Guo *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2023) 2023:16
https://doi.org/10.1186/s13636-023-00282-x

EURASIP Journal on Audio,
Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

Voice activity detection in the presence of transient based on graph

Xiao-Yuan Guo[1], Chun-Xian Gao[1] and Hui Liu[1*]

Abstract

Voice activity detection remains a significant challenge in the presence of transients since transients are more dominant than speech, though it has achieved satisfactory performance in quasi-stationary noisy environments. This paper studies the differences between speech and transients in nonlinear dynamic characteristics and proposes a new method for accurately detecting speech and transients. Limited by algorithm complexity, previous research has proposed few detectors to model speech and transients based on contextual information and thus failing to detect transient frames accurately. To address this challenge, our study proposes to map features of audio signals to a time series complex network, a kind of graph data, analyzed by the Laplacian and adjacency matrix of graphs, then classified by the support vector machine (SVM) classifier. The proposed algorithm can analyze a more extended speech period, allowing the full utilization of contextual information of preceding and following frames. The experimental results show that the performance of this method has obvious superiority over other existing algorithms.

Keywords Voice activity detection, Transients, Time series complex networks, Nonlinear dynamic characteristics

1 Introduction

Voice activity detection divides a given audio signal into one part containing speech and the other containing only noise or interference, an essential component in many speech systems, such as speech recognition [1], speaker identification [2], and speech separation. Usually, voice activity detection is the initial block of voice systems.

Common methods of voice activity detection for noisy signals, as described in [3–6], track the statistics of signals with the recursive average in short time intervals. However, these methods assume that the noise spectrum varies slowly with time. Thus they fail to model non-stationary noise accurately, leading to the incorrect identification of noise as speech.

Transients are non-stationary noise, such as door knocking, keyboard tapping, and hammering. Usually, transients have a short duration and a broadband frequency spectrum, and their frequency spectrum varies even more rapidly with time than speech. Since transients have high energy and appear more dominant than speech, frames containing both speech and transient are typically misclassified as transient frames, which largely explains the degraded performance of traditional methods.

Several studies have also verified the favorable property of deep networks to model the inherent structure contained in the speech signal [7]. Zhang and Wu [8] proposed to extract acoustic features and feed them to a deep-belief network to obtain a more meaningful representation of the signal. Thomas et al. [9] proposed VAD based on convolutional neural networks (CNN). Modern methods rely on recurrent neural networks (RNN) to model temporal relations between consecutive time frames [10]. Ivry et al. [11] proposed to implicitly model the data without using an explicit noisy signal model. However, machine learning-based VAD methods require frame-level data for training, and some of them often misclassify frames that contain both speech and

*Correspondence:
Hui Liu
lh@xmu.edu.cn
[1] Department of Information and Communication Engineering, Xiamen University, Xiamen, China

transients as non-speech frames due to the dominance of transients over speech.

To successfully separate speech from audio signals with transients, many features have been proposed in the literature to facilitate the differentiation of the human voice and other voices, such as energy, pitch [12], formant structure [13], autocorrelation function [14], and spectral features [15, 16]. However, certain drawbacks are associated with using these algorithms formulated on different features. For instance, 4Hz modulation energy can only be used to distinguish between music and speech [17]. Similarly, being the main feature of vowels, the pitch is not an effective parameter to identify consonants, which results in the imprecise detection of speech frames.

Yoo et al. [13] proposed a voice activity detection algorithm developed on formants. Human speech is classified into vowels or consonants. Vowels have spectral peaks, also known as formants, which can be useful indicators to confirm the presence of vowels. Most formants may survive in the signal even when the signal is polluted by noise. It is assumed that consonants are accompanied by adjacent vowel sounds [18], so the preceding and following frames of detected speech frames are regarded as speech. The algorithm requires a small amount of computation and demonstrates its robustness against various environmental noises, such as the one produced by railway stations or airports. However, transients will destroy the speech formants and cannot be distinguished from speech.

To overcome the limitations of traditional methods, an assumption has been made in [19–21] that transients contain an underlying geometric structure, which can be inferred from the signal with manifold learning tools. The method presented in [21] assumes that transients and speech are generated by independent latent variables respectively and solves the problem of non-linear independent component analysis by using the modified Mahalanobis distance. It has been proved that the Mahalanobis distance between frames approximates the Euclidean distance between the generating variables, representing the similarity between frames. The similarity metric is then incorporated into the kernel-based manifold learning method to construct a low-dimensional signal representation. The detection only includes 15 preceding and 15 following frames of the current frame to determine its content. Using highly overlapping frames would lead to a considerable increase in the computational cost of the algorithm.

The context of audio signals is relevant, so contextual information can be factored in to improve the accuracy of voice activity detection. Graph signal processing (GSP) provides a new approach to processing data, which can illustrate the relationships between data in the form of nodes, edges, and weights of edges [22]. Graph signals are highly adaptable to different kinds of signals and can be reasonably designed according to the characteristics of the signal. In GSP, the weight matrices of edges representing the geometry of the graph consist of the graph Laplacian matrix [23] and graph adjacency matrix [24]. Xue et al. [25] had the features of audio signals mapped to the graph domain, and defined the graph Fourier basis vector through singular value decomposition of graph adjacency matrices to denoise speech.

Considering that speech signal is nonlinear, we map signal features to graph signals and propose a GSP-based method to distinguish between speech and transients. Audio signals are a time series. Traditionally, time series analysis often relied on statistical analysis methods. Recently, the complex network analysis theory has eliminated the dependence on the accuracy of mathematical models, a characteristic of traditional methods. The theory maps features to a time series complex network [26], one kind of graph. Then the Laplacian and adjacency matrix of graphs are extracted to present the characteristics of the nodes and edges of graphs. More continuous time frames are considered in this method, which effectively weakens the transient effect and reduces the impact of transients on speech frames. We demonstrate that time series complex networks of transients and speech display significant differences in both nodes and edges. The features extracted from graphs can effectively distinguish between speech and transients.

The remainder of the paper is organized as follows. In Section 2, we formulate the problem of voice activity detection and propose a method that distinguishes between speech and transients by constructing time series complex networks and uses the graph-based method to measure speech and transients. The extraction and classification of graph features are also discussed. Section 3 introduces a voice activity detection algorithm based on graphs and provides experimental results to demonstrate that the algorithm outperforms other competing methods. Section 4 concludes and discusses the proposed method.

2 Problem formulation

2.1 Voice activity detection feature

Voice activity detection aims to automatically detect signal frames containing speech from a continuous observation signal. This paper examines speech with transients collected by a single microphone, divided into frames to extract acoustic features to distinguish speech and transient frames. Let $x(t)$ the noisy signal, given by:

$$x(t) = s(t) + n(t) \tag{1}$$

where $s(t)$ and $n(t)$ are the pure speech signal and the transient signal respectively, our task is to divide the

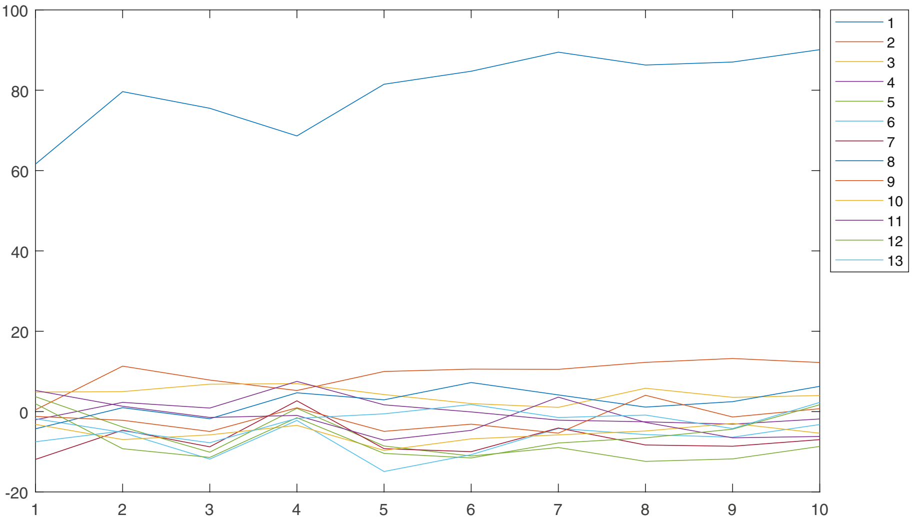


Fig. 1 13 MFCC of some audio frames

frames of the observed signal into speech and non-speech clusters.

Mel-frequency cepstral coefficients (MFCC) depend on the acoustic characteristics of human ears. The human auditory system perceives sounds of different frequencies with different sensitivities, corresponding nonlinearly to Hz [2].

In voice activity detection, transients are usually more dominant than speech due to their higher amplitude and broader bandwidth. Accordingly, the MFCC of frames containing speech and transients is often similar to those containing only transients. To extract nonlinear dynamic features of speech, we map MFCC to graph signals and identify the type of frames according to the feature of graphs. The input signal $x(t)$ is divided into N frames, and each frame is denoted as $\boldsymbol{x}_n$. $MFCC(\cdot)$ denotes the extraction process of MFCC of the input signal, and its output is $y_n \in \mathbb{R}^{1\times M}$ given by:

$$\boldsymbol{y}_n = \mathrm{MFCC}(\boldsymbol{x}_\mathrm{n}) \tag{2}$$

2.2 Dimension reduction

The number of triangular filters decides the number of MFCC's dimensions, generally 13, 26, or 39. High dimensionality poses tremendous challenges to mapping MFCC to graph signals. Mapping a high-dimensional time series to a time series complex network will produce a complex network with a complicated structure.

Moreover, analyzing more frames will moderately increase computation complexity, which increases the difficulty of feature extraction and thus generates the necessity for a low-dimensional representation of audio signals. Dimensionality reduction based on the correlation matrix, such as principal component analysis (PCA), factor analysis (FA), and canonical correlation analysis (CCA), is linear mapping to study the linear correlation between variables. Such methods are particularly beneficial for processing linear data but are not conducive to retaining the inherent characteristics of original data [27, 28]. Cosine similarity [29] and Pearson correlation coefficient are commonly used distance and similarity metric, but they are less discriminative for these coefficients.

It can be inferred from Fig. 1 that the variation trends of some coefficients are consistent. However, it is not easy to calculate the distance between each coefficient of each frame. In order to capture the low-dimensional geometric structure and retain nonlinear dynamical features of speech, we assume that some coefficients in MFCC have similar probability distributions and variation trends. Therefore, the similarity of the variation trends can be measured by the difference between the probability distribution of a large amount of data.

Kullback-Leibler (KL) divergence and Jensen-Shannon (JS) divergence are currently the most popular methods for measuring differences in probability distributions. However, these metrics are meaningless when two distributions do not overlap. The advantage of the Wasserstein distance [30] is that it provides a meaningful gradient even if two distributions do not overlap or barely overlap. It is being used increasingly in Statistics and Machine Learning, which arises from the idea of optimal transport. The method normalizes the histograms and computes the minimum transportation distance to measure the similarity between two histograms. It is defined by:

$$W\left(P_i,P_j\right) = \inf_{\gamma\sim\Pi\left(P_i,P_j\right)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|] \tag{3}$$

where $\Pi\left(P_i,P_j\right)$ represents all joint distributions γ for (X,Y) that have marginals distributions P_i and P_j. $\|x-y\|$ is the distance between two samples in the joint distribution γ. The expected value of the distance between samples in the γ distribution can be calculated, and the maximum lower bound of all expected values is the Wasserstein distance. The greater the Wasserstein distance between two probability distributions is, the greater their difference is. The function $WS(\cdot)$ is to calculate the distance between the probability distributions of every two coefficients, given by:

$$W = WS\begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ y(M) \end{pmatrix} \tag{4}$$

Let $y(m)$ denote the mth MFCC of all speech frames, and M denotes the number of MFCC's dimensions. Element $W_{i,j}$ in $W \in \mathbb{R}^{\mathrm{M}\times\mathrm{M}}$ indicates the distance between the probability distributions of the ith and jth coefficient.

$$\text{index } = \mathrm{Max}(\mathrm{W}) \tag{5}$$

The output of $\mathrm{Max}(\cdot)$ is the index value of the first K largest sum of elements in each column of the input matrix, and the corresponding coefficients are selected as the feature $\mathbf{z}_n$, given by:

$$\boldsymbol{z}_n = \boldsymbol{y}_{n,\,\text{index}} \tag{6}$$

2.3 Complex network construction

After dimension reduction, we map the feature to graph signals following the complex network theory. In the classical time series processing method, the parameter calculus of phase space reconstruction hinders the efficient construction of complex networks. The visibility graph [26] starts from the original data characteristics of the time series and examines the internal connections between data, significantly improving network construction efficiency. Inspired by the above ideas, we focus on exploring the internal connections of time series and use the pattern representation for time series to determine the nodes and edges of complex networks.

In this paper, the pattern representation of time series is realized by equal probability symbolization before the main patterns in time series are extracted as the nodes of networks using equivalent transformation. To establish a weighted directed complex network, the edge direction and weight between nodes are determined by the conversion relationship and frequency between different patterns.

We assume that there is a set of N observable frames and a K-dimensional time series $\mathbf{z}_n$. Each dimension time series is divided into P intervals with equal probability, and integers $\{1,2,\cdots,P\}$ are taken to represent different intervals. The time series is transformed into the symbolic series by:

$$c_{n,k} = \mathrm{symbol}\left(\mathrm{z_{n,k}}\right) \tag{7}$$

where $z_{n,k}(k=1,\cdots,K)$ denotes the kth coefficient of the speech feature $\mathbf{z}_n$, and $c_{n,k} \in \{1,2,\cdots,P\}$ denotes the symbolized value. Let the $symbol(\cdot)$ represent the process of equal probability symbolization, which maps time series to symbolic series.

The symbol c_n denotes the pattern of each frame which is regarded as the node of the graph signal, given by:

$$c_n = c_{n,1}c_{n,2}\cdots c_{n,K} \tag{8}$$

When modeling each frame of signals, $2J+1$ continuous time frames will be considered, and $\underline{\boldsymbol{z}}_n$ is defined by:

$$\underline{\boldsymbol{z}}_n = \left[\mathbf{z}_{n-J},\ldots,\mathbf{z}_n,\ldots,\mathbf{z}_{n+J}\right] \tag{9}$$

Moreover, the symbolic series is given by:

$$\underline{c}_n = \left[c_{n-J},\cdots,c_n,\cdots,c_{n+J}\right] \tag{10}$$

where J is the number of past and future time frames, the multidimensional time series is transformed into pattern series represented by symbols.

Different patterns are regarded as the nodes of time series complex networks, and the conversion direction and frequency between them determine their connecting directions and weights. If the following pattern is the same as the current one, the node remains unchanged; otherwise, an edge exists connecting the two nodes. The direction of the edge is set to be from the current node to the next one, and the weight of the connecting edge increases by one. Following the rule above, we can transform the symbolic series into a weighted, directed, complex network.

2.4 Graph signal and feature

A complex network is a typical graph signal and unstructured data representation. Classical digital signal processing addresses temporal signals, implying a highly regular data structure; each sample point is arranged chronologically. However, attempts are generally required to process non-regular structured data, such as traffic and social networks, including numerical information and structural correlations between data. Graph signal processing is dedicated to processing signals on the graph structure, which are not limited to those with regular structures, as addressed in classical digital signal processing.

We define a weighted directed graph signal as $O_{\mathcal{G}}$, given by:

$$O_{\mathcal{G}} = \Gamma\left(\underline{c}_n\right) \tag{11}$$

where $\Gamma(\cdot)$ denotes the mapping of symbolic series $\underline{c}_n \in \mathbb{R}^{K\times(2J+1)}$ to graph signal. Let v_k be the node of the graph, and we assume that

$$O_{\mathcal{G}} = \left[O_{v_1}, O_{v_2}, \ldots, O_{v_M}\right]^T \in \mathbb{R}^M, M \leq 2J+1 \tag{12}$$

$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ denotes the topology of the graph signal $O_{\mathcal{G}}$, and $\mathcal{V} = [v_1, v_2, \ldots, v_M]^{\mathrm{T}}$ denotes the set of nodes on the graph corresponding to $O_{\mathcal{G}}$, which means that the value of node v_k is O_{v_k}. $\mathcal{E} = \left\{e_{i,j}\right\} \in \mathbb{R}^{M\times M}$ denotes the edge matrix of graph signals, in which $e_{i,j} \in \{0, 1\}$. If an edge points from v_j to v_i, $e_{i,j} = 1$, otherwise $e_{i,j} = 0$. $\mathcal{W} = \left\{w_{i,j}\right\}_{(e_{i,j}\in\mathcal{E})} \in \mathbb{R}^{M\times M}$ is the edge weight matrix of graph signals, in which $w_{i,j}$ denotes the weight of edge $e_{i,j}$. The larger the value of $w_{i,j}$ is, the closer the connection between node v_j and v_i is.

Many concepts in the graph signal theory are generalized from classical digital signal processing. The gradient and divergence of a graph function are also generalized from classical functions. In graph signals, the gradient $\nabla O_{\mathcal{G}}$ of a graph signal is edge-specific, and the gradient of each edge is defined as the difference between the values of the nodes at both ends of the edge multiplied by the weights. $\nabla O_{\mathcal{G}}$ denotes the partial derivative value of a graph signal on the edge from the node v_j to v_i, given by:

$$\nabla O_{\mathcal{G}} = \left(O_{v_i} - O_{v_j}\right) \times w_{i,j} \tag{13}$$

Relevant elements are defined as follows: (1) the incidence matrix $\boldsymbol{K} = \left(k_{\mathrm{ij}}\right)_{\boldsymbol{M}\times\boldsymbol{Q}}$, (2) the graph adjacency matrix $\boldsymbol{A} = \left(a_{\mathrm{ij}}\right)_{\boldsymbol{M}\times\boldsymbol{Q}}$, (3) the nodes of the graph are $v_1, v_2, \ldots, v_M$, (4) the edges are $e_1, e_2, \ldots, e_Q$, and (5) the weight of each edge is $w_1, w_2, \ldots, w_Q$ [31].

$$k_{ij} = \begin{cases} 1 & \text{if } v_i \text{ is the head of } e_j \\ -1 & \text{if } v_i \text{ is the tail of } e_j \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

$$a_{ij} = k_{ij} \times w_j \tag{15}$$

Each row of the adjacency matrix represents a node, each column an edge, and each element the weight of the edge. The gradient of a graph signal can be defined as:

$$\nabla O_{\mathcal{G}} = A^T O_{\mathcal{G}} \tag{16}$$

The graph Laplacian matrix is defined as:

$$L = AK^T \tag{17}$$

The graph adjacency matrix is not necessarily symmetric, whereas the Laplacian matrix is. The divergence of the gradient of a graph signal is a characteristic of nodes, which is defined as:

$$\nabla^2 O_{\mathcal{G}} = LO_{\mathcal{G}} \tag{18}$$

The Laplace matrix is a semi-positive definite real symmetric matrix whose eigenvalues are all non-negative real numbers. Since the row sum of the Laplace matrix is always zero, its minimum eigenvalue is also always zero and the eigenvalues satisfy $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M = 0$. The eigenvalue of the Laplacian matrix is also the frequency of graph signals so that the graph Laplacian matrix can describe the extent of changes in a graph.

According to the singular value decomposition (SVD) theorem, there are M-order orthogonal matrices $U = (u_1, u_2, \cdots, u_M)$ and $V = (v_1, v_2, \cdots, v_M)$, which satisfy:

$$L = UDV^T \tag{19}$$

The diagonal matrix $\boldsymbol{D} = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_{\mathrm{M}}\}$, $\lambda_i (i = 1, 2, \cdots, M)$ is the singular value of matrix L. All eigenvalues are formed into an eigenvector $f = (\lambda_1, \lambda_2, \cdots, \lambda_M)$, which can describe the characteristics of the Laplacian matrix. Therefore, it can be used to characterize graph signals. The following parameters are defined according to the eigenvectors:

i. The maximum eigenvalue f_1 is one of the critical indicators of changes in signal energy.

$$f_1 = \max(\boldsymbol{f}) = \max\left(\lambda_1, \lambda_2, \cdots, \lambda_M\right) \tag{20}$$

ii. The variance of the eigenvector f_2 reflects the fluctuation of the Laplacian matrix.

$$\begin{aligned} f_2 &= \frac{1}{M}\sum_{i=1}^{M}\left(\lambda_i - \bar{f}\right)^2 \\ \bar{f} &= \frac{1}{M}\sum_{i=1}^{M}\lambda_i \end{aligned} \tag{21}$$

iii. The energy of the eigenvector f_3 represents the energy of elements in the Laplacian matrix.

$$f_3 = \frac{1}{M}\sum_{i=1}^{M}\lambda_i^2 \tag{22}$$

These three parameters with the gradient $\nabla O_{\mathcal{G}}$ and the divergence $\nabla^2 O_{\mathcal{G}}$ are used together to represent the characteristics of graph signals and identify the signals of different voice activities.

The proposed algorithm for voice activity detection is summarized in Algorithm 1. It grows in complexity with higher P and K. It carries $\mathcal{O}(PKN)$ computational cost. The data is easily visualized when K is equal to 2.

Even if K increase, the audio signal can still be mapped to graph. However, considering the computational complexity and real-time performance, P value would be adjusted as needed.

```
Require: Preprocessing: the original speech without noise signal $s(t)$ and divide it into frames $\boldsymbol{s}_n(n=1,2\cdots,N)$
Extract MFCC for each frame $\boldsymbol{y}_n = MFCC\left(\boldsymbol{s}_n\right)$
Compute the probability distribution of each coefficient of MFCC and calculate Wasserstein distance
    $W = WS\begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ y(M) \end{pmatrix}$
Find the index value of the first K largest sum of elements in each column of matrix W $index = \mathrm{Max}(W)$
Select the corresponding coefficients as speech feature $\boldsymbol{z}_n = \boldsymbol{y}_{n,\text{ index}}$.
Conduct equal probability symbolization of speech features $\boldsymbol{z}_n$ and map time series to symbolic series $c_{n,k} = \mathrm{symbol}\left(z_{n,k}\right)$.
Require: Voice activity detection: noisy speech signal $x(t) = s(t) + n(t)$ and divide it into frames $x_n(n=1,2,\cdots,N)$
Extract MFCC for each frame $\boldsymbol{y}_n = MFCC\left(\boldsymbol{x}_n\right)$
Reduce the dimensions $\boldsymbol{z}_n = \boldsymbol{y}_{n,\text{ index}}$
for $n = 1 : N$ do
    $\underline{\boldsymbol{z}}_n = [\boldsymbol{z}_{n-J},\ldots,\boldsymbol{z}_n,\ldots,\boldsymbol{z}_{n+J}]$
    Symbolize $\underline{c}_n = \mathrm{symbol}\left(\underline{\boldsymbol{z}}_n\right)$
    Map to graph signal $O_{\mathcal{G}} = \Gamma\left(\underline{c}_n\right)$
    Compute the Laplacian matrix $L$ and the adjacency matrix $A$ of graph signal
    Calculate the gradient $\nabla O_{\mathcal{G}} = A^T O_{\mathcal{G}}$, the divergence $\nabla^2 O_{\mathcal{G}} = LO_{\mathcal{G}}$, and the parameters $f_1, f_2, f_3$ of Laplacian matrix of graph.
    The above features are normalized and classified using the SVM classifier.
end for
```

Algorithm 1 Voice activity detection

3 Experimental result

3.1 Implementation

To evaluate the performance of the proposed method, we use speech from the Librispeech dataset [32].The sampling rate is 16kHz, and the signals are divided into 512 sampling points in each frame with a 50% overlap between adjacent frames. As shown in the flow chart of Algorithm 1, there are two phases. In the phase of preprocessing, we select 160 min of speech from 200 speakers to calculate the probability distributions of MFCC and reduce dimension, while in the second phase, we use 200 s of speech from 20 speakers for the training of the SVM classifier.

We also select ten types of transients from online free corpora [33], including door knocking, keyboard tapping, doorbells and so on. We obtain 100 s of data for each type of transient, a total of 1000 s as the transient dataset for subsequent experiments.

Speech and transients were normalized to have the same maximum value [19, 20, 34], which was more convenient than normalizing them according to their energy since transients often have low energy due to their short duration.

In order to maintain the nonlinear characteristics of the original data, MFCC was extracted after the signal was divided into frames and the probability distributions of 13 coefficients were calculated. The probability distributions of the second to thirteenth coefficients are approximate to a Gaussian distribution. Figure 2 depicts the probability distributions of the first to fourth coefficients, which proves the validity of the assumption that the several coefficients share the similarity in their probability distributions and variation trends.

The Wasserstein distance was used to calculate the distance between probability distributions to measure their difference to reduce time series dimensions when constructing time series complex networks. The heat map of the distance matrix is shown in Fig. 3, in which the darker the color is, the higher the value is. A higher Wasserstein distance between two probability distributions indicates more significant differences between them. It can be judged from Fig. 3 that the distributions of the first and the third coefficient differ significantly from those of other coefficients. Therefore, these two coefficients were selected for the pattern representation and complex network construction.

Both the selected coefficients were symbolized with equal probability. Figure 4 shows the probability distribution of the first and the third coefficient, where the red line is the threshold of equal probability symbolization.

For each frame of speech signal x_n, we considered the continuous frames $x_{n-J}, x_{n-J+1}, \cdots, x_n, \cdots\cdots, x_{n+J}$ within a specific period, where J was set to be 50. This allowed us to fully leverage the contextual information the preceding and following frames provided to distinguish speech and transients better. Figure 5 shows the time series complex networks of speech without transients, transients, and speech with transients. There are some significant differences between them. The nodes of speech without transients concentrate in the upper part of the graph, while those of transients concentrate in the lower right area. Compared with the other two categories, speech with transients has more node types; its nodes have a wider distribution range and mainly concentrate in the right region.

Complex networks also show differences in node conversion. Most of the nodes of speech without transients converse to their adjacent or adjacent diagonal nodes, while node conversion in transients is more complex. With different SNRs, the types and conversions of nodes of speech with transients do not show distinct differences, but they are different from those of speech without

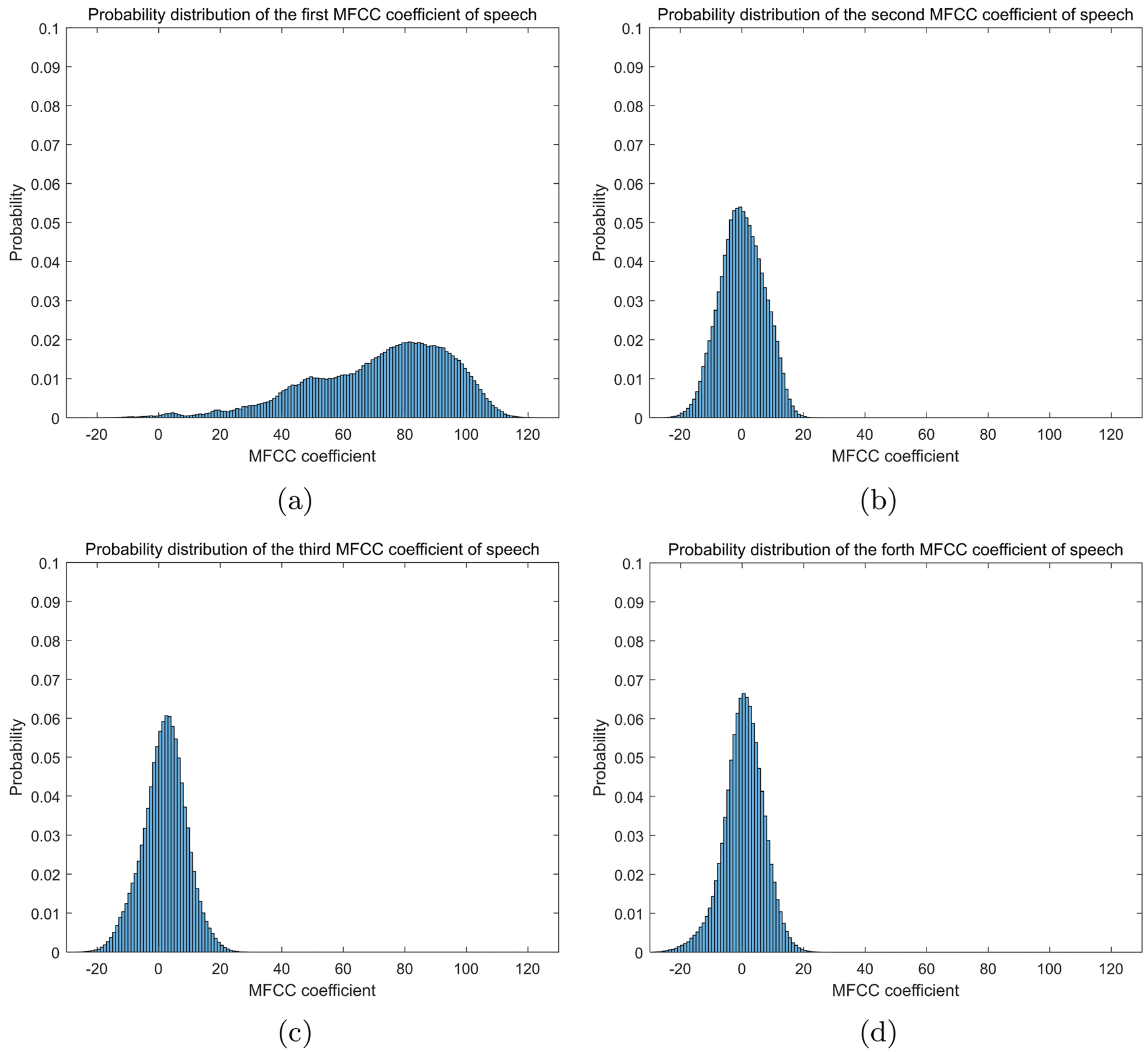
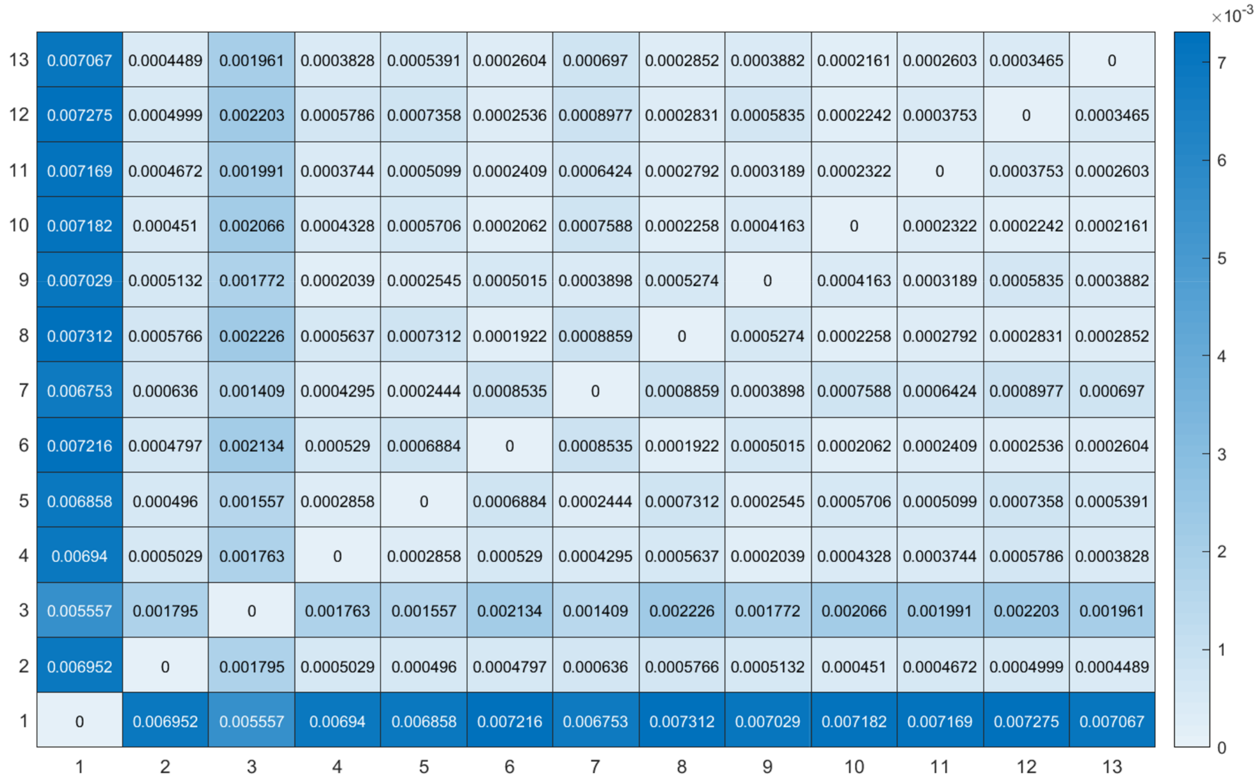


Fig. 2 The probability distributions of the first to fourth coefficients

transients and those of transients. Hence, we can differentiate these three types of voice activities based on time series complex networks.

Since transients usually are more dominant than speech, many algorithms will flag frames containing both speech and transients as ones containing only transients, or identify frames containing transients as speech frames. Compared with traditional algorithms, the graph-based feature improves the clustering effect of audio frames. Figure 6 is a scatter plot of graph signals' gradient and divergence characteristics in which dots of three different colors represent three voice activities. It can be seen that frames containing speech and transients are closer to speech frames, proving that the graph-based feature effectively distinguishes speech frames from transient frames.

Considering that the SVM classifier can distinguish the features of speech and transients better, we select 200 s of speech from 20 speakers and then randomly add transients with different SNRs to it, producing a 1000-s dataset. The dataset is used for feature extractions and the training of the SVM classifier. The ratio of the train data and test data is 8:2.



	1	2	3	4	5	6	7	8	9	10	11	12	13
13	0.007067	0.0004489	0.001961	0.0003828	0.0005391	0.0002604	0.000697	0.0002852	0.0003882	0.0002161	0.0002603	0.0003465	0
12	0.007275	0.0004999	0.002203	0.0005786	0.0007358	0.0002536	0.0008977	0.0002831	0.0005835	0.0002242	0.0003753	0	0.0003465
11	0.007169	0.0004672	0.001991	0.0003744	0.0005099	0.0002409	0.0006424	0.0002792	0.0003189	0.0002322	0	0.0003753	0.0002603
10	0.007182	0.000451	0.002066	0.0004328	0.0005706	0.0002062	0.0007588	0.0002258	0.0004163	0	0.0002322	0.0002242	0.0002161
9	0.007029	0.0005132	0.001772	0.0002039	0.0002545	0.0005015	0.0003898	0.0005274	0	0.0004163	0.0003189	0.0005835	0.0003882
8	0.007312	0.0005766	0.002226	0.0005637	0.0007312	0.0001922	0.0008859	0	0.0005274	0.0002258	0.0002792	0.0002831	0.0002852
7	0.006753	0.000636	0.001409	0.0004295	0.0002444	0.0008535	0	0.0008859	0.0003898	0.0007588	0.0006424	0.0008977	0.000697
6	0.007216	0.0004797	0.002134	0.000529	0.0006884	0	0.0008535	0.0001922	0.0005015	0.0002062	0.0002409	0.0002536	0.0002604
5	0.006858	0.000496	0.001557	0.0002858	0	0.0006884	0.0002444	0.0007312	0.0002545	0.0005706	0.0005099	0.0007358	0.0005391
4	0.00694	0.0005029	0.001763	0	0.0002858	0.000529	0.0004295	0.0005637	0.0002039	0.0004328	0.0003744	0.0005786	0.0003828
3	0.005557	0.001795	0	0.001763	0.001557	0.002134	0.001409	0.002226	0.001772	0.002066	0.001991	0.002203	0.001961
2	0.006952	0	0.001795	0.0005029	0.000496	0.0004797	0.000636	0.0005766	0.0005132	0.000451	0.0004672	0.0004999	0.0004489
1	0	0.006952	0.005557	0.00694	0.006858	0.007216	0.006753	0.007312	0.007029	0.007182	0.007169	0.007275	0.007067

Fig. 3 The heat map of the Wasserstein distance matrix

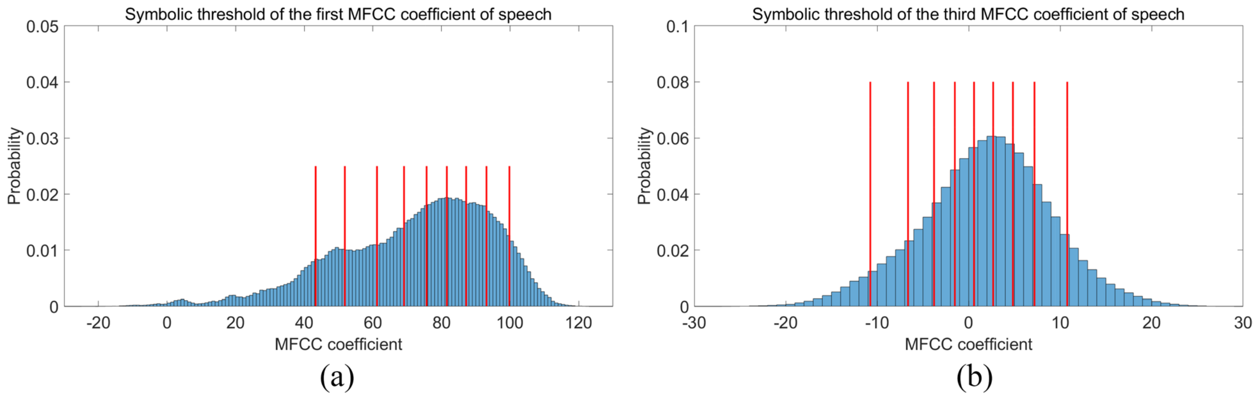


Fig. 4 **a** The probability distribution of the first coefficient. **b** The probability distribution of the third coefficient. The red line represents the threshold for symbolization, divided into ten intervals and represented by 1–10, respectively

3.2 Performance comparison

Ten different speakers are selected from the dataset, each with about 10s speech without transients and 10s speech with random transients, a total of 200 s as test data. We have implemented sets of experiments to test the detection accuracy for speech, transient and speech with transients.

The length of the context window controls the trade-off between correct detection and false alarm rates. As shown in Table 1, when $J = 50$, the correct detection rate of speech is the highest and the false alarm rate is the second lowest. Therefore, the parameter J is set to 50 in the subsequent experiments.

The performance of the proposed method was compared with two voice activity detection algorithms, which are based on formant [13] and kernel [21] respectively and represented as "PND" and "Kernel." The algorithm

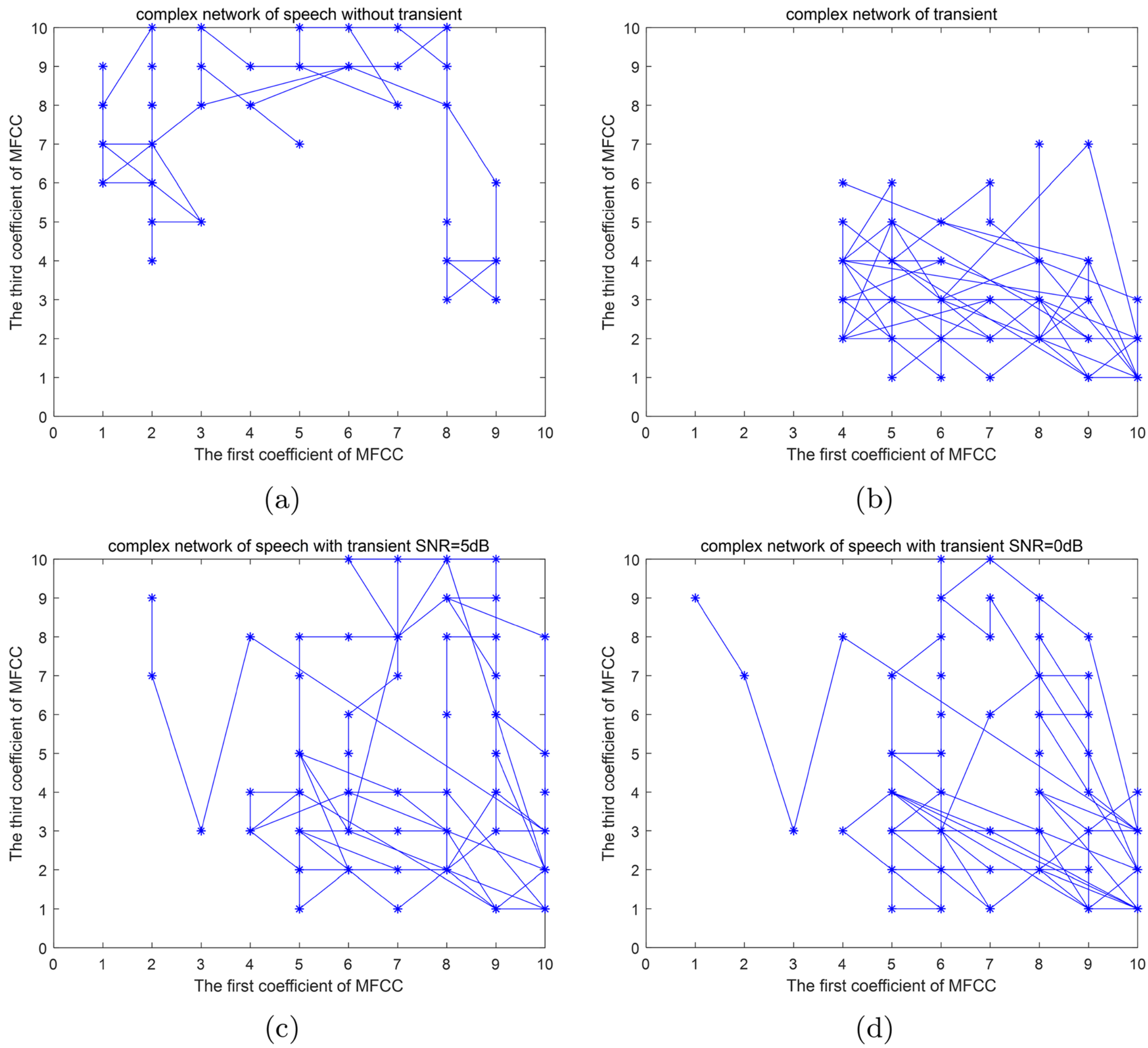


Fig. 5 The complex network of **a** speech without transient; **b** transient; **c** speech with transient, SNR=5dB; and **d** speech with transient, SNR=0dB. Take the symbolized value of the first coefficient as *X*-axis and that of the third as *Y*-axis

proposed in this paper is represented by "Proposed." Figure 7 shows the waveform of audio signals contaminated with door-knocking and keyboard-tapping transients. Figures 8 and 9 present the qualitative assessment of the proposed VAD, compared with detectors "PND" and "Kernel." Table 2 shows the accuracy of voice activity detection of three algorithms in two transient environments.

In voice activity detection, the silent frames less than 100 ms between speech frames are considered as parts of speech, and only speech with a duration of more than 250 ms is considered as speech. "PND" detects each frame independently, and disregards the contextual information of the preceding and following frames, though it regards five preceding and following frames of the detected frame as speech frames for the presence of consonants. Limited by the computational complexity, "Kernel" can only consider 15 preceding and following frames of the detected frame. However, these two algorithms are sensitive to the silent frames with a brief duration between speech frames and flag them as non-speech frames which are outside the scope of our interest in this study. In addition, they will recognize transients with high energy as speech frames, and thus fail to identify frames containing only transients.

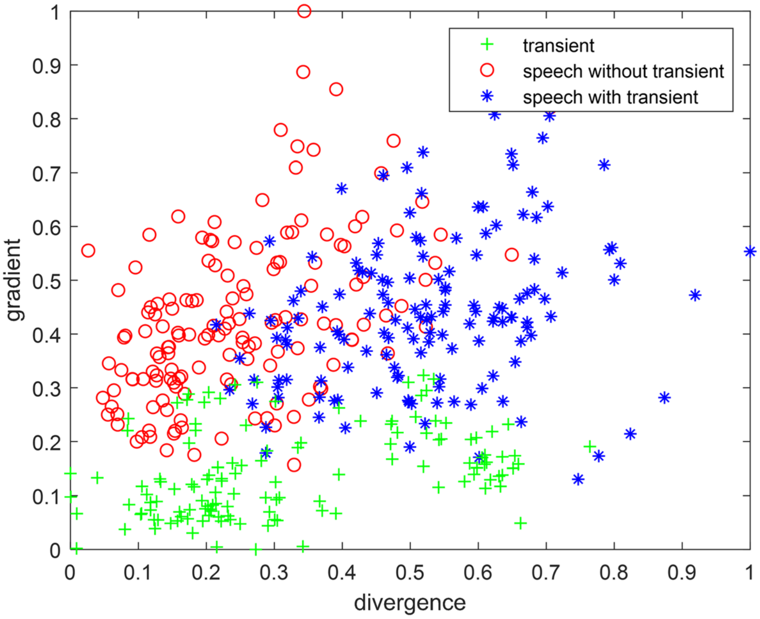


Fig. 6 The scatter plot of the gradient and divergence characteristics of graph signals. Red dots, blue dots, and green dots represent frames of speech without transients, speech with transients, and transients, respectively

Table 1 The accuracy of the proposed algorithm in different context window

J	20	30	40	50	60
Speech	71.30%	80.47%	83.52%	**87.73%**	85.46%
Transient	66.42%	75.58%	**82.91%**	81.54%	78.91%
Speech with transient	71.30%	78.03%	80.06%	88.25%	**89.55%**

Note: Entries in boldface are the best results

The results demonstrate that our proposed detector has a lower false alarm rate than the other two detectors, especially for transient frames after the 8th second in the keyboard-tapping transient environment. Our proposed method can include more context to distinguish silent frames of interest from the signal more accurately, instead of silent frames with a brief duration between speech frames. However, it will identify some frames containing only transients as speech frames. Mistakes often occur when speech frames are adjacent to those containing only transients. Nevertheless, these two kinds

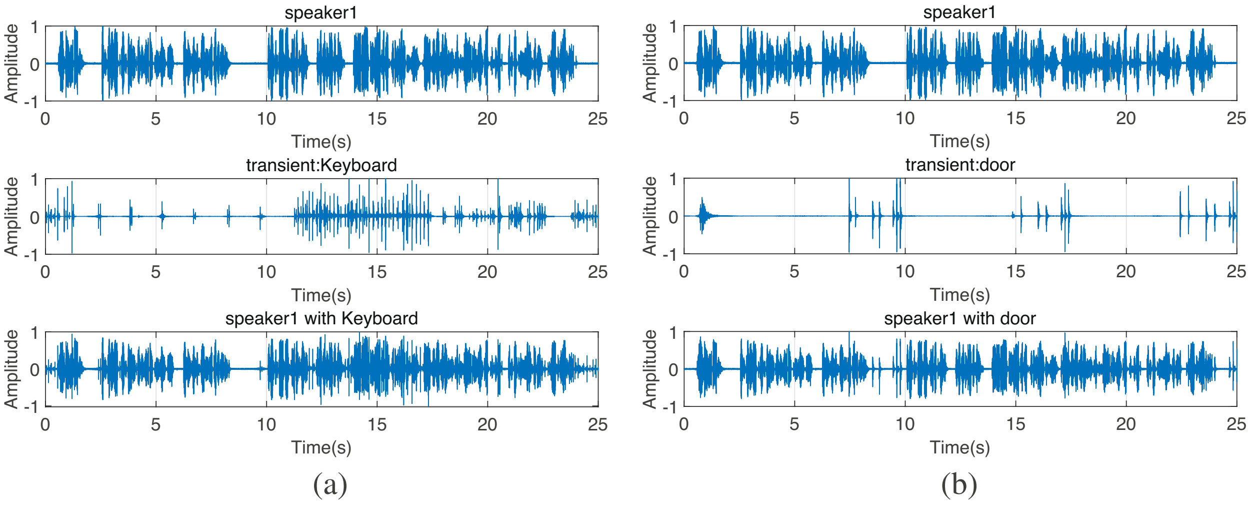


Fig. 7 The waveform of speech signal contaminated with door knocking and keyboard tapping

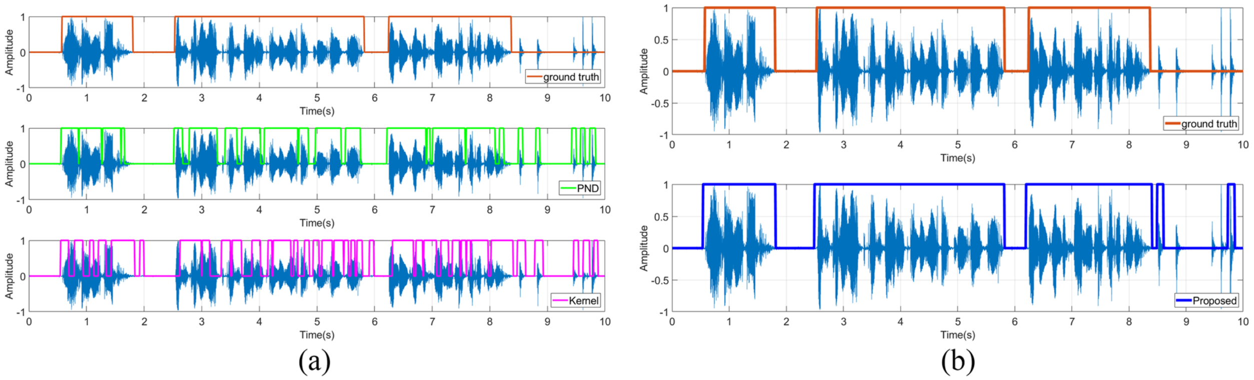


Fig. 8 The performance of the proposed detector, with a transient of keyboard tapping. The light blue line is the input audio signal with transient, and the red line is the ground-truth representing the frames containing speech. The green, magenta and dark blue line are the detection result of "PND," "Kernel," and "Proposed," respectively

Table 2 The accuracy of voice activity detection of three algorithms in two transient environments

	Knocking	Keyboard taps
PND	84.20%	87.34%
Kernel	74.02%	70.45%
Proposed	**95.63%**	**92.71%**

Note: Entries in boldface are the best results

of frames can be better distinguished based on the proposed method generally, which is based on graphs, the new method for VAD.

Table 3 displays the detection accuracy and the real-time performance of these three algorithms in different scenarios. In one scenario, the material was divided into two categories: speech frames and non-speech frames. The category of speech frames consisted of speech with and without transients. Transient and silent frames were classified into non-speech frames. In the other scenario, the material was divided into three types, where speech frames were subdivided into speech with transients and speech without transients.

"PND" utilizes formant frequency information to detect the presence or absence of speech, and the formants may also be detected when the signal includes additive noise. For this reason, though transients without formant structures can be identified, it is difficult to find an appropriate threshold to distinguish speech with transients from speech without transients. Furthermore, some transients also contain similar formant structures, leading to a further deterioration of the performance. "Kernel" performs better than "PND" and when the material is divided into two categories, it slightly outperforms our proposed algorithm. However, "Kernel" is of high complexity, and it requires a much longer running time than the proposed one.

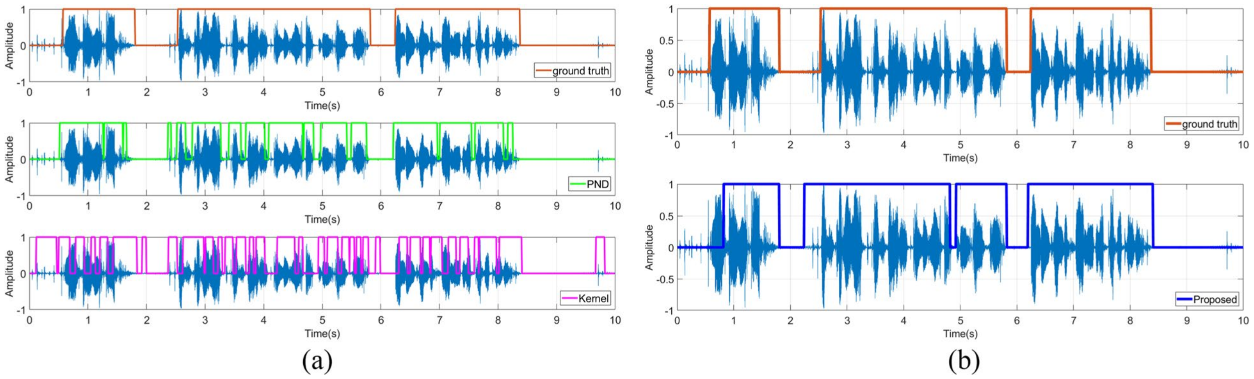


Fig. 9 The performance of the proposed detector, with a transient of door knocking. The light blue line is the input audio signal with transient, and the red line is the ground-truth representing the frames containing speech. The green, magenta and dark blue line are, respectively, the detection result of "PND," "Kernel," and "Proposed"

Table 3 The average accuracy and processing time per frames of three algorithms in different scenarios

Algorithm		Two categories	Three categories	Processing time per frame
PND	Speech	83.77%	82.42%	0.0012 s
	Transient	61.98%	61.98%	
	Speech with transient	**91.59%**	27.06%	
Kernel	Speech	**88.18%**	66.95%	0.2216 s
	Transient	40.56%	40.56%	
	Speech with transient	86.87%	**72.98%**	
Proposed	Speech	87.73%	**88.75%**	0.0093 s
	Transient	**81.54%**	**79.35%**	
	Speech with transient	88.25%	71.73%	

Note: Entries in boldface are the best results

When severely corrupted by transients, the proposed algorithm is superior to “Kernel” and “PND.” With the proposed method, the characteristics of transients can be extracted. Transients can be successfully distinguished without being regarded as speech frames, even when they are more dominant than speech. However, due to the slight similarity in the node distribution and edge conversion of graph signals mapped from speech with and without transients, the method performs less satisfactorily in deciding whether the speech contains transients.

Experiments show that the proposed algorithm averages 0.0093 s processing time per frame. The algorithm complexity is low and allows real-time processing.

4 Conclusion and discussion

Voice activity detection is particularly challenging in the presence of transients, which are usually more dominant than speech due to their short duration, high amplitude, and rapid spectrum changes over time. This paper discusses voice activity detection in the presence of transients and proposes to apply the complex network analysis theory to investigate the differences between speech and transients in nonlinear dynamic characteristics. Moreover, a new method is introduced to analyze complex networks in the form of graph signal processing. With a low complexity, the algorithm is adequate for real-time detection.

The speech signal is mapped to the graph signal using the symbolic representation of time series to preserve the nonlinear dynamic characteristics of speech. A series of symbolic patterns are extracted as the nodes of time series complex networks. Then the association between the symbolic patterns is used to determine the weight and direction of the network edges. The time series symbolization allows partial noise elimination and fast information extraction. Therefore, the structure and characteristics of complex networks are very sensitive, enabling them to identify and analyze time series quickly.

With the proposed method, transients can be successfully distinguished from speech frames instead of being regarded as the latter, even when transients dominate in the frame. Considering that the method allows us to cluster frames according to the presence of speech rather than transients, it is advisable for voice activity detection. The major limitation of this method is its failure to distinguish the speech frames containing transients from speech frames. Therefore, a natural progression of this work is to find a better feature to build a complex network to improve the accuracy of voice activity detection.

Abbreviations

FA	Factor analysis
KL	Kullback-Leibler
JS	Jensen-Shannon
SVM	Support vector machine
VAD	Voice activity detection
GSP	Graph signal processing
PCA	Principal component analysis
CCA	Canonical correlation analysis
SNR	Signal-to-noise ratio
MFCC	Mel-frequency cepstral coefficients

Acknowledgements
Not applicable.

Authors’ contributions
Xiaoyuan Guo conducted the research and performed the experiments. Chunxian Gao and Hui Liu guided and supervised the project. The authors read and approved the final manuscript.

Authors’ information
Not applicable.

Funding
Not applicable.

Availability of data and materials
Speech dataset is from Librispeech database [32] and transient is from online free corpora [33].

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 23 June 2022 Accepted: 25 March 2023
Published online: 20 April 2023

References

1. B. Schuller, M. Wöllmer, T. Moosmayr, Recognition of Noisy Speech: A Comparative Survey of Robust Model Architecture and Feature Enhancement. J Audio Speech Music Proc. **2009**, 942617 (2009)
2. K. Veena, D. Mathew, in *2015 International Conference on Power, Instrumentation, Control and Computing (PICC)*. Speaker identification and verification of noisy speech using multitaper mfcc and gaussian mixture models (IEEE 2015), pp. 1-4
3. N. Cho, E.-K. Kim, Enhanced voice activity detection using acoustic event detection and classification. IEEE Trans. Consum. Electron. **57**(1), 196–202 (2011)
4. J.-H. Chang, N.S. Kim, S.K. Mitra, Voice activity detection based on multiple statistical models. IEEE Trans. Sig. Process. **54**(6), 1965–1976 (2006)
5. J. Sohn, N.S. Kim, W. Sung, A statistical model-based voice activity detection. IEEE Sig. Process. Lett. **6**(1), 1–3 (1999)
6. J. Ramırez, J.C. Segura, C. Benıtez, A. De La Torre, A. Rubio, Efficient voice activity detection algorithms using long-term speech information. Speech Commun. **42**(3–4), 271–287 (2004)
7. G. Hinton et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
8. X.-L. Zhang, J. Wu, Deep belief networks based voice activity detection. IEEE Trans. Audio Speech Lang. Process. **21**(4), 697–710 (2013)
9. S. Thomas, S. Ganapathy, G. Saon, H. Soltau, Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions, 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2519-2523 (2014). https://doi.org/10.1109/ICASSP.2014.6854054
10. R. Tahmasbi, S. Rezaei, A soft voice activity detection using GARCH filter and variance gamma distribution. IEEE Trans. Audio, Speech, Lang. Process. **15**(4), 1129-1134 (2007)
11. A. Ivry, B. Berdugo, I. Cohen, in *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2. Voice Activity Detection for Transient Noisy Environment Based on Diffusion Nets (2019), pp. 254-264. https://doi.org/10.1109/JSTSP.2019.2909472
12. Kobayashi, H., Shimamura, T.: in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Proceedings (Cat. No. 00CH37100), vol. 3. A weighted autocorrelation method for pitch extraction of noisy speech (IEEE 2000), pp. 1307-1310
13. I.-C. Yoo, H. Lim, D. Yook, Formant-based robust voice activity detection. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(12), 2238–2245 (2015)
14. T. Kristjansson, S. Deligne, P. Olsen, Voicing features for robust speech detection. Entropy. **2**(2.5), 3 (2005)
15. S.O. Sadjadi, J.H. Hansen, Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Sig. Process. Lett. **20**(3), 197–200 (2013)
16. Y. Ma, A. Nishihara, Efficient voice activity detection algorithm using long-term spectral flatness measure. EURASIP J. Audio Speech Music Process. **2013**(1), 1–18 (2013)
17. E. Scheirer, M. Slaney, in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. Construction and evaluation of a robust multifeature speech/music discriminator (IEEE, 1997), pp. 1331-1334
18. D. Vlaj, Z. Kačič, M. Kos, Voice activity detection algorithm using nonlinear spectral weights, hangover and hangbefore criteria. Comput. Electr. Eng. **38**(6), 1820–1836 (2012)
19. R. Talmon, I. Cohen, S. Gannot, Single-channel transient interference suppression with diffusion maps. IEEE Trans. Audio Speech Lang. Process. **21**(1), 132–144 (2012)
20. R. Talmon, I. Cohen, S. Gannot, R.R. Coifman, Supervised graph-based processing for sequential transient interference suppression. IEEE Trans. Audio Speech Lang. Process. **20**(9), 2528–2538 (2012)
21. D. Dov, R. Talmon, I. Cohen, Kernel method for voice activity detection in the presence of transients. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(12), 2313–2326 (2016)
22. M. Petrovic, R. Liegeois, T.A. Bolton, D. Van De Ville, Community-aware graph signal processing: Modularity defines new ways of processing graph signals. IEEE Sig. Process. Mag. **37**(6), 150–159 (2020)
23. E. Pavez, A. Ortega, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Generalized laplacian precision matrix estimation for graph signal processing (IEEE, 2016), pp. 6350-6354
24. A. Hiruma, K. Yatabe, Y. Oikawa, in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Separating stereo audio mixture having no phase difference by convex clustering and disjointness map (IEEE, 2018), pp. 266-270
25. X. Yan, Z. Yang, T. Wang, H. Guo, An iterative graph spectral subtraction method for speech enhancement. Speech Commun. **123**, 35–42 (2020)
26. X. Li, D. Yang, X. Liu, X.M. Wu, Bridging time series dynamics and complex network theory with application to electrocardiogram analysis. IEEE Circ. Syst. Mag. **12**(4), 33–46 (2012)
27. H. Trang, T.H. Loc, H.B.H. Nam, in *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*. Proposed combination of pca and mfcc feature extraction in speech recognition system (IEEE, 2014), pp. 697-702
28. D. R. Hardoon, S. Szedmak, J. Shawe-Taylor, in *Neural Computation*, vol. 16, no. 12. Canonical Correlation Analysis: An Overview with Application to Learning Methods (2004), pp. 2639-2664. https://doi.org/10.1162/0899766042321814
29. X. Peipei, Z. Li, L. Fanzhang, Learning similarity with cosine similarity ensemble[J]. Inf. Sci. **307**(C): 39-52 (2015)
30. V.M. Panaretos, Y. Zemel, Statistical aspects of wasserstein distances. (2018). arXiv preprint arXiv:1806.05500
31. M. Mesbahi, M. Egerstedt, in *Graph Theoretic Methods in Multiagent Networks*. Graph theoretic methods in multiagent networks (Princeton University Press, 2010)
32. V. Panayotov, G. Chen, D. Povey, S. Khudanpur, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Librispeech: an asr corpus based on public domain audio books (IEEE, 2015), pp. 5206-5210. https://ieeexplore.ieee.org/document/7178964
33. F. Font, G. Roma, X. Serra, Freesound technical demo[C]//Proceedings of the 21st ACM international conference on Multimedia. 411-412 (2013). Transients source: http://www.freesound.org/
34. S. Mousazadeh, I. Cohen, Voice activity detection in presence of transient noise using spectral clustering. IEEE Trans. Audio Speech Lang. Process. **21**(6), 1261–1271 (2013)

Publisher's Note

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