Signal, Image and Video Processing (2023) 17:1785–1792
https://doi.org/10.1007/s11760-022-02389-8

ORIGINAL PAPER

Significance of voiced and unvoiced speech segments for the detection of common cold

Pankaj Warule[1] · Siba Prasad Mishra[1] · Suman Deb[1]

Received: 25 August 2022 / Revised: 22 October 2022 / Accepted: 2 November 2022 / Published online: 15 November 2022

Abstract

This work investigates the significance of the voiced and unvoiced region for detecting common cold from the speech signal. In literature, the entire speech signal is processed to detect the common cold and other diseases. This study uses a short-time energy-based approach to segment the voiced and unvoiced region of the speech signal. Then, frame-wise mel frequency cepstral coefficients (MFCC) features are extracted from the voiced and unvoiced segments of each speech utterance, and statistics (mean, variance, skewness, and kurtosis) are calculated to get the feature vector for each speech utterance. The support vector machine (SVM) is utilized to analyze the performance of features extracted from the voiced and unvoiced region. Result shows that the feature extracted from voiced segments, unvoiced segments, and complete active speech (CAS) gives almost similar results using the MFCC features and SVM classifier. Therefore, rather than processing the CAS, we can process the unvoiced speech segments, which have fewer frames compared to CAS and voiced regions of speech. The processing of solely unvoiced segments can reduce the time and computation complexity of a speech signal-based common cold detection system.

Keywords Cold speech · Mel frequency cepstral coefficients · Support vector machine · Unvoiced region · Voiced region

1 Introduction

The speech signal contains the linguistic information that the speaker wants to transmit and paralinguistic information such as emotional and health state, age, and gender of the speaker [1]. Current research efforts are being conducted to reliably and correctly identify a person's health and emotional state [2,3]. Computational paralinguistic analysis is gaining interest in evaluating various health conditions from the speech signal. Due to the intricacy of the speech production system and the involvement of physiological and cognitive systems, such as the respiratory system and the brain, slight changes in a speaker's physical and mental condition impact their ability to control their vocal apparatus [1]. Such adjustments may significantly affect the acoustic properties of the produced speech. Furthermore, since speech can be readily recorded, transmitted, and stored, speech-based analytic paradigms have the potential to become a new kind of non-invasive technology for a wide variety of health issues in the future. Cold speech is a sort of pathological speech produced by someone with a common cold or flu. The typical cold affects both the nasal and the esophagus. Speech is produced as a consequence of the vocal tract's linear filtering of stimulation source data. Because the vocal tract is engaged during speech production, the acoustic properties of cold speech vary from those of normal speech. The average amplitude of cold speech is higher than that of normal speech and the duration of cold speech is shorter than normal speech [4].

Upper Respiratory Tract Infections (URTIs) such as the Common Cold and flu are major public health problems, causing approximately 3 to 5 million cases of severe illness [1,5]. Social isolation and early diagnosis are two of the most effective strategies for reducing the spread of infectious diseases [1]. The analysis and classification of cold speech may be useful in the diagnosis of common cold and other related illnesses to stop the spread of these viral infections. The

✉ Pankaj Warule
d20ec007@eced.svnit.ac.in

Siba Prasad Mishra
ds20ec005@eced.svnit.ac.in

Suman Deb
sumandeb@eced.svnit.ac.in

[1] Department of Electronics Engineering, Sardar Vallabhbhai National Institute of Technology, Surat 395007, India

recent epidemic caused by the Covid-19 virus has brought to light the necessity of remote digital healthcare systems [6,7]. Diagnosing the common cold from a patient's speech may be beneficial for remotely monitoring patients' health. Speech-based screening system for the common cold and associated illnesses can be embedded into smart gadgets like smartphones and smart watches to monitor a person's health. Generally, speech recognition and speaker recognition systems are trained using normal speech. If these systems are tested using cold speech, system performance may degrade. Therefore, analysis of cold speech can be used to improve the performance of speech/speaker recognition and man–machine interaction [8–10].

In this study, we have analyzed the significance of voiced and unvoiced regions of speech for the classification of cold and healthy speech. Common cold symptoms include stuffy noses, hoarseness, coughing, or sneezing, which alter the speech signal by affecting the vocal cords and vocal tract [11]. Voiced and unvoiced regions of speech have distinct speech production processes and energy patterns [12]. We hypothesized that common cold affects both voiced and unvoiced speech segments. This motivates us to use the voiced and unvoiced regions independently to detect the common cold from the speech signal.

2 Database

This investigation makes use of the Upper Respiratory Tract Infection Corpus (URTIC) database, made available by the Institute of Safety Technology at the University of Wuppertal in Germany [13]. It comprises voice recordings from 630 participants, 111 of whom had a common cold and 519 healthy. The mean and standard deviation of the age of participants are 29.5 and 12.1 years, respectively. These speech samples are downsampled to 16 kHz after being recorded at 44.1 kHz.

Each participant completed a binary one-item measure based on the Wisconsin Upper Respiratory Symptom Survey (WURSS-24) in German throughout the recording process to assess their health state and rate their common cold symptoms. The global illness severity item was binarized at a threshold of 6 on a scale from 0 for not sick to 7 for very ill. The participants had to perform numerous activities that were shown on a computer monitor. The subjects were advised to read widely used phonetics short stories, such as *The North Wind and the Sun*, as well as a typical German reading passage, *Die Buttergeschichte*. In addition, the participants were instructed to provide voice instructions for the driving assistance control system as well as numbers ranging from 1 to 40. The spontaneous narrative speech was captured in addition to the scripted speech. Each participant was instructed to describe a picture or tell a story about their recent weekend, best vacation, etc. The number of tasks are not similar for all subjects. Recording session for each subjects is varying from 15 min to 2 h.

Table 1 The number of speech samples per class in the train, develop and, test partition of the URTIC database

Partition	Cold	Healthy	Total
Train	970	8535	9505
Develop	1011	8585	9596
Test	895	8656	9551

The recordings were subdivided into 28 652 samples ranging in length from 3 to 10 s. The total duration of the database is about 45 h.The number of segments per subject is not the same. The 28 652 samples were divided into three speaker-independent partitions (train, develop, and test) that were balanced by gender, age, and experimenter. Table 1 shows the details of the URTIC database. Each partition (train, develop, and test) contains recordings of 210 speakers, 37 of whom had a common cold and 173 healthy. The train partition has 9505 samples, the develop partition has 9596 samples, and the test partition has 9551 samples. There are a total of 25,776 samples of healthy speech and 2876 samples of cold speech.

3 Related work

A number of studies have been conducted on the impact of cold speech on the performance of the speaker recognition system, as well as the classification of cold speech and healthy speech. Tull and Rutledge [14] noticed that the formant frequency was significantly lower in cold speech compared healthy speech. Tull et al. [15] examined the common cold's effect on mel frequency cepstral coefficients (MFCC). Ai et al. [10] proposed a dual model updating strategy for speaker recognition in cold speech for home assistants. This method combined time domain and frequency domain features to classify cold and healthy speech. Then, corresponding Gaussian mixture model (GMM) was selected from two separate models, one trained using cold speech and another using healthy speech for speaker recognition. The INTERSPEECH 2017 Computational Paralinguistics Challenge ComParE-2017 addressed the cold and healthy speech classification task using the URTIC database [13]. The baseline of the ComParE-2017 challenge utilized an end-to-end learning strategy using a convolutional neural network (CNN) and long short-term memory (LSTM), a 6373-dimensional ComParE-2013 feature set, and a 130-dimensional bag-of-audio-words (BoAW) features. Cai et al. [16] used perception-aware MFCC, and constant Q cepstral coefficients (CQCC) features to detect the common cold. Suresh et al. [17] used phoneme state posteriorgram (PSP) features for classifying cold and healthy speech. Wagner

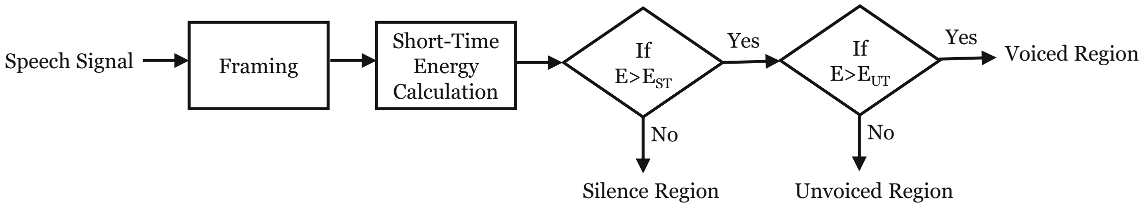

Fig. 1 Flowchart for the segmentation of the voiced and unvoiced region

et al. [18] investigated the effect of common cold at phonetic level. Author derived a phonetic transcription from an automatic speech recognizer (ASR), and based on phonetic transcription, the author trained a classification model for each phonetic class using different low-level features such as MFCC, invariant-integration features (IIF), and constrained maximum likelihood linear regression (CMLLR). Deb et al. [4] used variational mode decomposition approach to extract various statistics, entropy, and energy features from a voice signal. For classification, a mutual information-based weight assignment technique and the SVM classifier were used. Kao et al. [19] utilized discriminative autoencoders and MFCC features for cold and healthy speech classification. Vicente et al. [20] developed a Fisher vector for identifying cold speech using MFCC features and a generative Gaussian mixture model. In our previous study [21], we have utilized only vowel-like region segments of speech for cold and healthy speech classification. Vowel-like regions (VLR) were separated from speech signals by identifying vowel-like region onset points and endpoints using the Hilbert envelope of linear prediction residual signal and zero-frequency filtering methods. Then, MFCC features are extracted from vowel-like regions of speech signals. When only a vowel-like region is considered, the number of frames reduced dramatically during feature extraction. Deb et al. [22] employed MFCC and linear predictive coding (LPC) features and a deep neural network classifier to classify cold and healthy speech.

4 Segmentation of voiced and unvoiced region

Speech comprises various phonemes, which are produced by the vocal cords and the vocal tract. The state of the vocal cord, as well as the positions, shape, and size of various articulators, determine which phonemes are produced [12]. There are several methods to categorize events in a speech. The most straightforward method is to use the state of the source of speech production (vocal cord). In this method, the speech signal is categorized into voiced, unvoiced, and silence regions. The voiced sounds are produced when the vocal cord vibrates, and unvoiced sounds are produced when the vocal cord does not vibrate. During the silence, speech is not produced. Voiced and unvoiced speech may be distinguished because voiced speech waveform is quasi-periodic, and unvoiced speech waveform is aperiodic or random in nature [12].

Segmentation of voiced and unvoiced speech is a fundamental and important process for various speech processing applications. Cai [23] proposed a method based on the wavelet-based frequency distribution of the average energy, zero-crossing rate (ZCR), and short-time energy (STE) of the speech signal. Atal and Rabiner [24] utilized ZCR, STE, the correlation between adjacent speech samples, LPC analysis, and LPC error for the segmentation of voiced and unvoiced speech. Ijitona et al. [25] proposed a method based on the combination of linear prediction error variance (LPEV), STE, and ZCR for the segmentation of voiced, unvoiced, and silence regions. In this study, we employed STE to categorize each speech frame into voiced and unvoiced frames. If the STE of a speech frame is greater than the threshold energy, it is considered a voiced frame; otherwise, it is considered an unvoiced frame. Figure 1 shows the flowchart for segmentation of voiced and unvoiced region of speech signal. First, each Speech signal is segmented into small frames of 20 ms duration and 10 ms overlapping. Then, STE is calculated for each speech frame. The STE of kth speech frame is given by

$$E_k = \sum_{n=1}^{N} s_k^2(n) \tag{1}$$

where N is the total number of samples in each speech frame. The average energy for all speech frames in an utterance is calculated as

$$E_{\text{avg}} = \frac{1}{K} \sum_{k=1}^{K} E_k \tag{2}$$

where K represents the total number of frames in an utterance. Then, global thresholds are decided for silence region (E_{ST}) and unvoiced region (E_{UT}). The thresholds (E_{ST} & E_{UT} are calculated as

$$E_{\text{ST}} = \alpha * E_{\text{avg}} \tag{3}$$

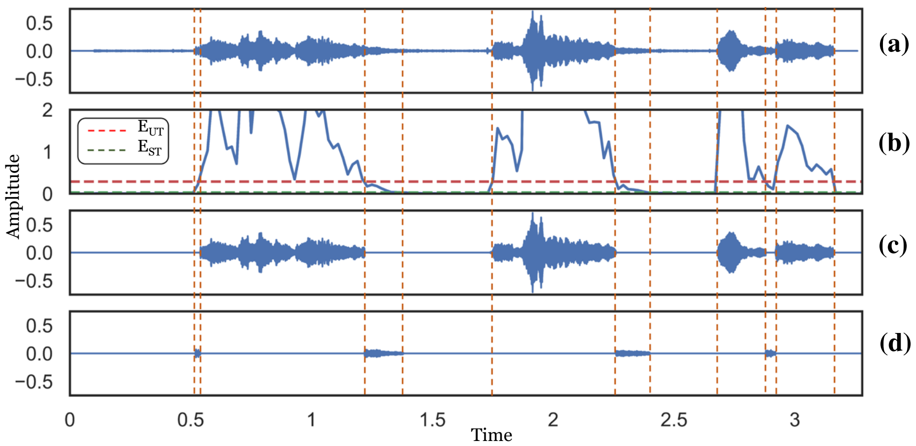

Fig. 2 Segmentation of voiced and unvoiced region of speech using STE. **a** Speech signal *devel*_0015.*wav* form the URTIC database. **b** STE of speech signal. **c** Detected voiced region. **d** Detected unvoiced region

$$E_{\text{UT}} = \beta * E_{\text{avg}} \tag{4}$$

where α and β are the constants. To select the values of α and β, we have analyzed the segmentation of 50 (25 cold and 25 healthy) randomly chosen utterances from the URTIC database. The analysis shows that $\alpha = 0.02$ and $\beta = 0.12$ give more accurate segmentation of voiced and unvoiced regions. Hence, we have selected $E_{\text{ST}} = 0.02 * E_{\text{avg}}$ and $E_{\text{UT}} = 0.12 * E_{\text{avg}}$ for the segmentation of voiced and unvoiced regions.

The speech frame having STE less than E_{ST} is considered as a silence frame and discarded. The resultant speech segments after removing silence is known as complete active speech (CAS). It comprises both voiced and unvoiced speech regions.

The speech frame having STE greater than E_{UT} is considered as a frame from the voiced region and the speech frame having an energy between E_{ST} and E_{UT} is considered as a frame from the unvoiced region of speech. We have also tried to detect voiced and unvoiced regions using zero-crossing rate (ZCR) and zero-frequency filtered signal (ZFFS) approaches [26]. The highest segmentation performance is achieved using STE-based detection of voiced and unvoiced regions.

Figure 2 shows the segmentation of voiced and unvoiced regions for speech sample *devel_0015.wav* from the URTIC database. Figure 2a shows the input speech signal. Figure 2b shows the STE for the speech signal with threshold levels E_{ST} and E_{UT} for voiced and unvoiced region segmentation. Figure 2c shows the voiced regions, and Fig. 2d shows the unvoiced regions of speech signal.

5 Classification method

In the previous section, we have segmented the speech signal into voiced and unvoiced regions using STE-based method. This section will analyze the significance of the voiced and unvoiced regions for classifying cold and healthy speech. Figure 3 shows the block diagram of proposed system for classification of cold and healthy speech using voiced and unvoiced speech regions. First, the speech signal is segmented into voiced and unvoiced regions, and MFCC features are extracted from them. Also, MFCC features are extracted from the CAS before segmenting it into voiced and unvoiced regions.

5.1 Feature extraction

This study employs MFCC features to classify cold and healthy speech. The MFCC features have been employed for detection of various pathological conditions [1,16,27,28].

In this study, 13 MFCC coefficients, 13 deltas, and 13 delta-delta coefficients are extracted from each overlapping speech frame. Then, these frame-level MFCC features extracted from voiced, unvoiced, or CAS regions of each speech utterance are converted into utterance-level features by calculating four statistics (mean, variance, skewness, and kurtosis) to form 156-dimensional feature vector per utterance. These 156-dimensional utterance-level feature vectors are utilized for training and testing the support vector machine (SVM).

5.2 Support vector machine (SVM)

The SVM is commonly utilized in speech pathology and emotion detection [29–34]. The SVM is a mathematical technique for maximizing a certain mathematical function with respect to a given set of data [35]. It makes use of kernel functions to map the original feature space into a space with a higher dimension so that it can be separated linearly. SVM employs convex optimization, which is useful for achieving a globally optimum solution. In this investigation, we have used SVM with radial basis function (RBF) kernel for binary (cold and healthy speech) classification.

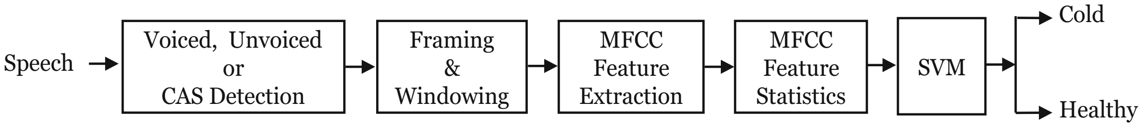

Fig. 3 Block diagram of proposed voiced and unvoiced speech-based classification system

Table 2 Performance of MFCC feature statistics extracted from voiced, unvoiced and CAS for classifying cold and healthy speech

Speech segment	% UAR	
	Develop	Test
CAS	66.12	64.92
Voiced	65.98	64.85
Unvoiced	66.15	64.69

Table 3 Performance comparison of proposed framework with the classification results reported using phoneme segmentation on the URTIC database

Speech segment	Feature	% UAR
Vowel [18]	IIF	60.8
Glide [18]	IIF	60.3
Consonant [18]	IIF	62.7
Liquid [18]	IIF	62.3
Stops [18]	IIF	62.2
Fricative [18]	IIF	62.2
Nasal [18]	IIF	59.6
Vowel-like Region [21]	MFCC	61.93
Voiced Region	**MFCC**	**65.98**
Unvoiced Region	**MFCC**	**66.15**

The bold values indicate the results achieved in our study

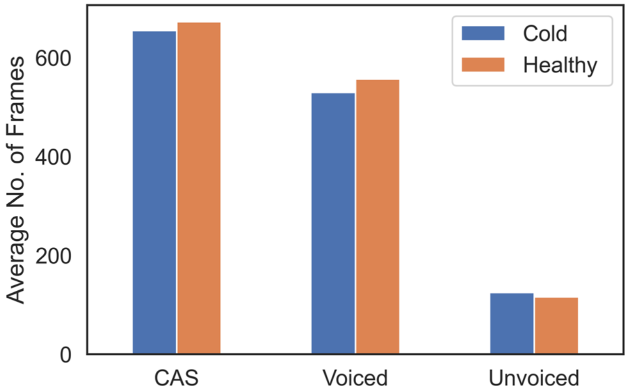

Fig. 4 Average no. of frames per utterance utilized for features extraction to classify cold and healthy speech using voiced, unvoiced and CAS segments

6 Results and discussion

The URTIC database is used in this study to analyze the significance of voiced and unvoiced regions to classify cold and healthy speech. Table 2 shows the performance of MFCC feature statistics extracted from voiced, unvoiced, and CAS segments of speech signal for cold and healthy speech classification. The MFCC feature statistics extracted from the voiced region give the UAR of 65.98% on the develop partition and 64.85% on the test partition of the URTIC database. The MFCC feature statistics extracted from the unvoiced region achieve the UAR of 66.15% on the develop partition and 64.69% UAR on the test partition of the URTIC database. Conversely, the MFCC feature statistics extracted from the CAS achieve UAR of 66.12% and 64.92%, respectively, on the develop and test partitions of the URTIC database.

The results show that performance achieved using the statistics of MFCC features gives almost similar results for voiced, unvoiced, and CAS on both test and develop partition of the URTIC database using SVM. For the develop partition, the highest classification performance of 66.15% UAR is achieved using the MFCC features extracted from the unvoiced region of speech, while for the test set, the highest classification performance of 64.92% UAR is achieved using the MFCC features extracted from the CAS using SVM.

Limited studies had utilized various speech segments or phonemes for the detection of the common cold from the speech signal. Table 3 shows the performance comparison of the proposed framework with the state-of-the-art methods using various phoneme segmentation on the URTIC database. Wagner et al. [18] analyze the effect of common cold speech on a phonetic level. Phonemes are grouped into vowel, glide, consonant, liquid, stops, fricative, and nasal sounds, and the classification performance is evaluated using MFCC, invariant-integration features (IIF), and constrained maximum likelihood linear regression (CMLLR) feature. They achieved the highest score using IIF features, as given in Table 3. They concluded that consonant articulation is more impaired than vowel articulation. They achieved the highest UAR of 62.70% for the consonants group of phonemes. In our previous study [21], we achieved 61.93% UAR using the MFCC features extracted from the vowel-like region segments of speech for cold and healthy speech classification.

The performance of the proposed framework is compared with the classification results reported in state-of-the-art methods on the URTIC database, as shown in Table 4. The URTIC dataset was utilized in the INTERSPEECH 2017 ComParE Challenge's Cold Sub-Challenge, and the base-

Table 4 Performance comparison of the proposed framework with state-of-the-art methods

Research work	Speech region	%Reduction in frames	Feature	Feature dimension	% UAR	
					Develop	Test
Schuller et al. [13]	CAS	–	ComParE	6373	64.00	70.20
			BoAW	130	64.20	67.30
Cai et al. [16]	CAS	–	MFCC	39	64.80	–
			CQCC		65.40	–
Suresh et al. [17]	CAS	–	PSP	5160	64.00	61.09
Deb et al. [4]	CAS	–	VMD	50	66.84	–
Kao et al. [19]	CAS	–	MFCC	60	65.81	66.00
Vicente et al. [20]	CAS	–	MFCC	39	63.98	66.12
Warule et al. [21]	VLR	50.76	MFCC	39	61.93	–
Deb et al. [22]	CAS	–	MFCC	39	65.64	–
			LPC	32	59.78	–
Proposed	CAS	–	MFCC Statistics	156	66.12	64.92
	Voiced	17.41			65.98	64.85
	Unvoiced	82.59			66.15	64.69

line results are presented in [13]. The UAR achieved using the proposed methods is greater than the baseline results for the develop partition of the URTIC database. The proposed framework is not as good as the baseline results on the test partition. Here, we wish to emphasize that the baseline findings were obtained using a 6373-dimensional ComParE-2013 feature set. Conversely, we have used only 156-dimensional features in the proposed framework to achieve comparable results with the baseline. The proposed framework gives higher results than state-of-the-art methods on the develop partition and is much in line with the results of the test partition of the URTIC database. In this study, the results achieved only using the unvoiced region of speech give higher UAR compared to the results of state-of-the-art methods on the develop partition of the URTIC database. Compared to the CAS segments, the number of frames utilized for feature extraction from the voiced and unvoiced regions is reduced by 17.41% and 82.59%, respectively. The results achieved only using the unvoiced region of speech give higher UAR and 82.59% reduction in the total number of frames. Figure 4 shows the average number of frames that need to be processed during feature extraction of training partition to classify cold and healthy speech using the voiced, unvoiced, and CAS speech segments.

The unvoiced region of speech signal has very few frames compared to the CAS and voiced regions of speech. We have used a simple STE-based approach for the segmentation of voiced, unvoiced, and silence regions of speech. In state-of-the-art methods, all the frames of CAS are processed for feature extraction. In our previous study [21], we used a VLR of speech which reduces the number of frames by 50.76% for the classification of cold and healthy speech. But UAR achieved using MFCC features extracted from VLR is low compared to the state-of-the-art methods. Wagner et al. [18] examined the effects of common cold on phonetic level. They concluded that phoneme-level cold and healthy speech classification was not worthwhile. Compared to the segmentation of various phonemes and vowel-like regions, segmentation of voiced and unvoiced regions is simple and quick. Also, the performance achieved for classifying cold and healthy speech using the voiced and unvoiced segments is high compared to the performance achieved using vowel-like regions segments and a group of phonemes. In this study, the results achieved only using the unvoiced region of speech give higher UAR and 82.59% reduction in the total number of frames compared to the results of state-of-the-art methods on the develop partition of the URTIC database. So, if someone is merely interested in common cold detection with the minimum effort and complexity, they can evaluate only the unvoiced portion of the speech. This system can be implemented in smart devices with limited memory to detect and monitor common cold and related disorders, which may be useful for remotely monitoring the patient's health and preventing the spread of these illnesses.

7 Conclusion

In this work, we have investigated the significance of the voiced and unvoiced speech segments for detecting the common cold. A short-time energy-based approach is used for segmenting speech signals into the voiced and unvoiced regions. After segmentation, 39-dimensional MFCC features are extracted from the voiced and unvoiced segments, and statistics (mean, variance, skewness, and kurtosis) are calculated to get a 156-dimensional feature vector for each speech utterance from the all frame-wise MFCC features of that utterance. The SVM classifier is used to analyze the performance of MFCC features extracted from the voiced and unvoiced speech region. The results show that the feature extracted from voiced and unvoiced segments shows the same discrimination capability for cold and healthy speech. The processing of only unvoiced segments to detect the common cold can also serve the purpose. The processing of the only unvoiced frame for common cold detection reduced the number of frames by 82.59% without significant change in the system performance compared to CAS. This study concludes that unvoiced speech segments include pathological information and can be utilized to diagnose common cold and associated disorders.

Author contributions PW conducted research and wrote the paper. SPM and SD participated in the writing and preparation of the paper. All authors reviewed the results and approved the final version of the manuscript.

Funding Not applicable.

Availability of data and materials The data that support the findings of this study are not openly available. Data may be available (DOI:https://doi.org/10.21437/Interspeech.2017-43) upon reasonable request.

Declarations

Ethical approval Not applicable.

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships.

References

1. Cummins, N., Baird, A., Schuller, B.W.: Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. Methods **151**, 41–54 (2018)
2. Shilandari, A., Marvi, H., Khosravi, H., Wang, W.: Speech emotion recognition using data augmentation method by cycle-generative adversarial networks. Signal Image Video Process. **2022**, 1–8 (2022)
3. Sun, L., Huang, Y., Li, Q., Li, P.: Multi-classification speech emotion recognition based on two-stage bottleneck features selection and mcjd algorithm. Signal Image Video Process. **2022**, 1–9 (2022)
4. Deb, S., Dandapat, S., Krajewski, J.: Analysis and classification of cold speech using variational mode decomposition. IEEE Trans. Affect. Comput. **11**(2), 296–307 (2017)
5. World Health Organization (2022). https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)
6. Cowie, M.R., Lam, C.S.: Remote monitoring and digital health tools in cvd management. Nat. Rev. Cardiol. **18**(7), 457–458 (2021)
7. Jnr, B.A.: Use of telemedicine and virtual care for remote treatment in response to Covid-19 pandemic. J. Med. Syst. **44**(7), 1–9 (2020)
8. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. **44**(3), 572–587 (2011)
9. Calvo, R.A., D'Mello, S.: Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans Affect. Comput. **1**(1), 18–37 (2010)
10. Ai, H., Wang, Y., Yang, Y., Zhang, Q.: An improvement of the degradation of speaker recognition in continuous cold speech for home assistant. In: International Symposium on Cyberspace Safety and Security. Springer, pp. 363–373 (2019)
11. Tyrrell, D., Cohen, S., Schilarb, J.: Signs and symptoms in common colds. Epidemiol. Infect. **111**(1), 143–156 (1993)
12. Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice-Hall Inc, Upper Saddle River (1993)
13. Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M. et al.: The interspeech 2017 computational paralinguistics challenge: addressee, cold & snoring. In: Computational Paralinguistics Challenge (ComParE), Interspeech 2017, pp. 3442–3446 (2017)
14. Tull, R.G., Rutledge, J.C.: Analysis of "cold-affected" speech for inclusion in speaker recognition systems. J. Acoust. Soc. Am. **99**(4), 2549–2574 (1996)
15. Tull, R.G., Rutledge, J.C., Larson, C.R.: Cepstral analysis of "cold-speech"for speaker recognition: a second look, Ph.D. thesis, Acoustical Society of America (1996)
16. Cai, D., Ni, Z., Liu, W., Cai, W., Li, G., Li, M., Cai, D., Ni, Z., Liu, W., Cai, W.: End-to-end deep learning framework for speech paralinguistics detection based on perception aware spectrum. In: INTERSPEECH, 2017, pp. 3452–3456 (2017)
17. Suresh, A.K., KM, S.R., Ghosh, P.K.: Phoneme state posteriorgram features for speech based automatic classification of speakers in cold and healthy condition. In: INTERSPEECH, 2017, pp. 3462–3466 (2017)
18. Wagner, J., Fraga-Silva, T., Josse, Y., Schiller, D., Seiderer, A., André, E.: Infected phonemes: how a cold impairs speech on a phonetic level (2017)
19. Kao, Y.-Y., Hsu, H.-P., Liao, C.-F., Tsao, Y., Yang, H.-C., Li, J.-L., Lee, C.-C., Lee, H.-S., Wang, H.-M.: Automatic detection of speech under cold using discriminative autoencoders and strength modeling with multiple sub-dictionary generation. In: 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE vol. 2018, pp. 416–420 (2018)
20. José Vicente, E.L., Gosztolya, G.: Using the fisher vector approach for cold identification. Acta Cybern. **25**(2), 223–232 (2021)
21. Warule, P., Mishra, S.P., Deb, S.: Classification of cold and non-cold speech using vowel-like region segments. In: 2022 IEEE International Conference on Signal Processing and Communications (SPCOM). IEEE, pp. 1–5 (2022)
22. Deb, S., Warule, P., Nair, A., Sultan, H., Dash, R., Krajewski, J.: Detection of common cold from speech signals using deep neural network. Circuits Syst. Signal Process. **2022**, 1–16 (2022)
23. Cai, R.: A modified multi-feature voiced/unvoiced speech classification method. In: 2010 Asia-Pacific Conference on Power Electronics and Design. IEEE, pp. 68–71 (2010)
24. Atal, B., Rabiner, L.: A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. IEEE Trans. Acoust. Speech Signal Process. **24**(3), 201–212 (1976)
25. Ijitona, T., Yue, H., Soraghan, J., Lowit, A.: Improved silence-unvoiced-voiced (suv) segmentation for dysarthric speech signals using linear prediction error variance. In: 2020 5th International Conference on Computer and Communication Systems (ICCCS). IEEE, 2020, pp. 685–690 (2020)
26. Ramteke, P.B., Koolagudi, S.G.: Phoneme boundary detection from speech: a rule based approach. Speech Commun. **107**, 1–17 (2019)
27. Islam, R., Tarique, M., Abdel-Raheem, E.: A survey on signal processing based pathological voice detection techniques. IEEE Access **8**, 66749–66776 (2020)

28. Muguli, A., Pinto, L., Sharma, N., Krishnan, P., Ghosh, P.K., Kumar, R., Bhat, S., Chetupalli, S.R., Ganapathy, S., Ramoji, S. et al.: Dicova challenge: Dataset, task, and baseline system for Covid-19 diagnosis using acoustics, arXiv preprint arXiv:2103.09148 (2021)
29. Jain, M., Narayan, S., Balaji, P., Bhowmick, A., Muthu, R.K. et al., Speech emotion recognition using support vector machine, arXiv preprint arXiv:2002.07590 (2020)
30. Schuller, B., Rigoll, G., Lang, M.: Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1. IEEE, pp. I–577 (2004)
31. Deb, S., Dandapat, S.: Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification. IEEE Trans. Cybern. **49**(3), 802–815 (2018)
32. Shahbakhi, M., Far, D.T., Tahami, E.: Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine. J. Biomed. Sci. Eng. **07**(04), 147–156 (2014)
33. Pishgar, M., Karim, F., Majumdar, S., Darabi, H.: Pathological voice classification using mel-cepstrum vectors and support vector machine, arXiv preprint arXiv:1812.07729 (2018)
34. Gil, D., Manuel, D.J.: Diagnosing Parkinson by using artificial neural networks and support vector machines. Glob. J. Comput. Sci. Technol. **9**(4), 2009 (2009)
35. Noble, W.S.: What is a support vector machine? Nat. Biotechnol. **24**(12), 1565–1567 (2006)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.