical spatiotemporal modeling for various applications, including speech and image recognition. She received the Best Paper Award from the Institute of Electronics, Information, and Communication Engineers of Japan, in 1993.

...