Automatic detection of the upper respiratory tract infection from speech using intrinsic mode multi-domain feature fusion

Pankaj Warule, Shubham Anjankar, Vishal Aher,
Sudhansu Sekhar Nayak, Siba Prasad Mishra

Abstract

Creating non-invasive diagnosis procedures using speech is a very promising research area in biomedical engineering. Screening for an upper respiratory tract infection (URTI) or common cold using speech signals may be advantageous in terms of preventing its spread. In this study, we have proposed the various intrinsic mode multi-domain feature fusions for diagnosing the URTIs. First, each active speech frame is decomposed into several intrinsic mode functions (IMFs) using variational mode decomposition to capture the non-linear and non-stationary characteristics of pathological speech. Then spectral and entropy domain features are extracted from each IMF and used as features for classification. The URTIC and the newly recorded PREC-RU URTIS database are employed to evaluate the efficacy of the proposed features. A transformer-based framework is employed to evaluate the discriminative capacity of the features, focusing on their capability to simulate long-range relationships with features. The proposed features outperform state-of-the-art methods, achieving a UAR of 70.82% and 72.24%, respectively, on the URTIC database's development and test partitions and 76.06% on the PREC-RU URTIS database. The results emphasize the efficacy of integrating intrinsic mode spectral and entropy features for reliable URTI identification.

Keywords: Common cold, Intrinsic mode entropy features, Intrinsic mode spectral features, Transformer model, Upper respiratory tract infection, Variational mode decomposition.

MSC 2020: 68T10, 68T07, 92C32.

©2026 by Computer Science Journal of Moldova
doi:10.56415/csjm.v34.04

1 Introduction

The speech signal carries a wealth of information regarding the speaker. It contains both the linguistic element of the message that the speaker desires to communicate and their paralinguistic aspects, such as current emotional state, health condition, age, and gender [1]. The current research explores speech signals for correctly identifying these paralinguistic aspects presented in speech signals. The acoustic-prosodic characteristics of speech are modulated with various health-related implications because of the complex nature involved in their production. Currently, diagnosing a wide range of medical disorders using speech signals is gaining popularity because of their non-invasive nature and the simplicity of remote transmission. With the complicated nature of the speech production mechanism and the importance of cognitive and physiological systems like the brain and respiratory system to human health and wellness, slight variations in a speaker's psychological and physical state may affect their ability to regulate their vocal apparatus. These changes caused by various health conditions alter the acoustic characteristics of the resulting speech. These alterations in speech can be measured by extracting suitable features for diagnosing the person's health condition.

Upper respiratory tract infections (URTI), such as the common cold, are humans' most prevalent infectious illnesses. Its symptoms include cough, sore throat, sneezing, nasal congestion, and discharge. These symptoms affect the speech produced during the URTIs. As a result, we can diagnose the URTI from the speech signal. The speech of an individual during a URTI is termed "URTI-affected speech," while the speech of a healthy individual is termed "healthy speech."

The URTIs, including influenza and the common cold, represent a significant global health burden, with annual estimates of 3–5 million severe cases and approximately 290,000–650,000 respiratory-related fatalities worldwide [1]. The two most efficient ways to stop the spread of these infectious illnesses are early diagnosis and social isolation [1]. The common cold and other associated disorders can be diagnosed early using speech signals. It may provide useful information for remotely monitoring patients' health. Speaker and speech recognition systems are often trained using healthy speech. When these systems are evalu-

ated with URTI-affected speech, their performance may be degraded. Thus, URTI-affected speech analysis can be utilized to improve speaker and speech recognition systems.

Some research has been carried out to detect the URTIs from the speech. Tull et al. [2] examined the effect of the common cold on Mel frequency cepstral coefficients (MFCC). They found that the second and third coefficients show significant differences for cold and healthy speech. Deb et al. [3] noticed that cold speech has a higher average amplitude than healthy speech, and cold speech has a shorter duration than healthy speech. Also, cold speech has differing formants and pitch frequency values compared to healthy speech. The ComParE-2017 Challenge for the identification of cold speech utilized the URTIC database [4]. The baseline of this challenge was derived using bag-of-audio-words (BoAW) and ComParE-2013 feature sets. Cai et al. [5] employed constant Q cepstral coefficients (CQCC) and MFCC for classifying URTI-affected and healthy speech. Huckvale and Beke [6] assess modulation spectra, spectral distribution, vowel spectra, and voice quality features for classifying URTI-affected and healthy speech. Kao et al. [7] proposed discriminative autoencoders for discriminating between URTI-affected and healthy speech. Vicente et al. [8] created a Fisher vector (FV) based on the generative GMM and 39-dimensional MFCC features. Deb et al. [3] used variational mode decomposition (VMD)-based features for classifying URTI-affected and healthy speech. Deb et al. [9] classified URTI-affected and healthy speech using MFCC, linear predictive coding (LPC) features, and a DNN classifier. The SMOTE-Tomek links were utilized to address data imbalance. Warule et al. [10] established the discriminative power of voiced/unvoiced speech regions in separating healthy and URTI-affected speakers, while their subsequent work [11] demonstrated the efficacy of sinusoidal model-based features with DNN classifiers. In another study, Warule et al. [12] illustrated the application of various spectral components and a transformer-based model employing a focal loss function for the classification of URTI-affected and healthy speech signals.

Speech signals are often non-stationary as their statistical properties change over time [1]. The VMD is particularly useful for analyzing

non-stationary signals because it adapts to the local properties of the signal [13]. This adaptability helps capture the time-varying characteristics of speech and enables improved representation of subtle acoustic variations present in pathological speech signals. The VMD decomposes non-stationary signals into a finite set of band-limited intrinsic mode functions (IMFs). This decomposition facilitates the analysis of different frequency components of the speech signal separately, which can enhance the extraction of discriminative acoustic features. The VMD has been applied to various signal-processing tasks, including speech analysis. In the context of speech processing, this property helps in isolating meaningful frequency bands associated with speech production mechanisms. Speech signals consist of various frequency components that correspond to different linguistic and acoustic information. Each IMF represents a specific frequency component and captures its variation over time. By decomposing the signal into IMFs, VMD separates these components and allows for the analysis of each frequency component individually. The URTI-affected speech has more energy in its low-frequency components and less energy in its high-frequency components compared to healthy speech [14]. So, we have considered the hypothesis that analyzing these IMFs can reveal important temporal and spectral characteristics of individual frequency components of the speech signal for classifying URTI-affected and healthy speech. These IMFs may provide a flexible and adaptive approach for feature extraction for the detection of URTI using speech and improve the ability of the proposed framework to capture URTI-related acoustic variations.

In this study, we have proposed the VMD-based intrinsic mode multi-domain feature fusion for diagnosing the URTI. The multi-domain feature fusion combines pertinent data from speech representations compared to single-domain methods. Then spectral features, such as MFCC, Mel frequency magnitude coefficients (MFMC), Mel-spectrogram, and entropy features, such as approximate entropy (AE), increment entropy (IE), and permutation entropy (PE), are extracted from each IMF to generate multi-domain features for classification. The spectral features capture formant and harmonic distortions in speech, while entropy-based measures measure non-linear dynamics, like vocal

instability. The proposed approach models both localized distortions and systemic speech degradation caused by URTIs by combining these domains. This fusion reduces feature bias and makes the results more generalizable, since no particular domain fully captures the complexity of the pathology from speech. This kind of hybrid strategy may be important for URTIs because of their wide range of symptoms. The performance of the proposed intrinsic mode multi-domain feature fusion method is evaluated on the URTIC database and the newly recorded PREC URTI and RU URTI databases. The following are the significant contributions of this paper: (1) A non-invasive diagnostic method for the upper respiratory tract infections (URTI) using speech signals; (2) Recording of new PREC-RU URTIS database; (3) The application of IMFs to model anomalous patterns in a speech during URTI; (4) The proposal of intrinsic mode multi-domain feature fusion to extract useful information for classifying URTI-affected and healthy speech; (5) The optimization of transformer architecture for the classification of URTI-affected and healthy speech.

The paper is structured as follows: The URTIC and PREC-RU URTIS databases used in this investigation are explained in Section 2. The proposed methodology for discriminating between URTI-affected and healthy speech is discussed in Section 3. The results of the proposed features and discussion are given in Section 4. The study's conclusion is provided in Section 5.

2 Databases

This analysis employed three distinct databases. The initial database is the URTIC database. In addition to the URTIC database, we have recorded the PREC-RU URTIS database. These databases are discussed below:

2.1 URTIC Database

We utilized the upper respiratory tract infection corpus (URTIC) database from INTERSPEECH Computational Paralinguistics Challenge ComParE-2017 [4]. It has been provided by the Institute of Safety Technology, University of Wuppertal, Germany. It comprises 28652 speech samples of 382 males and 248 females (630 subjects).

Out of 630 subjects, 111 had a cold, and 519 were in good health. The participants were asked to carry out various activities provided on a computer screen. The participants were directed to read short stories. Participants were also asked to speak 1 to 40 numbers and driving assistance commands. Also, the participants were asked to describe a recent weekend, a favorite vacation, a picture, etc. The session duration was 15 minutes to 2 hours, with a different number of exercises for each subject. The recordings were separated into 28,652 samples, each lasting between 3 and 10 seconds. These 28,652 samples are grouped into 3 speaker-independent partitions as described in Table 1.

Table 1. Details of the URTIC database

Partition	URTI	Healthy	Total
Train	8,535	970	9,505
Development	8,585	1011	9,596
Test	8,656	895	9,551

2.2 PREC-RU URTIS Databases

This upper respiratory tract infection speech (URTIS) database is recorded jointly at Pravara Rural Engineering College (PREC), Loni, Ahilyanagar, Maharashtra, India, and Ramdeobaba University (RU), Nagpur, Maharashtra, India. It consists of two classes of speech: URTI-affected speech and healthy speech. The URTI-affected speech has been recorded from an individual infected with URTI or the common cold. The healthy speech is recorded from the same individuals without pathology and without any stressful situations. The URTI-affected speech is captured initially, followed by the healthy speech from the same individual after recovery from the URTI (after 10 to 12 days). The data is collected from 40 individuals (28 males and 12 females). The age range of the individuals is from 16 to 40 years. An omnidirectional microphone with an adjustable position is utilized for recording the voice files. The data is captured through Audacity software. The data is captured at a sampling rate of 48 kHz and subsequently downsampled to 16 kHz. The final database has 960 speech files for each

class (24 samples of 40 individuals). Out of these 40 individuals, 20 are from PREC, and 20 are from RU. Each speech file has a duration of around 4 seconds. For the purpose of recording, 24 predefined English sentences [15] have been selected as mentioned below:

1. The three story house was built of stone.
2. Ship maps are different from those for planes.
3. Four hours of steady work faced us.
4. Thieves who rob friends deserve jail.
5. The sky that morning was clear and bright blue.
6. Hold the hammer near the end to drive the nail.
7. The streets are narrow and full of sharp turns.
8. The water in this well is a source of good health.
9. We don't get much money, but we have fun.
10. The price is fair for a good antique clock.
11. His shirt was clean but one button was gone.
12. Where were they when the noise started.
13. Every word and phrase he speaks is true.
14. The way to save money is not to spend much.
15. Footprints showed the path he took up the beach.
16. A round hole was drilled through the thin board.
17. The chair looked strong but had no bottom.
18. Next Sunday is the twelfth of the month.
19. Wood is best for making toys and blocks.
20. The houses are built of red clay bricks.
21. The train brought our hero to the big town.
22. Use a pencil to write the first draft.
23. It's easy to tell the depth of a well.
24. The pencil was cut to be sharp at both ends.

These two databases ensure the robustness and generalizability of our proposed classification system; speech samples were collected from two distinct geographical locations in Maharashtra, India: Ahilyanagar and Nagpur. These cities differ in climate conditions and demographic factors, which may influence vocal characteristics in individuals with URTIs. There are subtle accentual and dialectal variations in these two cities. While Ahilyanagar's speech patterns reflect a rural influence,

Nagpur's urban population exhibits the typical Vidarbha sub-dialect. By incorporating these accent-specific nuances from two regions, we ensure that the proposed features are robust to phonetic variability, avoiding bias toward a homogeneous accent group. This design choice aligns with real-world scenarios where URTI screening tools must generalize across linguistic diversity.

3 Method

The diagram presents the proposed method for detecting URTIs in Fig. 1. It consists of pre-processing of speech, proposed intrinsic mode multi-domain feature extraction, and classification using a transformer model.

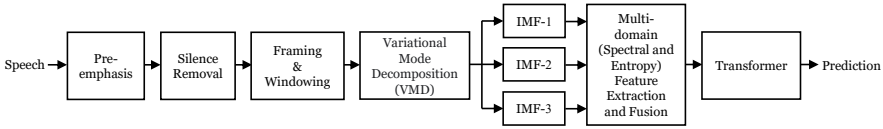

Figure 1. The proposed method for detecting URTIs

3.1 Pre-processing

Pre-processing comprises pre-emphasis, silence removal, framing, and windowing. First, the speech signal is emphasized with a pre-emphasis filter. Speech is generated when the glottal excitation is filtered by the vocal tract. The nose and throat are impacted by the common cold. The vocal tract is impacted during the production of speech since the nose and throat are key components of the vocal tract. First, the glottal source has a magnitude attenuation of around 12dB/octave [16]. Then it is countered by a 6dB/octave rise owing to the influence of lip radiation, resulting in an overall fall of the speech's excitation spectrum, which is approximately 6dB/octave at frequencies greater than 1kHz. Hence, with an overall decline of 6dB/octave, higher frequency information gets attenuated. A pre-emphasis filter can enhance the higher frequency information by compensating for the overall decline of 6dB/octave. It produces amplitudes that are comparable for all formants. A pre-emphasis filter is just a high-pass filter that flattens the signal's spectrum by enhancing the spectrum at higher frequencies [17].

In the next stage, silence regions are removed from the enhanced speech signal using short-term energy in order to obtain active speech segments, which include both voiced and unvoiced components. Then the active speech segments are subdivided into a number of frames with a duration of 20 ms and an overlap of 10 ms. In a broad range, the speech signal is non-stationary, but its characteristics are presumed to be stable over short durational frames. Hence, the active speech is divided into short durational frames. Choosing the proper frame duration is critical because the signal properties within the frame will significantly change with a longer frame duration. Finally, each frame is processed with the Hamming window to eliminate signal discontinuities at frame ends.

3.2 Variational mode decomposition (VMD)

The VMD is a multiresolution, non-recursive analysis approach that is used to express a non-stationary signal as a number of sub-signals or intrinsic mode functions (IMF) [13]. Each IMF is determined relative to the central frequency (ω_i). The initial constraint optimization problem for estimating the number of IMF and center frequencies is given by [13] as follows:

$$\min_{\{m_i\},\{\omega_i\}} = \left\{ \sum_{i=1}^{I} \left\| \frac{\delta}{\delta n} \left[(\delta(n) + \frac{j}{\pi n}) * m_i(n) \right] e^{-j\omega_i n} \right\|_2^2 \right\} \quad \text{subject to} \sum_{i=1}^{I} m_i(n) = s(n), \tag{1}$$

where $s(n)$ is the input speech frame, $m_i(n)$ and $\omega_i(n)$ represent the IMF signal and center frequency corresponding to i^{th} IMF, and I is the number of IMFs. The optimization problem in Eq. (1) can be solved by using augmented Lagrangian (L) as follows [13]:

$$L(\{m_i\},\{\omega_i\},\gamma(n)) = \alpha \sum_{i=1}^{I} \left\| \frac{\delta}{\delta n} \left[(\delta(n) + \frac{j}{\pi n}) * m_i(n) \right] e^{-j\omega_i n} \right\|_2^2 + \left\| s(n) - \sum_{i=1}^{I} m_i(n) \right\|_2^2 + \left\langle \gamma(n), s(n) - \sum_{i=1}^{I} m_i(n) \right\rangle, \tag{2}$$

where γ is the Lagrangian multiplier and α is the quadratic penalty factor. The IMFs and center frequencies for speech frames are obtained as

$$\tilde{m}_i(n) = \underset{\{m_i\}}{arg\ min} \left\{ \alpha \left\| \frac{\delta}{\delta n} \left[(\delta(n) + \frac{j}{\pi n}) * m_i(n) \right] e^{-j\omega_i n} \right\|_2^2 \right\} + \left\| s(n) - \sum_{i=1}^{I} m_i(n) + \frac{\gamma(n)}{2} \right\|_2^2, \tag{3}$$

$$\tilde{\omega}_i = \underset{\{\omega_i\}}{arg\ min} \left\{ \left\| \frac{\delta}{\delta n} \left[(\delta(n) + \frac{j}{\pi n}) * m_i(n) \right] e^{-j\omega_i n} \right\|_2^2 \right\}. \tag{4}$$

Fourier isometry and spectrum analysis can be used to solve Eq. 3. It is given by [13]:

$$\tilde{m}_i(\omega) = \underset{\{m_i(\omega)\}}{arg\ min} \left\{ \alpha \| j\omega \left[(1 + sgn(\omega + \omega_i)) * m_i(\omega + \omega_i) \right] \|_2^2 \right\} + \left\| s(\omega) - \sum_{i=1}^{I} m_i(\omega) + \frac{\gamma(\omega)}{2} \right\|_2^2. \tag{5}$$

The IMFs are determined iteratively by solving Eq. (5), which is given by [13]:

$$\tilde{i}_{\tilde{i}}(\omega) = \frac{s(\omega) - \sum_{i \neq \tilde{i}} m_i(\omega) + \frac{\gamma(\omega)}{2}}{1 + 2\alpha(\omega - \omega_i)^2}. \tag{6}$$

The IMF center frequency is updated by [13] as:

$$\tilde{\omega}_i = \frac{\sum_\omega \omega \left| m_i(\omega) \right|^2}{\sum_\omega \left| m_i(\omega) \right|^2}. \tag{7}$$

The VMD uses several parameters for the decomposition of the signal into various IMFs as follows: the number of modes or IMFs ($I = 3$), mode center frequencies (ω_i) initialized uniformly, the balancing parameter of the data-fidelity constraint ($\alpha = 120$), and the tolerance of the convergence criterion ($tol = 10^{-7}$). To determine the optimal number of IMFs, different values (3, 4, and 5 IMFs) were evaluated. The results showed that three IMFs provided the best classification performance, while increasing the number of IMFs did not improve the results and sometimes introduced redundant information.

3.3 Intrinsic mode multi-domain feature extraction

In this study, we have extracted the various intrinsic mode multi-domain features for diagnosing the URTIs. First, the voiced speech frame is decomposed into several IMFs using VMD. Then spectral features such as MFCC, MFMC, and Mel-spectrogram coefficients are extracted from each IMF and termed as intrinsic mode Mel frequency cepstral coefficients (IMMFCC), intrinsic mode Mel frequency magnitude coefficients (IMMFMC), and intrinsic mode Mel-spectrogram coefficients (IMMSC). Similarly, entropy features, such as AE, PE, and IE, are extracted from each IMF and termed as intrinsic mode approximate entropy (IMAE), intrinsic mode permutation entropy (IMPE), and intrinsic mode increment entropy (IMIE) and used as features for classification. These extracted features are discussed as follows:

3.3.1 Intrinsic mode Mel frequency cepstral coefficients (IMMFCC)

The 20-dimensional MFCC features are extracted from each IMF and termed them as intrinsic model Mel frequency cepstral coefficients (IMMFCC). The MFCC is crucial in numerous audio-related applications, such as voice and speaker recognition, music genre categorization, and sound event detection [18]. They capture the fundamental acoustic attributes of a sound from its frequency spectrum. The MFCC algorithm emulates the frequency perception of the human auditory system, exhibiting greater sensitivity to lower frequencies compared to higher ones. The configuration of the vocal tract modulates human speech, which can be represented by the envelope of the short-time power spectrum. The MFCC encapsulates this envelope. This procedure requires applying a discrete Fourier transform (DFT) to each frame to derive a spectrum, thereafter converting it to the Mel scale. It is a perceptual frequency scale that categorizes frequencies according to their perceived resemblance to human auditory perception. The transformation from the Hertz scale to the Mel scale is expressed by the equation Eq. (8):

$$M(f) \;=\; 1125\ln(1+\frac{f}{700}), \tag{8}$$

where f denotes the frequency in Hertz. In the Mel scale, 40 uniformly distributed triangular band-pass filters are designed with a response of 1 at the center frequency, tapering linearly to 0 at the center frequencies of neighboring filters. The filter banks are reverted to the Hertz scale and normalized. The logarithm of the power spectrum is further processed by the normalized Mel filter bank to derive the filter bank coefficients. Ultimately, the discrete cosine transform (DCT) is employed to obtain the decorrelated MFCC coefficients.

We have extracted 20 MFCC coefficients from three IMFs of the speech frame to get 60-dimensional IMMFCC features for each speech frame. These IMMFCC features are represented as [IM_1MFCC_1, IM_1MFCC_2, ..., IM_1MFCC_{20}, IM_2MFCC_1, IM_2MFCC_2, ..., IM_2MFCC_{20}, IM_3MFCC_1, IM_3MFCC_2, ..., IM_3MFCC_{20}].

3.3.2 Intrinsic mode Mel frequency magnitude coefficients (IMMFMC)

In this study, we have extracted MFCC features from each IMF and termed them as IMMFMC features. The DFT of each IMF is obtained to analyze the spectral properties of each IMF. The different frequency components present in the speech frame are evaluated using the DFT. The magnitude of the Fourier coefficients is calculated in the subsequent stage. As the MFCC is calculated using the squared amplitude, large signal components are affected by significant variations [19]. So, instead of using the squared magnitude as for MFCC, we utilized the magnitude of DFT coefficients while deriving MFMC features.

The magnitude spectrum is then converted to the Mel spectrum. Next, the Mel spectrum is divided into M equal-width bands that overlap by 50% on the Mel scale, spanning from zero to half of the sampling frequency in terms of frequency limits. The uniform M bands of the Mel scale are converted back into a linear frequency scale [19]. On the linear frequency scale, the consistent intervals of the Mel scale transform into uneven ones. Specifically, lower frequency bands occupy narrower bandwidths, while higher frequency bands span wider ranges. This discrepancy in bandwidth allocation results in enhanced resolution at lower frequencies compared to higher frequencies [19]. Triangular windows are used to modify these non-uniform bands. Since there is a 50%

overlap between subsequent windows, the left vertex of the window corresponds to the middle vertex of the preceding window. A triangular window attenuates the frequency component's magnitudes on each side of the middle frequency without changing the middle frequency component's magnitude. In the human auditory perceptional process, a triangular filter is employed to mimic how the center frequency of a critical band masks other frequencies [19].

The mathematical operation that corresponds to the response of human hearing is the logarithm. Finally, the logarithm of the sum of Mel band magnitudes gives IMMFMC features. We have extracted 20 MFMC coefficients from three IMFs of the speech frame to get 60-dimensional IMMFMC features for each speech frame. These IMMFMC features are represented as [IM_1MFMC_1, IM_1MFMC_2, ..., IM_1MFMC_{20}, IM_2MFMC_1, IM_2MFMC_2, ..., IM_2MFMC_{20}, IM_3MFMC_1, IM_3MFMC_2, ..., IM_3MFMC_{20}].

3.3.3 Intrinsic mode Mel spectrogram coefficients (IMMSC)

The 40 coefficients of the Mel-spectrogram are extracted from each IMF and termed intrinsic model Mel-spectrogram coefficients (IMMSC). The frequency domain representation of a speech signal lacks the representation of temporal information. Similarly, there is no availability of data regarding the frequency characteristics of the temporal domain signal. Consequently, a spectrogram is utilized to do an analysis in both the temporal and spectral domains. The spectrogram visually represents the variations in speech intensity concerning frequency and time. The Mel-spectrogram bears a resemblance to the sound frequency reception pattern intrinsic to the human auditory system [20]. Human frequency perception is non-linear, indicating that individuals excel at differentiating frequencies in low-frequency ranges rather than in high-frequency ranges. For instance, humans can readily differentiate between frequencies of 500 and 1000 Hz, while they find it challenging to discern between 10000 and 10500 Hz. The frequencies are transformed to the Mel scale to generate the Mel spectrogram. To extract these features, voice signals are divided into frames, and each frame is decomposed into IMFs using VMD. Then the DFT is applied to each IMF to get the frequency spectrum. The frequency spectrum is divided

into uniform intervals according to frequency to derive the Mel scale of the signal frame.

We have extracted 40 Mel-spectrogram coefficients from three IMFs of the speech frame to get 120-dimensional IMMSC features for each speech frame. These IMMFMC features are represented as [IM_1MSC_1, IM_1MSC_2, ..., IM_1MSC_{40}, IM_2MSC_1, IM_2MSC_2, ..., IM_2MSC_{40}, IM_3MSC_1, IM_3MSC_2, ..., IM_3MSC_{40}].

3.3.4 Intrinsic mode approximate entropy (IMAE)

In this study, we have extracted AE features from each IMF and termed them as IMAE features. The AE measures the complexity and regularity of time series data [21]. It is utilized in numerous applications, such as cardiovascular time series signal analysis, differentiation of voiced and unvoiced speech segments, identification of speech emotion [22], classification of complex time series systems [21], and detection of modulation in communication systems. The AE delivered from each IMF may distinguish between URTI-affected and healthy speech by examining complexity and regularities.

Let IMF vectors be represented by a discrete time sequence ($s(n) = [s(1), s(2), \ldots, s(M)]$, where M denotes the sample size of the sequence $s(n)$. The total number of d-dimensional embedded vectors for the sequence $s(n)$ can be expressed as follows [21]:

$$s_p(l) = [s(l), s(l+1), ..., s(l+d-1)], 1 \leq l \leq M-d+1. \quad (9)$$

The absolute difference between the two sequences, $s_p(l)$ and $s_p(k)$, is defined as:

$$d[s_p(l), s_p(m)] = \max_{k=1,2,\ldots,d} |s_p(l+k-1) - s_p(m+k-1)|. \quad (10)$$

Let $W_l^d(\nu)$ represent the ratio of the quantity of vectors satisfying the condition $d[s_p(l), s_p(m)] \leq \nu$, where $\nu \geq 0$, to the overall quantity of vectors. Mathematically, $W_l^d(\nu)$ is expressible as

$$W_l^d(\nu) = \frac{M_d}{M-(d+1)\tau}. \quad (11)$$

In this context, M_d represents the total count of vectors that satisfy the condition $d[s_p(l), s_p(m)] \leq \nu$, where τ and ν denote the time delay

and threshold value, respectively. By utilizing logarithms on Eq. (11) and subsequently calculating the average for the value of l, we obtain

$$\phi^d(\nu) = [M - (d+1)\,\tau]^{-1} \sum_{l=1}^{M-(d+1)\tau} ln(W_l^d(\nu)). \tag{12}$$

The AE for the s(n) is now provided by:

$$AE\,(d, \nu, M, \tau) = \phi^{d+1}(\nu) - \phi^d(\nu). \tag{13}$$

In this study, the IMAE is computed using embedding dimension $d = 2$, time delay $\tau = 1$, and tolerance parameter $\nu = 0.4\sigma$, where σ represents the standard deviation of the signal. We calculated three-dimensional AE features from all three IMFs to get a 9-dimensional IMAE feature. This feature vector is denoted as [IM_1AE_1, IM_1AE_2, IM_1AE_3, IM_2AE_1, IM_2AE_2, IM_2AE_3, IM_3AE_1, IM_3AE_2, IM_3AE_3].

3.3.5 Intrinsic mode permutation entropy (IMPE)

In this study, we have extracted PE features from each IMF and termed them as IMPE features. The PE features are utilized by many researchers in diverse applications, such as speech emotion recognition [22], image classification, biomedical signal processing, and bearing fault detection. It correlates directly with signal randomness. The variety in randomness or uncertainty of the IMFs among URTI-affected and healthy speech signals may result in equivalent changes in the computed PE among these signals. We anticipated that the utilization of the PE would markedly improve classification performance. Let $s(n) = [s(1), s(2), \ldots, s(M)]$ denote a discrete time sequence corresponding to each IMF, with M being the maximum number of samples in $s(n)$. The quantity of possible d-dimensional embedded vectors that can be produced using $s(n)$ is as follows [22]:

$$s(d) = [s(\tau), s(\tau+1), ..., s(\tau+(d-1))], \tag{14}$$

where τ represents the time delay and d denotes the embedded vector dimension, respectively. A pattern $g(g = [l_0, l_1, ..., l_{d-1}])$ is formed by arranging the $s(d)$ in ascending order as outlined in [23]:

$$s(\tau + l_0) \leq s(\tau + l_1), ..., \leq s(\tau + l_{d-2}) \leq s(\tau + l_{d-1}). \tag{15}$$

At last, the PE is calculated as

$$PE(d, \tau, M) = -\sum_{j=1}^{d!} g(j) \log_2 g(j). \tag{16}$$

In this work, the embedding dimension is set to $d = 4$ and the time delay is set to $\tau = 2$ for the computation of IMPE. We calculated three-dimensional PE features from all three IMFs to get a 9-dimensional IMPE feature. This feature vector is denoted as [IM_1PE_1, IM_1PE_2, IM_1PE_3, IM_2PE_1, IM_2PE_2, IM_2PE_3, IM_3PE_1, IM_3PE_2, IM_3PE_3].

3.3.6 Intrinsic mode increment entropy (IMIE)

In this study, we have extracted IE features from each IMF and termed them as IMIE features. The IE is a quantitative measure used to evaluate the complexity of time series data. The unpredictability of a signal is closely correlated with its information entropy. When a signal demonstrates significant unpredictability, the incremental entropy value is heightened; conversely, when the signal is more predictable, it is reduced. The IE primarily emphasizes quantifying the signal's increments, which encapsulate the characteristics of the signal's dynamic fluctuations. Let $s(n)$, consisting of M samples. An increment vector $p(l)$, for $1 \leq l \leq M-1$, can be derived from $s(n)$, expressed as $p(l) = s(l+1) - s(l)$. The total quantity of d-dimensional embedded vectors generated by $p(l)$ [24] is:

$$p(k) = [p(k), p(k+1), ..., p(k+d-1)], 1 \leq k \leq M-d. \tag{17}$$

Every component in $p(k)$ corresponds to q_{k+j} and a_{k+j}, where $j = 1, 2, ..., M-d$, conveying information regarding the sign and magnitude of adjacent elements. Each time series vector $p(k)$ is correlated to a 2D word $w_k = \bigcup j = 0^{d-1} qk + ja_{k+j}$, and there exists a $M-d$ word corresponding to each time series vector $p(k)$. Given the specified values of d and D, a word including $2d$ letters has $(2D+1)^d$ potential variations, where D is the quantization resolution. Let w_n represent one of the distinct words in w_n; then, the probability of each w_n can be determined as follows [24]:

$$p(w_n) = \frac{Q(w_n)}{M-d}. \tag{18}$$

Finally, IE is calculated as [24]:

$$IE(d, D) = -\sum_{n=0}^{(2D+1)^d} p(w_n) log(p(w_n)). \tag{19}$$

In this work, to calculate IMIE, the embedding dimension and quantization resolution are set to $d = 3$ and $D = 4$, respectively. We calculated one-dimensional IE features from all three IMFs to get a 3-dimensional IMIE feature. This feature vector is denoted as [IM_1IE, IM_2IE, IM_3IE].

The extracted features consist of 60-dimensional IMMFCC, 60-dimensional IMMFMC, 120-dimensional IMMSC, 9-dimensional IMAE, 9-dimensional IMPE, and 3-dimensional IMIE for each frame in the utterance. The frame-level features are averaged over all frames of the utterance to produce a 261-dimensional utterance-level feature vector, which is subsequently applied to the classifier for distinguishing between URTI-affected and healthy speech signals.

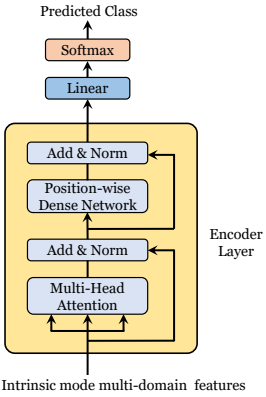

Figure 2. The transformer encoder model

3.4 Classification

In this work, we used the transformer encoder and DNN classifiers to discriminate between URTI-affected and healthy speech. The intrinsic mode multidomain feature vectors are applied to classifiers during the training and testing of the model.

3.4.1 Transformer encoder

The transformer was initially presented by Vaswani et al. [25] and has subsequently gained appreciable prominence in many natural language processing (NLP) tasks. It was developed using the attention mechanism, which enables parallelization and creates a thorough link between the input and output [25]. The diagram illustrating the suggested transformer encoder model for URTI speech categorization is presented in Fig. 2. The incorporation of multi-head attention in the transformer encoder layer is intended to augment the model's ability to comprehend long-term dependencies. The multi-head attention mechanism entails the concurrent evaluation of information from multiple positions within the retrieved features. The encoder layer of the transformer architecture features an FFNN that utilizes the rectified linear unit (ReLU) activation function and layer normalization. The transformer architecture used in this study utilized four encoder layers, each with eight self-attention heads. Each was succeeded by an FFNN with 128 dimensions and layer normalization. The final result underwent further processing using the linear and softmax layers to detect URTI. The extracted utterance-level feature vectors are first normalized and then used as input to the classifier. The Adam optimizer, configured with a learning rate of 0.0001 and utilizing sparse categorical cross-entropy loss, is employed for training. The model is trained using mini-batches with a batch size of 64 for multiple epochs, and the best-performing model is selected based on the development set performance.

3.4.2 Deep neural network (DNN)

A DNN is a type of artificial neural network that has several hidden layers. The neuron unit is the fundamental organizing element of DNN. The DNN has extremely high computing capabilities for handling large amounts of data. In our experiment, we utilized a DNN with three hidden layers. The ReLU and sigmoid activation functions are used in the hidden and output layers of the DNN, respectively. The tuned hyperparameters of the DNN are detailed in Table 2. The performance is evaluated using unweighted average recall (UAR) or balanced accuracy because recognition accuracy of both classes is dominant.

Table 2. Details of tuned hyperparameters of the DNN

Parameter	Value
No. of hidden layers	3
Hidden layer neurons	512, 256, 64
Loss function	Cross-entropy loss
Optimization algorithm	Adam
Learning rate	0.0001

4 Results & Discussion

This section explores the results achieved using the intrinsic mode multi-domain features for discriminating between URTI-affected and healthy speech. To analyze the statistical significance of the intrinsic mode multi-domain features, a T-test is conducted on the intrinsic mode multi-domain features. A T-test provides a means to compare the means of two groups. In a T-test, for each feature, both the t-value and the p-value are computed. Features with higher t-values and lower p-values (less than 0.0001) are deemed more pertinent for classification. The t-values and p-values between URTI and healthy speech classes are detailed in Table 3. These values are computed for the randomly selected intrinsic mode multi-domain features extracted from the PREC-RU URTIS database. The acquired t-values are high, and the associated p-values are below 0.0001. This reveals that the intrinsic mode multi-domain features may be effectively categorized into URTI and healthy speech classes and can be used for classification.

The effectiveness of the intrinsic mode multi-domain features is assessed on the URTIC database as well as the newly developed PREC-RU URTIS database using the transformer encoder model and DNN classifiers. The results in Table 4 show the performance of intrinsic mode multi-domain features for detecting URTI using the DNN and Transformer classifiers on the URTIC database. The IMMFMC features attended the highest test-set UAR score of 69.37% with DNN and 68.96% with Transformer. The IMMS features worked better with the Transformer classifier. The transformer model with IMMS achieved

Table 3. The results of the T-test for the IMMFMC features

Feature	t-value	p-value
IM_1MFCC_2	5.2306	<0.0001
IM_1MFCC_{11}	5.5010	<0.0001
IM_2MFCC_7	5.2108	<0.0001
IM_3MFCC_5	15.5068	<0.0001
IM_3MFCC_{17}	7.7068	<0.0001
IM_3MFCC_4	9.9076	<0.0001
IM_1MFMC_8	5.4556	<0.0001
IM_1MFMC_{15}	4.4361	<0.0001
IM_2MFMC_7	4.0475	<0.0001
IM_2MFMC_{14}	10.0964	<0.0001
IM_3MFMC_3	6.0012	<0.0001
IM_1MSC_{12}	7.0863	<0.0001
IM_1MSC_{28}	6.4216	<0.0001
IM_2MSC_5	5.1928	<0.0001
IM_3MSC_{19}	7.1250	<0.0001
IM_3MSC_{33}	5.4556	<0.0001
IM_1AE_3	4.2492	<0.0001
IM_2AE_2	5.8335	<0.0001
IM_1PE_2	4.8610	<0.0001
IM_2IE	4.5668	<0.0001

67.54% and 67.87% UARs on the development and test sets of the URTIC database. Conversely, the DNN achieved 66.35% and 65.74% UARs on the development and test sets. The intrinsic mode entropy features (IMAE+IMPE+IMIE) achieved UARs of 65.12% and 66.31% on the development and test sets. The multi-domain feature fusion method, which combined spectral and entropy features, got the best results, with UAR scores of 70.82% and 72.24% with the transformer model on the development and test sets, respectively. These results are higher compared to the DNN classifier. The DNN achieved UARs of 69.24% and 70.41% on the development and test sets, respectively.

Table 5 shows the performance evaluation of intrinsic mode multi-domain features on the newly recorded PREC-RU URTIS database. It shows the UAR percentages for both DNN and Transformer classifiers. We have performed speaker-independent fivefold cross-validation while reporting these results. The IMMFMC achieved the highest performance compared to the other spectral features, with 74.61% UAR using DNN and 73.23% using Transformer. The IMMS and IMMFCC features also achieved strong results, with UARs above 72% for both classifiers, indicating that spectral representations are highly discriminative for URTI detection. The intrinsic mode entropy features (IMAE+IMPE+IMIE) achieved UARs of 72.37% with DNN and 70.42% with Transformer. The multi-domain feature fusion technique,

which incorporated both spectral and entropy characteristics, obtained the highest overall results, with 76.06% UAR with DNN and 75.86% UAR with Transformer.

In this work, we have proposed intrinsic mode multi-domain features for discriminating URTI-affected and healthy speech. The results demonstrated that the proposed intrinsic mode multi-domain features efficiently distinguish between URTI-affected and healthy speech. Table 6 compares the proposed framework with the existing methods. In the INTERSPEECH 2017 ComParE Cold Sub-Challenge, the URTIC database was used, and the baseline results were presented as in [4]. The baseline results are produced utilizing the 6373-dimensional ComParE-2013 features and 130-dimensional BoAW features. The baseline UARs for development and test partitions are 64% and 70.2% using the ComParE feature set and 64.20% and 67.30% using BoAW features. On the URTIC database's development and test partitions, the proposed features achieved 70.82% and 72.24% UAR, respectively. Also, the intrinsic mode multi-domain feature dimension is much less compared to the baseline ComParE features. Compared to existing methods, the proposed framework using the intrinsic mode multi-domain features gives the highest results on the development as well as test partitions. This method improves the accuracy of diagnosis and facilitates practical applications in remote healthcare monitoring and early disease identification.

5 Conclusion

This investigation proposes the intrinsic mode multi-domain features for discriminating between URTI-affected and healthy speech. The effectiveness of the proposed features is examined using the URTIC database and the newly recorded PREC-RU URTIS database. The results indicate that the proposed features perform better than the baseline features, such as ComParE-2013 and BoAW features. The proposed framework demonstrates higher classification performance in comparison to state-of-the-art methods. The proposed intrinsic mode multi-domain features achieved a UAR of 70.82% and 72.24%, respectively, on the URTIC database's development and test partitions and

76.06% on the PREC-RU URTIS database. The efficacy of the method across datasets from several global regions (URTIC and PREC-RU)

Table 4. Performance of the intrinsic mode multi-domain Features on the URTIC database

Feature Domain	Features	% UAR Using DNN		% UAR Using Tranformer	
		Development Set	Test Set	Development Set	Test Set
Intrinsic Mode Spectral Features	IMMFCC	67.12	66.82	67.03	66.74
	IMMFMC	67.18	69.37	67.82	68.96
	IMMS	66.35	65.74	67.54	67.87
Intrinsic Mode Entropy Features	IMAE+IMPE+IMIE	65.12	66.31	64.97	64.42
Multi-domain Feature Fusion (Spectral + Entropy)	IMMFCC + IMMFMC + IMMS IMAE + IMPE + IMIE	69.24	70.41	70.82	72.24

Table 5. Performance of the Intrinsic mode multi-domain Features on the PREC-RU URTIS database

Feature Domain	Features	% UAR Using DNN	% UAR Using Tranformer
Intrinsic Mode Spectral Features	IMMFCC	72.61	71.23
	IMMFMC	74.61	73.23
	IMMS	73.57	72.12
Intrinsic Mode Entropy Features	IMAE+IMPE+IMIE	72.37	70.42
Multi-domain Feature Fusion (Spectral + Entropy)	IMMFCC + IMMFMC + IMMS IMAE + IMPE + IMIE	76.06	75.86

Table 6. Comparison of results of state-of-the-art approaches and proposed method

Research work	% UAR	
	Development	Test
Schuller et al. (ComParE + SVM) [4]	64.00	70.20
Schuller et al. (BoAW + SVM) [4]	64.20	67.30
Cai et al. (MFCC+GMM) [5]	64.80	-
Cai et al. (CQCC +GMM) [5]	65.40	-
Huckvale and Beke (MOD+DNN) [6]	67.95	62.10
Deb et al. (VMD+SVM) [3]	66.84	-
Kao et al. (MFCC+GMM) [7]	65.81	66.00
Vicente et al. (MFCC+GMM) [8]	63.98	66.12
Deb et al. (LPC+MFCC+DNN) [9]	67.71	-
Warule et al. (MFCC Statistics) [10]	66.12	64.92
Warule et al. (Sinusoidal model-based features+DNN) [11]	69.16	65.22
Warule et al. (Spectral features + Transformer) [12]	69.55	70.48
Proposed intrinsic mode multi-domain features + Transformer	70.82	72.24

demonstrates its capability to accommodate variations in accent and context. The results indicate significant potential for developing effective speech-based URTI screening tools, particularly for telehealth applications.

Future research may focus on enhancing population diversity, improving real-time system performance, and exploring the integration of additional biomarkers to augment accuracy and clinical utility. This study advances voice-based health diagnostics and establishes a foundation for employing speech analysis to identify further respiratory issues without the necessity of intrusive procedures.

References

[1] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[2] R. G. Tull, J. C. Rutledge, and C. R. Larson, "Cepstral analysis of "cold-speech"for speaker recognition: a second look," Ph.D. dissertation, Acoustical Society of America, 1996.

[3] S. Deb, S. Dandapat, and J. Krajewski, "Analysis and classification of cold speech using variational mode decomposition," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 296–307, 2017.

[4] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017, pp. 3442–3446.

[5] D. Cai, Z. Ni, W. Liu, W. Cai, G. Li, and M. Li, "End-to-end deep learning framework for speech paralinguistics detection based on perception aware spectrum." in *INTERSPEECH*, 2017, pp. 3452–3456.

[6] M. A. Huckvale and A. Beke, "It sounds like you have a cold! testing voice features for the interspeech 2017 computational paralinguistics cold challenge." International Speech Communication Association (ISCA), 2017.

[7] Y.-Y. Kao, H.-P. Hsu, C.-F. Liao, Y. Tsao, H.-C. Yang, J.-L. Li, C.-C. Lee, H.-S. Lee, and H.-M. Wang, "Automatic detection of speech under cold using discriminative autoencoders and strength modeling with multiple sub-dictionary generation," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 416–420.

[8] E. L. José Vicente and G. Gosztolya, "Using the fisher vector approach for cold identification," *Acta Cybernetica*, vol. 25, no. 2, pp. 223–232, 2021.

[9] S. Deb, P. Warule, A. Nair, H. Sultan, R. Dash, and J. Krajewski, "Detection of common cold from speech signals using deep neural network," *Circuits, Systems, and Signal Processing*, pp. 1–16, 2022.

[10] P. Warule, S. P. Mishra, and S. Deb, "Significance of voiced and unvoiced speech segments for the detection of common cold," *Signal, Image and Video Processing*, pp. 1–8, 2022.

[11] P. Warule, S. P. Mishra, S. Deb, and J. Krajewski, "Sinusoidal model-based diagnosis of the common cold from the speech signal," *Biomedical Signal Processing and Control*, vol. 83, 2023.

[12] P. Warule, S. Chandratre, S. P. Mishra, and S. Deb, "Detection of the common cold from speech signals using transformer model and spectral features," *Biomedical Signal Processing and Control*, vol. 93, 2024.

[13] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE transactions on signal processing*, vol. 62, no. 3, pp. 531–544, 2013.

[14] Y. Shan and Q. Zhu, "Speaker identification under the changed sound environment," in *2014 International Conference on Audio, Language and Image Processing*. IEEE, 2014, pp. 362–366.

[15] S. Deb and S. Dandapat, "Fourier model based features for analysis and classification of out-of-breath speech," *Speech Communication*, vol. 90, pp. 1–14, 2017.

[16] A. N. Ince, *Digital Speech Processing: Speech Coding, Synthesis and Recognition*. Springer Science & Business Media, 1991, vol. 155.

[17] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993.

[18] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[19] J. Ancilin and A. Milton, "Improved speech emotion recognition with mel frequency magnitude coefficient," *Applied Acoustics*, vol. 179, 2021.

[20] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, 2020.

[21] S. M. Pincus, "Approximate entropy as a measure of system complexity." *Proceedings of the national academy of sciences*, vol. 88, no. 6, pp. 2297–2301, 1991.

[22] S. P. Mishra, P. Warule, and S. Deb, "Improvement of emotion classification performance using multi-resolution variational mode decomposition method," *Biomedical Signal Processing and Control*, vol. 89, 2024.

[23] X. Li, G. Ouyang, and D. A. Richards, "Predictability analysis of absence seizures with permutation entropy," *Epilepsy research*, vol. 77, no. 1, pp. 70–74, 2007.

[24] X. Liu, A. Jiang, N. Xu, and J. Xue, "Increment entropy as a measure of complexity for time series," *Entropy*, vol. 18, no. 1, p. 22, 2016.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

Pankaj Warule,
Shubham Anjankar, Vishal Aher,
Sudhansu Sekhar Nayak, Siba Prasad Mishra

Received December 29, 2025
Revised March 12, 2026
Accepted March 20, 2026

Pankaj Warule
ORCID: https://orcid.org/0000-0001-8201-7663
Pravara Rural Engineering College
Loni, Maharastra, India
E–mail: warulepc@pravaraengg.org.in

Shubham Anjankar
ORCID: https://orcid.org/0000-0002-1057-7343
Shri Ramdeobaba College of Engineering and Management
Nagpur, Maharashtra, India
E–mail: anjankarsc1@rknec.edu

Vishal Aher
ORCID: https://orcid.org/0009-0000-7042-3920
Pravara Rural Engineering College
Loni, Maharastra, India
E–mail: aherva@pravaraengg.org.in

Sudhansu Sekhar Nayak
ORCID: https://orcid.org/0009-0007-9804-7710
Sardar Vallabhbhai National Institute of Technology
Surat, Gujarat, India
E–mail: d21ec005@eced.svnit.ac.in

Siba Prasad Mishra
ORCID: https://orcid.org/0000-0001-8076-8295
Amrita Vishwa Vidyapeetham
Bengaluru, India
E–mail: m_sibaprasad@blr.amrita.edu